OMICS in Biomedical Sciences

Lecture notes

Biomedische Wetenschappen

Studiejaar 2020 – 2021

Edition 9.01

OMICS in Biomedical Sciences

Lecture notes

Bio-medische Wetenschappen

Studiejaar 2020 – 2021

Edition 9.01

OMICS in Biomedical Sciences

Lecture notes
Bio-medische Wetenschappen
Studiejaar 2020 – 2021
Edition 9.01

Coördinatoren:
Antoine van Kampen
(1) Bioinformatics Laboratory
Academic Medical Center
(2) Biosystems Data Analysis group
Swammerdam Institute for Life Sciences
Contact: a.h.vankampen@amsterdamumc.nl

Martijs Jonker
RNA Biology & Applied Bioinformatics
Swammerdam Institute for Life Sciences
Contact: M.J.Jonker@uva.nl

University of Amsterdam


STUDY IS A PLEASURE RATHER THAN A TASK

Contents

Contributors

Dr. Rob Dekker	RNA Biology and Applied Bioinformatics, Swammerdam Institute for Life Sciences, University of Amsterdam
Prof. dr. Raoul Hennekam, MD	Pediatrics and Translational Genetics, department of Pediatrics, AMC
Dr. Huub Hoefsloot	Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam.
Dr. Martijs Jonker	RNA Biology and Applied Bioinformatics, Swammerdam Institute for Life Sciences, University of Amsterdam
Prof. dr. Antoine van Kampen	Medical Bioinformatics, Bioinformatics Laboratory, AMC.
Dr. Johan Westerhuis	Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam
Dr. F. M. Vaz	Laboratory Genetic Metabolic Disease (F0-224), Departments of Clinical Chemistry and Pediatrics, AMC
Prof. dr. Stanley Brul	Molecular Biology & Microbial Food Safety, Swammerdam Institute for Life Sciences, University of Amsterdam
Dr. Dave Speijer	Proteomics, Medical Biochemistry, AMC.

Acknowledgements

Special thanks to dr. Aldo Jongejan, dr. ir. Perry Moerland, dr. Mia Pras-Raves, Barbera van Schaik, Angela Luyf, and prof. dr. Arthur Verhoeven for proofreading this syllabus.

Used resources and acknowledgements

A large fraction of the text in these lecture notes was copied from scientific literature, books and internet resources. Text from various resources has been mixed, and (sometimes) modified by removing irrelevant parts and/or adjusting the text. Text from the lecturers has been added. Consequently, instead of precisely indicating the used resources in the text itself, we have listed these resources below. Figures that were copied from internet resources include the URL in their caption. Although we tried to refer to original sources as much as possible, the text also contains material from internet sites (e.g., Wikipedia) and internet documents (e.g., powerpoint presentations) that don't have references in the syllabus (sometimes because the original source could not be found anymore). However, I fully acknowledge all the authors, institutes and sources of this material. These lecture notes comply with UvA/Stichting UvO regulations (https://www.stichting-uvo.nl/) with respect to using material from scientific publications and books.

The following resources were used in this (or previous) editions of the syllabus:

Introduction to OMICS in Biomedical Sciences

- Chapter is partially based on Van Kampen, Moerland (2015) Taking Bioinformatics to Systems Medicine. In: Systems Medicine: Methods and Protocols (Eds, Olaf Wolkenhauer, Ulf Schmitz), Springer. In preparation.
- Figure copied from Lubrano-Berthelier (2003) Diabetes, 52, 2996
- Figure copied from Martínez, et al (2012) Proceedings of the National Academy of Sciences of the United States of America 109, 2672.
- Part of the text about sequence alignment is copied from Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press, Cambridge.
- Table copied from Hood and Affray (2013) Genome Medicine, 5(12), 110

Sanger sequencing

- Text and some figures copied from Strachan and Read (2011) Human Molecular Genetics, 4th edition, Garland Science, NY.

Next Generation Sequencing

- Text copied from Mardis ER: The impact of next-generation sequencing technology on genetics. Trends in genetics: TIG 2008, 24:133–41.
- Some text and figures were copied from http://seqanswers.com/forums/showthread.php?t=10.
- Some text and figures were copied from Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature reviews. Genetics 17, 333-351.
- Some text was copied Robasky, K., Lewis, N.E., and Church, G.M. (2014). The role of replicates for error mitigation in next-generation sequencing. Nature reviews. Genetics 15, 56-62
- Figure copied from Kou et al (2016). Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. PloS one 11, e0146638
- Text copied from Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature reviews. Genetics 14, 618-630.
- Figure copied from Navin, N.E. (2014). Cancer genomics: one cell at a time. Genome biology 15, 452.

DNA capture methods

- Some text copied from Strachan and Read (2011) Human Molecular Genetics, 4th edition, Garland Science, NY.

Exome sequencing

- Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. a, & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nature reviews. Genetics, 12(11), 745–55.
- Manolio TA, et al (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747

Types of DNA mutations

- http://ghr.nlm.nih.gov/handbook/mutationsanddisorders

Aanalyzing Transcriptomics, Proteomics and Metabolomics data

- Draghici S. Data Analysis Tools for DNA Microarrays, ISBN 1-58488-3.5-4. 2003. Boca Raton: Chapman & Hall/CRC.
- Gibson G, Spencer VM. A Primer of Genome Science, third edition, ISBN 978-0-87893-236-8. 2009. Sunderland, Massachusetts: Sinauer Associates, Inc Publishers.
- Pevsner J. Bioinformatics and Functional Genomics, ISBN 0-471-21004-8. 2003. Hoboken, New Jersey: John Wiley and Sons.
- Baxevanis AD, Francis Ouellette BF. Bioinformatics, ISBN 0-471-47878-4. 2005. Hoboken, New Jersey: John Wiley and Sons

Transcriptome and Transcriptomics

- Draghici S. Data Analysis Tools for DNA Microarrays, ISBN 1-58488-3.5-4. 2003. Boca Raton: Chapman & Hall/CRC.
- Gibson G, Spencer VM. A Primer of Genome Science, third edition, ISBN 978-0-87893-236-8. 2009. Sunderland, Massachusetts: Sinauer Associates, Inc Publishers.
- Cristianini N, Hahn MW. Introduction to Computational Genomics: A Case Studies Approach, ISBN 9780521856034. 2006. Oxford University Press
- Brown TA. Genomes 3 ISBN 9780815341383. 2006. Garland Science: New York.

Introduction to Genome Wide Association studies

- Parts of this chapter were inspired by: Gibson G, Muse SV. A Primer of Genome Science, Third Edition, Chapter 3, ISBN: 978-0-87893-236-8. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts USA.

Metabolomics

- Adapted from Vaz FM, Pras-Raves M, Bootsma AH, van Kampen AH. Principles and practice of lipidomics. J Inherit Metab Dis. 2015 Jan;38(1):41-52.

Proteomics

- Aebersold R., Mann M. (2016) Mass-spectrometric exploration of proteome structure and function. Nature 537:347-55.
- Van Oudenhove L., Devreese B. (2013) A review on recent developments in mass spectrometry instrumentation and quantitative tools advancing bacterial proteomics. Appl Microbiol Biotechnol 97:4749-62

Metabolomics pre-processing

- Smith CA, Want EJ, Maille GO, Abagyan R, Siuzdak G: XCMS : Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment , Matching , and Identification. Trial 2006, 78:779–787.
- The section about noise and noise filtering was partially copied from
    - the word document found at http://webpage.pace.edu/dnabirahni/cnld_instrumentation.htm
    - Figure copied from Agilent technical report http://www.chem.agilent.com/Library/technicaloverviews/Public/5990-8341EN.pdf
    - Figure copied from Thomas Coor (1968) J. Chem. Educ., 1968, 45 (7), p A533.
    - Figure copied from Savitzky and Golay (1964) Anal. Chem., 36, 1627.
- Section about peak identification:
    - Figure copied from Danielsson et al (2002) Analytica Chimica Acta, 454:167–184.
    - Figure copied from Katajamaa and Oresic (2007) Journal of Chromatography A, 1158, 318.
- Section about peak matching:
    - Figure copied from Christensen et al (2005) J Chromatogr A, 1062(1), 113

Introduction Systems Biology

- Text and figures copied from Voit EO: Biological Systems. In A first course in systems biology. First Edition. New York: Garland Science, Taylor & Francis Group, LLC; 2013:1–18.
- Figure copied from Keelan (2011) Journal of Reproductive Immunology, 88(2), 176
- Figure copied from Luo (2012) Genes & Development, 26(5), 417

- Parts of text copied from Bruggeman et al (2006) Trends Microbiol., 15(1), 45.
- Figure copied from de Jong et al (2012) PloS one, 7(6), e39498.
- Figure copied from Peskov (2012) The FEBS journal, 279, 3374.
- Figure copied from Van Riel (2006) Briefings in bioinformatics, 7,364.
- Figure copied from Stelling (2004) Current opinion in microbiology, 7, 513
- Photo and text copied from http://med.stanford.edu/ism/2012/march/snyder.html
- Figures copied from De Graaf AA (2009) PLoS Comput Biol. 5(11), e1000554; Hall KD (2006) American journal of physiology, 291, E23–37; Sedaghat (2002) American journal of physiology. Endocrinology and metabolism, 283, E1084

Mathematical Formulation of Elementary Reactions

- Text and figures copied from Voit EO: Biological Systems. In A first course in systems biology. First Edition. New York: Garland Science, Taylor & Francis Group, LLC; 2013:1–18.

Information management: Biological databases

- Part of the text has been copied from Jin Xiong (2006) Essential Bioinformatics. Cambridge University Press, Cambridge.
- Figure copied from Fleming (2011) Nature Chemical Biology, 7, 9-17
- Figure copied from von Mering (2005) Nucleic Acids Research, 33, D433.
- Figure copied from Callaway (2011) Nature, 475, 435-437.
- Figures copied from Segal (2004) Nat Genet. 36(10), 1090-1098

Preliminary

- This syllabus contains colored text and figures. Therefore, it is recommended to also use the electronic version of this syllabus.

- In these lecture notes words that are shown with this typesetting are listed in the Glossary at the end of this syllabus.

- Topics not covered in this syllabus will be provided by the lecturers separately.

- Instructions for the 'laptop colleges' are distributed through Canvas

- All complementary teaching material is distributed through Canvas

- This is **Version 9.01** of the syllabus for OMICS in Biomedical Sciences. All comments and suggestions for improvement are welcome (a.h.vankampen@amsterdamumc.nl). If you have comments, please be specific.

- For information about OMICS in Biomedical Sciences contact the course coordinators: Antoine van Kampen (a.h.vankampen@amsterdamumc.nl) or Martijs Jonker (M.J.Jonker@uva.nl)

1 Introduction to OMICS in Biomedical Sciences

Lecturer: Prof. dr. Antoine van Kampen (AMC)

After reading this chapter you should understand the concepts of

- High throughput experimentation, OMICs, OMEs, and the different OMICS disciplines, genome-wide
- Bioinformatics, systems biology, systems medicine, information management, and e-Science
- Personalized medicine

Contents

1.1 Recap biology

Don't hesitate to refresh your minds! OMICS in Biomedical Sciences assumes that you have good knowledge of molecular biology and biochemistry:

- Sadava D, Hillis DM, Heller HC, Berenbaum MR: Life. The science of biology. Sunderland, Mass.: Sinauer Associates, Inc; W.H. Freeman and Company; 2013, Ninth Edition.
- Molecular Biology of the Cell, Alberts et al., Taylor & Francis Inc (sixth edition).

1.2 From traditional to data-driven biology

Biomedical research is basic, applied, or translation research conducted to increase our understanding of (molecular) processes in health and disease. Biomedical research will increase our fundamental understanding of biology, and will help developing improved diagnostics tests, and to develop (personalized) treatment. Molecular biology is an essential ingredient in biomedical research.

The last decades, molecular biology has made enormous advances. We no longer only study small parts of our DNA, single genes, proteins or metabolites. Instead, new experimental technologies (**OMICS technologies**) made it possible to study all (or at least many) genes, proteins, and metabolites simultaneously for many samples. Therefore, these experimental technologies are referred to as '**high-throughput**' technologies. Since these OMICS experiments produce large amounts of data we need **bioinformatics** for the management and (statistical) analysis of the data.

In **OMICS in Biomedical Sciences** you will get acquainted with many new and state-of-the-art wet-lab and computational technologies that currently are widely used in nowadays biomedical research and in the clinic. You will also see many applications of OMICS.

1.3 High-throughput experimental technologies (OMCIS)

The development of **high-throughput experimental technologies** allows wet-lab researchers to perform large scale measurements (Box 1). These measurements include, for example,

(i) whole genome sequences (and gene variants) using next-generation sequencing technologies (NGS) (Metzker 2010);
(ii) measuring gene expression with DNA microarrays (Brown and Botstein 1999);
(iii) measuring RNA (e.g., gene expression) with NGS;
(iv) identifying and quantifying proteins and metabolites with NMR or LC-MS (Lindon and Nicholson 2008);
(v) measuring epigenetic changes such as methylation with bisulfide sequencing (Mensaert et al 2014).

These, 'omics' technologies, are capable of simultaneously measuring the many molecular building blocks that determine our (patho)physiology. Measuring complete DNA/RNA sequences, or all genes, proteins, or metabolites are sometimes referred to as **genome-wide measurements.** Omics has significantly advanced our fundamental understanding of the molecular biology of health and disease but has also contributed to new (commercial) diagnostic and prognostic tests (van 't Veer et al 2002; Zanotti et al 2014) and the selection and development of (personalized) treatment (Paik et al 2004).

OMICS technologies are used in many applications. Examples are:

- Identification of gene mutations that cause (rare) disorders.
- Elucidation of biological networks.
- Identification of biomarkers (e.g., genes, proteins, metabolites) that are predictive for disease outcome.
- Study of immune responses.
- Evolution of species and genes (phylogenetics).
- Identification of genes and/or regulatory elements in genomes of organisms.
- Understanding of the human microbiome (the genome of all our microbes).

Figure 1.1 shows several OMICS technologies at the levels of the central dogma.

1.4 Types of OMICS measurements

There are four major types of high-throughput measurements that are commonly performed: **genomic** measurements (e.g., the large-scale genotyping of single nucleotide polymorphisms; SNPs), **transcriptomic** measurements (i.e., the measurement of the expression of all genes in a cell or tissue), **proteomic** measurements (i.e., the identification and quantification of all proteins present in a cell or tissue type), and **metabolomic** measurements (i.e., the identification and quantification of all metabolites present in a cell or tissue type). Each of these four is distinct and offers a different perspective on the processes underlying disease initiation and progression as well as on ways of predicting, preventing, or treating disease. However, keep in mind that there exist other omics technologies for measuring, for example, DNA methylation (e.g., using bisulfide sequencing or methylation arrays), binding of transcription factors to DNA (e.g., ChIPseq).

1.4.1 Genomics

Genomics comprises the measurement and analysis of complete or partial (e.g., only the genes) DNA sequences. Genomics makes use of **Next Generation Sequencing (NGS) technologies** that determine the nucleotide order of the DNA sequence. Once this sequence is available it can be analysed for its genes, regulatory elements, SNPs, or other features. Genomic SNP genotyping measures a person's genotypes. The genotyping technology is quite accurate, but the SNPs themselves offer only limited information. These SNPs tend to be quite

common (with typically at least 5% of the population having at least one copy of the less frequent allele), and not strictly causal for a disease. Rather, SNPs can act in unison with other SNPs and with environmental variables to increase or decrease a person's risk of a disease. This makes identifying important SNPs difficult. Genome Wide Association Studies (**GWAS**) is one approach to associate SNPs to a disease. As an alternative one may use whole **exome sequencing** (WES).

1.4.2 Transcriptomics

Transcriptomic measurements are the oldest and most established of the high-throughput methodologies. One approach is to use commercially produced "oligonucleotide arrays" (i.e., **DNA microarrays** or **Affymetrix gene chips**) which have hundreds of thousands of small probes for each gene. RNA that has been extracted from cells is then labelled and hybridized to the chip, and the expression level of ~25,000 different mRNAs can be assessed simultaneously. Gene expression levels influence phenotypes more directly than SNPs. While transcriptomic measures are not as useful for pre-disease prediction (because a person's gene expression level may be quite different before the onset of disease), they are very well-suited for either early identification of a disease (i.e., finding people who have gene expression levels characteristic of a disease but who have not yet manifested other symptoms) or classifying patients with a disease into subgroups (by identifying gene expression levels that are associated with either better or worse outcomes or with higher or lower values of some disease phenotype). As an alternative to microarrays research nowadays mostly use NGS to determine gene expression levels. In this application, NGS is used to sequence the mRNA molecules (**RNAseq**). By counting the number of mRNA molecules (from the sequences obtained with NGS, one obtains a measure for the gene expression).

1.4.3 Proteomics

Proteomics is used to identify and quantify proteins in a samples. Proteins in a sample are typically separated using chromatography, 2 dimensional protein gels (which separate proteins based on charge and then size), or 1 dimensional protein gels (which separate based on size alone), and digested, typically with trypsin (which cuts proteins after each arginine and lysine), and then run through **mass spectroscopy**. The mass spectrometer identifies the size of each of the peptides, and the proteins can be identified by comparing the size of the peptides created with the theoretical digests of all know proteins in a database. Like transcriptomic measures, though, proteomic measures are excellent for early identification of disease or classifying patients into subgroups.

1.4.4 Metabolomics

Metabolomics, the high-throughput measurement of the metabolites present in a cell or tissue. Samples are typically separated by **chromatography** prior to **NMR** or **mass spectroscopy** to identify and quantify metabolites. Metabolomics is newer and less frequently used than the other technologies. Measurements of metabolites are dynamic as are gene

expression levels and proteins, and so are best suited for either early disease detection or disease subclass identification.

1.5 Bioinformatics

The traditional role of bioinformatics in biomedical research involves basic and applied research to augment our understanding of (molecular) processes in health and disease. Large part of bioinformatics is devoted to the **management and (statistical) analysis of OMICS data**. See also Box 2.

A **bioinformatician** is a biomedical researcher who uses computational approaches (opposed to wet-lab techniques) as major tools to study biology. An increasing number of wet-lab biomedical researchers get involved with bioinformatics that is required for the analysis and interpretation of their data. However, bioinformaticians may also come from other disciplines such as informatics, statistics, or mathematics. They all bring (and combine) unique expertise to solve complex biomedical questions.

Overall, bioinformatics is involved with the analysis of data obtained from wet-lab OMICS experiments and/or data retrieved from **public biological databases** for, e.g., the generation of new hypotheses (Figure 1.3). In addition, bioinformatics can help to **design wet-lab experiments** to ensure that data can be analysed in such a way that it contributes to answering the biological question. **Information management** (or research data management) is part of bioinformatics and deals with the organization of data in (public) databases, development of standards to describe, store, and exchange data, and integration of data (with in-house produced data).

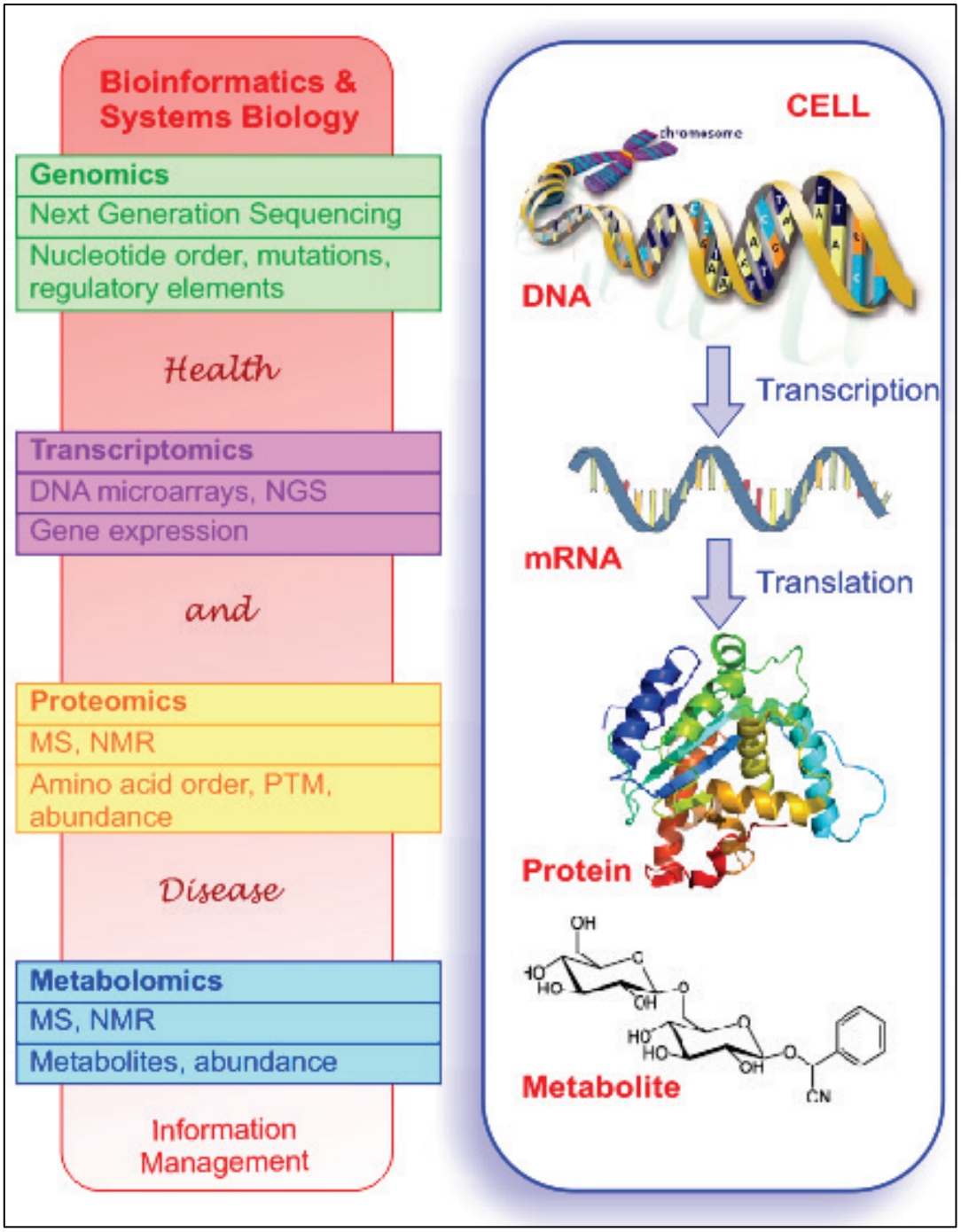


Figure 1.1. The central dogma, OMICS, Bioinformatics and Systems Biology in health and disease. NGS=Next Generation Sequencing; MS=Mass Spectroscopy; PTM=Post Translational Modification; NMR=Nuclear Magnetic Resonance; In addition to the omics technologies mentioned here, there also exist additional omics technologies to measure, for example, epigenetic characteristics (e.g., histon or DNA methylation).

***Box 1. High throughput wet-lab technologies**. These technologies allow determining the full nucleotide order of the DNA of organisms, and measuring a large number of genes, proteins and metabolites simultaneously. In OMICS in Biomedical Sciences you will be introduced to a selection of these methodologies and how these are applied in research and clinic.*

In general, OMICS refers to data-driven biology. OMICS technologies are high-throughput experimental (=wet-lab) technologies to measure molecules inside or outside a cell.

- *"**experimental**": Use in the laboratory (wet-lab)*
- *"**high-throughput**": many samples, genes, proteins, etc are measured simultaneously.*
- *"**molecules**": DNA, mRNA, proteins, metabolites, etc*

OMICS technologies allow the measure at every level of the central dogma (Figure 1.1). Various experimental technologies (e.g., Next Generation Sequencing, DNA microarrays, mass-spectroscopy) are available for measurements of these molecules. A range of molecular properties and events can be measured. For example, nucleotide order of DNA, mutations in DNA, expression levels of genes, concentrations of metabolites and proteins, post translational modifications (PTM) of proteins, etc.

OMICS informally refers to a field of study in biology ending in -omics, such as genomics, transcriptomics, proteomics or metabolomics:

- *genomics study of (full) genomes (DNA sequences)*
- *transcriptomics study of gene expression (mRNA)*
- *proteomics study of proteins (e.g., content, expression)*
- *metabolomics study of metabolites (e.g., content, concentrations)*

The different omics levels are shown in Figure 1.2.

The omics domains produce "ome's":

- *Genome DNA sequence*
- *Transcriptome mRNA species and expression*
- *Proteome protein species (and expression)*
- *Metabolome metabolite species (and expression)*

*Note: 'genomics' is often used to refer to all these fields. The term '**genome-wide**' is used to indicate the measurements of all or many genes, proteins and metabolites in parallel by any of the omics technologies.*

Box 2. What is bioinformatics? *There are many definitions of bioinformatics. The definitions below provide a good idea about this scientific field:*

- *Advancing the scientific understanding of living systems through computation*
- *Bioinformatics is conceptualizing biology in terms of molecules and applying "informatics techniques" (derived from disciplines such as applied mathematics, computer science and statistics) to understand and organize the information associated with these molecules, on a large scale. In short, bioinformatics is a management information system for molecular biology and has many practical applications.*
- ***Extraction of biomedical knowledge from complex data**.*

Bioinformatics is used for a wide-range of applications. For example:

- *Quality control, pre-processing and statistical analysis of OMICs data;*
- *Nucleotide / protein sequence comparison (sequence alignment);*
- *Identification of genes and regulatory elements in full genomes (DNA);*
- *Modelling of 3D-protein structures;*
- *Biological pathway analysis (e.g., understanding molecular mechanisms);*
- *Evolutionary analysis (phylogenetics);*
- *Determining genotype – phenotype relationships;*
- *Biomarker identification (molecular disease signature) for diagnostics*
- *Text mining of literature and electronic patient records.*

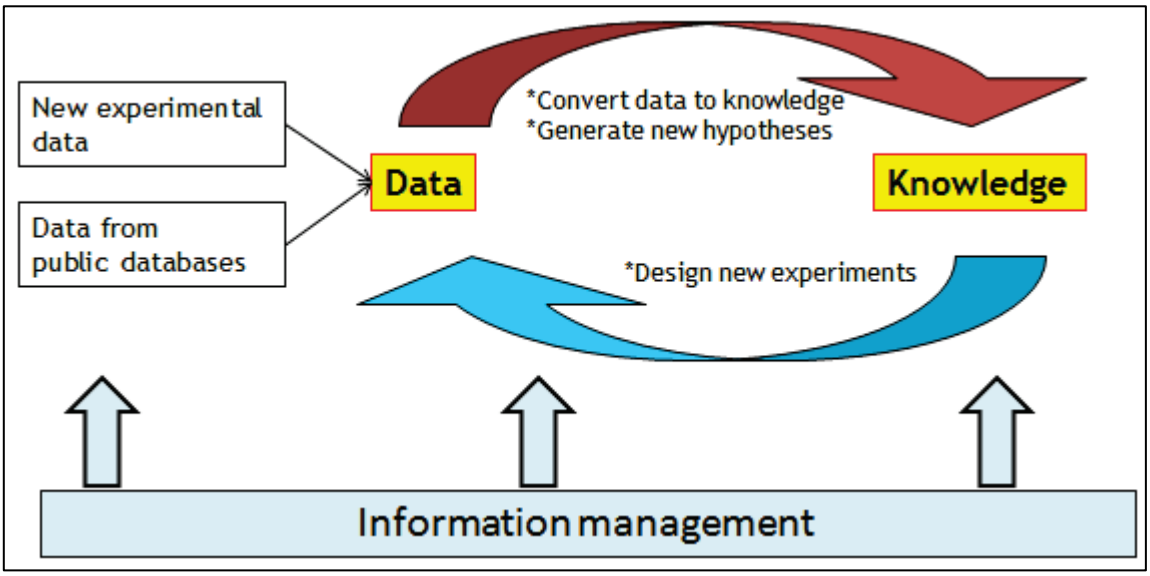


Figure 1.2. Aspects of bioinformatics: knowledge extraction from data, experimental design, and information management.

1.6 Systems medicine

Systems medicine finds its roots in **systems biology**, the scientific discipline that aims at a *systems level* understanding of, for example, biological networks, cells, organs, organisms, and (cell) populations. It generally involves a combination of wet-lab (omics) experiments and computational (bioinformatics) approaches. Systems medicine extends systems biology by focussing on the application of systems-based approaches to clinically relevant applications in order to improve patient health or the overall wellbeing of (healthy) individuals (Wolkenhauer et al 2013).

Box 3. Systems Biology. *Systems biology is a biology-based inter-disciplinary field of study that focuses on complex (dynamic) non-linear interactions within biological systems (e.g., pathways, cells, organs, organisms), using a more holistic perspective, instead of the more traditional reductionism approach, to biological and biomedical research. It involves a combination of wet-lab experiments and computational approaches. It brings together experimental information about the interplay of the components of the systems in computational models that generate testable hypothesis, allow prediction of the behaviour of such systems, and discover emergent properties. OMICS data can be used to construct statistical or mathematical models of systems.*

Example of a biological system *(not for examination)*

Figure 1.3 *shows a simple biological system comprising three transcription factors that are involved in the differentiation of centrocytes (a type of B-cell) to plasma cells. This differentiation takes place in the germinal centre, which is an anatomical structure involved in affinity maturation of B cells. These three transcription factors make up a gene regulatory network with positive and negative feedback loops. The structure of this network results in a so-called bistable system with two steady-states. One steady state is represented by the centroblasts and centrocytes (type of B cells), while the other state is represented by the plasma cells. The differentiation to plasma cells takes place if the centroblast/centrocyte receive external signals through binding to the antigen and binding to a T-cell. The differentiation to plasma cells is irreversible.*

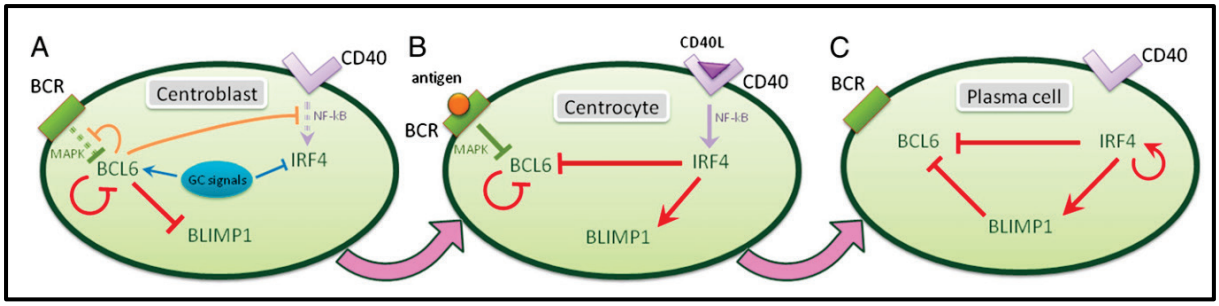


Figure 1.3. Example of a biological system at the molecular level. Time-dependent gene regulatory network of germinal center (GC) B cells. **(A)** Regulatory network at the centroblast stage. Upstream signals promote the expression of BCL6, a transcriptional repressor that controls the regulatory program of the GC. BCL6 directly represses BLIMP1, a key regulator necessary for plasma cell establishment. **(B)** At the centrocyte stage, the B cells compete for survival signals delivered by the B-cell receptors (BCRs) though binding to the antigen, and T cells through binding to CD40L, which lead to degradation of BCL6 protein and up-regulation of IRF4. **(C)** In the plasma cell stage, BLIMP1 and IRF4 are expressed and contribute to the transcriptional silencing of BCL6. The cell is locked in this terminally differentiated stage by a self-positive regulatory loop on IRF4. (Figure copied from Martínez, et al (2012) Proceedings of the National Academy of Sciences of the United States of America 109, 2672).

Systems medicine is expected to change health care practice in the coming years. It will contribute to new therapeutics through the identification of novel disease genes that provide drug candidates less likely to fail in clinical studies (Hood and Auffray 2013; Schneider and Klabunde 2013). It is also expected to contribute to fundamental insights into molecular networks perturbed by disease, improved prediction of disease progression (i.e., improved **molecular disease signatures**), **stratification** of disease subtypes, **personalized treatment**, and **prevention** of disease.

To enable systems medicine it is necessary to characterize the patient at various levels and, consequently, to collect, integrate and analyse various types of data including clinical (phenotype) and molecular data, but also information about cells (e.g., disease-related alterations in organelle morphology), organs (e.g., lung impedance when studying respiratory disorders such as asthma or chronic obstructive pulmonary disease), and even social networks. The full realization of systems medicine, therefore, requires the integration and analysis of environmental, genetic, physiological, and molecular factors at different temporal and spatial scales, which currently is very challenging. It will require large efforts from various research communities to overcome current experimental, computational, and information management related barriers.

1.7 Personalized medicine

Individualized or personalized medicine (PM) seems to lack a clear definition and is open to interpretation. Several definitions can be found in literature (referred to in Schleidgen et al 2013):

One definition is that PM is not a new concept at all since medicine has always been individualized. Other definitions say that PM is holistic health care centred around the needs of the individual patient, or that PM is treatment targeted at stratified subgroups.

However, for several reasons these definitions are not adequate. For example, medicine has always been individualized but literature shows that this is not what PM implies. Similarly, holistic health care is not an adequate definition of PM.

A better definition of personalized medicine is given by (Schleidgen et al 2013):
Personalized medicine seeks to improve stratification and timing of preventive and therapeutic measures by utilizing biological information and biomarkers on the level of molecular disease pathways, genetics, proteomics as well as metabolomics.

This definition makes clear that PM is about the detection of sub-groups (**stratification**) of patients that benefit from a certain measure. Thus, PM is not about individual patients. Instead, an individual patient is assigned to a certain sub-group of patients that appear to respond particularly well to a specific intervention. These sub-groups are identified by studying large groups of patients at the molecular level: patients with similar molecular characteristics (e.g., mutations) will form a sub-group. Each sub-group may require different interventions.

For example Table 1.1 shows different subtypes of breast cancer. Most breast cancers are luminal tumours. Luminal tumour cells look the most like cells of breast cancers that start in the inner (luminal) cells lining the mammary ducts. Of the four subtypes, luminal A tumours tend to have the best prognosis, with fairly high survival rates and fairly low recurrence rates. Because luminal A tumours tend to be ER-positive, treatment for these tumours often includes hormone therapy (such as tamoxifen).

Subtype	These tumors tend to be*	Prevalence (approximate)
Luminal A	• ER-positive and/or PR-positive • HER2-negative • Low Ki67	30-70%
Luminal B	• ER-positive and/or PR-positive • HER2-positive (or HER2-negative with high Ki67)	10-20%
Triple negative/basal-like	• ER-negative • PR-negative • HER2-negative	15-20%
HER2 type	• ER-negative • PR-negative • HER2-positive	5-15%

Table 1.1. Breast cancer subtypes. These are the most common profiles for each subtype. However, not all tumors within each subtype have all these features. ER-positive = estrogen receptor-positive; ER-negative = estrogen receptor-negative; PR-positive = progesterone receptor-positive; PR-negative = progesterone receptor-negative; HER2-positive = HER2 receptor-positive; HER2-negative = HER2 receptor-negative. Table copied from http://ww5.komen.org/BreastCancer/SubtypesofBreastCancer.html

1.8 New dimensions in biomedical research

The developments in experimental technologies that led to high-throughput experimental technologies have led to challenges that require additional expertise and new skills for biomedical researchers:

- **Information management (research data management)**. Modern biomedical research projects typically produce large and complex omics data sets, sometimes in the order of hundreds of gigabytes to terabytes of which a large part has become available through public databases (Baxevanis 2011; Fernandez-Suarez et al 2014). This contributes to

knowledge dissemination and facilitates re-analysis and meta-analysis of data, evaluation of hypotheses that were not considered by the original research group, and development and evaluation of new bioinformatics methods. The use of existing data can in some cases even make new (expensive) experiments superfluous. Alternatively, one can integrate publicly available data with data generated in-house for more comprehensive analyses, or to validate results (Rung and Brazma 2013). In addition, the obligation of making raw data available may prevent fraud and selective reporting. The management (transfer, storage, annotation, and integration) of data and associated meta-data is one of the main and increasing challenges in bioinformatics that needs attention to safeguard the progression of systems medicine.

- **Data analysis and interpretation**. Bioinformatics and statistical data analysis and interpretation of omics data has become increasingly complex, not only due to the vast volumes and complexity of the data but also as a result of more challenging research questions. Many of the methods developed in these areas are of direct relevance for systems medicine.

- **Experimental design**. Once the biological question is formulated the researcher has to decide on a sampling scheme. To successfully design the experiment the researcher has to overlook the complete process of data acquisition, from the biological experiment up to and including the measurement, to identify possible factors (e.g. temperature) that can disturb a proper measurement. These factors are usually not of experimental interest, but introduce noise and/or bias. There are three basic principles: 1) replication, 2) randomization, 3) blocking, that enable the researcher to deal with these factors. The aim of the experimental design is to ensure reliable measurements free from bias.

- **e-Science**. The large volumes of data, complexity of the data analysis, increased compute and storage requirements, and the emergence of large multi-disciplinary and geographically distributed research consortia have prompted computer science and bioinformatics groups to develop and/or apply advanced (high performance) ICT infrastructures such as grid and cloud, workflow engines, semantic web for e.g., development of biomedical ontologies, provenance, and Science Gateways to make computational services accessible to researchers and medical doctors. Such approaches are collectively known as e-Science and have led to several applications, for example, in DNA sequence analysis and cancer systems biology.

Clearly, the new experimental technologies have to a large extent **turned biomedical research in a data- and compute-intensive endeavour**. It has been argued that production of omics data has nowadays become the ‘easy’ part of biomedical research, whereas the real challenges currently comprise information management and bioinformatics analysis.

Consequently, next to the wet-lab, the computer has become one of the main tools of the biomedical researcher. **For all these reasons it is important that the next-generation of biomedical scientists have not only sufficient skills to design and perform wet-lab experiments but also master expertise about computational and statistical approaches to manage and analyse their own data**.

1.9 What will you learn?

In this text you have probably encountered many new aspects of biomedical research that you were not aware of. In addition, some of the terminology may be new or still confusing but this will become clearer during the coming lectures and computer labs.

OMICS of Biomedical Sciences will introduce you into:

- High throughput wet-lab technologies (OMICS)
- Large data volumes & information management (public biological databases).
- Analyzing and interpreting large data volumes (bioinformatics).
- Understanding biological systems (systems biology & systems modelling).

References

Baxevanis AD (2011) The importance of biological databases in biological discovery. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* Chapter 1: Unit 1 1.

Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. *Nature genetics* 21: 33-37.

Fernandez-Suarez XM, Rigden DJ, Galperin MY (2014) The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic acids research* 42: D1-6.

Hood L, Auffray C (2013) Participatory medicine: a driving force for revolutionizing healthcare. *Genome medicine* 5: 110.

Lindon JC, Nicholson JK (2008) Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics. *Annual review of analytical chemistry* 1: 45-69.

Lusis AJ, Attie AD, Reue K (2008) Metabolic syndrome: from epidemiology to systems biology. *Nature reviews Genetics* 9: 819-830.

Mensaert K, Denil S, Trooskens G, Van Criekinge W, Thas O, De Meyer T (2014) Next-generation technologies and data analytical approaches for epigenomics. *Environmental and molecular mutagenesis* 55: 155-170.

Metzker ML (2010) Sequencing technologies - the next generation. *Nature reviews Genetics* 11: 31-46.

Paik S, Shak S, Tang G, et al (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England journal of medicine* 351: 2817-2826.

Rung J, Brazma A (2013) Reuse of public genome-wide gene expression data. *Nature reviews Genetics* 14: 89-99.

Schleidgen S, Klingler C, Bertram T, Rogowski WH, Marckmann G (2013) What is personalized medicine: sharpening a vague term based on a systematic literature review. *BMC medical ethics* 14: 55.

Schneider HC, Klabunde T (2013) Understanding drugs and diseases by systems biology? *Bioorganic & medicinal chemistry letters* 23: 1168-1176.

van 't Veer LJ, Dai H, van de Vijver MJ, et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415: 530-536.

Wolkenhauer O, Auffray C, Jaster R, Steinhoff G, Dammann O (2013) The road from systems biology to systems medicine. *Pediatric research* 73: 502-507.

Zanotti L, Bottini A, Rossi C, Generali D, Cappelletti MR (2014) Diagnostic tests based on gene expression profile in breast cancer: from background to clinical use. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* 35: 8461-8470.

2 Next Generation Sequencing

Lecturer: Prof. dr. Antoine van Kampen (AMC)

After reading this chapter you should understand

- Difference between Sanger sequencing and Next Generation Sequencing (NGS)
- Applications of NGS
- Experimental errors, quality scores,
- Coverage (read depth, breadth)
- DNA capture methods
- Sample multiplexing,
- Single-end, paired-end, mate-pairs, long read sequencing
- Molecular barcodes (UMIs)
- Single-cell sequencing
- Pre-processing of sequence data

Contents

2.1 (Next Generation) Sequencing applications

DNA/RNA sequencing refers to the determination of the precise nucleotide order in nucleotide sequences (DNA and RNA). Until very recently, much of the modern DNA sequencing was based on an enzymatic sequencing method (so-called Sanger sequencing[1]) first developed in the 1970's, in which a DNA polymerase was used to synthesize new DNA chains by using a cloned single-stranded DNA template, consisting of millions of identical copies of a specific DNA sequence. Initially, DNA sequencing was slow and laborious and output was limited, but the need to scale up to conduct large-scale sequencing of library clones drove technological improvements. This resulted in Next Generation Sequencing (NGS) technologies, also known as massively parallel sequencing, which largely increased the throughput of DNA sequencing.

DNA and RNA sequencing has many applications. For example:

- **Full genomes**. The determination of the complete DNA sequence (genome) of an organism such as human, animals, bacteria, and viruses. The determination of a complete genome is, in general, not an aim by itself. Instead, it enables some the applications described below.

- **Variant detection**. By comparing DNA or RNA sequences to a reference sequence it becomes possible to determine small differences (variants) between the sequences. Such variants include Single Nucleotide Polymorphisms (SNPs) and small insertions and deletions (indels). A reference sequence can be obtained from a public database, or from an individual that serves as a control. Detection of (novel) variants is important since these may cause or contribute to disease.

- **Structural variation**. Also known as genomic structural variation. This is the variation in structure of an organism's chromosome. It consists of many kinds of variation in the genome of one species, and usually includes microscopic and submicroscopic types, such as deletions, duplications, copy-number variants, insertions, inversions and translocations. Typically a structure variation affects a sequence length about 1Kb to 3Mb, which is larger than SNPs and smaller than chromosome abnormality (though the definitions have some overlapping). Many structural variations are associated with disease.

- **Splice variant detection**. Alternative splicing is a regulated process during gene expression that results in a single gene coding for multiple proteins. In this process, particular exons of a gene may be included within, or excluded from, the final, processed messenger RNA produced from that gene. Through comparison of RNA

[1] Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74: 5463-5467.

sequences to other RNA sequences of the same gene or to a reference genome, one may identify different splice variants.

- **RNA-seq**. Is used to measure gene expression through the sequencing and counting of mRNA molecules. RNA-seq is an alternative for DNA microarrays.

- **Chip-seq**. Is used to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. It can be used to map global binding sites precisely for any protein of interest.

- **Exome sequencing**. This application allows the detection of gene variants in the coding part of the genome

- **DNA methylation**. By treating DNA with bisulfite and subsequently sequencing the DNA it is possible to determine its pattern of methylation. DNA methylation was the first discovered epigenetic mark, and remains the most studied. In animals it predominantly involves the addition of a methyl group to the carbon-5 position of cytosine residues of the dinucleotide CpG, and is implicated in repression of transcriptional activity.

- **Metagenomics**. This is the study of metagenomes (multiple genomes in parallel), genetic material recovered directly from environmental samples (e.g., from the gut). By sequencing all DNA in such sample, one obtains a largely unbiased view of all genes from all the members (e.g., genomes bacteria and viruses) of the sampled communities.

You should be aware of these applications and have basic knowledge of how sequencing is utilized in these applications.

2.2 Sample processing

Prior to the sequencing of DNA or RNA molecules we need to extract DNA or RNA from blood, cell or tissue. For Sanger sequencing, the DNA or RNA is amplified by recombinant DNA technologies or PCR. DNA libraries of DNA fragments are constructed. These pre-sequencing steps are beyond the scope of this text. If you want to refresh your mind then have a look at Chapter 8 of Molecular Biology of the Cell[2].

In the next sections 'Sanger sequencing' and 'Next Generation Sequencing' are explained in more detail. You may also want to read chapter 8 (e.g., Panel 8-1, pg478) of 'Molecular Biology of the Cell'[2].

2.3 Sanger sequencing (not for examination)

The DNA sequencing method that was used to sequence the human genome and many other genomes was an enzymatic sequencing method pioneered by Fred Sanger (Figure 2.1) [1]. It relies on random inhibition of chain elongation, creating newly synthesized DNA strands of various lengths that can be separated by size. The DNA needs to be in a single-stranded form that will act as a template for making a new complementary DNA strand in vitro by using a DNA polymerase.

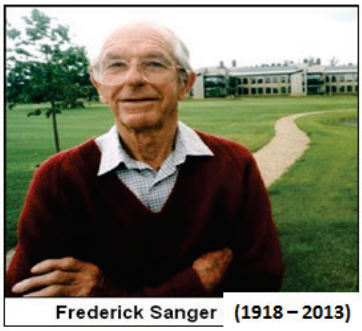


Figure 2.1. Frederick Sanger.

Single stranded sequencing template

The substrate for DNA sequencing was often a recombinant DNA that would be denatured so that a strand-specific sequencing primer could be used to direct new strand synthesis, or DNA fragments would be cloned into phagemid vectors that were manipulated to produce single-stranded recombinant DNA. Alternatively, and increasingly commonly, DNA produced by PCR amplification is used and converted to a single stranded form to act as a sequencing template. The final product of either method is a population of many identical copies of the DNA to be sequenced.

The use of ddNTPs for termination of chain elongation

Sequencing is conducted in four parallel reactions, each containing the four dNTPs (dATP, dCTP, dGTP, dTTP) plus a small proportion of one of the four analogous dideoxynucleotides (ddNTPs) that will serve as a base-specific chain terminator. A ddNTP is closely related to its dNTP counterpart but lacks a hydroxyl group (Figure 2.2). It can be incorporated into the growing DNA chain. However, because it lacks a 3' hydroxyl group, any ddNTP that is incorporated into a growing DNA chain cannot participate in phosphodiester bonding at its 3'

[2] Molecular Biology of the Cell, Alberts et al., Taylor & Francis Inc (sixth edition).

carbon atom. Once a ddNTP has been incorporated, therefore, it caused the abrupt termination of chain elongation.

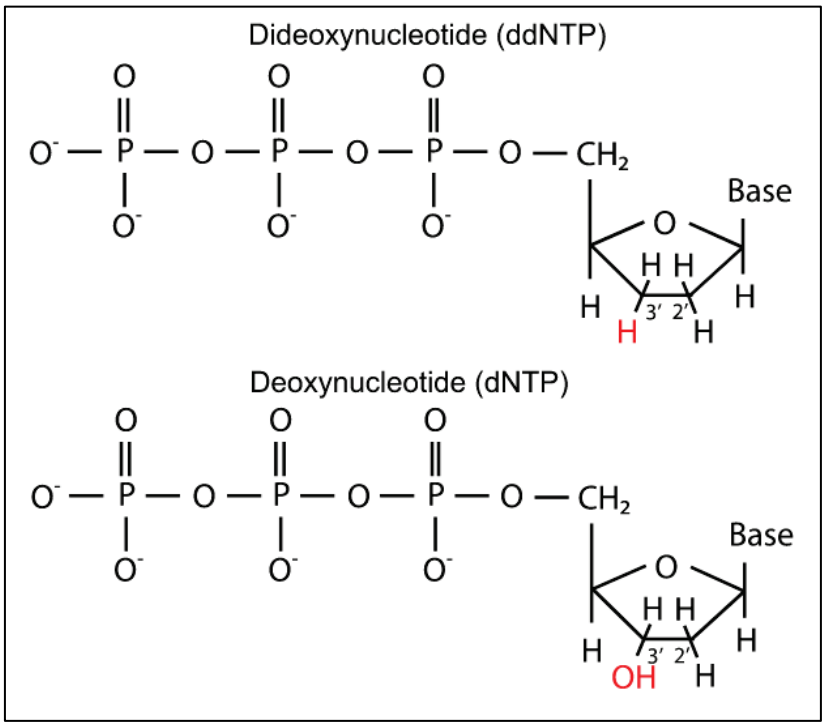


Figure 2.2. Structure of a dNTP and ddNTP. The hydroxyl group attached to carbon 3' in normal nucleotides is replaced by a hydrogen atom. (Figure copied https://binf.snipcademy.com/lessons/dna-sequencing-techniques/sanger-dideoxynucleotide). (Not for examination).

Sequencing reaction

By setting the concentration of the ddNTP to be very much lower than that of the corresponding dNTP analog, there will be competition between a specific ddNTP and its dNTP counterpart for inclusion in the growing DNA chain. The dNTP is present in excess; when it is incorporated, chain elongation continues, but occasionally the ddNTP will be incorporated in the growing chain, ending polymerization and so causing chain termination. Each reaction is therefore a partial reaction because chain termination will occur randomly at one of the possible choices for a specific type of base in any one DNA strand.

Because the DNA sample is a population of identical molecules, each of the four base-specific reactions will generate a collection of labeled DNA fragments of different lengths. Each of the fragments in one reaction will have a common 5'-end (defined by the sequencing primer). However, the 3' ends are variable because the insertion of the selected ddNTP occurs randomly at one of the many different positions that will accept that specific base (Figure 2.3). Fragments that differ in size by even a single nucleotide can be size-fractionated on a gel according to their molecular masses.

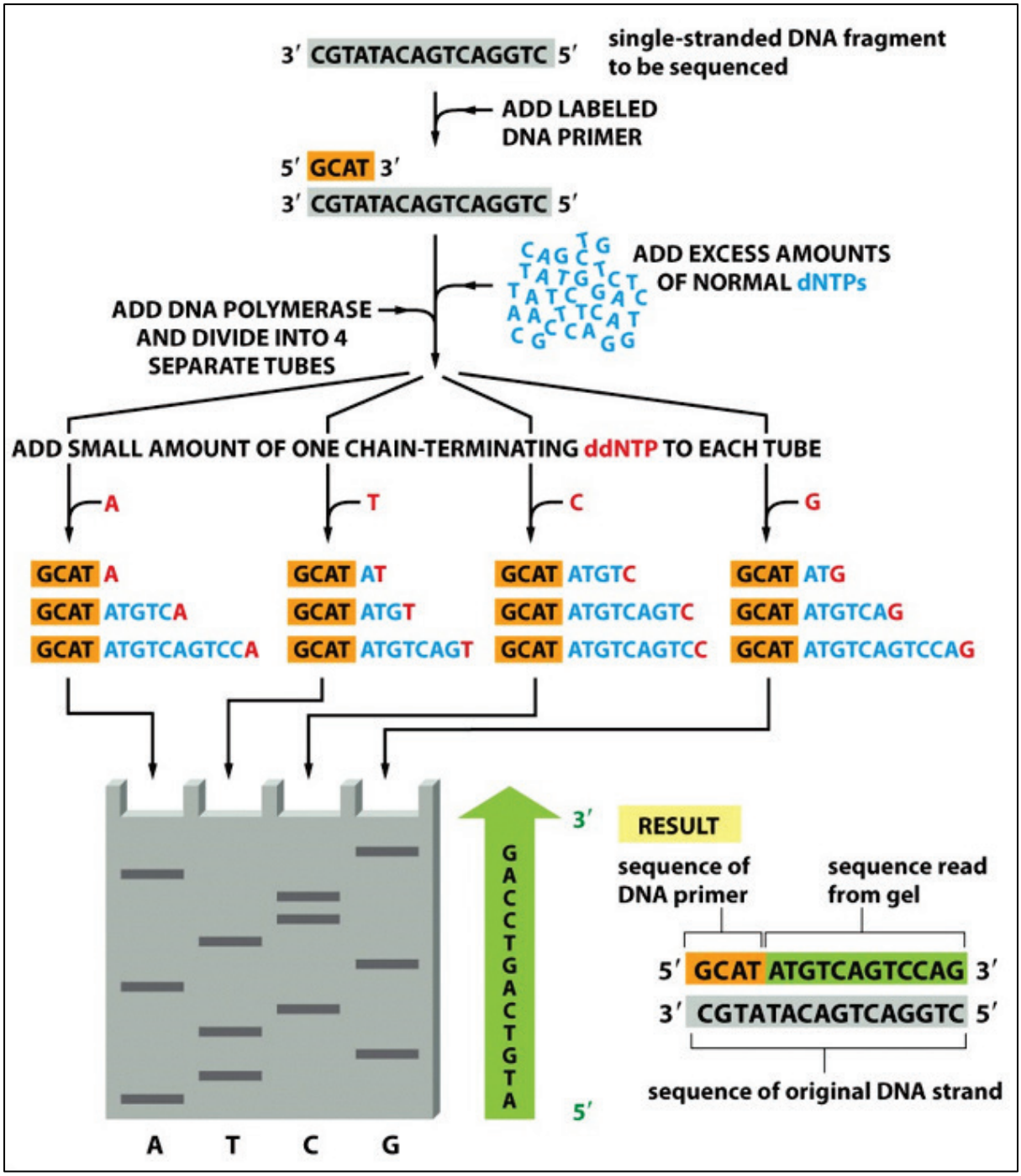


Figure 2.3. Dideoxy sequencing. With the use of a primer about 20 nucleotides long, a complementary sequence is synthesized from a single stranded DNA template. Four base-specific reactions are performed. Size fractionation on a polyacrylamide gel enables the sequence to be determined giving the sequence. (Figure copied Alberts B., Johnson A., Lewis J., Raff M., Roberts K., Walter P. (2014) Molecular biology of the cell. Sixth edition ed. Garland Science, New York)

Radioactive labeling

By ensuring that one of the four dNTPs or the primer is labeled with a radioactive label (Figure 2.4), the growing DNA strand becomes labeled. Subsequent, separation on a gel allows the determination of the nucleotide order Figure 2.5.

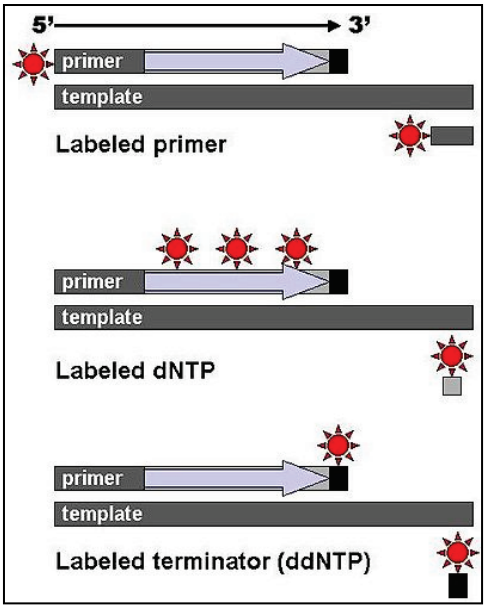


Figure 2.4. Three labeling methods

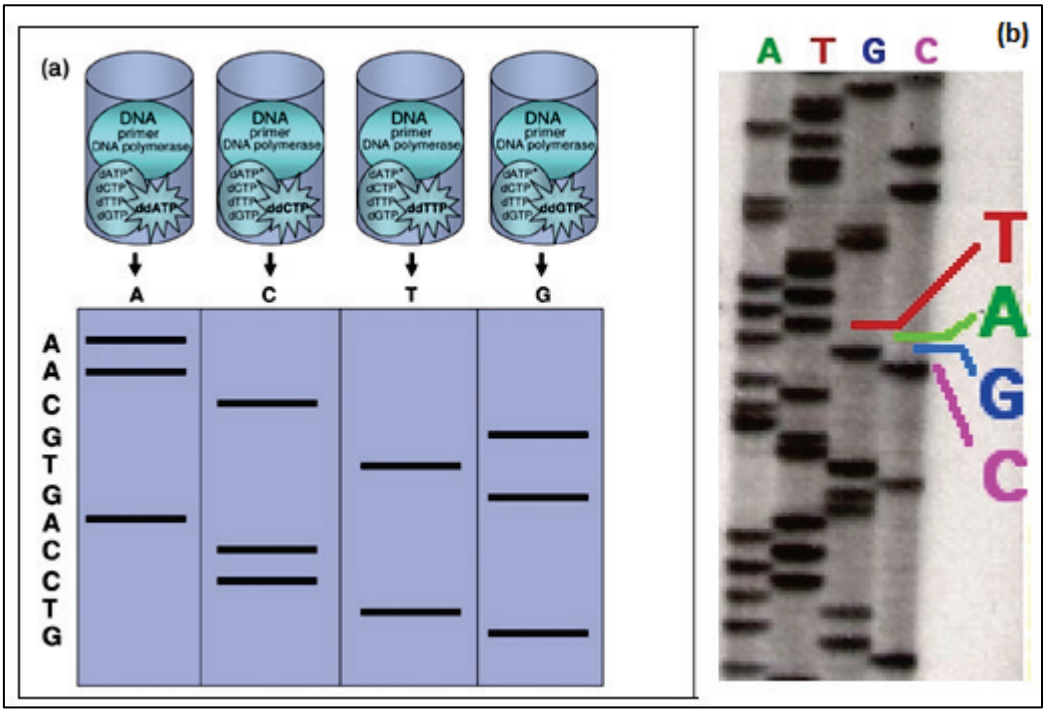


Figure 2.5. Schematic principle of the Sanger sequencing method. (A) Four separate reactions (four tubes). dNTPs or primer is radioactively labeled. Radioactive products are then separated through four lanes of a gel and scored according to their molecular mass. (Figure (A) copied from http://www.wiley-vch.de/books/sample/3527320903_c01.pdf. Figure (b) copied from http://en.wikipedia.org/wiki/Sanger_sequencing)

Automation of dideoxy DNA sequencing

Automated DNA sequencing machines that used fluorescence labeling (instead of radioactive labeling) of DNA were developed in the early 1990s. Four different fluorescent dyes are used

in the four base-specific reactions (Figure 2.7). By selecting dyes with different emission wavelengths, all four reactions could be loaded into a single sample well on the gel. During electrophoresis, the DNA fragments pass by an excitation source such as a laser, while a monitor detects and records the fluorescence signal as the DNA passes through a fixed point in the gel (Figure 2.6). This allows an output in the form of intensity profiles for each of the differently colored fluorophores while simultaneously storing the information electronically.

Early automated DNA sequencers used polyacrylamide gels, but greater DNA sequencing capacity became possible with capillary sequencing. In this technique, DNA samples migrate through long and thin glass capillary tubes containing a gel. Also these capillary machines read the base sequence as DNA moves through the gel, but a higher degree of automation can be achieved.

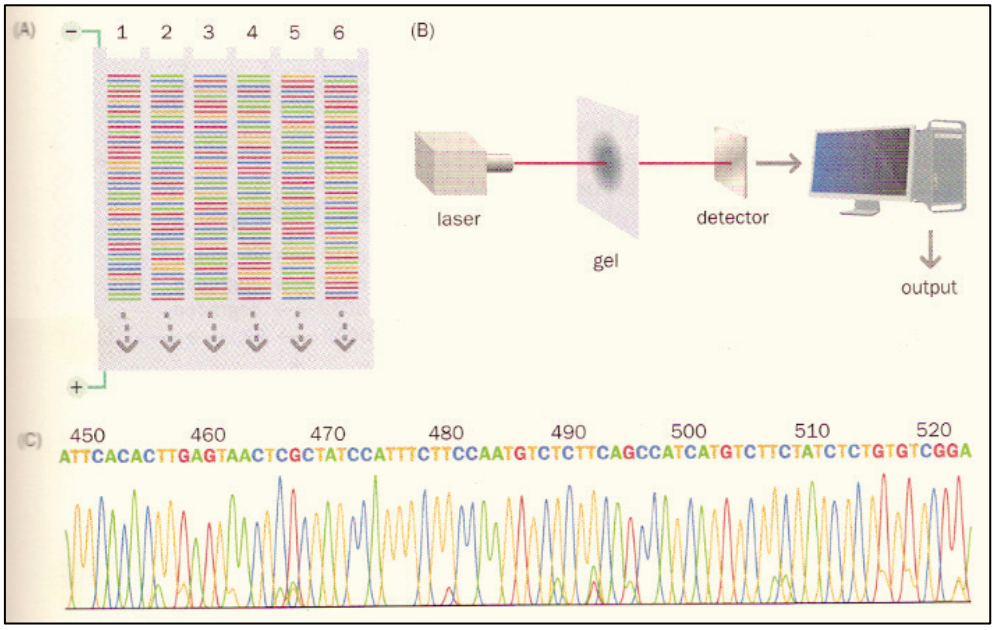


Figure 2.6. Automated DNA sequencing with fluorescent primers. (A) Four separate fluorescent dyes are used as labels for the base-specific reactions (label can be attached to primer or ddNTP). Samples are size fractionated on a gel. (B) while the fragments are migrated downward during the electrophoresis run, a laser beam is focused at a specific constant position on the gel. As the individual DNA fragments migrate past this position, the laser cause the dyes to fluoresce. The information is recorded electronically and the interpreted sequence is stored in a computer database. (C) Example of DNA sequence output, showing a succession of dye-specific (and therefore base-specific) intensity profiles. (Figure copied from Strachan and Read (2011) Human Molecular Genetics, 4th edition, Garland Science, NY)

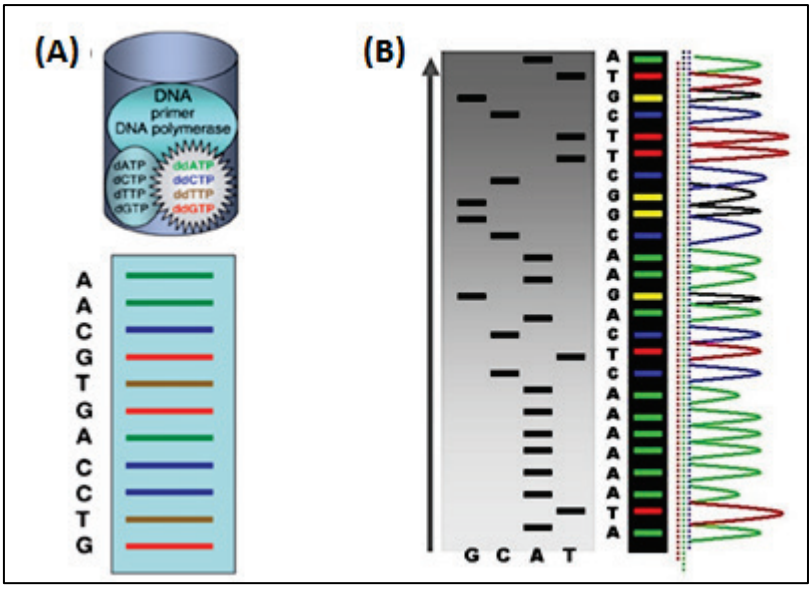


Figure 2.7. (A) In this case, instead of adding radioactive dATP, all four ddNTPs are labeled with different fluorescent dyes in a single tube. The extension products are then electrophoretically separated in a single glass capillary filled with a polymer. DNA bands move inside the capillary according to their masses. Fluorophores are excited by the laser at the end of the capillary. The DNA sequence can be interpreted by the color that corresponds to a particular nucleotide. (B) a comparison between radioactive and fluorescent labeling. (Figure A copied from http://www.wiley-vch.de/books/sample/3527320903_c01.pdf. Figure B copied from http://en.wikipedia.org/wiki/Sanger_sequencing)

2.4 Next generation sequencing (NGS)

With minor changes, the dideoxysequencing method underpinned molecular genetics for three decades. The method is, however, disadvantaged by relying on gel electrophoresis to fractionate newly synthesized DNA fragments. Not only does this make the method a laborious one, but more importantly it makes it difficult to sequence large number of DNA fragments at a time. Fundamentally different DNA sequencing technologies that did not require gel electrophoresis were not developed until the early to mid-2000s.

An important breakthrough was to develop methods that could record the DNA sequence while a DNA strand was being synthesized by a polymerase from a single-stranded DNA template. That is, the sequencing method was able to monitor the incorporation of each nucleotide in the growing DNA chain and to identify which nucleotide was being incorporated at each step. At the heart of the first such approach was a method known as pyrosequencing. This approach was used by the 454 Life Sciences sequencer, which now has been replaced by newer sequencers. Many different massively parallel sequencing platforms have recently been developed that can carry out up to billions of sequencing reactions in parallel (see Table 1).

NGS technologies detect bases from measuring light intensity if a nucleotide is incorporated in the DNA chain. The detection of bases is generally referred to as **base-calling**.

Note that Next Generation sequencing (NGS), massively parallel sequencing (MPS) and high-throughput sequencing all refer to the same technologies. A recent overview of NGS technologies is given in[3]. A table from this paper with a comparison of sequencing technologies (platforms) is given at the end of this chapter.

2.4.1 Differences between NGS and capillary-based sequencing

- **Library construction**. NGS sequence reads (nucleotide sequences) are produced from fragment 'libraries' that have not been subject to the conventional cell-based DNA cloning and amplification used in capillary sequencing. NGS also does not rely on gel electrophoresis to fractionate newly synthesized DNA fragments. The workflow to produce next-generation sequence-ready libraries is straightforward: DNA fragments that may originate from a variety of front-end processes (e.g., sonication) are prepared for sequencing by ligating specific adaptor oligos to both ends of each DNA fragment (see Appendix 1. 'Sequencing by Ligation' for a specific example). Importantly, relatively little input DNA is needed to produce a library.

- **Parallelism**. NGS methods have the ability to process millions of sequence reads in parallel rather than 96 (capillaries) at a time with Sanger sequencing. These sequence reads are derived from the sequence fragments (i.e., the actual DNA) produced during library construction.

- **Read lengths and sequence error rates**. After three decades of gradual improvement, the Sanger biochemistry can be applied to achieve read-lengths of up to ~1,000 bp, and per-base 'raw' accuracies as high as 99.999%. In the context of high-throughput shotgun genomic sequencing. In comparison NGS sequences are shorter (50-300 bp) and the accuracy of NGS ranges from ~85 - 99.9%[3,4].

- **Sequencing costs.** Sanger sequencing costs on the order of $0.50 per kilobase. . Assuming a full genome sequence for $1000, the NGS sequence costs would be $0.0003 per kilobase.

2.4.2 Experimental errors in NGS

Sequencing methods are imperfect. NGS applications such as whole-genome sequencing, targeted capture (see below), high-throughput RNA sequencing (RNA-seq) and chromatin immunoprecipitation followed by sequencing (ChIP–seq), are prone to errors that result in miscalled bases, leading to for example misalignment of short reads (**Question: why?**). A 'miscalled' base is an incorrect nucleotide that was obtained from the measurement.

[3] Goodwin et al (2016) Nature Reviews Genetics, 17,333
[4] Shendure and Hanlee (2008) Nature Biotechnology, 26, 1135

Thus, the sequencing error rate is defined as the percentage of bases that are incorrectly called. If 0.8% error rate then for every 1000 bases coming off the sequencer, 8 of them will report the incorrect base. When considered alone, an error is indistinguishable from a sequence variant. This problem can be overcome by increasing the number of sequencing reads. If sequencing results in a 1% variant-error rate then the combination of 8 identical reads that cover the location of the variant will produce a strongly supported variant call with an associated error rate of 10^{-16} (=0.01^8). Increased depth of coverage therefore 'rescues' inadequacies in sequencing methods.

Even for methods (Sanger sequencing) that have the lowest reported error rates, the absolute numbers of miscalled genomic variants remain large (there might be thousands of false-positive variants in a fully sequenced human genome). Furthermore, miscalled bases are mistaken as rare and somatic variants, thereby obfuscating true variants of clinical interest. Known sources for experimental errors can be grouped by their occurrence in the sequencing workflow; that is, during sample preparation, library preparation, or sequencing and imaging[5]. These are not further discussed here.

Sequencing errors can affect the detection of rare somatic mutations. A somatic mutation is an alteration in the DNA that occurs after conception. Somatic mutations can occur in any of the cells of the body except the germ cells (sperm and egg) and therefore are not passed on to children. These alterations can (but do not always) cause cancer or other diseases.

Sequencing errors can also affect the detection of SNPs (variants that are passed on to the children) if the coverage (see below) is not high enough.

2.4.3 Single-end, paired-end, and mate-pair

When sequencing you can choose between single-end (SE) or paired-end (PE) sequencing. When sequencing, we chop up our DNA into small fragments, and then ligate some adaptors. Then, for SE, we only sequence *one* end of a DNA fragment. PE sequencing (Figure 8) involves sequencing both ends of the DNA fragments in a sequencing library and aligning the forward and reverse reads as read pairs. The sequenced fragments can be separated by a certain number of bases (insert size) or can be overlapping, giving rise to a contiguous longer single-end fragment after merging. The uses of paired-end reads can improve the accuracy of reads mapping onto a reference genome. The typical fragment size is **200bp** to **500bp.** In addition to producing twice the number of reads for the same time and effort in library preparation, sequences aligned as read pairs enable more accurate read alignment and the ability to, for example, detect indels (insertion and deletions) which is not possible with single-read data.

A mate-pair is different from paired-end in the sense of how the sequence library is made. In mate-pair sequencing, **2-5kb** fragments are selected and sequenced from both end, thus

[5] Robasky, K., Lewis, N.E., and Church, G.M. (2014). The role of replicates for error mitigation in next-generation sequencing. Nature reviews. Genetics 15, 56-62.

giving information how nucleotides that are far apart belong together. Mate-pairs are, for example, used for structural variant (SV) detection across a widened SV size-spectrum, and to resolve repetitive areas during genome assembly.

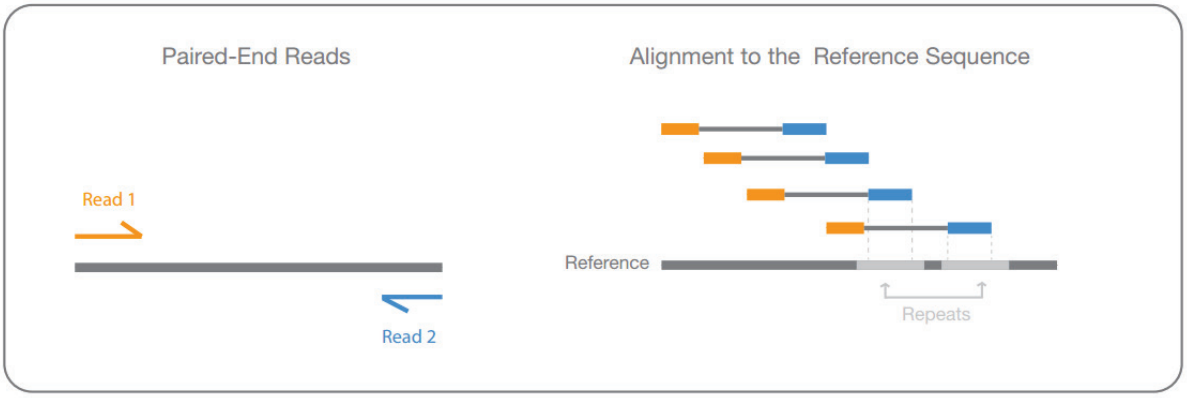


Figure 2.8. Paired-End Sequencing and Alignment—Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome. (Figure copied from https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf).

2.4.4 Long-read sequencing

(Human) genomes are highly complex with many long repetitive elements, copy number alterations, and structural variations that are relevant to evolution, adaptation and disease. Many of these complex elements are so long that short-read paired-end sequencing is insufficient to resolve them. Long-read sequencing delivers reads in excess of several kilobases, allowing for the resolution of these large structural features. Such long-reads can span complex or repetitive regions with a single continuous read, thus eliminating ambiguity in the positions or size of genomic elements. Long-reads can also be useful for transcriptomic research, as they are capable of spanning entire mRNA transcripts, allowing researchers to identify the precise connectivity of exons and discern gene isoforms.

Currently, there are two main types of long-read technologies: **single-molecule** real-time sequencing approaches and **synthetic approaches** that rely on existing short-read technologies to construct long reads *in silico*. The single-molecule approaches differ from short-read approaches in that they do not rely on a clonal population of amplified DNA fragments (e.g. bridge amplification in Illumina) to generate detectable signal, nor do they require chemical cycling for each dNTP added. Alternatively, the synthetic approaches do not generate actual long-reads; rather, they are an approach to library preparation that leverages barcodes to allow computational assembly of a larger fragment.

2.1 Multiplexing

In order to maximize sequencing capacity and reduce workflow of sample preparation, a single sequencing run containing multiple biological samples is sometimes preferred. To this end, a multiplexing method with so-called barcodes has been developed for sequencing platforms. Barcodes are unique 5-10 base sequences that are added at the 3'end of the template. Sets up to 96 barcodes have been designed and can be assigned to up to 96 individual samples.

Figure 2.16 shows the general principle of multiplexing. (A) Shows the situation without multiplexing. Eight samples (libraries; indicated by different colors) are independently processed and deposited on separate areas (lanes) of the flow-cell (slide). In (B) the libraries are pooled after barcodes are added. A single emulsion PCR, enrichment, and deposition is performed using the pooled library sample. Sequencing templates from different samples are no longer separated on the slide. (C) Shows how the barcode is attached to the sequence template. During sequencing the barcode and target DNA are obtained as two independent sequence reads. During data pre-processing the sequences from the different samples are separated.

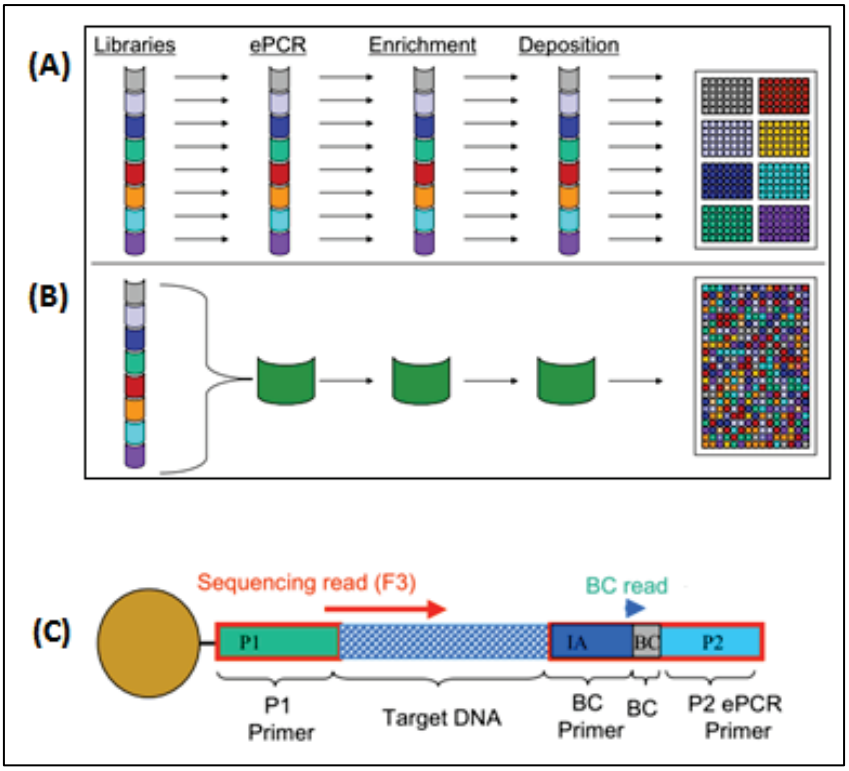


Figure 2.9. The principle of multiplexing with barcodes (BC). ePCR= emulsion PCR.

2.2 DNA capture methods

Notwithstanding the huge recent progress in genome sequencing, many current applications of massively parallel DNA sequencing are focused on target sequences that collectively constitute a small fraction of a genome. For example, screening for cancer gene susceptibility could involve sequencing all exons, exon-intron boundaries, and known regulatory elements for all known cancer genes. There are many hundreds of known cancer genes, including many new genes identified by international programs such as The Cancer Genome Atlas (TCGA). PCR amplification of what may be hundreds of sequence elements in each cancer gene is both tedious and time-consuming. A second example is the screening of the coding part of the human genome for SNPs that cause disease. This so-called '**exome sequencing**' requires the screening of about 1.5% of the human genome (all gene exons).

For such applications it is not (yet) desirable to sequence the complete genome because this still requires more wet-lab work, could complicate data analysis, and is more expensive. Consequently, so-called capture tools (also known as target capture or target enrichment) have been developed to enrich for the desired sequences that are then submitted for high-throughput sequencing.

Figure 2.17 shows two alternative capture technologies. In this example, we are interested in sequencing the colored regions (e.g., the exons) of the Human genome. The genomic DNA is fragmented and sequencing adapters are attached. Subsequently, the regions of interest are captured, and these regions are amplified and sequenced. Two capture approaches are

- Microarray capture
- Solution capture

In both approaches we use probes that are complementary to the regions that we aim to capture. If we are interested in the coding regions of the genome (the gene exons) then we require probes that are complementary to these exons. The Human genome reference sequence or the Consequences Coding Sequence (CCDS) database provide us with information about all known exon sequences. Based on these sequences, complementary probes can be synthesized that are unique for specific exons. In general, multiple probes for a single exon are synthesized.

In microarray capture, the synthesized probes are attached to a solid surface, while for the solution capture the probes are in solution. In case of the microarray capture, the fragment library is loaded on the microarray, which will result in a hybridization of the exon fragments to the complementary probes. The non-hybridized fragments (e.g., non-exon sequences) are washed of and the resulting exon fragments are submitted to sequencing. In case of the solution capture, the fragment library is hybridized to the probes in solution. Subsequently, streptavidin beads are used to pull down the complex of capture probes and hybridized

genomic DNA fragments. Again, unbound fragments are removed by washing and the remaining exon sequences are submitted to sequencing.

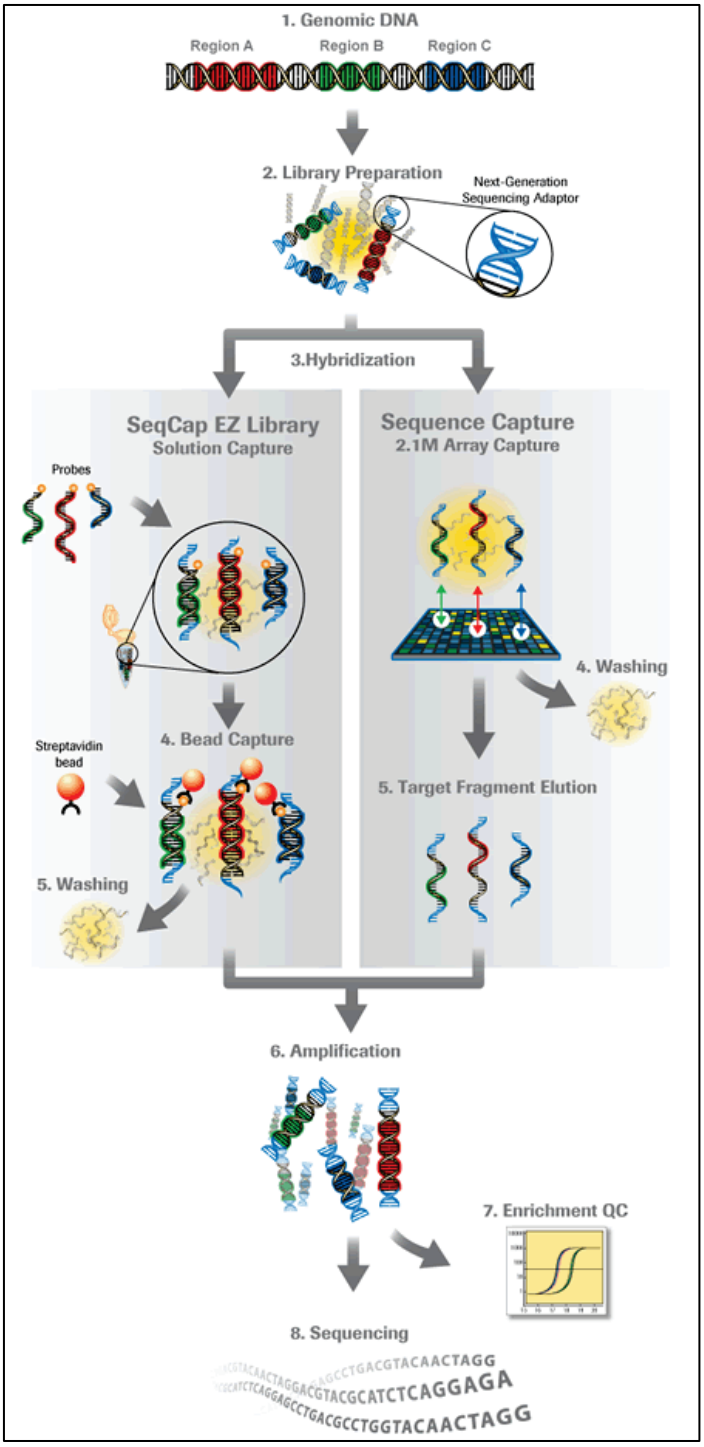


Figure 2.10. DNA capture technologies. This figure shows two alternative capture methodologies: solution capture and array capture.

2.5 Unique Molecular Identifiers (UMIs)

NGS yield vast numbers of short sequences (reads) from a pool of DNA fragments. A wide variety of sequencing applications have been developed which estimate the abundance of a

particular DNA or RNA fragment by the number of reads in a sequencing experiment (read counting) and then compare these abundances across biological conditions. Perhaps the most widely used read counting approach is RNA-seq, which seeks to compare the number of copies of each mRNA transcript in different cell types of conditions. Prior to sequencing, a PCR amplification step is normally performed to ensure sufficient DNA/RNA for sequencing. However, PCR may include biases that lead to an over-representation of specific sequences. Moreover, it may also lead to the incorporation of wrong nucleotides. These biases and errors propagate to the final library that will be sequenced and, thus, may affect the quantification of DNA/RNA abundances.

Determining the relative abundance of two different molecular species or the absolute number of molecules in a single sample is challenging. **Unified Molecular Identifiers** (UMIs; also known as unique identifiers (UIDs)) are short random nucleotide sequences that can be used as an absolute counting method. In this method, each molecule in a population is first made unique by adding an UMI prior to PCR sequencing. The UMIs in the library acts as a molecular 'memory' of the number of molecules in the starting sample. Upon deep sequencing, each UMI will be observed multiple times and the number of original DNA molecules can be determined simply by counting each UMI only once (i.e., counting the unique UMIs).

Through a UMI, identical copies arising from distinct molecules (having distinct UMIs) can be distinguished from those arising through PCR-amplification of the same molecule (having identical UMIs). UMIs also allow to determine sequencing or PRC errors that may have been included during the process.

The use of UMIs lead to a better quantification of gene expression with RNA-seq or to an improved detection of low-frequent mutations. Figure 2.1 shows how UMIs help in the detection of true sequence variants. In this example, UMIs were short sequences comprising 22 random nucleotides. This results in $4^{22} = 1.76*10^{13}$ possible UMIs, which far exceeds the primer molecule number available in a 20 µl reaction with 200 nM primers: $6.02*10^{23} * 20*10^{-6} * 200*10^{-9} = 2.4*10^{12}$ (note: 6.02*1023 is number of Avogadro). Therefore, 22 nucleotides is likely to ensure each template molecule obtains a unique UMI regardless of the template number[6].

In the first step primers (grey) were extended with a stretch of 22 random nucleotides as UMIs (different colors), and partial P5 (light green) or P7 (orange) adaptors to facilitate barcoded libraries construction for Illumina sequencing. Two cycles of PCR are performed resulting in P5/P7 tagged amplicons that include a UMI. Note that the adapters allow the library fragment (grey) to attach to the flow cell surface. Then a second stage of amplification is performed which results in the library that is sequenced.

[6] I will not ask for such calculation during the examination

In this example we start with 2 molecules (transcripts of a specific gene) which contains a mutation (red star). Each molecule is assigned a unique UMI (dark green, dark gray). The mutation will be present in each read with the same UMI. Sequencing or PCR errors (blue stars) that are introduced at a later stage (after the initial 2 PCR cycles) will not be present in all reads with the same UMI. Thus, reads with the same UMI should be identical. All nucleotides that are not the same are sequencing/PCR errors.

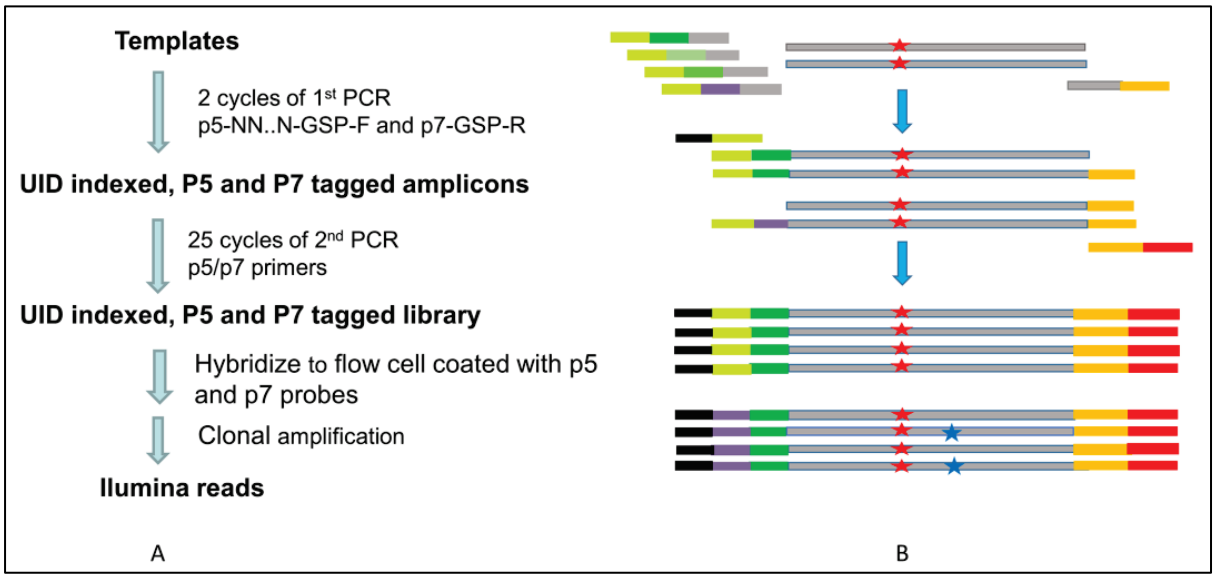


Figure 2.11. UMI-targeted DNA sequencing workflow and the principle in distinguishing errors from true mutation. (A) Illustration of UMI-targeted DNA sequencing workflow. (B) True mutation from errors introduced during PCR and sequencing. A true mutation (illustrated as red star) is expected to be present in all the reads carrying the same UMI (or derived from the same template molecule), while an error (illustrated as blue star) is expected in some but not all the reads carrying the same UMI. (Figure copied from Kou et al (2016). Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. PloS one 11, e0146638.)

2.3 Analysis of sequence data

Once the sequencing is performed the data analysis starts. The precise data analysis steps depend on the application, e.g., the analysis of SNPs will require other analysis than the analysis of gene expression measured with RNAseq.

2.4 The data

To give an impression of the amount of data that is produced by the ABI Solid 5500 sequencing machine, consider the following numbers:

Amount of data (not for examination)

Until recently the AMC used two Solid machines in the Genomics Facility. Each sequencer holds a slide containing the sample(s). Slides are divided in 12 lanes which contain different samples if no multiplexing is used.

If both machines do a sequencing run then the following amount of data is produced:

- Compressed raw data (2 machines x 12 lanes on slide x 15 Gb ~ 360 Gigabytes);
- Uncompressed raw data (2 x 12 x 60 Gb ~1.44 Terabytes)
- During analysis (intermediate) results such as alignment produce another 1.2 Terabyte (2 x 12 x 50Gbyte).

Thus in one year we can produce about 60 Terabyte of data with two sequencers. For exome sequencing we used three quarters of a lane, which produces about 75 million reads. This corresponds to about ¾ * (15+60+50) ~94 Gigabyte of data for one exome sequencing sample.

Data format (not for examination)

The current Solid 5500 sequence machines produce 1 XSQ file per lane of the slide. XSQ files have the so-called HDF5 format which is a kind of directory structure. In the computer exercises we will work with data from an older sequencer: the Solid 4 sequencer. This machine produces so called csFasta (color space fasta files) and separate quality files.

2.5 Data pre-processing

The first steps of the data preprocessing involve several steps required to get the sequence data ready for further, more specialized, data analysis such as identification of SNPs (exome sequencing), pathway analysis, and gene expression analysis. These first steps are called 'data pre-processing'. Note that the distinction between pre-processing and further downstream data analysis is to some extent arbitrary and a matter of definition. Here we show several pre-processing steps that are often encountered.

Data conversion

The raw data from the sequencers is first converted to the fastq format. Fastq is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores.

Quality control

Quality scores

Sequencing machines do not only produce sequences but also quality values for every nucleotide. These are often expressed as Phred Scores *Q* which are defined as

$$Q = -\log_{10} p$$

where *p* is the probability of an sequencing error. Table 2.1 shows how different values of the Phred score are interpreted.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

Table 2.1. Phred quality scores and associated base call accuracies.

Figure 2.18 shows the average quality scores at every position of the sequence reads (length of reads = 40). For each read position a Box-Whisker type plot is drawn. The elements of these plots are determined from all reads that were obtained for a sample. The elements of the plot are as follows:

- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality

The y-axis on the graph shows the quality scores. The higher the score the better the base call. The background of the graph divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read. If the qualities at the end of the reads become too low, then one may decide to remove the corresponding nucleotides from the reads, but this may affect the quality of a sequence alignment (why?).

Note: you don’t have to memorize this definition of a Box-Whisker plot for the examination but it is important that you understand what it is representing since this is a very commonly used plot.

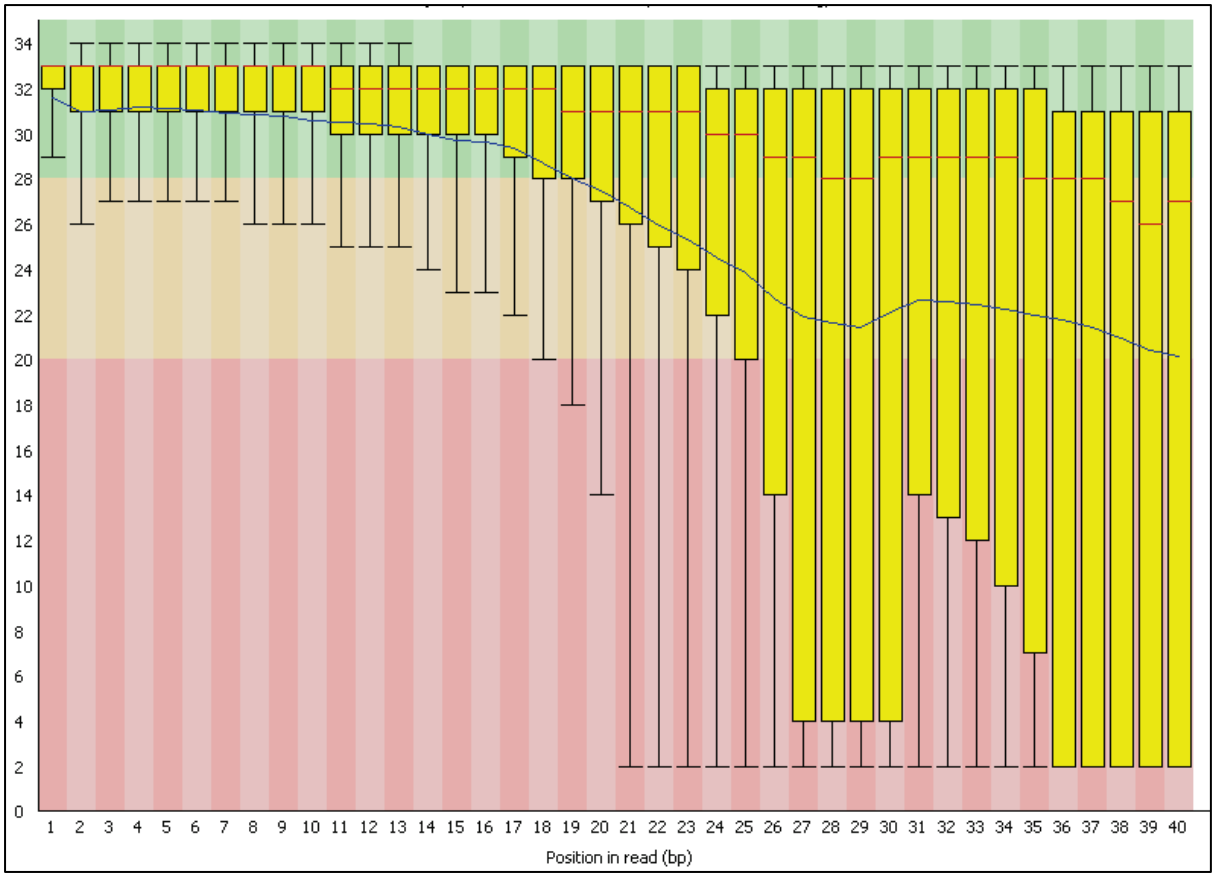


Figure 2.12. Quality control: base qualities across the sequence reads. (Figure copied from http://www.bioinformatics.babraham.ac.uk/projects/fastqc)

Sequence alignment

It is important that you **fully understand the concept of 'sequence alignment'**. If you don't know what a sequence alignment is then Google or https://en.wikibooks.org/wiki/Next_Generation_Sequencing_(NGS)/Alignment

For many applications (including exome sequencing, which in the next chapter) one of the first steps is to perform a sequence alignment of the reads against a reference sequence. Figure 2.13 and Figure 2.14 are examples of sequence alignments. Several tools are available for aligning sequence reads against a reference sequence. It is possible to use BLAST, but other methods such as BWA are better suitable and faster for alignment of short sequence reads from NGS.

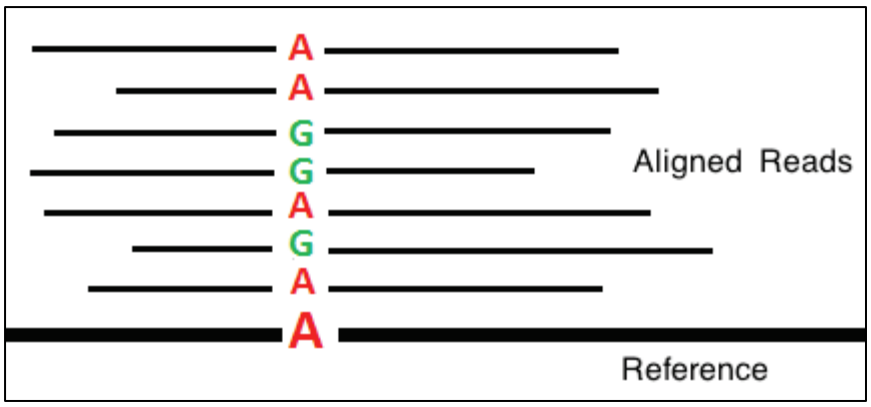


Figure 2.13. Sequence alignment Copied from: https://en.wikipedia.org/wiki/SNV_calling_from_NGS_data

It is important to realize that most NGS technologies produce short sequence reads (~50-150bp), which are more difficult to align than then long sequences obtained with Sanger sequencing. Alignment of a sequence read may result in several outcomes:

- Unique alignment. In this case the sequence read is aligned to exactly one position of the reference sequence. This is the ideal situation.
- Non-unique alignment. Sometimes a read can be mapped on multiple positions of the reference sequence. This may happen because the sequence is too short to have a unique position. Alternatively, the read may correspond to a low-complexity region or pseudogene and is therefore mapped at multiple positions. In general, such reads are removed because we don't know where the read originated from.
- Low confidence alignments. We may also have a unique alignment that has a low quality (confidence) because there were (too) many mismatches with the reference genome.
- No alignment. Sometimes a read cannot be aligned at all. Also these reads are discarded.

2.5.1 Coverage (read depth)

The term 'Coverage' is confusing. Sometimes it refers to the percentage of the DNA (or RNA that is covered by the reads obtained with sequencing. This is called breadth of coverage. However, in most cases we refer to 'read depth' if we talk about coverage.

Coverage (read depth)

Coverage or read depth is the average number of reads representing a given nucleotide in the reconstructed (or reference) sequence. A high coverage is generally desired to increase reliability in the results (e.g., identified SNPs).

The average coverage can be predicted from the length of the original genome (G), the number of reads (N), and the average read length (L) as N*L/G . For example, a hypothetical genome with 2,000 base pairs reconstructed from 8 reads with an average length of 500

nucleotides is expected to have a two-fold coverage. The actual (average) coverage can be determined once the reads are alignment to the reference sequence Figure 2.19.

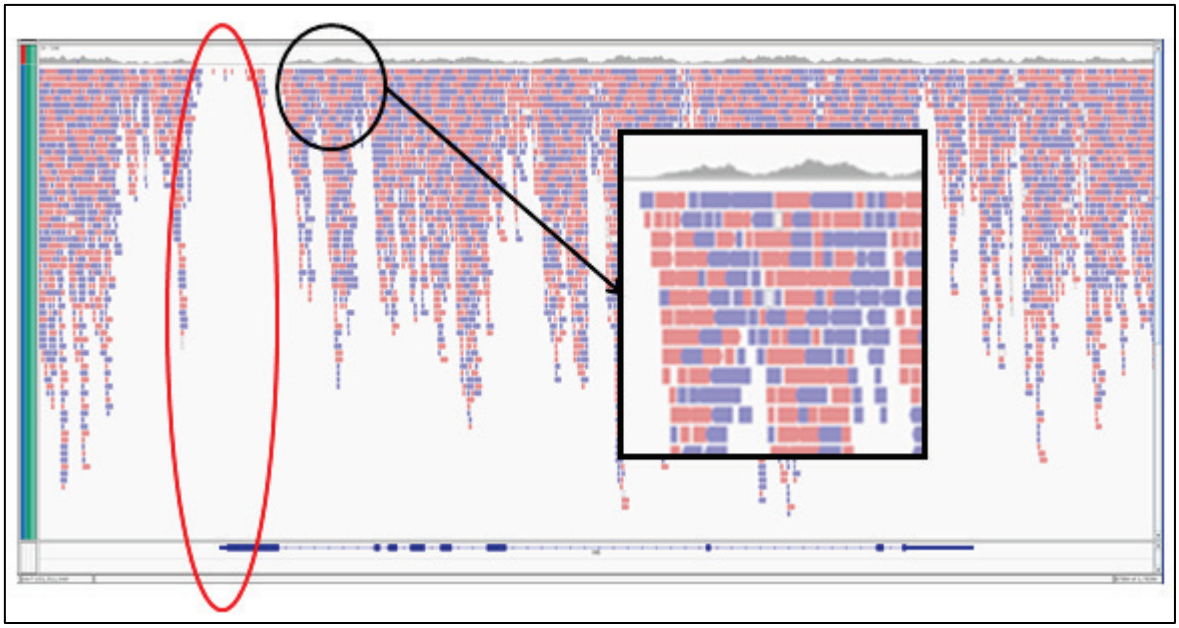


Figure 2.14. Visualization of the coverage of a gene region by sequence reads. At the bottom of the figure the exons (blue boxes) of a single gene are shown. The sequence reads are shown by red and blue arrows (see inset) corresponding to sense and anti-sense alignments. The grey 'mountains' above the sequence reads indicate the coverage at each position. In this figure we see that there is a problem with the coverage of the first exon: almost no reads correspond to this exon. In addition, we note that the coverage is not uniform across this section of the genome. (Figure provided by Martin Haagmans, Genomics Facility, AMC).

2.6 Downstream analyses

Once the sequence data is pre-processed and low-quality sequences (or other artifacts) are removed one can continue with the further application-specific analysis. There are many applications of sequencing (see beginning of this chapter). In the next chapter one specific application will be discussed in more detail: exome sequencing.

2.6 Single-cell sequencing

DNA sequencing has undergone constant improvement since its inception in the 1970s. Today, next-generation sequencing (NGS) approaches are accelerating in speed and decreasing in cost more quickly than Moore's law. DNA sequencing technologies have improved in precision and throughput, and have enabled the sequencing of entire genomes of species and individuals. An increasing number of questions can be addressed by DNA-sequencing-based technologies. In particular, transcriptomic, epigenomic and proteomic analyses are being carried out using methods that reduce a specific analysis problem to a DNA-sequencing problem.

DNA sequencing technology has not only scaled up rapidly in throughput but has also scaled down in terms of the amount of DNA that is required for analysis, to the point at which it is now feasible to **analyze the DNA content of individual cells**[7]. Single cells are the fundamental units of life. Therefore, single-cell analysis is not just one more step towards more-sensitive measurements, but is a jump to a more-fundamental understanding of biology. It opens up a wealth of previously impossible applications in both basic research and clinical science. Examples are: the study of microorganisms that cannot be cultured using direct single-cell genome sequencing; transcriptome analysis of rare, tumor cells circulating in blood; characterization of the earliest differentiation events in human embryogenesis; the investigation of transcriptional noise and stochastic fate choice; and the study of tumor heterogeneity and microevolution. Figure 14 shows examples of processes in cancer that can be studied with single cell sequencing.

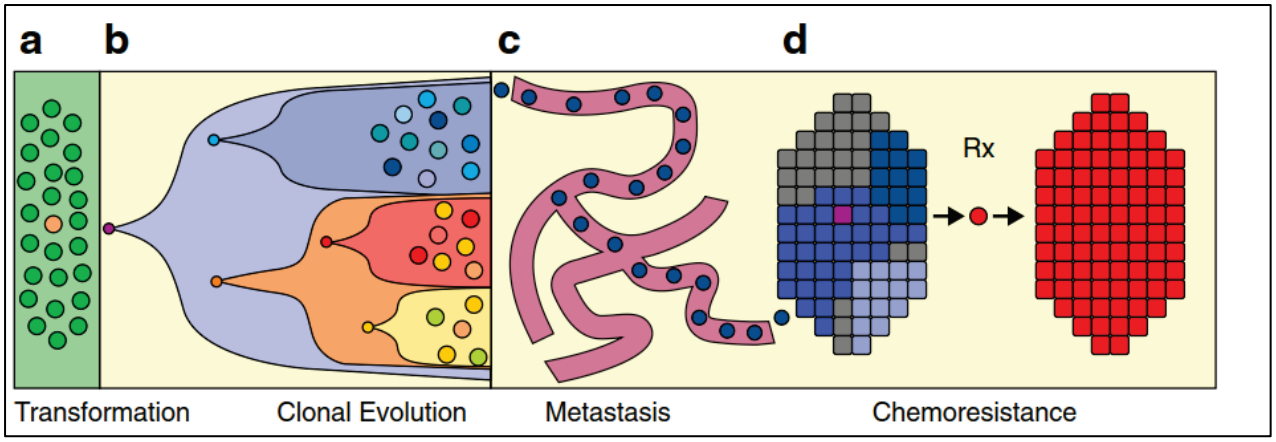


Figure 2.15. Single-cell processes in cancer. Although single cancer cells interact with their neighbors and the adjacent stromal cells, there are many biological processes that occur through the actions of individual cancer cells, shown in this illustration. These complex biological processes in human cancers include: (a) transformation from a single normal somatic cell into a tumor cell; (b) clonal evolution that occurs through a series of selective sweeps when single cells acquire driver mutations and diversify, leading to intratumor heterogeneity; (c) single cells from the primary tumor intravasate into the circulatory system and extravasate at distant organ sites to form metastatic tumors; and (d) the evolution of chemoresistance that occurs when the tumor is eradicated but survived by single tumor cells that harbor resistance mutations and expand to reconstitute the tumor mass. (Figure copied from Navin, N.E. (2014). Cancer genomics: one cell at a time. Genome biology 15, 452).

[7] Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature reviews. Genetics 14, 618-630.

Table 1 | **Summary of NGS platforms**

Platform	Read length (bp)	Throughput	Reads	Runtime	Error profile	Instrument cost (US$)	Cost per Gb (US$, approx.)
Sequencing by ligation							
SOLiD 5500 Wildfire	50 (SE)	80 Gb	~700 M*	6 d*	≤0.1%, AT bias‡	NA§	$130‡
	75 (SE)	120 Gb					
	50 (SE)*	160 Gb*					
SOLiD 5500xl	50 (SE)	160 Gb	~1.4 B*	10 d*	≤0.1%, AT bias‡	$251,000‡	$70‡
	75 (SE)	240 Gb					
	50 (SE)*	320 Gb*					
BGISEQ-500 FCS[155]	50–100 (SE/PE)*	8–40 Gb*	NA‖	24 h*	≤0.1%, AT bias‡	$250 (REF. 155)	NA‖
BGISEQ-500 FCL[155]	50–100 (SE/PE)*	40–200 Gb*	NA‖	24 h*	≤0.1%, AT bias‡	$250,000 (REF. 155)	NA‖
Sequencing by synthesis: CRT							
Illumina MiniSeq Mid output	150 (SE)*	2.1–2.4 Gb*	14–16 M*	17 h*	<1%, substitution‡	$50,000 (REF. 118)	$200–300 (REF. 118)
Illumina MiniSeq High output	75 (SE)	1.6–1.8 Gb	22–25 M (SE)*	7 h	<1%, substitution‡	$50,000 (REF. 118)	$200–300 (REF. 118)
	75 (PE)	3.3–3.7 Gb	44–50 M (PE)*	13 h			
	150 (PE)*	6.6–7.5 Gb*		24 h*			
Illumina MiSeq v2	36 (SE)	540–610 Mb	12–15 M (SE)	4 h	0.1%, substitution‡	$99,000‡	~$1,000
	25 (PE)	750–850 Mb	24–30 M (PE)*	5.5 h			$996
	150 (PE)	4.5–5.1 Gb		24 h			$212
	250 (PE)*	7.5–8.5 Gb*		39 h*			$142‡
Illumina MiSeq v3	75 (PE)	3.3–3.8 Gb	44–50 M (PE)*	21–56 h*	0.1%, substitution‡	$99,000‡	$250
	300 (PE)*	13.2–15 Gb*					$110‡
Illumina NextSeq 500/550 Mid output	75 (PE)	16–20 Gb	Up to 260 M (PE)*	15 h	<1%, substitution‡	$250‡	$42
	150 (PE)*	32–40 Gb*		26 h*			$40‡
Illumina NextSeq 500/550 High output	75 (SE)	25–30 Gb	400 M (SE)*	11 h	<1%, substitution‡	$250‡	$43
	75 (PE)	50–60 Gb	800 M (PE)*	18 h			$41
	150 (PE)*	100–120 Gb*		29 h*			$33‡
Illumina HiSeq2500 v2 Rapid run	36 (SE)	9–11Gb	300 M (SE)*	7 h	0.1%, substitution‡	$690‡	$230
	50 (PE)	25–30 Gb	600 M (PE)*	16 h			$90
	100 (PE)	50–60 Gb		27 h			$52
	150 (PE)	75–90 Gb		40 h			$45
	250 (PE)*	125–150 Gb*		60 h*			$40‡
Illumina HiSeq2500 v3	36 (SE)	47–52 Gb	1.5 B (SE)	2 d	0.1%, substitution‡	$690‡	$180
	50 (PE)	135–150 Gb	3 B (PE)*	5.5 d			$78
	100 (PE)*	270–300 Gb		11 d*			$45‡
Illumina HiSeq2500 v4	36 (SE)	64–72 Gb	2 B (SE)	29 h	0.1%, substitution‡	$690‡	$150
	50 (PE)	180–200 Gb	4 B (PE)*	2.5 d			$58
	100 (PE)	360–400 Gb		5 d			$45
	125 (PE)*	450–500 Gb*		6 d*			$30‡
Illumina HiSeq3000/4000	50 (SE)	105–125 Gb	2.5 B (SE)*	1–3.5 d*	0.1%, substitution‡	$740/$900 (REF. 156)	$50
	75 (PE)	325–375 Gb					$31
	150 (PE)*	650–750 Gb*					$22 (REF. 157)

Table 1 (cont.) | **Summary of NGS platforms**

Platform	Read length (bp)	Throughput	Reads	Runtime	Error profile	Instrument cost (US$)	Cost per Gb (US$, approx.)
Sequencing by synthesis: SNA (cont.)							
Illumina HiSeq X	150 (PE)*	800–900 Gb per flow cell*	2.6–3 B (PE)*	<3 d*	0.1%, substitution‡	$1,000‡,¶	$7.0‡
Qiagen GeneReader	NA‖	12 genes; 1,250 mutations[22]	NA‖	Several days[22]	Similar to other SBS systems[22]	NA‖	$400–$600 per panel[22]
Sequencing by synthesis: SNA							
454 GS Junior	Up to 600; 400 average (SE, PE)*	35 Mb*	~0.1 M*	10 h*	1%, indel‡	NA§	$40,000‡
454 GS Junior+	Up to 1,000; 700 average (SE, PE)*	70 Mb*	~0.1 M*	18 h*	1%, indel‡	$108,000‡	$19,500‡
454 GS FLX Titanium XLR70	Up to 600; 450 mode (SE, PE)*	450 Mb*	~1 M*	10 h*	1%, indel‡	NA§	$15,500‡
454 GS FLX Titanium XL+	Up to 1,000; 700 mode (SE, PE)*	700 Mb*	~1 M*	23 h*	1%, indel‡	$450,000‡	$9,500‡
Ion PGM 314	200 (SE)	30–50	400,000–550,000*	23 h	1%, indel‡	$49‡	$25–3,500‡
	400 (SE)	60–100 Mb*		3.7 h*			
Ion PGM 316	200 (SE)	300–500 Mb	2–3 M*	3 h	1%, indel‡	$49‡	$700–1,000‡
	400 (SE)*	600 Mb–1 Gb*		4.9 h*			
Ion PGM 318	200 (SE)	600 Mb–1 Gb	4–5.5 M*	4 h	1%, indel‡	$49‡	$450–800‡
	400 (SE)*	1–2 Gb*		7.3 h*			
Ion Proton	Up to 200 (SE)	Up to 10 Gb*	60–80 M*	2–4 h*	1%, indel‡	$224‡	$80‡
Ion S5 520	200 (SE)	600 Mb–1 Gb	3–5 M*	2.5 h	1%, indel‡	$65 (REF. 158)	$2,400*
	400 (SE)*	1.2–2 Gb*		4 h*			$1,200*
Ion S5 530	200 (SE)	3–4 Gb	15–20 M*	2.5 h	1%, indel‡	$65 (REF. 158)	$950*
	400 (SE)*	6–8 Gb*		4 h*			$475*
Ion S5 540	200 (SE)*	10–15 Gb*	60–80 M*	2.5 h*	1%, indel‡	$65 (REF. 158)	$300*
Single-molecule real-time long reads							
Pacific BioSciences RS II	~20 Kb	500 Mb–1 Gb*	~55,000*	4 h*	13% single pass, ≤1% circular consensus read, indel‡	$695‡	$1,000‡
Pacific Biosciences Sequel	8–12 Kb[69]	3.5–7 Gb*	~350,000*	0.5–6 h*	NA‖	$350 (REF. 69)	NA‖
Oxford Nanopore MK 1 MinION	Up to 200 Kb[159]	Up to 1.5 Gb[159]	>100,000 (REF. 159)	Up to 48 h[160]	~12%, indel[159]	$1,000*	$750*
Oxford Nanopore PromethION	NA‖	Up to 4 Tb*	NA‖	NA‖	NA‖	$75*	NA‖
Synthetic long reads							
Illumina Synthetic Long-Read	~100 Kb synthetic length*	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500 (possible barcoding and partitioning errors)	No additional instrument required	~$1,000*
10X Genomics	Up to 100 Kb synthetic length*	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500	See HiSeq 2500 (possible barcoding and partitioning errors)	$75 (REFS 72,161)	See HiSeq 2500 +$500 per sample[161]

Approx., approximate; AT, adenine and thymine; B, billion; bp, base pairs; d, days; Gb, gigabase pairs; h, hours; indel, insertions and deletions; Kb, kilobase pairs; M, million; Mb, megabase pairs; NA, not available; PE, paired-end sequencing; SBS, sequencing by synthesis; SE, single-end sequencing; Tb, terabase pairs. *Manufacturer's data. ‡Rounded from Field Guide to next-generation DNA sequencers[160] and 2014 update. §Not available as this instrument will be discontinued or only available as an upgraded version. ‖As this product has been developed only recently, this information is not available. ¶Not available as a single instrument.

3 Exome sequencing

Lecturer: Prof. dr. Antoine van Kampen (AMC)

After reading this chapter you should understand

- Exome sequencing (principle, wet-lab procedure, data analysis)
- Filtering (exome comparison, stratification) approaches in exome sequencing

Contents

3.1 Introduction

Elucidation of the genetic basis of human diseases and other health-related traits has commonly relied on the oversimplified but nevertheless useful dichotomy between 'monogenic, simple and rare' and 'multigenic, complex and common' diseases (Figure 3.1). Primarily through linkage mapping and candidate gene resequencing, loci underlying about one-half to one-third (~3,000) of all known or suspected Mendelian disorders (e.g. cystic fibrosis, sickle cell anemia) have been discovered. However, there is a substantial gap in our knowledge about the genes that cause many (very) rare Mendelian phenotypes. Several factors limit the power of traditional gene-discovery strategies. For example

- Availability of only a small number of cases or families to study;
- Reduced penetrance;
- Locus heterogeneity;
- Substantially diminished reproductive fitness;
- Responsible mutation may be *de novo* mutation (not inherited from parent).

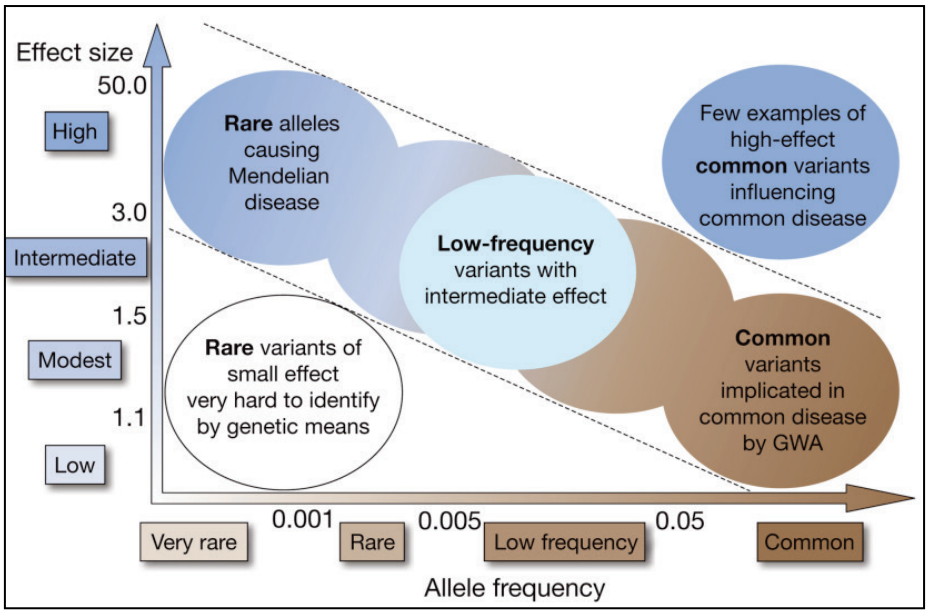


Figure 3.1. With exome sequencing we aim to identify rare alleles causing Mendelian disorders. GWA: Genome Wide Association. (Figure copied from Manolio TA, et al (2009). Nature, 461(7265), 747).

The development of methods for coupling targeted capture and massively parallel DNA sequencing has made it possible to determine cost effectively nearly all of the coding variation present in an individual human genome, a process termed **'exome sequencing'** [1,2]. Whole genome sequencing it is still rather expensive to achieve sufficient sequence reads to be applied on a routine basis (e.g., for diagnostics). Exome sequencing requires that we sequence <2% of the human genome (only the exons). This technique has become a powerful new approach for identifying genes that underlie Mendelian disorders in circumstances in which conventional approaches have failed.

The principle of exome sequencing is shown in Figure 3.2. The exons (colored boxes) of a limited number of patients are sequenced. Subsequently, gene variants (red stars) are identified in the exons. The gene (exon; yellow box) that is affected in all patients is likely to cause the disorder.

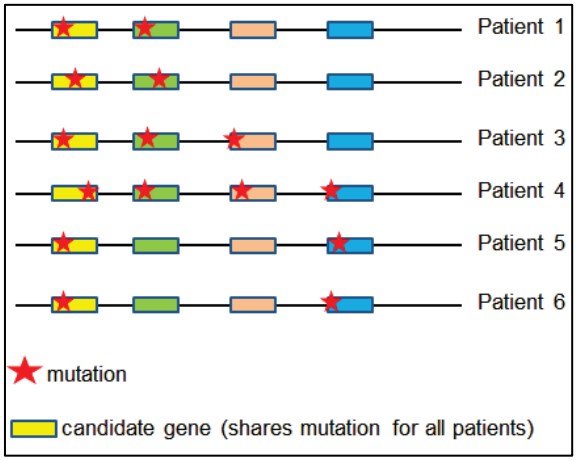


Figure 3.2. Principle of exome sequencing.

3.2 Why exome sequencing?

Despite the fundamental limitation that exome sequencing does not currently assess the impact of non-coding alleles, it is a well-justified strategy for discovering rare alleles underlying Mendelian phenotypes and perhaps complex traits as well:

- Positional cloning studies that are focused on protein-coding sequences have, when adequately powered (sufficient individuals included), proved to be highly successful at identifying variants for monogenic diseases;
- Most alleles that are known to underlie Mendelian disorders disrupt protein-coding sequences;
- A large fraction of rare, protein-altering variants, such as missense or nonsense single-base substitutions or small insertion-deletions (Table 3.1), are predicted to have functional consequences and/or to be deleterious.
- Splice acceptor and donor sites are also enriched for highly functional variation and are also targeted by exome sequencing (although the capture probes represent exons, the captured DNA exon fragments may still contain part of the intron, which results in the sequencing of the intron-exon boundaries).

As such, the exome represents a highly enriched subset of the genome in which to search for variants with large effect sizes.

3.3 Defining the exome

One particular challenge for applying exome sequencing has been how best to define the set of targets (probes for the exon capture) that constitute the exome. Considerable uncertainty remains regarding which sequences of the human genome are truly protein coding. When

sequence capacity was more limiting, initial efforts at exome sequencing were on the conservative side (for example, by targeting the high-confidence subset of genes identified by the Consensus Coding Sequence (CCDS) Project). Commercial kits (e.g., from Nimblegen) now target, at a minimum, all of the RefSeq collection and an increasingly large number of hypothetical proteins. Nevertheless, all existing targets have limitations:

- Our knowledge of all truly protein-coding exons in the human genome is still incomplete, so current capture probes can only target exons that have been identified so far;
- The efficiency of capture probes varies considerably, and some sequences fail to be targeted by capture probe design altogether;
- Not all templates are sequenced with equal efficiency, and not all sequences can be (uniquely) aligned to the reference genome.

These caveats aside, exome sequencing is rapidly proving to be a powerful new strategy for finding the cause of known or suspected Mendelian disorders for which the genetic basis has yet to be discovered.

3.4 Overall wet-lab workflow for exome sequencing

There has been tremendous progress in the development of diverse technologies for capturing arbitrary subsets of a mammalian genome. To capture all protein-coding sequence, the field has largely converged on the aqueous-phase, capture-by-hybridization approach.

The basic steps required for exome sequencing are shown in Figure 3.3.

1. Genomic DNA is randomly sheared, and several micrograms are used to construct an in vitro shotgun library;
2. The library fragments are flanked by adaptors (not shown);
3. The library is enriched for sequences corresponding to exons (dark blue fragments) by aqueous-phase hybridization capture: the fragments are hybridized to biotinylated DNA (orange fragments);
4. Recovery of the hybridized fragments by biotin-streptavidin-based pulldown and washing the non-hybridized fragments.
5. Amplification and massively parallel sequencing of the enriched, amplified library
6. Mapping (alignment) and calling of candidate causal variants. This step comprises the bioinformatics analysis.

Key performance parameters include the

- Degree of enrichment prior to sequencing (using qPCR of control loci prior and after capture).
- Uniformity with which targets are captured (no bias for specific targets)
- Molecular complexity of the enriched library (large diversity of targets)

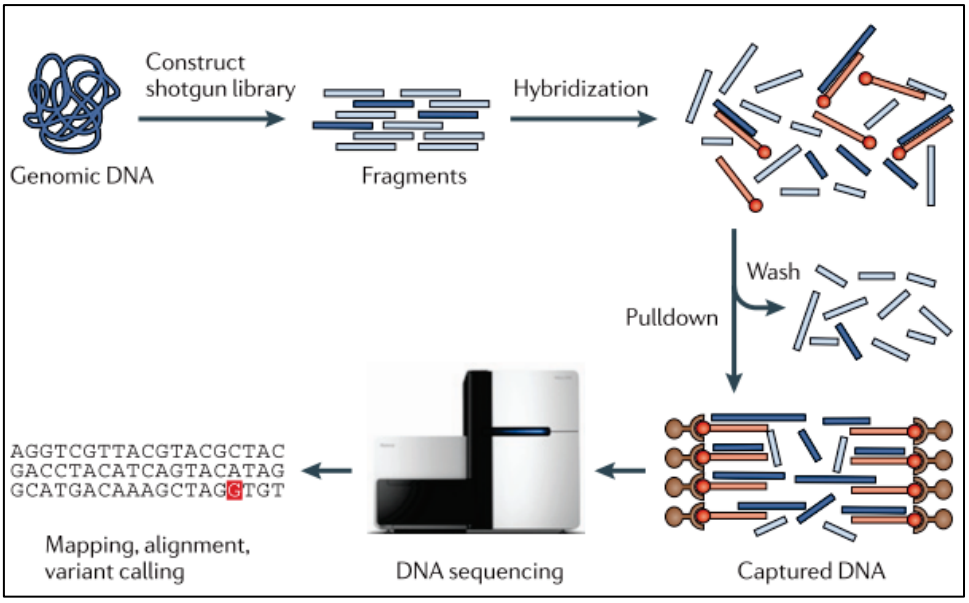


Figure 3.3. Overall workflow for exome sequencing. Dark blue bars represent exon fragments. (Figure copied from Bamshad (2011) Nature reviews. Genetics, 12(11), 745).

3.5 Bioinformatics analysis

The bioinformatics analysis (Figure 3.4) comprises several steps, which are explained below.

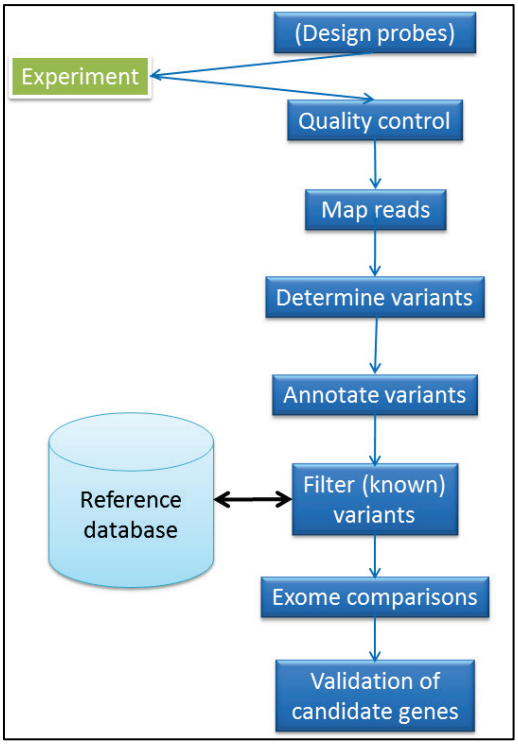


Figure 3.4. Overall bioinformatics workflow for the analysis of exome sequence data.

1. **Probe design**. In this step probes are designed to capture the exon fragments. The challenge is to design probes that are unique for the different exons and which are

efficient during the experiment. Generally, the design and synthesis is done by companies like Nimblegen and Agilent.

2. **Quality Control**. Prior to the analysis the quality of the sequence data is inspected. For example, the nucleotide frequencies should be about equal (~25%) across the sequence reads. The base qualities across the reads should be sufficiently large. (Software: FastQ)

3. **Map reads**. This step is also called 'sequence alignment'. The mapping of the sequence reads (obtained from the DNA fragments (exons) of the patient) against a reference genome is done by specialized algorithms (e.g., Burrows-Wheeler Aligner; BWA which is an alternative for BLAST). The result of such mapping is shown in Figure 3.5. The aim of this mapping is to identify the genomic segments from which the reads were derived. In general, the sequence read (from the patient) will be highly similar to the corresponding genomic region of the reference genome. Mapping is a crucial (and time consuming step) in the data analysis and several problems can occur:
    a. Not all reads can be mapped against the reference. Unmapped reads are discarded.
    b. Some reads are mapped against multiple genomic regions. This reads are also discarded since it is impossible to determine which region (exon/gene) is the correct position.
    c. Some reads can only be mapped with low quality (many mismatches between the nucleotides of the read and the nucleotides of the reference)
    d. Some reads can be mapped uniquely and with high quality to the reference. This is the best situation.

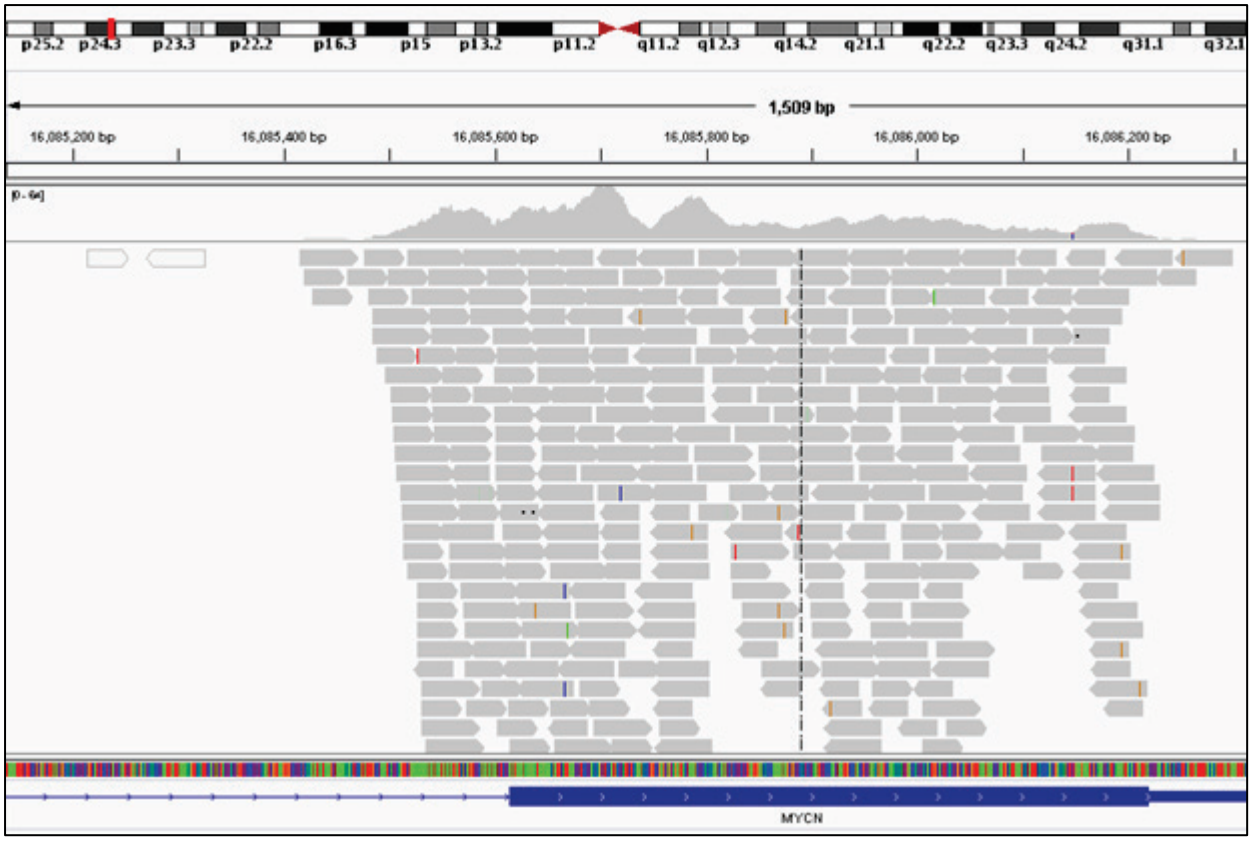


Figure 3.5. Mapping of sequence reads against a reference genome. In the top part (1509 nucleotides) of the reference genome sequence is shown (part of chromosome 2). The red line shown on the chromosome indicates which area is amplified in the panel below. The grey arrow bars are the sequence reads obtained from exome sequence and indicate if a read is aligned to the sense or anti-sense strand of the DNA. The colored lines in these reads indicate the differences (variants) with the reference genome (note that the black line represents the position of the cursor). The grey 'mountains' that are shown above the reads indicate the coverage along this region. Thus a high mountain indicates that many reads were sequenced and aligned at that position. At the bottom one of the exons of the MYCN gene is shown (blue box).

4. **Determine variants**. Once the sequence reads (patient) have been aligned to a reference genome (from the public databases) one can determine all positions at which the nucleotide of the sequence read is different from the reference genome. Such difference is a potential variant (e.g., a SNP or indel) but can also be a sequencing error. Multiple (statistical) approaches have been developed to identify true gene variants. The program 'VarScan' follows a straightforward approach by filtering out all differences that do not fulfill three criteria (Figure 3.6):
    a. At the position of the putative variant we should observe at least *N* reads (default value N=8). Otherwise the coverage at that position is too low for the variant to be trusted.
    b. From these *N* reads at least *K* reads (default value *K*=2) should contain the variant. Thus, if only one read contains the putative variant, it is probably a sequence error.
    c. The average base quality at the position of the variant over all reads should at least be *Q* (detault value *Q*=15). If the quality is too low, then there is a large chance that we are looking at a sequencing error.

In an ideal situation this filtering provides us with only true gene variants.

5. **Annotate variants**. To facilitate further filtering and interpretation of the variants, these are annotated by using the program 'Annovar'. During annotation, each variant is assigned various properties such as gene name, region (intron, exon, splice site), nucleotide position, type of variant (e.g., SNP, indel), type of mutation (e.g., non-synonymous, frameshift), variant (e.g., A is replaced with T), number of reads in which variant is observed, quality, segmental duplications, etc. (see below)

6. **Filter known variants**. Based on the annotation of the variants it is possible to filter out all variants that are unlikely to cause the disorder (non-pathogenic polymorphisms and SNPs related to other disorders). For example, we remove the synonymous variants (why?) and variants that are already present in the public SNP databases (why?) or in an in-house reference database. Such filtering steps may significantly reduce the number of variants at the risk of removing the gene variant of interest.

7. **Exome comparison**. The exome comparison comprises the comparison of the exomes of different patients to find one or more affected genes in each of these patients. Note that the patients do not necessarily have to share the precise same variant within the gene(s). It is sufficient if they share a gene that has a variant in each patient. This will further reduce the number of candidate genes and variants. In case of large heterogeneity one could try to find a subgroup of patients that share an affected gene(s).

8. **Validation of candidate genes**. Once only one or a few candidate genes are left over these can be validated in the wet-lab (e.g., Sanger sequencing) or by computational approaches such as comparison to other sets of exomes and genomes.

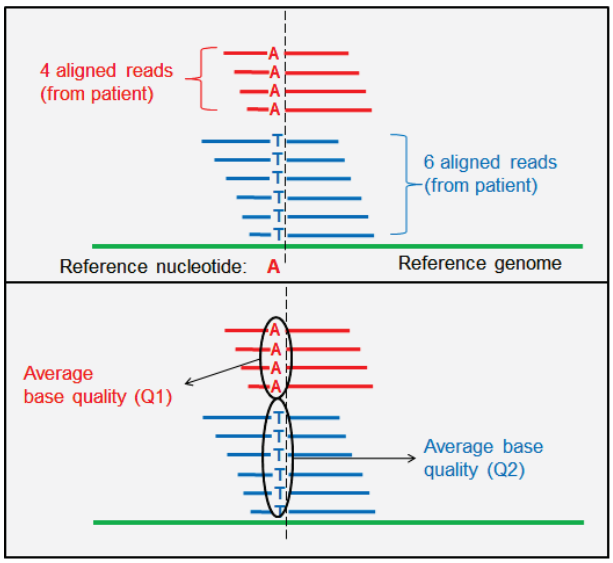


Figure 3.6. Information used by VarScan to detect variants. Here we see 10 reads that are mapped at this specific location of the reference genome. 6 of these read (VarFreq=6/10=0.6) contain a variant (T instead of a A). This is a heterozygous situation in which the patient inherited a different allele from each of the parents. The average base quality is calculated separately for the two sets of sequence reads.

3.6 Identifying causal alleles

A key challenge of using exome sequencing to find novel disease genes for either Mendelian is how to identify disease-related alleles among the background of non-pathogenic polymorphism and sequencing errors. For example, on average, exome sequencing identifies 24,000 single nucleotide variants (SNVs or SNPs: Single Nucleotide Polymorphism) in African American samples and 20,000 in European American samples. More than 95% of these variants are already known as polymorphisms in human populations. Strategies for finding causal alleles against this background vary, depending on factors such as:

- The mode of inheritance of a trait (e.g., dominant, recessive);
    - For recessive disorders, an individual has two variant alleles in a single gene (one from both parents) while for dominant disorders an individual only has a single variant allele in a gene. Therefore, intuitively, exome comparison should be more efficient for recessive disorders (that is, they require sequencing of fewer cases) than for dominant disorders, because the genome of any given individual has around 50-fold fewer genes with two, rather than one, novel protein-altering alleles per gene.
- The pedigree or population structure;
- Whether a phenotype arises owing to de novo or inherited variants;
    - If we know that there are no other disease cases in the family, then the disease is probably not inherited. In such case, the variant that is identified in the patient should not be present in the parents (i.e., a de novo variant).
- The extent of locus heterogeneity for a trait.
    - If a disorder can be caused by mutations in different genes then many more subjects will need to be included in an exome sequencing study.

Such factors influence both the sample size needed to provide adequate power to detect trait-associated alleles and the bioinformatics approach.

The overall approach to identify the causal allele is to filter the data in two directions (Figure 3.7):

1. Discrete filtering: comparing variants among individuals and against public databases and controls;
2. Stratification of variants.

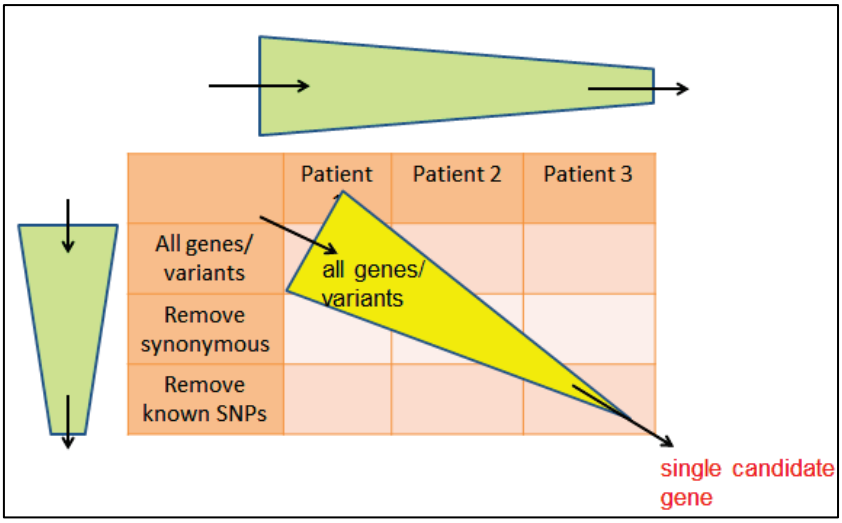


Figure 3.7. Filtering of exome data to identify the causal allele. The top horizontal green funnel shows the reduction of variants by comparison of several patients (and controls). The vertical green funnel shows the stratification of variants (i.e., the removal of variants according to different criteria such as 'known SNPs' from the public databases). The combination of both funnels (filters) will largely reduce the number of candidate genes, and hopefully results in a single candidate that can be validated.

3.6.1 Filtering: comparing variants among individuals and against public databases and controls

Most of the exome studies have, to varying extents, relied on comparisons with exome sequences and variants that are found in a small number of unrelated or closely related affected individuals to find rare alleles or novel alleles in the same gene shared among affected individuals (Figure 3.8).

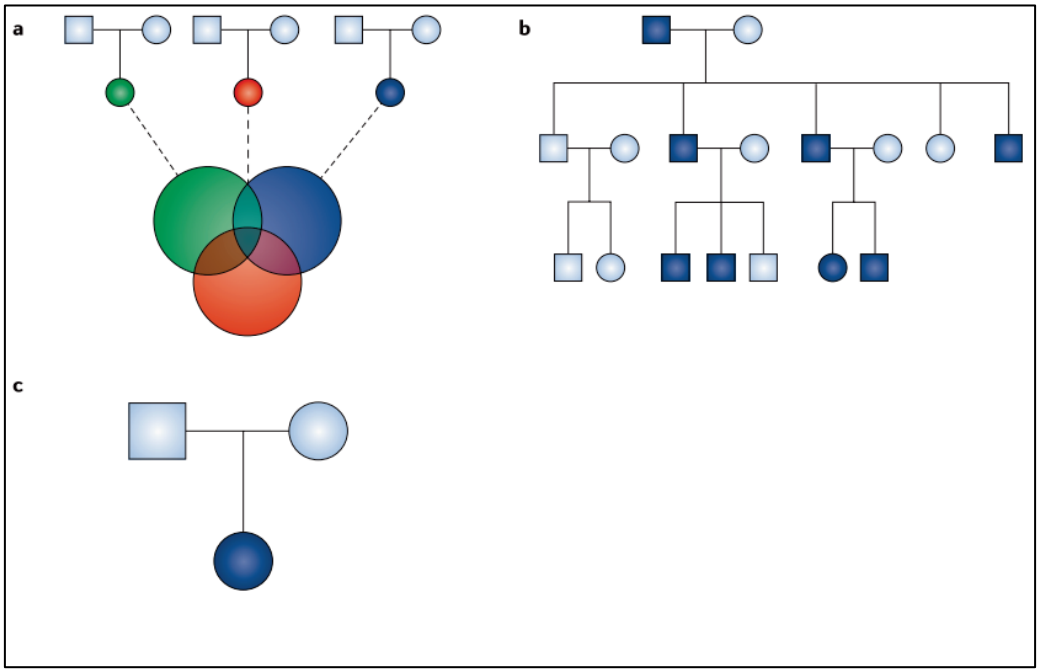


Figure 3.8. Strategies for finding disease-causing rare variants using exome sequencing. The main strategies are illustrated. **A**. Sequencing and filtering across multiple unrelated, affected individuals (indicated by the three colored circles). This approach is used to identify novel (de novo) variants in the same gene (or genes), as

indicated by the shaded region that is shared by the three individuals in this example. **B.** Sequencing and filtering among multiple affected individuals from within a pedigree (shaded circles and squares) to identify a gene (or genes) with a novel variant in a shared region of the genome. **C**. Sequencing parent-child trios for identifying de novo mutations.

In these cases, novelty is assessed by filtering the variants:

- Against a set of polymorphisms that are available in public databases (for example, dbSNP and 1000 Genomes Project) and/or
- Those found in a set of unaffected individuals (that is, controls).

This filtering step is used to eliminate candidate genes by assuming that any allele found in the 'filter set' cannot be causative (e.g., we are dealing with (very) rare disorders, thus the chance that the causative SNP is in any of the public databases or controls is very small).

This approach is powerful in part because only a small fraction (~2% on average) of the SNVs identified in an individual by exome sequencing is novel. The sequencing of only a modest number of affected individuals, and then applying filtering to the data to reduce the number of candidate genes to a minimum number of high priority candidates (if not a single one) is an important advantage that exome-sequencing approaches have over conventional approaches.

Underlying this method is the assumption that the control set contains no alleles from individuals with the phenotype being studied. This assumption can be problematic for two reasons.

- dbSNP is 'contaminated' with a small but appreciable number of pathogenic alleles.
- As the number of sequenced exomes and genomes increase we will filter out an increasing number of variants that occur in these control exomes/genomes and are, therefore, assumed not to play a role in the rare disease under investigation. However, filtering in this manner that does not take into account that some pathogenic alleles have higher minor allele frequency (MAF). That is, the pathogenic gene variant that we are hunting for may also occur in the control exomes/geneomes. Thus, simply filtering all known variants runs the risk of eliminating truly pathogenic alleles. This risk is especially relevant for recessive disorders (e.g. cystic fibrosis), in which carrier status of such more common allele will not result in a phenotype that might otherwise exclude an individual from the 'control' population of exomes/genomes (for example, 1000 Genomes Project). Thus, instead of removing all known variants a better strategy might be to remove the variants with a MAF above a certain threshold. However, because we will remove fewer known variants, this may result in more false positives. Consequently, larger sample sizes (more patients) may be needed.

3.6.2 Stratifying candidates

By mutation type

Candidate alleles can be stratified on the basis of their predicted impact or deleteriousness. Alleles can be stratified by their mutation type by giving greater weight to frameshifts, stop codon mutations and disruptions of splice sites than to missense variants (see Table 3.1 for definitions). Synonymous mutations that do not change the amino acid are removed.

By segmental duplications

Segmental duplications [3] are segments of DNA (1 to 400kb in length) with near identical (>90% sequence similarity) sequences. Sequences reads (from the patients) that are similar to these segments will be aligned to one or both segments. In both cases, in case of a variant, it will be difficult to decide to which segment (and thus which gene in our case!) a variant belongs. Therefore, variants found in segmental duplications are discarded.

By pseudogenes

Pseudogenes are dysfunctional relatives of genes that have lost their protein-coding ability. Reads aligned to pseudogenes are discarded (for reasons similar to 'segmental duplications').

By function

Candidate alleles can be stratified by existing biological or functional information about a gene: for example, its predicted role (or roles) in a biological pathway or its interactions with genes or proteins that are known to cause a similar phenotype.

By functional impact (sequence conservation)

Exploit the observation that regions of genes in which mutations are deleterious tend to show high sequence conservation. Such approaches that stratify non-synonymous alleles (for example, SIFT and PolyPhen) identify functional sites at which observed variants are more likely to affect the protein and thus the phenotype (i.e., cause a disorder)

Mutation type	Description
Missense	This type of mutation is a change in one DNA base pair that results in the substitution of one amino acid for another in the protein made by a gene.
Nonsense	A nonsense mutation is also a change in one DNA base pair. Instead of substituting one amino acid for another, however, the altered DNA sequence prematurely signals the cell to stop building a protein (stop gain mutation). This type of mutation results in a shortened protein that may function improperly or not at all. Alternatively, there may be a stop loss mutation (loss of stop codon)
Insertion	An insertion changes the number of DNA bases in a gene by adding a piece of DNA. As a result, the protein made by the gene may not function properly.
Deletion	A deletion changes the number of DNA bases by removing a piece of DNA. Small deletions may remove one or a few base pairs within a gene, while larger deletions can remove an entire gene or several neighboring genes. The deleted DNA may alter the function of the resulting protein(s).

InDel	(small) insertion or deletion
(segmental) duplication	A duplication consists of a piece of DNA that is abnormally copied one or more times. This type of mutation may alter the function of the resulting protein.
Frameshift	This type of mutation occurs when the addition or loss of DNA bases changes a gene's reading frame. A reading frame consists of groups of 3 bases that each code for one amino acid. A frameshift mutation shifts the grouping of these bases and changes the code for amino acids. The resulting protein is usually nonfunctional. Insertions, deletions, and duplications can all be frameshift mutations.

Table 3.1. Types of DNA mutations (copied and adjusted from ghr.nlm.nih.gov/handbook/mutationsanddisorders). See also Chapter 15 (Gene mutation and molecular medicine) in Sadava D, Hillis DM, Heller HC, Berenbaum MR: Life. The science of biology. Sunderland, Mass.: Sinauer Associates, Inc; W.H. Freeman and Company; 2013, Ninth Edition.

3.7 Limitations

The most successful reports of the identification of a novel disease gene by exome sequencing have relied on filtering, often with the aid of mapping data (from linkage studies or GWAS). However, it is difficult to know how often this approach has failed, as negative results are rarely reported. Failure can result for many reasons:

Technical failure can occur because:

- Part or all of the causative gene is not in the target definition (for example, it is not known to be a gene or there is a failure in the probe design);
- There is inadequate coverage of the region that contains a causal variant (for example, because of poor capture or poor sequencing);
- The causal variant is covered but not accurately called (for example, in the presence of a small but complex indel;
- False variants in a gene are called because of mismapped reads or errors in alignment.

Improvements on current methods to overcome these weaknesses are being investigated, so technical failures are likely to diminish rapidly over the next few years.

Failures can also follow from the limitations and assumptions of filtering. Perhaps the major limitation of discrete filtering is that its power is substantially reduced by genetic heterogeneity. For example, if alleles of one gene account for only a fraction of cases, no single gene will be found to have disease-causing alleles in all cases. In addition, false-positive calls will result in detection of candidate genes that cannot logically be eliminated by filtering alone. False-positive calls are frequently observed in segmental duplications and processed pseudogenes (for example, CDC27).

3.8 Application of (exome) sequencing to clinical diagnostics

Discovery of variants that underlie Mendelian (and complex traits) will naturally lead to a much deeper understanding of disease mechanisms that should, in turn, facilitate development of improved diagnostics, prevention strategies and targeted therapeutics. Some of these improvements will soon be realized (for example, better diagnostics for Mendelian disorders and disorders of unknown etiology), whereas others (e.g., drug development) are likely to be a more distant realization.

3.8.1 Research

Exome sequencing was initially used in research, especially to detect genes in Mendelian disorders. Various strategies were used but the main ones are:

1. Autosomal dominant disorders: performing linkage analysis within a single family of sufficient size to allow for linkage studies; subsequently exome sequencing in which the regions that were linked within the family to the phenotype were analyzed
2. Autosomal recessive disorders: mainly consanguineous families were used; in these SNP arrays were used to search for regions of which the affected individuals were homozygous; subsequent exome sequencing of either the whole exome or only the regions that show homozygosity, to detect pathogenic homozygous variants
3. *De novo* variants in individuals in which the phenotype allows reliable recognition of individuals with the same disorder: gathering at least 4 individuals with this phenotype, and perform whole exome sequencing in all four, assuming that they should all have a variant in exomes of one and the same gene
4. *De novo* variants in individuals with unknown entities: perform exome sequencing in both parents and the affected individual, assuming that a pathogenic variant that is present only in the affected individual and not in the parents might be causing the phenotype.

In each of the above strategies exome sequencing has proven its value. However, there is one major caveat: finding a pathogenic variant in a limited number of individuals is often not considered to be sufficient proof that the gene indeed is causing the disorder. Usually either a large number of individuals should have a variant in the same gene that the chance this is occurring by coincidence is small, or one has to perform functional analysis showing that the variant is causing such a functional abnormality that this is likely to cause the phenotype. Often exome sequencing is performed in infrequently occurring disorders, and, thus, gathering large series of affected individuals may be extremely difficult. On the other hand functional studies are not possible for each gene and functional studies are expensive, which may make it difficult to perform the studies especially if it is a rare disorder. This is recognized by the international genetic community and therefore initiatives are at present running to establish a world-wide database containing both results of fine phenotyping of patients and the results of exome sequencing, with the goal to facilitate finding more than a single patient

with comparable phenotype and a variant in the same gene (a combination of the Human Variome Project and the Human Phenome Project).

A major challenge for clinicians is making a specific diagnosis in individuals with novel phenotypes or those with phenotypes that are difficult to differentiate into etiologically distinct categories (for example, autism or global developmental delay). Recent applications of exome sequencing to identify de novo variants in children with idiopathic (arising spontaneously or from an unknown cause) intellectual disabilities and children with sporadic (not familial) autism (caused by *de novo* mutations) suggest that such phenotypes could be tractable to genome-wide screening for protein-coding variants that are predicted to have deleterious effects. But the major problem also in these groups of patients remains to proof that the finding is truly causative. Especially for these groups the international initiative for a combined genome and phenome database will be essential.

3.8.2 Diagnosis

The main difference between research and clinical diagnostics is that the result of the exome sequencing itself should be absolutely reliable in an individual. This may cause significant problems in some applications of exome sequencing in patient care.

One of the main applications of exome sequencing in the near future will be facilitating the accurate diagnosis of individuals with well-known Mendelian disorders. Patients visiting their physicians with a complaint of which it is known that this can be caused by disorders that are in part or completely caused by genetic changes (such as cataract, diabetes mellitus, cancer or hypertension), are likely to be investigated first by exome sequencing. This way the genetic component or even the complete cause can be detected. In the exome sequencing not all exomes are analyzed: only the exomes of genes of which it is known they can cause the disorder will be analyzed by the bioinformatician (targeted analysis). The reason for this approach is that it decreases significantly finding variants in other genes that have nothing to do with the complaint (secondary findings). We foresee that this application will be extremely widely used in medical care in just a limited number of years. It will also facilitate diagnosing individuals with atypical disorders or with disorders that otherwise would ask for extensive and/or invasive procedures to detect the cause.

For example, exome sequencing was used to discover a novel Cys203Tyr variant in X-linked inhibitor of apoptosis (XIAP) in a young boy with severe inflammatory bowel disease in whom a definitive diagnosis was elusive, despite a comprehensive evaluation (i.e., Nicholas Volker story; http://www.jsonline.com/features/health/111224104.html). Mutations in XIAP are a known cause of X-linked lymphoproliferative syndrome type 2 (XLP2), but severe colitis is an unusual symptom of XLP2. Furthermore, the diagnosis of XLP2 suggested a specific course of treatment (namely, allogeneic haematopoietic progenitor cell transplant) that had not been considered previously and appears to have been, at least in the short term, successful.

3.8.3 Challenges

Widespread, useful, convenient and cost effective use of exome sequencing and eventually whole-genome sequencing for clinical diagnosis or screening will necessitate overcoming a number of major challenges that currently limit broad applicability. These challenges can be divided into those that are related to technical considerations and those that pose challenges to implementation.

There are several technical hurdles.

- Sequencing and genome assemblies will have to be highly accurate to avoid misdiagnosis;
- Algorithms for annotating variants will need to be automated:
    - Approaches for characterizing the functional impact of rare and novel variants will have to be improved.
    - Relevance to disease-related risk will require comparison to a well-curated catalogue of variants that are known to influence risk of disease.
    - General databases, such as the Human Gene Mutation Database (HGMD), and locus-specific databases are currently used with caution. Efforts are therefore underway to create comprehensive collections of validated associated variants (for example, the Human Variome Project).
- Strategies for interpreting the use of variants (for example, clinical, reproductive or personal use) need to be developed and tested.
- Standards and guidelines for exome or whole-genome sequence testing and reporting in clinical laboratories will need to be established.

On a national level it will be essential to determine the specific scenarios in which personal exome or whole-genome data are useful in prevention and/or clinical management (for example, diagnosis and treatment) in such a way that benefits outweigh costs. Another challenge is the need to train health-care providers to incorporate genomic information into their practice. Similarly it is unclear how the limited number of geneticists and genetic nurses in the Netherlands will be able to effectively communicate such a volume and scope of exome and genome sequencing results to affected individuals and their families. Furthermore, the interpretation of information will change over time as new risks are reported and others are refuted, as the magnitude of risks change and as interactions among variants and interactions with environmental factors are discovered. Despite all of these challenges, exome (and in the future genome) sequence information will eventually become part of the routine clinical evaluation of all persons suspected of having a disorder with a significant genetic component and may eventually be used to provide personalized health-care profiling.

3.9 Ethical considerations

The use of exome sequencing for disease gene discovery generates new manifestations of several long-standing ethical issues in human genetics research. Two areas of research ethics that require consideration in particular are the limitations of the current consent process and management of individual research results. In addition, there are several other important ethical issues that should be considered in the context of exome-sequencing studies but these are not discussed here. They include issues surrounding data sharing, the return of test results to individuals over time and the necessity of validation of exome-sequencing protocols.

3.9.1 Informed consent for exome sequencing

There are several important ethical challenges in exome-sequencing research related to informed consent. Many studies that incorporate exome sequencing may use banked samples collected using consent documents that did not specifically anticipate, let alone describe, exome sequencing. This raises questions about what type of information is needed to make informed decisions about participation in exome-sequencing research and whether and how this information differs from standard information about the risks and benefits of genetic research. In many cases, the answers are complex and contextual. Furthermore, in many ways, the goals of exome sequencing are similar to the targeted sequencing approaches already applied in genetic analysis. However, there are possible risks that can be considered:

- Exome-sequencing approaches increase the chance of uncovering clinically useful results that are unrelated to the primary aim of the study (for example, identification of a disease gene). The need to describe the increased chance of returning results to the tested individual will need to be balanced with the desire to avoid unrealistic participant expectations and potential therapeutic misconception about possible benefits of participation.
- The risks of sharing individual-level genotype or raw sequence data from exome sequencing studies in databases should be assessed, which is essential for the development of a set of appropriate data-sharing policies and protections.

3.9.2 Return and management of results

Researchers and policy makers continue to struggle to develop a framework and guidelines for the return of results from genetic studies. Although there is not a clear consensus, several practices have emerged that generally minimize the need to return results unless:

- they are identified in the course of routine research analysis;
- they have been validated;
- they are determined to be clinically useful and to be actionable.

Exome sequencing also challenges existing assumptions for the return of results, particularly regarding the nature of so-called 'incidental findings'. It is no longer a question of whether

clinically useful results will be found in any research participant, but rather how many such results will be identified in each participant.

Given the new realities of exome sequencing, researchers who consider returning results will have to contend with several major issues:

- They will need to identify 'known' variants that are associated with health-related traits and interpret their clinical importance.
- They will need to consider what kinds of health-related results (for example, carrier status, cancer predisposition or drug response) to return to participants.
- Researchers will need to consider participant expectations about re-contact and develop an ethically appropriate, context-specific plan for the return of results.
- Whereas the return of results by conventional means (for example, face-to-face genetic counseling) is the gold standard, it is also expensive, especially given the added cost imposed by returning a larger number of results. Although exome sequencing will identify a much larger number of clinically useful results than other genetic research approaches, it does not follow that there ought to be a mandatory review and return of all such results to participants in all studies. The decision about whether and how to return results must take into account factors such as: the commitments made at the time of informed consent and the resources available both to analyze variant data

3.10 Future directions

Because of our poor ability to make sense of non-coding variation, the analytical components of most 'whole genome' studies have disproportionately focused on variation within the exome. As the cost of sequencing continues to fall, the field will probably gradually move from exome to whole-genome sequencing. However, taking advantage of these more comprehensive data for disease gene discovery and molecular diagnostics in patients crucially depends on the development of analytical strategies for making sense of non-coding variation. This is as much an opportunity as it is a challenge.

There are several specific areas in which focused efforts are likely to advance the field substantially. These include:

- The proper curation of phenotypes, particularly in the context of Mendelian disorders. To this end, there are hundreds, perhaps thousands, of poorly defined familial phenotypes that are rare or unique. Development of repositories in which descriptive information about such phenotypes and an accompanying DNA sample could be banked by clinicians would facilitate both delineation of new Mendelian disorders and discovery of the underlying genes.
- We need improved technical, statistical and bioinformatic methods for: reducing the rate of false-positive and false-negative variant calls; calling indels; prioritizing

candidate causal variants; and predicting and annotating the potential functional impact for disease gene discovery or molecular diagnostics.

- To realize fully the potential of sequencing for clinical diagnostics and personal genomic profiling, we need to address the challenges posed by ethics and policy issues. Nevertheless, exome and even genome sequencing are likely to be introduced in the clinical setting before these challenges are fully resolved owing in part to their ability to facilitate diagnosis and inform therapy.

3.11 References

1. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272-276.
2. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nature genetics 42: 30-35.
3. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al. (2005) Segmental duplications and copy-number variation in the human genome. American journal of human genetics 77: 78-88.

4 Transcriptome and Transcriptomics

Lecturers: **Dr. Martijs Jonker (RNA Biology and Applied Bioinformatics, UvA)**
Dr. Rob Dekker (RNA Biology and Applied Bioinformatics, UvA)

After reading this chapter you should understand

- To summarize the different types of RNA and their function.
- To give examples of a biological research questions that can be studied using transcriptomics [further explained in the lectures].
- To explain the principles of RNA-seq and microarray analyses.
- To describe why and how Principal Components Analysis is used in transcriptome analysis, and to interpret the results [further explained in the lectures].
- To explain how differentially expressed genes are identified and why multiple testing correction is applied [further explained in the lectures].
- To describe gene set testing.
- To describe at least two data bases or webtools that help with the biological interpretation of transcriptomics results.

Please note that these are further specifications of the Teaching and Learning Goals formulated in the Course Manual.

Sources:

- Korpelainen E, Tuimala J, Somervuo P, Huss M, Wong G. RNA-seq Data Analysis. A Practical Approach, ISBN 978-1-4665-9500-2. 2015. Boca Raton: Chapman & Hall/CRC.
- Draghici S. Data Analysis Tools for DNA Microarrays, ISBN 1-58488-3.5-4. 2003. Boca Raton: Chapman & Hall/CRC.
- Gibson G, Spencer VM. A Primer of Genome Science, third edition, ISBN 978-0-87893-236-8. 2009. Sunderland, Massachusetts: Sinauer Associates, Inc Publishers.
- Cristianini N, Hahn MW. Introduction to Computational Genomics: A Case Studies Approach, ISBN 9780521856034. 2006. Oxford University Press
- Brown TA. Genomes 3 ISBN 9780815341383. 2006. Garland Science: New York.

Contents

4.1 TRANSCRIPTOME & TRANSCRIPTOMICS

4.1.1 What is the transcriptome?

The central dogma of molecular biology describes the flow of genetic information from genes to proteins. This comprises a two-step process: first, DNA, the permanent, heritable, genetic information repository, is transcribed by the RNA polymerase enzymes into RNA, a short-lasting information carrier; second, a subset of RNA, the messenger RNAs, mRNAs, are translated into protein. Atypical mammalian cell contains 10–30 picogram RNA representing only 1% of the cell as a whole. The assortment of **all RNA molecules** (also called transcripts) that are present in the cell at a given point in time is called the **transcriptome**, *i.e.* all that has been transcribed from the nuclear and mitochondrial genome. The transcriptome of a cell can thus be seen as a highly diverse repertoire of RNA molecules that have been transcribed from their respective genes in that particular cell. It plays a central role in establishing cell type and the ability of cells to adequately respond to changes in its environment.

The best way to understand the RNA content of a cell is to divide it into categories and subcategories depending on function. The primary division is between **coding RNA** and **noncoding RNA**. The coding RNA is made up of just one class of molecule, the **messenger RNAs** (**mRNAs**), which are transcripts of protein-coding genes and hence are translated into protein in the second stage of genome expression (translation). The total amount of mRNA in a cell typically makes up less than 5% of the total RNA content.

The second type of RNA is referred to as **noncoding** as these molecules are not translated into protein. There are several diverse types of noncoding RNA, the most important being as follows:

Ribosomal RNAs (**rRNAs**) are present in all organisms and are usually the most abundant RNAs in the cell, making up over 90% of the total RNA. These molecules are components of ribosomes, the structures on which protein synthesis takes place.

Transfer RNAs (**tRNAs**) are small molecules that are also involved in protein synthesis and, like rRNA, are found in all organisms. The function of tRNAs is to carry amino acids to the ribosome and ensure that the amino acids are linked together in the order specified by the nucleotide sequence of the mRNA that is being translated.

Small nuclear RNAs (**snRNAs**) are found in the nuclei of eukaryotes. These molecules are involved in splicing, one of the key steps in the processing events that convert the primary transcripts of protein-coding genes into mRNAs.

Small nucleolar RNAs (**snoRNAs**) are found in the nucleolar regions of eukaryotic nuclei. They play a central role in the chemical modification of rRNA molecules by directing the enzymes that perform the modifications to the specific nucleotides where alterations, such as addition of a methyl group, must be carried out.

MicroRNAs (miRNAs) and **short interfering RNAs (siRNAs)** are small RNAs that regulate the expression of individual genes by specifically binding and degrading their respective mRNA sequences.

Non-coding RNAs can and often do play roles in human diseases such as cancer, cardiovascular, and neurological disorders. While the study of the transcriptome is most commonly applied to the mRNAs, it also provides important data regarding the content of the cell's noncoding RNAs. miRNAs, for example, are of particular interest for studying human disease.

4.1.2 Gene expression

Gene expression is defined as the conversion of information encoded in a gene into protein or noncoding RNA structures, *i.e.* synthesis of a functional final gene product. Thus, expressed genes include genes that are transcribed into mRNA and then translated into protein, as well as genes that are transcribed into RNA but not translated into protein, *e.g.* ribosomal and transfer RNA. Because mRNAs play an intermediary 'messenger' role between genes and proteins, the level of an mRNA is commonly considered as a relative indicator of gene expression. Of course there are exceptions were the above does not hold true. If there is more mRNA from a particular gene, this does not necessarily mean that it is being translated as efficiently into protein as before, or that the protein is processed, folded or degraded at the same rate. However, transcriptome studies indicate that a relative change in the mRNA level of a gene brings about a similar change in the expression of the protein gene product.

mRNA is relatively short-lived, its degradation being highly regulated. Most eukaryotic mRNAs are degraded a few hours after synthesis. This rapid turnover means that the composition of the transcriptome is not fixed and can quickly be restructured by changing the rate of synthesis of individual mRNAs. The level of a particular mRNA is always the resultant of *de novo* gene transcription and RNA degradation. A transcript that is present at a constant level in time has a balanced synthesis and degradation rate. Lowering the rate of transcription for a given gene will disturb the current transcription-degradation balance because the existing transcript is continuously being degraded without being replenished at a similar rate, thereby lowering the mRNA level in the cell. Conversely, increasing transcription will result in an increased mRNA level until a new balance has been set.

Although the mRNA fraction of the transcriptome usually makes up less than 5% of the total cell RNA, it is the most significant component because it contains the coding RNAs that are used in the next stage of genome expression, *i.e.* it codes for the proteome. It is important to note that the transcriptome is never synthesized *de novo*. Every cell receives part of its parent's transcriptome when it is first brought into existence by cell division, and (dynamically) maintains a transcriptome throughout its lifetime. Even quiescent cells in bacterial spores or in the seeds of plants have a transcriptome, although translation of that transcriptome into protein may be completely switched off. Transcription of individual protein-coding genes does not, therefore, result in *synthesis* of the transcriptome but instead *maintains* the transcriptome by replacing mRNAs that have been degraded, and brings about *changes* to the composition of the transcriptome via on-off switching of different sets of genes.

Even in the simplest organisms, such as bacteria and yeast, many genes are active at any one time. Transcriptomes are therefore complex, containing copies of hundreds, if not thousands, of different mRNAs. The transcriptome is thus diverse in two ways: (1) it comprises mRNAs with differing sequences, transcribed from different genes (*sequence diversity*); and (2) each of these unique mRNA sequences can be present in different quantities (*quantitative diversity*). The specific combination of many different RNA molecules and their respective quantities constitutes a unique transcriptome that determines a cell's phenotype. For instance, some genes are only being actively transcribed in cells that have highly specialized biochemistries, reflected by transcriptomes in which one or a few mRNAs are exclusive to the cell or tissue type (tissue-specific expression). Beta cells in the pancreas, for example, are the only cells in the human body that actively transcribe the insulin gene and thus contain the mRNA that codes for insulin.

4.1.3 What is transcriptomics?

Transcriptomics is defined as the study of (changes in) the composition of the transcriptome, under

specific circumstances or in a specific cell, using high-throughput methods. These methods enable a researcher to measure the relative expression levels of all RNAs in the transcriptome simultaneously; an approach that is also known as *transcription profiling*. Transcription profiling, which thus follows changes in behavior of a cell *in toto*, not of a single gene or just a few genes, is used throughout diverse areas of biomedical research, including disease diagnosis, biomarker discovery, risk assessment of new drugs or environmental chemicals etc. Transcription profiling can be applied to loss- and gain-of-function mutants to identify the changes associated with the mutant phenotype. The transcriptomic techniques have been particularly useful in identifying the functions of genes. Transcriptomics also allows identification of pathways that respond to or ameliorate environmental stresses. Moreover, relatively new transcription profiling techniques such as RNA-seq (more about this later) can also identify disease-associated gene fusions, single nucleotide polymorphisms and even allele-specific expression.

4.1.4 How can transcriptome data be used?

Transcriptomics is most commonly used to compare specific pairs of samples. The differences may be due to different external environmental conditions, e.g., hormonal effects or toxins. More commonly, healthy and disease states are compared. For example, in cancer, transcriptomics analyses address classification, the mechanisms of pathogenesis and even outcome prediction. Transcriptome studies can classify cancer beyond anatomical location and histopathology. Outcome predictions can establish gene-based benchmarks to predict tumor prognosis and therapy response. These approaches are already in use for personalized medicine, individualized cancer patient therapies.

Organisms and tissues at various stages of development can be molecularly characterized. The transcriptomes of stem cells help to understand the processes of cellular differentiation or embryonic development. Because of its very broad approach transcriptome analysis is a great source for identifying targets for treatment. By analyzing the entire collection of RNAs in a cell researchers can determine when and where each gene is turned on or off (or anything in between) in the cells and tissues of an organism. Depending on the technique used, it is even possible to "count the number of transcripts" relatively by sequencing (RNA-seq) to determine the amount of gene activity in a certain cell or tissue type. In humans and other multi-cell creatures, nearly every cell contains the same genes, but different cells show different patterns of gene expression. These differences are responsible for the many different types of properties and behaviors seen among various cells and tissues, both in health and disease. By collecting and comparing transcriptomes of different types of cells, researchers can gain a deeper understanding of what constitutes a specific cell type, how that type of cell normally functions, and how changes in the normal level of gene activity may reflect or contribute to disease. Furthermore, by aligning the transcriptome of each cell type to the genome, it is possible to generate a comprehensive, genome-wide picture of what genes are active in which cells.

4.2 TRANSCRIPTOME ANALYSIS

4.2.1 Transcription profiling methodologies

Imagine that you could make a snapshot of a cell, and all the molecules it contains, to capture the thousands of mRNA molecules present at a single point in time. These mRNA sequences, or transcripts, are a record of the proteins the cell intends to produce at this moment in time. If you could count how many molecules of each type are present, you would measure the level of activity of each gene. How would you count all those molecules? The early approach to study whole transcriptomes used

microarrays, a set of defined sequences arranged on a solid substrate. Microarrays almost exclusively represented mRNAs. Nowadays the microarray approach is supplanted by high-throughput RNA sequencing (RNA-seq), which detects all transcripts in a sample, including the regulatory siRNA and lncRNA transcripts. We will focus on the latest RNA-seq methods for the remainder of this chapter, but because of the vast amount of published transcriptome data that has been generated using microarrays, we will briefly look at what a microarray is and how it works in detecting relative mRNA levels.

4.2.2 Microarrays

In 1995, Pat Brown's lab at Stanford University introduced microarrays to the world. At the time, the technology behind the microarray was not new. However, by combining and improvingvarious pre-existing concepts and methods, microarrays made it possible to simultaneously study the expression of many genes in a large number of samples.

Microarray technology uses an important characteristic of nucleic acids that can be used to distinguish among different sequences: complementarity. Remember that DNA (but also RNA) can be double-stranded because it tightly binds to its complementary sequence. The important point here is that it does so with high specificity. We can imagine constructing a short single-stranded DNA sequence (the probe) that is exactly complementary to part of a particular mRNA transcript (the target). If this target mRNA is present in a cell it will hybridize to our short DNA probe sequence; if it is not present we get no hybridization. However, when a gene is expressed in a cell or tissue it will not be a single molecule, but many thousands of mRNA molecules. To quantify the number of mRNA molecules we need to be sure that an excess of identical probe molecules is present (*i.e.* much more probe molecules than target molecules), so that all mRNA molecules are able to hybridize. When this condition is met, we could in principle quantify the number of transcripts present by measuring the number of probes that have been hybridized. An important restriction to keep in mind is that the way in which we can detect hybridization does not allow quantification by counting the number of times a probe hybridized to an mRNA. Rather, hybridization is measured as a relative value that is not equal to the number of mRNA molecules, but proportional to it. These values do not tell us much on their own except that a gene is expressed to some extend (because hybridization occurred, the target mRNA must be present). What we can do, however, is compare these relative values between for example different tissues, giving us another relative value that is the actual change in expression of a particular gene in different situations, *e.g.* healthy versus tumor tissue. This difference in gene expression is called *fold-change*. For example, if the fold-change in expression of gene X is 2.5 between tumor and healthy tissue, then tumor tissue contains 2.5 times as much of the mRNA of gene X than healthy tissue does. Now we know that gene X is upregulated in cancer, but we do not know how much of the mRNA was exactly present in both cases. From a biological perspective, knowing fold-changes in gene expression is in most cases sufficient to understand what is happening in a cell, especially when this can be done simultaneously for all mRNAs that are present in the cells or tissues of interest.

There are more than 20,000 different genes in multicellular eukaryotes, such as humans. How can we quantify the expression of all of these genes simultaneously? This is were the actual meaning ofthe word *microarray* comes into play. For each gene in the genome, specific probes can be synthesized so that all mRNAs that make up the transcriptome can be hybridized to their own matching probe. When a probe-target hybridization occurs we need to measure this independently for all probe sequences. This can be done by "printing" the unique probe sequences on a solid support, such as a glass microscope slide, and chemically linking them to the surface. By organizing the probes as spots in a

regular pattern, a so-called microarray is constructed. The prefix *micro* is used because a small slide of only a few cm^2 typically contains tens, if not hundreds of thousands of spots. Each spot in the array contains a large excess of the same probe sequence that was designed to specifically hybridize to a particular mRNA only. The remaining challenge is now to detect if hybridization has occurred on each one of these spots independently.

This is done by using fluorescent molecules such as the cyanine dyes Cy3 and Cy5. If we extract all mRNA molecules from a tissue and attach a fluorescent dye to each of them (a so-called fluorescent labeling reaction), these labeled mRNAs will hybridize to their complementary probes on the microarray and generate a fluorescent signal. When we measure the fluorescence of each spot on the array independently, we will obtain values that are proportional to the number of mRNA molecules that were bound and thus are a measure of the relative number of mRNA molecules present in the tissues at the time of harvesting the RNA.

This general description begs two further questions: how do we label mRNAs with a fluorescent dye, and how do we measure the fluorescence on each spot individually? The first problem is solved by transcribing the mRNA into DNA using the DNA polymerase *reverse trancriptase*. The resulting single-stranded DNA copy of the mRNA is called cDNA. During this polymerization reaction new nucleotides (dATP, dCTP, dGTP and dTTP) are incorporated into the growing cDNA strand. If we substitute one of these nucleotides with a fluorescently-labeled variant, the reverse transcriptase will build these into the growing cDNA strand, effectively labeling the entire cDNA copy with numerous fluorescent labels. Such fluorescently labeled nucleotides are chemically synthesized in advance and are commercially available. As any molecular biologist knows, polymerases cannot polymerize DNA out of nothing. They need a starting point that is already double-stranded, meaning that the template (DNA or RNA) should be *primed* first. We therefore need a *primer* for the reaction to occur, but how can this be done if every mRNA has a different sequence? With some exceptions, mRNA molecules typically contain a stretch of multiple A nucleotides on their 3'-end, known as the *poly A tail*. This is frequently as a priming site in these labeling reactions. By adding a primer made out of ~15 consecutive T nucleotides [denoted oligo$(dT)_{15}$], all mRNAs can be primed for reverse transcription in a single reaction. Next, the resulting labeled cDNAs are purified and incubated with the probes on the microarrays' surface under conditions that favor optimal hybridization (salt strength, pH etc). Afterwards, unbound cDNAs are washed from the slides using a careful washing procedure. At this point we have an array of thousands of spots, each spot having captured the fluorescently-labeled cDNA copies of the original mRNA molecules from the tissue under study. Finally, the fluorescence of each spot is measured using a dedicated device called a microarray scanner. These scanners use lasers to excite the fluorophores on the microarray and read the array surface point-by-point using a photomultiplier tube. The greater the number of mRNA molecules in a cell, and therefore hybridization of the labeled cDNA to the spot, the greater the intensity of fluorescence at the spot containing the complementary probe (Figure 4.1).

Microarrays allowed scientists to monitor the activity of all the genes in a genome in a single experiment. The data produced by microarray experiments – expression levels for every gene – present many computational and statistical challenges, but the potential of this method is considerable. As such, the analysis of gene expression has become one of the fundamental tasks of computational genomics (as well as an economic force in biotechnology), on par with the analysis of DNA sequences.

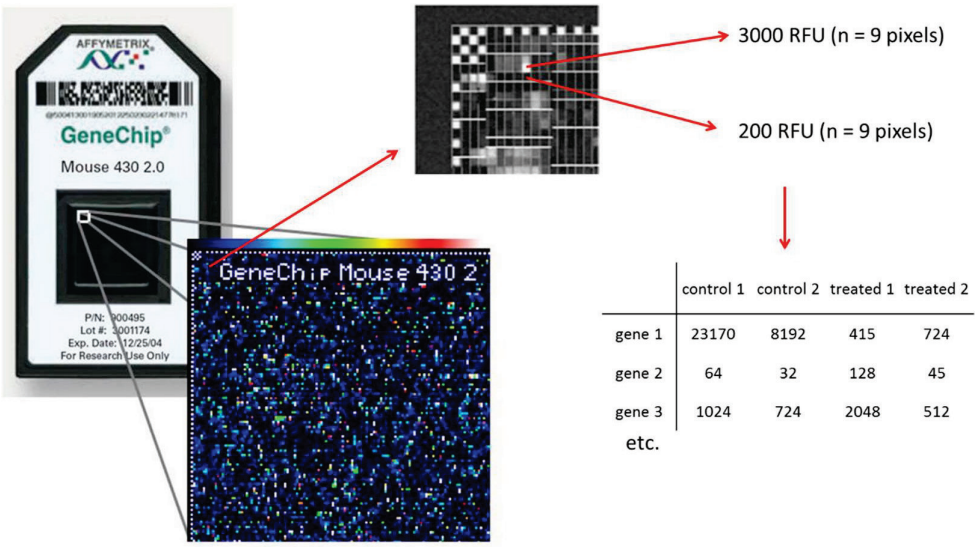


Figure 4.1 Visualization of the computation of the spot intensities of a microarray. The chip contains millions of spots (also called features), and each feature contains probes with the same sequence. Different features contain probes with different sequences. When labeled target is hybridized to the chip, each feature will produce a slightly different signal intensity. After scanning, a grid is laid over the image and each feature is quantified, which results in a gene expression matrix.

4.2.3 RNA-seq

Around 2005, various next-generation sequencing (NGS) technologies became commercially available to the scientific community. Initially this new high-throughput technology was primarily used for sequencing genomic DNA. Soon the technique was adapted for use as a more powerful alternative to transcription profiling by microarray, called *RNA-seq*. Briefly, by isolating RNA from a tissue, converting all mRNAs to cDNA and by sequencing each one of these cDNAs, we can get a very thorough overview of the transcriptome at the sequence level (this procedure will be discussed in detail in the next paragraph). Keep in mind that each of the large number of resulting sequences (called *reads*) is originally derived from a single transcript. Various algorithms are available to identify every sequence by comparing it to a large database of all known mRNA sequences (called *mapping*). Now that we have the identity of each read, we can 'count' how many mRNAs were present in the tissue for each gene separately. It is important to note that these transcript counts are, similar to the fluorescence values obtained in a microarray analysis, relative values that are related to the level of gene expression but do not represent the actual number of mRNA molecules that were present in the original tissue or cells. As with microarray, however, these count values can still be compared between different tissues to obtain a fold change of gene expression.

4.2.4 Preparing RNA-seq libraries

A typical workflow for RNA-seq comprises RNA isolation, library preparation, massively parallel sequencing and data analysis. First, as for genomic DNA sequencing (DNA-seq), sequencing libraries have to pe prepared to make the RNA compatible for sequencing by the NGS machine. This basically means that special adapter sequences, typically 50 to 75 nucleotides in length, have to be attached to each end of every DNA or RNA fragment for the sequencer to "recognize" them. Two different types

of adapter sequences are used and each fragment will have one adapter type to one end and the other adaptor type to the other end. The adaptors serve multiple purposes: (1) they can be used to amplify the library by PCR if there is very little starting material available, (2) they are needed to attach each library molecule spatially separated from each other to the flow cell of the sequencing device in order to be able to sequence them separately, (3) they contain specific binding sites for the sequencing primer to bind to, and (4) they allow the introduction of accessory sequences such as barcodes (also called indexes) enabling the use of many different samples in the same sequence run (more about his later). When sequencing genomic DNA, we can directly add the adapters after fragmenting the DNA to sizes that the sequencer can handle (<1000 bp). Since current NGS platforms cannot directly sequence RNA, we first need to convert RNA into DNA. Similar to the fluorescent labeling procedure used for microarrays, we can achieve this by using a reverse transcriptase to make cDNA from our RNA first. Also, we need to make sure that our sequencing library (cDNA + adapters) is made up of fragments that are compatible with the sequencer. All this can be achieved in numerous ways and below we will describe some of the most used methods.

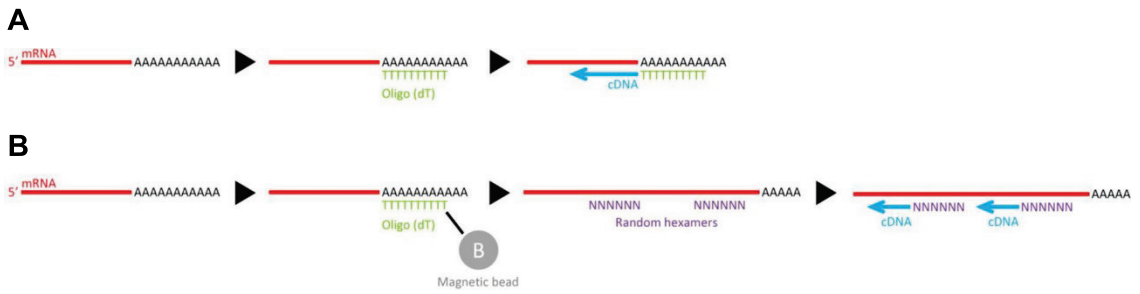


Figure 4.2 Synthesis of cDNA from mRNA using different priming approaches. (A) Direct priming of mRNA using oligo(dT). cDNA is synthesized starting from the annealed oligo(dT) primer using a reverse transcriptase. This can be performed in total RNA because only mRNA molecules will be primed and converted into cDNA. The resulting cDNA will be biased for the 3' end of the mRNA. (B) Random priming of preselected mRNA. First, mRNA is separated from other RNA types (rRNA etc.) using oligo(dT) primers attached to magnetic beads. Next, the selected mRNAs are converted into cDNA using random hexamer primers (N6). The resulting cDNA will be derived from the entire mRNA and will be less biased for either of its ends, compared to method(A).

Similar to microarrays, the synthesis of cDNA from RNA can be done using oligo dT primers that anneal to the mRNA's poly A tail (Figure 4.2A). Since polymerases rarely work perfectly, most of the time they will not reach the 5'-end of the mRNA resulting in only part of the mRNA being converted into cDNA. This is called a *3' bias* since the 3'-end where the poly A resides is more efficiently reverse transcribed and thus overrepresented in the sequence data. As a result, libraries that are made this way are less suitable to study alternative splicing since the transcript is not fully sequenced. In many cases, it is not necessary for a researcher to have detailed information on gene splicing as it is of less biological relevance to the research question and makes data analysis much easier. Since it is often sufficient to quantify gene expression with a single relative value that includes all mRNA (splicing) variants for a particular gene, just as was the case for microarrays, various library preparation protocols have been developed by the scientific community. The main advantage is that when a library is prepared in such a way that one mRNA molecule generates only one library fragment molecule (and therefore one sequence read), less sequence capacity is wasted by getting sequence reads dispersed over the entire transcript. Getting only a single read from each mRNA molecule can be achieved in various ways using methods such as *TagSeq*, *3'Tag RNA-Seq*, *Digital RNA-seq*, and *Quant-Seq*. It is beyond the scope of

this course to discuss these in detail. Just remember that these techniques are available and what the advantages and disadvantages are when using them.

What if we are actually interested in obtaining the entire transcript sequence with as little positional bias as possible? In addition to being able to study gene splicing, this information could also be used to put together a complete overview of the sequence composition of each transcript in the transcriptome of, for example, a species that has not previously been sequenced (known as *de novo* transcriptome assembly). There are various approaches to generate so-called full-length or whole transcriptome sequencing (WTS) libraries. One solution is to add random hexamers, a mixture of short oligonucleotides of 6 nucleotides long containing all 4^6 (=4,096) possible base combinations, to the reverse transcription reaction. These hexamers, denoted oligo$(dN)_6$ or in short N6, can function as primers for reverse transcriptase and will (at least theoretically) anneal in a random pattern over the entire length of the RNA transcript (Figure 4.2B). Reverse transcription of the RNA – hexamer mixture will result in short cDNA fragments generated over the entire length of the original transcript. Similar to oligo(dT) priming, however, these fragments will unfortunately not cover the entire transcript equally so there will still be a bias, but now towards the 5'-end of the mRNA. By simultaneously priming cDNA synthesis with random hexamers and oligo (dT), a more equal distribution of fragments over the entire length of the transcript can be achieved. The size of the resulting cDNA fragments can be controlled by the amount of hexamers added. Adding too many will in the end completely occupy the entire transcript so that reverse transcription is inhibited. The random hexamer approach solves two of the challenges in a single step: (1) priming the conversion of RNA to cDNA, and (2) creating short fragments that can be handled by the sequencer.

4.2.5 Getting rid of ribosomal RNA

As previously discussed, less than 5% of the total cellular RNA content is mRNA. The rest is mostly rRNA (5S, 5.8S, 18S and 28S rRNA subunits) and to a lesser extent other non-coding RNAs such as tRNA, lncRNA and miRNA. By using the mRNA's poly A tail to prime cDNA synthesis, we have the advantage of only generating library fragments from the mRNAs without being bothered by the vast excess of rRNA. However, when we use random hexamers for priming cDNA synthesis we get into trouble because the large majority of library fragments will be made up from rRNA sequences. When sequencing these libraries, easily >90% of all reads will be rRNA-derived, which is not very interesting unless you are studying rRNA sequence variation. There are two ways to circumvent this problem:

(1) **Poly A^+ enrichment.** Before starting the library preparation procedure, exclusively the polyadenylated mRNAs (with poly A tail) can be easily isolated using oligo dT sequences attached to small magnetic beads (Figure 4.2B). So instead of priming the cDNA synthesis, oligo dT is now used to capture mRNA molecules on magnetic beads, which can be easily separated from the rest of the RNA solution containing the rRNA and other non-coding RNAs. This so-called *poly-A enriched* fraction contains all poly adenylated mRNAs that can now be used as input for the library preparation. Note that: (1) mRNAs that do not have a poly A tail (or only a very short one) will not be isolated and sequenced, and (2) non-coding RNAs that have internal stretches of multiple A nucleotides are still isolated and sequenced, but in practice this results only a minor fraction of the total number of reads obtained after sequencing the library.

(2) **Ribodepletion.** In stead of fishing out the mRNA from the entire cellular RNA content, we could also do the opposite: fish out the unwanted rRNA molecules. This can be done by synthesizing single-stranded DNA sequences that are complementary to the four types of rRNA and linking these to magnetic beads, similar to the procedure used for poly A enrichment. When these beads are incubated with the total cellular RNA content, all rRNA molecules will bind to there complementary sequences on the magnetic beads, enabling their efficient removal using a simple magnet. The mRNA that is left, called the *ribodepleted RNA fraction*, can be directly used as input for the library preparation procedures described above. This approach has important advantages: (1) Unpolyadenylated mRNAs are retained and sequenced, in contrast to the poly A enrichment method. Remember that prokaryotes do not have polyadenylated mRNAs, so this technique is indispensable for performing RNA-seq in prokaryotes. (2) Non-coding RNAs are retained and sequenced. There is currently a large interest in the role of long non-coding RNAs (lncRNA) in disease. By using this method we can include theseand other non-coding RNA in our transcriptomics dataset.

4.2.6 Small RNA sequencing

Small RNAs constitute a separate class of RNA molecules that are solely defined by their size being shorter than 200 nucleotides in length. Even though some mRNAs and 5/5.8S rRNA are shorter than 200 nucleotides, these by definition also belong the to group of small RNAs. One of the most well studied small RNAs at the moment are the microRNAs (miRNA), which are between 20 and 25 nucleotides in length, depending on the species. Sequencing small RNAs has an interesting advantage: their sequences are so short that they can be sequenced directly without any fragmentation. Adapters can be directly ligated onto the small RNA molecules and the resulting sequence composition and length is exactly that of the molecule as it was present in the cell. This in marked contrast to RNA-seq, where larger RNA molecules are digested into small fragments first, which later have to be put together by mapping the reads to a reference sequence. Some small RNAs, particularly miRNAs, are known to be processed after transcription and can therefore vary in length (isomiRs) . By sequencing them without fragmentation the data can directly inform us if any such processing has occurred *in vivo*. During the library preparation procedure, size selection steps are included to get rid of the larger 5S and 5.8S rRNA and tRNAs, which are not of interest for most transcriptomics studies, but could consume a lot of “uninteresting” reads.

4.2.7 Single-cell transcriptome sequencing (scRNA-seq)

One of the most recent additions to the transcriptome analysis repertoire is the ability to isolate and sequence the transcriptome of single cells. Previous gene expression measurements have been performed on bulk samples. Conventional bulk-based RNA sequencing can provide a full view of all gene expression, and by example, in combination with flow cytometry, be useful to investigate a tumor’s microenvironment. However, such a blended gene expression analysis of multiple cells and cell types together, might mask the minor cell population that may be the actual origin of tumor progression. To overcome this problem, RNA sequencing methods that can analyze mRNA expression at the single-cell level from thousands of individual cells are required. The fundamentally necessary steps to take for single-cell RNA sequencing are: (1) single-cell isolation with a high survival rate, (2) cell lysis to obtain mRNA, (3) conversion of mRNA into cDNA, (4) specific amplification of cDNA, (5)

cDNA fragmentation process, and (6) creation of high-quality sequencing libraries (Figure 4.3). The biggest challenge is isolating intact single-cells from a tissue or body fluid like blood (step 1). Firstly, the best method of achieving a suspension of intact single cells, and separating these into single cells, is strongly dependent on the tissue type. Isolating single cells from blood, for example, is a lot easier than isolating single cells from the islets of Langerhans in the pancreas. Cells in dense tissues with a lot of connective tissue and extracellular matrix will require treatment with proteases to detach the cells. Secondly, the procedure should be performed quickly as cells will immediately start changing the expression of (some of) their genes when they are manipulated and taken out of their original environment.

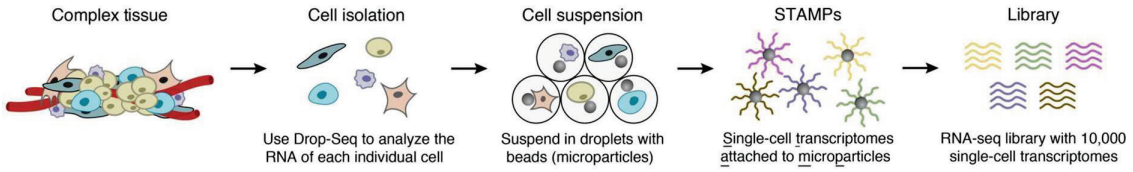


Figure 4.3 Workflow for single-cell RNA-seq (scRNA-seq). An example of a possible workflow using the Drop-seq method (discussed below) is shown.

After preparing a single-cell suspension, the cells have to be separated and for each individual cell an RNA-seq library needs to be prepared. Currently available single-cell technologies solve these two challenges differently. There are some innovate single-cell transcriptome analysis methods for generating high-quality RNA-seq libraries from very little starting RNA. Keep in mind that approximately 10^5–10^6 mRNA molecules are present in a typical single mammalian cell, and up to 10,000 different genes may be expressed. The total cellular RNA content (including rRNA etc.) is about 10–30 picograms so different procedures are necessary compared to regular RNA-seq. The approaches that are available vary from technically challenging protocols that can only be properly executed by highly-skilled researchers, to ready-for-use fool-proof equipment that is made commercially available by companies such as *10x Genomics*. The strategy used for capturing single cells and making the libraries determines the throughput, how the cells can be selected as well as what kind of additional information besides the sequencing that can be obtained. The available methods fall into three categories: *microwell-based*, *droplet-based*, and *microfluidic device*.

Microwell-based. For well-based platforms, cells are isolated using for example a pipette, laser capture or automated cell sorter, and placed in microfluidic wells (Figure 4.4). One advantage of well-based methods is that they can be combined with fluorescent activated cell sorting (FACS), making it possible to select cells based on surface markers. This strategy is thus very useful for situations when one wants to isolate a specific subset of cells for sequencing. Another advantage is that one can take pictures of the cells. The image provides an additional modality and a particularly useful application is to identify wells containing damaged cells or doublets. The main drawback of these methods is that they are often low-throughput and the amount of work required per cell may be considerable.

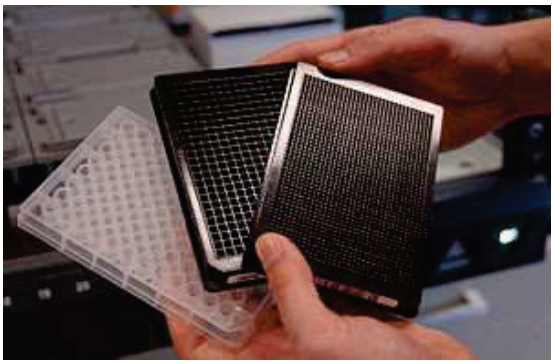


Figure 4.4 Microwell plates are available as 96, 384, or 1536 individual wells (image taken from Wikipedia). Various methods are available to split a suspension of single cells (prepared from a solid or liquid tissue) into microwell plates in such a way that most wells contain a single cell only. The most basic approach is hand-picking single cells using a micropipettor, but cell sorters such as FACS can also be used for this purpose.

Droplet-based. The idea behind droplet-based methods is to encapsulate each individual cell inside a nanoliter droplet together with a bead (Figure 4.5). The bead is loaded with the reagents required to construct the library. In particular, each bead contains a unique barcode which is attached to all of the reads originating from that cell. Thus, all of the droplets can be pooled, sequenced together and the reads can subsequently be assigned to the cell of origin based on the barcodes. Droplet platforms typically have the highest throughput since the library preparation costs are relatively low. Instead, sequencing costs often become the limiting factor and a typical experiment the coverage is low with only a few thousand different transcripts detected.

Example: Drop-seq (https://www.youtube.com/watch?v=vL7ptq2Dcf0)

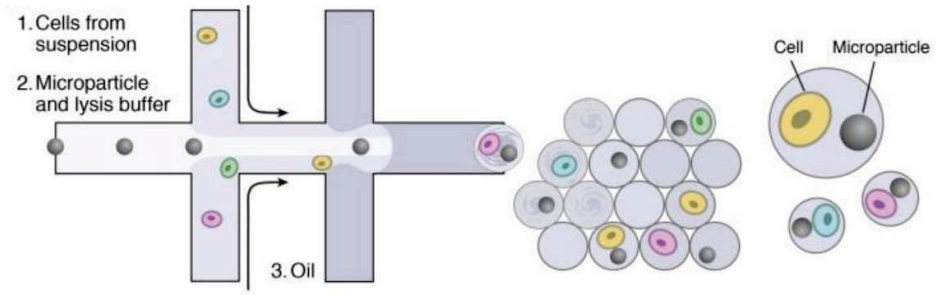


Figure 4.5 Schematic overview of the Drop-seq method. An emulsion of water droplets in oil is made in such a way that (most) droplets will contain one cell, one microparticle (bead) containing adapters and all reagents (cell lysis buffer, polymerases etc.) needed to convert mRNA into cDNA attached to the surface of the bead. Since the mRNA from each individual cell now has been converted into cDNA, tagged with a unique barcode (index) and is attached to its own bead, all droplets can now be broken and the contents pooled together to amplify and finish preparation of the sequencing library (not shown).

Microfluidic device. Microfluidic platforms, such as Fluidigm's C1, provide a more integrated system for capturing cells and for carrying out the reactions necessary for the library preparations (Figure 4.6). Thus, they provide a higher throughput than microwell-based platforms. These systems are highly automated ready-for-use commercial solutions, which require very little handling by the operator. The

whole procedure, from capturing individual cells to preparing a ready-for-sequencing barcoded library, is all performed inside the device. However, typically only around 10% of cells are captured in a microfluidic platform and thus they are not appropriate if one is dealing with rare cell-types or very small amounts of input. Moreover, the chip is relatively expensive, but since reactions can be carried out in a smaller volume money can be saved on reagents.

Example: Fluidigm C1 (https://www.youtube.com/watch?v=TF4NJRE4Xg4)

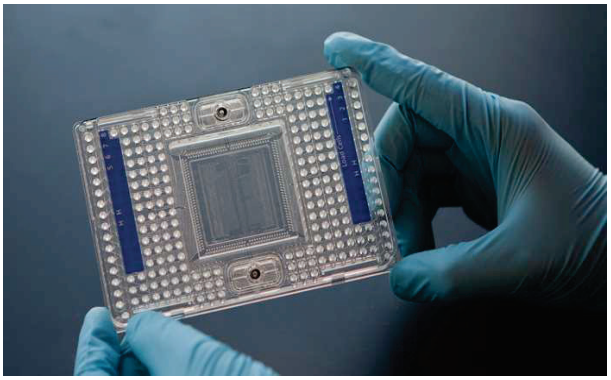


Figure 4.6 Image of a 96-well Fluidigm C1 chip (image taken from Fluidigm). This is an example of a commercially available technology that offers a completely integrated solution from cell suspension to read-for-use RNA sequencing libraries. Starting from a suspension of single cells, cells are individually captured on a chip's dedicated micro reactor, lysed and RNA is converted into a barcoded library.

4.2.8 Unique molecular identifiers (UMIs)

Preparing RNA-seq libraries involves multiple steps where RNA and DNA are manipulated using state-of-the-art tricks from the molecular biology toolbox. Any modification of a collection of RNA or DNA molecules, such as ligation or polymerization, introduces a bias in the composition of this mixture, because molecules with a different DNA sequence are modified by these enzymes with different efficiencies. For example, based on the sequence and length of a DNA/RNA fragment, adapters will ligate poorly and/or some fragments will be poorly copied by the polymerase because of secondary structure. For many applications, such as molecular cloning, this is less of an issue, but for transcriptomics, where quantitative comparisons are made between different transcriptomes, this is problematic. Sequence bias is the cause of loss of exact quantitative information during transcriptomics lab procedures. For example, when a library is sequenced and 20 of these sequences are derived from the same gene, then this does not mean that 20 RNA molecules from this gene were originally present in the cells or tissue under study, *i.e.* absolute quantification is not possible. Similarly, sequence bias also plays a role in microarrays, where there is no direct absolute quantitative link between number of RNA molecules and fluorescent signal on the array. What we can do, in the best case scenario, is compare the expression of the same gene between different biological samples because the sequence, and thus its detection bias, stays the same between these samples. At least, this is an assumption that has been made for bulk transcriptomics for a long time. For scRNA-seq this assumption is no longer safe because the amount of total RNA that is obtained from a single cell is very low, of which only about 1 picogram is mRNA. The material therefore needs to be amplified to such an extent (*e.g.* by PCR) that sequence bias increases substantially.

The inclusion of dedicated sequences, called unique molecular identifiers (UMIs), in the sequencing library adapters, can help to (partly) restore quantitative sequencing. UMIs are a collection of short

random sequences of around 5 nucleotides long, giving a total of 4^5 = 1024 possible sequence combinations. During preparation of a sequence library, a random one of these UMIs is attached to each library fragment, either by including them in the adapters or by ligating them to the fragments independently from the adapters (see Figure 4.7: colored boxes attached to the curly lines/RNA). Each library fragment molecule is now tagged with an UMI as well as a barcode, which can be confusing. Remember that library fragments that are derived from the same single cell carry the same barcode sequence. During sequencing many single-cell libraries are combined into a single sequencing run, but later during the data analysis stage are separated again using their barcode (this is called *demultiplexing*). In contrast to the barcode, however, a UMI sequence is not associated with the original cell, but rather with the original RNA fragment molecule because they have been randomly attached. A fragment's UMI sequence can therefore be used to determine if multiple identical sequence reads are derived from the same original RNA fragment molecule (*i.e.* they are PCR duplicates of the same original fragment) or from another transcript molecule transcribed from the same gene. This is how it works: If two sequence reads have the same sequence then they are obviously derived from the same gene, but could also be identical because they are copies from the same original RNA fragment (actually cDNA) made by the PCR. As explained above, PCR efficiency and adapter ligation are influenced by the sequence of the target, so if the PCR for a particular RNA fragment behaves more optimally in one sample compared to another, the difference in the number of reads obtained between these samples is not necessarily the exclusive result of a change in gene expression, but possibly more the result of a difference in PCR efficiency. By looking at the sequence of the UMI, we can easily conclude if two identical reads are the result of PCR duplication. If both have the same UMI, then they are PCR duplicates, if not then they are derived from another transcript molecule that was transcribed from the same gene. This is true because UMIs are randomly attached to the each library fragment and since there are 1024 possible UMI sequences, the chance that the same UMI sequence is attached to two identical RNA molecules twice is negligible. Instead of counting all reads that are derived from the same gene, we discard identical reads with the same UMI and only score these as one count. Figure 4.7 illustrates this concept by tracking a few mRNA molecules throughout the library preparation procedure. If you count the total number of reads obtained for each of the three example mRNAs (blue, black and purple), then you will find that these numbers no longer correlate with the original number of mRNA molecules that were present in the single cells. By discarding reads with the same mRNA sequence and the same UMI, this bias is corrected. UMIs can be used for DNA sequencing in a similar way when quantitation is important.

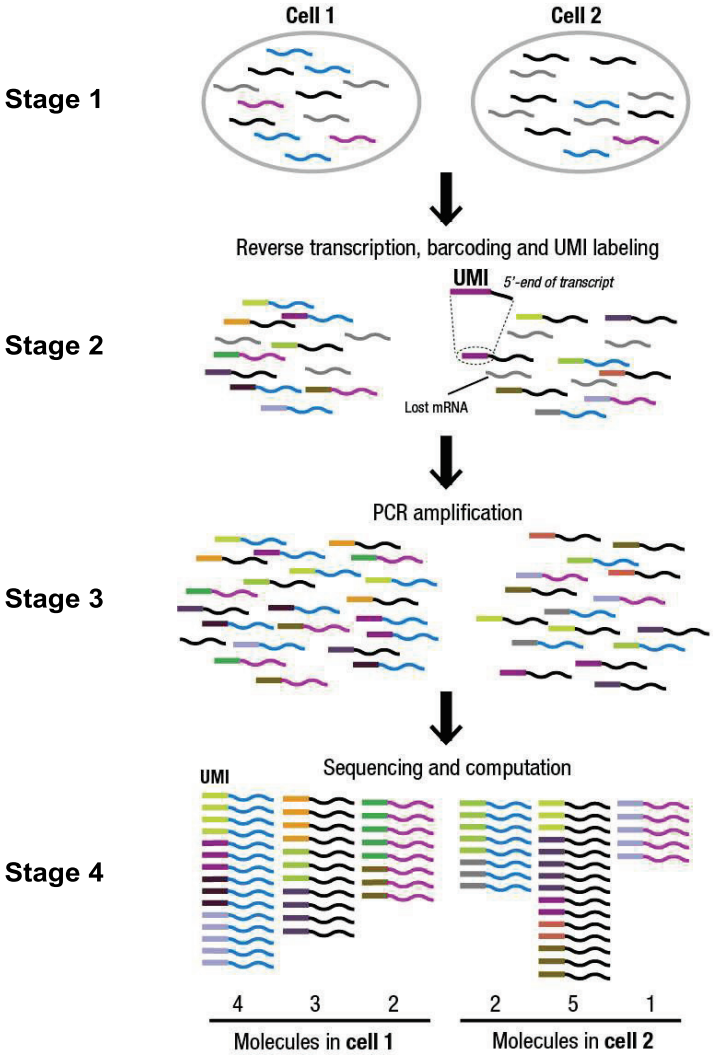


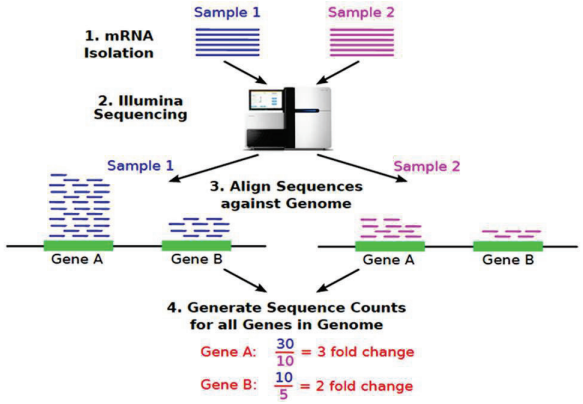
Figure 4.7 Use of Unique Molecular Identifiers (UMIs) in single-cell RNA-seq. Shown is the library preparation workflow of a typical transcriptomics experiment where single cells are compared to find differentially expressed genes. To demonstrate the principle behind the use of UMIs, only 2 cells are shown. During the procedure, each cell receives a unique barcode that is used to identify which read belongs to which single cell. In addition, to correct for sequence bias, which results from the library preparation procedure and PCR amplification, UMIs are attached to each library molecule to enable the identification of unique transcript molecules during data analysis. Curly lines of the same color depict transcripts derived from the same gene. UMIs are indicated by boxes, colored according to their sequence. Example: **(Stage 1)** *4 identical copies of the blue mRNA are present in cell 1, but 15 corresponding reads are counted after sequencing* **(Stage 4)**, *which is the result of PCR duplication* **(Stage 3)**. *During UMI labeling* **(Stage 2)** *these 4 molecules are tagged with a different UMI because having the same UMI attached to the same mRNA twice is statistically improbable. The only probable way in which the same read can have the same UMI attached* **(Stage 4)** *is when these are PCR duplicates, which should not count for quantitation of gene expression.*

4.3 Data analysis

4.3.1 Quantification of gene expression levels through aligning and counting reads

A sequencing run usually generates millions of short sequences (the reads) for each sample. The next step is to translate these reads into values representing a gene expression level. This basically occurs through counting the numbers of reads that originate from each and every gene. So the challenge is

A



B

	control 1	control 2	treated 1	treated 2
gene 1	825	935	5014	4512
gene 2	64	45	128	32
gene 3	2096	724	1025	645

etc.

Figure 4.8 Visualization of the alignment and counting of sequencing reads. (A) After mapping the reads to a reference genome, the number of reads that align to a certain gene are counted, and this count quantifies the expression level of the genes. (B) The counts for each gene for multiple samples are frequently organized as a gene expression matrix.

to find out where each read comes from. Biomedical research often involves human samples, such as human cell lines or organoids. Or otherwise model organisms are used such as Mouse (*Mus musculus*), Zebrafish (*Danio rerio*), Yeast (*Saccharomyces cerevisiae*), Fruit fly (*Drosophila melanogaster*), or the Nematode worm (*Caenorhabditis elegans*). Bacterial research usually focusses on model species like *Escherichia coli*, *Bacillus cereus* or *Staphylococcus aureus*. For these species reference genomes exist that are pretty well described and annotated. So, to determine the origin for each read, it is often decided to align, or map, the reads to a reference genome. But please note that a well-known reference genome important for this step to be successful. This mapping is achieved by comparing the sequence of the read to that of the reference sequence. A mapping algorithm will try to locate a (hopefully unique) location in the reference sequence that matches the read, while tolerating a certain amount of mismatch and splicing events. Some more details of read alignment is discussed in the lectures and elsewhere in this syllabus, but in-depth knowledge about the algorithms is beyond the scope of this course.

After mapping the reads to a reference genome, the number of reads that align to a certain gene can be counted, and this count quantifies the expression level of the genes. A high number of reads results in a large count-value, and this means that many RNA molecules were present in the sample, indicating a high expression level. A low number of reads results in a low count-value, which means that only a few RNA molecules were present in the sample, indicating a low expression level. The data from multiple samples is usually organized as an expression matrix, with the samples as columns and the genes as rows. Figure 4.8 shows a typical RNA-seq data set for four samples (two control and two treated samples), and three genes. In reality you will have thousands of rows, namely one row for every gene. Researchers usually calculate differences in gene expression between conditions, expressed as fold changes. Of note, the first gene in this example seems upregulated due to the treatment.

Another strategy to work towards gene expression values is to map the reads to a reference transcriptome instead of a reference genome. A reference transcriptome consists of a large number of sequences that represent each gene. A reference transcriptome contains mature mRNA's without introns and also contains the known isoforms for each gene. Hence, an advantage of this strategy is that spicing does not need to be taken into account during the mapping, and that isoforms can be analyzed. You don't need to use the gene coordinates of the genome whilst counting reads, so you will have a direct read-out of transcript abundance. A disadvantage of this strategy is that mapping is complicated by the sequence similarity between isoforms. Note that isoforms of a gene can have very similar sequences, so it may be difficult to decide whether a read comes from one isoform or the other. Yet, methods do exist that take this mapping uncertainty into account in the statistics (Pimentel et al., 2017). Another challenge is that the researcher needs to decide how isoform abundances are translated into gene abundances. It may happen that isoform abundances differ between conditions, but that gene abundances are similar. In practice, it appears that gene expression level analysis at the level of transcripts (isoforms of a gene) adds another level of complexity, such that researchers often do not address this. It has however been shown that relevant biological phenomena are missed when splicing and isoforms are not taken into account in the gene expression analysis (Yi et al., 2017). Nevertheless, mapping on a transcriptome also results into an expression matrix such as indicated in Figure 4.8B, with estimates for expression levels for every transcript, for every sample. So the data can then be subjected to differential expression analysis.

Reference transcriptomes are usually far from complete. The reason is that splicing is condition specific, and thus also isoform abundance is condition specific, and that a reference transcriptome is usually constructed on the basis of a limited number of conditions. It may happen that the experimental condition under study results in new isoforms that yet need to be discovered! This can be addressed in the analysis, and depending on the situation and organism under study, different strategies exist: a reference-based strategy, a de novo strategy or a combined strategy that merges the two. These strategies will be briefly discussed in this section based on a review by Martin and Wang (2011).

When a reliable reference genome for the target transcriptome is available, the transcriptome assembly can be built upon it. In general, this strategy, which is known as 'reference-based' assembly, involves three steps. First, RNAseq reads are aligned to a reference genome using a splice-aware aligner. Second, overlapping reads from each locus are clustered to build a graph representing all possible isoforms. The final step involves resolving the individual isoforms by combining the merged exons.

The 'de novo' transcriptome assembly strategy does not use a reference genome. Note that a sequencing run causes every transcript to be sequenced multiple times. De novo transcriptome assembly leverages this redundancy to find overlaps between the reads and assembles them into longer stretches of sequence (contigs). Some assembly algorithms analyze the overlaps to directly assemble each isoform. Other types of algorithms assemble the data set multiple times to reconstruct transcripts, and then post-process the assemblies to merge contigs and remove redundancy. Whereas most of the short-read de novo assemblers created to date were developed and optimized using short-read data sets, longer second-generation reads, such as Minion reads, can also be integrated into de novo transcriptome assemblies, which improves the ability to resolve alternative isoforms.

Reference-based and de novo strategies can be combined to create a more comprehensive transcriptome. By bringing together these two complementary strategies, one can take advantage of the high sensitivity of reference-based assemblers while leveraging the ability of de novo assemblers

to detect novel and trans-spliced transcripts. Generally, the combined assembly strategy can be carried out by first either aligning the reads to the reference genome and then by de novo assembling the unaligned reads (Align-then-assemble). Alternatively, a de novo assembly could be performed first followed by alignment of the contigs to the reference to extend the contigs into transcripts.
After the transcriptome assembly process, the reads are mapped back onto the transcript sequences, which also results into an expression matrix such as indicated in Figure 4.8B, with estimates for expression levels for every transcript, for every sample. One should be aware of the fact that many sequences may be novel and hence unknown. So the next challenge is to give functional interpretation to these discoveries. The data is then ready to estimate fold changes of genes and differential gene expression analysis.

Bulk RNAseq and Single Cell RNAseq. The explanation of the processing of sequence reads was so far a general description and primarily based on bulk RNAseq data. This technology measures the average expression level for each gene across a large population of input cells. This is very useful for comparative transcriptomics and disease studies. However, in some cases one may want to have more insight in cell heterogeneity, for instance in early development studies, or when dealing with complex tissues (brain). Then it may be more useful to perform a single cell RNAseq analysis. Single cell RNAseq can also give insight into the stochastic nature of gene expression.

Single cell RNAseq measures, as the name suggests, the distribution of expression levels for each gene in individual cells. Datasets range from 10^2 to 10^6 cells and increase in size every year. This technology allows to study new biological questions in which cell-specific changes in the transcriptome are important, e.g. cell type identification, heterogeneity of cell responses, stochasticity of gene expression, inference of gene regulatory networks across the cells. Currently there are several different protocols in use, e.g. SMART-seq2 (Picelli et al. 2013), CELL-seq (Hashimshony et al. 2012) and Drop-seq (Macosko et al. 2015). There are also commercial platforms available, including the Fluidigm C1, Wafergen ICELL8 and the 10X Genomics Chromium. Note that the exact methods for data analysis depend on the platform that is used. But in general several computational analysis methods from bulk RNAseq can be used. Depending on the situation, also adaptations of the existing methods or are required or development of new ones. But overall, experimental scRNAseq protocols are similar to the methods used for bulk RNAseq. Some specific aspects of scRNAseq preprocessing steps, are the the following:

1) Demultiplexing. This involves identifying and removing the cell-barcode sequence from the (paired end) reads. Note, when UMIs are used, then also these barcodes should be removed from the read sequence. A common practice is to add the barcode to the read name. Demultiplexing is done differently depending on the protocol used, and described above, and will not be discussed further detail in the course.
2) Selecting cell-containing droplets/microwells. For droplet based methods only a fraction of droplets contains both beads and an intact cell. So reaction containers (droplets/microwells) without intact cells have to be removed from the data set through a specialized quality control procedure.

3) Analysis for gene 'dropouts', in which a gene is observed at a moderate expression level in one cell but is not detected in another cell; an effect caused by low starting amounts of transcripts since the RNA comes from one cell only.

One final remark is that scRNAseq data sets are typically quite large. So bioinformatics researchers are developing methods to speed up mapping algorithms.

4.3.2 Similarities and differences between RNA-seq and microarray data analysis

Microarray technology and RNA sequencing technology are very different, but in many cases they are used for the same goal: to look for genes that display a difference in expression level due to experimental conditions. Hence, in many cases, the data generated from an RNA sequencing platform and from a microarray platform are very similar: both technologies result in an expression matrix. In an expression matrix, the columns indicate the samples measured in an experiment, and the rows indicate the genes, and for each gene, a number indicate the expression level of this gene for each particular sample. A high number indicates a high expression level, and a low number indicates a low expression level (see Figure 4.8). So, although the technologies might differ with respect to dynamic range, sensitivity and accuracy, the analyses should lead to similar biological conclusions. Although microarray technology is nowadays largely replaced by RNA sequencing, we decided to still discuss it in this course for three reasons: 1) we try to teach transcriptomics concepts transcending the technology is used, 2) for historic reasons it is still part of –omics and biomedical sciences, and you have to be able to read scientific literature where microarrays were used, 3) there are still thousands of microarray datasets available in public databases that are reanalyzed by biomedical researchers to answer novel research questions.

Please note that there are also cases where array and sequencing technologies result in very different types of data sets. For instance, if the analysis is focused on reconstruction or characterization of the transcriptome (for instance, detecting splice variants, isoforms, RNA editing), or if other types of RNA molecules are analyzed (such as microRNA's). However, many concepts apply generally to transcriptome analysis independently of the platform used, especially when the interest is differential analysis. In this course we will try to teach you those general concepts, but we try to omit going too much into technical details. Such details are usually platform dependent and outdated in five years' time. Hence, platform specific details are omitted as much as possible.

4.4 WORKFLOW FOR DIFFERENTIAL GENE EXPRESSION ANALYSIS

Depending on the experimental context, differential gene expression analysis can be used for generating and testing hypotheses, classifying samples, annotating genes, studying developmental and evolutionary processes, and performing clinical assessments. There are arguably four basic steps that are relevant for virtually all transcriptomics experiments:

1) Experimental design and data collection. Frame a biological question. Choose a transcriptomics platform. If sequencing is chosen, then one has to decide on the sequencing protocol. Identify noise factors and design the experiment. Execute the experiment.

2) Data quality control (pre)processing. Quality control of the raw data. Trim sequence reads.

Calculate the gene expression values (mapping, assembly). Address batch effects and normalize the data to remove biases introduced by sampling and measurement.

3) Data analysis. Perform exploratory data analysis. Analyze the results of assembly and/or mapping. Perform hypothesis testing: statistical tests to find significant differences between groups.

4) Biological interpretation. Interpret transcriptome differences in relation to experimental conditions. Analyze the response of sets of genes.

The goal of this chapter is to present some of the transcriptomics specific aspects of the data analysis pipeline.

4.5 EXPERIMENTAL DESIGN AND DATA COLLECTION

4.5.1 Frame a biological question

The first step in the experimental design is to formulate a biological question. The number of biological questions is in principle endless, but there are in transcriptomics arguably four main types of objectives that each require a different type of experimental design: 1) Detection of responsive genes under controlled experimental conditions (perturbation study), 2) Detection of biomarkers and 3) Identification of regulatory or mechanistic relationships between genes, 4) analyzing heterogeneous tissues and cell typing. Because of time, we will only be able to briefly address each of these objective during this course. In general, studies executed with the second, third or fourth objective contain many more samples than a perturbation study. Formulating the biological question is extremely important as it determines the complete execution of the experiment.

4.5.2 Choose a transcriptomics platform

One relevant decision that needs to be made in gene expression profiling experiments is on the type of platform and/or sequencing protocol. The factors that are relevant for this choice is the research question, the species that needs to be investigated and the availability of a reference genome. If human cells are going to be profiled, then the obvious choice is a sequencing run where reads are mapped to the human genome or transcriptome. If the researcher is only interested in mRNA levels of “established” genes, then a 3-prime biased assay could sufficient and a very affordable option. If one is interested in splice variants of genes, then a paired end random RNA sequencing protocol is an obvious choice, preferably improved with a long reads sequencing run. If the researcher is more interested in small RNA’s, then it is recommended to select the small fraction of RNA’s and adjust the data analysis protocol. The choice between sequencing platforms and protocols is also determined by the budget. As indicated, 3’ biased protocols are less expensive than random protocols. The sensitivity of a sequencing analysis is determined by the sequencing depth, reliable analyses including low expressed genes require a large sequencing depth and are therefore more expensive. For microRNA analyses on the other hand, sequencing is very cost-effective.

4.5.3 Identify noise factors and design the experiment

The aim of identifying noise factors is to categorize the factors that can hinder reliable measurements. These factors may introduce random noise or may cause bias: a structural error in the measurement. An example of bias is measuring a consistently higher expression level in some samples, because of a

more effective microarray-hybridization process or because more sequencing-reads have been generated. Recognizing these factors enables the researcher to take them into account in the experimental design, to avoid erroneous measurements. The identification of noise factors starts with evaluating the biological experiment. For experiments with plants or mice, climate rooms (or breeding facilities) used may introduce noise. For experiments with cell cultures, the use of different batches may introduce noise. These kind of noise factors are obviously very experiment specific.

Once the biological question is formulated and the platform is chosen, and noise factors have been identified, the researcher has to decide on a sampling scheme and design the experiment. Then the classical basic principles of experimental design become important. There are three basic principles: 1) replication, 2) randomization, 3) blocking. These three tools enable the researcher to deal with factors that are not of experimental interest, but introduce noise, and/or can influence the outcome of an experiment. The aim of the experimental design is to ensure reliable measurements free from bias. This is a general aim and also discussed elsewhere in the syllabus.

4.5.4 Execute the experiment

The execution of the complete experiment involves many steps, and usually multiple people are involved. Typically: 1) Obtaining the biological material (cells or mice or making knock-outs etc), 2) setting up the experiment, 3) running the experiment, 4) harvest the biological material, 5) RNA extraction, 6) library preparation, 7) sequencing. It is important to be meticulous, but a good communication between the people involved is as relevant as accurate work.

4.5.5 Data quality control and (pre)processing.

As indicated above, transcriptomics experiments contain biological and technological variability (or noise) which need to be taken into account. The challenge appears when comparing different experimental conditions. Is the variation of a particular gene due to noise or is it a genuine difference between the different conditions tested. Furthermore, when looking at a specific gene, how much of the measured variance is due to the gene regulation and how much is due to noise? In addition, some "noise" is random, but "noise" can also introduce bias. The noise is an inescapable phenomenon. The noise can be handled by smart design of the experiment, good experimental practice, and a proper statistical analysis.

A very important aspect is the quality control. The goal of a quality control procedure is to infer whether the data are of high enough (technical) quality to allow for biological conclusions. The results of a quality control procedure can indicate which of the sources of variability may have an impact on the results, and whether there is a risk for biases. This impact of technical noise (variability) can then be accounted for by normalization or by omitting low quality data points. There are many ways to perform quality control, and very often they involve platform specific metrics or analyses. The quality control of sequencing reads has been described elsewhere in the syllabus, and will not be discussed in much detail here. But in general, a very important aspect of a quality control is visualizing the raw and preprocessed and normalized data. There are quite a number of plots and tools that can be used for data visualization. You can find examples of some important visualization tools in the lectures and Figure 4. Some examples:

Quality control of the sequence reads. This is obviously specific for RNAseq experiments. An important quality metric is the Phred score. A Phred score is a measure of the quality of the identification of the

nucleotides generated by a sequencing machine. This score (Q) is a value that is logarithmically related to the base-calling error probability (P): $Q = -\log10(P)$. For example, if Phred assigns a quality score of 10 to a base, the chances that this base is called incorrectly are 1 in 10. For Illumina sequencing data a Phred score of 30 is the norm, which translates to chances of a base being called incorrectly of 1 in 1000. Phred scores are assigned to each base in each read, and these are stored in the fastq raw data files. The distribution of Phred scrores for each read position in a complete data set is usually depicted as a boxplot (Figure 4.9A). There are also other quality metrics that need to be considered, such as GC content, the per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, adapter sequences etc. A handy piece of software to use is FASTQC. You can find examples on https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. But an in-depth knowledge about sequence quality control is not required for this course.

Quality control of the mapped reads. Bad quality reads are usually discarded, and pieces of reads with low Phred scores are often removed. This process is called trimming and is not further discussed here. The reads can then be mapped to a reference genome, as described above. There are several ways to check the quality of the mapping result, but those are beyond the scope of this course. Yet, one "easy" mapping quality metric that should generally be checked is the percentage of mapped reads. For instance, if human RNAseq data is mapped on a human reference genome, then one would expect a >85% mapping rate.

Once expression values are calculated, and count tables are generated (see Figure 4.8B), then a lot of quality control occurs by "looking at" the data. However, large data files are difficult to an overview of, but it is handy to make plots. There are many types of plots that are useful. Some of which are:

Box plots/Violin plots. The boxplots can indicate the distribution of the gene expression values, indicated on the y-axis (Figure 4.9B). A boxplot is a plot that represents graphically several descriptive statistics of a given data sample. The boxplot usually has a box including a central line and two tails. The central line in the box shows the position of the median. The upper and lower boundaries of the box show the location of the upper quartile and lower quartile respectively. The upper and lower quartiles are the 75th and 25th percentiles respectively. Thus, the box will represent the interval that contains the central 50% of the data. The interval between the upper and lower quartiles is called the interquartile range (IQR). The length of the tails is usually 1.5×IQR. Violin plots are similar to box plots, except that they also show the probability density of the data (Figure 4.9C).

Histograms/density plots. A histogram or density plot is a graph that shows the frequency distribution of the thousands of expression values from a given sample, i.e. from one sequencing run for example. The horizontal axis of the plot spans the entire range of expression values. The vertical axis shows the frequency of each value. Usually, a histogram is represented as a bar graph. Histograms provide information about the shape of the distribution that generated the data. The histogram may be used as an empirical probability density function. Two experiments can be compared by constructing a histogram of the log2 ratio's, which represents the difference between the experiments. Density plots are "smoothed versions" of the histogram (Figure 4.9D). Yet, as they are plotted as "line-graphs", they are more suitable to visualize the distributions of expression values for many samples.

Scatter plots/MA plots. The scatter plot is probably the simplest tool that can be used to analyze gene expression levels. In a scatter plot, each axis corresponds to an experiment, and each expression level

corresponding to an individual gene is represented as a point. Scatter plots are very useful to convey information about direct comparisons (such as between samples), and they are frequently used in scientific papers. If a gene has an expression level of *x* in the first experiment and that of *y* in the second experiment, the point representing this gene will be plotted at coordinates (*x*,*y*) in the scatter plot. Genes with similar expression levels will appear somewhere on the first diagonal (the line *y*=*x*) of the coordinate system. A gene that has an expression level that is very different between the two experiments will appear far from the diagonal. Therefore, it is easy to identify such genes quickly. Scatter plots can also visualize certain characteristics of the data. For instance, for sequencing data has the property that the variability at low expression levels is higher than at high expression levels. This can be visualized by plotting the log foldchange between conditions (see below) on the y-axis against the average expression level on the x-axis. This is called an MAplot (Figure 4.9E), or else also sometimes called a ratio-intensity plot.

Principal component analysis. One very common difficulty in many problems is the large number of dimensions (= genes). In gene expression experiments each gene and each sample may represent one dimension. For instance, a set of 10 samples involving 20,000 genes may be conceptualized as 20,000 data points (genes) in a space with 10 dimensions (samples), or as 10 data points (samples) in a space with 20,000 dimensions (genes). Both situations are well beyond the capabilities of current visualization tools. A natural approach is to reduce the number of dimensions and thus, hopefully, the complexity of the problem, by eliminating those dimensions that are not "important". Of course, the problem now shifts to defining what an important dimension is. A common statistical approach is to pay attention to those dimensions that account for a large percentage of the variance in the data, and to ignore the dimensions in which the data do not very much. This is the approach used by Principal component analysis (PCA). PCA is for instance used to plot the samples against axes which explain most variance (PC1) or second most variance (PC2) in the gene expression data. By doing so, the number of dimensions have been reduced from 20,000 to two, and still most dynamic features in the data is visualized. PCA is therefore especially useful to check whether strange phenomena, grouping between samples and or variables occurs in the data (Figure 4.9F, see lectures). It worthwile to realize that principal components analysis is extremely useful to identify batch effects, and therefore extremely important for single cell RNA sequencing data analysis. Single cell RNA sequencing data analysis is often complicated by batch effects.

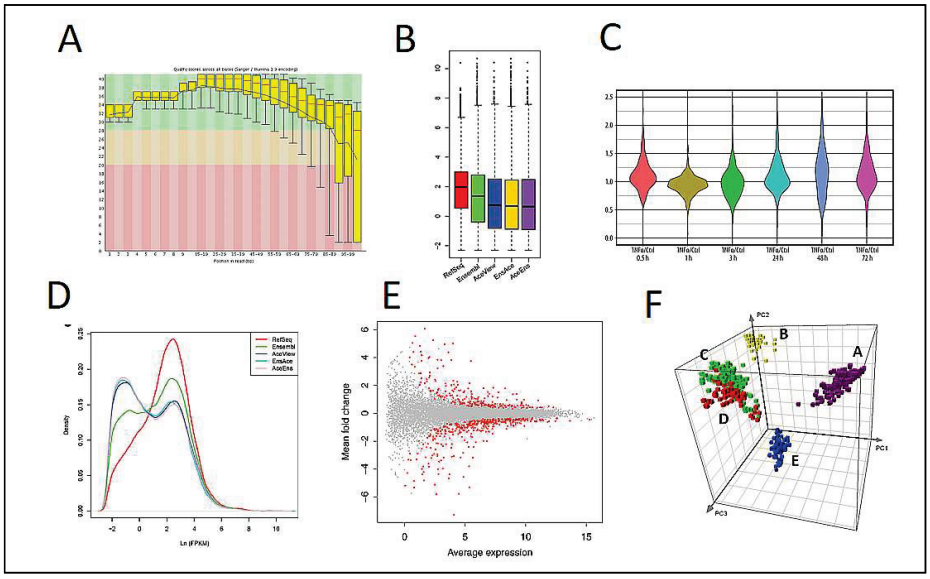


Figure 4.9 Visualization of transcriptomics data. Several examples of gene expression data visualization are shown here: The distribution of Phred scrores for each position in the read (A), boxplot (B), violin plot (C) density plot (C), MA plots (D) and a PCA plot (E). The PC's explain 41, 21 and 6% of the total variance, and the dots indicate samples from different tissues (purple: liver, yellow: blood, green: spleen, red: lymph node and blue: bladder). See text for more detailed descriptions.

4.5.6 Perform data processing and normalization

The aim of data preprocessing and normalization is to remove biases introduced by sampling and measurement. Data preprocessing involves removing genes that were measured with low numbers of reads, (logarithmic) transformation, eliminating outliers and data imputation. Only the logarithmic transformation will be discussed here. The logarithmic function has been used to pre-process transcriptomics data from the very beginning. There are several reasons for this. First of all: convenience. The difference between a treatment and a control is usually expressed as a ratio, or fold change. For each transcript the following is calculated: [treatment]/[control]. This leads to strongly asymmetric values: twofold induction leads to a value of 2 and a twofold repression leads to a value of 0.5. All values for the down-regulated genes are compressed in values between 0 and 1. Taking the log-transform with base 2 leads to symmetric values that are easy to interpret. A twofold induction leads to a value of +1 and a twofold repression leads to a value of -1. The distribution of log2(ratio's) is usually bell shaped, and that is nice to work with. A second reason is statistical. In raw data the variability depends on the average expression level: low expression values are measured with a small variability and high expression values are measured with a large variability. The log-transform stabilizes the variance. And finally, without log transformation, many analyzes will be dominated by a few highly expressed genes, obscuring (maybe relevant) differences between lowly expressed genes. The log transform reliefs this issue, because the log transform can be considered as a way to "stretch out" low range values and "compress" high range values.

There are three important notes to make, which concern RNAseq data. First, for RNA sequencing data one often uses a shifted logarithm $\log 2(n + 1))$, where n indicates the read count. This is because genes may have zero read counts. Second, there are other (variance-stabilizing) transformations that

are frequently used, but those are beyond the scope of the course. Finally, most methods for differential gene expression analysis are designed to work on counts (see below), which means that it is not necessary to transform the data for those statistical operations.

Data normalization is performed in order to remove technical, experimental biases and variability, while the biological variability should be maintained. Normalization procedures differ between transcriptomics platforms, and is still being researched for scRNAseq applications. For RNAseq there are a number of different types of normalization.

For RNA-seq data, the most prevalent bias is an overall difference in "sequencing depth", or in other words, an overall difference in a total number of (mapped) reads between the samples. The problem is shown in table 4.1. Please note that most genes have a higher expression level in sample 2 than sample 1. But that may be caused by the fact that sample 2 has more reads overall. There are different ways to normalize for this effect. One way is to perform a calculation to transform the number of reads for each gene by a value called "Reads per million". First, add up the total mapped reads in a sample, and this value represents the sequencing depth, and serves as a normalization factor. This value differs for each sample. Then, divide the read counts of each gene by this normalization factor. You have basically calculated the fraction of the total number of reads that has been attributed to each gene, and hence normalized for sequence depth. This value is usually quite a small number, so you can multiply it with a large number, typically 10^6, to get values that are easy to work with, and these values can be called reads per million (RPM). In principle this should normalize for differences in sequencing depth between samples. Although this is intuitively a correct approach, it can introduce biases, which is explained in the lectures. In practice, other more sophisticated methods can be used to normalize for differences in sequencing depth between samples.

Table 4.1. Normalizing read counts towards reads per million. The most left table shows the original read counts and the total number of mapped reads. Dividing the counts by this number yields corrected counts. Multiplying this number with 10^6 yields reads per million values.

	Original counts		Corrected counts		Reads per million	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
gene 1	10	25	0.0000215	0.0000397	21.5	39.7
gene 2	12	15	0.0000258	0.0000238	25.8	23.8
gene 3	14	66	0.0000301	0.0001047	30.1	104.7
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
gene 20000	10	11	0.0000215	0.0000175	21.5	17.5
Total Sum	465321	630215				

4.6 DATA ANALYSIS

4.6.1 Exploratory data analysis and batch effects.

The normalized gene expression matrix is usually used to perform exploratory data analyses. The same visualization tools as in the data quality control are used here: box-plots, scatter plots, MA plots, histograms, density plots and principal component analysis (or other multivariate analysis techniques). The goal of exploratory data analysis is to see whether strange phenomena, grouping between samples and or variables occurs in the data. In fact, it is common practice to go back to the

quality control and preprocessing and normalization phase if unexplainable artifacts are detected. The researcher may decide to discard samples, or use other normalization techniques to deal with these artifacts. Hence, data quality control, (pre)processing, normalization, and exploratory data analysis should be considered as one iterative procedure to work towards a reliable data set. The researcher will get to know the data very well during this process. The exploratory data analysis step is therefore of high importance.

One important thing to check for is batch effects. A batch effect occurs when non-biological factors in an experiment cause changes in the data produced by the experiment. Such effects can lead to inaccurate conclusions, and this needs to be addressed. Dimension reduction methods, like principal component analysis, can be used to recognize groups of samples in the data, which are indications for batch effects. In single cell RNA seq data, batch effects are an unavoidable property. Single-cell data is often compiled from multiple experiments with differences in capturing times, handling personnel, reagent lots, equipments, and even technology platforms. These differences lead to large variations or batch effects in the data, and can confound biological variations of interest during data integration. As such, effective batch-effect correction is essential. There are two ways to correct for batch effects: 1) account for batch effects in the statistical analysis for differentially expressed genes, and 2) correct for the batch effects during the (pre)processing of the data. In general it is most correct to use the first option and to include batch effects in the statistical test (as indicated below), hence this is usually done in the analysis of bulk RNAseq data. However, scRNAseq data sets are very large and and batch effects are often non-linear and complex, so scRNAseq data are usually batch corrected before the statistical analyses.

4.6.2 Differentially expressed genes

When the the data is considered trustworthy, a next analysis step is frequently to determine which genes are differentially expressed between different conditions. Once the researcher is investigating thousands of genes, it is straightforward to perform a fold-change cut-off. One could declare genes that are (on average) twofold upregulated, or twofold downregulated in response to a treatment, to be differentially expressed. It is however more elegant to perform a statistical test. Given the fact that thousands of genes are investigated, the researcher needs to perform gene-by-gene thousands of t-tests or thousands of ANOVA tests, depending on the experimental design. This may, however, have low power because the sample (number of replicates) size is often small. One other thing to consider is that the data usually is not distributed according to a normal distribution, making “standard” statistical tests in principle impermissible. RNA-seq data for instance, is frequently analyzed using specialized approaches that take into account that the data consists of discrete counts that follow a negative binomial distribution.

Another consequence of the low number of replicates is that the parameter estimates (averages, variances etc.) that are required for calculating test statistics, are not stable. To improve the reliability of the analysis, and improve the power, modified statistical tests are usually used for –omics data. For transcriptomics experiments, researchers generally use so called “moderated” statistical tests. These approaches are similar to, but not the same as, robust statistics, which are more resistant to outliers. They are designed to deal with data containing a low number of replicates that measure an extreme large number of variables. For this course you don’t need to know how these statistical tests are calculated, but please remember that for gene expression profiling modified t-tests and modified F-tests should be used that are optimized for genome wide gene expression data. Established software

packages are DESeq2 (Love at al., 2014) , edgeR (Robinson et al., 2010), and limma (Smyth 2004). Note that these software packages enable you to include batch effects, by setting up statistical models in an ANOVA-kind of way, taking into account multiple experimental factors. We will elaborate a bit more on this subject in the lectures.

The result of a differential gene expression analysis is a p-value for each gene, indicating the probability of your finding, given that this gene is NOT different between the experimental conditions. The researcher usually decides on a certain significance level which determines differential regulation. You may recall from your statistics lessons that this significance level is called α. The results of the statistical test can easily be visualized in a volcanoplot of MAplot (Figure 4.10).

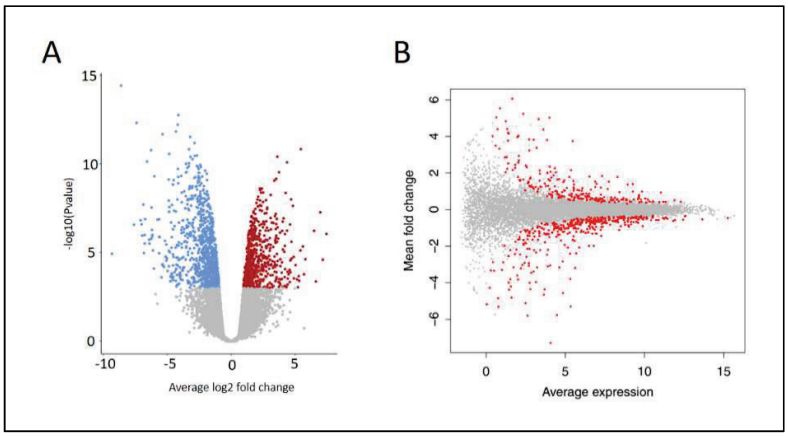


Figure 4.10 Visualizations of differentially expressed genes. Volcano plots, where the negative log10-transformed p-values are plotted against the log2 fold change (log ratios) in a two-group experiment (A), and the MAplot, where the log2 fold changes are plotted against the log2 average expression. The dots indicate genes, and colored dots represent the significantly differentially expressed genes based on an α cutoff.

4.6.3 Multiple testing correction

The significance level α is defined as the acceptable probability of making a Type I error (typically 5%, but this is an arbitrary choice). This corresponds to a situation in which the null hypothesis, no difference, is rejected when it is in fact true. The variables (transcripts, genes) that are called differentially expressed when in fact they are not are called false positives.

Now let us think in terms of hypothesis testing. When the t-statistic is more extreme than the threshold t_α, we will call this variable differentially regulated. However, the t-statistic may have extreme values just due to random effects. This will happen with probability α. If this happens and we will call this variable differentially regulated, we will be making an erroneous decision. Thus, the probability of making a mistake of this kind is α. However, if we do not make a mistake, we will be drawing the correct conclusion for that given variable. This will happen with probability:

Prob(correct) = 1-p

As indicated, transcriptomics experiments generate values for many variables, so now we have to take into consideration the fact that there are many tests performed: one for each variable. Suppose that we have performed a microarray experiment. Let us consider there are R genes. For each of them we

will follow the same statistical reasoning as described above. However, in the end we would like to draw the correct conclusion for all of them. This means we would like to have the correct conclusion for the first gene AND the second gene AND the third gene AND ⋯ AND the last gene. Obviously, the probability of such an event is the multiplication of the probabilities of the individual events. Therefore, the probability of drawing the correct conclusion for all tested genes is:

$$\text{Prob(globally correct)} = (1-p)\times (1-p) \cdots (1-p) = (1-p)^R$$

We can now calculate the probability of being wrong somewhere. This would be one minus the probability of being correct for all tests.

$$\text{Prob(wrong somewhere)} = 1 - \text{Prob(globally correct)} = 1 - (1-p)^R$$

In this situation, being wrong means drawing the wrong conclusion for at least one gene. This is in fact the α value for the whole –omics experiment. Table 1 shows the values of this probability for various significance levels and various sizes of an array. An array with as few as 20 genes generates a probability of 87.84% of having at least one false positive if the gene level test is performed at a significance level of 0.1. For an array with 100 genes, the probability becomes 99.99%. Although this is worrisome, the table does not paint the whole picture. After all, having a false positive from time to time may be deemed to be acceptable. The question becomes how many such false positives are acceptable for a given array size and gene level significance. For instance, table 2 shows that a small array with 5000 genes on which the gene level significance is performed at 0.01 significance level is expected to produce 500 false positives. These 500 false positives mix up with whatever true positives there are in a given condition. So, the task is to control the global or experiment level significance level. This is the probability of having a Type I error anywhere. This probability is also known as the family wise error rate.

Table 1. The probability of making a type I error (false positives)

# genes \ p-value	0.01	0.05	0.1	0.15
10	0.10	0.40	0.65	0.80
50	0.39	0.92	0.99	1.00
100	0.63	0.99	1.00	1.00
500	0.99	1.00	1.00	1.00
1000	1.00	1.00	1.00	1.00
5000	1.00	1.00	1.00	1.00
10000	1.00	1.00	1.00	1.00

Table 2. The expected number of false positives

# genes \ p-value	0.01	0.05	0.1	0.15
10	<1	<1	1	1.5
50	<1	2.5	5	7.5
100	1	5	10	15
500	5	25	50	75
1000	10	50	100	150
5000	50	250	500	750
10000	100	500	1000	1500

In conclusion, technologies that generate high dimensional data such as transcriptomics are prone to generating many false discoveries. It is therefore desirable to apply statistical procedures that can control these errors. The “basic” methods are Sidak correction, Bonferroni correction and Benjamini Hochberg False Discovery Rate correction, but many more methods exist.

4.6.3.1 The Sidak correction

The sidak correction is one method to control the overall probability of making a Type I error. This probability is equal to the probability of making at least one such mistake. Recall that

$$\text{Prob(wrong somewhere)} = 1 - (1-p)^R$$

Which can be rewritten as:

$\alpha_e = 1- (1- \alpha_c)^R$

where α_e is the probability of a Type I error at the experiment level, and α_c is the probability of a Type I error at the gene level (single comparison). The task is to calculate the α level that we need to use for individual genes (α_c) in order to ensure that the global, or experiment level type I error is less or equal to α_e. Using simple algebraic manipulations, we can extract α_c from the equation as:

$\alpha_c = 1- \sqrt[R]{(1- \alpha_e)}$

This is the Sidak correction for multiple comparisons.

4.6.3.2 The Bonferroni correction

Bonferroni noted that for small values of p, the overall a probability of making Type I error can be approximated (using binomial expansion) as follows:

$\alpha_e = 1- (1- \alpha_c)^R \approx R \times \alpha_c$

Using this approximation, the gene level α_c to control the overall a probability of making Type I error can be calculated as:

$\alpha_e = R \times \alpha_c \rightarrow \alpha_c = \alpha_e / R$

This is the Bonferroni correction for multiple comparisons. This is a very simple formula but it is only an approximation of the exact value given by Sidak. Bonferroni starts to depart from the exact values even for as few as 20 genes. In general, both Bonferroni and Sidak are not really suitable for microarray studies. Because for large numbers of R, the required significance at the gene level becomes very small very quickly. If an experiment level α_e of 0.05 is required for an array with only 1000 genes, the gene specific significance cutoff α_c is 0.00005. At such stringent significance levels, the hypothesis testing approach will not be able to reject the null hypothesis for many genes. Bonferroni and Sidak are conservative methods in the sense that if a gene is significant after either Bonferroni or Sidak adjustments, then the gene is truly different between the groups. However, among the non-significant genes many may still be truly different. In statistical terms, Bonferroni and Sidak are sufficient but not necessary conditions.

4.6.3.3 The false discovery rate (FDR) correction

Stepwise methods such as the false discovery rate correction allow for less conservative adjustments of the p-values. These methods order the genes in increasing order of their p-values and make successive smaller adjustments. The FDR procedure was initially proven for independent variables, but was recently extended to allow for some dependencies. This is helpful, as genes are actually known to be involved in complex dependencies and regulatory mechanisms.

The false discovery rate correction procedure works in principle as follows:

1) Choose the experiment level significance level α_e.

2) Order the variables (transcripts/genes) in the increasing order of individual p-values:

Variables	g_1	g_2	⋯	g_k	⋯	g_R

Increasing p-values	p_1	p_2	⋯	p_k	⋯	p_R

3) Compare the p-values of each variable with a threshold that depends on the position of the variable in the list of ordered values. The thresholds are as follows: $1/R \times \alpha_e$ for the first variable, $2/R \times \alpha_e$ for the second, etc.

Variables	g_1	g_2	⋯	g_k	⋯	g_R
Increasing p-values	p_1	p_2	⋯	p_k	⋯	p_R
Test	$p_1<\alpha_e/R$	$p_2<2\alpha_e/R$	⋯	$p_k<k\alpha_e/R$	⋯	$p_R< \alpha_e$

4) Reject the null hypotheses for those variables that have a p-value lower than their corresponding threshold. These genes are then considered different between the two groups at a chosen α_e significance level.

4.6.3.4 Interpretation of the outcome

In transcriptomics thousands of individual tests are performed in parallel. In order to avoid an overwhelming amount of false discoveries (variables that are called differentially expressed when in fact they are not) some sort of multiple testing correction is required. You have seen that there are different procedures available, and some are stringent (Bonferroni) and some are more lenient (FDR). It should be noted that Bonferroni and Sidak are fundamentally different from the FDR methods. Bonferroni and Sidak control what is called the family-wise error rate (FWER). The FWER is the probability of at least one Type I error. You can choose an FWER method if high confidence in all selected genes is desired. If a certain proportion of false positives is tolerable, then procedures based on FDR are more appropriate. The interpretation of the FDR is: the expected proportion of Type I errors among the rejected hypotheses. For instance, if you find 100 differentially expressed genes after performing FDR correction with an α_e=0.05, then you can expect to have five false positives among these 100 genes. The FDR is most often used in gene expression analysis.

4.7 BIOLOGICAL INTERPRETATION

4.7.1 Cluster analysis

Cluster analysis is very useful to find patterns in the data that enable biological interpretation. For instance, one obvious thing to investigate is the correlation in expression (co-expression) of genes across the experimental conditions. They may represent modules associated with certain biological functions. Cluster-analysis algorithms are helpful, as they group objects on the basis of some sort of similarity metric that is computed for one or more 'features' or variables. For example, k-means clustering groups genes into classes on the basis of the similarity in their expression profiles across tissues, cases or conditions. Hierarchical cluster analysis graphically presents results in a tree diagram (dendrogram), and is probably the most common unsupervised classification algorithm in microarray analysis. Nevertheless, the researcher should be aware of the fact that results from cluster analysis depend on assumptions made by the algorithm and the results may therefore not be consistent. Non-hierarchical clustering methods, such as k-means clustering, divide the cases (samples or genes) into

a predetermined number of groups in a manner that maximizes a specific function (for example, the ratio of variability between and within clusters). Cluster-analysis approaches entail making several choices, such as which metric to use to quantify the distance or similarity among pairs of objects, what criteria to optimize in determining the cluster solution, and how many clusters to include in the solution. The results depend on these choices, and no consensus or clear guidelines exist to guide these decisions. Cluster analysis always produces clustering, but whether a pattern observed in the sample data characterizes a pattern present in the population remains an open question (Allison et al., 2006).

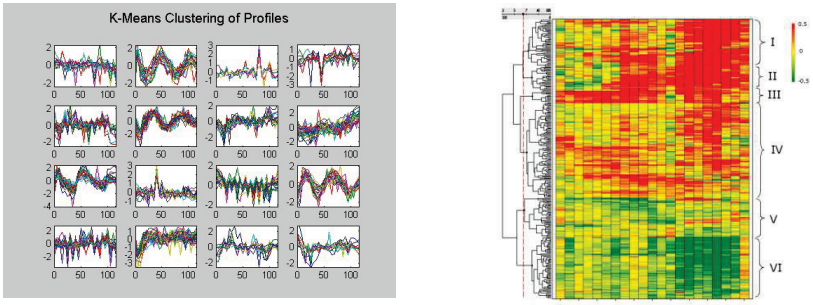


Figure 4.11 Examples of k-means clustering (left) and hierarchical clustering (right). The k-means picture is adapted from http://www.computational-genomics.net/case_studies/cellcycle_demo.html

4.7.2 Analyzing the response of sets of genes.

Extracting biological information from a transcriptomics can be challenging given the high number of data points. It can be helpful to analyze if a set of genes, which are known to have a specific biological functional relationship, are changing together in response to the experimental condition(s) tested. In transcriptomics, such sets of genes could be a set of target genes of a particular transcription factor, or a set of genes involved in a particular biological pathway or process, such as nucleotide excision repair. If the genes in a set are changing in concert, or in other words, if many genes from this set of genes are differentially expressed, then the conclusion could be that the biological process is involved in the transcriptomics response.

4.7.2.1 Functional relationships in transcriptomics.

Knowledge of functional relationships between genes, between proteins or between metabolites is basically a result from life sciences research and can be extracted from scientific literature. However, suppose you had performed a RNAseq experiment, it would then be a major task to identify all functionally related sets of genes from literature that could be important for your transcriptome experiment. It is then worthwhile to make use of the knowledge of functional relationships that is compiled by other researchers. There is a lot of information available of relationships between genes/proteins/metabolites via the internet in (public) data bases. Some information is free and for some information you need to pay. Some information is browsable, and some is not. Either way, it is always advisable to inform yourself where the information comes from and with what rationale functional relationships are defined. Two well-known data bases, KEGG and the Gene ontology, are

discussed here as examples.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway data base contains curated metabolic and signaling pathways. Information on enzymetic reactions, enzymes, small molecules and genes is also available from KEGG. Pathways are available as searchable and clickable images called maps. Pathway maps can depict metabolism, regulatory pathways, and large complexes such as the ribosome. Most metabolic pathways maps are reference maps that depict generalized pathways. Generalized pathways are not species specific, thus may never be found in their entirety in a single species. You can however select to highlight the enzymes on the generalized map that are present in an organism of interest.

To understand the Gene Ontology (GO) one needs to realize what an ontology is: a controlled vocabulary of terms. The GO consortium is a project that compiles a dynamic, controlled vocabulary of terms related to different aspects of genes and gene products (proteins). The consortium was begun by scientists associated with three model organism databases: the Saccharomyces Genome Database, the Drosophila genome database ((FlyBase), and the Mouse Genome Informatics data bases. Subsequently, data bases associated with many other organisms have joined the GO Consortium. The GO data base is not centralized per se, but instead relies on external data bases in which each gene or gene product is annotated with GO terms. Thus it represents an ongoing, cooperative effort to unify the way genes and gene products are described. There are three main organizing principles of GO: 1) molecular function, 2) biological process, and 3) cellular compartment. Molecular function refers to the tasks performed by the individual gene products. For example, a protein can be a transcription factor or a carrier protein. Biological process refers to the broad biological goals that a gene product is associated with, such as mitosis or purine metabolism. Cellular component refers to the subcellular localization of a protein. Examples include nucleus and lysosome. Any protein may participate in more than one molecular function, biological process and/or cellular component. Genes and gene products are assigned to GO categories through a process of annotation. The author of each GO annotation supplies an evidence code that indicates the basis for that annotation (see lectures).

4.7.2.2 Gene set tests

The data bases yield sets of genes that probably have a functional relationship. Hence, it is possible that a researcher wants to test the response of all apoptosis related genes to the conditions tested. Quantifying the association between the expression of the genes of the given gene set and the studied experimental conditions, requires a statistical test. There are several different gene-set tests, and many of those tests are similar to the hypergeometric test, based on the hypergeometric distribution. Basically, the hypergeometric test tests whether genes from a certain set are overrepresented in your list of differentially expressed genes. For instance, given the list of differentially expressed genes in an experiment, are there more DNA-repair genes differentially expressed than expected by chance alone. Please note, there are many types of gene set tests available through the internet, and although these types of gene set tests may look like one another, they cab in fact be very different and they do not necessarily yield the same results.

Let´s explain the test based on the hypergeometric distribution. Let us consider there are N genes on the array that are measured. Any given gene is in either functional category F or not. In other words,

the N genes are of two categories: F and non-F (NF). This is similar to having an urn filled with N balls of two colors, such as red (F) and green (NF). M of these balls are red and N-M are therefore green, or similarly, M of these genes are F and N-M are therefore NF. Suppose that K genes are differentially expressed (fdr corrected p-value ≤ 0.05), and that x of these K genes are belonging to functional category F. Basically, we want to determine the probability of this happening by chance. The probability that a category F occurs exactly x times just by chance in the list of K differentially expressed genes is appropriately modeled by the hypergeometric distribution with parameters (N,M,K). This distribution can be used to calculate a p-value, indicating whether the x genes of category F are overrepresented in the list of differentially expressed genes or not. If the p-value is below a certain threshold, then more genes of category F are differentially expressed than expected by chance alone (Huang et al., 2009).

A disadvantage of this calculation procedure is the cutoff, which needs to be set to determine the list of differentially expressed genes (K). Basically, any cutoff is arbitrary. It may also happen that established significance levels, such as fdr corrected p-value ≤ 0.05, no differentially expressed genes are found. It may then be useful to apply less conservative methods such as gene set enrichment analysis (Subramanian et al., 2005), which is not discussed in this course.

4.7.2.3 Remarks on gene set testing

Over the past few years many different types of genes set tests have been published and this is still subject to active methodological research. One issue is for instance that the gene sets are not independent: many gene sets are a subset of another gene set. For instance, in the Gene Ontology it is obvious that “positive regulation of apoptosis” is a subset of: “regulation of apoptosis”. Finally it is worth noting that if many (thousands) of gene sets are tested that officially a multiple testing correction needs to be applied. However, the dependencies between gene sets make it hard to calculate exact fdr corrected p-values. Nevertheless, the methodology has proven to be very useful for biological interpretation of the results from a gene expression profiling experiment.

4.8 REFERENCES

- Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. 2006. Nat Rev Genet. 7(1):55-65.
- Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. 2003. Genome Biol. 4(4):210.
- Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. 2009. Nucleic Acids Res. Jan;37(1):1-13.
- Schulze A, Downward J. Navigating gene expression using microarrays--a technology review. 2001. Nat Cell Biol. 3(8):E190-5.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. 2005. Proc Natl Acad Sci 102(43):15545-50
- This list is unfortunately not complete, due to time constraints.

5 An introduction to Genome Wide Association Studies (GWAS)

Lecturer: Dr. Martijs Jonker (SILS)

This chapter will be distributed through Canvas

6 Proteomics technologies and the study of disease

Lecturer: Dr. Dave Speijer (Amsterdam University Medical Centers, AMC)

After reading this chapter you should

- Have a good understanding of the basic techniques of mass spectrometry (MS)
- Know about the specific processes of ionization and detection in MS
- Understand the specific problems in quantitative analyses of proteins

Contents

This chapter is complemented by two papers that are added to the end of this chapter:

- Van Oudenhove L., Devreese B. (2013) A review on recent developments in mass spectrometry instrumentation and quantitative tools advancing bacterial proteomics. Appl Microbiol Biotechnol 97:4749-62.

- Aebersold R., Mann M. (2016) Mass-spectrometric exploration of proteome structure and function. Nature 537:347-55.

6.1 Introduction: proteomics technologies. The use of mass spectrometry in the quantitative analysis of complex protein samples

During this course a few lectures will deal with the complex analysis of biological samples containing highly complex protein mixtures. The only technique allowing large-scale analysis and **quantification** of protein contents is **mass spectrometry (MS)**. Thus, an introduction to the basic techniques in use is a prerequisite to understand **proteomics**. Using these basic techniques with proteins will only work for specific forms of **ionization (ESI and MALDI).** To be able to get accurate information regarding complex mixtures of rather large molecules (whether peptides or proteins) needs very high **resolution** (highly accurate mass separation and sensitive detection). Therefore, **the mass separation/detection methods** used, will also be (briefly) discussed. Apart from the fact that large (!) ions have to be detected, the specific chemical properties of complex protein mixtures pose their own specific challenges. To understand these, some knowledge regarding such mixtures is crucial. In this context, **dynamic range** and **post-translational modifications (PTM's)** are important. Going from relatively simple, "old-fashioned", protein identification of purified proteins, using **peptide mass fingerprinting (PMF)**, we will travel to full-scale proteomics using automated **LC-MS/MS analysis**. In the last example, the second round of MS analyses the fragments obtained from the ion measured in the prior analysis, generated by controlled **fragmentation**, because very little (structural) information can be gained from the simple mass spectrum of the ion on its own. All the basic concepts will be illustrated with real life examples and the students will be challenged to interpret mass-spectrometric data themselves. To stress the (clinical) significance most examples discussed will come from experiments performed in the areas of HIV-1/AIDS and Neisseria meningitidis/Meningitis research.

6.2 Proteomics

Proteomics, the large-scale study of proteins, uses mass spectrometry to identify and quantify proteins in complex biological mixtures. These mixtures are referred to as proteomes. In a more restricted sense, the proteome represents the entire set of proteins to be found in an organism. This already tells us that it is a highly dynamic concept, as what is produced or modified by an organism, at any one moment, can differ tremendously. In a less restricted sense, the proteome represents the entire set of proteins that can be found in any system. Therefore, we can talk about e.g. human plasma proteomics, mitochondrial proteomics or even chicken feather microbiome proteomics. The further technological advances in proteomics have enabled the identification of ever-increasing protein numbers. However, it should be stressed that a really complete (!) analysis of all proteins, their modifications and differently processed forms, as well as their absolute concentrations in any given somewhat complex mixture is not possible. Proteomics is an interdisciplinary field that has grown in conjunction with the genetic information of many genome projects, of course especially the Human Genome Project. Of note, proteomics generally refers to large-scale experimental

analysis of proteins and proteomes, but most often is used specifically to refer to protein purification (which will only be discussed in passing) followed by mass spectrometry.

6.3 Mass spectrometry

Mass spectrometry (MS) is a technique to analyse ionized chemical species that sorts ions based on their mass-to-charge (M/Z) ratio. Thus, a mass spectrum gives the masses within a sample. Mass spectrometry is used in physical, (bio) chemical and medical analyses and can be used to study pure as well as complex samples. When analysing the latter some kind of separation (mostly chromatographic) is needed prior to analysis. A mass spectrum plots ion signals on a Y-axis (absolute or relative amounts of ions) as a function of the mass-to-charge (M/Z) ratio (indicators of mass; X-axis). In the live sciences, such spectra can be used to determine the masses of biologically relevant particles and molecules, and to elucidate their chemical structures.

In an MS experiment, a sample, either solid, liquid, or gas, is ionized using an ionization technique. Actually, getting molecules into an ionized state to fly of into a partial vacuum (at a well-timed moment) so their behaviour can be measured in real time to obtain their (M/Z) ratio can be quite challenging. The older ionization methods (for example bombarding the sample with electrons, electron impact (EI), or with atoms, fast atom bombardment (FAB)) may cause some of the sample's molecules to break into charged fragments. In the life sciences, “softer” ionization methods are needed: matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI). Next, the resulting ions are separated according to their (M/Z) ratio by subjecting them to electric/magnetic fields: ions of the same mass-to-charge ratio will undergo the same amount of deflection/acceleration. Following this, ions can be detected by mechanisms capable of detecting charged particles, such as electron multipliers. Results are given by spectra as described above. The molecules in the sample can be identified by correlating known masses (of the entire molecule and/or its fragments) to the ones found in the mass spectrum (most often, using characteristic fragmentation patterns).

6.3.1 Ionization (ESI and MALDI)

ESI

As mentioned above, for protein identification and analysis, “soft” ionization methods are needed: matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI). Electrospray ionization (ESI) is the technique used in MS to produce ions using an electrospray in which a high voltage is applied to a liquid to create an aerosol of tiny (nano) droplets. With further evaporation they reach the so-called Rayleigh limit and ions enter the gas phase to be analysed. It is a crucial method for producing ions from macromolecules in protein study (and other biomolecules) as it overcomes the tendency of such molecules to fragment when ionized (remember, ESI is a so-called “soft” ionization technique). ESI differs from other ionization processes (such as matrix-assisted laser desorption/ionization; MALDI) as it mostly

produces multiple-charged ions (M/Z with Z equal to or > 2), thereby extending the mass range of MS machinery. Using this method, the kDa-MDa (10^3 -10^6) range of MW encountered in proteins (and the polypeptide fragments derived from them) can easily be analysed.

MALDI

The other important technique in protein mass spectrometry is matrix-assisted laser desorption/ionization (MALDI). It is an ionization technique using laser energy absorbing matrix molecules to form ions from large molecules, again with minimal fragmentation. In most cases small organic, acidic compounds are used (e.g. α-Cyano-4-hydroxycinnamic acid). It can be used for the analysis of all kinds of biomolecules (apart from proteins and peptides, also biopolymers such as DNA and sugars can be analysed). All such molecules tend to be fragile, fragmenting when ionized by the older ionization methods. In contrast to ESI, MALDI typically produces hardly any multi-charged ions (Z = 1). MALDI methodology consists of three steps. To begin with, samples are mixed with suitable matrix compounds (dissolved in organic solvents) and allowed to dry on a metal plate. Next, a pulsed laser (using light of a wavelength which will very efficiently excite matrix molecules) irradiates the sample, triggering ablation/desorption of both sample and matrix molecules. Finally, the molecules to be analysed (in proteomics: proteins and peptides) are ionized by being protonated (in some cases, using other matrix compounds, deprotonated) in the high temperature plume of ablated gases. The charged particles (cations in the positive mode; anions in the negative mode) are then accelerated into the mass spectrometer to be measured.

6.3.2 Mass separation/detection methods

Mass analysers separate ions according to their M/Z ratio. Two laws can be applied to the behaviour of ions in electric and magnetic fields in a "vacuum":

$F = Q (E + v \times B)$ The so-called Lorentz force law;

$F = ma$ Newton's famous second law.

In this context, F = force applied to the ion, m = mass of the ion, a = acceleration, Q = ion charge, E = electric field, and v × B = the "vector cross product of the ion velocity and the magnetic field"[1].

Using both of the above expressions for F applied to ions gives us:

$(m / Q) a = E + v \times B.$

[1] for those interested, see e.g.: https://en.wikipedia.org/wiki/Maxwell%27s_equations

How these expressions are used to calculate the (M/Z) ratio of the ions depends on the type of machine used. I will only very briefly discuss the most commonly used (for proteomic studies) types here.

6.3.3 Time-of-flight (TOF)

The TOF analyser uses an electric field (e.g. a voltage difference of 15 kV between plate and detector) to accelerate ions through this potential, measuring the time they take to reach the detector. If particles have the same charge their acceleration will depend only on their masses. Ions with lower masses will hit the detector (e.g. a micro-channel plate; MCP) first. In other words, from F = ma we can deduce that, because the potential and charge are equal F is identical for all ions with the same charge, Z, the mass difference has to lead to acceleration differences that are proportional to the mass. However, this proportionality is only maintained if ions do not differ in initial velocities. Even particles with the same m/z might arrive at different times at the detector, because of differences in initial velocities. To fix this, so-called delayed extraction (all ions starting to move at the "same" time) is standard in ion sources with TOF-MS.

6.3.4 Quadrupole

A quadrupole (Q) mass analyser is a type of analyser consisting of four cylindrical rods, in a parallel configuration. Ions are separated in a quadrupole based on the stability of the trajectories of the selected ions from the sample due to their mass-to-charge ratio (M/Z) in oscillating electric fields which can be applied to the rods. Hexapoles and octupoles have six and eight rods respectively.

6.3.5 Ion trap

Ion traps use combinations of electric and/or magnetic fields to capture ions. The three most commonly used ion traps are the so-called Penning trap, using a potential made from a combination of electric and magnetic fields, the Kingdon trap (see below), and the Paul trap (Wolfgang Paul; Nobel prize 1989) forming its potential using a combination of static and oscillating electric fields. This kind of ion trap can be thought of as a three-dimensional quadrupole mass analyser.

6.3.6 Orbitrap

The Orbitrap is the most recent (and highly successful) development in (proteomic) mass spectrometry. It is based on a Kingdon trap arrangement. This trap, originally with a very thin central wire, uses an outer cylindrical electrode and (isolated) cap electrodes at the ends. Static applied voltages result in radial logarithmic potentials between the electrodes. In the Orbitrap configuration, invented by Alexander Makarov, an outer barrel-like electrode and a coaxial inner spindle-like electrode trap ions in orbital motions around the central spindle.

The currents generated by trapped ions are detected and converted to mass spectra using the Fourier transform of their frequency signals (compare Fourier-Transform Ion Cyclotron Resonance Spectrometry; FTICR, which uses Penning traps and strong magnetic fields instead).

6.3.7 Coupled (hybrid) MS instruments

In modern day Proteomics most instruments used are multiples of mass spectrometers (if different types are combined we talk of hybrids). Exceptions to the rule are found in the single ion-trap or TOF machine. The advantage of multiples is that one can use a low to medium resolution (see below) MS machine to select ions for further characterization by high resolution mass analysis of fragments resulting from controlled fragmentation (see below) using the next machine. The stereotypical example is a Q-TOF.

6.4 Peptide mass fingerprinting (PMF)

Peptide mass fingerprinting (PMF) is a (nowadays somewhat old-fashioned) analytical technique for protein identification. Unknown proteins of interest are purified and cleaved into smaller peptides (mostly using trypsin). The masses of the (tryptic) digest are measured with a high-resolution mass spectrometer such as a MALDI-TOF or ESI-TOF. The peptide masses obtained are then compared to databases of protein sequences derived from predictions based on known genomes (obtained with genomics techniques), using programs that:

1. translate the genome of the organism studied into proteins,
2. theoretically cut the proteins into peptides (considering proteolytic preferences of the protease used; in the case of trypsin, cutting after K (lysine) or R (arginine) unless the next residue is a P (proline),
3. calculate the absolute masses of the peptides resulting from such an “in-silico digest” of each protein predicted in the organism studied,
4. and compare the measured masses of the peptides derived from the unknown protein to the predicted theoretical peptide masses of each protein encoded by its genome.

6.5 Dynamic range

Protein identification using MS technology, as described succinctly above, has emerged as a very powerful tool for analysing large-scale proteomes. But quite a number of challenges remain for proteomics research. Probably, the very high degree of protein complexity and, especially, the huge dynamic range of proteins expressed in the complex biological mixtures, will remain problematic in the (near) future. The dynamic range, here referring to the difference in abundance of the most numerous protein(s) as compared with the least abundant one(s), certainly exceeds six orders of magnitude in cells and ten orders of magnitude in body fluids. Alas, quite a few highly important proteins have low expression

levels, so methods to detect them, even in complex samples, are necessary. Mostly this is undertaken by depleting the samples of the most numerous protein(s), or by specifically purifying low-abundant proteins of interest, with mixed results.

6.6 Post-translational modifications (PTM's)

An extra source of difficulties for proteomic MS techniques is found in the very frequent presence of post-translational modifications (PTMs). These refer to covalent modification of specific amino acids in proteins after protein biosynthesis. Thus, after protein synthesis by ribosomes translating mRNA into polypeptides, they very frequently undergo PTM during maturation, (in)activation, and/or breakdown. Also, different kinds of processing can occur (e.g. removal of N-terminal signal peptides of proteins targeted to mitochondria). A physiologically important example is the regulation of protein components in cell signalling, e.g. via phosphorylation of S (serine), T (threonine) and/or Y (tyrosine) on target signal cascade intermediates or when prohormones are converted to hormones. The regulation of gene expression is influenced by histone modifications such as acetylation and methylation. These examples are just the tip of the iceberg[2].

6.7 Fragmentation and LC-MS/MS analysis

It might seem a bit of irony, that after taking care not to have the molecules fragment during the ionization process, now effort is put into obtaining fragmentation. The difference is of course that controlled fragmentation of measured ions is the goal here, as these fragments (product/daughter ions) allow highly confident identification of the original precursor/parent ion. The desired fragmentation is obtained in the collision zone of the mass spectrometer. The mechanisms are studied as part of gas phase ion chemistry[3].

For proteomics 3 types of mass fragmentation are most important: collision-induced dissociation (CID), electron-capture dissociation (ECD), and electron-transfer dissociation (ETD). Especially the first one is standard. A highly pure collision gas (e.g. argon or nitrogen) is allowed into the collision zone. In this fashion one gets collisions in the energy range that will most likely give (on average) one break in the peptide backbone, resulting in a mixture of mostly b-ions (containing the N-terminus) and y-ions (containing the C-terminus). From this mixture the sequence of the fragmented peptide can be reconstructed. This, in essence, constitutes the important MS peptide sequencing technique.

The most commonly used technique in proteomic analysis uses liquid chromatography–mass spectrometry (LC-MS or LC-MS/MS). Here complex protein or peptide mixtures (obtained via digestion) are analysed by combining techniques for physical separation (i.e. liquid chromatography; FPLC or HPLC) with mass analysis on a directly coupled mass spectrometer

[2] Much more information can be found in: https://en.wikipedia.org/wiki/Post-translational_modification
[3] see e.g.: https://en.wikipedia.org/wiki/Fragmentation_(mass_spectrometry)

(the MS or MS/MS part, in the latter instance fragmentation and subsequent measurement is involved). Coupling occurs mostly by having the eluate of the column (mostly using reverse phase chromatography, with the peptides becoming more and more hydrophobic as higher concentrations of organic solvents are needed to release them from the column) spray into the MS system, using nano-spray ESI. The techniques enhance each other synergistically, getting information about > 3000 proteins in the mixture with a (typical) thirty minutes run.

6.8 Resolution

In MS, resolution (resolving power) "R", refers to measures reflecting the ability to distinguish between peaks of slightly different (M/Z) ratios ΔM, in mass spectra. Let's assume for simplicity's sake that Z =1. R = M/ΔM. A medium-resolution machine might be able to resolve peaks between 5000 and 5001 Dalton (Da); meaning it has an R of 5000, while a high-resolution machine can even resolve a one Da difference at 20 kDa (R = 20000) or higher.

6.9 Quantification

After identification, quantitative proteomics is needed to determine the (relative) amounts of the identified proteins in a sample. Methods for protein identification are the same as those used in general, qualitative, proteomics. However, to be able to include quantification as an additional source of information, extra conditions have to be fulfilled. For clinical and biochemical applications absolute quantitative measurements are in general not important (though feasible). Relative measurements, however, can be crucial. Does protein "X" abundance change significantly in a certain disease condition or upon a specific cellular signal are the kind of questions quantitative proteomics tries to answer. Thus, instead of just providing extensive lists of proteins identified, quantitative proteomics gives information regarding physiological differences between two or more biological samples. Because the amount of proteins or peptides seen in mass spectrometry is mostly determined by ionization efficiency (the ease with which a certain protein/peptide picks up charge), and only somewhat by the abundance in the sample, quantitative analysis is challenging. These problems can be circumvented by labelling techniques (e.g. using mass isotopes that can identify from which sample condition a certain ion results, using e.g. ^{13}C, ^{15}N, ^{18}O) and/or applying high quality label-free AMRT ("accurate mass retention time") approaches. These modifications of standard proteomic techniques will be explained during the lectures.

6.10 Concluding remarks

MS analysis of proteins (proteomics) is a complex, rapidly developing field, in which new approaches and possibilities come up with surprising speed. Some of these new developments (e.g. the application of bio-orthogonal amino acids, such as azidohomoalanine, which allows the specific (!) purification of newly translated protein products upon cells receiving a signal, or complex applications of ITRAQ labelling) will come up at the end of the

lectures and in the problem-solving group challenges. These will also focus on obtaining critical understanding of the applicability and limits of certain types of MS approaches to specific research questions.

REVIEW

doi:10.1038/nature19949

Mass-spectrometric exploration of proteome structure and function

Ruedi Aebersold[1,2] & Matthias Mann[3,4]

Numerous biological processes are concurrently and coordinately active in every living cell. Each of them encompasses synthetic, catalytic and regulatory functions that are, almost always, carried out by proteins organized further into higher-order structures and networks. For decades, the structures and functions of selected proteins have been studied using biochemical and biophysical methods. However, the properties and behaviour of the proteome as an integrated system have largely remained elusive. Powerful mass-spectrometry-based technologies now provide unprecedented insights into the composition, structure, function and control of the proteome, shedding light on complex biological processes and phenotypes.

Collectively, proteins catalyse and control essentially all cellular processes. They form a highly structured entity known as the proteome, the constituent proteins of which carry out their functions at specific times and locations in the cell, in physical or functional association with other proteins or biomolecules. A proliferating *Schizosaccharomyces pombe* cell contains about 60 million protein molecules, which have abundances that range from a few copies to 1.1 million copies per expressed gene[1]. Across the species, proteins constitute about 50% of the dry mass of a cell and reach a remarkable total concentration of 2–4 million proteins per cubic micrometre or 100–300 mg per ml (ref. 2). The extensive proteome network of the cell adapts dynamically to external or internal (that is, genetic) perturbations and thereby defines the cell's functional state and determines its phenotypes. Describing and understanding the complete and quantitative proteome as well as its structure, function and dynamics is a central and fundamental challenge of biology.

Two strategies that differ in principle have been used to study the proteome and the molecular mechanisms that it mediates. Conventionally, specific proteins are isolated and then analysed with respect to their structure and function through the established methods of biochemistry and biophysics. But it has also become possible to perform large-scale, systematic measurements of proteomes to generate biological insights from the computational analysis of proteomic datasets, either on their own or in combination with other 'omics' types of data. Both approaches have been transformed fundamentally by the development of powerful mass-spectrometry-based methods. Such techniques have the capability to identify conclusively and quantify accurately almost any protein that has been expressed. They can also systematically identify and localize modified amino acids in the polypeptide chain as well as determine the composition, stoichiometry and topology of the subunits of multiprotein complexes and even contribute to determining their structure.

The annotated genome identifies the entire proteome of an organism. However, the literature has focused on the small fraction of the proteome for which measurement assays are readily available[3]. This set of intensely studied proteins has remained surprisingly constant over the past few decades. Robust mass-spectrometry-based methods now enable most proteins to be measured reliably, which vastly extends the range of the classic, mechanism-focused analyses of specific components of the proteome. They also make possible the systematic analysis of the proteome to an extent that had been predicted previously[4,5].

Underlying reasons for the success of mass spectrometry in proteomics include its inherent specificity of identification, the generic nature of the proteomics workflow and its potential for extreme sensitivity that, in principle, extends to the single ion. In practice, it has been challenging to realize the full potential of the technique, and ingenious ways of implementing mass spectrometry as a universal detector of protein identity, abundance, precise chemical state and cellular context and localization are still being devised. At present, no single mass-spectrometry-based system or method can determine by itself these diverse dimensions for proteome data.

This Review highlights the achievements of mass-spectrometry-based proteomics and the challenges that remain. Efforts to catalogue systematically the proteomes of an array of species and to transform these catalogues into highly specific assays that can quantify any component are described. The analysis of post-translational modifications is discussed, especially with regard to completeness of measurement and how the research community might assign functions to the tens of thousands of modified sites that have been discovered in the past decade. The state of mass spectrometry is reviewed in the context of the study of functional modules, in which components of the proteome come together stably or temporarily in complexes to carry out a biochemical function. Last, mass-spectrometry-based techniques that are capable of quantifying thousands of proteins across collections of large numbers of samples with a high degree of reproducibility are described; these generate large datasets that can be mined by statistical machine-learning tools to determine the state of the proteome and its response to perturbations. Such datasets start to uncover systemic malfunctions at the cellular and organismal levels in diseases that have been difficult to reach through classic protein-based or nucleic-acid-based research.

The identification and quantification of the proteome

The ability to identify reliably any component of the proteome is a requirement both for mechanistic, hypothesis-driven investigations and for large-scale, omics-type studies. A comprehensive and reliable mass-spectrometry-based proteome map is also a prerequisite for the development of targeted mass spectrometry techniques, as well as for data-independent acquisition (DIA) strategies (Fig. 1 and Box 1); these rely on information from pre-existing high-quality spectral libraries. The importance of accurate quantification in proteomics is hard to overstate,

[1]Institute of Molecular Systems Biology, Department of Biology, ETH Zürich, 8093 Zürich, Switzerland. [2]Faculty of Science, University of Zürich, 8093 Zürich, Switzerland. [3]Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany. [4]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark.

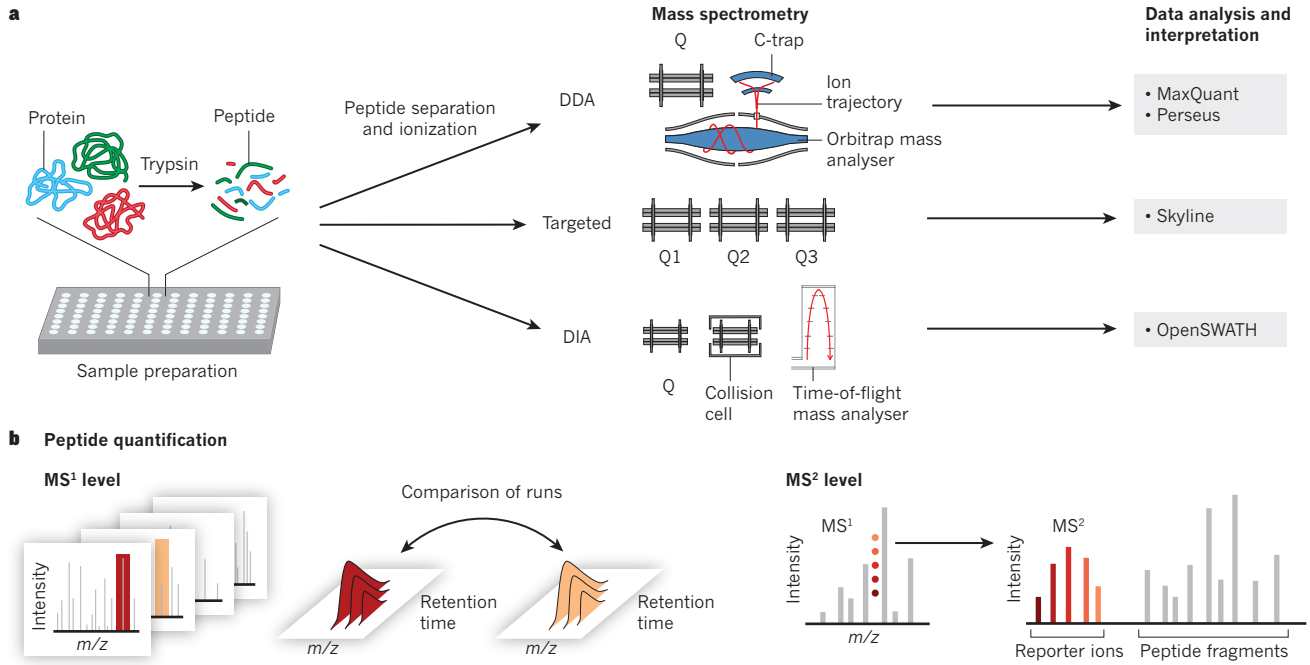


Figure 1 | Bottom-up proteomics workflows. **a**, All bottom-up proteomics workflows begin with a sample-preparation stage in which proteins are extracted and digested by a sequence-specific enzyme such as trypsin. Present methods of protein preparation are highly efficient and can be performed in 96-well plates with robotic assistance. Peptides are then separated by means of chromatography and electrosprayed, after which they are introduced into the vacuum of a mass spectrometer. Three classes of methods are shown. In DDA methods, a full spectrum of the peptides (at the MS^1 level) is acquired, followed by the collection of as many fragmentation spectra (at the MS^2 level) as possible, within a cycle time of about 1 second. A quadrupole–orbitrap mass analyser is depicted, although other types of analyser are also used in DDA. Results are interpreted using software packages such as MaxQuant[100] and the downstream Perseus environment[101]. In targeted analysis, a peptide of known mass-to-charge ratio (m/z) is selected in the first quadrupole, then the peptide is fragmented and several fragments are monitored over time. These transitions are multiplexed and their specificity is checked using software packages such as SkyLine[102]. In DIA methods, which are exemplified by sequential window acquisition of all theoretical fragment-ion spectra (SWATH)–MS[103], ranges of m/z values (that typically span 25 m/z units) are selected and peptides are fragmented, followed by the acquisition of the fragments in a time-of-flight mass spectrometer. The instrument rapidly and seamlessly cycles through the entire mass range within a few seconds. The multiplexed fragment spectra are interpreted — often with the help of known fragment spectra from large spectral libraries — by software such as OpenSWATH[104]. **b**, Peptide quantities can be determined at the MS^1 level by integrating the signal from peaks of the precursor ions that elute from the high-performance liquid chromatography column. An arbitrary number of runs (stacked mass spectra, left) can be compared using sophisticated alignment and normalization procedures. Quantitative comparison of the isotopic cluster of the same peptide over two runs can be performed. Peptide identities can also be transferred when the peptide is fragmented in only one of the runs but matches precisely the mass and elution time of an aligned peak (known as the 'match between runs' feature in MaxQuant[100]). Absolute quantities can be estimated by adding up the peak volumes of all peptides that identify a particular protein then determining the proportion of the (known) total proteome mass that has been analysed. Peptides can also be subjected to label-free quantification at the MS^2 level (right). In this case, the fragment-ion intensities that are unique to a specific peptide are used for quantification, in a way that is analogous to the use of precursor-ion signal intensities for quantification using MS^1-level data. In multiplexed shotgun proteomics, up to ten samples are labelled differentially so that they release reporter ions that can be distinguished in the MS^2 spectra. In DIA-based methods, the intensities of fragments that belong to the same precursor ion are extracted to yield a measure of peptide abundance[104,105]. Q, quadrupole.

and this has become a crucial requirement for almost all functional studies in the past 10 years.

The preferred method for proteome discovery is data-dependent acquisition (DDA) (Fig. 1) and the past decade has seen striking advances in this area. Whereas the first description of a complete model proteome[6] and the identification of more than 10,000 different proteins in human cell lines[7,8] were technological tours de force, a similar depth of coverage can now be achieved within hours and with minimal sample-preparation steps[9,10]. These developments, although still confined to a few specialized laboratories, will make proteomics increasingly applicable to everyday cell biology and biochemical research, which overwhelmingly uses classic antibody-based techniques such as western blotting. In addition to its exquisite specificity, other advantages of DDA-based proteomics include that it is unbiased and free from hypotheses; that is, the researcher does not need to know the identity of the expected proteins in advance. Furthermore, in a DDA-based proteomics experiment all proteins can be interrogated at once. As well as helping to answer a specific question, proteomics can therefore turn every experiment into a global discovery study, which enables the detection of new and unexpected molecules and connections, providing fresh biological insights. These developments are supported by publicly accessible bioinformatics tools for processing and interpreting the large amounts of data that are generated in complex projects (Fig. 1). The continued development of highly streamlined and robust proteomics workflows, including robust and economical mass spectrometers, is advocated to usher in an age of complete, accurate and ubiquitous proteomes[11], in analogy to what the introduction of next-generation sequencing has provided for genomics-related fields.

Present technology already enables analysis of the complete protein inventory of biological systems, including cell-type-specific proteomes of mammalian organs[12–14]. One outcome of in-depth proteomics studies has been a demonstration of the extent to which diverse cellular systems have similar proteomes, with few proteins being uniquely detectable in specific situations[15]. This surprising finding is supported by the Human Protein Atlas, a large-scale antibody-based study that also reports ubiquitous expression[16]. The identity of cells and tissues therefore seems to be determined primarily by the abundance at which they express their constituent proteins, and perhaps by the manner in which the proteins are organized in the proteome, rather than the presence or absence of certain proteins.

The application of DDA-based proteomics to a collection of human

BOX 1

Bottom–up proteomics

Proteins can be studied as intact entities by mass spectrometry, an approach called top-down proteomics[21]. This has the advantage that all modifications that occur on the same molecule can, in principle, be measured together, enabling identification of the precise proteoform[107]. However, bottom-up proteomics, in which peptides are generated by the enzymatic digestion of proteins, has been experimentally and computationally more tractable and is the most widespread proteomic workflow. A number of bottom-up techniques exist; each has a specific purpose, a performance profile and a range of utility. In all of the techniques, proteins are extracted from the source material then digested into peptides by a sequence-specific enzyme such as trypsin. The resulting mixture of peptides is separated by reverse-phase chromatography, which is coupled online to electrospray ionization (Fig. 1). The peptide ions are then transferred to the vacuum of a mass spectrometer, where they are fragmented in the gas phase to generate MS/MS (MS^2) spectra that contain the information to identify and quantify specific peptides. Almost always, collision-induced dissociation or higher-energy collisional dissociation[108] are used for fragmentation, but alternative methods are becoming more widely available. One such method, electron transfer dissociation[109], is particularly beneficial for the fragmentation of large and modified peptides. The resulting data are analysed by mass-spectrometry-specific computational pipelines as well as general downstream systems-biology solutions that are tailored to proteomics[101].

Three main approaches are used in bottom-up proteomics: discovery (or shotgun) proteomics by means of DDA, aimed at achieving unbiased and complete coverage of the proteome; targeted proteomics using selected reaction monitoring, aimed at the reproducible, sensitive and streamlined acquisition of a subset of known peptides of interest; and multiplexed fragmentation of all peptides that elute from the high-performance liquid chromatography column by DIA, aimed at generating comprehensive fragment-ion maps for a sample (Fig. 1a–c).

In DDA-based methods, mass spectra of all the ion species that co-elute at a specific point in the gradient elution (that is, precursor-ion spectra) are recorded at the MS^1 (or full-scan) level. The instrument alternates between the acquisition of full-scan data and the acquisition of fragment-ion spectra, in which as many precursors as possible are sequentially isolated and fragmented (at the MS^2 level). Of many possible instrument configurations, quadrupole–orbitrap analysers[110] dominate DDA proteomics but time-of-flight instruments also have unique promise. In typical 'top *N*' cycles (in which '*N*' denotes the number of MS^2 spectra that follow), an MS^1 scan is followed by about ten fragment-ion scans. Contemporary instruments transfer ions into the vacuum with greatly improved efficiency, which results in very bright beams (of more than 10^9 ions per second). The resolution of orbitraps has improved several fold, enabling very fast top *N* cycles at high resolution. However, the capacity of orbitraps is still limited to about 1 million ions, which restricts the dynamic range that can be achieved in MS^1 spectra.

In targeted proteomics, the proteins of interest are predetermined and known. Using pre-existing information, characteristic (proteotypic) peptides are selectively and recursively isolated and then fragmented over their chromatographic elution time. This is done by setting the first quadrupole of a triple quadrupole instrument to the expected precursor ion *m/z* ratio and the third quadrupole to the *m/z* ratio of an abundant fragment ion that is specific for the targeted peptide. (The second quadrupole houses the collision chamber.) To achieve selectivity, the process is multiplexed to several fragments per peptide (known as multiple reaction monitoring, MRM), and throughput is increased by multiplexing it to many peptides[111]. Alongside the robust and economical triple quadrupole instruments, high-resolution instruments such as quadrupole orbitraps are used increasingly for targeted analysis, a variant known as parallel reaction monitoring because it utilizes the entire MS^2 spectrum[112].

In DIA-based methods[113] such as SWATH[103], entire ranges of precursors are fragmented at the same time. The peptide fragmentation information is retrieved from the multiplexed MS^2 spectra either by targeted signal extraction on the basis of previously acquired single-peptide fragmentation spectra[112] or by the generation of 'pseudo' fragment-ion spectra constructed directly from the DIA data that are then subjected to classic database searching[105]. The advantage of this approach is that the entire range of possible precursor-ion masses can be analysed seamlessly and in rapid succession, which eliminates the missing value problem of DDA (in which peptides are only measured in some of a set of liquid chromatography–mass spectrometry (LC–MS^2) runs), at least within the dynamic range that is achieved in the experiment. At present, DIA is limited to a dynamic range of 4–5 orders of magnitude and it requires the *a priori* construction of fragment-ion spectra for the query peptides to deconvolve these peptides from the DIA data[104,105,114].

Each of these approaches has advantages and limitations; hybrid methods that combine the best aspects will therefore probably emerge in the near future. Entirely new methods will also be created. For instance, in the past year it has become possible to store several precursor ions in parallel in a trapped-ion mobility device, which can then be followed by serial fragmentation. Known as parallel accumulation–serial fragmentation (PASEF), this method promises to increase the speed and sensitivity of fragmentation several fold[115].

Metabolic and chemical labelling strategies have matured and can now be used for precise quantification, but they can still suffer from limitations to their accuracy and dynamic range[116–118]. Improvements in the resolution that can be achieved, combined with advances in algorithms, are making label-free quantification increasingly useful for DDA[119], selected reaction monitoring[120] and DIA[104,105] methods.

tissues, combined with the integration of data from the community, has resulted in two draft human proteomes[17,18]. Mass-spectrometric evidence for 84% (ref. 17) or 92% (ref. 18) of protein-coding sequences was reported. However, re-analysis of the data using standard and community-approved false-discovery rates for peptides and proteins leads to much lower coverage and the removal of proteins not thought to be expressed in the sampled tissues[19,20]. Extensive peptide pre-fractionation has been combined with digestion by various enzymes and peptide fragmentation methods to reach a depth of proteome coverage that should soon be on par with the comprehensiveness to which the transcriptome can be probed by next-generation sequencing[13]. Comprehensive characterization of the proteome is therefore feasible and we predict that it will soon become routine[11]. The coverage of identified proteins with sequenced peptides has also been improving, which makes it increasingly realistic to distinguish between and quantify proteoforms, the different molecular forms of a protein that originate from the same gene. A complete inventory of proteoforms cannot yet be achieved and will be a challenge to attain because of the combinatorial explosion of proteoforms that are created by even a moderate number of modifications. Top-down proteomics characterizes the actual combination of modification events for each proteoform[21]. Although attractive in principle, top-down mass spectrometry is experimentally and computationally challenging because

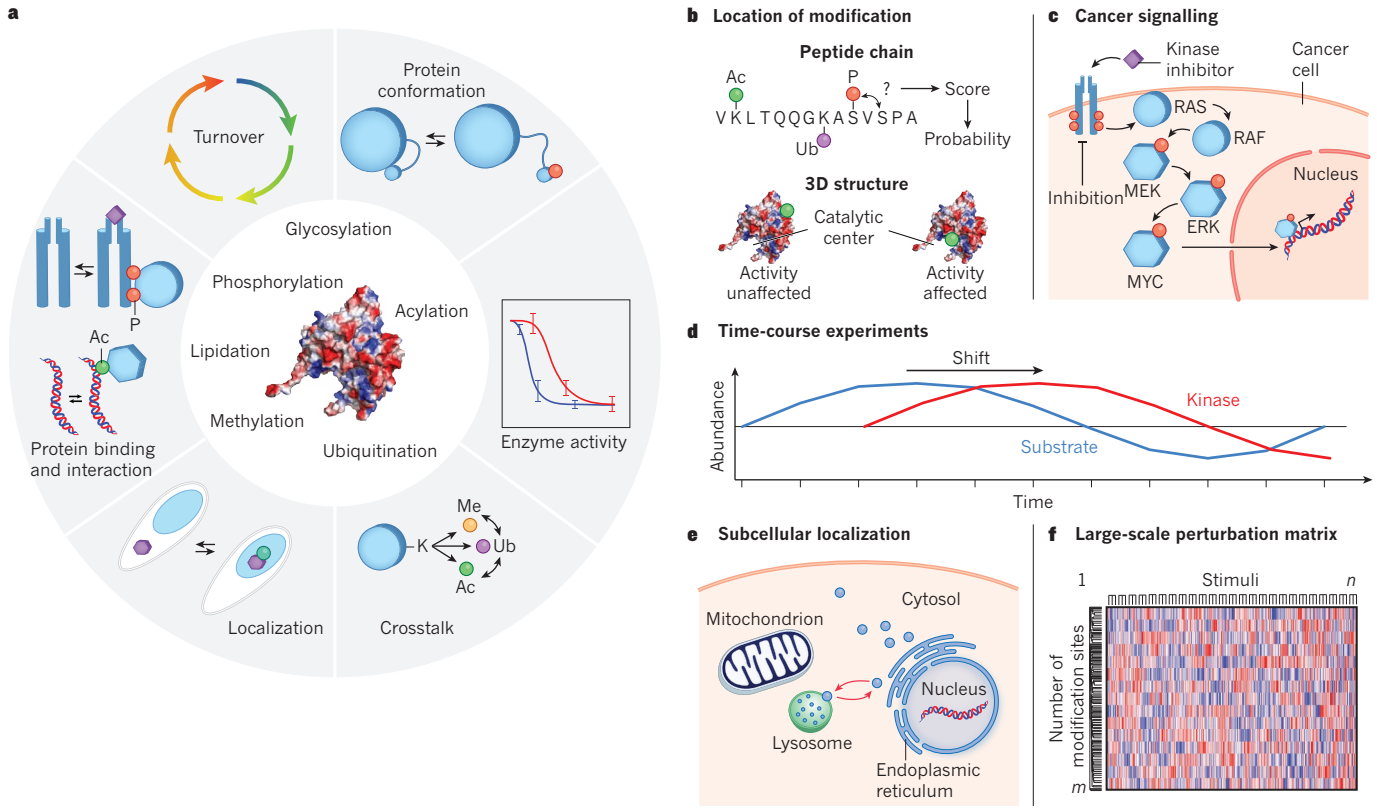


Figure 2 | Analysis of post-translational modifications. **a**, In post-translational modification, proteins are modified through the attachment of a chemical moiety such as a phosphate group, usually by a dedicated and highly specific system of enzymes. The most commonly studied post-translational modifications are listed (centre) and these are accompanied by hundreds of other less-well-studied or unknown types of modifications. Such modifications can lead to: alterations in protein conformation (through phosphorylation) and subsequent allosteric regulation; changes in enzyme activity; crosstalk that results from the same amino-acid residue being targeted by more than one type of modification; alterations in the subcellular localization of proteins; changes in protein binding; and alterations in protein lifetimes (for example, through the attachment of the small protein ubiquitin). Ac, acetyl; ERK, extracellular signal-related kinase; Me, methyl; MEK, mitogen-activated protein kinase kinase; MYC, transcription factor cMYC; P, phosphate; RAF, RAF kinase; RAS, RAS GTPase; Ub, ubiquitin. **b**, After a modified peptide has been identified from the fragment spectra, the amino acid in the peptide chain to which the post-translational modification is attached must be determined. The location of the modification within the three-dimensional structure of the protein can often also be determined, which provides clues about function. **c**, Global interrogation of the changes in a signalling pathway can be achieved readily by quantitative phosphoproteomics. For example, the suppression of aberrant signalling in cancer cells by drugs known as kinase inhibitors can be followed. **d**, Detailed time-course experiments yield information on the temporal ordering of events such as the activation of a kinase upstream of one of its substrates. The proportion of proteins that are modified by a particular post-translational modification (also termed the occupancy or stoichiometry) can change drastically depending on the biological conditions (not shown). It can be derived from the changes in protein level and the levels of the modified and unmodified peptide in two cellular states[106]. **e**, The modification of a protein often determines its subcellular localization — that is, whether it is found in the nucleus or the cytosol, for instance. Many types of stimuli can be applied to biological systems, after which the level of a particular post-translational modification can be determined. **f**, The structure of the perturbation matrix that results reveals the regulated sites and how they correlate between stimuli, as indicated by hot spots in the heat map. *m*, number of modification sites quantified; *n*, number of stimuli applied.

of the greater difficulty in analysing proteins in comparison with peptides and because each protein is distributed as multiple proteoforms that might or might not differ functionally. The array of modern mass spectrometry techniques has also been deployed to analyse unique types of sample with biological and clinical importance, including secreted proteins in the context of immunology[22], the peptidome of body fluids such as cerebrospinal fluid[23], the immunopeptidome[24] and the extracellular matrix[25].

Proteomics is sufficiently advanced to warrant the in-depth characterization of a great variety of biological systems. Along with other important information, this enables protein copy numbers or concentrations to be determined on a proteome-wide scale[26–28], which helps to improve understanding of the underlying biology.

Characterizing protein modifications and cell signalling

Mass-spectrometry-based proteomics is well suited to the study of post-translational modifications because such changes lead to characteristic shifts in mass and can be located with the resolution of a single amino acid through peptide-fragment ion spectra (Fig. 2). The only deviation from the DDA-based proteomic workflow that is used to identify unmodified peptides is the addition of an enrichment step for peptides that carry the modification of interest. Post-translational modifications that are particularly labile, such as *O*-linked β-*N*-acetylglucosamine (*O*-GlcNAc), benefit from the use of electron transfer dissociation as the fragmentation method, and certain classes of modifications, including glycosylations with large glycans and nucleotide modifications, can also be challenging to detect using mass spectrometry. The most frequently studied types of post-translational modifications are phosphorylation, ubiquitylation, the addition of ubiquitin-like proteins, glycosylation, methylation, acetylation and other types of acylation. For these, present technology enables the identification of thousands of sites of modification and their accurate quantification between proteomic states[29]. The main surprise has been the number and diversity of these post-translational modifications as well as how many of them seem to be involved in cellular regulation. For example, more than 50,000 phosphorylation events on at least 75% of the proteome have been documented in a single cell line[30]. Phosphoproteomics is used routinely to quantify the response of cells to

stimuli and such studies have reached a remarkable level of detail and sophistication. As well as providing large catalogues of sites, they have led to the discovery of sites of regulation with pivotal roles in determining the state of biological processes[31–35]. A streamlined protocol has made it possible to analyse *in vivo* signalling events with high temporal resolution[36]. This revealed that insulin signalling in the liver is unexpectedly fast: maximal phosphorylation was reached within a few seconds at many sites and transcription factors were phosphorylated fully within 30 seconds. Another message emerging from phosphoproteomics is that the proportion of sites that are functional seems to be high. This is suggested by high stoichiometry (that is, the fraction of proteins that are phosphorylated at a specific site), a large number of highly regulated sites in diverse processes, and by the tight temporal correlation of many uncharacterized sites with sites that are known to be functional. Conversely, lysine acetylation behaves very differently: the stoichiometry is extremely low for most sites and often these modifications seem to be of a non-enzymatic origin, which is also true for acylations such as succinylation[37,38]. Lysine is the most frequently modified amino-acid residue and the specific target of ubiquitylation, a modification that can be enriched efficiently and studied in a linkage-specific manner by mass spectrometry. Effective strategies also exist for characterizing SUMOylation and modification with other ubiquitin-like proteins, and these have revealed unique insights into their large-scale behaviour[39]. Histone modifications and their regulators (proteins known as 'writers', 'readers' and 'erasers' that make, recognize and edit epigenetic marks) are of great interest and specific methods have been devised for their detection[40,41].

Mass spectrometry also enables the characterization of hundreds of exotic or unknown modifications[42–44]. This emerging area builds on new instrumentation, innovative methods of fragmentation and fresh protocols for enrichment but faces the challenge of devising enrichment methods that are specific for each post-translational modification of interest. As the proteome is probed to ever increasing depths, the analysis of modifications without their enrichment is becoming more feasible, and this is already possible for methylation and phosphorylation.

Post-translational modifications and proteolytic processing events, in particular, can also be analysed using chemical proteomics approaches. These use compounds that bind to engineered small-molecule binding pockets[45] or probes that label the freshly created N termini of proteins after cleavage[46,47]. The deep, quantitative and time-resolved analysis of specific types of modifications in many systems and species has already provided a wealth of biological insights. These data also indicate that specific modification systems intersect and cooperate to generate a specific cellular state. The comprehensive analysis of proteoforms that differ in their state of modification, the determination of the functional significance of such proteoforms and the elucidation of the processes that catalyse and control their homeostasis remain challenges for the future.

Protein modules, networks and cellular functions

Proteins rarely function alone; instead, they depend on the association of various components into macromolecular complexes. The concept of modular biology, proposed by Leland Hartwell and his colleagues, states that the biological functions of the cell are carried out by multicomponent modules[48], and the modularity of the proteome has been impressively demonstrated by several classic studies[49]. An array of mass-spectrometry-based strategies, the best established of which is interaction proteomics, has made considerable contributions to integrative or hybrid approaches to yield the composition, topology and structure of specific complex macromolecular assemblies[50].

Interaction proteomics involves a pull-down assay of a bait protein with its binding partners followed by mass-spectrometric analysis, known as affinity-purification mass spectrometry (AP–MS)[51] (Fig. 3a). Thousands of proteins can be detected in such experiments owing to the high sensitivity of mass spectrometry and the propensity of the samples to contain unspecific contaminants. Proteins that bind with specificity to the bait can be distinguished effectively from the contaminants through the quantitative comparison of samples with control assays, preferably using rigorous statistical controls[52,53]. Without the ability to distinguish background binding, the reported interactomes of specific proteins often contain hundreds of purported binders with little biological importance. Versions of this basic AP–MS workflow have been implemented robustly to support large-scale mapping of the wiring diagrams of the human cellular proteome[54]. Taking advantage of the relative abundance levels of prey proteins and the endogenously expressed bait, and adding copy numbers of the entire cellular proteome, provides a human interactome in three quantitative dimensions and enables the estimation of binding stoichiometries. This helps to classify interactions into stable, regulatory

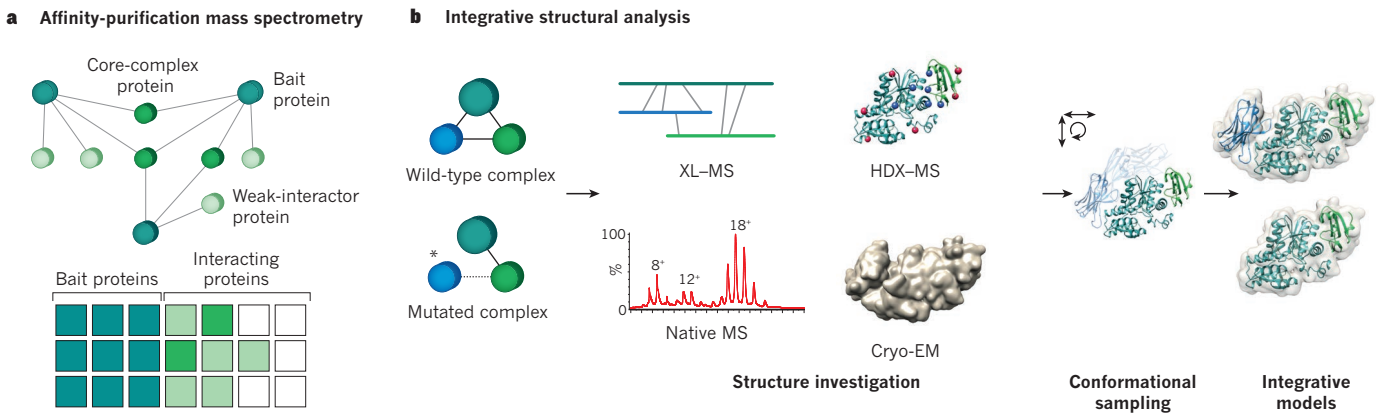


Figure 3 | Interaction proteomics and structural proteomics. a, Schematic representations of a protein interaction network with bait proteins (teal), core complex members (dark green) and weak interactors (light green). A bait protein is precipitated with its interaction partners and is measured in replicates by one of the workflows described in Fig. 1. By considering the interaction stoichiometry (the molar ratio of prey proteins and the bait protein expressed under endogenous control) and the relative cellular abundances of the proteins, stable core complexes can be distinguished from weak interactions and unspecific interactions, as well as from asymmetric interactions between proteins of different abundances[55]. **b**, A wild-type protein complex and the same complex with mutations (*) are investigated using complementary structural techniques, collectively termed integrative or hybrid structural analysis. For example, XL–MS can reveal information about subunit topology and direct domain–domain interactions. Hydrogen–deuterium exchange mass spectrometry (HDX–MS) is able to determine the interaction surfaces and solvent-exposed regions. Native mass spectrometry (native MS), in which entire protein complexes are electrosprayed into the mass spectrometer, can infer the stoichiometry and the assembly pathway of such complexes, and cryo-EM can obtain their overall shape and their density maps. The heterogeneous structural restraints are integrated in a common computational framework that evaluates subunit configurations (known as conformational sampling). Consensus models that represent the structures of the wild-type and mutated complexes can then be derived.

or transient ones and even captures client interactions such as proteins being folded by chaperone complexes[55]. This work established that networks of cells are surprisingly dominated by a large number of weak interactions and that the number of stable core complexes is limited. The emerging picture of a modular proteome in which modules have variable stoichiometric robustness is also supported by a study in which the relative changes of bona fide protein components of 182 complexes were determined in 11 cell types and 5 temporal states[56]. The covariance of the co-expression profiles for complex subunits varied considerably, which suggests that dynamic subunit associations fine-tune the composition and function of specific cellular modules[56].

Modified peptides, oligonucleotides and small molecules have also been used with success as bait proteins for AP–MS experiments[51]. For instance, transcription-factor complexes that are crosslinked to DNA can be analysed readily, as can protein complexes that are recruited to specific DNA lesions[57]. Other approaches to capture protein interactions include enzyme-meditated proximity labelling in cells followed by pull-down assays of the labelled proteins[58,59] and the accurate measurement of co-fractionation patterns[60–62]. Such measurements are also the basis of organellar proteomics, which aims to determine the subcellular location and dynamics of the proteome[63–66], a valuable complement to imaging-based technologies.

Although AP–MS and related methods indicate the composite population of proteins that is associated with a particular bait protein, other mass-spectrometry-based methods can also identify the subunit interfaces, topology, conformation and structure of protein complexes (Fig. 3b), as shown by the analysis of the nuclear pore complex[67].

Native mass spectrometry, which is the direct analysis of macromolecular assemblies by mass spectrometry, has been used both by itself[68] and as part of an integrative approach[69] to gain insights into the subunit stoichiometry, topology and structure of macromolecular assemblies. When applied to membrane protein complexes, the technique revealed an unappreciated structural role for lipids in respiratory protein complexes[70].

Integrative or hybrid approaches complement X-ray crystallography and nuclear magnetic resonance, methods that are central to structural biology, and mass spectrometry has become an essential component of the hybrid structural-biology toolbox[71]. Distance restraints that are generated by chemical crosslinking and the mass-spectrometry-based identification of crosslinked residues (an approach termed XL–MS) have proven helpful for determining the structure of large complexes[72], particularly in combination with single-particle cryo-electron microscopy (cryo-EM) data. XL–MS and cryo-EM have been used to solve longstanding problems in structural biology[71], to identify the substrate binding sites in molecular chaperones[73] and to detect steric alterations in complexes in different functional states[74]. XL–MS has also been used to analyse protein–RNA interfaces[75], to identify receptor–ligand pairs directly[76], to map physical interactions between different types of biomolecules and to identify the ligands of orphan receptors.

Integrative structural-biology methods are being adapted for use with the microgram amounts of protein complexes that are isolated by affinity purification, and this advance has been applied to mapping the organization of the protein phosphatase 2A (PP2A) enzyme system in HEK293 cells[77]. Using the two catalytic subunits, the scaffold subunit and most of the 15 regulatory subunits from which trimeric PP2A structures are combinatorially assembled as bait proteins, XL–MS identified the protein–protein interfaces, the actual subunit composition of the PP2A complexes that are concurrently expressed in the cell and their associated proteins to establish a high-granularity protein interaction network consisting of more than 150 proteins[77].

Notably, XL–MS is beginning to be used on a proteomics scale[78,79]. Although the crosslinks that are identified in such studies come primarily from highly expressed complexes, they highlight a path towards the direct measurement of protein–protein interfaces in the cell. The combination of AP–MS and XL–MS was recently refined so that chemical crosslinks could be identified from samples containing only a few million cells[80,81]. Complexes that are isolated by AP–MS can also be used to generate cryo-EM single-particle data, which opens up the possibility of linking the atomic structure and function of macromolecular assemblies that have been isolated from cells in a particular functional state. Results from cryo-electron tomography studies further extend this perspective towards the possibility of observing specific macromolecular modules by template matching *in situ*[82,83].

In a similar way to their composition, the conformation of the subunits of protein complexes can adapt to the state of the cell. Mass spectrometry techniques can detect changes in protein conformation and protein interfaces and then relate these observations to functional alterations in particular proteins. Hydrogen–deuterium exchange mass spectrometry is a classic method for determining alterations in the conformation, structure and interfaces of specific complexes[84]. By contrast, the hydroxyl radical footprinting method predominantly labels solvent-exposed side chains and is not affected by back exchange of the labelled residues[85]. The different conformations of a protein can vary in thermal stability, an observation that has been used to probe conformational changes at a proteomic scale[86]. Cells treated with a cancer drug were subjected to different temperatures, after which heat-denatured proteins were removed and the remaining soluble proteins were analysed by mass spectrometry. This pinpointed both expected and unexpected binding partners of the drug. A conceptually similar technique used the fact that conformational changes in proteins can be detected using protein digestion patterns generated under conditions of limited proteolysis[87]. Structural features of more than 1,000 yeast proteins were concurrently monitored by targeted mass spectrometry and altered conformations for about 300 proteins on a change in nutrients were detected[87]. Such examples demonstrate how structural proteomics techniques are helping to tackle the challenge of detecting often weak interactions between proteins, small-molecule ligands and co-factors on a global scale, as well as the structural effects of ligand binding.

Proteotype states and cellular phenotypes

In the 1940s, Linus Pauling established that a structural alteration in haemoglobin was related causally to a disease phenotype[88]. In that particular case, the structural variation was caused by a single amino acid change in one of the haemoglobin chains, the result of a mutation in the gene that encodes the chain. The extension of this fundamental principle of biology to the level of proteome networks suggests that genetic or external perturbations change the state of the proteome network and that such changes cause or correlate with altered phenotypes (Fig. 4). The state of a proteome that is associated with a specific phenotype can be described as a proteotype. The association between a proteotype and its corresponding phenotype can be investigated by means of two mass-spectrometry-based approaches that differ in principle. The first approach attempts to describe a phenotype mechanistically using the aggregated structure and function of the proteins or modules that constitute the underlying processes. The second approach associates a phenotype with its proteotype through advanced statistical machine-learning tools (known collectively as 'big data' analytics) but does not necessarily reach a causal or mechanistic understanding of the underlying processes. Both approaches have been greatly advanced by mass-spectrometry-based technology. In particular, the big data approach based on statistical associations has become possible only through the development of mass spectrometry techniques that are capable of quantifying sets of proteins with a high degree of reproducibility across large collections of samples, generating large data matrices of proteins measured across various samples with minimal missing values. Mass-spectrometry techniques that are used to generate such matrices include the matching of MS^1 intensity maps, using their retention time versus mass-to-charge ratio, from collections of samples and DIA-based methods, and the targeted mass spectrometry of smaller numbers of proteins (Box 1).

In a demonstration of these concepts, a yeast genetic reference panel was used to quantify the effect of genetic perturbations on a metabolic network[89]. Selected-reaction-monitoring targeted mass spectrometry measured 50 metabolic proteins in 96 genetically well-defined strains of yeast. Parental strains acquired independent genetic variations that

consistently affected levels of proteins from the same module or pathway that selective pressures favoured for the acquisition of sets of polymorphisms that maintain the stoichiometry of complexes and pathways. Similarly, 192 proteins that constituted a metabolic network were quantified by selected reaction monitoring of liver samples in two metabolic states from 40 strains of mice from a genetic reference strain compendium[90], enabling genetic and environmental perturbations to be probed effectively[91]. This established a direct mechanistic link between alleles of the gene *Dhtkd1* (a protein quantitative trait locus (pQTL)), the quantity of 2-aminoadipate (a metabolite that is controlled by Dhtkd1) and a disease risk for type 2 diabetes. Mechanistic and data-driven approaches can therefore converge to enhance understanding of complex phenotypes if multilevel omics data are integrated at the level of modular networks. Repeating the proteomics measurements of the liver samples using DIA-based mass spectrometry techniques quantified more than 2,600 proteins across the collection of samples, which led to the detection of hundreds of pQTLs as well as mechanistic insights into inborn errors of metabolism and the determination of a molecular basis for respiratory super-complex formation[92].

These examples and analogous ones from the proteogenomics of cancer[93] establish a link through association studies between genetic loci and the network state, as well as between the network state and disease phenotypes. The mass spectrometry methods of bottom-up proteomics (Box 1) represent a general experimental framework for systematically probing the proteotype at ever increasing levels of completeness and precision to support the association of proteotypes and phenotypes.

In the context of translational medicine, proteins that consistently alter their abundance in correlation with a disease phenotype are considered to be biomarker candidates for the phenotype of interest. Typically, a small number of study participants are investigated in depth to extract potential biomarkers that can be validated in larger cohorts[94,95]. Although attractive in principle, biomarker discovery using mass-spectrometry-based methods is extremely challenging in practice. However, data-driven approaches are opening fresh avenues to associating protein-expression patterns with disease states.

In particular, the detection of protein biomarkers in blood plasma as a window to the physiological state of a person has been an important goal of protein science since before the advent of mass spectrometry. Experience gained over the past decade in plasma proteome analysis by mass spectrometry has demonstrated the enormous challenges of this approach, which are rooted in the complexity of the plasma proteome, its inherent variability across a population and the prevalence of factors that affect its composition, including age, gender and lifestyle. However, several studies[94,96,97] have shown that the highly reproducible mass spectrometry techniques used for proteotype measurements in tissues can be applied to plasma proteins. Fast and reliable measurements of plasma samples will therefore be possible in collections that consist of hundreds of samples. The systematic measurement of plasma proteins in twin populations has already been used to associate observed changes in abundance in the plasma proteome with genotype[98]. Furthermore, the plasma proteome can now be probed in a broad and high-throughput manner with the aim of extracting as much information about the health or disease state of an individual as possible, effectively enabling high-throughput phenotyping of people[96]. Continuing advances in mass spectrometry technology might therefore enable the future discovery of clinically actionable protein biomarker patterns.

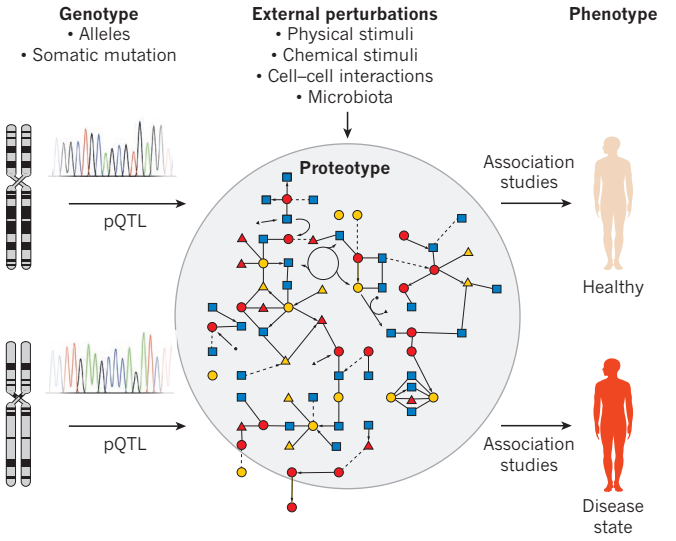


Figure 4 | Proteotype states and phenotypes. The proteotype, which is the acute state of the proteome, is shown as a modular network of interacting protein entities (coloured shapes). The composition of the proteotype and the organization of individual proteins into functional modules and interaction networks are determined by the combined effects of genotype and external perturbations, which include physical or chemical stimuli, cell–cell interactions or the microbiota. Genotypic differences such as allele differences or somatic mutations might perturb the proteotype. The relationship between genetic loci and the abundance of a protein can be described by a pQTL. These are identified by associating the abundance of a specific protein with particular alleles in genetically characterized sample populations such as genetic reference panels. In turn, the proteotype determines phenotypes, including clinical phenotypes. Association studies can identify relationships between proteotypes and phenotypes. Establishing such associations requires the generation of quantitatively accurate and highly reproducible datasets in which the same proteins are quantified across a large number of samples (for example, genetic reference panels or cohorts of patients). Datasets that support such association studies can now be generated using various mass spectrometry techniques.

Outlook

Over the past decade, mass-spectrometry-based proteomics has matured from a largely technology-driven field of research into a mainstream analytical tool for the life sciences. It is a versatile approach that supports the analysis of many aspects of proteins, including sequence, quantity, state of modification, structure and macromolecular context. It also accommodates a variety of research approaches, such as mechanism-oriented exploration for determining causal relationships and big-data strategies that rely on statistical associations to discover biological relationships.

Further, dramatic improvements in the core technology of mass spectrometry are probable and will open up the field of proteomics to even more applications. Aside from a focus on signalling and structural applications, important goals for proteomics will be to build comprehensive and quantitative catalogues of proteins under many conditions and perturbations and to organize these proteoforms into a modular proteome of the cell. This will improve understanding of processes across many areas of biology and diseases and will constitute an excellent starting point for modelling the cell. For this to occur, proteomics must be tightly integrated with other technologies and it should address challenges such as single-cell analysis, an approach that was pioneered by mass cytometry[99]. The integration of different types of data is already far advanced in the case of next-generation sequencing technologies (for example, RNA sequencing, chromatin immunoprecipitation followed by sequencing (ChIP-seq) and ribosome profiling) and metabolomics, and the integration of data from structural biology and imaging-based technologies is advancing at a rapid pace. There are also considerable opportunities for bringing proteomics together with increasingly efficient tools for editing the genome — in particular, CRISPR–Cas9. We envision this to work in an iterative manner in which proteomics findings are interrogated by deleting, tagging and point-mutating one or more genes of importance, followed by further rounds of proteomics measurements to determine the effects of the genetic alterations on the proteome. This will address the fundamental question of how genotypic variability is mechanistically translated into phenotypic variability. The integration of various omics approaches and many perturbations will generate exponential flows of disparate data types. This will necessitate commensurate advances in bioinformatics and computational proteomics, which will be powered increasingly by

machine-learning technologies while retaining their ability to generate biological insights. In this regard, the journey from single-protein analysis to a true understanding of the proteome and the importance of proteotypes will be long, challenging and exciting. ■

Received 11 January; accepted 15 July 2016.

1. Marguerat, S. *et al.* Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* **151,** 671–683 (2012).
2. Milo, R. What is the total number of protein molecules per cell volume? A call to rethink some published values. *BioEssays* **35,** 1050–1055 (2013).
3. Edwards, A. M. *et al.* Too many roads not taken. *Nature* **470,** 163–165 (2011).
4. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).
5. Cravatt, B. F., Simon, G. M. & Yates, J. R. The biological impact of mass-spectrometry-based proteomics. *Nature* **450,** 991–1000 (2007).
6. de Godoy, L. M. F. *et al.* Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455,** 1251–1254 (2008).
This paper demonstrates that complete proteomes of a model organism can be obtained and quantified in different biological states.
7. Beck, M. *et al.* The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7,** 549 (2011).
8. Nagaraj, N. *et al.* Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.* **7,** 548 (2011).
9. Hebert, A. S. *et al.* The one hour yeast proteome. *Mol. Cell. Proteomics* **13,** 339–347 (2014).
10. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nature Methods* **11,** 319–324 (2014).
11. Mann, M., Kulak, N. A., Nagaraj, N. & Cox, J. The coming age of complete, accurate, and ubiquitous proteomes. *Mol. Cell* **49,** 583–590 (2013).
12. Azimifar, S. B., Nagaraj, N., Cox, J. & Mann, M. Cell-type-resolved quantitative proteomics of murine liver. *Cell Metab.* **20,** 1076–1087 (2014).
13. Richards, A. L., Merrill, A. E. & Coon, J. J. Proteome sequencing goes deep. *Curr. Opin. Chem. Biol.* **24**, 11–17 (2015).
14. Sharma, K. *et al.* Cell type- and brain region-resolved mouse brain proteome. *Nature Neurosci.* **18,** 1819–1831 (2015).
15. Lundberg, E. *et al.* Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* **6,** 450 (2010).
16. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).
This paper provides an integrative analysis of the human proteome through large-scale antibody localization and transcriptomics; the findings are organized in an accompanying database.
17. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–581 (2014).
18. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).
This study aggregates data on diverse human proteomes from the authors and the research community and, like ref. 17, argues that a large part of the genome is accessible to mass-spectrometric detection.
19. Ezkurdia, I., Vázquez, J., Valencia, A. & Tress, M. Analyzing the first drafts of the human proteome. *J. Proteome Res.* **13,** 3854–3855 (2014).
20. Omenn, G. S. *et al.* Metrics for the Human Proteome Project 2015: progress on the human proteome and guidelines for high-confidence protein identification. *J. Proteome Res.* **14,** 3452–3460 (2015).
21. Tran, J. C. *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480**, 254–258 (2011).
22. Meissner, F., Scheltema, R. A., Mollenkopf, H.-J. & Mann, M. Direct proteomic quantification of the secretome of activated immune cells. *Science* **340,** 475–478 (2013).
23. Secher, A. *et al.* Analytic framework for peptidomics applied to large-scale neuropeptide identification. *Nature Commun.* **7,** 11436 (2016).
24. Caron, E. *et al.* Analysis of major histocompatibility complex (MHC) immunopeptidomes using mass spectrometry. *Mol. Cell. Proteomics* **14,** 3105–3117 (2015).
25. Schiller, H. B. *et al.* Time- and compartment-resolved proteome profiling of the extracellular niche in lung injury and repair. *Mol. Syst. Biol.* **11,** 819 (2015).
26. Malmström, J. *et al.* Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*. *Nature* **460,** 762–765 (2009).
27. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).
A pioneering investigation of the degree of correlation between the transcriptome and the proteome — a question that is still unresolved.
28. Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol. Cell. Proteomics* **13,** 3497–3506 (2014).
29. Doll, S. & Burlingame, A. L. Mass spectrometry-based detection and assignment of protein posttranslational modifications. *ACS Chem. Biol.* **10,** 63–71 (2015).
30. Sharma, K. *et al.* Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep.* **8,** 1583–1594 (2014).
31. Hsu, P. P. *et al.* The mTOR-regulated phosphoproteome reveals a mechanism of mTORC1-mediated inhibition of growth factor signaling. *Science* **332,** 1317–1322 (2011).
32. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143,** 1174–1189 (2010).
33. Olsen, J. V. *et al.* Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127,** 635–648 (2006).
34. Francavilla, C. *et al.* Functional proteomics defines the molecular switch underlying FGF receptor trafficking and cellular outputs. *Mol. Cell* **51,** 707–722 (2013).
35. Steger, M. *et al.* Phosphoproteomics reveals that Parkinson's disease kinase LRRK2 regulates a subset of Rab GTPases. *eLife* **5,** e12813 (2016).
This study used a combination of genetics, chemical proteomics and cutting-edge phosphoproteomics to reveal genuine, *in vivo* substrates of the Parkinson's disease kinase LRRK2, opening the way to clinical trials.
36. Humphrey, S. J., Azimifar, S. B. & Mann, M. High-throughput phosphoproteomics reveals *in vivo* insulin signaling dynamics. *Nature Biotechnol.* **33,** 990–995 (2015).
37. Weinert, B. T. *et al.* Acetyl-phosphate is a critical determinant of lysine acetylation in *E. coli*. *Mol. Cell* **51,** 265–272 (2013).
38. Choudhary, C., Weinert, B. T., Nishida, Y., Verdin, E. & Mann, M. The growing landscape of lysine acetylation links metabolism and cell signalling. *Nature Rev. Mol. Cell Biol.* **15,** 536–550 (2014).
39. Hendriks, I. A. *et al.* Uncovering global SUMOylation signaling networks in a site-specific manner. *Nature Struct. Mol. Biol.* **21,** 927–936 (2014).
40. Huang, H., Lin, S., Garcia, B. A. & Zhao, Y. Quantitative proteomic analysis of histone modifications. *Chem. Rev.* **115,** 2376–2418 (2015).
41. Zheng, Y., Huang, X. & Kelleher, N. L. Epiproteomics: quantitative analysis of histone marks and codes by mass spectrometry. *Curr. Opin. Chem. Biol.* **33,** 142–150 (2016).
42. Savitski, M. M., Nielsen, M. L. & Zubarev, R. A. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol. Cell. Proteomics* **5,** 935–948 (2006).
43. Jungmichel, S. *et al.* Proteome-wide identification of poly(ADP-ribosyl)ation targets in different genotoxic stress responses. *Mol. Cell* **52,** 272–285 (2013).
44. Chick, J. M. *et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature Biotechnol.* **33,** 743–749 (2015).
45. Rix, U. & Superti-Furga, G. Target profiling of small molecules by chemical proteomics. *Nature Chem. Biol.* **5,** 616–624 (2009).
46. Gawron, D., Ndah, E., Gevaert, K. & Van Damme, P. Positional proteomics reveals differences in N-terminal proteoform stability. *Mol. Syst. Biol.* **12,** 858 (2016).
47. Kleifeld, O. *et al.* Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates. *Nature Protocols* **6,** 1578–1611 (2011).
48. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402** (suppl.), C47–C52 (1999).
49. Pawson, T. Protein modules and signalling networks. *Nature* **373,** 573–580 (1995).
50. Ward, A. B., Sali, A. & Wilson, I. A. Integrative structural biology. *Science* **339,** 913–915 (2013).
51. Dunham, W. H., Mullin, M. & Gingras, A.-C. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics* **12,** 1576–1590 (2012).
52. Choi, H. *et al.* SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nature Methods* **8,** 70–73 (2011).
53. Keilhauer, E. C., Hein, M. Y. & Mann, M. Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Mol. Cell. Proteomics* **14,** 120–135 (2015).
54. Huttlin, E. L. *et al.* The BioPlex network: a systematic exploration of the human interactome. *Cell* **162,** 425–440 (2015).
A large-scale investigation of proteins binding to tagged constructs to establish a human interactome.
55. Hein, M. Y. *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163,** 712–723 (2015).
This paper describes the characterization of a human interactome using bait proteins that are expressed under endogenous control; its analysis in several quantitative dimensions revealed a preponderance of weak interactions.
56. Ori, A. *et al.* Spatiotemporal variation of mammalian protein complex stoichiometries. *Genome Biol.* **17,** 47 (2016).
57. Räschle, M. *et al.* Proteomics reveals dynamic assembly of repair complexes during bypass of DNA cross-links. *Science* **348,** 1253671 (2015).
58. Roux, K. J., Kim, D. I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J. Cell Biol.* **196,** 801–810 (2012).
59. Rhee, H.-W. *et al.* Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* **339,** 1328–1331 (2013).
60. Havugimana, P. C. *et al.* A census of human soluble protein complexes. *Cell* **150,** 1068–1081 (2012).
61. Kristensen, A. R., Gsponer, J. & Foster, L. J. A high-throughput approach for measuring temporal changes in the interactome. *Nature Methods* **9,** 907–909 (2012).
62. Wan, C. *et al.* Panorama of ancient metazoan macromolecular complexes. *Nature* **525,** 339–344 (2015).
63. Christoforou, A. *et al.* A draft map of the mouse pluripotent stem cell spatial proteome. *Nature Commun.* **7,** 8992 (2016).
64. Larance, M. & Lamond, A. I. Multidimensional proteomics for cell biology. *Nature Rev. Mol. Cell Biol.* **16,** 269–280 (2015).

65. Yates, J. R., Gilchrist, A., Howell, K. E. & Bergeron, J. J. M. Proteomics of organelles and large cellular structures. *Nature Rev. Mol. Cell Biol.* **6,** 702–714 (2005).
66. Itzhak, D. N., Tyanova, S., Cox, J. & Borner, G. H. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* **5,** e16950 (2016).
67. Alber, F. *et al.* The molecular architecture of the nuclear pore complex. *Nature* **450,** 695–701 (2007).
68. Marcoux, J. & Robinson, C. V. Twenty years of gas phase structural biology. *Structure* **21,** 1541–1550 (2013).
69. Politis, A. *et al.* A mass spectrometry-based hybrid method for structural modeling of protein complexes. *Nature Methods* **11,** 403–406 (2014).
70. Zhou, M. *et al.* Mass spectrometry of intact V-type ATPases reveals bound lipids and the effects of nucleotide binding. *Science* **334,** 380–385 (2011).
An elegant demonstration of native mass spectrometry in structural studies of intact membrane complexes.
71. Leitner, A., Faini, M., Stengel, F. & Aebersold, R. Crosslinking and mass spectrometry: an integrated technology to understand the structure and function of molecular machines. *Trends Biochem. Sci.* **41,** 20–32 (2016).
72. Liu, F. & Heck, A. J. Interrogating the architecture of protein assemblies and protein interaction networks by cross-linking mass spectrometry. *Curr. Opin. Struct. Biol.* **35,** 100–108 (2015).
73. Joachimiak, L. A., Walzthoeni, T., Liu, C. W., Aebersold, R. & Frydman, J. The structural basis of substrate recognition by the eukaryotic chaperonin TRiC/CCT. *Cell* **159,** 1042–1055 (2014).
74. Walzthoeni, T. *et al.* xTract: software for characterizing conformational changes of protein complexes by quantitative cross-linking mass spectrometry. *Nature Methods* **12,** 1185–1190 (2015).
75. Kramer, K. *et al.* Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nature Methods* **11,** 1064–1070 (2014).
76. Frei, A. P. *et al.* Direct identification of ligand-receptor interactions on living cells and tissues. *Nature Biotechnol.* **30,** 997–1001 (2012).
77. Herzog, F. *et al.* Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science* **337,** 1348–1352 (2012).
This study pioneered the use of chemical crosslinking to reveal the topology of an important phosphatase complex.
78. Liu, F., Rijkers, D. T. S., Post, H. & Heck, A. J. R. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nature Methods* **12,** 1179–1184 (2015).
79. Navare, A. T. *et al.* Probing the protein interaction network of *Pseudomonas aeruginosa* cells by chemical cross-linking mass spectrometry. *Structure* **23,** 762–773 (2015).
80. Makowski, M. M., Willems, E., Jansen, P. W. T. C. & Vermeulen, M. Cross-linking immunoprecipitation-MS (xIP-MS): topological analysis of chromatin-associated protein complexes using single affinity purification. *Mol. Cell. Proteomics* **15,** 854–865 (2016).
81. Shi, Y. *et al.* A strategy for dissecting the architectures of native macromolecular assemblies. *Nature Methods* **12,** 1135–1138 (2015).
82. Aufderheide, A. *et al.* Structural characterization of the interaction of Ubp6 with the 26S proteasome. *Proc. Natl Acad. Sci. USA* **112,** 8626–8631 (2015).
83. Mahamid, J. *et al.* Visualizing the molecular sociology at the HeLa cell nuclear periphery. *Science* **351,** 969–972 (2016).
84. Engen, J. R. Analysis of protein conformation and dynamics by hydrogen/deuterium exchange MS. *Anal. Chem.* **81,** 7870–7875 (2009).
85. Wang, L. & Chance, M. R. Structural mass spectrometry of proteins using hydroxyl radical based protein footprinting. *Anal. Chem.* **83,** 7234–7241 (2011).
86. Savitski, M. M. *et al.* Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **346,** 1255784 (2014).
In this paper, isobaric chemical labelling was used to measure the proportion of proteins that bound to a drug as a function of temperature, on a proteome-wide scale.
87. Feng, Y. *et al.* Global analysis of protein structural changes in complex proteomes. *Nature Biotechnol.* **32,** 1036–1044 (2014).
88. Pauling, L., Itano, H. A., Singer, S. J. & Wells, I. C. Sickle cell anemia, a molecular disease. *Science* **110,** 543–548 (1949).
89. Picotti, P. *et al.* A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **494,** 266–270 (2013).
90. Andreux, P. A. *et al.* Systems genetics of metabolism: the use of the BXD murine reference panel for multiscalar integration of traits. *Cell* **150,** 1287–1299 (2012).
91. Wu, Y. *et al.* Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population. *Cell* **158,** 1415–1430 (2014).
92. Williams, E. G. *et al.* Systems proteomics of liver mitochondria function. *Science* **352,** aad0189 (2016).
A demonstration of the combined use of proteomics and genetics to interrogate mitochondrial function.
93. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534,** 55–62 (2016).
This analysis of breast cancer tissues revealed that proteomics is almost on a par with transcriptomics in terms of achievable depth of coverage of gene expression.
94. Carr, S. A. *et al.* Targeted peptide measurements in biology and medicine: best practices for mass spectrometry-based assay development using a fit-for-purpose approach. *Mol. Cell. Proteomics* **13,** 907–917 (2014).
95. Rifai, N., Gillette, M. A. & Carr, S. A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nature Biotechnol.* **24,** 971–983 (2006).
96. Geyer, P. E. *et al.* Plasma proteome profiling to assess human health and disease. *Cell Syst.* **2,** 185–195 (2016).
97. Surinova, S. *et al.* Prediction of colorectal cancer diagnosis based on circulating plasma proteins. *EMBO Mol. Med.* **7,** 1166–1178 (2015).
98. Liu, Y. *et al.* Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.* **11,** 786 (2015).
99. Bandura, D. R. *et al.* Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81,** 6813–6822 (2009).
100. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnol.* **26,** 1367–1372 (2008).
101. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **13,** 731–740 (2016).
102. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26,** 966–968 (2010).
103. Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **11,** O111.016717 (2012).
104. Röst, H. L. *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnol.* **32,** 219–223 (2014).
105. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods* **12,** 258–264 (2015).
106. Olsen, J. V. *et al.* Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.* **3,** ra3 (2010).
107. Smith, L. M., Kelleher, N. L. & The Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nature Methods* **10,** 186–187 (2013).
108. Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods* **4,** 709–712 (2007).
109. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl Acad. Sci. USA* **101,** 9528–9533 (2004).
110. Zubarev, R. A. & Makarov, A. Orbitrap mass spectrometry. *Anal. Chem.* **85,** 5288–5296 (2013).
111. Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nature Methods* **9,** 555–566 (2012).
112. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol. Cell. Proteomics* **11,** 1475–1488 (2012).
113. Chapman, J. D., Goodlett, D. R. & Masselon, C. D. Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom. Rev.* **33,** 452–470 (2014).
114. Rosenberger, G. *et al.* A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data* **1,** 140031 (2014).
115. Meier, F. *et al.* Parallel accumulation–serial fragmentation (PASEF): multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *J. Proteome Res.* **14,** 5378–5387 (2015).
116. Ow, S. Y. *et al.* iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly". *J. Proteome Res.* **8,** 5347–5355 (2009).
117. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature Methods* **8,** 937–940 (2011).
118. Wühr, M. *et al.* The nuclear proteome of a vertebrate. *Curr. Biol.* **25,** 2663–2671 (2015).
119. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13,** 2513–2526 (2014).
120. Ludwig, C., Claassen, M., Schmidt, A. & Aebersold, R. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. *Mol. Cell. Proteomics* **11,** M111.013987 (2012).

Acknowledgements We thank M. Faini and R. Ciuffa for help in preparing the figures and Y. Liu for help in compiling the literature citations. M. Hein provided inspiration for this Review, read the manuscript critically and helped with preparing the figures, as did F. Hosp, P. Geyer and S. Beck. We thank members of our groups for critical discussions.

Author information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this paper at go.nature.com/2bqo2n8. Correspondence should be addressed to R.A. (aebersold@imsb.biol.ethz.ch) or M.M. (mmann@biochem.mpg.de).

Appl Microbiol Biotechnol (2013) 97:4749–4762
DOI 10.1007/s00253-013-4897-7

MINI-REVIEW

A review on recent developments in mass spectrometry instrumentation and quantitative tools advancing bacterial proteomics

Laurence Van Oudenhove · Bart Devreese

Received: 19 December 2012 / Revised: 29 March 2013 / Accepted: 3 April 2013 / Published online: 27 April 2013

Abstract Proteomics has evolved substantially since its early days, some 20 years ago. In this mini-review, we aim to provide an overview of general methodologies and more recent developments in mass spectrometric approaches used for relative and absolute quantitation of proteins. Enhancement of sensitivity of the mass spectrometers as well as improved sample preparation and protein fractionation methods are resulting in a more comprehensive analysis of proteomes. We also document some upcoming trends for quantitative proteomics such as the use of label-free quantification methods. Hopefully, microbiologists will continue to explore proteomics as a tool in their research to understand the adaptation of microorganisms to their ever changing environment. We encourage them to incorporate some of the described new developments in mass spectrometry to facilitate their analyses and improve the general knowledge of the fascinating world of microorganisms.

Keywords Mass spectrometry · Quantitative proteomics · Multidimensional chromatography · Microbiology

Introduction

The term "proteome" was created in 1994 by Marc Wilkins to indicate all time- and condition-specific proteins that are simultaneously produced by a cell or a tissue (Anderson and Anderson 1998; Wilkins 2009a). Studying this proteome poses an analytical challenge. The large diversity in protein size and properties as well as in posttranslational modifications makes the proteome much more complex than the genome or the transcriptome. Moreover, (micro-)organisms adapt rapidly to changes in the environment resulting in a highly dynamic protein composition. Proteomics aims to use state-of-the-art protein analysis tools to reveal particular features in the cellular system, including the identification of (subcellular) proteins, the changes in abundance of proteins as well as in their maturation, posttranslational modifications, and degradation of those proteins in response to a certain challenge. Protein networks and their dynamics are resolved in addition to the structure of proteins to allow their functional annotation. Proteomics can thus be considered as a field where researchers provide insights into cellular processes and function by regrouping different pre-fractionation methods, quantification methods, mass spectrometry (MS), and bioinformatics. The recent development of high-throughput proteomic techniques can help in the microbiologist's quest to identify and characterize microorganisms and to study their evolution and origin as well as their interaction with the environment.

The main focus in today's bacterial proteomics consists of the use of quantitative tools to analyze changes in protein abundance in laboratory experiments aiming to measure the effect of changing culture conditions (temperature, nutrients, chemical (antibiotic) treatment). Traditionally, these studies combine two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) and matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS) for protein identification. However, in this review, we will focus on novel non-gel-based, mass spectrometric methods for quantitative proteomics. Thanks to developments in genome sequencing, the scope of microbial proteomics has broadened. Next-generation nucleotide sequencing and automatic annotation pipelines have had a tremendous impact on the number of microbial genomes that are publically available today. This

L. Van Oudenhove · B. Devreese (✉)
Laboratory for Protein Biochemistry and Biomolecular Engineering (L-ProBE), Department of Biochemistry and Microbiology, Ghent University, K.L. Ledeganckstraat 35, 9000 Ghent, Belgium
e-mail: bart.devreese@ugent.be

genomic information is not only crucial for the interpretation of the MS spectra and identification of the proteins but, reversely, proteomics can also aid in a better annotation of the genome by providing proof of existence of predicted proteins and generating better functional annotation (proteogenomics) (Armengaud 2012). Additionally, the collective proteome of microbial communities can be studied as a meta-organism (metaproteomics) under either a controlled environment in the laboratory or in their natural environment (Hettich et al. 2012).

The application areas of microbial proteomics range from fundamental understanding of bacterial physiology (system-wide or specific environmental stress responses, adaptation) to more practical problems including wastewater treatment problems or the effects of metabolic engineering for fermentations (Lacerda and Reardon 2009). Microbial communities are also extensively studied in complex human environments, such as the gastrointestinal system (VerBerkmoes et al. 2009). Proteomics is used in clinical microbiology to study pathogenicity factors by comparing the proteins synthetized by virulent and avirulent strains grown under similar conditions. It is also used to support the development of monoclonal antibodies, serological tools for diagnosis, and vaccine design by identifying immune-reactive proteins (Bensi et al. 2012). Additionally, the development of new antibiotics is increasingly profiting from proteomics for the identification of new targets and the understanding of the mechanisms of action of existing drugs and of antibiotic resistance (Fournier and Raoult 2011). Some examples of the contribution of proteomics to antibiotic drug discovery are the building of a compendium of protein profiles covering mechanisms of action of known antibiotics to achieve a classification of the compounds, to support antibiotic structure improvement programs, to identify toxic effects of possibly new antibiotics, and to support target-based antibiotic discovery (Wecke and Mascher 2011; Wenzel and Bandow 2011).

Proteomics involves multiple techniques and is still an evolving discipline. Here, we will focus on the challenges and the more recent developments in sample preparation and mass spectrometry for the quantitative analysis of microbial proteomes.

Proteomics—basics, opportunities, and limitations

A typical proteomics experiment consists of different stages (Fig. 1). First, the sample preparation stage aims to isolate the proteins from cell lysates or subcellular compartments. Then, this complex mixture of intact proteins is separated using chromatographic or electrophoretic techniques. Individual fractions can then be directly analyzed by mass spectrometry (top-down approach). While significant improvements of this approach have been established, the number of true applications in microbial physiology studies is limited and we therefore refer to specialized literature. More widespread is the use of electrophoretic techniques, either sodium dodecyl sulfate (SDS)-PAGE or 2D-PAGE to separate proteins that are then digested in-gel with a specific protease like trypsin, followed by mass spectrometric analysis of the peptides (bottom-up approach). Alternatively, the proteins can be digested first and the peptide mixture is consequently subjected to a chromatographic separation in order to diminish its complexity before it reaches the mass spectrometer (shotgun approach) (Fig. 1).

Sample preparation and pre-fractionation in bacterial proteomics

A prerequisite for optimal mass spectrometric analysis is the availability of a sample of well-dissolved proteins or peptides, devoid from interfering compounds such as peptidoglycan or extracellular polymeric substances. Typically for microbial proteomics, the cell wall is disrupted first by mechanical means such as sonication or bead milling, by enzymatic digestion with lysozyme, or by the use of detergents. Usually, a combination of these methodologies is employed to obtain proteins in solution (Bhaduri and Demchick 1983; Herbert et al. 2006; Cañas et al. 2007; Abram et al. 2009).

Protein complexes have to be disintegrated and the interactions with other proteins or other molecules have to be broken to obtain soluble proteins which are ready for direct 2D-PAGE, LC-MS analysis, or further proteolytic digestion into peptides. This can be achieved by using chaotropes such as urea and guanidinium hydrochloride, combined with detergents. Meanwhile, proteins have to be protected from proteolysis and modification to reflect the proteome as it was at the time of the cell collection, by the addition of protease and phosphatase inhibitors. Nucleic acids that were released during the extraction can interfere and are preferably removed by the addition of RNAse and DNAse in the lysis buffer. Some of these chaotropes, proteins, enzymes, buffer salts, and detergents can be detrimental to enzymatic digestion or further fractionation and have to be removed. Different (commercial) clean-up methods exist and the choice of the method should be carefully considered, taking into account possible losses of proteins, costs, and foremost the purity grade of the used products as impurities can seriously interfere with LC-MS analysis.

An alternative way to reduce interference of other biomolecules is to perform a subcellular fractionation prior to protein extraction. Different protocols exist to specifically isolate proteins from a subcellular compartment, such as the secreted proteins, outer or inner membrane proteins, and periplasmic and cytoplasmic proteins (Cordwell et al. 2001; Thein et al. 2010). Cellular fractionation has another major advantage: it

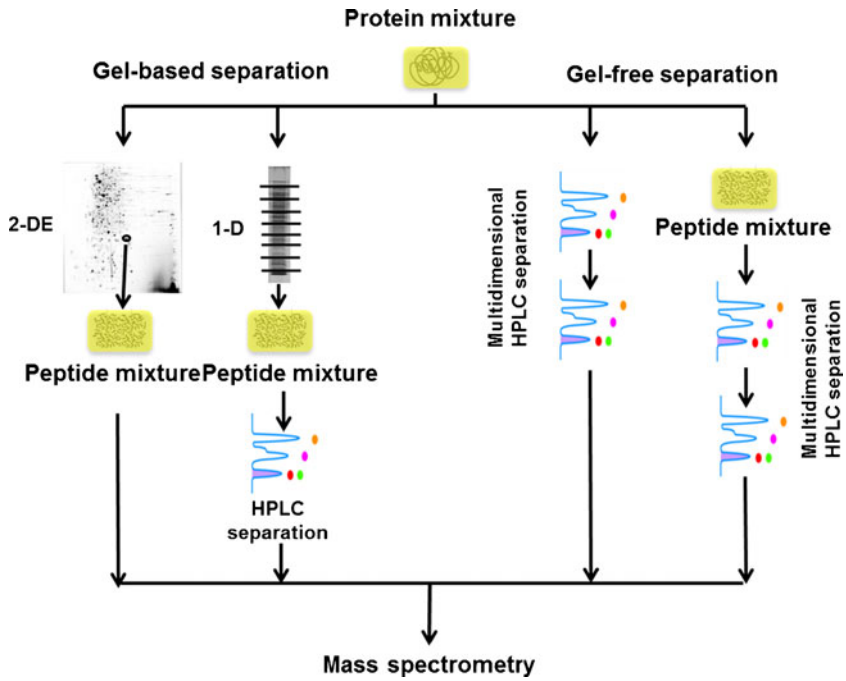


Fig. 1 Workflow of a typical proteomic experiment. First, the proteins are extracted from a sample and then subjected to fractionation before being enzymatically digested into a peptide mixture and identified by mass spectrometry. This reduction in complexity can either be done by using gel-based separation methods such as two-dimensional gel electrophoresis (*2-DE*) and geLC or by using multidimensional HPLC separations of the proteins before mass spectrometric identification

strongly reduces the complexity of the protein mixture and therefore less subsequent separation steps are required. Most commonly, cell fractionation in bacteria is meant to isolate the membrane compartment (Fischer et al. 2006; Hahne et al. 2008; Poetsch and Wolters 2008). Samples obtained from detergent-based dissolving of cell pellets may still contain large amounts of cytosolic proteins. Specific collection of membrane proteins therefore requires more substantial separation steps using two-phase partitioning and density centrifugation (Norling et al. 1998). In classical biochemical protocols, commonly used detergents include SDS, Triton X-100, and amidosulfobetaine 14, but these can be detrimental to subsequent LC-MS approaches. Recently, some manufacturers have solved this problem providing novel acid-labile detergents (Chen et al. 2007). These detergents help in solubilizing proteins throughout the sample preparation protocol, but are cleaved by acid treatment releasing a non-interfering polar group and an insoluble hydrophobic part that can be removed by centrifugation prior to LC-MS analysis.

Advances in analytical chromatography

Most commonly, protein samples are enzymatically digested by trypsin and the resulting peptide mixtures are extensively fractionated by chromatographic separations before being introduced to a mass spectrometer for MS/MS analysis (Fig. 1). Nanoscale reversed-phase high pressure liquid chromatography (RP-HPLC) can easily be hyphenated to a mass spectrometer because of its compatibility of flow rate, solvents, its high resolving power, and reproducibility. However, the resolving power of a single chromatographic separation is often not enough for the very complex peptide mixtures encountered in shotgun proteomics (Nilsson and Davidsson 2000; Shi et al. 2004). Indeed, despite the advances in mass spectrometric instrumentation, both at the level of sensitivity and resolution (see further), undersampling in mass spectrometry is often observed. This can be attributed to, amongst others, limitations in peak capacity at the level of the chromatography, matrix suppression, saturation effects, or MS instrument dwell time. Optimized column dimensions, lengths, and gradient conditions as well as separation temperature are all steps towards higher separation efficiencies (Eeltink et al. 2010; Horie et al. 2012). Exploring the use of sub-2 μm particles for packed RP-LC columns in ultra-high pressure liquid chromatography (UPLC), of silica core–shell particles, or the use of monolithic columns has led to high-throughput separations with improved peak capacities and consequently an increased proteome coverage (Patel et al. 2004; de Villiers et al. 2006; Luo et al. 2007; Sandra et al. 2008; Iwasaki et al. 2010; Rozenbrand et al. 2011). Despite these improvements in single dimension LC-MS, further fractionation of the peptide mixture is typically needed to reduce the complexity and consequently minimizing undersampling during the mass spectrometric analysis (Motoyama and Yates 2008).

The field was revolutionized by the introduction of "multidimensional protein identification technology (MUDPIT)," where the strong cation exchange (SCX), RP-HPLC, and MS analysis were performed in an online hyphenation (Washburn et al. 2001). The separation of peptides is based on two orthogonal methods, i.e., the samples are separated according to unrelated molecular properties to be able to increase the peak capacity and the resolving power of the separation as much as possible (Shi et al. 2004; Gilar et al. 2005a; Motoyama and Yates 2008; Horvatovich et al. 2010).

The combination of SCX gradient elution, fraction collection, and subsequent RP-HPLC separation (offline) followed by electrospray ionization (ESI)-MS measurements has long been the standard approach (Shi et al. 2004; Vollmer et al. 2004). Alternatively, the peptides can be separated in an orthogonal fashion using the combination of two RP-HPLC separations at different pH. First, the peptides are separated at pH 10 followed by a classical separation at pH 3 coupled to ESI-MS (Gilar et al. 2005b; Nakamura et al. 2008). This approach has similar orthogonal properties as the SCX-RP-HPLC, due to the different behavior of the peptides on the RP stationary phase at basic and acidic pH (Gilar et al. 2005a). RP-HPLC at different pH was shown to outperform the SCX-RP-HPLC approach when comparing protein identification numbers (Dowell et al. 2008). In our laboratory, an offline RP/RP-HPLC shotgun approach together with MALDI-TOF/TOF MS was successfully used for the identification of membrane proteins that showed a change in abundance upon antibiotic challenge in the opportunistic pathogen *Stenotrophomonas maltophilia* (Van Oudenhove et al. 2012). Online 2D-RP-UPLC at different pH was applied for the characterization of the proteome of *Methylocella silvestris* grown with methane, succinate, or propane as their carbon source (Patel et al. 2012). The authors showed that performing a two-dimensional separation results in almost a doubling of the identified proteins compared to single LC-MS, in addition to a significant enhancement of the sequence coverage. Another example is the in-depth analysis of the cytosolic proteins in *Corynebacterium glutamicum* (Lasaosa et al. 2009).

Advances in mass spectrometry

There are different types of mass analyzers used in the proteomic field with each its advantages and limitations regarding the sensitivity, accuracy, dynamic range, resolution, and speed of analysis (Domon and Aebersold 2006; Thelen and Miernyk 2012). The basic types of mass analyzers are the quadrupole (Q), the time-of-flight (TOF) analyzer, the ion trap, the Fourier transform ion cyclotron resonance (FT-ICR MS), and the Orbitrap (Fig. 2). They can stand alone or be placed in tandem to take advantage of their individual strengths (reviewed in Aebersold and Mann 2003; Graham et al. 2007). The quadrupole is mostly used as an ion guide to focus ions in an ion trap (Q-TRAP) or reflector TOF (Q-TOF) mass spectrometer and in MS/MS analysis for high resolution selection of peptide ions to be fragmented by collision-induced dissociation (CID). A major advantage of the Q-TOF configuration is the high speed of analysis, allowing state-of-the-art equipment to take MS/MS spectra at 20 Hz rate, dramatically increasing the number of proteins identified in single LC-MS runs (Andrews et al. 2011). The TOF mass analyzer remains a widely applied, versatile, and sensitive component in many mass spectrometers. It is widely used in MALDI-TOF MS analysis for clinical and microbiological diagnosis, where protein profiling or peptide mapping is used as a distinctive tool. When peptide sequence information is aimed, a TOF/TOF instrument can be used. Here, the first TOF analyzer is used as a timed ion gate to select the precursor ions of interest for MS/MS analysis, while the second one separates the fragment ions prior to detection. This instrument became the standard for analysis of 2D-PAGE spots (Vanrobaeys et al. 2003). Recently, the (Q-)TOF was combined with ion mobility devices, where ionized molecules are separated based on their different behavior in a carrier buffer gas. Though ion mobility MS is mostly used in structural biology, it also has been applied recently as an extra dimension of separation in an LC-MS setup, further increasing the proteome coverage (Valentine et al. 2011). It leads to an enhanced and more accurate quantification because of the more accurate interpretation of chimeric MS/MS spectra. This is due to a diminished interference of fragment ions from precursor ions that were present in the selection window, but not intended to be fragmented.

A widely used MS analyzer is the ion trap, which can perform multiple fragmentation cycles (MS^n), where the ions are trapped, fragmented, and analyzed several times after each other. This is an interesting feature for the detection of phosphorylated peptides, for example, where the neutral loss of the phosphate group can trigger an additional round of MS/MS (MS^3) to improve the peptide sequence information and subsequently the identification of the phosphorylated peptide. The more recent linear ion traps that replaced the traditional three-dimensional ion traps offer several advantages. Examples are the faster scan rates and enhanced sensitivity, while a better trapping efficiency and capacity are also achieved. The system can easily be coupled to hybrid devices such as Fourier transform-based mass spectrometers (FTMS) to obtain an ultimate performance in resolution and sensitivity. FTMS is nowadays dominated by Orbitrap mass analyzers, since these instruments show a high mass resolution and accuracy as well as a dynamic range greater than 10^3 at a much lower cost than the classical FT-ion cyclotron resonance instruments (Hu et al. 2005). An improved MS sensitivity was demonstrated for shotgun proteomics using a hybrid linear ion trap Orbitrap instrument. Subparts per million precursor as well as product mass accuracy are achieved after internal calibration (Olsen et al. 2009; Wenger et al. 2010). The technology has been refined, and the current benchtop quadrupole-Orbitrap instrument (Q-Exactive) outperforms other configurations in terms of the numbers of peptide and protein identifications (Michalski et al. 2011b). An example of the use of this instrument in microbial proteomics is the large-scale proteomic analysis of *Mycobacterium tuberculosis* to

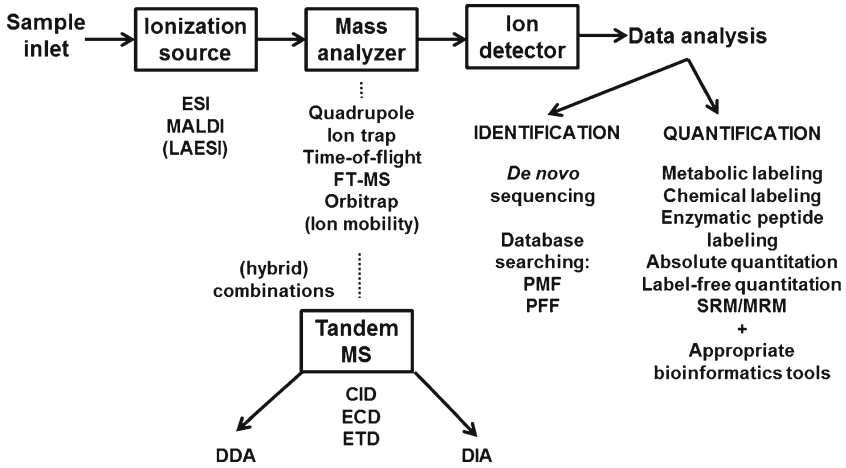


Fig. 2 Schematic overview of the principal components of a mass spectrometric-centered proteomic setup. The abbreviations used in the overview are for electrospray ionization (*ESI*), matrix-assisted laser desorption ionization (*MALDI*), laser ablation electrospray ionization (*LAESI*) for imaging MS, Fourier transform mass spectrometry (*FTMS*), data-dependent mode of acquisition (*DDA*), data-independent mode of acquisition (*DIA*), collision-induced dissociation (*CID*), electron capture dissociation (*ECD*) and electron transfer dissociation (*ETD*), peptide mass fingerprinting (*PMF*), peptide fragment fingerprinting (*PFF*), selected reaction monitoring (*SRM*), and multiple reaction monitoring (*MRM*). The quantification of peptides and proteins can be done using different labeling, label-free, or absolute quantification strategies

improve gene annotations from the Sanger and The Institute for Genomic Research databases (de Souza et al. 2008).

Another upcoming trend is the use of alternative dissociation methods to provide additional sequence information, complementary to that obtained by CID. Electron capture dissociation (ECD) and electron transfer dissociation (ETD) are combined with high mass accuracy MS instruments, such as the LTQ-Orbitrap or SYNAPT G2 QTOF, for high-throughput posttranslational modification analysis (Zubarev et al. 1998; Syka et al. 2004).

Finally, an emerging trend is to replace the traditional data-dependent mode of acquisition (DDA) by a data-independent mode of acquisition (DIA) (Fig. 2). The DDA serial approach for fragmentation typically lets the mass spectrometer cycle through an MS survey scan and then uses automated acquisition software to make a decision on which peptide precursor ions, detected in the MS survey scan, will be selected for MS/MS fragmentation. Ion intensity is one of the key parameters in this decision process, usually selecting precursor ions in a serial manner from the highest to the lowest intensities before these ions are excluded for a limited period of time to allow other less intense precursor ions to be selected (dynamic exclusion). This leads to a bias towards the selection of the most abundant peptide ions in real complex biological samples and to the advent of product ion spectra which are composed of fragment ions from different isobaric and nearly co-eluting peptide ions, resulting in identification difficulties during the database searching (chimeric spectra) (Michalski et al. 2011a). These limitations can largely be overcome with a DIA, where parallel measurement and fragmentation of all peptide precursor ions that are present at that time point is performed. Purvine et al. demonstrated that parallel precursor and fragment acquisition with in-source CID on a TOF MS and subsequent manual alignment followed by SEQUEST peptide identification was feasible (Purvine et al. 2003). DIA by sequential narrowband selection and fragmentation of precursor windows of 10 *m*/*z* within an ion trap has been utilized by the Yates group for the qualitative and quantitative analysis of metabolically labeled yeast (Venable et al. 2004). In LC-MS^E, a HPLC or UPLC separation is combined with a Q-TOF mass spectrometer in which the quadrupole functions as a guide to transfer all ions in the collision cell. The collision energy is continuously switched from low (MS) to high (MS/MS) at a high frequency throughout the analysis (MS^E) (Bateman et al. 2002; Silva et al. 2005, 2006b; Chakraborty et al. 2007). Sophisticated post-acquisition software can align the chromatographic profiles of the precursor and product ions based on retention time and accurate mass measurements to enable subsequent database searching and peptide (protein) identification (Geromanos et al. 2009; Li et al. 2009). This strategy provided an overall protein coverage ranging from 10 to 80 % for an unfractionated *Escherichia coli* proteome (Silva et al. 2006a). Several groups confirmed the dramatic improvement in proteome coverage and protein identification, especially for the lowest abundant proteins, using an LC-MS^E DIA experiment compared to a DDA approach (Geromanos et al. 2009; Patel et al. 2009; Blackburn et al. 2010; Levin et al. 2011). Similarly, the DIA method referred to as precursor acquisition independent from ion count, in which narrow isolation windows (*m*/*z* channels) are sequentially scanned in an ion trap mass spectrometer, regardless of whether a precursor ion is observed or not, resulted in the identification of 70 % or more

of the expressed proteins from *Pseudomonas aeruginosa* without any prior protein fractionation or enrichment method other than the RP-UPLC separation coupled to the LTQ-Orbitrap (Panchaud et al. 2009). Mann et al. used a stand-alone Orbitrap mass spectrometer instead, to allow alternation between MS acquisition and "all-ion fragmentation" MS/MS acquisition in a high-energy collisional dissociation (HCD) cell (Geiger et al. 2010a). Owing to the high resolution and mass accuracy of this instrument, the fragmentation peaks are assigned to their precursor ions on the basis of co-elution profiles. Furthermore, the Aebersold group presented the SWATH MS acquisition method, where a high resolution Q-Q-TOF MS repeatedly cycles through 32 consecutive precursor isolation windows of 25 Da (swaths) for the time-resolved acquisition of fragment ions. SWATH combines this DIA approach with a data analysis method for targeted data extraction resulting in the confident identification of yeast peptides over 4 orders of magnitude (Gillet et al. 2012).

Relative quantitation in proteomics

Proteomics was defined by Anderson and Anderson (1998) as "the use of quantitative protein-level measurements of gene expression to characterize biological processes (e.g., drug effects) and decipher the mechanisms of gene expression control." The study of the changes in abundance of proteins upon certain perturbations such as gene mutations and chemical and environmental variables will help us in understanding what their function are, in addition to elucidating the mechanisms of either action or reaction of these perturbations. Hence, quantitative proteomics is an essential component of "systems biology," which is the attempt to systematically study all concurrent physiological processes in a cell by global measurement of differentially perturbed states (Aebersold and Mann 2003).

Quantitative proteomics was dominated for a long time by 2D-PAGE, particularly after the introduction of immobilized pH gradients and of the difference fluorescent labeling method (DIGE). Although the method has still a number of advantages, including the ability to discriminate posttranslational modified forms of proteins, we will focus here on more recently introduced MS-driven quantitative approaches (Ong and Mann 2005; Bantscheff et al. 2007; Domon and Aebersold 2010; Otto et al. 2012).

Relative quantitation of proteins using metabolic labeling

For metabolic labeling for quantitative proteomics, the cells are grown under a particular condition in media supplemented with either a light or heavy stable isotope of a nutrient, typically a nitrogen source or an amino acid. The proteins are extracted and combined prior to enzymatic digestion, which decreases the experimental error introduced in the sample, highlighting the major advantage of this quantitative technique (Fig. 3). The quantitation is performed using the MS signal, where two peaks are detected for the same peptide with an *m*/*z* interval corresponding to the difference between the light and heavy isotope forms. Oda et al. (1999) initiated the idea of growing mutant yeast supplemented with ^{15}N, added as ammonium salt in the medium, comparing to wild-type yeast grown in normal medium. The method is particularly of interest for the study of protein dynamics, as the incorporation of the isotope will be faster in proteins with a high turnover (Bunai et al. 2005; Rao et al. 2008). Labeling with elementary nitrogen, however, is challenging since all nitrogen, including backbone amide groups, are labeled, and therefore, the resulting mass difference is peptide sequence and size dependent. This challenges the forthcoming data analysis. The use of stable isotope-labeled amino acids is more popular. This so-called stable isotope labeling by amino acids in cell culture (SILAC) strategy was developed by the group of Mann in 2002 (Ong et al. 2002). Here, "essential" amino acids, labeled with a heavy or natural occurring isotope, are supplemented in the amino acid-deficient growth medium and allow for the incorporation of these amino acids in all proteins as they are synthetized (Fig. 3). Typically, $^{13}C/^{15}N$-labeled lysine and/or arginine is used, resulting in a fixed mass difference when trypsin or Lys-C endopeptidase is used. A serious disadvantage is that this metabolic labeling is limited to those organisms that are (made) auxotrophic to these particular amino acids, typically requiring genetic engineering of the lysine or arginine biosynthetic pathway. Therefore, only few applications of this method in microbial proteomics appeared. Soufi et al. (2010) used SILAC on an auxotrophic *Bacillus subtilis* strain to compare the gluconeogenetic growth on succinate with growth under phosphate starvation. They also reported successful identification and quantitation of Ser/Thr/Tyr phosphorylation using this approach. Other applications are described in *E. coli* (Sommer et al. 2010), *Salmonella typhimurium* (Yu and Guo 2011), and recently in *Neisseria gonorrheae* biofilm studies (Phillips et al. 2012). Meanwhile, the SILAC approach was further developed to overcome some of its early limitations. Applications were all in the area of higher eukaryotes, but they might be applicable in prokaryotes. The group of Gevaert combined SILAC with differential sample mixing to overcome the singleton detection problem, where only the light or heavy form of the peptide is detected and thus hampering correct quantification (Impens et al. 2010). Geiger et al. showed recently how SILAC can be expanded to multiplex comparisons with Super-SILAC as well as the use of SILAC as a spiked standard in quantitative proteomics (Geiger et al. 2010b, 2011). pSILAC or pulsed stable isotope labeling by amino acids in cell culture takes the method even further to compare protein dynamics, namely protein translation rates (Schwanhäusser et al. 2009).

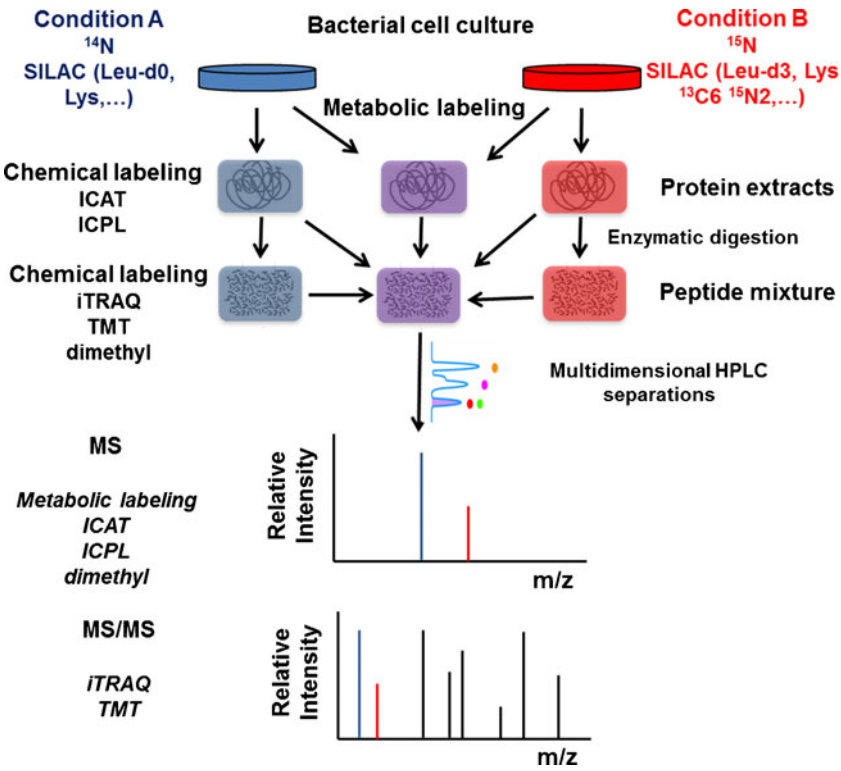


Fig. 3 Schematic overview of some popular metabolic and chemical labeling approaches for quantitative proteomics. A bacterial cell culture grown under two different conditions is compared here (*blue and red colors*). Different labeling strategies are shown on the level of the bacterial growth, extracted proteins, or peptide mixtures. Depending on the method chosen, the quantification is either done using the MS spectrum or the MS/MS spectrum

Relative quantitation of proteins using chemical labeling

As an alternative to metabolic labeling, differential analysis can be achieved using chemical labeling. In 1999, isotope-coded affinity tags (ICAT) were developed to covalently label cysteines in extracted proteins with either a light or heavy (deuterium containing) form of the ICAT reagent (Fig. 3, Gygi et al. 1999). The proteins of the two samples are mixed and digested together before further affinity purification of the ICAT-labeled peptides, reducing the sample complexity towards mass spectrometric analysis. The MS analysis then reveals the peptide intensity ratios corresponding to the quantity of these peptide pairs (mass shift of 8 Da for 2^+ charged peptides) in the two samples. This method was also further developed to overcome some of its initial limitations, as was explained by Goshe and Smith (2003), but the limitation to cysteine containing peptides has substantial disadvantages in terms of missing proteins and the fact that quantitation is often based on a low number of peptides per protein. The use of ICAT seems nowadays to focus on measuring redox states of cells (Sethuraman et al. 2004).

Chemical isotope labeling strategies are nowadays dominated by the use of multiple isobaric tags that are predominantly applied on whole proteome tryptic digests. Several commercial products are available, e.g., isobaric tags for relative and absolute quantification (iTRAQ) or tandem mass tags (TMT) (Thompson et al. 2003; Ross et al. 2004). In both methods, tryptic peptides from different samples are labeled at their N-terminus and lysine side chains, using isobaric tags. A major advantage is that several samples can be multiplexed together for the detection of differences in protein expression (Fig. 3) (Thompson et al. 2003). Moreover, in contrast to ICAT, the reliability of protein identification and quantification as well as the proteome coverage are improved by tagging almost all peptides, by reducing the MS complexity as well as by quantifying at the MS/MS level by the generated reporter ions at specific *m/z* values only (Fig. 3). The Lottspeich group developed a similar strategy, denoted as isotope-coded protein label (ICPL; Schmidt et al. 2005). As the name suggest, it is promoted to be used at the protein level, labeling both N-termini and lysine side chains. This has the advantage that labeled samples can be mixed and an earlier stage, but tryptic digestion is then restricted to arginine side chains resulting in less and larger peptides per protein. Technically spoken, ICPL can also be used at the peptide level (Leroy et al. 2010) and iTRAQ and TMT at the protein level since the basic chemistry for derivatization is the same (succinimide-based amine labeling).

Out of the many applications of iTRAQ and TMT, we list a few examples in the area of the bacterial resistance against antibiotics. Yun et al. (2011) detected common and antibiotic-specific protein responses to tetracycline and imipenem in a clinical *Acinetobacter baumannii* strain. The membrane protein profile of *E. coli* stimulated with an antimicrobial peptide or of *S. maltophilia* upon imipenem challenge was interrogated by combining iTRAQ with a 2D

LC-MS/MS approach (Zhou and Chen 2011; Van Oudenhove et al. 2012)

Isobaric chemical labeling techniques suffer mainly from the advent of chimeric MS/MS spectra, which result in a diminished accuracy of quantification (Altelaar et al. 2012; Evans et al. 2012). Recently, it was demonstrated that an MS^3-based analysis can be used to eliminate this interference problem and holds a solution for the continuation of using isobaric chemical labeling in today's requirements for quantitative proteomics (Ting et al. 2011). Finally, labeling the N-terminus and lysine side chains of peptides by reductive amination or "dimethyl labeling" is also gaining popularity for the quantification of protein abundances (Fig. 3) (Hsu et al. 2003; Boersema et al. 2009).

Label-free quantitation of proteins

Isotope labeling is associated with a number of technical difficulties ranging from the specific requirements for metabolic labeling to reproducibility problems in chemical labeling approaches. Therefore, recent improvements in high throughput and automation of LC-MS instruments and especially the development of novel algorithms dealing with LC-MS data, quantitative proteomics using label-free approaches attracted a lot of interest in the proteomics community. The label-free techniques for bottom-up proteomics can be divided into two groups depending on their correlation with protein abundances in the sample: (1) spectral counting, which counts the number of peptides assigned to a protein in an MS/MS experiment and (2) chromatographic peak area under the curve (AUC) or signal intensity measurement of the precursor ion MS spectra (reviewed by Neilson et al. 2011). Spectral counting relies on the observation that more abundant proteins will be selected more often for fragmentation in a DDA experiment and will thus produce more MS/MS spectra (Liu et al. 2004). The challenges, limitations, and further developments of spectral counting were recently reviewed elsewhere (Lundgren et al. 2010). The performance of spectral counting was compared with that of metabolic labeling and iTRAQ/TMT labeling on a LTQ-Orbitrap Velos using a *Pseudomonas putida* strain (Li et al. 2012). The technique showed improved proteome coverage, but did not outperform the quantification of label-dependent strategies owing to reproducibility problems in the quantitation.

Alternatively, the protein abundance index (PAI) or the number of observed unique peptides divided by the theoretical number of tryptic peptides for each protein that can be observed within a *m*/*z* range is able to estimate the abundance relationship between proteins in a sample (Rappsilber et al. 2002). This technique was further improved using the exponentially modified PAI (emPAI), which is directly proportional to the protein amount in a sample (Ishihama et al. 2005). The emPAI method, however, suffers from saturation when highly abundant proteins are present in the sample, as well as from a decreased correlation with real protein abundances when low resolution mass spectrometers are used. The method of absolute protein expression (APEX) takes the probability of detection of the peptides by MS into account, as was demonstrated in *E. coli* and yeast when estimating the contributions of transcriptional and translational gene regulation (Lu et al. 2007). Because this method uses a machine learning classification algorithm for peptide length and composition, the selection of an appropriate training set can be challenging when facing many unknown proteins in your sample. Finally, Asara et al. (2008) showed that the spectral total ion chromatogram (TIC), which is the average of the total ion count for a protein, can be used as a quantitative value that eliminates the bias towards larger proteins because they generate more tryptic peptides. It also expands the dynamic range of quantification compared to basic spectral counting methods.

In addition to the spectral counting-based label-free quantification methods, the chromatographic peak area or ion intensity (ion count) for a given peptide at a specific retention time in an LC-MS run can be used to quantify proteins in a sample because this measure is linearly correlated to the concentration of that peptide in the sample (Bondarenko et al. 2002; Chelius and Bondarenko 2002). Indeed, ESI generates multiple charged ions with a signal strength proportional to the concentration of the ion (Graham et al. 2007). Under well-standardized LC-MS conditions, peak intensities of a peptide can thus be compared between multiple LC-MS runs. However, this AUC quantitation heavily relies on a good LC resolution and reproducibility and should be performed on MS instruments with a high mass accuracy and resolution for the correct assignment of *m*/*z* values (TOF, Orbitrap). Moreover, this method only holds with the appropriate software analysis for the alignment of the retention times of the different LC-MS runs covering all samples, the peak picking, normalization of peak abundances, and statistics to detect real biological differences in protein abundance (America and Cordewener 2008). Thanks to recent developments in the analysis of label-free data such as dealing with peptides shared amongst proteins, minimizing false discovery rates, data normalization, and appropriate statistical analysis, and label-free proteomics is more reliable than some years ago (Podwojski et al. 2010; Neilson et al. 2011).

As previously discussed in this review, the method called $LC\text{-}MS^E$ solves the alignment issue by using a high resolution mass spectrometer (Q-TOF) which cycles between MS and MS/MS (DIA), enabling the registration of changes in peptide signal response from each accurate mass measurement and retention time (AMRT) value and thus reflecting the concentration of that peptide in a sample compared to another one (Fig. 4) (Silva et al. 2005; Richardson et al. 2012). This label-free method can even allow for absolute

quantitation, when the peak ion count for the three most intense peptides by electrospray ionization and protein concentration is used as a correlation. The inclusion of a protein digest with a known concentration as an internal standard allows a response factor to be calculated from this correlation, which is then applied to proteins with minimally three peptides observed (Silva et al. 2006a). The advantage of this label-free LC-MS^E method, in terms of sample requirement, LC-MS instrument time, and a higher protein coverage compared to the gel-based or iTRAQ-based quantitation methods was clearly demonstrated with *M. silvestris* proteomics (Patel et al. 2009). These advantages are responsible for the promising future of this method for the accurate quantitative analyses of, e.g., many environmental and clinical samples with low sample amount available and for experiments where minimal sample preparation is crucial for the detection and quantification of transient modifications. This approach was also recently validated for accurate quantitation in simple as well as complex samples by Levin et al. (2011).

MS-based validation methods for proteomic results

Unlike the relative quantitative strategies that are mostly used in hypothesis-generating or discovery-driven proteomic strategies, absolute protein quantities can be obtained in targeted proteomic (hypothesis-driven) strategies. Gerber et al. (2003) proposed an absolute quantification (AQUA) strategy for proteins as well as their posttranslationally modified forms, in which synthetic proteotypic peptides with incorporated stable isotopes are used as the ideal internal standard. These internal standard peptides are then used to measure the absolute quantity of a protein of interest after digestion using selected reaction monitoring (SRM) MS measurements. This commercially available approach is the most commonly employed one for absolute quantitation with SRM. An alternative to the expensive and sometimes difficult to synthetize AQUA peptides are the artificial genes encoding a concatenation of isotopically labeled tryptic peptides for the absolute quantification of multiple proteins (QCAT, QconCAT). This QconCAT strategy is more suited for highly multiplexed absolute quantification experiments because the QconCAT can encode between 10 and 30 target proteins with two proteotypic peptides per protein (Beynon et al. 2005; Simpson and Beynon 2012). As with AQUA, the efficiency of tryptic digestion has to be assessed for accurate quantification. Alternatively, the protein standard absolute quantification method (PSAQ) that spikes in isotope-labeled full length target proteins can be incorporated with the samples at the very beginning of the sample preparation workflow, circumventing these problems (Brun et al. 2007; Dupuis et al. 2008). Evidently, the limitations of this approach are the capacity of expressing, purifying, and quantifying the native proteins.

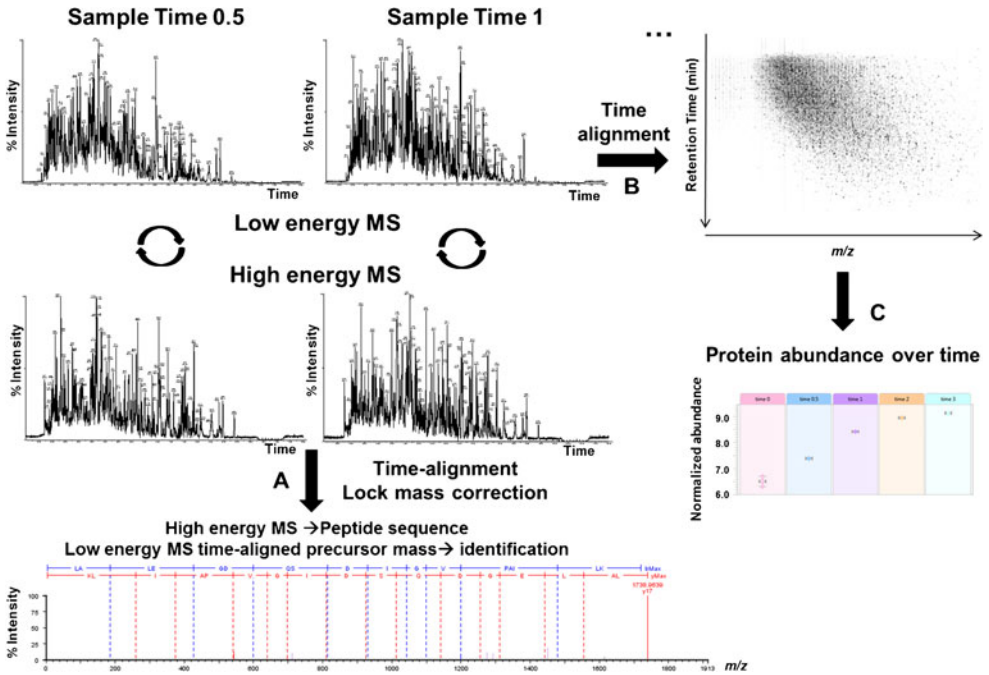


Fig. 4 Example of an LC-MS^E label-free relative quantitation analysis. Proteins from *S. maltophilia* at different time points after antibiotic stimulation were analyzed. The SYNAPT HDMS (Q-TOF) low-energy (MS) and high-energy (MS/MS) base peak ion chromatograms for two samples are shown. Both spectra resulting are time aligned and the masses of the ions are corrected using internal mass calibration for peptide identification (**a**). The precursor peak intensities from similar mass-to-charge ratios (*m*/*z*) measured at a specific retention time are aligned through all the LC-MS runs (**b**). The peak intensities are compared throughout the different samples, and after statistical analysis, changes in protein abundance at different time points before and after antibiotic challenge are observed (**c**)

More information on the recent developments in and differences between these different isotope dilution strategies for absolute quantification can be found in the reviews of Brun et al. (2009) as well as of Picotti and Aebersold (2012).

Malmström et al. developed a groundbreaking strategy combining the absolute quantification of proteins using those isotope-labeled reference peptides (AQUA), the MS intensity-based label-free quantitation, and high-throughput sequencing with LC-MS to obtain the average number of protein copies per cell for a significant portion of the *Leptospira interrogans* proteome (Malmström et al. 2009). Schmidt et al. (2011) used the same idea with a directed MS strategy and AQUA to detect and absolutely quantify tryptic peptides from *L. interrogans* in 25 different conditions, resulting in one of the most complete proteome abundance profile comparisons so far.

SRM is frequently used as a validation technique to detect and quantify targeted proteins with a high precision across a high number of samples. It is used in a LC-MS system mostly using a triple quadrupole or Q-TRAP mass spectrometer, which specifically monitors an analyte ion and one or several predetermined fragment ions generated by CID (SRM transitions). Several SRM transitions can be sequentially measured and thus quantification of multiple analytes can be done across the same LC-MS run, termed multiple reaction monitoring (MRM) (Yocum and Chinnaiyan 2009). Selection of target peptides and transitions is based on previous knowledge and computational tools (Cham Mead et al. 2010). The application of SRM in proteomics together with relative or absolute quantification strategies, advances, pitfalls, and future directions was recently reviewed by Picotti and Aebersold (2012). SRM is also used to study protein modifications such as phosphorylation (Cox et al. 2005). Currently, researchers are exploring possibilities to increase the multiplexing capabilities of SRM. Intelligent SRM, e.g., monitors intense transitions for a peptide's precise quantification. When a preset threshold is exceeded, additional transition signals are acquired in a DDA manner to confirm the peptide identity (Kiyonami et al. 2011). Another example is the use of fragment ion spectra from all precursors using SWATH MS followed by targeted SRM to uniquely identify peptides in the DIA fragment ion maps (Gillet et al. 2012). Increased SRM sensitivity and specificity could be obtained by coupling SRM to ion mobility separation. Moreover, improved bioinformatics tools for the prediction of SRM transitions as well as the data evaluation will increase the specificity of SRM assays in the future.

Future applications for microbial quantitative proteomics

Improvements in sensitivity, mass accuracy, and MS/MS capabilities of mass spectrometers had a tremendous impact on the field of proteomics since its inception some 20 years ago. For quantitative proteomics, we have observed that gel-free shotgun proteomics methods are replacing 2-DE and that label-free quantitative approaches growingly become more popular than the labeling techniques. However, the measurement of protein abundances by 2-DE or shotgun proteomics provides an overview of the expressed protein abundances in a cell at a single time point, but mostly no insight into the dynamic changes (Wilkins 2009b; Doherty and Whitfield 2011). Measuring protein turnover on a proteome-wide scale as response to changes in the environment will be necessary to improve a more complete view of a biological system. Similarly, increasing the sequence coverage of proteins by digestion procedures complementary to the traditional trypsin digestion will have to be further exploited for a more comprehensive view of the "complete" proteome (Thelen and Miernyk 2012). In the future, microbial proteomics will also have to evolve from profiling and expression studies towards the comprehensive analysis of the role of different posttranslational modifications in certain biological processes, the function of protein complexes, and interactions instead of individual proteins as well as the spatial localization of proteins in bacterial cells or even multicellular structures, such as biofilms, by imaging MS (Blaze et al. 2012).

Further developments in tools for data mining, protein functional annotating, and finding meaningful answers to the posed biological questions are needed. All too often, proteomics experiments fail to provide concluding results as a result from an inadequate experimental design. Therefore, a close collaboration with bioinformaticians and statisticians will be needed. Additionally, the use of publically available data repositories for acquired proteomic data as well as a better cooperation of different institutes and databases to create a clear set of gene names, symbols, functional annotation, and predicted localization as well as possible modifications will be beneficial for the proteomic community to continue to thrive.

Finally, a more transparent reporting of proteomic results will help in the correct interpretation of the available proteomic data sets and improve the comparison of the existing quantitative proteomic approaches as well as the development and testing of new ones (Taylor et al. 2007; Mead et al. 2009). MS-based quantitative proteomics is still evolving rapidly and will continue to be a tremendously important tool for deciphering complex biological processes such as microbial communities and their creative adaptation mechanisms towards environmental changes.

Acknowledgments The authors are indebted to the Belgian Federal Government's Interuniversity Attraction Pole Action P7/44, to the "Bijzonder Onderzoeksfonds" from Ghent University for a concerted action grant, and to the Hercules Foundation (grant AUGENT019).

Conflict of interest The authors have declared no conflict of interest.

References

Abram F, Gunnigle E, O'Flaherty V (2009) Optimisation of protein extraction and 2-DE for metaproteomics of microbial communities from anaerobic wastewater treatment biofilms. Electrophoresis 30:4149–4151

Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422:198–207

Altelaar AFM, Frese CK, Preisinger C, Hennrich ML, Schram AW, Timmers HTM, Heck AJR, Mohammed S (2012) Benchmarking stable isotope labeling based quantitative proteomics. J Proteomics. doi:10.1016/j.jprot.2012.10.009

America AH, Cordewener JH (2008) Comparative LC-MS: a landscape of peaks and valleys. Proteomics 8:731–749

Anderson NL, Anderson NG (1998) Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis 19:1853–1861

Andrews GL, Simons BG, Bryce Young J, Hawkridge AM, Muddiman DC (2011) Performance characteristics of a new hybrid quadrupole time-of-flight tandem mass spectrometer (TripleTOF 5600). Anal Chem 83:5442–5446

Armengaud J (2012) Microbiology and proteomics, getting the best of both worlds! Environ Microbiol. doi:10.1111/j.1462-2920.2012.02811.x

Asara JM, Christofk HR, Freimark LM, Cantley LC (2008) A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen. Proteomics 8:994–999

Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B (2007) Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem 389:1017–1031

Bateman RH, Carruthers R, Hoyes JB, Jones C, Langridge JI, Millar A, Vissers JP (2002) A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation. J Am Soc Mass Spectrom 13:792–803

Bensi G, Mora M, Tuscano G, Biagini M, Chiarot E, Bombaci M, Capo S, Falugi F, Manetti AG, Donato P, Swennen E, Gallotta M, Garibaldi M, Pinto V, Chiappini N, Musser JM, Janulczyk R, Mariani M, Scarselli M, Telford JL, Grifantini R, Norais N, Margarit I, Grandi G (2012) Multi high-throughput approach for highly selective identification of vaccine candidates: the group A *Streptococcus* case. Mol Cell Proteomics 11:M111.015693

Beynon RJ, Doherty MK, Pratt JM, Gaskell SJ (2005) Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. Nat Methods 2:587–589

Bhaduri S, Demchick PH (1983) Simple and rapid method for disruption of bacteria for protein studies. Appl Environ Microbiol 46:941–943

Blackburn K, Mbeunkui F, Mitra SK, Mentzel T, Goshe MB (2010) Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation. J Proteome Res 9:3621–3637

Blaze MTM, Aydin B, Carlson RP, Hanley L (2012) Identification and imaging of peptides and proteins on *Enterococcus faecalis* biofilms by matrix assisted laser desorption ionization mass spectrometry. Analyst 137:5018–5025

Boersema PJ, Raijmakers R, Lemeer S, Mohammed S, Heck AJR (2009) Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. Nat Protoc 4:484–494

Bondarenko PV, Chelius D, Shaler TA (2002) Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. Anal Chem 74:4741–4749

Brun V, Dupuis A, Adrait A, Marcellin M, Thomas D, Court M, Vandenesch F, Garin J (2007) Isotope-labeled protein standards : towards absolute quantitative proteomics. Mol Cell Proteomics 6:2139–2149

Brun V, Masselon C, Garin J, Dupuis A (2009) Isotope dilution strategies for absolute quantitative proteomics. J Proteomics 72:740–749

Bunai K, Nozaki M, Kakeshita H, Nemoto T, Yamane K (2005) Quantification of de novo localized ^{15}N-labeled lipoproteins and membrane proteins having one and two transmembrane segments in a *Bacillus subtilis* secA temperature-sensitive mutant using 2D-PAGE and MALDI-TOF MS. J Prot Res 4:826–836

Cañas B, Piñeiro C, Calvo E, López-Ferrer D, Gallardo JM (2007) Trends in sample preparation for classical and second generation proteomics. J Chromatogr A 1153:235–258

Chakraborty AB, Berger SJ, Gebler JC (2007) Use of an integrated MS-multiplexed MS/MS data acquisition strategy for high-coverage peptide mapping studies. Rapid commun mass spectrom 21:730–744

Cham Mead JA, Bianco L, Bessant C (2010) Free computational resources for designing selected reaction monitoring transitions. Proteomics 10:1106–1126

Chelius D, Bondarenko PV (2002) Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. J Proteome Res 1:317–323

Chen EI, Cociorva D, Norris JL, Yates JR III (2007) Optimization of mass spectrometry-compatible surfactants for shotgun proteomics. J Proteome Res 6:2529–2538

Cordwell SJ, Nouwens AS, Walsh BJ (2001) Comparative proteomics of bacterial pathogens. Proteomics 1:461–472

Cox DM, Zhong F, Du M, Duchoslav E, Sakuma T, McDermott JC (2005) Multiple reaction monitoring as a method for identifying protein posttranslational modifications. J Biomol Tech 16:83–90

de Souza GA, Målen H, Søfteland T, Sælensminde G, Prasad S, Jonassen I, Wiker HG (2008) High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using *Mycobacterium tuberculosis* as an example. BMC Genomics 9:316. doi:10.1186/1471-2164-9-316

de Villiers A, Lestremau F, Szucs R, Gélébart S, David F, Sandra P (2006) Evaluation of ultra performance liquid chromatography. Part I. Possibilities and limitations. J Chromatogr A 1127:60–69

Doherty MK, Whitfield PD (2011) Proteomics moves from expression to turnover: update and future perspective. Expert Rev Proteomics 8:325–334

Domon B, Aebersold R (2006) Mass spectrometry and protein analysis. Science 312:212–217

Domon B, Aebersold R (2010) Options and considerations when selecting a quantitative proteomics strategy. Nat Biotechnol 28:710–721

Dowell JA, Frost DC, Zhang J, Li L (2008) Comparison of two-dimensional fractionation techniques for shotgun proteomics. Anal Chem 80:6715–6723

Dupuis A, Hennekinne JA, Garin J, Brun V (2008) Protein Standard Absolute Quantification (PSAQ) for improved investigation of staphylococcal food poisoning outbreaks. Proteomics 8:4633–4636

Eeltink S, Dolman S, Detobel F, Swart R, Ursem M, Schoenmakers PJ (2010) High-efficiency liquid chromatography-mass spectrometry separations with 50 mm, 250 mm, and 1 m long polymer-based monolithic capillary columns for the characterization of complex proteolytic digests. J Chromatogr A 1217:6610–6615

Evans C, Noirel J, Ow SY, Salim M, Pereira-Medrano AG, Couto N, Pandhal J, Smith D, Pham TK, Karunakaran E, Zou X, Biggs CA, Wright PC (2012) An insight into iTRAQ: where do we stand now? Anal Bioanal Chem 404:1011–1027

Fischer F, Wolters D, Rögner M, Poetsch A (2006) Toward the complete membrane proteome: high coverage of integral membrane proteins through transmembrane peptide detection. Mol Cell Proteomics 5:444–453

Fournier PE, Raoult D (2011) Prospects for the future using genomics and proteomics in clinical microbiology. Annu Rev Microbiol 65:169–188

Geiger T, Cox J, Mann M (2010a) Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. Mol Cell Proteomics 9:2252–2261

Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M (2010b) Super-SILAC mix for quantitative proteomics of human tumor tissue. Nat Methods 7:383–385

Geiger T, Wisniewski JR, Cox J, Zanivan S, Kruger M, Ishihama Y, Mann M (2011) Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. Nat Protoc 6:147–157

Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci U S A 100:6940–6945

Geromanos SJ, Vissers JP, Silva JC, Dorschel CA, Li GZ, Gorenstein MV, Bateman RH, Langridge JI (2009) The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. Proteomics 9:1683–1695

Gilar M, Olivova P, Daly AE, Gebler JC (2005a) Orthogonality of separation in two-dimensional liquid chromatography. Anal Chem 77:6426–6434

Gilar M, Olivova P, Daly AE, Gebler JC (2005b) Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. J Sep Sci 28:1694–1703

Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, Bonner R, Aebersold R (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics 11(O111):016717. doi:10.1074/mcp.O111.016717

Goshe MB, Smith RD (2003) Stable isotope-coded proteomic mass spectrometry. Curr Opin Biotechnol 14:101–109

Graham RLJ, Graham C, McMullan G (2007) Microbial proteomics: a mass spectrometry primer for biologists. Microb Cell Fact 6:26. doi:10.1186/1475-2859-6-26

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nature Biotechnol 17:994–999

Hahne H, Wolff S, Hecker M, Becher D (2008) From complementarity to comprehensiveness—targeting the membrane proteome of growing *Bacillus subtilis* by divergent approaches. Proteomics 8:4123–4136

Herbert BR, Grinyer J, McCarthy JT, Isaacs M, Harry EJ, Nevalainen H, Traini MD, Hunt S, Schulz B, Laver M, Goodall AR, Packer J, Harry JL, Williams KL (2006) Improved 2-DE of microorganisms after acidic extraction. Electrophoresis 27:1630–1640

Hettich RL, Sharma R, Chourey K, Giannone RJ (2012) Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. Curr Opin Microbiol 15:373–380

Horie K, Sato Y, Kimura T, Nakamura T, Ishihama Y, Oda Y, Ikegami T, Tanaka N (2012) Estimation and optimization of the peak capacity of one-dimensional gradient high performance liquid chromatography using a long monolithic silica capillary column. J Chromatogr A 1228:283–291

Horvatovich P, Hoekman B, Govorukhina N, Bischoff R (2010) Multidimensional chromatography coupled to mass spectrometry in analysing complex proteomics samples. J Sep Sci 33:1421–1437

Hsu JL, Huang SY, Chow NH, Chen SH (2003) Stable-isotope dimethyl labeling for quantitative proteomics. Anal Chem 75:6843–6852

Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Cooks RG (2005) The Orbitrap: a new mass spectrometer. J Mass Spectrom 40:430–443

Impens F, Colaert N, Helsens K, Ghesquière B, Timmerman E, De Bock PJ, Chain BM, Vandekerckhove J, Gevaert K (2010) A quantitative proteomics design for systematic identification of protease cleavage events. Mol Cell Proteomics 9:2327–2333

Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics 4:1265–1272

Iwasaki M, Miwa S, Ikegami T, Tomita M, Tanaka N, Ishihama Y (2010) One-dimensional capillary liquid chromatographic separation coupled with tandem mass spectrometry unveils the *Escherichia coli* proteome on a microarray scale. Anal Chem 82:2616–2620

Kiyonami R, Schoen A, Prakash A, Peterman S, Zabrouskov V, Picotti P, Aebersold R, Huhmer A, Domon B (2011) Increased selectivity, analytical precision, and throughput in targeted proteomics. Mol Cell Proteomics 10(M110):002931. doi:10.1074/mcp.M110.002931-1

Lacerda CM, Reardon KF (2009) Environmental proteomics: applications of proteome profiling in environmental microbiology and biotechnology. Brief Funct Genomic Proteomic 8:75–87

Lasaosa M, Delmotte N, Huber CG, Melchior K, Heinzle E, Tholey A (2009) A 2D reversed-phase x ion-pair reversed-phase HPLC-MALDI TOF/TOF-MS approach for shotgun proteome analysis. Anal Bioanal Chem 393:1245–1256

Leroy B, Rosier C, Erculisse V, Leys N, Mergeay M, Wattiez R (2010) Differential proteomic analysis using isotope-coded protein-labeling strategies: comparison, improvements and application to simulated microgravity effect on *Cupriavidus metallidurans* CH34. Proteomics 10:2281–2291

Levin Y, Hradetzky E, Bahn S (2011) Quantification of proteins using data-independent analysis (MS^E) in simple andcomplex samples: a systematic evaluation. Proteomics 11:3273–3287

Li GZ, Vissers JP, Silva JC, Golick D, Gorenstein MV, Geromanos SJ (2009) Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. Proteomics 9:1696–1719

Li Z, Adams RM, Chourey K, Hurst GB, Hettich RL, Pan C (2012) Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ Orbitrap Velos. J Proteome Res 11:1582–1590

Liu H, Sadygov RG, Yates JR III (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76:4193–4201

Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nature Biotechnol 25:117–124

Lundgren DH, Hwang SI, Wu L, Han DK (2010) Role of spectral counting in quantitative proteomics. Expert Rev Proteomics 7:39–53

Luo Q, Page JS, Tang K, Smith RD (2007) MicroSPE-nanoLC-ESI-MS/MS using 10-μm-i.d. silica-based monolithic columns for proteomics. Anal Chem 79:540–545

Malmström J, Beck M, Schmidt A, Lange V, Deutsch EW, Aebersold R (2009) Proteome-wide cellular protein concentration of the human pathogen *Leptospira interrogans*. Nature 460:762–765

Mead JA, Bianco L, Bessant C (2009) Recent developments in public proteomic MS repositories and pipelines. Proteomics 9:861–881

Michalski A, Cox J, Mann M (2011a) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. J Proteome Res 10:1785–1793

Michalski A, Damoc E, Hauschild JP, Lange O, Wieghaus A, Makarov A, Nagaraj N, Cox J, Mann M, Horning S (2011b) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop

quadrupole Orbitrap mass spectrometer. Mol Cell Proteomics 10: M111.011015
Motoyama A, Yates JR III (2008) Multidimensional LC separations in shotgun proteomics. Anal Chem 80:7187–7193
Nakamura T, Kuromitsu J, Oda Y (2008) Evaluation of comprehensive multidimensional separations using reversed-phase, reversed-phase liquid chromatography/mass spectrometry for shotgun proteomics. J Proteome Res 7:1007–1011
Neilson KA, Ali NA, Muralidharan S, Mirzaei M, Mariani M, Assadourian G, Lee A, van Sluyter SC, Haynes PA (2011) Less label, more free: approaches in label-free quantitative mass spectrometry. Proteomics 11:535–553
Nilsson CL, Davidsson P (2000) New separation tools for comprehensive studies of protein expression by mass spectrometry. Mass Spectrom Rev 19:390–397
Norling B, Zak E, Andersson B, Pakrasi H (1998) 2D-isolation of pure plasma and thylakoid membranes from the cyanobacterium *Synechocystis* sp. PCC 6803. FEBS Lett 436:189–192
Oda Y, Huang K, Cross FR, Cowburn D, Chait BT (1999) Accurate quantitation of protein expression and site-specific phosphorylation. Proc Natl Acad Sci U S A 96:6591–6596
Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, Denisov E, Lange O, Remes P, Taylor D, Splendore M, Wouters ER, Senko M, Makarov A, Mann M, Horning S (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. Mol Cell Proteomics 8:2759–2769
Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 1:376–386
Ong SE, Mann M (2005) Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol 1:252–262
Otto A, Bernhardt J, Hecker M, Becher D (2012) Global relative and absolute quantitation in microbial proteomics. Curr Opin Microbiol 15:364–372
Panchaud A, Scherl A, Shaffer SA, von Haller PD, Kulasekara HD, Miller SI, Goodlett DR (2009) Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. Anal Chem 81:6481–6488
Patel KD, Jerkovich AD, Link JC, Jorgenson JW (2004) In-depth characterization of slurry packed capillary columns with 1.0-μm nonporous particles using reversed-phase isocratic ultrahigh-pressure liquid chromatography. Anal Chem 76:5777–5786
Patel NA, Crombie A, Slade SE, Thalassinos K, Hughes C, Connolly JB, Langridge J, Murrell JC, Scrivens JH (2012) Comparison of one- and two-dimensional liquid chromatography approaches in the label-free quantitative analysis of *Methylocella silvestris*. J Proteome Res 11:4755–4763
Patel VJ, Thalassinos K, Slade SE, Connolly JB, Crombie A, Murrell JC, Scrivens JH (2009) A comparison of labeling and label-free mass spectrometry-based proteomics approaches. J Proteome Res 8:3752–3759
Phillips NJ, Steichen CT, Schilling B, Post DM, Niles RK, Bair TB, Falsetta ML, Apicella MA, Gibson BW (2012) Proteomic analysis of *Neisseria gonorrhoeae* biofilms shows shift to anaerobic respiration and changes in nutrient transport and outer membrane proteins. PLOS One 7(6):e38303
Picotti P, Aebersold R (2012) Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods 9:555–566
Podwojski K, Eisenacher M, Kohl M, Turewicz M, Meyer HE, Rahnenführer J, Stephan C (2010) Peek a peak: a glance at statistics for quantitative label-free proteomics. Expert Rev Proteomics 7:249–261
Poetsch A, Wolters D (2008) Bacterial membrane proteomics. Proteomics 8:4100–4122
Purvine S, Eppel JT, Yi EC, Goodlett DR (2003) Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. Proteomics 3:847–850
Rao PK, Marcela Rodriguez G, Smith I, Li Q (2008) Protein dynamics in iron-starved *Mycobacterium tuberculosis* revealed by turnover and abundance measurement using hybrid-linear ion trap-Fourier transform mass spectrometry. Anal Chem 80:6860–6869
Rappsilber J, Ryder U, Lamond AI, Mann M (2002) Large-scale proteomic analysis of the human spliceosome. Genome Res 12:1231–1245
Richardson K, Denny R, Hughes C, Skilling J, Sikora J, Dadlez M, Manteca A, Jung HR, Jensen ON, Redeker V, Melki R, Langridge JI, Vissers JPC (2012) A probabilistic framework for peptide and protein quantification from data-dependent and data-independent LC-MS proteomics experiments. OMICS 16:468–482
Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 3:1154–1169
Rozenbrand J, de Jong GJ, van Bennekom WP (2011) Comparison of monolithic and 1.8-μm RP-18 silica capillary columns using chromatographic data and mass spectrometric identification scores for proteins. J Sep Sci 34:2199–2205
Sandra K, Moshir M, D'hondt F, Verleysen K, Kas K, Sandra P (2008) Highly efficient peptide separations in proteomics Part 1. Unidimensional high performance liquid chromatography. J Chromatogr B Analyt Technol Biomed Life Sci 866:48–63
Schmidt A, Kellermann J, Lottspeich F (2005) A novel strategy for quantitative proteomics using isotope-coded protein labels. Proteomics 5:4–15
Schmidt A, Beck M, Malmström J, Lam H, Claassen M, Campbell D, Aebersold R (2011) Absolute quantification of microbial proteomes at different states by directed mass spectrometry. Mol Syst Biol 7:510. doi:10.1038/msb.2011.37
Schwanhäusser B, Gossen M, Dittmar G, Selbach M (2009) Global analysis of cellular protein translation by pulsed SILAC. Proteomics 9:205–209
Sethuraman M, McComb ME, Heibeck T, Costello CE, Cohen RA (2004) Isotope-coded affinity tag approach to identify and quantify oxidant-sensitive protein thiols. Mol Cellul Prot 3:273–278
Shi Y, Xiang R, Horváth C, Wilkins JA (2004) The role of liquid chromatography in proteomics. J Chromatogr A 1053:27–36
Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li GZ, McKenna T, Nold MJ, Richardson K, Young P, Geromanos S (2005) Quantitative proteomic analysis by accurate mass retention time pairs. Anal Chem 77:2187–2200
Silva JC, Denny R, Dorschel C, Gorenstein MV, Li GZ, Richardson K, Wall D, Geromanos SJ (2006a) Simultaneous qualitative and quantitative analysis of the *Escherichia coli* proteome: a sweet tale. Mol Cell Proteomics 5:589–607
Silva JC, Gorenstein MV, Li GZ, Vissers JPC, Geromanos SJ (2006b) Absolute quantification of proteins by LCMS^{E}: a virtue of parallel MS acquisition. Mol Cell Proteomics 5:144–156
Simpson DM, Beynon RJ (2012) QconCATs: design and expression of concatenated protein standards for multiplexed protein quantification. Anal Bioanal Chem 404:977–989
Sommer U, Petersen J, Pfeiffer M, Schrotz-King P, Morsczeck C (2010) Comparison of surface proteomes of enterotoxigenic (ETEC) and commensal *Escherichia coli* strains. J Microbiol Methods 83:13–19
Soufi B, Kumar C, Gnad F, Mann M, Mijakovic I, Macek B (2010) Stable isotope labeling by amino acids in cell culture (SILAC) applied to quantitative proteomics of *Bacillus subtilis*. J Proteome Res 9:3638–3646

Syka JEP, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. Proc Natl Acad Sci U S A 101:9528–9533

Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJR, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL, Souda P, Tsugita A, Vandekerckhove J, Vondriska TM, Whitelegge JP, Wilkins MR, Xenarios I, Yates JR III, Hermjakob H (2007) The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol 25:887–893

Thein M, Sauer G, Paramasivam N, Grin I, Linke D (2010) Efficient subfractionation of Gram-negative bacteria for proteomics studies. J Proteome Res 9:6135–6147

Thelen JJ, Miernyk JA (2012) The proteomic future: where mass spectrometry should be taking us. Biochem J 444:169–181

Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AK, Hamon C (2003) Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 75:1895–1904

Ting L, Rad R, Gygi SP, Haas W (2011) MS3 eliminates ratio distortion in isobaric labeling-based multiplexed quantitative proteomics. Nat Methods 8:937–940

Valentine SJ, Ewing MA, Dilger JM, Glover MS, Geromanos S, Hughes C, Clemmer DE (2011) Using ion mobility data to improve peptide identification: intrinsic amino acid size parameters. J Proteome Res 10:2318–2329

Van Oudenhove L, De Vriendt K, Van Beeumen J, Mercuri PS, Devreese B (2012) Differential proteomic analysis of the response of *Stenotrophomonas maltophilia* to imipenem. Appl Microbiol Biotechnol 95:717–733

Vanrobaeys F, Devreese B, Lecocq E, Rychlewski L, De Smet L, Van Beeumen J (2003) Proteomics of the dissimilatory iron-reducing bacterium *Shewanella oneidensis* MR-1, using a matrix-assisted laser desorption/ionization-tandem-time of flight mass spectrometer. Proteomics 3:2249–2257

Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods 1:39–45

Verberkmoes NC, Russell AL, Shah M, Godzik A, Rosenquist M, Halfvarson J, Lefsrud MG, Apajalahti J, Tysk K, Hettich RL, Jansson JK (2009) Shotgun proteomics of the human distal gut microbiota. ISME J 3:179–189

Vollmer M, Horth P, Ngele E (2004) Optimization of two-dimensional off-line LC/MS separations to improve resolution of complex proteomic samples. Anal Chem 76:5180–5185

Washburn MP, Wolters D, Yates JR III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature Biotechnol 19:242–247

Wecke T, Mascher T (2011) Antibiotic research in the age of omics: from expression profiles to interspecies communication. J Antimicrob Chemother 66:2689–2704

Wenger CD, McAlister GC, Xia Q, Coon JJ (2010) Sub-part-per-million precursor and product mass accuracy for high-throughput proteomics on an electron transfer dissociation-enabled Orbitrap mass spectrometer. Mol Cell Proteomics 9:754–763

Wenzel M, Bandow JE (2011) Proteomic signatures in antibiotic research. Proteomics 11:3256–3268

Wilkins M (2009a) Proteomics data mining. Expert Rev Proteomics 6:599–603

Wilkins MR (2009b) Hares and tortoises: the high- versus low-throughput proteomic race. Electrophoresis 30:S150–S155

Yocum AK, Chinnaiyan AM (2009) Current affairs in quantitative targeted proteomics: multiple reaction monitoring-mass spectrometry. Brief Funct Genomic Proteomic 8:145–157

Yu J, Guo J (2011) Quantitative proteomic analysis of *Salmonella enterica* serovar *Typhimurium* under PhoP/PhoQ activation conditions. J Prot Res 10:2992–2302

Yun SH, Choi CW, Kwon SO, Park GW, Cho K, Kwon KH, Kim JY, Yoo JS, Lee JC, Choi JS, Kim S, Kim SI (2011) Quantitative proteomic analysis of cell wall and plasma membrane fractions from multidrug-resistant *Acinetobacter baumannii*. J Proteome Res 10:459–469

Zhou Y, Chen WN (2011) iTRAQ-coupled 2-D LC-MS/MS analysis of membrane protein profile in *Escherichia coli* incubated with apidaecin IB. PLoS One 6:e20442. doi:10.1371/journal.pone.0020442

Zubarev RA, Kelleher NL, McLafferty FW (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. J Am Chem Soc 120:3265–3266

7 Metabolomics

Lecturers: **Dr. Michel van Weeghel (Lab. For Genetic Metabolic Disease, AMC)**

After reading this chapter you should understand

- The basic principles of metabolomics and lipidomics.
- The metabolomics workflow
- Different approaches towards data acquisition

This chapter is adapted from Vaz FM, Pras-Raves M, Bootsma AH, van Kampen AH. (2015) Principles and practice of lipidomics. J Inherit Metab Dis., 38(1):41-52.

Contents

Abbreviations

- APCI Atmospheric pressure chemical ionization
- ESI Electrospray ionization
- FWHM Full width at half maximum
- LC Liquid chromatography
- MALDI Matrix-Assisted Laser Desorption/Ionization
- MDMS Multi-dimensional MS
- MRM Multiple reaction monitoring
- (U)HPLC (Ultra)-High performance chromatography
- Q Quadrupoles
- QC Quality Control
- QqQ Triple quadrupole instrument
- TOF Time-of-flight

7.1 Introduction

Lipids are highly diverse molecules which are traditionally best known for their role in the formation of biological membranes in cellular systems and as a way to store energy. In the last decade, lipids have outgrown this rather dull image and have taken center stage in apoptosis, cell signaling, inflammation, immunity and, last but not least, inborn errors of metabolism. Lipids are sometimes simply defined as molecules that are insoluble in water and soluble in organic solvents. As this is not true for all lipids, a more generally accepted definition has been described by the International Lipid Classification and Nomenclature Committee of lipids: “hydrophobic or amphipathic small molecules that may originate entirely or in part by carbanion-based condensation of thioester (fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, saccharolipids, and polyketides) and/or by carbocation-based condensations of isoprene units (prenol lipids and sterol lipids)”. This designation describes eight main categories that are subdivided based on their chemical properties. This classification also has been used by the LIPID Metabolites And Pathways Strategy (LIPID MAPS) consortium, a multi-institutional effort to identify and quantitate lipid species in mammalian cells (Fahy et al 2009). This classification system has been widely accepted and the LIPID MAPS database contains more than 37.500 unique structures for biologically relevant lipids from mammals, plants, bacteria, fungi, algae, and marine organisms ((Fahy et al 2009), Table 7.1).

Lipids were traditionally analyzed by thin-layer chromatography, gas chromatography and mass spectrometry. Technical advances in mass spectrometry have paved the way for the realization of a new type of metabolomics: lipidomics. Lipidomics aims to study the pathways and networks of cellular lipids by characterization and quantitation of all lipids present in a biological system. Especially the development of “soft” ionization techniques as electrospray ionization (ESI) and exact mass resolution, high resolution mass spectrometers have greatly propelled the field of lipidomics. In addition, new bioinformatics tools have been developed to cope with the increasing amounts of raw data and extract relevant information to yield biological insight. The parallel rise of next-generation sequencing and concomitant identification of new inborn errors of metabolism in genes encoding lipid modifying enzymes has brought about the desire to fully characterize the lipidome, further boosting the development and refinement of lipidomic techniques. Lipidomics is being applied to find biomarkers

suitable for diagnosis, follow-up and ideally prognosis in order to characterize the course of the patient's disorder. In addition to linking new monogenetic inborn errors of lipid metabolism there is a growing number of links being uncovered between lipid metabolism and complex genetic traits as obesity, diabetes, atherosclerosis and cancer (Hyotylainen and Oresic 2014). The corresponding research communities increasingly use lipidomics underscoring the broad application of this technique for research in general.

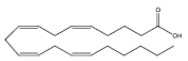
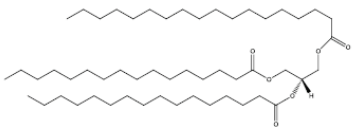
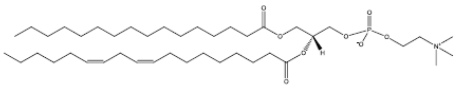
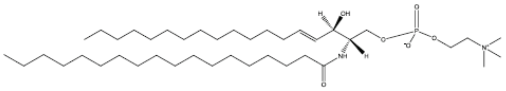
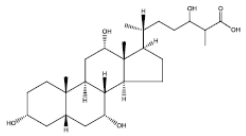
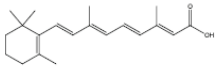
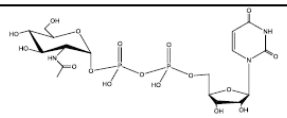
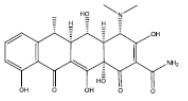
Lipid categories	Example	Main classes (selection)	Biological role	IEM examples**
Fatty acyls	 5Z,8Z,11Z,14Z-eicosatetraenoic acid, arachidonic acid	Fatty Acids and Conjugates (unsaturated fatty acids, epoxy fatty acids) Octadecanoids (Jasmonic acids) Eicosanoids (lipids derived from arachidonic acid: prostaglandins, leukotrienes, thromboxanes) Docosanoids (similar to eicosanoids but derived from docosahexaenoic acid)	Major components of glycerolipids, glycerophospholipids and sphingolipids. Important source of cellular energy. Derivatives (eicosanoids/docosanoids) are important (generally pro-inflammatory) signaling molecules.	• Fatty acid α- and β-oxidation defects (numerous) • X-linked adrenoleukodystrophy (*ABCD1*) • Stargardt macular degeneration (ABCA4/ELOVL4) • CYP2U1 deficiency (*CYP2U1*)
Glycerolipids	 1,2-dihexadecanoyl-3-octadecanoyl-sn-glycerol, TG(16:0/16:0/18:0)	Monoradylglycerols (monoacylglycerols, monoalkylglycerols) Diradylglycerols (diacylglycerols, mono/di-alkylglycerols) Triradylglycerols (triacylglycerols, alkyldiacylglycerols)	Storage of energy, precursors for glycerophospholipids, second messengers in signal transduction.	• Chanarin-Dorman Syndrome, α/β-Hydrolase 5 (*ABHD5*) • Diacylglycerol O-acyltransferase (*DGAT1*) • Adipose triglyceride lipase (*PNPLA2*) • Sengers syndrome, acylglycerol kinase (*AGK*) • PHARC syndrome, α/β-Hydrolase 5 (*ABHD12*)
Glycerophospholipids	 1-hexadecanoyl-2-(9Z,12Z-octadecadienoyl)-sn-glycero-3-phosphocholine, PC(16:0/18:2ω6)	Glycerophosphocholines (PC, LPC, PC-plasmalogens) Glycerophosphoethanolamines (PE, LPE, PE-plasmalogens) Glycerophosphoserines (PS, LPS) Glycerophosphoglycerols (CL, PG, BMP, LPG) Glycerophosphoinositols (PI, LPI) Glycerophosphates (PA, LPA)	Major structural components of biomembranes, assist/take part in vesicle fusion/fission and signal transduction.	• Barth syndrome (*TAZ*) • MEGDEL syndrome (*SERAC1*) • Phosphatidate phosphatase (LPIN1) • Spondylometaphyseal dysplasia with cone-rod dystrophy, Choline-phosphate cytidylyltransferase A (PCYT1A)
Sphingolipids	 N-(octadecanoyl)-sphing-4-enine-1-phosphocholine, SM(d18:1/18:0)	Sphingoid bases (sphinganines, lysosphinganines) Ceramides (ceramides, dihydroceramides, phytoceramides) Phosphosphingolipids (ceramide phosphocholines (sphingomyelins), ceramide phosphoethanolamines) Neutral glycosphingolipids (globosides, other non sialic acid-containing oligoglycolipids) Acidic glycosphingolipids (gangliosides, sulfatides)	Like glycerolphospholipids, sphingolipids are major structural components of biomembranes, assist/take part in vesicle fusion/fission and signal transduction. Found in high	• Sphingolipidoses (including Gaucher (GBA), Fabry (*GLA*), Krabbe (*GALC*) etc. • Hereditary spastic paraplegia, glucosyl ceramidase (*GBA2*)
Sterol Lipids	 3α,7α,12α,24-Tetrahydroxy-5β-cholestan-26-oic acid, varanic acid	Sterols (cholesterol and derivatives, steryl esters) Steroids (estrogens, androgens, gluco/mineralocorticoids) Bile acid and derivatives (C_{24}-bile acids, C_{27}-bile acids) Steroid conjugates (glucoronides, sulfates) taurines	Components of biomembranes (sterols), solubilisation of fats (bile acid), signaling (hormones).	• Wolman/Cholesteryl ester storage disease (*LIPA*) • Smith Lemli Opitz syndrome (*DHCR7*) • Tangier disease (*ABCA1*) • Bile acid synthesis defects (numerous) • Defects in steroidogenesis (numerous)
Prenol Lipids	 3,7-dimethyl-9-(2,6,6-trimethylcyclohexen-1-yl)nona-2E,4E,6E,8E-tetraenoic acid, retinoic acid	Isoprenoids (terpenes, retinoids) Quinones and hydroquinones (ubiquinones, vitamin E, vitamin K) Polyprenols (phytoprenols, dolichols, bactoprenols)	Sterol lipid precursors, membrane anchors of proteins, part of electron transfer chain (ubiquinone), vitamins.	• Mevalonic aciduria/Hyper IgD syndrome (MVK) • Selection of congenital disorders of glycosylation involving dolichoylphosphate-mannose anchors.
Saccharolipids	 UDP-N-acetyl-α-D-glucosamine, UDP-GlcNAc	Acylaminosugars (mono-, di-, triacylaminosugars) Acylaminosugar glycans Acyltrehaloses	Lipids where the glycerol backbone has been substituted for a sugar. Mainly of bacterial/fungal origin. Example: lipid A component of the lipopolysaccharides in Gram-negative bacteria.	None at this time
Polyketides	 Doxycycline	Linear polyketides Polyenes Flavonoids (flavans, flavanols and leucoanthocyanidins, rotenoid flavonoids) Aromatic polyketides (naphthalenes and naphthoquinones, anthracenes and phenanthrenes)	Diverse family of natural products with biological activities and pharmacological activities mainly found in bacteria, fungi and plants. Many commonly used antimicrobial, antiparasitic, and anticancer agents are polyketides or polyketide derivatives. Examples: erythromycins and tetracylines.	None at this time

Table 7.1. Lipid categories according to LIPID MAPS*

*As published by (Fahy et al 2009) with examples of main classes and examples members of the subclasses in parentheses (adapted from (Brugger 2014;Li et al 2014)).

**This column, IEM examples, lists one or more examples of known inborn errors of metabolism (IEM) in the particular main class. The word "deficiency" has been omitted for clarity/space. The gene involved is shown in between parentheses

This section provides an introduction to the major methodological strategies currently used for lipidomics, and by no means attempts to provide a complete and in-depth overview. Instead, we frequently refer the reader to excellent reviews for further reading. We focus on lipidomic techniques that employ mass spectrometry and the pre-analytical/analytical phase but will also discuss the data (pre)-processing and analysis, which has become an important part of lipidomics workflow. The latter emphasizes that metabolomics in general is becoming a multidisciplinary field that requires input and knowledge from many specialists to succeed, including physicians, technicians, analytical chemists, bioinformaticians and clinical biochemists.

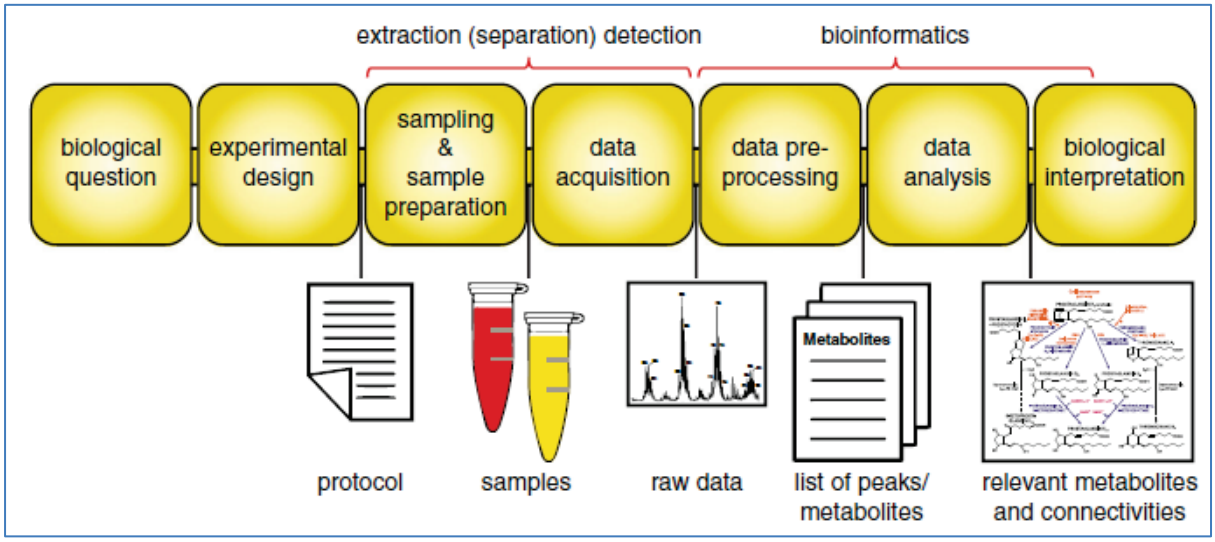


Figure 7.1. Lipidomics workflow. The different steps in the lipidomics experiment

7.2 Experimental design

The lipidomics experiment comprises of several steps which are depicted in Figure 7.1. After defining the biological question, the experiment is carefully formulated into a protocol which 1. takes into account the statistical considerations, 2. describes the collection of samples and 3. describes sample preparation and data acquisition (the actual measurement). In the post-analytical phase, bioinformatics approaches are used for data visualization, data pre-processing to translate the raw data into a list of detected and quantified peaks, (multivariate) statistical analysis for the comparison of samples, and various types of downstream analysis depending on the specific biological question.

Like any metabolomics experiment, a lipidomics experiment must be meticulously designed from beginning to end in order to avoid pitfalls and to ensure that the required information can be obtained from the data. Many factors before, during and after the actual analysis can greatly (and negatively!) influence the outcome of the experiment. Typically, experimental design involves decisions about the number of samples, group size (i.e., technical and biological replicates), and quality control samples (Box et al 2005). A main requirement for experimental design is its ability to account for unwanted effects (experimental bias) such as contamination of the chromatographic column during a study. Randomization and blocking of samples are two vehicles towards this requirement. In addition, choices concerning sample collection (timing, procedure, processing, storage), sample preparation (extraction, derivatization), and analytical method (method/reagents, internal standards, type of instrument) greatly influence the end result and require careful consideration. In the post-analytical

bioinformatics phase data pre-processing (peak detection and integration, isotope correction, normalization, compound identification) and (statistical) data analysis are key steps that need to be meticulously prepared and controlled.

Conducting a power analysis on metabolomics data is difficult, since in general, the effect size and variance of metabolites across samples are unknown a priori and depend on differences in metabolite concentrations between experimental groups (e.g. control versus diseased) as well as biological and technical variation of the samples within a group. Sometimes this information can be obtained from previous similar studies or from conducting a small pilot study before investing time and effort into a large scale study. For longitudinal studies or studies with a large number of samples, it is prudent to include a number of Quality Control (QC) samples in the study design, i.e. samples of a consistent composition which are included in the measurements repeatedly and which allow for the correction of run to run effects (Hendriks et al 2011). Depending on the study objective, other experimental design issues may need consideration. For example, to determine fluxes in metabolic pathways (e.g., flux balance analysis (FBA)) one should ensure that the measured metabolites sufficiently constrain the putative fluxes (Orth et al 2010).

7.3 Sample preparation/workup

Extraction of lipids from the sample is the first step towards their isolation, enrichment and concentration but also serves to remove salts and proteins, which interfere with separation techniques as liquid chromatography and subsequent mass spectrometric analysis. Liquid-liquid extraction is mostly used but solid phase extraction is also employed. Lipid liquid-liquid extraction usually makes use of a two-phase system (organic/aqueous) where lipids partition into the organic phase whereas water soluble molecules remain in the aqueous phase, and proteins are precipitated. Two established methods have been described long ago by Folch et al. (Folch et al 1957) and Bligh and Dyer (Bligh and Dyer 1959) and these procedures are still widely used, but alternatives have been developed and compared to these established standards (Reis et al 2013). For example, the use of methyl-tert-butyl ether is gaining in popularity (Matyash et al 2008; Abbott et al 2013) as it has a lower density than water (making it the upper layer in contrast to chloroform-based procedures) which makes it more suitable for high-throughput applications and automation. Given the high level of complexity and physicochemical diversity of lipid molecules it is as yet impossible to extract all lipid species using a single extraction procedure. This makes the measurement of the complete lipidome in a single analysis, which of course is the Holy Grail of lipidomics and metabolomics in general, impossible at this time. The choice for an extraction method therefore depends on the aim of the experiment; either target/optimize for a specific lipid category or try a general method that extracts as many lipids as possible. Yet, researchers have experimented with serial extractions to create fractions that partition the different lipid categories as efficiently as possible (Han et al 2012).

Another important aspect of the sample preparation and for the lipidomics experiment itself is the addition of stable isotope labeled compounds that serve as internal standards. One or more of these internal standards can be added for each class of lipids and be used to perform ratiometric comparisons with the molecular species of that lipid class to allow (semi)quantitative analysis of these analytes. In addition these stable isotope labeled molecules can be used for pulse/chase experiments to perform fluxomics studies (Mueller and Heinzle 2013).

Lastly, care should be taken during sample preparation and storage to protect lipids from being chemically or enzymatically modified as this obviously negatively influences the outcome of the

lipidomics experiment. Prior to extraction samples should preferably be stored at -80°C. Lipids in tissues and cells are relatively protected by natural antioxidant systems and compartmentalization. After sample homogenization, however, cellular content is mixed and unavoidably diluted, which renders lipids more prone to chemical or enzymatic modification. Some lipids are very stable (sterols, bile acids) while others are more prone to chemical oxidation (plasmalogens, lipids containing polyunsaturated fatty acids) or oxidation by light (7-dehydrocholesterol, ergosterol) (Wolf and Quinn 2008). Enzymatic modification can be prevented by working at temperatures close to 0°C and adding a small percentage of organic solvent to the homogenization buffer. After extraction, lipids are much more prone to chemical oxidation and extracts therefore should preferably be stored in glass vials, solubilized in sufficient organic solvent at -80°C and care should be taken to eliminate air/oxygen by flushing with inert gasses and storing in full containers.

Box 1. Mass Spectroscopy lexicon

Amu	Stands for atomic mass unit, which is the same as Dalton (Da).One (unified) atomic mass unit is defined as one twelfth of the mass of an unbound neutral atom of ^{12}C in its nuclear and electronic ground state and is equivalent to 1 g/mol.
Acquisition rate	Amount of scans that can be performed per second for a certain mass range.
Accurate mass	The measured exact mass.
Data-dependent scanning	Automated real-time selection of ions for MS^n analysis.
Exact mass	The calculated mass of anion based on the sum of the (monoisotopic) masses of each atom in the molecule.
Ion suppression/enhancement	The negative (or positive) effect on the ionization of the molecules of interest by the presence of other endogenous or exogenous molecules.
Linear dynamic range	This is the range over which ion signal is linear with the analyte concentration m/z mass (m) to charge (z) ratio, i.e., the mass of the ionized molecule divided by its charge.
Mass range	The range of m/z that can be covered by a certain mass spectrometer
Mass accuracy	The difference between the exact mass and the measured mass divided by the exact mass (e.g., ¦(exact mass – accurate mass) ¦/exact mass). In other words, how close is the measured mass to the exact mass. Mass accuracy is usually measured in parts per million (ppm) or amu/Da.
MS^n	Mass spectrometry where after fragmentation, fragments are either analyzed of fragmented again n-times. Tandem mass spectrometry as performed by QqQ machines is termed MS^2.
Resolution	Resolution is defined as the m/Δm, where m is the mass of the peak and Δm is the peak width at half height and is a measure of the ability to distinguish two peaks of slightly different mass-to-charge ratios, in a mass spectrum. The higher the value the better the resolution. Also sometimes indicated as “full width at half maximum” = FWHM.
Sensitivity	Amount of moles (usually femtomol/attomol range) that can be detected by the instrument.

7.4 Data acquisition

Common techniques to separate lipids before MS analysis include high/ultra-high performance liquid chromatography (HPLC/UHPLC) and capillary electrophoresis. Alternatively, direct infusion of the lipid extract into the mass spectrometer (e.g. no prior chromatographic separation), so called “shotgun lipidomics”, is also frequently used.

For convenience, several technical terms in relation to mass spectrometry are listed in Box 1. To appreciate the different approaches in lipidomics one must first understand the ionization process and the types of mass spectrometers and techniques that are used to separate/filter ions. Basically, there are three types of mass filters namely Quadrupoles (Q), Time-of-flight (TOF) and Ion traps. By using combinations of these mass filters, different mass spectrometers can be constructed which are called hybrids. Frequently used types of machines include triple quadrupole instruments (QqQ), Quadrupole time-of-flight (QTOF) and ion traps.

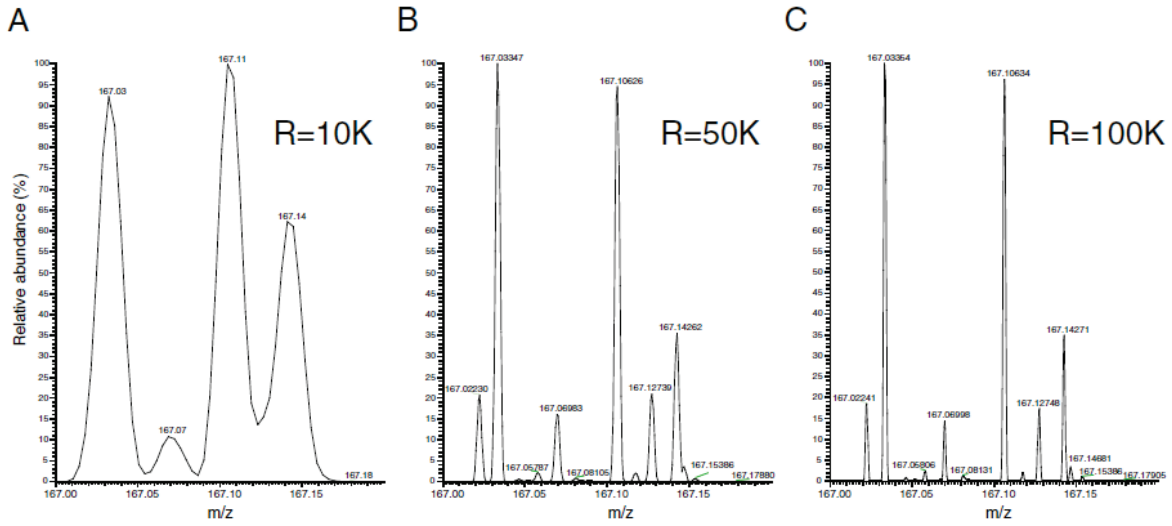


Figure 7.2: Mass resolution. One sample measured using three different resolutions; 10.000 (10K), 50.000 (50K) and 100.000 (100K) on a Thermo Q-Exactive. Increasing resolution enables the detection of more ions at higher mass accuracy which makes identification of the corresponding compounds more dependable.

7.4.1 Triple quadrupole instruments (low resolution)

Triple quadrupole instruments (QqQ) are commonly used in the targeted quantification of metabolites in biological samples. The two quadrupoles (Q) are separated by a collision cell (q) which is used to fragment ions originating form the first quadrupole, making it possible to perform different types of scans (Han et al 2012). Typically, a combination of constant neutral loss scans, precursor ion scans and product ion spectra is used to detect metabolites in biological matrices (Liu 2012). The limited resolution of the quadrupole (up to 7500, FWHM) is not sufficient for metabolite identification based on the measured mass in contrast to high resolution instruments as will be dealt with next. At higher acquisition rates and when scanning a wide mass range, sensitivity is considerably reduced, making QqQ instruments less suitable for biomarker discovery. QqQ instruments are, however, the best option for sensitive targeted quantitative analysis (especially when combined with UHPLC) of a limited amount of metabolites using their tandem MS capabilities in multiple reaction monitoring (MRM).

7.4.2 Time-of-flight (TOF), ion traps and hybrids (high resolution)

TOF mass spectrometers combine high-resolution with mass accuracy, which increases the possibility of determining elemental compositions of molecules and also provides high specificity of detection. By combining TOF's with quadrupoles, hybrid machines have been developed allowing novel scan modes together with high resolution capabilities. Machines as QqTOFs are, just like QqQ instruments, capable of selecting and fragmenting ions that can be subsequently separated by the TOF and detected at high resolution. The resolution (up to 70.000, FMWH) is sufficient for metabolite identification and TOF instruments can cope with a high acquisition rate which combines well with fast chromatographical systems as UHPLC. Resolution is an important parameter of mass spectrometry and is explained in more detail in Figure 7.2.

Ion traps, Orbitraps in particular, have developed into mass filters with very high resolution (up to 450.000 at m/z 200) making it possible to determine the exact mass of the compounds but also to resolve isobaric species and reveal isotope fine structures. Another advantage of some ion trap machines is that these instruments have MSn capability, meaning that ion fragments can be further fragmented and characterized, which is particularly useful for structural elucidation of molecules. When compared to TOF instruments, ion traps have a relatively slow acquisition rate, especially at high resolution and resolution declines at higher m/z values. Ion traps are therefore less suitable for fast chromatography. By interfacing Orbitraps with linear ion traps or quadrupole mass filters many of these limitations have partially been overcome.

An interesting and useful feature that is especially useful when using hybrid machines is the capability to perform so called data-dependent scanning. This technique uses specific criteria to select one or more ions of interest for subsequent fragmentation, meaning that a product ion scan is performed for these selected ions providing more structural information of the compound. Scanning in the data-dependent mode starts with a survey scan to identify ions and their abundances in the sample. This survey scan is then followed by the acquisition of a fragment spectrum from the automatically selected precursors. In this way, product ion scans are obtained (or in some ion traps even MSn spectra) that can aid in the identification of the precursor ions. Although this type of scanning is biased towards the more abundant ions it provides valuable structural information of the ions (Bhattacharya 2013).

7.4.3 Ionization techniques (ESI, APCI and MALDI)

Before any molecule can be analyzed by the mass spectrometer it first has to be ionized so that it can be manipulated using electric fields and mass filters and detected by the detector. For lipidomics, three ionization techniques are most used, 1. Electrospray ionization (ESI), 2. Atmospheric pressure chemical ionization (APCI) and 3. Matrix-Assisted Laser Desorption/Ionization (MALDI). There are many other ways to ionize analytes (Li et al 2014) but these will not be discussed here. The three ionization techniques mentioned above are all so called "soft ionization" techniques as very little fragmentation occurs during the ionization process and mainly monocharged molecular ions, which is ideally suited for metabolomics purposes.

ESI is most widely used for thermally labile and mostly non-volatile molecules and therefore can be applied to almost all lipid categories. The sample is nebulized through a highly charged capillary using heated nitrogen gas, producing a fine aerosol. This results in evaporation of the solvent and ionization of the molecules after which the ions enter the mass spectrometer.

With APCI, the sample is nebulized and heated so that both solvent and analytes are in the gas phase followed by a corona discharge which ionizes the solvent molecules that subsequently also ionize the

analyte molecules. APCI generally yields monocharged ions and is mainly used with small thermally stable nonpolar molecules (<1500 Da). For lipidomics, APCI is mostly applied for neutral lipids including triglycerides, sterols and fatty acid esters (Byrdwell 2001; Li et al 2014).
For MALDI, the sample is mixed with a matrix that readily forms crystals that aids the ionization process. The fluid mixture of sample and matrix is spotted on a MALDI plate and allowed to dry. To ionize the analyte molecules, a laser is fired at the matrix crystals in the dried-droplet spot, which absorbs the laser energy resulting in desorption and ionization. The ionized matrix molecule then transfers its charge to the analyte, thus ionizing the analyte.

7.4.4 Direct infusion MS, chromatography- and MALDI-based MS analysis

For the data acquisition, essentially two strategies can be chosen, MS preceded by a chromatographic separation (usually LC-MS) or direct infusion, which is also called "shotgun MS". Both approaches have their advantages and limitations.
In shotgun lipidomics, the lipid extract is introduced directly into the MS and acquisition can proceed for prolonged periods of time to acquire multiple spectra with good signal to noise ratios. Especially when combined with nanomate devices, reproducible direct infusion into the mass spectrometer can be accomplished. As there is no prior separation, analysis time is relatively fast and when using suitable internal standards, the shotgun approach is surprisingly reproducible, even in complex matrices (Jung et al 2011). Multiple analyses are done on multiplex extractions of the same sample using different types of scans on QqQ machines (neutral loss scan, precursor ion scan and product ion scans) and this information combined is so called multi-dimensional MS (MDMS) which allows identification and quantification of individual lipid species (Han et al 2012). As sensitivity, mass resolution and accuracy, acquisition rate and dynamic range of mass spectrometers have all improved rapidly, shotgun MS is becoming a considerably attractive method of choice for lipidomics. Despite these advances, however, shotgun lipidomics is hampered by ion-suppression/enhancement effects, the inability to distinguish certain isobars/isomers, unreliable quantification of low abundant lipids and difficulty in identifying unknown lipids (Hyotylainen and Oresic 2014). ESI-QTOF instruments are frequently used for shotgun lipidomics as they combine the ability to perform fragmentation with accurate mass measurement. Like hybrid instruments containing ion traps, however, these instruments are limited in their capabilities to perform tandem MS scans typically used in QqQ instruments. As they capture all precursors and their fragments in parallel and in a single scan, the analyses of the data pose a great challenge to relate the fragments to their precursors (Bhattacharya 2013).
LC-MS, which is the most frequently used hyphenated mass spectrometric technique, has the advantage that separation of the different lipid classes/species lowers the complexity of the sample, reduces ion suppression/enhancement effects and therefore allows a more specific identification of lipid species. Another advantage is that isobaric/isomeric lipids sometimes can be separated and quantified, which is not possible using shotgun MS. For example, bismonoacylglycerolphosphates and phosphatidylglycerols are isobaric/isomeric molecules that cannot be distinguished using MS and have to be separated by chromatography to be detected and quantified separately. Another advantage of chromatographical separation is that some lipids can undergo in source fragmentation that "synthesizes" other lipids. For instance, phosphatidylserine can lose its headgroup and form phosphatidic acid which can be recognized when phosphatidylserine-derived phosphatidic acid is separated from endogenous phosphatidic acid by liquid chromatography prior to MS detection

(Knittelfelder et al 2014). When using LC-MS, however, the acquisition rate needs to be sufficient to fully designate an eluting peak with enough sensitivity to allow reproducible quantification and identification. This can be problematic when using ion traps as the technical setup of these instruments is such that at higher resolutions acquisition rates are significantly reduced. And although ion suppression/enhancement is considerably less when using LC, the different elution time of internal standards and analytes calculated using those internal standard can introduce variation because of variable amounts of ion suppression during the chromatographical run. The LC step also necessitates extra data pre-processing efforts as retention time correction and correct peak grouping needs to be implemented in the bioinformatics pipeline.

MALDI-based MS analysis in lipidomics is mostly used to directly detect lipids from surfaces including tissue sections and TLC plates, but spotted lipid extracts can also be analyzed (Berry et al 2011; Ellis et al 2013). The relatively simple sample preparation and the fact that considerable amounts of impurities are tolerated make MALDI-based MS useful for the analysis of a large number of samples. The MALDI ionization, however, renders it difficult to hyphenate with other techniques, UHPLC in particular, and metabolite quantification is not a strong suit of MALDI-based MS (Fuchs et al 2010). Despite these disadvantages, MALDI-based MS is being used more and more in lipidomics because of the capability to scan tissue slices and provide a spatial distribution map of lipids within a sample, a technique called mass spectrometry imaging or MSI (Berry et al 2011). To be able to visualize the location of different lipids in tissue slices, and even at the cellular level, makes MSI a valuable addition to the techniques to characterize the lipidome.

7.5 Bioinformatics data pre-processing

Metabolomics experiments generate large amounts of data, which can be processed by various bioinformatics methods to detect and quantify metabolite peaks, assign compound names, and perform further downstream analysis such as statistical, biological pathway, or metabolic flux analysis. Pre-processing is the first part of this bioinformatics workflow and comprises the application of methods needed to generate a peak table for each sample which contains all detected peaks (positions) and their relative concentrations (peak intensities and/or areas). Subsequently, to facilitate comparative analysis, matching peaks representing the same metabolite from different samples are grouped together in a peak group list (Boccard et al 2010). Sometimes, chromatogram alignment is required to adequately group matching peaks (see below). In essence, the challenge of data pre-processing is to collect as many true metabolite signals from the data as possible, while at the same time minimizing the number of detected artefact peaks (e.g. noise, spikes). Highly abundant metabolites usually yield strong, unmistakable signals, but as the metabolite concentrations reach the detection limit, it becomes harder to automatically identify and quantify the peaks. In practice time-consuming curation and improvement of the peak table by mass spectrometry experts still remains necessary. Generally, a few rounds of pre-processing and curation are required to establish a final peak table. **Bioinformatics data pre-processing will be discussed in the next chapter of the syllabus**.

7.6 Conclusion

The technical advances in mass spectrometry, particularly the development of (ultra)-high-resolution/mass accuracy measurement capabilities in combination with refinement of soft ionization techniques, have increased the application and success of lipidomics to answer biological questions in relation to lipid metabolism. Together with other omics technologies, lipidomics has become an important tool to practice systems biology as lipids comprise a very significant part of the metabolome and play pleiotropic roles in cellular functions. As an increasing number of disorders are linked to lipid metabolism, lipidomics is used to search for biomarkers, understand disease mechanism and follow the efficacy of therapeutic options. Still, with a plethora of different techniques, each with their own strengths and weaknesses, it is clear that no single platform is sufficient to fully characterize the lipidome completely. Combining different lipidomics techniques, such as mass spectrometry imaging, shotgun/LC lipidomics, global profiling using LC-high resolution MS and quantitative targeted lipid analyses with LC-QqQ MS and GC-MS has the potential to yield the most complete lipidomics data sets (Brugger 2014; Hyotylainen and Oresic 2014). Unfortunately, only a few laboratories have the technical, financial and analytical capabilities to achieve this. Many examples in literature, however, have shown that even a single platform can yield insightful results, which is encouraging more researchers to setup lipidomics platforms for both research and diagnostics.

There is a great need for standardization and consolidation in the field of lipidomics, in terms of sample preparation, data collection methodology, and data management and analysis. The exchange of data and results between different laboratories through publicly accessible repositories (such as Metabolights (Salek et al 2013)), will allow the community to build a knowledge base to map the entire lipidome. The pre-processing of metabolomics data remains a major bottleneck and continues to require the time-consuming step of manual validation. Clearly, more bioinformatics research is required to develop improved algorithms that eventually allow to pre-process data with no or only minimal human intervention. However, this will require the large-scale availability of public metabolomics data. Another bottleneck is the assignment of compound names to identified peaks in the raw data. A solution to this problem should be a joint effort of experimental and bioinformatics approaches. A final challenge is the further development of statistical, bioinformatics and systems biology methods to enable biological interpretation. In particular, efforts towards the integration and interpretation of metabolomics data with genome data (e.g., SNPs), gene expression data, and proteomics data are expected to become more common and will enable new insights in living systems and the identification of biomarkers.

It is important to realize that in each phase of the metabolomics experiment different specialists play a dedicated role. When researching human disorders, physicians ensure that patient selection is accurate and together with clinical biochemists see to it that no pre-analytical trivialities occur. Technicians and analytical chemists/clinical biochemists need to develop a reproducible and robust method and bioinformaticians are responsible to assist in the statistical design of the experiment and provide reliable and validated data pre-processing and analysis methods to facilitate correct biological interpretation of the results. This requires the presence of dual-thinkers on the lipidomics team and emphasizes that research and development is based on constructive cooperation between different disciplines to obtain the best results.

7.7 References

- Abbott SK, Jenner AM, Mitchell TW, Brown SH, Halliday GM, Garner B (2013) An improved high-throughput lipid extraction method for the analysis of human brain lipids. Lipids 48: 307-318.
- Berry KA, Hankin JA, Barkley RM, Spraggins JM, Caprioli RM, Murphy RC (2011) MALDI imaging of lipid biochemistry in tissues by mass spectrometry. Chem Rev 111: 6491-6512.
- Bhattacharya SK (2013) Recent advances in shotgun lipidomics and their implication for vision research and ophthalmology. Curr Eye Res 38: 417-427.
- Bligh EG, Dyer WJ (1959) A rapid method of total lipid extraction and purification. Can J Biochem Physiol 37: 911-917.
- Box GEP, Hunter JS, Hunter WG (2005) Statistics for Experimenters: Design, Innovation, and Discovery: Wiley.
- Brugger B (2014) Lipidomics: analysis of the lipid composition of cells and subcellular organelles by electrospray ionization mass spectrometry. Annu Rev Biochem 83: 79-98.
- Byrdwell WC (2001) Atmospheric pressure chemical ionization mass spectrometry for analysis of lipids. Lipids 36: 327-346.
- Ellis SR, Brown SH, In Het Panhuis M, Blanksby SJ, Mitchell TW (2013) Surface analysis of lipids by mass spectrometry: more than just imaging. Prog Lipid Res 52: 329-353.
- Fahy E, Subramaniam S, Murphy RC, et al (2009) Update of the LIPID MAPS comprehensive classification system for lipids. J Lipid Res 50 Suppl: S9-14.
- Folch J, Lees M, Sloane Stanley GH (1957) A simple method for the isolation and purification of total lipides from animal tissues. J Biol Chem 226: 497-509.
- Fuchs B, Suss R, Schiller J (2010) An update of MALDI-TOF mass spectrometry in lipid research. Prog Lipid Res 49: 450-475.
- Han X, Yang K, Gross RW (2012) Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses. Mass Spectrom Rev 31: 134-178.
- Hyotylainen T, Oresic M (2014) Systems biology strategies to study lipidomes in health and disease. Prog Lipid Res 55C: 43-60.
- Jung HR, Sylvanne T, Koistinen KM, Tarasov K, Kauhanen D, Ekroos K (2011) High throughput quantitative molecular lipidomics. Biochim Biophys Acta 1811: 925-934.
- Knittelfelder OL, Weberhofer BP, Eichmann TO, Kohlwein SD, Rechberger GN (2014) A versatile ultra-high performance LC-MS method for lipid profiling. J Chromatogr B Analyt Technol Biomed Life Sci 951-952: 119-128.
- Li M, Yang L, Bai Y, Liu H (2014) Analytical methods in lipidomics and their applications. Anal Chem 86: 161-175.
- Liu ZY (2012) An introduction to hybrid ion trap/time-of-flight mass spectrometry coupled with liquid chromatography applied to drug metabolism studies. J Mass Spectrom 47: 1627-1642.
- Matyash V, Liebisch G, Kurzchalia TV, Shevchenko A, Schwudke D (2008) Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. J Lipid Res 49: 1137-1146.
- Mueller D, Heinzle E (2013) Stable isotope-assisted metabolomics to detect metabolic flux changes in mammalian cell cultures. Curr Opin Biotechnol 24: 54-59.
- Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nat Biotechnol 28: 245-248.
- Reis A, Rudnitskaya A, Blackburn GJ, Mohd Fauzi N, Pitt AR, Spickett CM (2013) A comparison of five lipid extraction solvent systems for lipidomic studies of human LDL. J Lipid Res 54: 1812-1824.
- Salek RM, Haug K, Steinbeck C (2013) Dissemination of metabolomics results: role of MetaboLights and COSMOS. Gigascience 2: 8.
- Wolf C, Quinn PJ (2008) Lipidomics: practical aspects and applications. Prog Lipid Res 47: 15-36.

8 From experiment to data

Lecturer: Prof. dr. Antoine van Kampen (AMC)

After reading this chapter you should

- Have basic knowledge of data acquisition
- Understand the concepts of signal, noise, and signal to noise ratio.
- Understand the concept of moving averaging to reduce noise

Contents

8.1 Introduction

In these chapter you will learn:

- How a biological measurement results in data points in an (excel) file
- What is measurement noise, and how to reduce this once you have the data
- The principle of using a moving window for data analysis. Here we use moving windows for noise reduction, but you also encounter them in the analysis of NGS data, gene finding, etc).

In OMICS experiments you almost always end up with one or more (excel) files that contain the results of the experiments: **data**. This data, for example, represent gene expression levels measured with microarrays, or protein/metabolite concentrations measured with mass spectroscopy. The files containing your results are the starting point for data analysis.

But where do these numbers come from? In other words, how is the biological signal (in your test tube) translated to a data value in a computer document? If you understand this, then you will also be able to understand that noise is almost always introduced in the data and that this can negatively affect your data.

Methods for noise reduction are, generally, applied in the background by the hardware of the measurement device or by the associated software. The process of noise reduction may therefore be invisible for you. In other cases you will need to explicitly apply noise reduction methods (for example, in the pre-processing of metabolomics data as you will see in another computer practicum).

We will briefly explain where signal and noise are coming from and show some basic approaches to remove noise from measured data.

8.2 Signals and noise

A **detector** (also called sensor) is a converter that measures a physical quantity and converts this quantity into a signal which can be read by an electronic instrument.

For example, a thermocouple (Figure 8.1) measures temperature and converts this to an output voltage which can be read by a voltmeter. A thermocouple consists of two dissimilar conductors in contact, which produces a voltage when heated. The size of the voltage is dependent on the difference of temperature of the junction to other parts of the circuit.

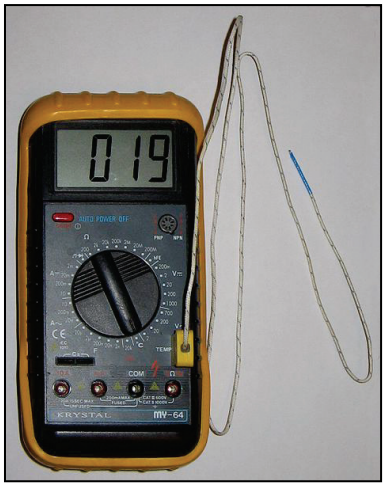


Figure 8.1. Thermocouple connected to a multimeter displaying room temperature. (Figure copied from http://en.wikipedia.org/wiki/Thermocouple).

In the thermocouple example, the electronic instrument is a multimeter. However, it can also be a computer. Figure 8.2 shows the general principle of measurements made by a detector. A signal is detected by a detector that, subsequently, generates an electrical signal. This 'analog' signal is digitized by an analog-to-digital convertor. AD conversion only converts the amplitude of the signal to a number that can be read in and stored by a computer.

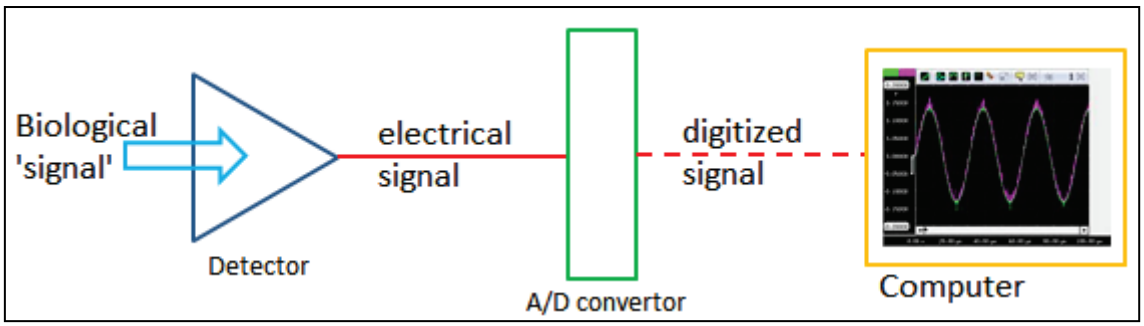


Figure 8.2. General principle of signal detection. A detector is used to measure a biological signal (e.g., metabolites) and converts this into an electrical signal (e.g. depending on concentration). This analog signal is digitized by an analog-to-digital (AD) convertor and stored in a computer (your data). At various stages in this chain, unwanted noise can be introduced.

Figure 8.3 shows a more concrete example for **mass spectroscopy used in proteomics and metabolomics**. A sample is vaporized and ionized. Ions are then accelerated and collide with a detector. When an ion hits the metal box, its charge is neutralized by an electron jumping from the metal on to the ion (right hand diagram). That leaves a space amongst the electrons in the metal, and the electrons in the wire shuffle along to fill it. A flow of electrons in the wire is detected as an electric current which can be amplified and recorded by a computer after the AD conversion. The more ions arrive, the greater the current. Note that in time-of-flight mass spectrometry the ion's mass-to-charge ratio is determined via a time measurement.

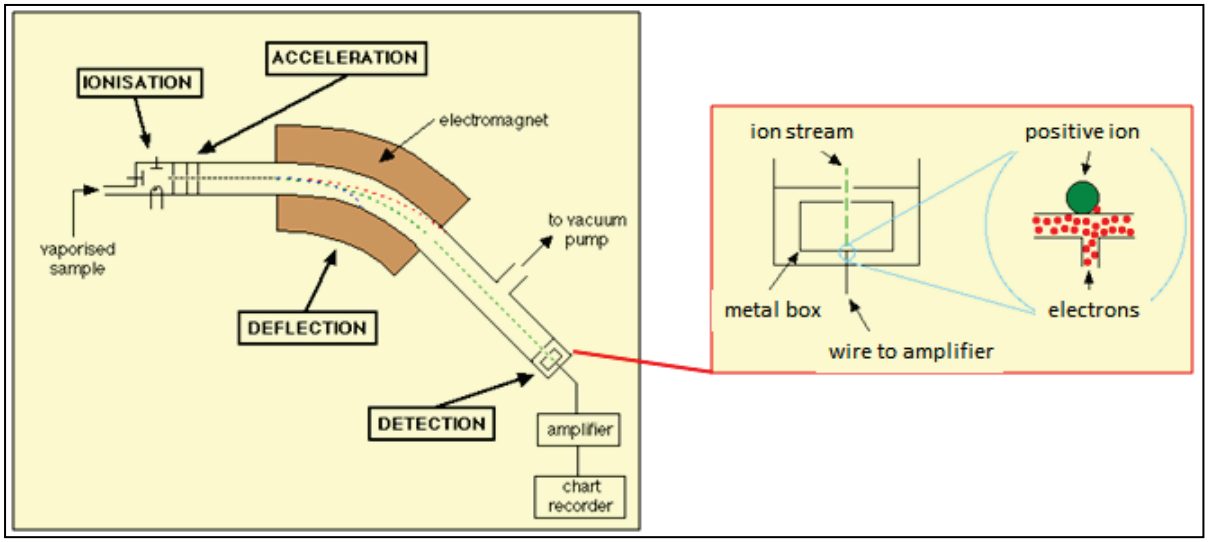


Figure 8.3. Detection of ions in a mass spectrometer. Note the red electron that has jumped to the green ion. (Copied from http://www.chemguide.co.uk/analysis/masspec/howitworks.html)

Given this general setup, every analytical measurement is made up of two components:

- **Signal**. Carries information about the analyte that is of interest to the experimenter (e.g., number of ions in mass spectroscopy);
- **Noise**. Is made up of extraneous information that is unwanted because it degrades the precision of an analysis and also places a lower limit on the amount of analyte that can be detected. See Figure 8.4.

As an example, consider an mp3 with your favorite music (the signal). Now listen to acoustic noise (http://www.youtube.com/watch?v=3sBIxqsCAE0) that you don't want to be present together with your music.

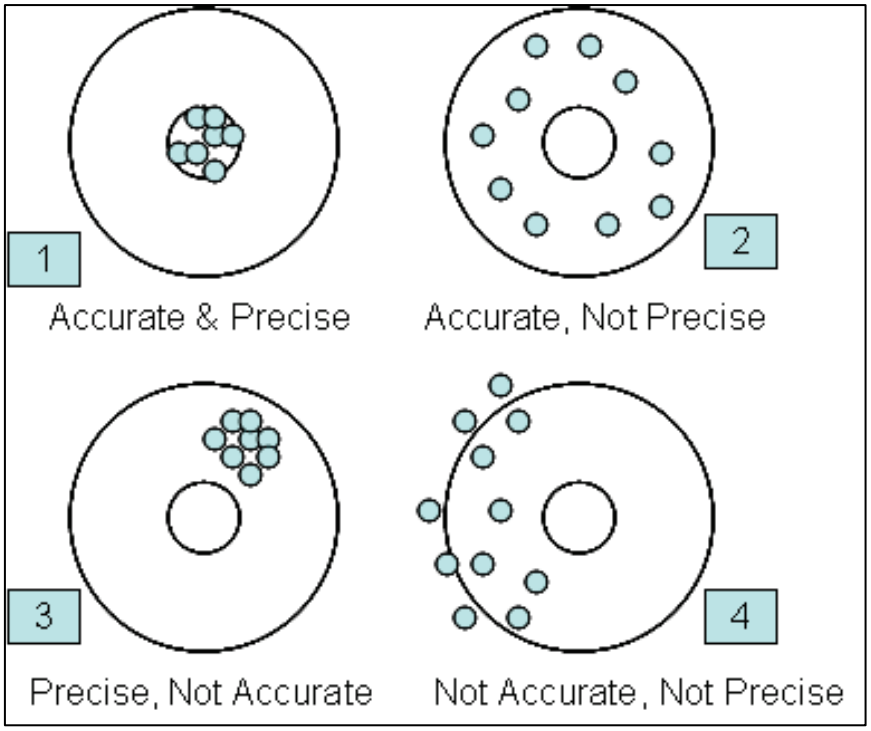


Figure 8.4. Precision and accuracy. Noise makes measurements less precise. Smoothing (see below) may make the measurements less accurate (i.e. introduces a measurement bias).

8.2.1 Sources of noise

It is important for the analyst who uses a particular instrumental method to be aware of the sources of noise because noise determines both the accuracy and detection limits of any measurement. Noise enters a measurement system from environmental sources external to the measurement system (Figure 8.5), or it appears as a result of fundamental, intrinsic properties of the (electronic) system. It is usually possible to identify the sources of environmental noise and to either reduce or avoid their effects on the measurement by, for example, moving the measurement device away from an AM radio or power line. Fundamental noise cannot be removed and therefore ultimately limits accuracy, precision and detection limits in every measurement. It is beyond the scope of this text to discuss the various sources of noise in detail, but question 2 will show the effect of noise on the appearance of signal.

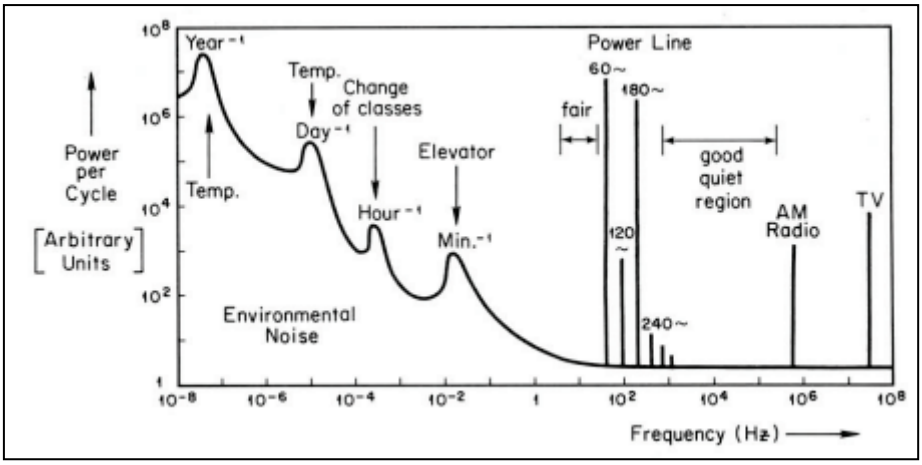


Figure 8.5. Sources of environmental noise. Noise at lower frequencies is also called drift. For example, the temperature that changes during the year may effect measurements. (Figure copied from Thomas Coor (1968) Signal to noise optimization in chemistry Part one, J. Chem. Educ., 1968, 45 (7), p A533

8.2.2 The signal to noise ratio (S/N):

The signal to noise ratio is a representative marker that is used in describing the quality of an analytical method of an instrument. In most measurements, the average strength (standard deviation) of the noise (N) is constant and independent of the magnitude of the signal (S). Thus, the effect of noise on the relative error of a measurement becomes greater and greater as the analytical quantity being measured decreases in magnitude. For this reason the signal to noise ratio is useful as a figure of merit for describing the quality of an analytical method.

The signal to noise ratio (SNR) can be defined as

$$S/N = \frac{\mu_S}{\sigma_N}$$

where μ_S is the mean signal and σ_N is the standard deviation of the noise.

Figure 8.6 shows a chromatographic peak (e.g., for a specific mass).

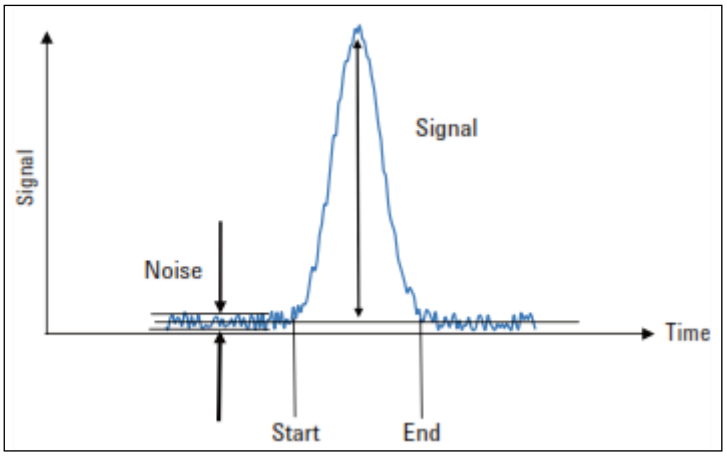


Figure 8.6. Chromatographic peak for a specific mass (m/z value). The start and end of the peak are indicated. The signal (S) is determined as the magnitude of the peak above the average noise level. Note that this signal (S) may also be calculated as the average value for a few points surrounding the position of the maximum value. The noise level (S) is determined by taking several measurement values before the peak start (the baseline, which is supposed to have zero signal) and, subsequently, determining the standard deviation of this signal. (Figure copied from Agilent techical report http://www.chem.agilent.com/Library/technicaloverviews/Public/5990-8341EN.pdf)

8.3 Signal to noise improvement

Many laboratory measurements require only minimal effort to maintain the signal to noise ratio at an acceptable level. Both hardware (e.g., shields and low-pass filters) and software methods are available for improving the signal to noise ratio of an instrumental method.

Here we will only consider "software filters", which are computer algorithms that remove or attenuate noise from the measured signal (data). There are several algorithms for removing noise. We will consider only one class of methods ('**smoothing**') to illustrate the principles and because this is used by xcms for the pre-processing of metabolomics data.

Smoothing methods for reducing noise include the following approaches:

- Moving average (also called boxcar averaging)
- Savitzky-Golay algorithm
- Moving Gaussian
- Moving negative second derivative of Gaussian

These methods can be used if the signal varies slowly with time, and assume that the average of a small number of adjacent points to express the signal is better (less noisy) than any single individual point. A limitation of these methods is that they may distort the signal.

8.4 What is smoothing?

Noise may affect the height of a measured (Gaussian) peak. Next we will see how we can remove this noise with 'smoothing' approaches.

In smoothing, the data points of a signal are modified so that individual points that are higher than the immediately adjacent points (presumably because of noise) are reduced, and points that are lower than the adjacent points are increased (Figure 8.7).This naturally leads to a smoother signal. As long as the true underlying signal is smooth (does only vary slowly with time), then the true signal will not be much distorted by smoothing, but only the noise will be reduced. We will first discuss moving averaging (smoothing).

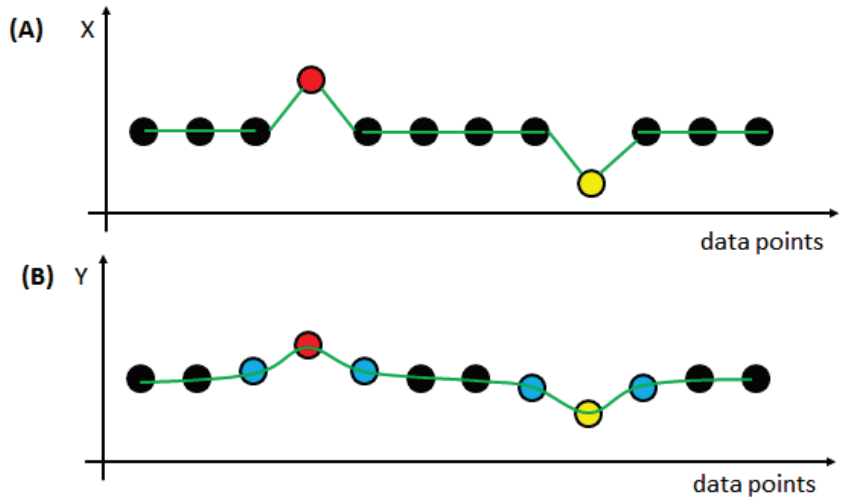


Figure 8.7. Principle of smoothing. Simplified example to demonstrate the principle of smoothing. The dots represent a selection of measured data points. (A) Here we see the original (raw) data (X) where the red point has a larger value than its neighboring black dots, and the yellow point has a smaller value than its neighbors. (B) Here we see the smoothed data (Y). The value of the red dot is decreased towards the value of its neighbors while at the same time its neighboring values are increased towards the red dots. For the yellow data point we see a similar effect. The result is a smoother green curve in (B) compared to (A). Thus, if the red and yellow dots represented noise fluctuations, then these have now been diminished.

8.4.1 Moving average algorithm

The principle of smoothing shown in Figure 8.7 is nothing else than taking the average value of neighboring points. Thus, the algorithm is very simple!

To smooth every point in a dataset you have to take the average values for every point and its neighbors. Thus, you start at the beginning and then move towards the last point in de dataset. This is what we call a 'moving average' (Figure 8.8).

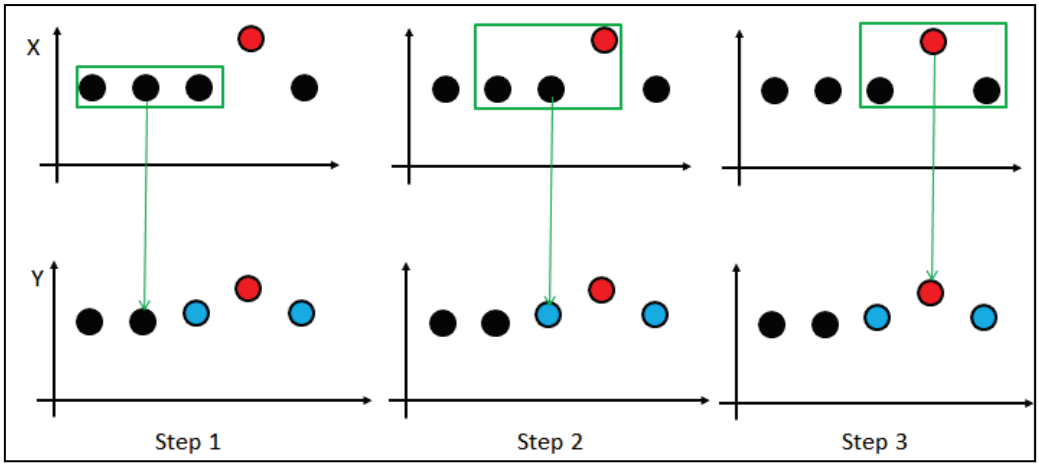


Figure 8.8. Principle of a moving average for the first few points of a data set. In the first step we take the average value of the first three points (green box). The average (smoothed) value (Y) is plotted for the center point in the lower panel. Next we move the box one point to the right and calculate the average value for the next three points. Again this average is plotted for the center point in the lower panel (blue point). In step 3 we move the box again on step to the right and calculate the average for the red point. We continue this procedure for the whole dataset. Thus, we only calculate averages and moving the box; hence the name 'moving average'. Note that we could not calculate an average for the first (and last point) of the dataset. The width of the green box is called window size. Here the window size = 3.

The most straightforward approach for smoothing signals consists of taking equidistant points and performing a moving average as we have already seen in Figure 8.8. Suppose we have made a measurement (chromatogram) then these measurements form a vector of raw noisy data X=$[x_1, x_2, ...x_n]$. The measurement values *x* can be converted to a new vector Y of smoothed data (thus measurements with reduced noise).

To calculate a smoothed point we just calculate the averages:

$$y[3] = \frac{x[1]+x[2]+x[3]+x[4]+x[5]}{5}$$

Thus, here measurement point number 3 is replaced by a smoothed point (*y[3]*) by taking the average of five measurement points.

Note:

- in contrast to Figure 8.8 we now have five points in the box.
- We start at y[3] and no average can be calculate for y[1] and y[2] because we cannot move the box further to the left.
- The size of the box (3 in the figure, and 5 in formula) is called the **window size**. The size of the window is always an odd number and is chosen by the user of the algorithm.

Next, we 'move' one point to the right and calculate the average for point 4 (note that the indices change):

$$y[4] = \frac{x[2] + x[3] + x[4] + x[5] + x[6]}{5}$$

We continue this procedure until we are at the end of the measurement data. Thus, if we have measured 100 points then the last average is calculated as:

$$y[98] = \frac{x[96] + x[97] + x[98] + x[99] + x[100]}{5}$$

Thus, in general, we smooth a data point *i* by calculating the following average:

$$y[i] = \frac{x[i-2] + x[i-1] + x[i] + x[i+1] + x[i+2]}{5}$$

$$y[i] = \frac{1}{5} \sum_{j=-2}^{+2} x[i+j]$$

To see what is happing for a real Gaussian signal (chromatographic peak), have a look at Figure 8.9 and Figure 8.10. The measured signal (100 points) is shown as a black line. This chromatographic peak clearly contains noise (random fluctuations), which need to be removed because otherwise the peak position and peak area cannot be determined with high precision. To remove the noise, a moving average of five points is applied, which starts a point *x=3* and ends at point *x=98*. Note, we cannot calculate the average for point 1 and 2 because for these two data points we don't have two additional points preceding these measurements. Similarly, we cannot calculate the moving average for point 99 and point 100. Thus we 'loose' four points of our measurements. This is generally not a problem since the start and end points of a measured signal are generally of no interest. The moving average is represented by the blue box. This box represents five measured data points from which the average is calculated. The calculated average is then shown as a red dot. By moving the box one point to the right, we can calculate the next average (red point). We see that much of the noise is has now been removed from the data in the red smoothed curve.

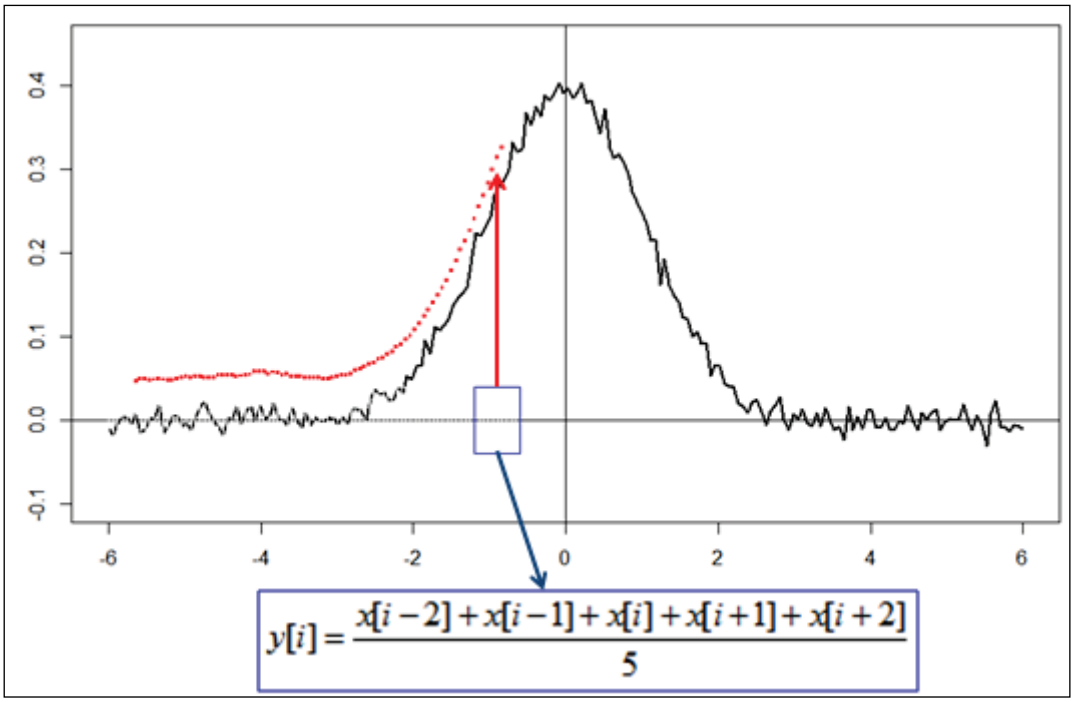


Figure 8.9. Graphical representation of 100 measured data points (chromatogram; black line). The signal (chromatographic peak) is clearly visible but we also observe a certain amount of noise (random fluctuations). The smoothed signal is shown by the red line. One iteration of the moving average is shown for point *i*. The average of these five points is depicted as a red point. Note that to enhance visualization, the y-values of the red curve are shifted upwards (Figure 8.10 shows final result of smoothing)

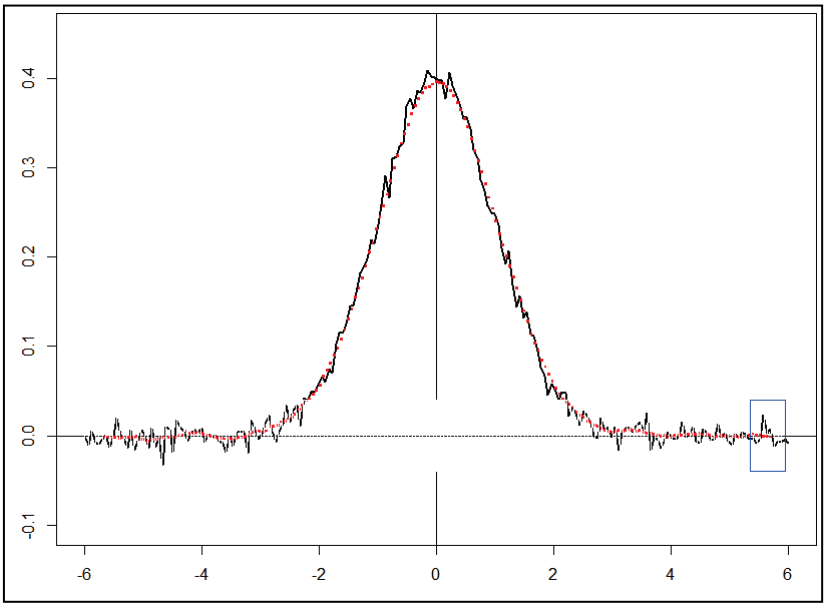


Figure 8.10. Result of smoothing. Measured (black) and smoothed (red) data points. Note that we have not shifted the red curve in this plot.

The formula for the moving average can be generalized to:

$$y[i] = \frac{1}{M} \sum_{j=-(M-1)/2}^{(M-1)/2} x[i+j]$$

Window size

Instead of five points for calculating the average we may take any other odd number (e.g., *M*=15 or *M*=51). This is called the window size. The greater the window size (*M*), the more intense is the smoothing effect. In Figure 8.11 we have taken a window size of *M*=45 (and increased the number of data points). Compared to a window size of *M*=5 in the previous figure, we see that the smoothed curve in Figure 8.11 has become smoother (more noise removed). However, we also see that details from the signal are lost i.e., it has become broader and the height decreased: the signal becomes distorted. Too much weight is given to points that are well removed from the central point. The moving average algorithm is particularly damaging when the box passes through peaks that are narrow compared to the window size. Thus, in practice we must balance between a good noise reduction and deterioration of the peak shape when choosing the window size.

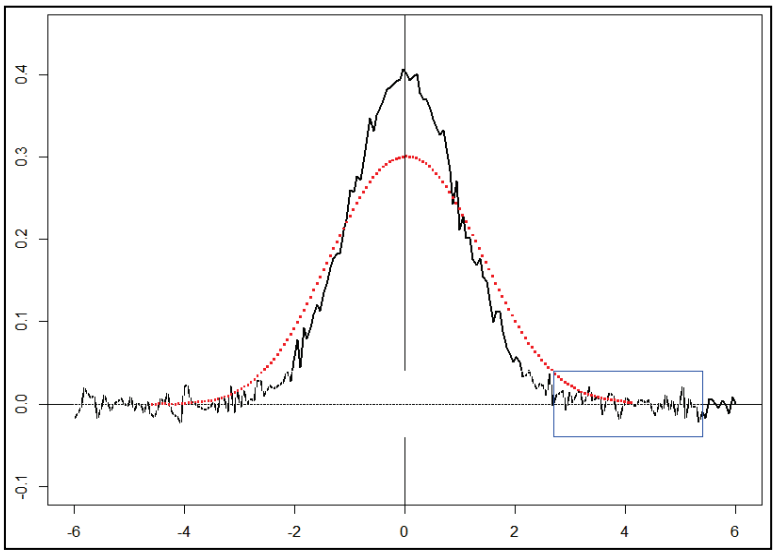


Figure 8.11. Smoothing with moving average and a large window size

8.4.2 Savitzky-Golay algorithm

8.4.2.1 The weighted average

To reduce distortion of the peak shape other smoothers have been developed such as the Savitzky-Golay algorithm. These algorithms do not assign the same weights to the data points for calculating an average. Instead they calculate a **weighted average**.

The formula for the moving average which we have seen in the previous section (with a window size of *W*=5) can now be rewritten by including weights:

$$y[i] = \frac{1}{N} \sum_{j=-(W-1)/2}^{(W-1)/2} w_j x[i+j] \qquad N = \sum w_j = 5$$

$$w_{-2} = w_{-1} = w_0 = w_1 = w_2 = 1$$

$$y[3] = \frac{w_{-2}x[1] + w_{-1}x[2] + w_0 x[3] + w_1 x[4] + w_2 x[5]}{\sum w_j}$$

$$y[3] = \frac{1x[1] + 1x[2] + 1x[3] + 1x[4] + 1x[5]}{5}$$

Thus we see in this example for data point 3, that each measurement is multiplied by a weight, which are in this case all set to one (each measurement is equally important). Thus, nothing changed by rewriting the formula!

Alternatively, we may decide that points that are further away from the central point receive less weight. In such case we may calculate a **weighted average**:

$$w_{-2} = 0.1$$

$$w_{-1} = 0.5$$

$$w_0 = 1$$

$$w_1 = 0.5$$

$$w_2 = 0.1$$

$$y[3] = \frac{w_{-2}x[1] + w_{-1}x[2] + w_0 x[3] + w_1 x[4] + w_2 x[5]}{2.2}$$

$$y[3] = \frac{0.1x[1] + 0.5x[2] + 1x[3] + 0.5x[4] + 0.1x[5]}{2.2}$$

8.4.2.2 Savitzky-Golay algorithm (not for examination)

The question is how to optimally choose the weights. Algorithms such as Savitzky-Golay choose the weights in a smart way.

A better procedure than simply averaging points is to draw (fit) a line through the data points in a window such that this line describes follows the data points as closely as possible. Examples of lines that you can draw are:

$$y(x) = b_0 + b_1 x \quad \text{(straight line)}$$
$$y(x) = b_0 + b_1 x + b_2 x^2 \quad \text{(quadratic; parabola)}$$
$$y(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3 \quad \text{(cubic)}$$

These lines are so-called 'polynomials'.

For example, in Figure 8.12 we see how you can draw a parabola through five data points in a window. The center point is smoothed by moving the point towards the line of the parabola.

The advantage of this approach is that you can choose a polynomial that reflects the shape of your data points and therefore results in less distortion of the signal in comparison to a simple average.

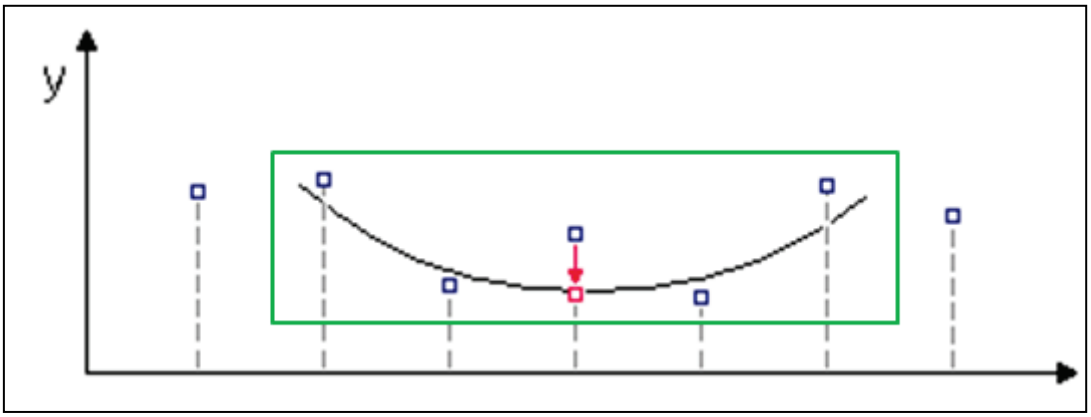


Figure 8.12. Smoothing by using a parabola (line) within a window of five points (green box). The center point is smoothed by moving it to the line of the parabola (red point). In the next step the window is moved one point to the right and the procedure is repeated.

Savitzky and Golay [38] showed that by choosing a specific normalization constant (N) and specific weights (*w*) for smoothing within the moving window one effectively performs a smoothing with a polynomial.

Thus, again, we can simply calculate a moving average in a similar manner as we have seen above (W is window size)

$$y[i] = \frac{1}{N} \sum_{j=-(W-1)/2}^{(W-1)/2} w_j x[i+j]$$

It is beyond the scope of this text to explain how *N* and the weights *w* are derived but if we want to use a parabola then we use the weights shown in Table 8.1. Thus, if the window size is five, then

$$y[i] = \frac{1}{N} \sum_{j=-(W-1)/2}^{(W-1)/2} w_j x[i+j]$$

$$y[i] = \frac{1}{35}\left[-3x[i-2] + 12x[i-1] + 17x[i] + 12x[i+1] - 3x[i+2]\right]$$

$$i = 10 \rightarrow y[10] = \frac{1}{35}\left[-3x[8] + 12x[9] + 17x[10] + 12x[11] - 3x[12]\right]$$

For data point 10 one smooth the data by taking a weighted average of data points 8, 9, 10, 11 and 12. The weights are such that this effectively fits a parabola to the data. Note that the normalization constant *N* is the sum of the weights.

	Window size (W)			
i	11	9	7	5
-5	-36			
-4	9	-21		
-3	44	14	-2	
-2	69	39	3	-3
-1	84	54	6	12
0	89	59	7	17
1	84	54	6	12
2	69	39	3	-3
3	44	14	-2	
4	9	-21		
5	-36			
Normalization (N)	**429**	**231**	**21**	**35**

Table 8.1. Weights to fit a quadratic polynomial (parabola) to data points within a window size of 5, 7, 9 or 11.

The smoothing effect of the Savitzky-Golay algorithm is not as aggressive as the moving average and the loss and/or distortion of vital information is comparatively limited. However, it should be stressed that both algorithms are "lossy", i.e. part of the original information is

lost or distorted. Figure 8.13 shows a comparison of the moving average and Savitzky-Golay smoothers. For example, when looking at the first peak in the raw data, then this could actually be two overlapping peaks (a small shoulder peak seems to be present at the left flank). This information has completely disappeared with the moving average, but is still visible after application of Savitzky-Golay.

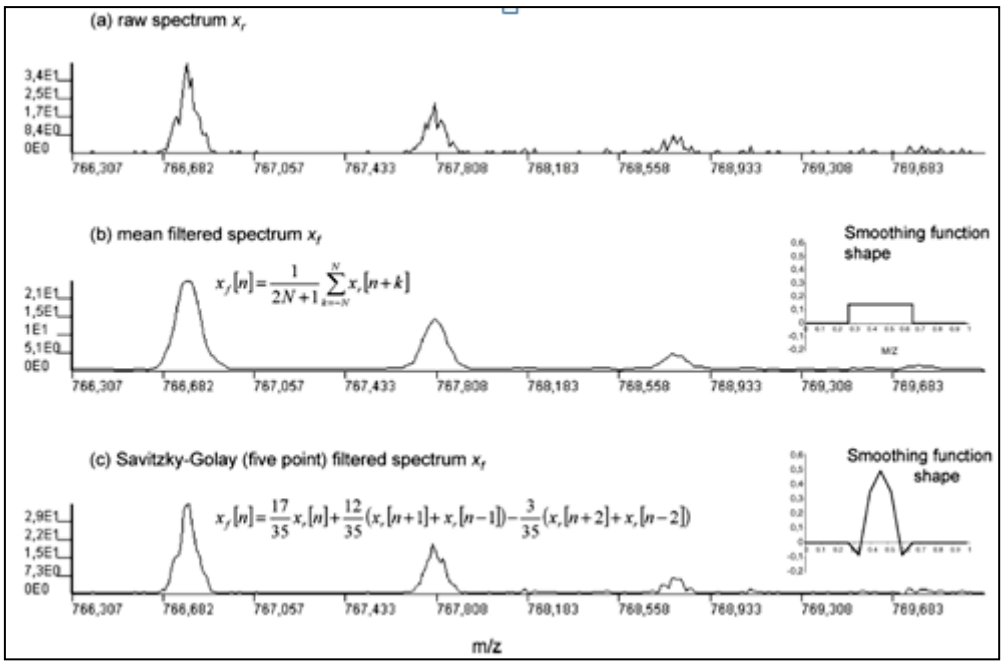


Figure 8.13. Smoothing of noisy raw data with a moving average and Savitzky-Golay filter. Note the 'shoulder' peak in the first peak of the raw spectrum. This shoulder erroneously was filtered out by the moving average, but is still visible after application of Savitzky-Golay.

8.5 Calculating the first and second derivative (not for examination)

In this section we will look at the effect of smoothing on the first and second derivatives of a Gaussian peaks.

The **first derivative** of a curve gives the rate at which a signal y changes with respect to a change in the independent input x. You may recall that the derivative at a point equals the slope of the tangent line to the graph of the function at that point (Figure 8.14).

The **second derivative** is the derivative of the derivative. Roughly speaking, the second derivative measures how the rate of change of a quantity is itself changing; for example, the second derivative of the position of a vehicle with respect to time is the instantaneous acceleration of the vehicle, or the rate at which the velocity of the vehicle is changing.

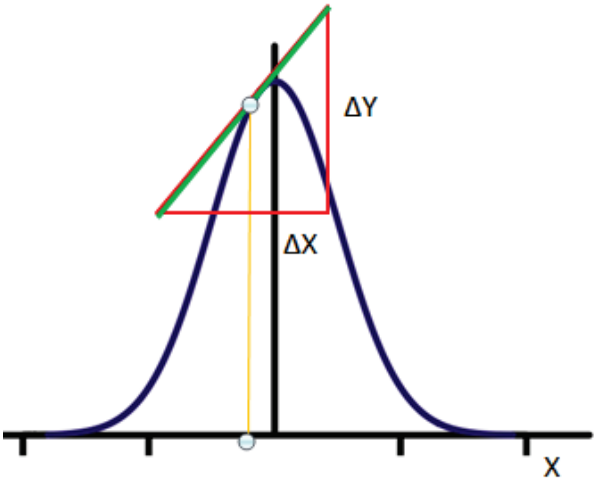


Figure 8.14. The first derivative. In this figure the first derivate at a point x (indicated by the yellow line) is equal to the slope of a tangent line (green line) at that point. The slope of this tangent line is given by first derivative=slope=ΔY/ΔX.

The first derivative and second derivatives can be used to detect the position of peaks and also the start and end of peaks. This is used by the software package XCMS to detect peaks in LC-MS metabolomics data as we will see in another practicum.

Unfortunately, the derivative is very sensitive to noise. Therefore, before calculating the derivative one should reduce the noise in the data.

8.6 Matched filtration: using Gaussians to smooth (not for examination)

Chromatographic peaks do not match polynomial functions since these are not e.g., a parabola. Therefore, applying Savitzky-Golay may cause peak distortion during smoothing.

Chromatographic peaks much more resemble Gaussian curves (although this is still a simplification). Therefore, instead of choosing weights to fit a polynomial, one may choose weights that would fit a Gaussian curve (Figure 8.15 shows the smoothing function shape i.e., the weights). Smoothing with a curve that matches the peak shape in the measured data is called **matched filtration**.

The negative second derivative of the Gaussian curve can be used to detect peaks and determine the peak start and peak stop. It is beyond the scope of this chapter to explain this. A nice trick is to perform smoothing with weights that are chosen according to such negative second derivative as shown in Figure 8.15. By doing this one may reduce noise and calculate the second derivative in a single step. Nice trick! This is used by XCMS for peak detection.

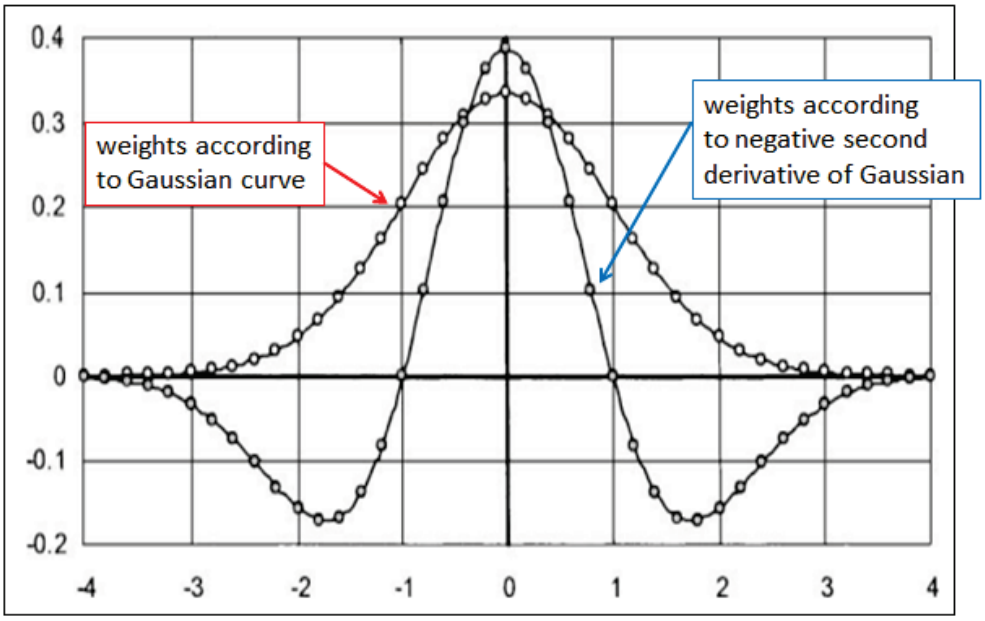


Figure 8.15. Weights according to a Gaussian curve and weights according to the negative second derivative of a Gaussian curve. (Figure copied from Danielsson et al (2002) Analytica Chimica Acta, 454:167–184).

8.7 References

1. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nature genetics 42: 30-35.
2. Savitzky A, Golay MJE (1964) Smoothing and differentition of data by simplified least squares procedures. Analytical Chemistry 36: 1.

9 Pre-processing of LC-MS data

Lecturer: Prof. dr. Antoine van Kampen (AMC)

After reading this chapter you should understand

- Peak table and its content
- Structure of LC-MS data
- TIC, EIC
- The different pre-processing steps and why they are necessary (you don't have to know the exact methods or algorithms)
- Note that the methods are explained in more detail in order to better understand the computer exercises.

Contents

9.1 Introduction

Metabolite profiling has gained popularity using a number of techniques including nuclear magnetic resonance (NMR) or different combinations of liquid chromatography (LC), gas chromatography (GC), and mass spectrometry (MS). One particularly popular platform for untargeted metabolite profiling is LC/MS. LC/MS resolves individual chemical components into separate peaks. Unlike GC/MS, it additionally detects nonvolatile compounds, which make up a large proportion of metabolites. Finally, LC separation provides a means for resolving isobaric compounds (compounds with the same weight).

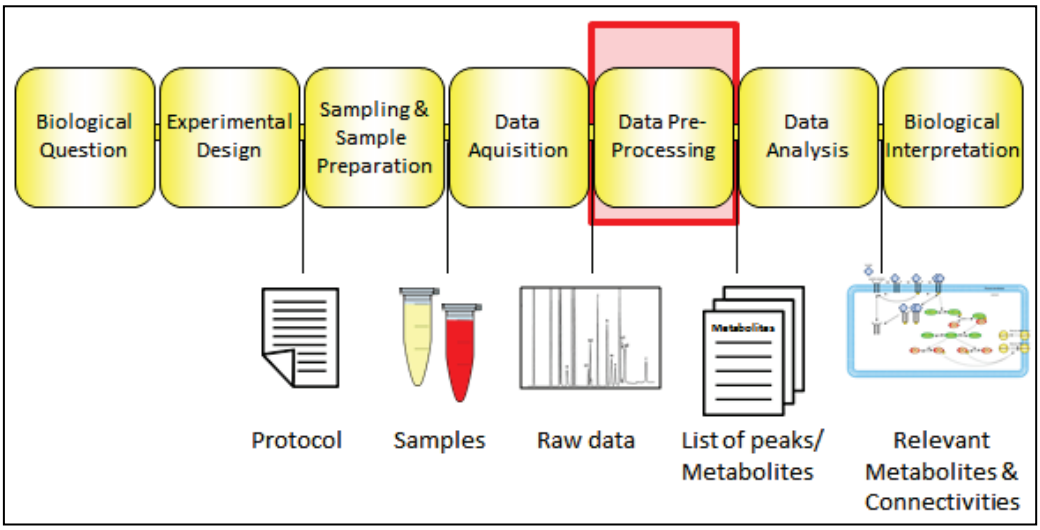


Figure 9.1. Overall metabolomics workflow. The data pre-processing step is highlighted and produces peak tables from the raw LC-MS data.

The raw LC-MS data needs to be pre-processed to produce lists of peaks (Figure 9.1) for each sample. Such peak lists (**peak tables**) provide information such as the m/z value, retention time, and intensity (or areas) of the peaks in a sample. Each peak corresponds to a specific metabolite and its isotopes, but for many peaks the assignment of a metabolite name is impossible because its mass alone do not always provide sufficient information for identification. The peak height and the peak area provide information about the relative concentration of the metabolite. Table 9.1 shows small part of a typical peak table, which normally contains hundreds to thousands of peaks. Peak tables provide the input for further (statistical) analysis. The generation of peak tables from raw data is still one of the difficult steps in the metabolomics pipeline, and in practice the peak tables are generally checked by mass-spectroscopy experts.

m/z	Rt (minutes)	Intensity	Start (minutes)	End (minutes)	Area	Metabolite
600.4	4	10000	3.5	4.4	50000	glucose
700.9	4.5	5000	3.6	4.6	10000	unknown
756.5	6	12056	5.6	6.4	34000	unknown
etc						

Table 9.1. Example of a peak table (not real values). One peak is identified as 'glucose'. The metabolites corresponding to the other peaks could not be identified. The peak intensity (height) and peak area are measures for the relative concentration of the metabolites. Start and End indicate the start en end of the chromatographic peak between which the area of the peak was calculated.

9.2 LC-MS data

Figure 9.2 shows a graphical representation of a LC-MS dataset, which contains numerous peaks. Each peak is characterized by a specific m/z-value and chromatographic retention time, and has certain intensity. The black profile represents the so-called **Total Ion Current** (TIC) chromatogram and is the sum of the intensities for all m/z values. This is also shown in Figure 9.3, which shows a TIC with two 'scans' indicated for 120 and 121 seconds. Thus, at 120 seconds all masses (in the chromatographic fraction) are measured and their intensities are summed. The summed intensity is plotted in the TIC graph. Compare this to Figure 9.4.The TIC

- Monitors a very large window often of several hundred mass-to-charge units (the entire range of masses being detected).
- It is useful in understanding the overall chromatography and hunting for new peaks.
- In complex samples the TIC can often be fairly useless as so many peaks elute over the run that the peaks are unresolved.
- The TIC is reconstructed after data collection of complete mass spectral data.

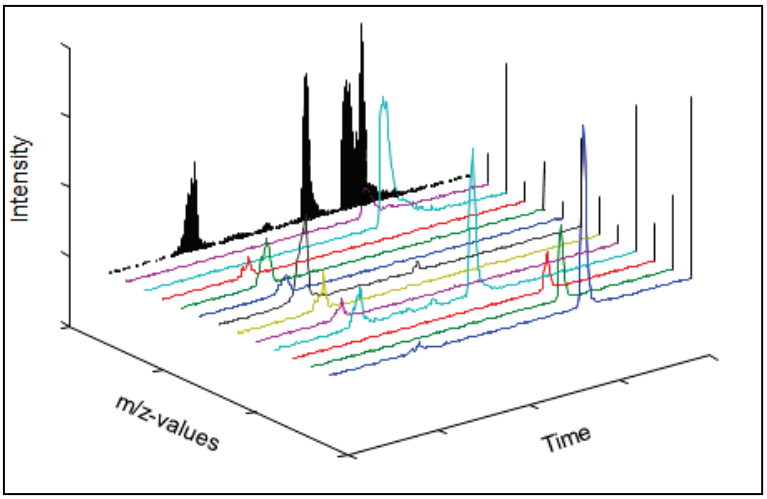


Figure 9.2. Example of a LC-MS dataset (copied from the NUGO workshop 2007, TNO).

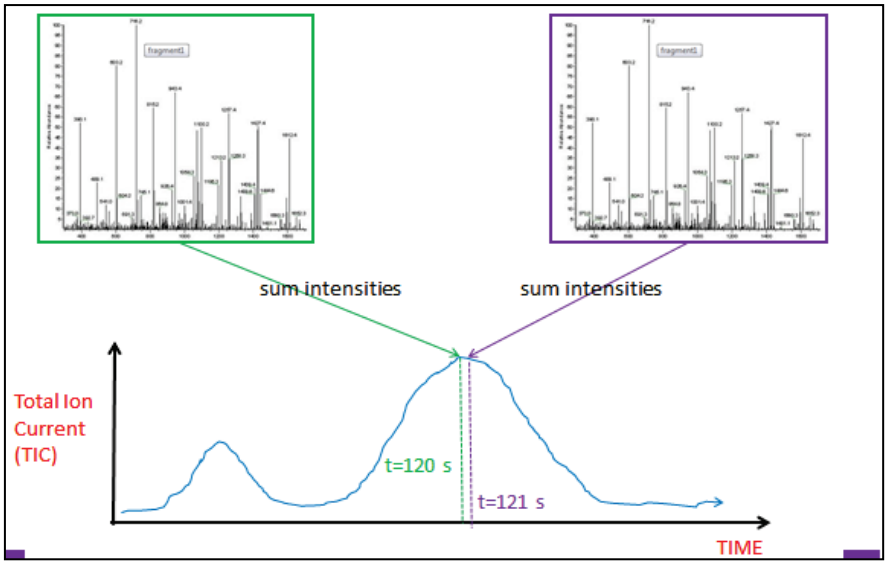


Figure 9.3. Total Ion Current (TIC) chromatogram. Two scans (mass spectra) are shown for t=120s and t=121s.

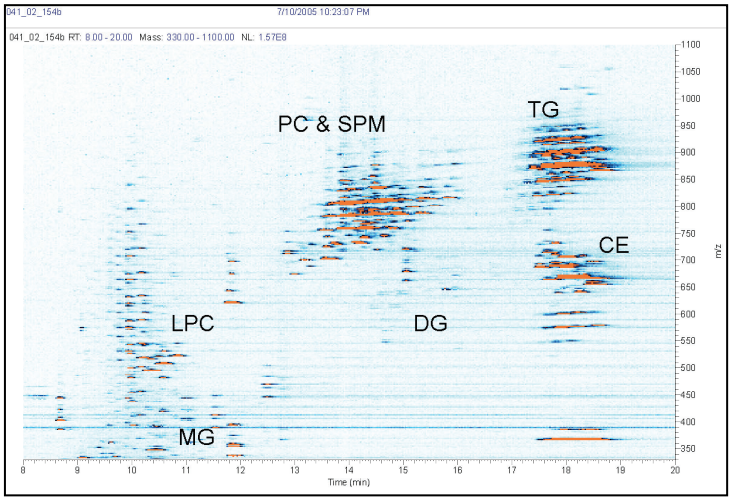


Figure 9.4. LC-MS dataset. Compare this to Figure 9.2. Here we have not visualized the dataset in three dimensions but as a so-called heatmap. The x-axis denotes the retention time. The y-axis is are the m/z values. The colors/intensities indicate the various peaks (in this case lipids). MG: Monoradylglycerolipids, DG: Diradylglycerolipids, CE: cholesteryl ester, TG: Triglycerides, PC: Phosphatidyl choline, LPC: lyso-Phosphatidyl choline. (Figure copied from the NUGO workshop 2007, TNO).

Extracted Ion Chromatograms (EIC or XIC) represent the sum of intensities for one or more m/z values that are recovered ('extracted') from the data set for a chromatographic run. Such EIC corresponds to one or multiple lines (mass spectra) in Figure 9.2. EIC's are useful for re-examining data to:

- Provide clean chromatograms of compounds of interest (i.e., for only one m/z value; see Figure 9.5)
- Detect previously-unsuspected metabolites. All m/z values are examined separately in an EIC and can easily be visually detected. In a TIC these peaks may be 'hidden'.
- Highlight potential isomers. Isomers have the same mass (m/z value) but may elute at different retention times and can thus be detected by looking at single masses.

- Resolve suspected co-eluting substances because only a single m/z value is considered.

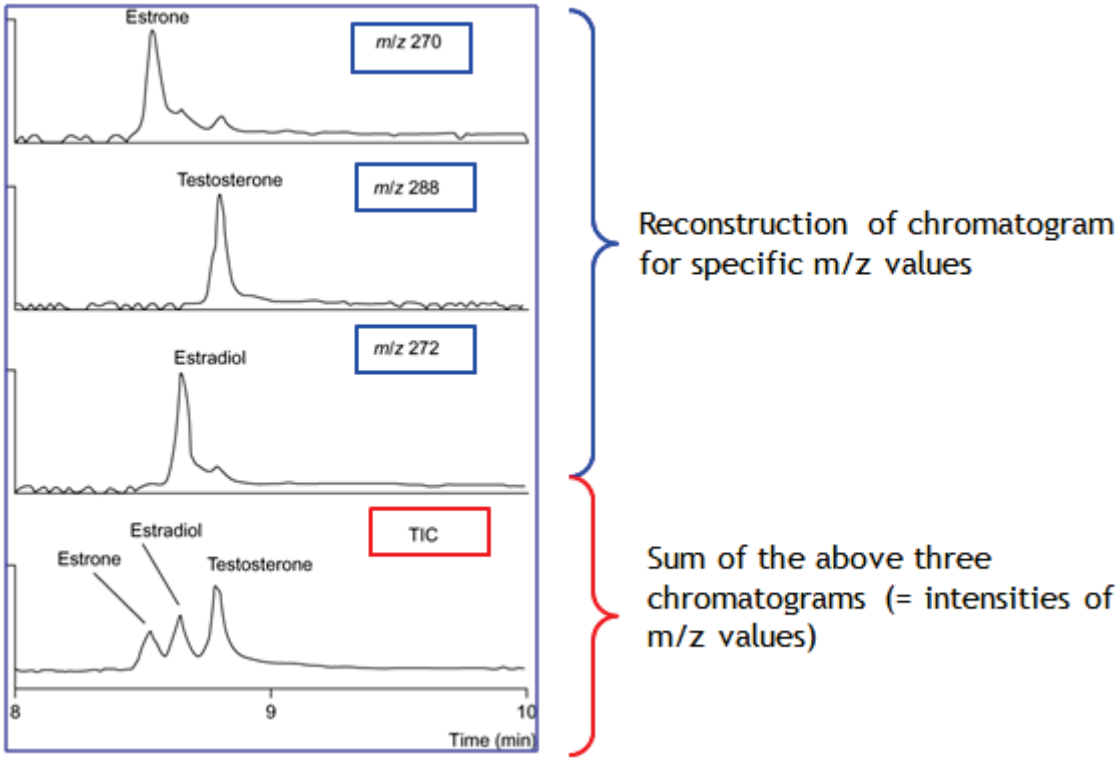


Figure 9.5. Three extracted ion chromatograms (EIC) and the corresponding TIC.

9.3 Overall outline of the main pre-processing steps

To illustrate several main pre-processing steps for LC-MS data, the open-source R package XCMS (an acronym for various forms (X) of chromatography mass spectrometry) is used as an example. XCMS will also be used for the computer exercises.

The following pre-processing steps will be discussed (Figure 9.6):

1. Filtering and identification of peaks
    - This step combines noise reduction, non-constant baseline removal, and identification of peaks from the raw data of a single sample (not yet the assignment of metabolite names to these peaks)
2. Match peaks across samples
    - If multiple samples are measured than in this step one determines which peaks from each sample correspond to each other. This matching is not obvious because corresponding peaks from different samples may have slightly different retention times.
3. Retention time correction
    - Based on the peak matching, corrections are made to retention times of the samples to make them more comparable.

4. Fill in missing peak data
    - In some cases for one or few samples, peaks may not be found. In this case one may decide to fill in the missing peak data.
5. Assignment of metabolite names to identified peaks
    - This step involves the matching of the peaks (m/z values) against the MetLin database.
6. Statistically analyze results
    - For example, t-tests to compare two groups of samples.
7. Visualization of peaks
    - Inspection of EIC's for classes of samples, and inspection of integration boundaries used to determine peak areas.

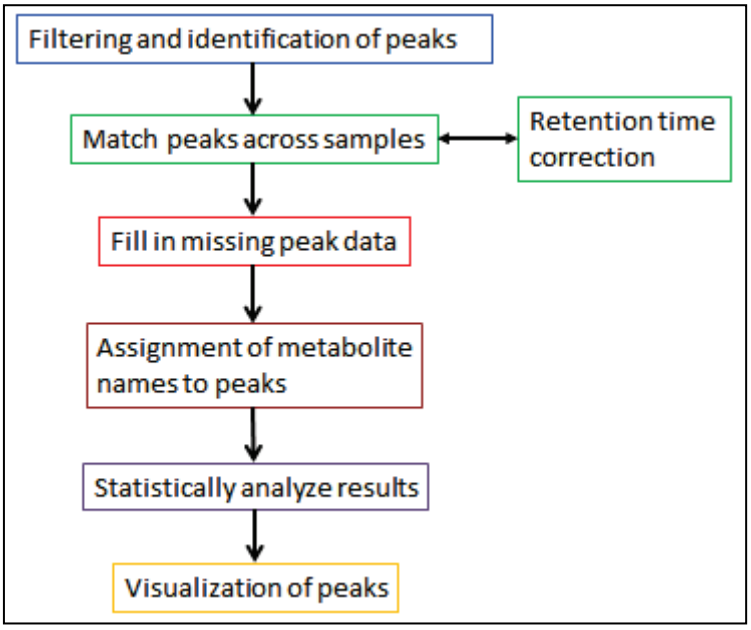


Figure 9.6. Flowchart showing the general strategy for pre-processing of LC-MS data.

9.3.1 Identification of peaks

The identification of peaks from raw LC-MS data of a single sample should result in the peak position (m/z and retention time values), peak height, peak area and integration boundaries (the start and end of the peak between which the area has been determined). Thus, in this step we generate the peak table. The peak height and peak area are a measure for the relative concentration of a metabolite within a sample. This step does not yet involve the assignment of metabolite names to the peaks, which is done at a later stage.

Many methods have been developed for peak identification, which is still a challenge due to noise, baseline drift, distorted peaks, overlapping peaks, etc. Recall from Figure 9.2 that the LC-MS produces a two-dimensional dataset with peaks defined by the m/z and retention time values. Having this in mind, three approaches can be taken towards peak identification (Figure 9.7):

- **Two directions**. Peaks can be detected along the m/z axis and along the retention time axis. Thus, one effectively projects the data on one axis and applies a peak identification method to determine the peaks. Only if a peak is detected at both the m/z and retention time axis the peak position has been identified.
- **XIC (=EIC) slices**. Here one divides the m/z axis in small slices (called **bins**). For every m/z slice one reconstructs the chromatogram in the retention time direction (this is an Extracted Ion Chromatogram (EIC), as explained above). Subsequently, peaks are identified in the EIC's of every m/z bin.
- **Model fitting**. One may also attempt to fit a two-dimensional Gaussian curve to the data directly.

We will only consider the XIC slice approach, which is implemented by the software package xcms.

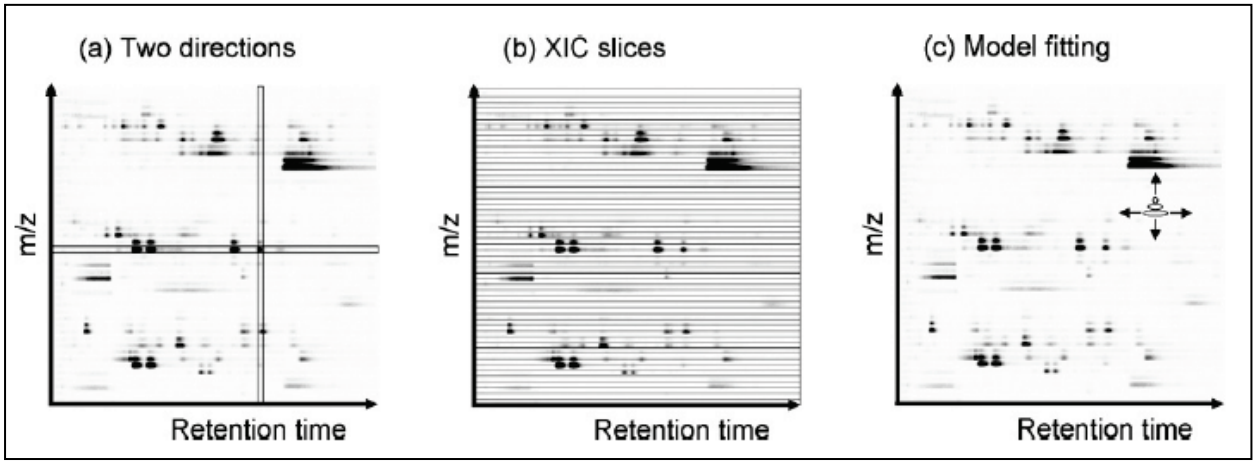


Figure 9.7. Peak identification approaches. Peaks are shown in a two-dimensional plane defined by the m/z and retention time values. The black/grey areas denote peaks where the intensities of these areas reflect the intensity of the peak (i.e., peak height). (Figure copied from Katajamaa and Oresic (2007) Journal of Chromatography A, 1158, 318)

9.3.2 Binning

The purpose of binning is to have a single peak intensity within a bin at each time point from which an extracted ion chromatogram (EIC) can be determined. Thus, an EIC is constructed for each bin. Subsequent peak detection (see below) is performed on each EIC.

Binning is the process of dividing the m/z axis in equidistant slices (e.g., 0.1 m/z). Within each slice, the signal is determined by taking the maximum intensity at each time point in the slice (Figure 9.8).

Binning has a number of drawbacks. If breaks between bins are chosen arbitrarily, the signal from a given analyte (metabolite) may get split between two adjacent bins. Additionally, multiple independent analytes may contribute to the signal of a single bin (thus two peaks

within one bin), thus decreasing the ability to discern significant differences in individual metabolites. In practice one, therefore, has to carefully set the width of the slice.

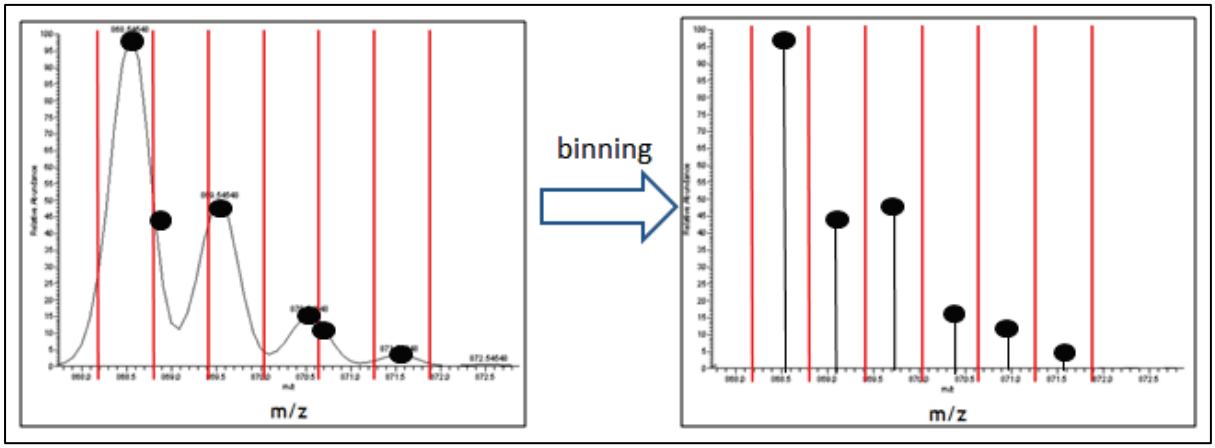


Figure 9.8. The process of binning. In this figure one mass-spectrum at an arbitrary retention time is shown (this is one scan). The m/z axis is divided in equidistant slices (red lines). We see that within each bin we have a range of intensity values (the m/z peaks). To simplify further processing we aim to assign a single intensity to each bin. Therefore, within each bin, we determine the maximum value of the m/z peak intensities (indicated by the black dots). Subsequently, these intensities are assigned to each bin (right panel).

Figure 9.9 shows the reconstruction of a chromatogram from one specific bin (m/z = 268.1). This is the so-called extracted ion base-peak chromatogram (EIBPC) or just EIC as defined above. For every scan (i.e., time point in the chromatographic direction) the intensity is reflected by the intensity of the signal in the selected bin. Thus, one projects the selected slice on the retention time axis. The EIC shows that the metabolite with m/z=268.1 elutes from the chromatographic column at about 3680 seconds. Although we can clearly see the peak, for software algorithms it is not so easy to determine the position, start, end, height and area of such peak. Looking at the EIC, you can understand why: the signal looks very noisy (erratic). For this reason we need to reduce the noise. XCMS uses matched filtration to do this.

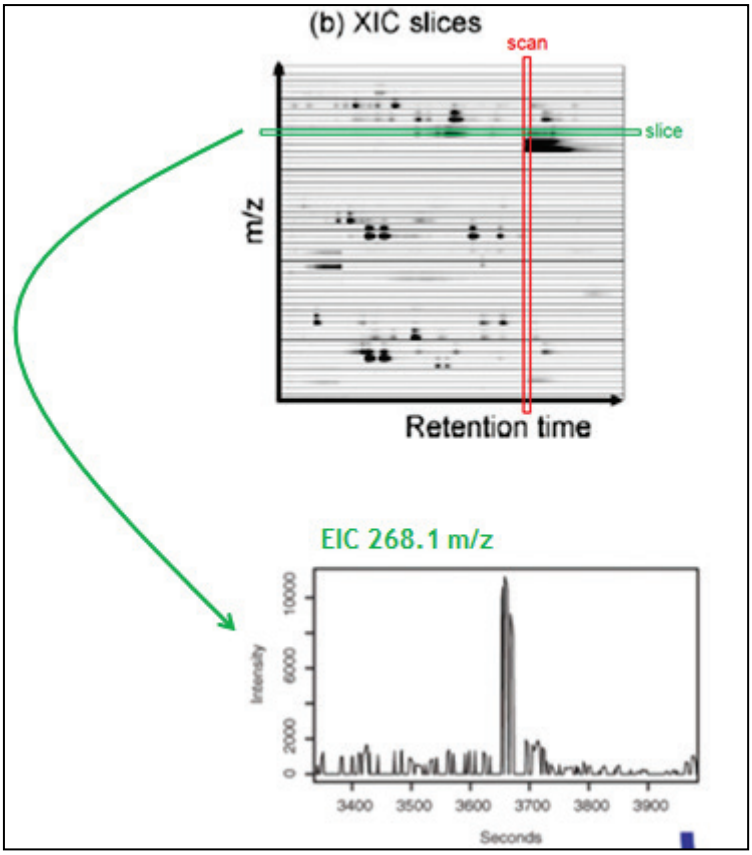


Figure 9.9. The process of binning and reconstruction of an Extracted Ion Chromatogram (EIC) from a specific splice (for m/z=268.1). In this example, one bin (slice; green box) is selected. The intensities of the peaks within this bin are used to calculate the chromatographic peak (in the time domain). The red box corresponds to a single scan (m/z spectrum at a certain time point). At the cross-point between the slice and scan the intensity of the peak is determined and plotted in the EIC.

9.3.3 Matched filtration for peak identification

In the previous step we have defined the bins, and for each bin an EIC can be constructed in which we need to identify the peaks. The method of 'matched filtration' is used for peak detection, which results in the position of each peak, the peak height, peak start, peak stop and the area between start and stop. Determination of the peak area is also called peak 'integration'.

In the computer practicum 'From experiment to data' we have learned about smoothing and first/second derivatives. Matched filtration is based on these principles. Matched filtration is the smoothing of the EICs with a moving window in which the weights are chosen according to a negative second derivative of a Gaussian curve. By doing this, noise reduction, baseline correction and peak detection are combined in a single step! This principle of matched filtration is shown in Figure 9.10.

For smoothing the weights of the moving filter are chosen according to the negative second-derivative Gaussian. This has several effects. First, since we are still smoothing (but with very special weights) the noise of the signal is reduced. Second, because we have chosen the

weights according to the second-derivative, the smoothed curve will reflect the negative second-derivative of a Gaussian curve. Finally, the baseline is (partially) removed due to application of the second-derivative.

The peak of the negative second-derivative coincides with the peak in the raw data. Thus, if the peak of the second-derivative is sufficiently large (several times above the noise level) then it effectively identifies the position of a peak. To determine the area of the peak, one needs to determine the start en end of the peak. This information is also provided by the second-derivative. The second-derivative transformation causes the filtered chromatogram (=second derivative) to cross the x-axis at the peak inflection points. Those zero-crossing points define the area of peak integration (it is beyond the scope of these lecture notes to demonstrate that these points indeed provide proper boundaries).

After application of matched filtration, we have obtained a peak table for a single sample.

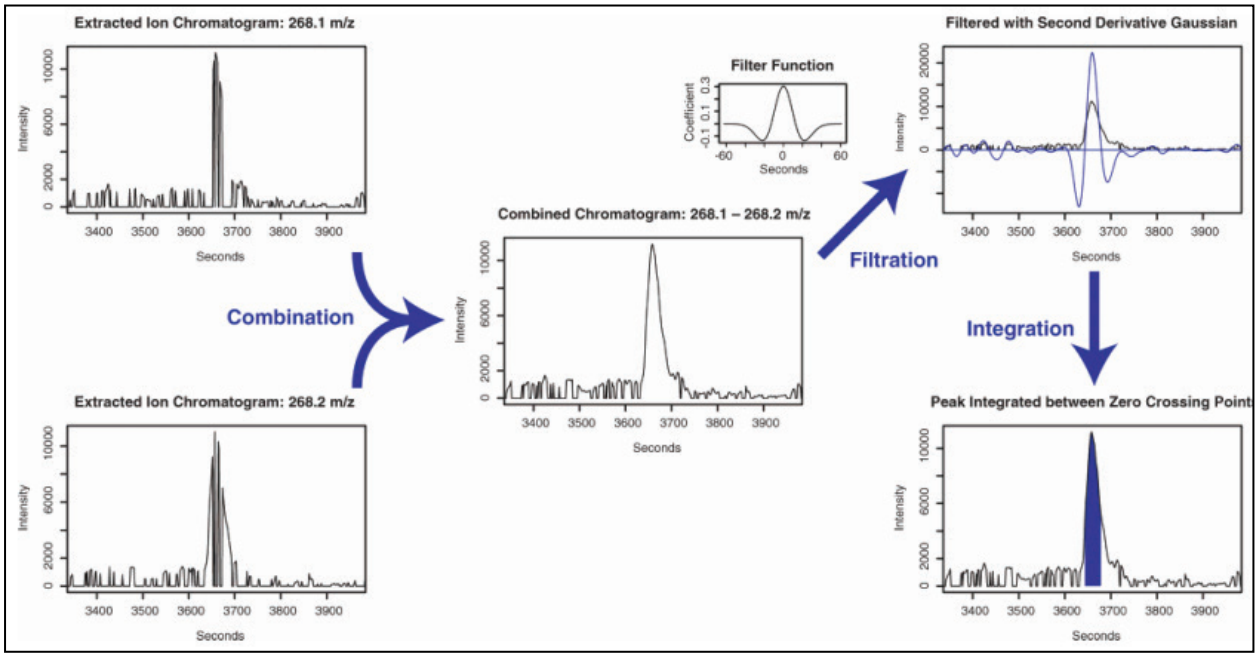


Figure 9.10. Illustration of the peak detection method using a single peak from LC-MS data data. The peak shown here is contained in two adjacent chromatographic bins at 268.1 and 268.2 m/z. Combining two EICs (bins) gives a less erratic chromatogram, which makes peak detection easier. The algorithm creates overlapping combined chromatograms (i.e., m/z 268.0/268.1, 268.1/268.2, 268.2/268.3, etc.) with the m/z 268.1/268.2 chromatogram showing a cleaner signal. The data are then processed with a matched filter (filter function = smoother) whose coefficients are equal to a negative second-derivative Gaussian function. The second-derivative transformation causes the filtered chromatogram to cross the x-axis at the peak inflection points (blue curve in top right panel). Those zero-crossing points define the area of peak integration (blue area shown in peak in bottom right panel).

9.3.4 Matching and alignment of peaks in multiple samples

During a chromatography experiment retention times measured for metabolites may (slightly) vary between samples for several reasons. For example, in Figure 9.11 ten replicate

samples are shown. The three large peaks (metabolites) do not have the precise same retention time. It is, therefore, necessary to establish which peaks from the different sample correspond to each other. Although, this might be clear from visual inspection, this is much more difficult for a computer program. However, once the corresponding peaks are identified, the retention times of all peaks may be corrected such that corresponding peaks are perfectly aligned at a single position. We need to take two steps:

- Peak matching: determine which peaks from different samples correspond
- Retention time alignment: correct the retention times for the peaks in the different samples.

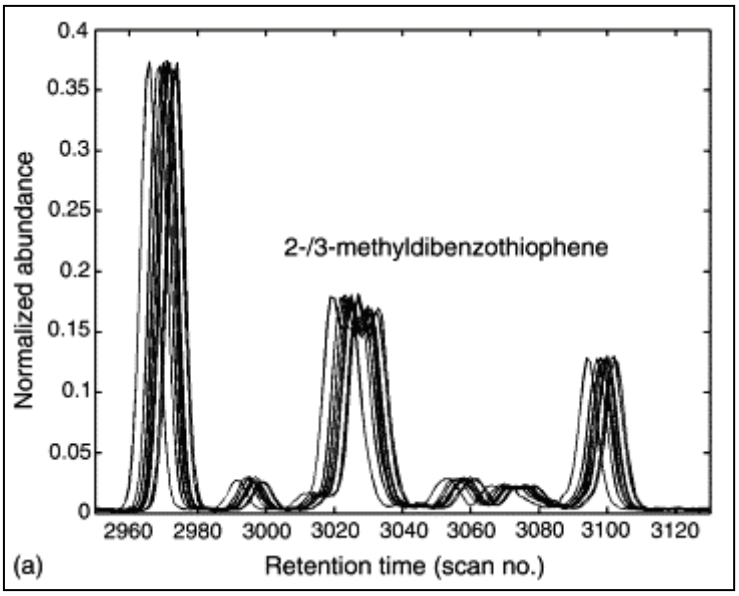


Figure 9.11. Section of the m/z 198 chromatogram for ten selected replicate laboratory reference samples. One identified peak (2-/3-methyldibenzothiophene and two unidentified peaks) (Figure copied from Christensen et al (2005) J Chromatogr A, 1062(1), 113).

9.3.4.1 Peak matching

The principle of peak matching is straightforward but there are several details involved which are omitted in this text. For illustration of the peak matching method we look at a simplified representation of the data (right panel of Figure 9.12).

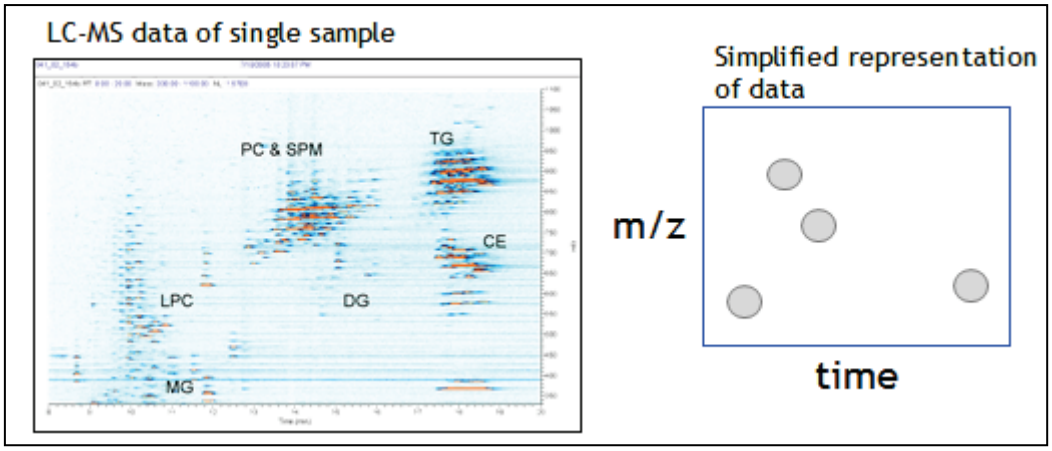


Figure 9.12. LC-MS data shown as heat map (left) and a simplified representation of a different heat map with only four peaks (grey circles).

The general principle of the peak matching method is shown in Figure 9.13. Four samples are shown as four squares (m/z and time direction). For simplicity we have only shown 2 peaks (grey and yellow) in each sample. The matching of corresponding peaks requires the identification of peaks that have similar m/z and similar retention time values. Therefore, as a first step, the m/z axis is divided into small overlapping bins. One such bin is shown by the dashed box. All peaks within a bin have, by definition, a similar m/z value. In a second step, for all peaks within a single bin the peaks are grouped according to retention time by using a so-called Gaussian Kernel (for the current purpose, this can be seen as a Gaussian curve) This kernel groups peaks with similar retention time as shown in the figure. Once we have completed these two steps we have found all corresponding peaks.

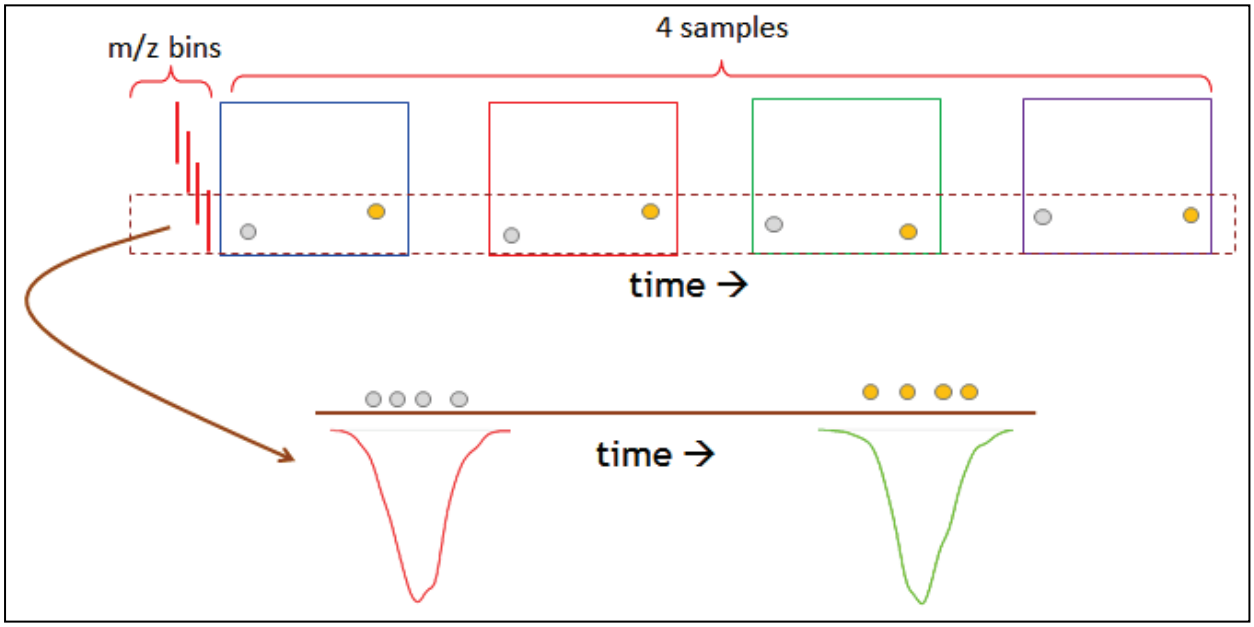


Figure 9.13. General principle of the peak matching method.

Unfortunately, we do not always have a nice situation as in Figure 9.13. For example, Figure 9.14 shows a situation in which not all the peaks occur in all samples (for example, the third

sample does not contain the grey and yellow peak). In this case we can again identify the corresponding peaks but we see that one of the groups only contain a single peak. We will remove this group because it may represent an artifact but, in addition, it cannot be used for the correction of the retention times. The second group contains three peaks which we consider sufficient (in this example) and call a 'well behaved peak group'.

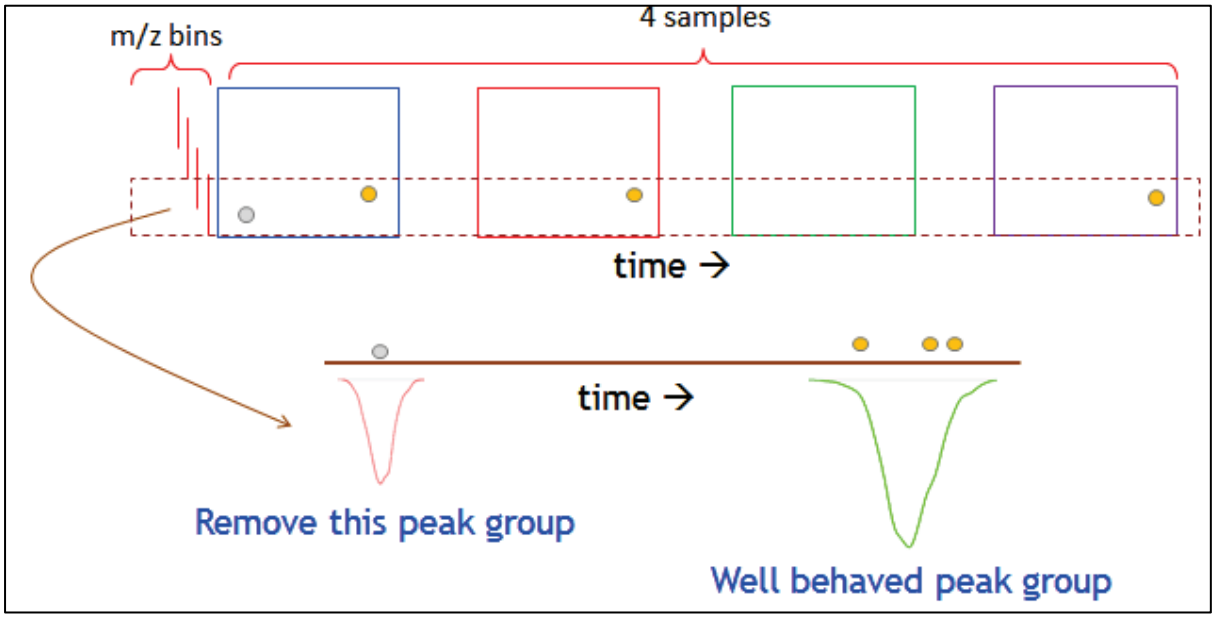


Figure 9.14. General principle of the peak matching method: a more problematic situation. The grey peak is only present in one of the four replicated samples. Therefore, this peak cannot be used for the subsequent time alignment step and is removed. The orange peak is missing in only the third sample and can be used in time alignment. In the 'peak filling' step (see below), the missing peaks are reconstructed from the raw data.

Figure 9.15 shows the results of the peak matching method. The table shows all groups that have been found. Each group is defined by its m/z and retention time boundaries (mzmin, mzmax, rtmin, rtmax) and has a median m/z and retention time value. This is schematically shown in the figure below the table. This particular example represents a study of six wild type and six knockout mice. Thus every metabolite can occur in a maximum of 12 samples. The number of peaks (max 12) is shown in the table as well how many of these peaks (metabolites) are found in the KO samples and in the WT samples.

mzmed	mzmin	mzmax	rtmed	rtmin	rtmax	npeaks	KO	WT
200.1000	200.1000	200.1000	2925.480	2876.967	2931.740	9	4	5
205.0000	205.0000	205.0000	2790.894	2784.635	2795.591	12	6	6
205.9927	205.9786	206.0023	2790.112	2784.635	2795.591	12	6	6
207.0850	207.0440	207.1000	2718.906	2712.647	2726.731	12	6	6
219.0848	219.0488	219.1000	2524.852	2518.592	2529.547	9	4	5
231.0236	231.0000	231.0812	2517.029	2509.202	2535.807	6	3	3
233.0371	233.0128	233.0652	3022.507	3016.247	3077.281	13	6	6
234.0277	234.0133	234.1000	3020.943	3019.378	3028.769	5	2	3
235.0550	235.0000	235.0938	2692.302	2667.264	2720.471	6	3	2
236.1018	236.0678	236.1188	2523.287	2518.592	2529.547	10	5	5

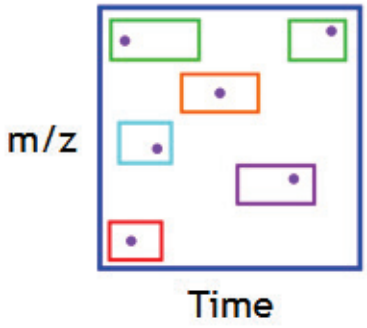


Figure 9.15. Results of the peak matching method. The blue square represents a single sample. The colored rectangles represent the peak groups that were determined from all samples. The purple dots denote the median m/z and retention time within each peak group as determined from all samples.

9.3.5 Retention time alignment

The peak matching step has established which peaks from the different samples correspond (i.e., should correspond to the same metabolite). However, the peak matching did not correct for the differences in retention time between the peaks from different samples. This is done by so-called 'retention time alignment', which is based on the established peak groups.

Retention time alignment is necessary to make the actual correction of retention time deviations. Moreover, it not only corrects the peaks in the peak groups but also all other peaks (based on the peak groups).

Retention time alignment is used to correct the (small) deviations in retention time for the same metabolite in different samples that have occurred during the experiment. The principle of alignment is more complex than the previous steps but we will explain the overall principle, which is shown in Figure 9.16.

In Figure A all the identified peak groups are shown. Consider a single peak group (the orange one), which is shown in Figure B. The purple dot indicates the position of the median retention time and median m/z values for all peaks in this group. The purple dashed line also indicates the median retention time. Now suppose that this peak group contains four samples (Figure

C). Thus the median of the retention times of these four red samples give the purple dot (mzmed). In retention time alignment one adjusts the retention time of the individual samples towards the median retention time (shown in Figure D).

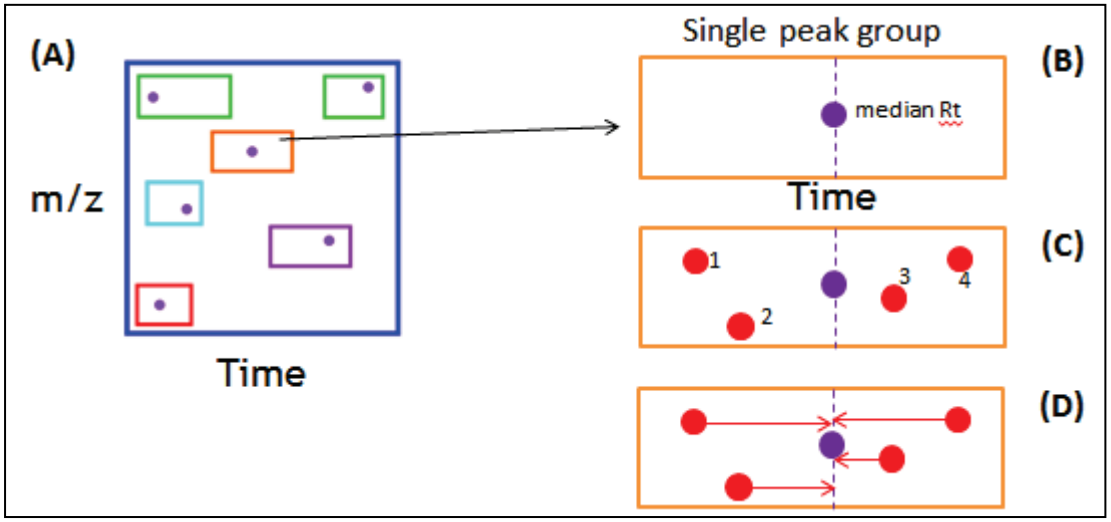


Figure 9.16. Principle of retention time alignment. Panel B, C and D show the effect of peak alignment for peaks within a single peak group. The median retention time (purple dot) was based on four peaks (red dots). During peak alignment, the red peaks are shifted towards the median retention time. Note, that the final result of alignment is that the peaks are closer to the median retention time, but not necessarily equal to the retention time.

The amount of correction is, however, not as simple as setting the retention time of each peak to the median retention time. Instead, a regression analysis is used. Figure D tells us that sample 1 and 2 have a retention time lower than the median and eluted from the chromatographic column before sample 3 and 4. We can also represent this in a different way as shown in Figure 9.17. For purpose of illustration, we selected two samples (1 and 2) in two peak groups (red and green). The green peak group has a larger median retention time than the red peak group, which is shown in figure B. The position of the peak group on the x-axis (retention time) is defined by the median retention time of the peak group. The deviation of sample 1 from the peak group medians corresponds to the y-value. We see that sample 1 has a retention time that is lower than the median retention time, which we define as a positive deviation on the y-axis. The blue arrow shows this correspondence. For sample 2 we see that the deviation is negative for both peak groups.

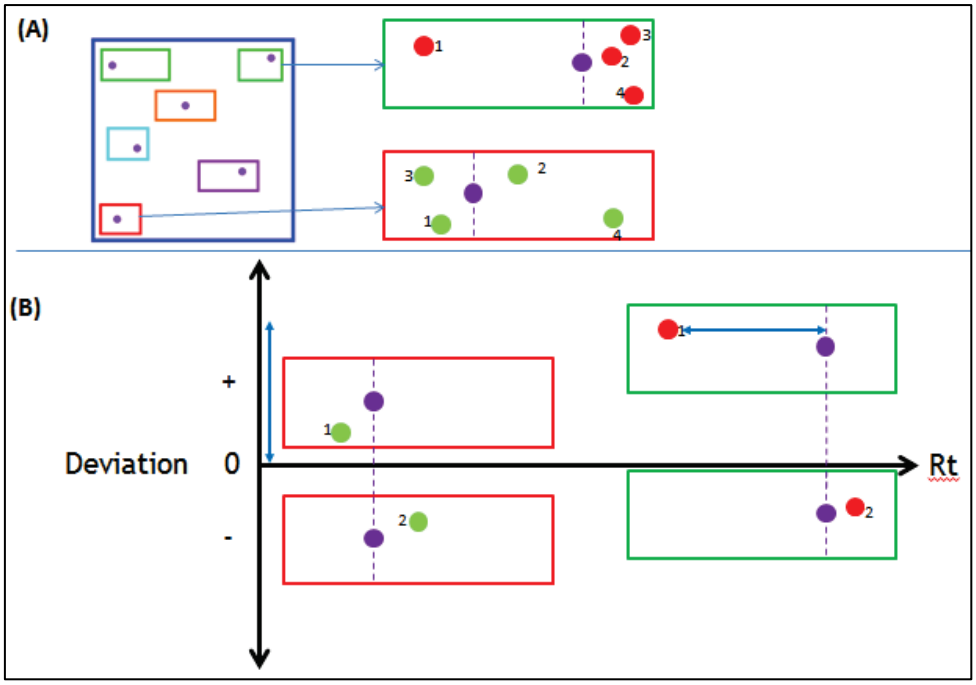


Figure 9.17. A different representation of two peak groups. In this example two peak groups shown in figure (A) are plotted as deviations from the median retention time in figure (B). Note that the green dots represent corresponding peaks (same metabolite). Also the red dots represent corresponding peaks.

We can now further simplify this figure as shown in Figure 9.18A. Here we removed the peak groups and now only show sample 1 and 2. The positive and negative deviations determine the amount of retention time correction (resulting in the blue and orange dots). Keep in mind that not only the peaks in the peak groups are corrected, but also the peaks between the peak groups. The amount of correction is determined by connecting samples from subsequent peak groups with a line (regression; figure B). The y-value of the line determines the amount of correction.

Note that the figures represent a simplification. As we will see with the computer exercises, these figures normally contain many more peak groups, and instead of a linear green line, the peak groups are connected by using lowess (locally weighted scatterplot smoothing), which is a so-called non-linear regression method.

Normally, the deviation of the corrected retention time is not necessarily zero (i.e., equal to the median retention time) and one may perform an additional iteration of peak matching and retention time correction (Figure 9.6). An example of the result of peak matching and retention time alignment is shown in Figure 9.19.

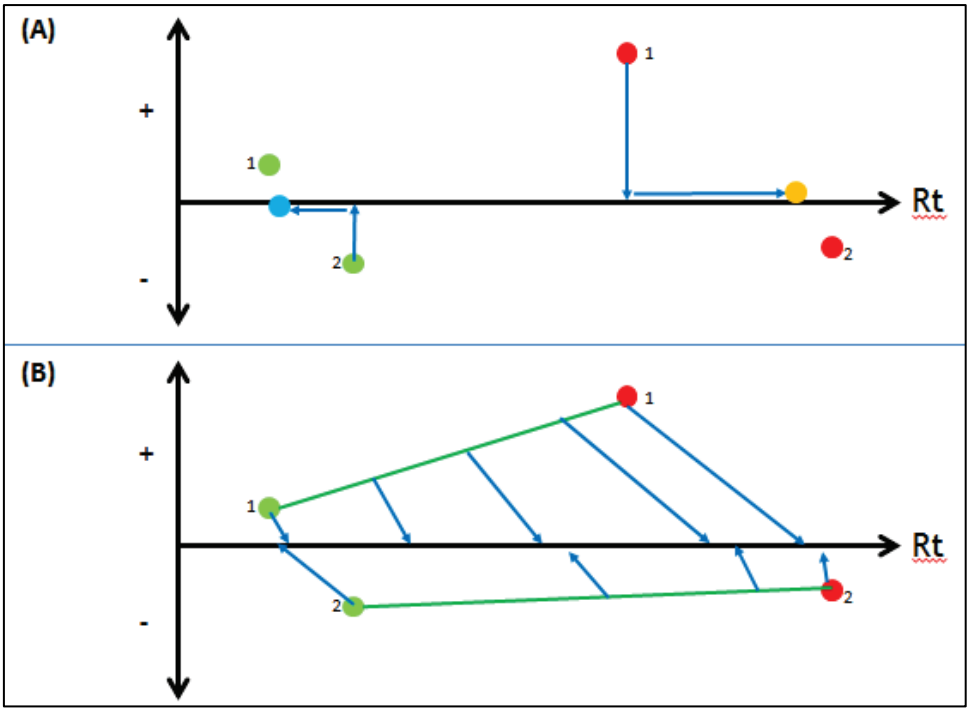


Figure 9.18. Retention time alignment by using regression analysis (simplified case). (A). The blue arrows denote the amount of correction. The deviation (y-axis) is added to the retention time in case of a positive deviation (resulting in the orange dot), or is subtracted from the retention time for negative deviations (resulting in the blue dot). (B) the correction is not only performed for peaks in peak groups but for also for all other peaks between the peak groups. The amount of correction is given by the green line which is determined by regression. Note that the correction indicated by the blue arrows result in (1) zero deviation from the median (blue and orange dot) and (2) the corresponding peaks (two green dots, and two red dots) move closer together.

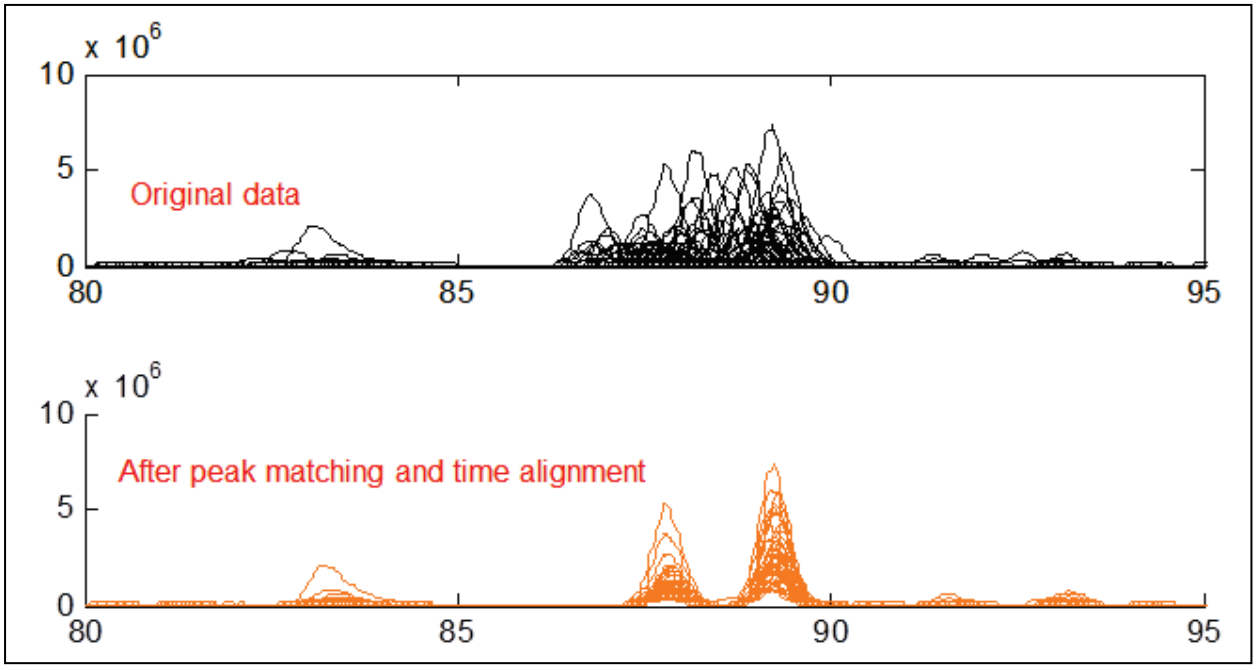


Figure 9.19. Example of peak matching and retention time alignment.

9.3.6 Fill in missing peaks

In principle, when we replicate measurements by measuring multiple samples, then we expect that each peak (metabolite) we will identify will be present in each sample. However, this might not be the case for several reasons (e.g., the metabolite may indeed not be present in the samples, or the peak identification algorithm was not able to identify it. This was shown in Figure 9.14 in which one or more peaks in a peak group could be missing. However, since we know the location of the other peaks in the same peak group we may go back to the raw data of the other samples and add the missing peaks by simply using the intensity values (which could be zero) observed at those positions.

9.3.7 Metabolite identification

It is important to realize that the peak identification algorithm only informs us about the peak position, area and height of the measured peaks in the LC-MS data. It does not tell use to which metabolite a peak corresponds. Thus, we need to assign metabolite names to every peak in the peak table. This is still a difficult step and in practice most metabolites cannot be identified and would require more precise or additional experiments. However, the most straightforward approach is to match each m/z value against a (public) database of metabolites of which m/z values are calculated. One such database is Metlin (http://metlin.scripps.edu/). Figure 9.20 shows a few examples from the Metlin database.

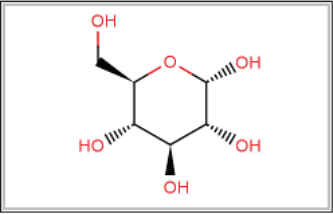
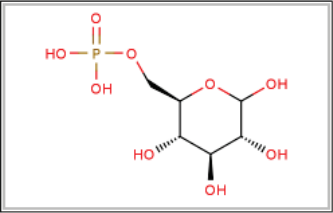
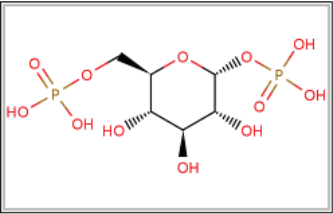
MID	Mass	Name	Formula	CAS	KEGG	MS/MS	Structure
133	180.0634	α-D-Glucose	$C_6H_{12}O_6$	492-62-6	C00267	View	
145	260.0297	D-Glucose 6-phosphate	$C_6H_{13}O_9P$	56-73-5	C00092	View	
154	339.9961	alpha-D-Glucose 1,6-biphosphate	$C_6H_{14}O_{12}P_2$	10139-18-1	C01231	NO	

Figure 9.20. Example entries from the Metlin database. Peaks identified by the peak identification method can be matched against this database.

9.4 Statistical analysis and visualization

One all peaks (metabolites) are identified we can compare different groups of samples. For example, in a study where data was obtained for six KO mice and six WT mice we can use a

statistical analysis to identify all metabolites that are differentially 'expressed' between these two groups. The results of such analysis are shown in Figure 9.21 and Figure 9.22.

name	fold	tstat	pvalue	mzmed	mzmin	mzmax	rtmed	rtmin	rtmax	npeaks	KO	WT
M300T3390	5.69359	14.4437	5.03E-08	300.19	300.171	300.2	3390.32	3386.76	3396.33	12	6	6
M301T3390	5.87659	15.5757	6.71E-08	301.188	301.166	301.195	3389.63	3386.76	3392.1	7	6	1
M298T3187	3.87092	11.9389	3.31E-07	298.151	298.105	298.159	3186.8	3184.12	3191.31	4	4	0
M491T3397	24.9757	16.8399	4.46E-06	491.2	491.188	491.206	3397.16	3367.12	3424.68	6	6	0
M348T3288	9.00507	17.2627	5.03E-06	348.162	348.123	348.174	3288.25	3284.66	3294.13	6	6	0
M423T3257	6.24617	10.8189	4.71E-05	423.15	423.106	423.157	3256.61	3254.85	3261.53	6	6	0
M327T3419	26.7484	11.0506	9.44E-05	327.199	327.168	327.2	3419.47	3412.92	3427.23	6	6	0
M326T3417	15.5444	10.4239	0.00013	326.2	326.17	326.2	3417.01	3411.31	3425.62	12	6	6
M410T3938	6.78734	9.11467	0.00017	410.266	410.212	410.299	3937.7	3932.87	3946.14	9	6	3

ko15	ko16	ko18	ko19	ko21	ko22	wt15	wt16	wt18
4534354	4980914	5290739	4564263	4733236	3931593	349661	491793	64552
962353	1047934	1109303	946943	984787	806171	86450.4	120097	1430(
180781	203927	191016	190627	156869	220289	16269.1	43677.8	54739.
432037	332159	386967	334951	294816	373578	7643.14	10519.9	26472.
165831	183665	150845	134637	136452	167008	24302.8	16631.4	19213.
236250	255169	212711	180691	191747	152861	29530.2	17037.2	35133.
1108851	950127	674223	677091	772290	1013978	58898.7	21991	2764
4809521	3931305	2913712	2819101	3284987	4346410	259229	314154	1619(

Figure 9.21. Part of the results from peak-identification, peak matching, time alignment and filling in missing peaks. This is a peak table that lists all identified peaks in 6 KO and 6 WT mice samples. The raw peak intensities are shown in the lower panel. A t-test is applied to compare the KO and WT mice for each peak. This results in a p-value that indicates if a result is significant (e.g., $p<0.01$) or not.

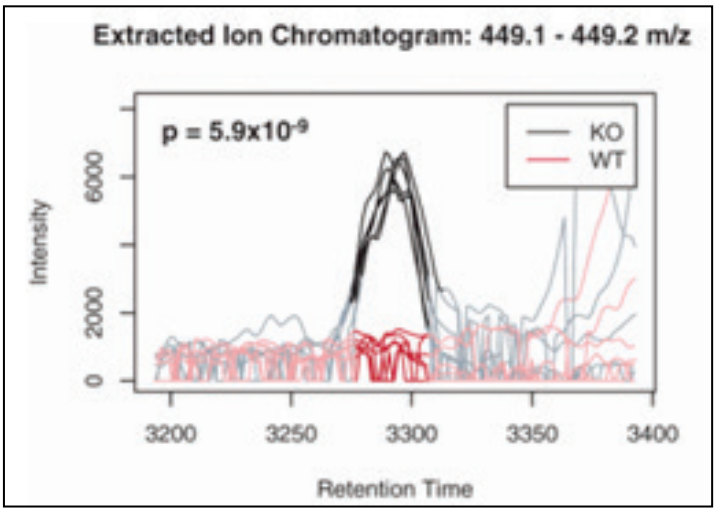


Figure 9.22. Visualization of chromatogram for one of the identified peaks (mz=449.1 / 449.2). The black curves show the data for the KO mice, while the red curves show the data for the WT mice. Clearly the peak is absent in the WT mice. Consequently, the t-test that compares KO and WT for this metabolite results in a highly significant p-value.

10 Introduction to Systems Biology

Lecturer: prof. dr. Antoine van Kampen (AMC)

After reading this chapter you should understand

- Definition of a system and systems biology
- Why systems biology is difficult and why multidisciplinary communication is important
- The concept of spatial and temporal scales; multiscale modeling; integrative systems biology
- Bottom-up and top-down systems biology
- Mechanistic and phenomenological models
- Different types of mathematical/statistical models (interaction networks, constrained based, mechanistic)

Note that the first few sections are a long introduction to systems biology to get you acquainted with a few basic ideas and terminology. Section 10.7 and beyond discuss some topics in more detail. Also have a look again at Chapter 1 of this syllabus.

Contents

10.1 Biological systems

When we think of biological systems, our minds may immediately wander to the Amazon rainforest, brimming with thousands of plants and animals that live with each other, compete with each other, and depend on each other. We might think of the incredible expanse of the world's oceans, of colorful fish swimming through coral reefs, nibbling on algae. Two-meter-high African termite mounds may come to mind, with their huge colonies of individuals that have their specific roles and whose lives are controlled by an intricate social structure. These examples are indeed beautiful manifestations of some of the fascinating systems nature has evolved. However, we don't have to look that far to find biological systems. Much, much smaller systems are in own bodies and even within our cells. Kidneys are waste-disposal systems. Mitochondria are energy-production systems. Ribosomes are intracellular machines that make proteins from amino acids. Bacteria are amazingly complicated biological systems. Viruses interact with cells in a well-controlled, systemic way. Even seemingly modest tasks often involve an amazingly large number of processes that form complicated control systems (see for example Figure 10.1). The more we learn about the most basic processes of life, such as cell division or the production of a metabolite, the more we have to marvel the incredible complexity of the systems that facilitate these processes. In our daily lives, we usually take these systems for granted and assume that they function adequately, and it is only when, for example, disease strikes that we realize how complex biology really is and how damaging the failure of just a single component can be.

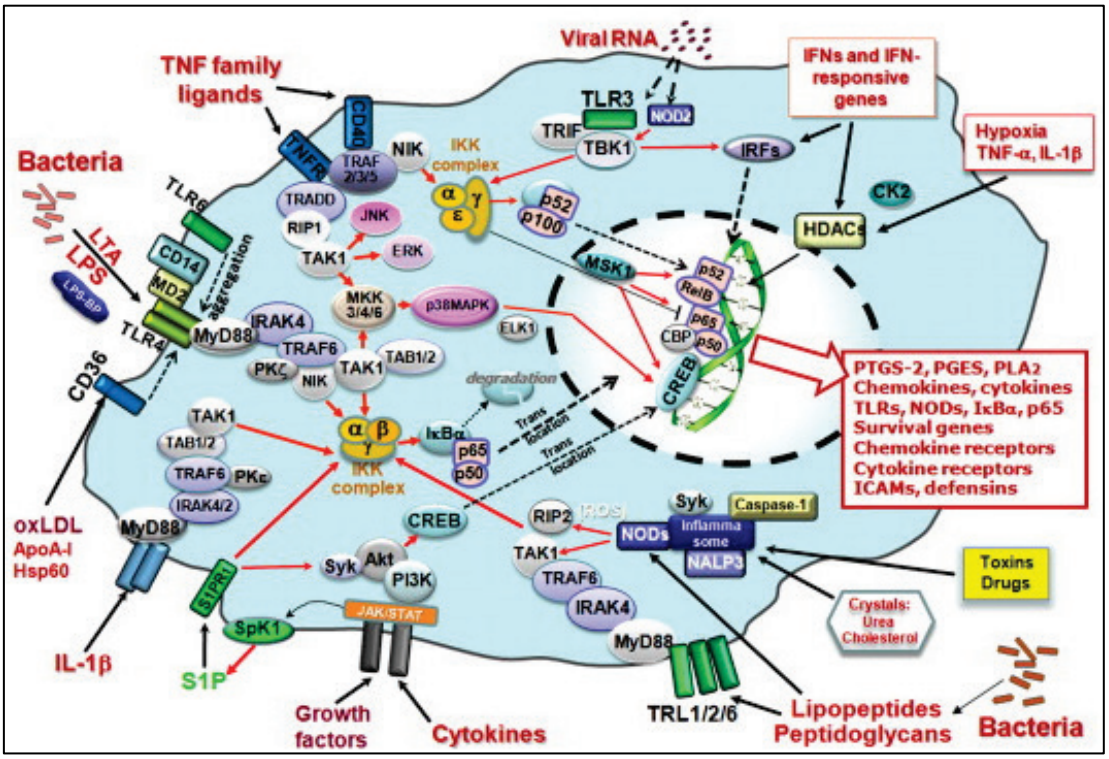


Figure 10.1. Major intracellular signaling pathways linking activation of the innate immune system with expression of inflammation-associated genes in human gestational membranes. Only a key subset of receptors, signaling pathways and gene products are shown. (Figure copied from Keelan (2011) Journal of Reproductive Immunology, 88(2), 176).

In spite of our long history of dealing with biological systems, our mastery of engineered systems far outstrips our capability to manipulate biological systems. We send spaceships successfully to faraway places and predict correctly when they will arrive and where they will land. We build skyscrapers exceeding by hundreds of times the sizes of the biggest animals and plants. Our airplanes are faster, bigger, and more robust against turbulence than the most skillful birds. Yet, we cannot create new human cells or tissues from basic building blocks and we are seldom able to cure diseases except with rather primitive methods like cutting into the body or killing a lot of healthy tissue in the process, hoping that the body will heal itself afterwards. We can anticipate that our grandchildren will only shake their heads at such medieval-sounding, draconian measures. We have learned to create improved microorganisms, for instance for the bulk production of industrial alcohol or the generation of pure amino acids, but the methods for doing so rely on bacterial machinery that we do not fully understand and on artificially induced random mutations rather than targeted design strategies.

Before we discuss the roots of the many challenges associated with understanding and manipulating biological systems in a targeted fashion, and our problems predicting what biological systems will do under yet untested conditions, we should ask whether the goal of a deeper understanding of biological systems is even worth the effort. The answer is a resounding "Yes!" Not only to satisfy our intellectual curiosity, but also because the potential and scope of advances that might develop from biological systems analyses. Applications that are already emerging on the horizon are personalized medical treatments with minimal side effects, pills that will let the body regain control over a tumor that has run amok, prevention and treatment of neurodegenerative diseases, and the creation of spare organs from reprogrammed stem cells. Reprogrammed microbes or nonliving systems composed of biological components will dominate the production of chemical compounds from prescription drugs to large-scale industrial organics. Modified viruses will become standard means for supplying cells with healthy proteins or replacement genes. The rewards of discovering and characterizing the general principles and the specifics of biological systems will truly be unlimited.

If it is possible to engineer very sophisticated machines and to predict exactly what they will do, why are biological systems so different and difficult? One crucial difference is that we have full control over engineered systems, but not over biological systems. As a society, we collectively know all details of all parts of engineered machines, because we made them. We know their properties and functions, and we can explain how and why some engineer put a machine together in a particular fashion. Furthermore, most engineered systems are modular, with each module being designed for a unique, specific task. While these modules interact with each other, they seldom have multiple roles in different parts of the system, in contrast to biology and medicine, where, for instance, the same lipids can be components of

membranes and have complicated signaling functions, and where diseases are often not restricted to a single organ or tissue, but may affect the immune system and lead to changes in blood pressure and blood chemistry that secondarily cause kidney and heart problems. A chemical refinery looks overwhelmingly complicated to a layperson, but for an industrial engineer, every piece has a specific, well-defined role within the refinery, and every piece or module has properties that were optimized for this role. Moreover, should something go wrong, the machines and factories will have been equipped with sensors and warning signals pinpointing problems as soon as they arise and allowing corrective action.

In contrast to dealing with sophisticated, well-characterized engineered systems, the analysis of biological systems requires investigations in the opposite direction. This type of investigation resembles the task of looking at an unknown machine and predicting what it does. Adding to this challenge, all scientists collectively know only a fraction of the components of biological systems, and the specific roles and interactions between these components are often obscure and change over time. Even more than engineered systems, biological systems are full of sensors and signals that indicate smooth running or ensuing problems (Figure 10.2), but in most cases our experiments cannot directly perceive and measure these signals and we can only indirectly deduce their existence and function. We observe organisms, cells, or intracellular structures as if from a large distance and must deduce from rather coarse observations how they might function or why they fail.

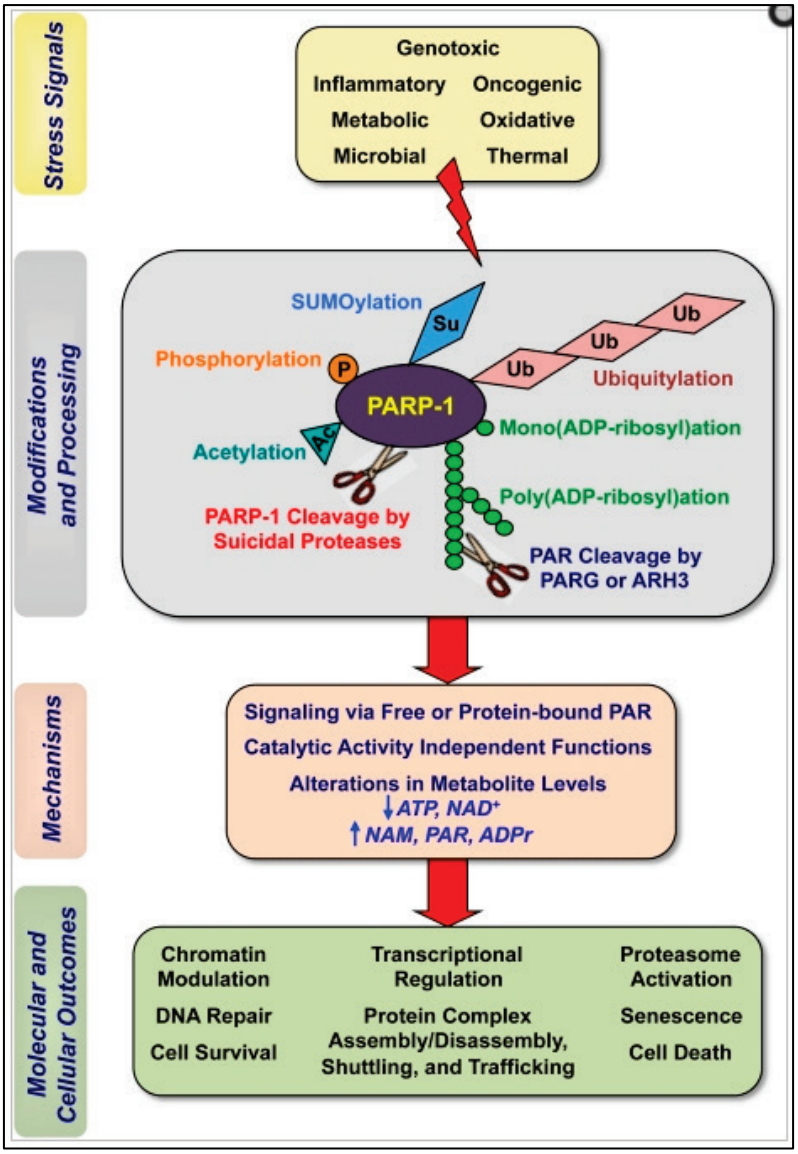


Figure 10.2. PARP-1 at the crossroads of cellular stress responses. Cells are regularly exposed to a wide variety of extrinsic and intrinsic stress signals, including those initiated by oxidative, nitrosative, genotoxic, oncogenic, thermal, inflammatory, and metabolic stresses. PARP-1 senses these stresses and propagates different stress signals to execute diverse downstream molecular and cellular actions. PARP-1 function at the intersection of converging stress signaling pathways. (Ac) Acetylation; (P) phosphorylation; (Su) SUMOylation; (Ub) ubiquitylation; (NAM) nicotinamide. (Figure copied from Luo (2012) Genes & Development, 26(5), 417)

What exactly is it that makes biological systems so difficult to grasp? It is certainly not just size. The size of a (biological) network or system is not necessarily correlated with its complexity. Biological systems often consist of large numbers of components, but they pose an additional, formidable challenge to any analysis because the processes that govern them are not linear. This is a problem, because we are trained to think in linear ways: if an investment of $100 leads to a return of$120, then an investment of $10,000 leads to a return of $12,000. Biology is different. If we fertilize our roses with 1 tablespoon of fertilizer and the rose bushes produce 50 blossoms, a little bit more fertilizer may increase the number of blossoms, but 100 tablespoons of fertilizer will not produce 5000 blossoms but almost certainly kill the plants. Now imagine that thousands of components, many of which we do

not know, respond in such a fashion, where a small input does not evoke any response, more input evokes a physiological response, and a little bit more input causes the component to fail or exhibit a totally different "stress" response. In addition, many components of a biological system (indirectly) interact. This makes the understanding of biological systems (non-linear coupled systems) so difficult.

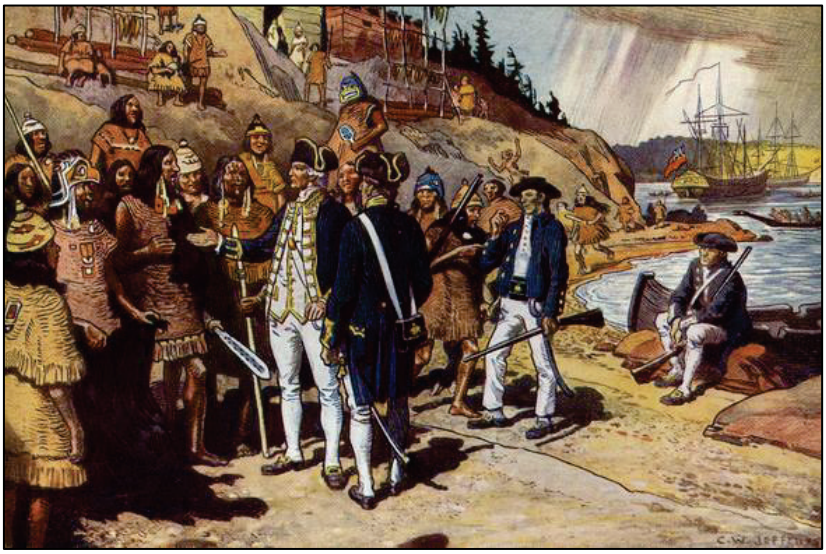


Figure 10.3. Collecting information is the first step in most systems analyses. The eighteenth-century Britisch explorer Captain James Cook sailed the Pacific Ocean and catalogued many plants and animal species that had never been seen before in Europe. (Figure copied from http://www.cmhg.gc.ca/cmh/page-331-eng.asp).

10.2 Reductionism and systems biology

So the situation is complicated. But because we humans are a curious species, our forebears did not give up on biological analysis and instead did what was doable, namely collecting information on whatever could be measured with the best current methods (Figure 10.3). By now, this long-term effort has resulted in an amazing list of biological parts and their roles. Initially, this list contained new plant and animal species, along with descriptions of their leaves, berries, and roots, or their body shapes, legs, and color patterns. These external descriptions were valuable, but did not provide specific clues on how plants and animals function, why they live, and why they die. Thus, the next logical step was to look inside-even if this required stealing bodies from the cemetery under a full moon! Cutting bodies open revealed an entirely new research frontier. What were all those distinct body parts and what did they do? What were organs, muscles, and tendons composed of? Not surprisingly, this line of investigation eventually led to the grand-challenge quest of discovering and measuring all parts of a body, the parts of the parts (... of the parts), as well as their roles in the normal physiology or pathology of cells, organs, and organisms. The implicit assumption of this reductionist approach was that knowing the building blocks of life would lead us to a comprehensive understanding of how life works. If we fast-forward to the twenty-first century, have we succeeded and assembled a complete parts catalog? Do we know the

building blocks of life? The answer is a combination of yes's and no's. The catalog is most certainly not complete, even for relatively simple organisms. Yet, we have discovered and characterized genes, proteins, and metabolites as the major building blocks.

The sequencing of the human genome was without any doubt an incredible achievement. But there is much more to a human body than genes. So, the race for building blocks extended to proteins and metabolites, toward individual gene variations and an assortment of molecules and processes affecting gene expression, which changes in response to external and internal stimuli, during the day, and throughout our lifetimes. As a direct consequence of these ongoing efforts, our parts list continues to grow at a rapid pace: A parts catalog that started with a few organs now contains over 20,000 human genes, many more genes from other organisms, and hundreds of thousands of proteins and metabolites along with their variants. In addition to merely looking at parts in isolation, we have begun to realize that most biological components are affected and regulated by a variety of other components. The expression of a gene may depend on several transcription factors, metabolites, and a variety of small RNAs, as well as molecular, epigenetic attachments to its DNA sequence. It is reasonable to expect that the list of processes within the body is much larger than the number of components on our parts list. Biologists will not have to worry about job security any time soon!

The large number of components and processes alone, however, is not the only obstacle to understanding how cells and organisms function. We can make very accurate predictions regarding a gas in a container, even if trillions of molecules are involved. If we increase the pressure on the gas without changing the volume of the container, we know that the temperature will rise, and we can predict by how much. Not so with a cell or organism. What will happen to it if the environmental temperature goes up? Nothing much may happen, the rise in temperature may trigger a host of physiological response processes that compensate for the new conditions, or the organism may die. The outcome depends on a variety of factors that collectively constitute a complex stress response system. Of course, the comparison to a gas is not quite fair because, in addition to their large number, the components of a cell are not all the same, which drastically complicates matters. Furthermore, as mentioned earlier, the processes with which the components interact are nonlinear, and this permits an enormous repertoire of distinctly different behaviors with which an organism can respond to a perturbation.

10.3 Nonlinear biological systems

It is easy to demonstrate how quickly our intuition can be overwhelmed by a few nonlinearities within a system. As an example, let's look at a simple chain of processes and compare it with a slightly more complicated chain that includes regulation. The simple case merely consists of a chain of reactions, which is fed by an external input (**Figure 10.4**A). It

does not really matter what *X, Y*, and *Z* represent, but, for the sake of discussion, imagine a metabolic pathway such as glycolysis, where the input, glucose, is converted into glucose-6-phosphate, fructose-1,6-bisphosphate, and pyruvate, which is used for other purposes that are not of interest here. For illustrative purposes, let's explicitly account for an enzyme *E* that catalyzes the conversion of *X* into *Y*.

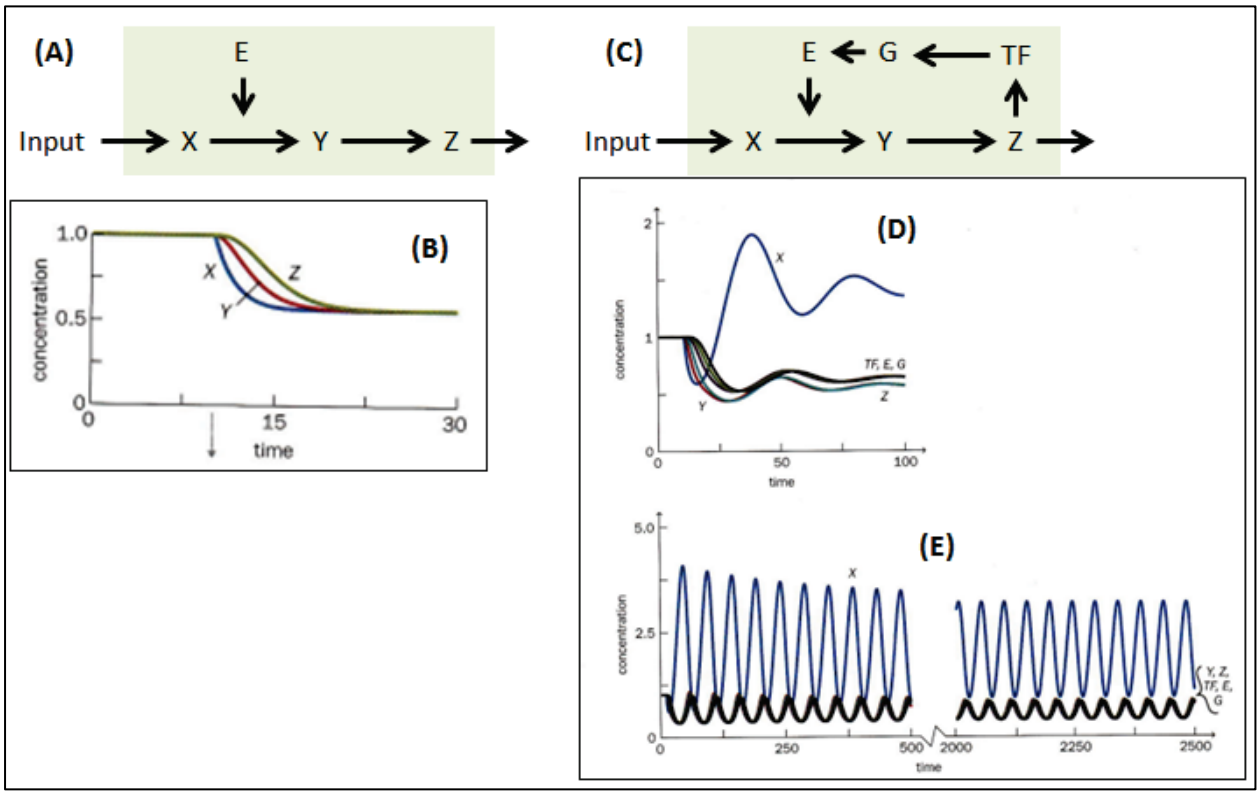


Figure 10.4. (A) In this simple pathway, an external input is converted sequentially into *X, Y*, and *Z*, which leaves the system. The conversion of *X* into *Y* is catalyzed by an enzyme *E*. It is easy to imagine that any increase in *input* will cause the levels of *X, Y*, and *Z* to rise. (B) Simulations with the system in confirm our intuition: reducing *input* in to 75% at time 10 (arrow) leads to permanent decreases in *X, Y*, and *Z*. (C) Even simple systems may not allow us to make reliable predictions regarding their responses to stimuli. Here, the linear pathway is embedded into a functional loop consisting of a transcription factor *TF* and a gene *G* that codes for enzyme *E*. The responses to changes in Input are no longer obvious. (D, E) Simulation results demonstrate that the looped system in may exhibit drastically different responses. (D) If the effect of *Z* on *TF* is relatively small, the functional feedback loop causes the system to go through damped oscillations before assuming a new stable state. (E) For stronger effects of *Z* on *TF*, the system response is a persistent oscillation.

We can formulate a model of such a pathway system as a set of differential equations. And while the details are not important here, it does not hurt to show such a model, which might read:

$$\dot{X} = \frac{dX}{dt} = input - aEX^{0.5}$$

$$\dot{Y} = \frac{dY}{dt} = aEX^{0.5} - bY^{0.5}$$

$$\dot{Z} = \frac{dZ}{dt} = bY^{0.5} - cZ^{0.5}$$

Here, *X, Y*, and *Z* are concentrations, *E* is the enzyme activity, and *a*, *b*, and *c* are rate constants that respectively represent how fast *X* is converted into *Y*, how fast *Y* is converted into *Z*, and how quickly material from the metabolite pool *Z* leaves the system. The dotted quantities on the left of the equal signs are differentials that describe the change in each variable over time, but we need not worry about them at this point. In fact, we hardly have to analyze these equations mathematically to get an idea of what will happen if we change the input, because intuition tells us that any increase in *input* should lead to a corresponding rise in the concentrations of the intermediates *X, Y*, and *Z*, whereas a decrease in Input should result in smaller values of *X*, *Y*, and *Z*. The increases or decreases in *X, Y*, and *Z* will not necessarily be exactly of the same extent as the change in Input, but the direction of the change should be the same. The mathematical solution of the above system of equations confirms this intuition. For instance, if we reduce *input* from 1 to 0.75, the levels of *X, Y*, and *Z* decrease, one after another, from their initial value of 1 to 0.5625 (Figure 10.4B).

Now suppose that *Z* is a signaling molecule, such as a hormone or a phospholipid that activates a transcription factor *TF* that facilitates the up-regulation of a gene *G* that codes for the enzyme catalyzing the conversion of *X* into *Y* (Figure 10.4C). The simple linear pathway is now part of a functional loop. The organization of this loop is easy to grasp, but what is its effect? Intuition might lead us to believe that the positive-feedback loop should increase the level of enzyme *E*, which would result in more *Y*, more *Z*, and even more *E*, which would result in even more *Y* and *Z*. Would the concentrations in the system grow without end? Can we be sure about this prediction? Would an unending expansion be reasonable? What will happen if we increase or decrease the input as before?

The overall answer will be surprising: the information given so far does not allow us to predict particular responses with any degree of reliability. Instead, the answer depends on the numerical specifications of the system. This is bad news for the unaided human mind, because we are simply not able to assess the numerical consequences of slight changes in a system, even if we can easily grasp the logic of a system as in Figure 10.4C.

To get a feel for the system, one can compute a few examples with an expanded model that accounts for the new variables (*G* and *TF*). Here, the results are more important than the technical details. If the effect of *Z* on *TF* is weak, the response to a decrease in *input* is essentially the same as in Figure 10.4B. This is not too surprising, because the systems in this case are very similar. However, if the effect of *Z* on *TF* is stronger, the concentrations in the system start to oscillate, and after a while these oscillations dampen away (Figure 10.4D). This behavior is not easy to predict. Interestingly, if the effect is further increased, the system enters a stable oscillation pattern that does not cease unless the system input is changed again (Figure 10.4E).

The hand-waving explanation of these results is that the increased enzyme activity leads to a depletion of *X*. A reduced level of *X* leads to lower levels of *Y* and *Z*, which in turn lead to a reduced effect on *TF*, *G*, and ultimately *E*. Depending on the numerical characteristics, the ups and downs in *X* may not be noticeable, they may be damped and disappear, or they may persist until another change is introduced. Intriguingly, even if we know that these alternative responses are possible, the unaided human mind is not equipped to integrate the numerical features of the model in such a way that we can predict which system response will ensue for a specific setting of parameters. A computational model, by contrast, reveals the answer in a fraction of a second.

The specific details of the example are not as important as the take-home message: If a system contains regulatory signals that form functional loops, we can no longer rely on our intuition for making reliable predictions. But essentially all realistic systems in biology are regulated-and not just with one, but with many control loops. This leads to the direct and sobering deduction that intuition is not sufficient and that we instead need to utilize computational models to figure out how even small systems work and why they might show distinctly different responses or even fail, depending on the conditions under which they operate.

The previous sections have taught us that biological systems contain large numbers of different types of **components that interact** in potentially complicated ways and are **controlled by regulatory signals**. What else is special about biological systems? Many answers could be given. For instance, two biological components are seldom 100% the same. They **vary** from one organism to the next and change over time. Sometimes these variations are inconsequential, at other times they lead to early aging and disease. In fact, most diseases do not have a single cause, but are the consequence of an unfortunate combination of slight alterations in many components. Another feature that complicates intuition is the **delay in many responses** to stimuli. Such delays may be of the order of seconds, hours, or years, but they require the analyst to study not merely the present state of a biological system but also its history. For instance, recovery from a severe infection depends greatly on the

preconditioning of the organism, which is the collective result of earlier infections and the body's responses.

Finally, it should be mentioned that different parts of biological systems may simultaneously operate at **different scales, with respect to both time and space**. These scales make some aspects of their analysis easier and some harder. Let's begin with the **temporal scale**. We know that biology at the most basic level is governed by physical and chemical processes. These occur on timescales of the order of milliseconds, if not faster. Biochemical processes usually run on a scale of seconds to minutes. Under favorable conditions, bacteria divide every 20-30 minutes. Our human lifespan extends to maybe 120 years, evolution can happen at the genetic level with lightning speed, for instance, when radiation causes a mutation, while the emergence of an entirely new species may take thousands or even millions of years. On one hand, the drastically different time scales make analyses complicated, because we simply cannot account for rapid changes in all molecules of an organism over an extended period of time. As an example, it is impossible to study aging by monitoring an organism's molecular state every second or minute. On the other hand, the differences in timescales justify a very valuable modeling "trick": if we are interested in understanding some biochemical process, such as the generation of energy in the form of adenosine triphosphate (ATP) by means of the conversion of glucose into pyruvate, we can assume that developmental and evolutionary changes are so slow in comparison that they do not change during ATP production. Similarly, if we study the phylogenetic family tree of species, the biochemical processes in an individual organism are comparatively so fast that their details become irrelevant. Thus, by focusing on just the most relevant timescale and ignoring much faster and much slower processes, any modeling effort is dramatically simplified.

Biology also happens on many **spatial scales**. All processes have a molecular component, and their size scale is therefore of the order of Ångströms and nanometers. If we consider a cell as the basic unit of life, we are dealing with a spatial scale of micrometers to millimeters (with some exceptions). As with the different temporal scales, and using analogous arguments, models of biological systems often focus on one or two spatial scales at a time. Nonetheless, such simplifications are not always applicable, and some processes, such as aging, may require the simultaneous consideration of several temporal and spatial scales. Such **multiscale** assessments are often very complicated and constitute a challenging frontier of current research.

10.4 Why now?

Many of the features of biological systems have been known for quite a while and, similarly, many concepts and methods of systems biology have their roots in its well-established parent disciplines, including physiology, molecular biology, biochemistry, mathematics, engineering, and computer science. In fact, it has been suggested that the nineteenth-century scientist

Claude Bernard might be considered the first systems biologist, as he proclaimed that the "*application of mathematics to natural phenomena is the aim of all science, because the expression of the laws of phenomena should always be mathematical*" [1, 2]. A century later, Ludwig von Bertalanffy reviewed in a book his three decades of attempting to convince biologists of the systemic nature of living organisms [3, 4]. At the same time, Mihajlo Mesarovic used the term "Systems Biology" and declared that "*real advance ... will come about only when biologists start asking questions which are based on systems-theoretic concepts*" [5]. The same year, a book review in Science envisioned "*... a field of systems biology with its own identity and in its own right*" [6]. A few years later, Michael Savageau proposed an agenda for studying biological systems with mathematical and computational means [7].

In spite of these efforts, systems biology did not enter the mainstream for several more decades. Biology kept its distance from mathematics, computer science, and engineering, primarily because biological phenomena were seen as too complicated for rigorous mathematical analysis and mathematics was considered applicable only to very small systems of little biological relevance. The engineering of biological systems from scratch was impossible.

So, why has systems biology all of the sudden moved to the fore? Any good detective will know the answer: motive and opportunity. The motive lies in the realization that reductionist thinking and experimentation alone are not sufficient if complex systems are involved. Reductionist experiments are very good in generating detailed information regarding specific components or processes of a system, but they often lack the ability to characterize, explain, or predict emergent properties that cannot be found in the parts of the system but only in their web of interactions. For instance, the emergence of oscillations in the example system represented by the equations above cannot be credited to a single component of the system but is a function of its overall organization. Although we had complete knowledge of all details of the model pathway, it was very difficult to foresee its capacity either to saturate or oscillate in a damped or stable fashion. Biology is full of such examples.

A few years ago, the assembly of a complete catalogue of single mutants in the bacterium *Escherichia coli* was completed. Yet, the scientific community is still not able to foresee which genes the bacterium will up or down-regulate in response to new environmental conditions. Another very challenging example of emergent system properties is the central nervous system. Even though we understand quite well how action potentials are generated and propagated in individual neurons, we do not know how information flows, how memory works, and how diseases affect the normal functioning of the brain. It is not even clear how information in the brain is represented. Thus, while reductionist biology has been extremely successful and will without any doubt continue to be the major driving force for future discovery, many biologists have come to recognize that the detailed pieces of information

resulting from this approach need to be complemented with new methods of system integration and reconstruction.

The opportunity for systems biology is the result of the recent confluence and synergism of three scientific frontiers:

- **Data.** The first is of course the rapid and vast accumulation of detailed biological information at the physiological, cellular, molecular, and submolecular levels. These targeted investigations of specific phenomena are accompanied by large-scale, high-throughput (omics) studies that were entirely infeasible just a couple of decades ago. They include quantification of genome-wide expression patterns, simultaneous identification of large arrays of expressed proteins, comprehensive profiling of cellular metabolites, characterization of networks of molecular interactions, global assessments of immune systems, and functional scans of nervous systems and the human brain. These exciting techniques are generating unprecedented amounts of high-quality data that are awaiting systemic interpretation and integration (Figure 10.5A).

- **Technologies**. The second frontier is the result of ingenuity and innovation in engineering, chemistry, and material sciences, which have begun to provide us with a growing array of technologies for probing, sensing, imaging, and measuring biological systems that are at once very detailed, extremely specific, and usable in vivo. Many tools supporting these methods are in the process of being miniaturized, in some cases down to the nanoscale of molecules, which allows diagnoses with minute amounts of biological materials and one day maybe biopsies of individual, living cells. Devices at this scale will allow the insertion of sensing and disease treatment devices into the human body in an essentially noninvasive and harmless fashion. It is even becoming feasible to use molecular structures, prefabricated by nature, for new purposes in medicine, drug delivery, and biotechnology (Figure 10.5B).

- **Computation**. The third frontier is the co-evolution of mathematical, physical, and computational techniques that are more powerful and accessible to a much wider audience than ever before. Imagine that only a few decades ago computer scientists used punch cards that were read by optical card readers (Figure 10.5C). Now, there are even specific computing environments, including R, Matlab, Maple and Mathematica, as well as different types of customized mark-up languages (XML), such as the systems biology mark-up language SBML (see section 10.7.6)

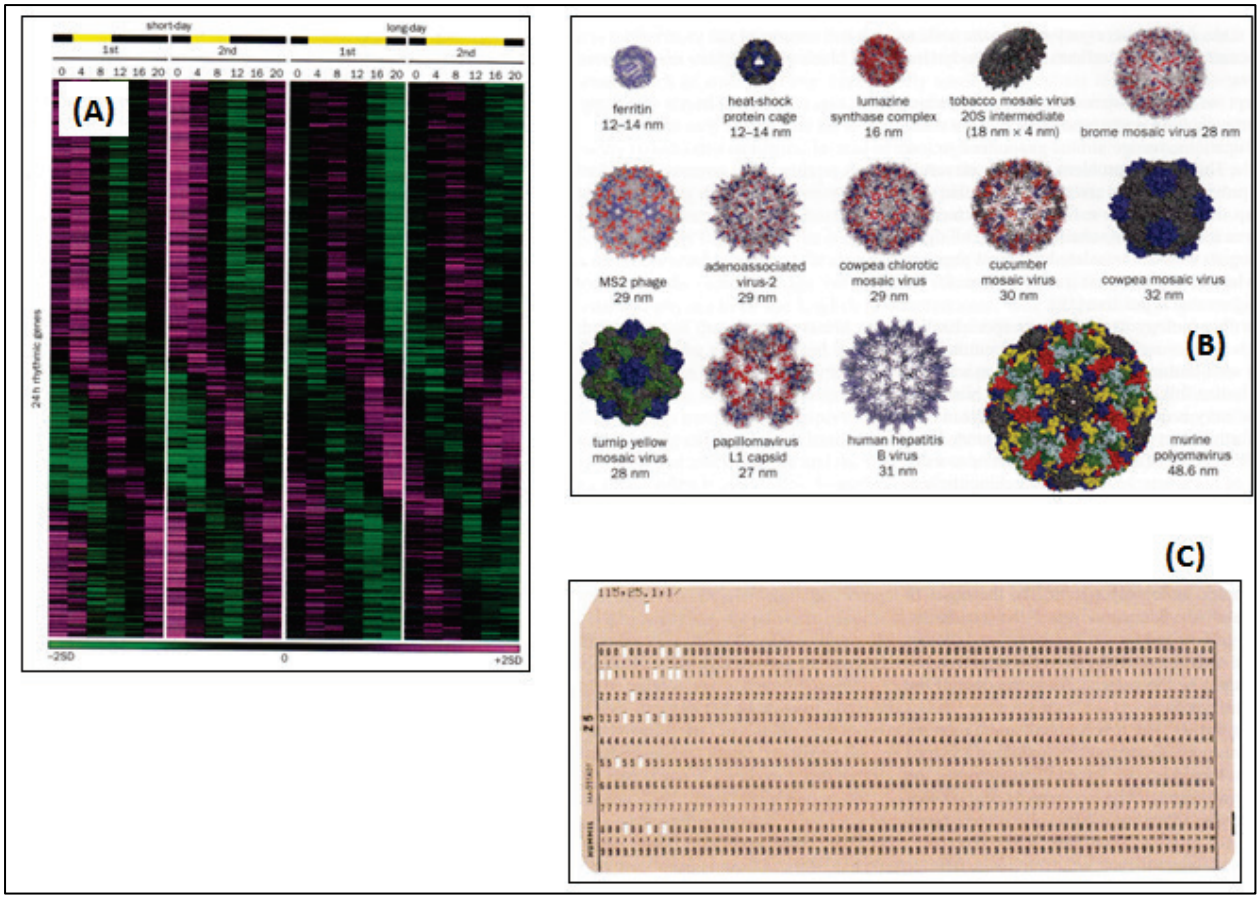


Figure 10.5. (A) Modern high-throughput methods of molecular biology offer data in unprecedented quantity and quality. As an example, the heat map shown here represents a genome-wide expression profile of 24-hour-rhythmic genes in the mouse under chronic short-day and long-day conditions. (From Masumoto et al (2010) Curr. Biol, 20, 2199). (B) "Protein cages" are particles that have applications in bionanotechnology and nanomedicine. These particles are very interesting biological building blocks because they self-assemble into a variety of different shapes. The features of these bionanoparticles can be genetically manipulated and fine-tuned for biomedical purposes, such as drug delivery, gene therapy, tumor imaging, and vaccine development (From Lee et al (2006) Nanomedicine, 2, 137). (C) Advances in computer power, accessibility, and user-friendliness over the past 30 years have been tremendous. Not too long ago, computer code had to be fed manually into the computer with punch cards.

Before today's much more effective computer science techniques were available, it was not even possible to keep track of the many components of biological systems, let alone analyze them. But over the past few decades, a solid theoretical and numerical foundation has been established for computational methods specifically tailored for the investigation of dynamic and adaptive systems in biology and medicine. These techniques are now at the verge of making it possible to represent and analyze large, organizationally complex systems and to study their emergent properties in a rigorous fashion. Methods of machine learning, numerical mathematics, and bioinformatics permit the efficient analysis of the data. Algorithmic advances permit the simulation and optimization of very large biological flux distribution networks. Computer-aided approximation approaches yield ever-finer insights into the dynamics of complex nonlinear systems, such as the control of blood flow in healthy and diseased hearts. New mathematical, physical, and computational methods are beginning

to make it possible to predict the folding of proteins and the binding between target sites and ligands. These predictions, in turn, suggest insights into specific molecular interactions and promise the potential of targeted drug interventions that minimize toxic side effects.

Motive and opportunity have met to make systems biology attractive and feasible. It has become evident that the relevant disciplines complement each other in unique ways and that the synergism among them will revolutionize biology, medicine, and a host of other fields, including biotechnology, environmental science, food production, and drug development.

10.5 Communicating systems biology

It is not a trivial task to talk succinctly about 25,000 genes and their expression state or about the many processes occurring simultaneously in response to a signal that a cell receives at its outer surface. Our minds are ill equipped to characterize numerical relationships, let alone discuss complicated mathematical functions, especially if these depend on many variables.

We may willingly or grudgingly accept the fact that we need mathematics, which comes with its own terminology, but communication is a two-way process. If system biologists start talking about eigenvalues and Hopf bifurcations, we are almost guaranteed to lose mainstream biologists, let alone lay people. This is a real problem, because systems biology results must be conveyed to biologists, who are providing the data, and to the public that pays for our research. The only true solution to this challenge is the bilingual education and nurturing of systems biologists who can translate biological phenomena into math and computer code and who can really explain the biology reflected by results from mathematical analysis.

Communication is not trivial even within biology itself, because specialization has progressed so far that different fields such as molecular biology, immunology, and nanomedicine have developed their own terminology and jargon. Let's look at this issue in the form of a parable from Indian folklore that describes six blind men exploring an elephant (Figure 10.6). The story has it that each of the blind men touched a different part of an elephant and came to a different conclusion concerning the object of his research. The man touching the side thought he was touching a wall, the one feeling the leg concluded he was touching a tree. The elephant's trunk gave the impression of a snake, the tusk that of a pointed spear, the tail felt like a rope, and the ear appeared to be like a large leaf or fan. It is not difficult to see the analogy to a complex biological system like the onset of Alzheimer's disease. The first scientist found "*the Alzheimer gene*", the second discovered "*a strong association between the disease and former head injuries*", another scientist detected "*problems with fatty acid metabolism in the brain*", and yet another suggested that "*aluminum in cookware might be the culprit*". As in the case of the elephant, the scientists were right, to some degree.

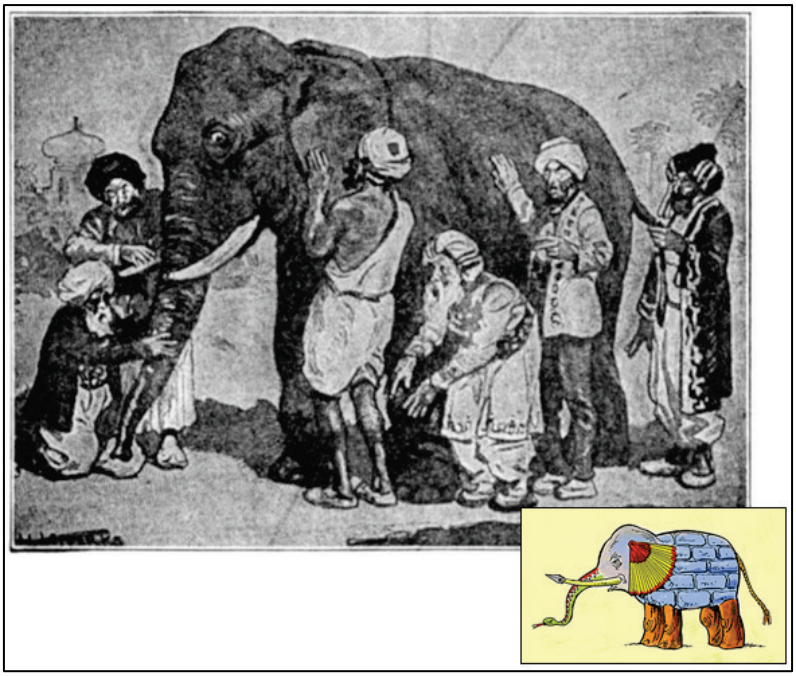


Figure 10.6. Information about isolated parts of a system alone does not always reveal the true nature of the system. An old story of six blind Indian men trying to determine what they touch is a parable for the dangers of scientific silos and the lack of good communication. (Figure copied from http://maddy06.blogspot.nl/2012/10/six-blind-men-and-elephant.html and http://wordinfo.info/unit/1).

Let's analyze the elephant story a little further. The first problem among the six blind men might have been the homogeneous pool of researchers. Including a female or a child might have provided additional clues. Also, we have to feel sorry for the Indian men for being blind. However, they were apparently not mute or deaf, so that a little discussion among them might have gone a long way. While all six were blind, it is furthermore fair to assume that they had friends with working vision, who could have set them straight. They could have used not just their hands but also their other senses, such as smell. Do tree trunks really smell like elephant feet? Finally, they apparently stayed in their one spot, thereby greatly limiting their experience base.

It is again easy to translate these issues into biology, especially when we think of purely reductionist strategies. Instead of a homogeneous pool of biologists analyzing biological systems, it is without doubt more effective to have a multidisciplinary team including different varieties of biologists, but also physicists, engineers, mathematicians, and chemists. Instead of only focusing on the one aspect right in front of our nose, communication with others provides context for singular findings. We don't know whether the Indian men spoke the same language, but we know that even if biologists, computer scientists, and physicists all use English to communicate, their technical languages and their views of the scientific world are often very different, so that communication may initially be superficial and ineffective. That is where multidisciplinary groups must engage in learning new terminologies and languages and include interdisciplinary translators. Just as the Indian men should have called upon their

seeing friends, investigators need to call in experts who master techniques that had not been applied to the biological problem at hand. The established scientific disciplines in the past have often become silos. Sometimes without even knowing it, researchers kept themselves inside these silos, unable or unwilling to break out and to see the many other silos around, as well as a whole lot of space between them.

Systems biology does not ask the six blind men to abandon their methods and instead to run circles around the elephant. By focusing on one aspect, the reductionist "elephantologists" are poised to become true experts on their one chosen body part and to know everything there is to know about it. Without these experts, systems biology would have no data to work on. Instead, what systems biology suggests is a widening of the mindset and at least rudimentary knowledge of a second language, such as math. It also suggests the addition of other researchers, assisting the reductionist by developing new tools of analysis, by telling them in their language what others had found, by closing the gap between the parts.

Well-trained systems biologists should be able to develop strategies for merging heterogeneous information into formal models that permit the generation of testable hypotheses. These hypotheses may be wrong, but they can nevertheless be very valuable, because they focus the scientific process on new, specific experiments that either confirm or refute the hypothesis.

The story tells us that effective communication can solve a lot of complex questions. In systems biology, such communication is not always easy and requires not only mastering the terminology of several parent disciplines but also internalizing the mindset of biologists and clinicians on the one hand and of mathematicians, computer scientists, and engineers on the other. So, let's learn about biology. Let's study laboratory data and information and explore the mindset of biologists. Let's study graphs and networks with methods from computer science. Let's see how mathematicians approach a biological system, struggle with assumptions, make simplifications, and obtain solutions that are at first incomprehensible to the non-mathematician but do have real meaning once they are translated into the language of biology.

10.6 Why do we need to understand biological systems

What does that really mean to understand biological systems? Generically, it means that we

1. should be able to explain how biological systems work and why they are constructed in the fashion as we observe them and not in a different one.
2. we should be able to make reliable predictions of responses of biological systems under yet-untested conditions.
3. we should be able to introduce targeted manipulations into biological systems that change their responses to our specifications.

10.7 Definitions, concepts and terminology

The sections above have defined biological systems and (the aims) of systems biology in general. In the sections below we will briefly summarize and discuss some of the concepts in more detail.

10.7.1 Definition of systems biology

Systems biology is a biology-based inter-disciplinary field of study that focuses on complex (dynamic) non-linear interactions within biological systems (e.g., pathways, cells, organs, organisms), using a more holistic perspective, instead of the more traditional reductionism approach, to biological and biomedical research. It involves a combination of wet-lab experiments and computational approaches. It brings together experimental information about the interplay of the components of the systems in quantitative mathematical models that generate testable hypothesis, allow prediction of the behavior of such systems, and discover emergent properties.

A typical systems biology cycle is shown in Figure 10.7. Given a new hypothesis a properly designed wet-lab experiment is performed to generate new data. This data, in conjunction with data that was previously generated and/or information from literature, is used to construct new mathematical/statistical models that can be analyzed to obtain new biological insights in the system under investigation. In a next iteration of the cycle, these insights can then be validated in new experiments followed by model refinement.

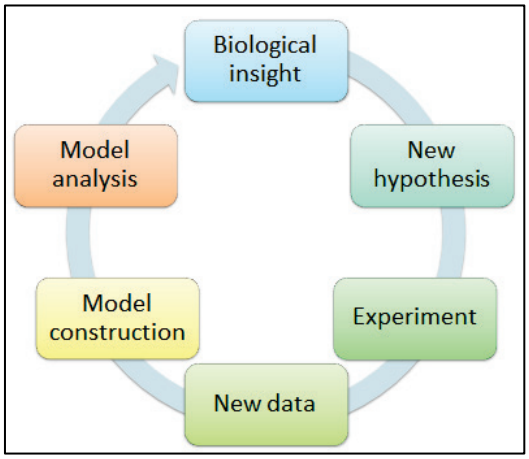


Figure 10.7. The systems biology cycle.

10.7.2 Top-down and bottom-up systems biology

We may distinguish between top-down and bottom-up systems biology approaches (Figure 10.8)

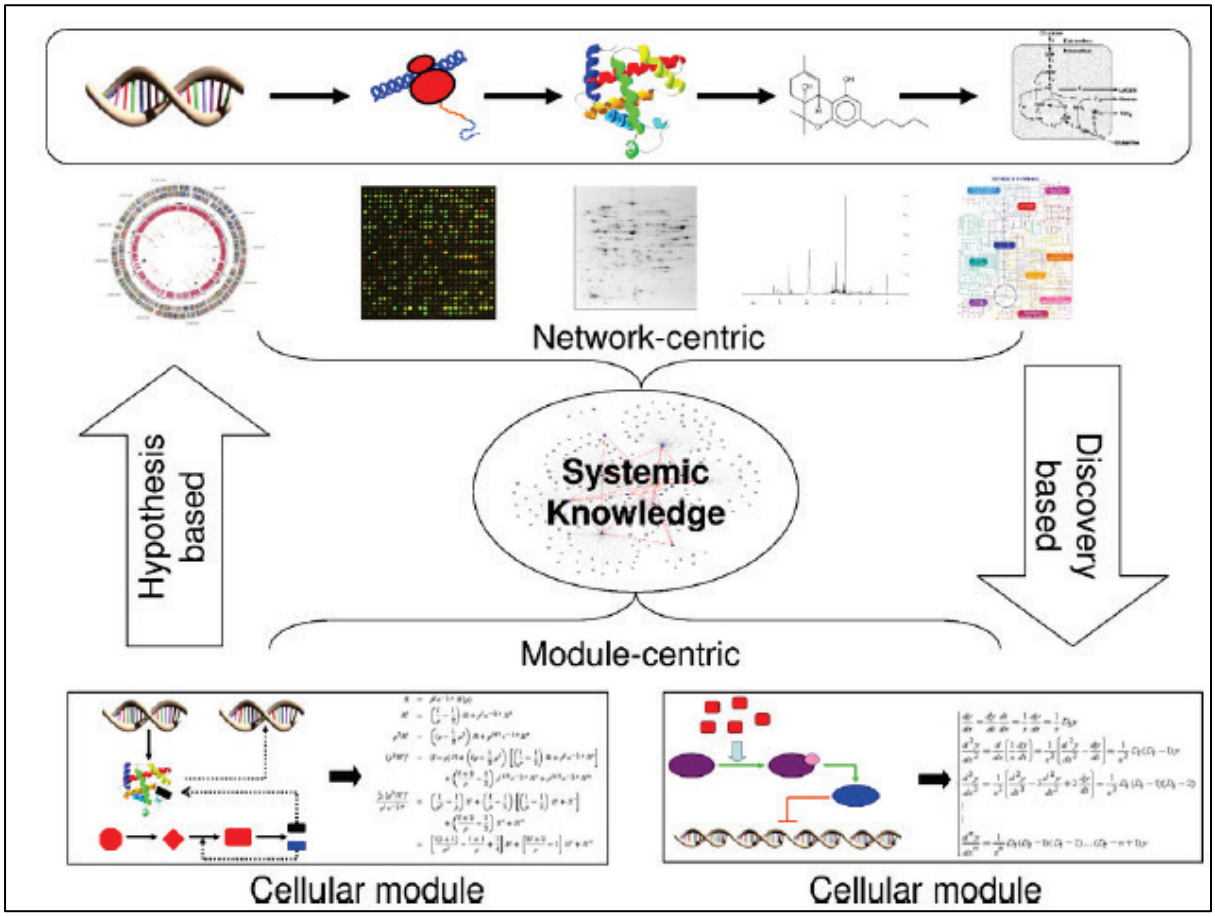


Figure 10.8. Top-down and bottom-up approaches to systems biology. The bottom-up approach combines well-defined mechanistic models (cellular modules described by mathematical equations) to describe the system of interest. The top-down approach starts from omics data to produce genome-wide networks of e.g., genes (e.g., correlations between genes) and proteins (e.g., protein interactions). (Figure copied from Petranovic (2009) J. Biotechnology, 144(3), 204).

Top-down systems biology

With the introduction of 'omics', the top-down approach emerged as a dominant method. It starts from a bird's eye view of the behavior of the system (from the top or the whole) by measuring genome-wide experimental data, and aims to discover and characterize biological mechanisms closer to the bottom (i.e., the parts and their interactions). It concerns the 'reverse engineering' of the structure of the molecular network underlying the system from genome-wide data sets. In top-down systems biology, the main objective is to obtain a better understanding of the relation between the components of the system from experimental data, (statistical) data analysis and data integration. Top-down approaches lead to new insights in relations between genes, proteins, and/or metabolites but don't reveal precise biological mechanisms but, generally, results in hypotheses concerning co- and inter-regulation of groups of those molecules. Of course, such hypotheses can eventually result in a mechanistic model through complementary experimental and computational approaches.

These hypotheses then predict new correlations, which can be tested in new rounds of experiments or by further biochemical analyses. The major strengths of top-down systems biology are that it is potentially complete (genome-wide) and that it addresses the metabolome, transcriptome, proteome and/or other ome's.

Figure 10.9 shows how we can construction gene co-expression networks from microarray data to identify relation between genes. The resulting co-expression network shows that genes A, B and C are connected, which implies that the expression levels of gene A, B and C are similar in the four samples (i.e., the expression profiles are correlated). However, from such analysis it remains unclear if, for example, these three genes are co-regulated by the same transcription factor. The reveal such mechanism, additional experiments are required.

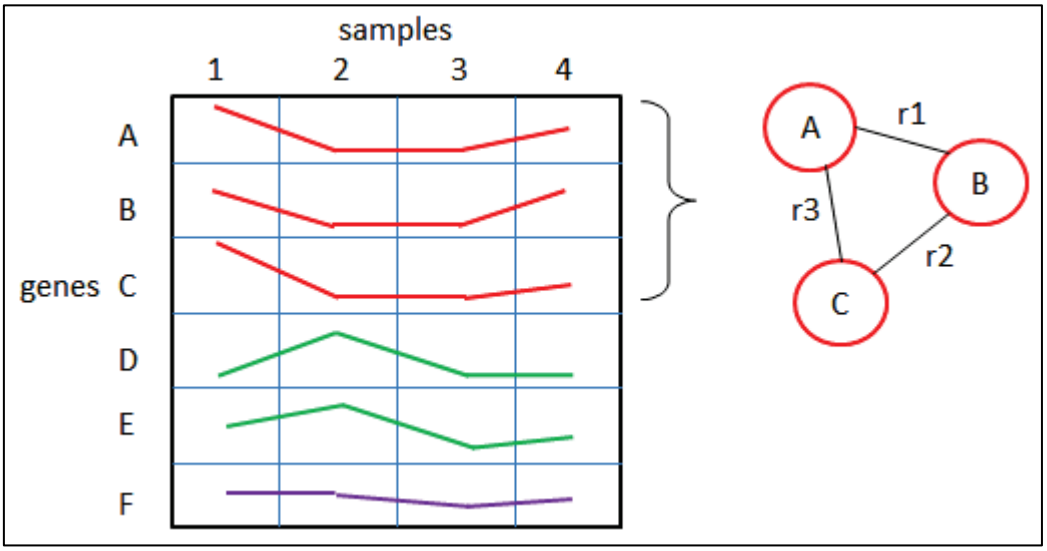


Figure 10.9. Construction of a co-expression network from microarray data. In this simplified example, the expression levels of 6 genes (A-F) have been measured in 4 samples. The expression levels are represented by the colored gene expression profiles. Profiles with a similar shape have the same color. Three groups of profiles are distinguished (red, green and purple). The profiles in such group have a high correlation, which can be expressed by a correlation coefficient (r) between pairs of profiles. To construct a gene co-expression network, the gene pairs that are highly correlated are connected. One co-expression network is shown for the three genes in in first group.

Figure 10.10 shows an example of a co-expression network that was constructed from a DNA microarray experiment (gene expression) in a study that attempted to identify genes underlying schizophrenia. This co-expression network is significantly enriched with brain-expressed genes and with genetic variants that were implicated in a previous GWAS study, which could imply a causal role in schizophrenia etiology. The most highly connected intramodular hub gene in this module turned out to be ABCF1. Interpretation of such co-expression networks is, in general, difficult but may lead to new insights in the underlying molecular mechanisms of disease.

Models resulting from top-down strategies often are 'interaction' models such as the co-expression networks. These models reveal functional of physical interactions between its

components but do not reveal the precise nature of such interaction. These models are **phenomenological** and not based on real biological mechanisms. A phenomenological model represents a hypothesized relationship between the variables in the data set (e.g., co-regulation of genes by the same transcription factors) [8]. In contrast **mechanistic models** represent relationships in terms of biological processes that are thought to have given rise to the data. An example, is shown below.

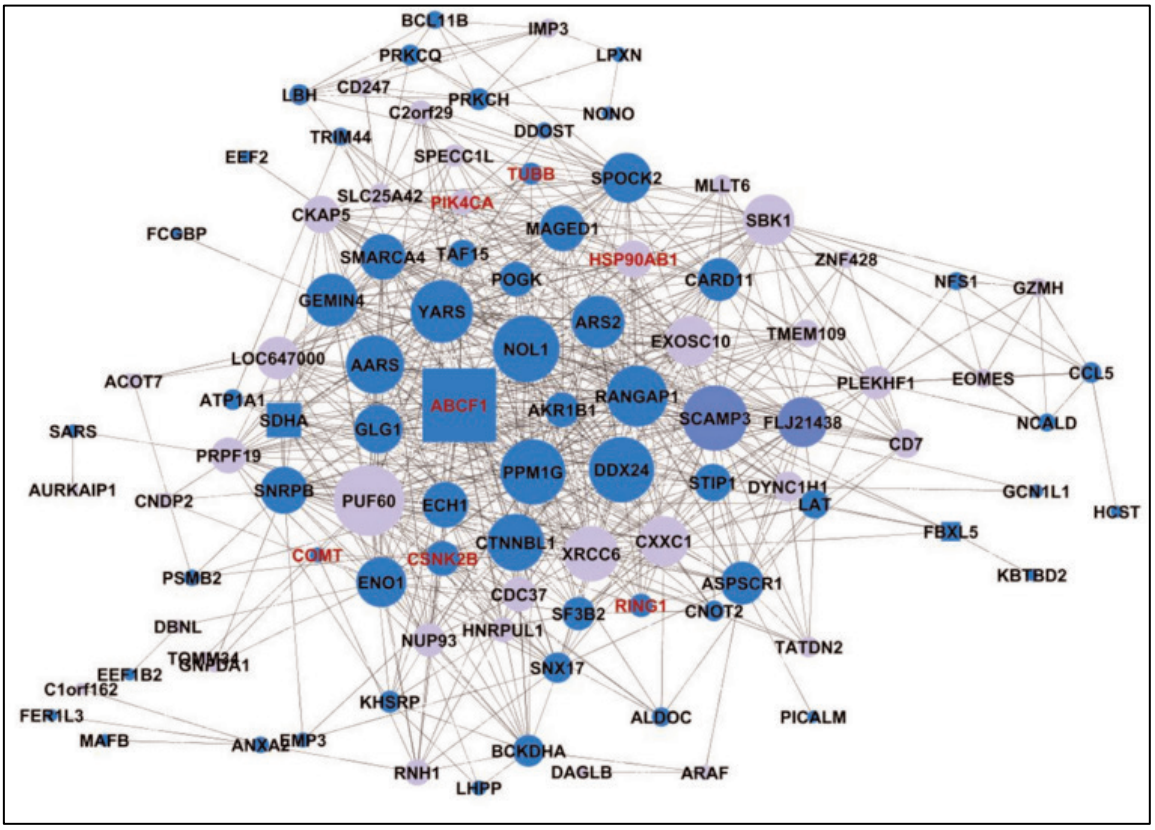


Figure 10.10. Top-down systems biology. A gene co-expression network was constructed from DNA microarray data of 92 medicated schizophrenia cases and 78 controls. Blue-colored nodes represent brain-expressed genes. Square-shape nodes indicate cis-regulation. Node size is related to the number of connections of that particular gene; a highly connected gene (i.e. 'hub gene') is therefore larger than genes with fewer connections. Red text indicates genes previously implicated in schizophrenia. (Figure copied from De Jong et al (2012) PloS ONE, 7(6), e39498).

Bottom-up systems biology

Bottom-up systems biology deduces the functional properties that could emerge from a subsystem that has been characterized to a high level of mechanistic detail using molecular methods. Bottom-up systems biology starts from the bottom (the constitutive parts) by formulating the behavior (e.g., rate equation) of each component process (e.g. enzymatic process) of a manageable part of the system. It then integrates these formulations to predict system behavior. The ultimate aim of this approach is to combine the subsystems into a model for the entire system level.

Bottom-up systems biology studies rely on:

1. Experimental studies that determine the precise properties of the components of a system. For example, studies that determine the kinetic properties of enzymes by studying these enzymes in isolation (in a tube) or in the context of their pathway (*in vivo*);
2. Data concerning responses of the subsystem to perturbations while it is in the context of the cell (in vivo). For example, one may knock-out a specific gene and investigate the effect on metabolic fluxes;
3. The construction of detailed mechanistic models using the data from (2) and subsequent simulation, validation and improvement;
4. The development of computational tools for model analysis and representation.

These models are mechanism-based rather than phenomenological.

As an example consider enzyme kinetics. Figure 10.11 shows an example of a mechanistic model of an enzyme-catalyzed reaction.

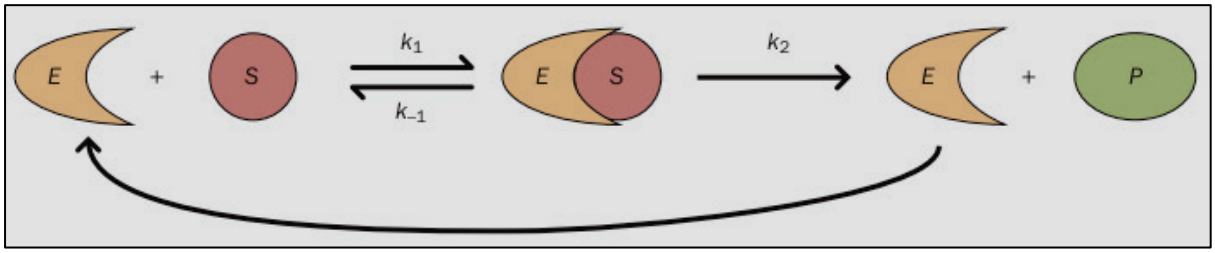


Figure 10.11. Conceptual mechanistic model of an enzyme-catalyzed reaction mechanism, as proposed by Michaelis and Menten. A substrate S and an enzyme E form a complex (ES), which may either return S and E or lead to the product P, thereby releasing free enzyme (Figure copied from Voit EO: Biological Systems. In: A first course in systems biology. First Edition. New York: Garland Science, Taylor & Francis Group, LLC; 2013).

The mechanistic model can be described by a mathematical model:

$$\frac{dS}{dt} = -k_1 SE + k_{-1}(ES)$$

$$\frac{d(ES)}{dt} = k_1 SE - (k_{-1} + k_2)(ES)$$

$$\frac{dP}{dt} = k_2(ES)$$

Here the *k*'s represent rate constants and S, E, SE, and P represent the concentrations of the substrate, enzyme, substrate-enzyme complex and product. These equations are easy to interpret. For instance, the first equation says that the change of substrate concentration in time is governed by two processes. In the first, $-k_1$SE, substrate and enzyme enter a bimolecular reaction with rate k_1. The process carries a minus sign, because substrate is lost while forming the ES complex. The second process, k_{-1}(ES), describes that some of the complex (ES) reverts to S and E, and this happens with rate k_{-1}. The sign here is positive, because the process augments the pool S. An equation for E is not formulated, because the

sum of free enzyme E and enzyme bound in the complex (ES) remains constant. Thus, once we know the dynamics of (ES), we can infer the dynamics of E.

If we want to determine the rate at which the product is formed, then we need to solve the above equations. This can be done if we make two assumptions: first that the substrate is in excess relative to the enzyme, and that (ES) does not appreciably change in concentration (quasi-steady-state assumption). If the quasi-steady-state assumption is accepted then the left-hand side of the second equation can be set equal to zero (because (ES) is assumed not to change) and one can algebraically simplify the equations so that ultimately the generation of product can be expressed as:

$$v_p = \frac{V_{\max} S}{K_m + S}$$

Thus function is the Michaelis–Menten rate law. It contains two parameters: the maximum velocity V_{max} and the Michaelis constant K_M. V_{max} quantifies the fastest possible speed with which the reaction can proceed and is defined as:

$$V_{\max} = k_2 E_{total} = k_2[E + (ES)]$$

K_m is defined as:

$$K_M = \frac{k_{-1} + k_2}{k_1}$$

and corresponds to the substrate concentration for which the reaction speed is exactly half the maximum velocity. In many cases, K_M is close to the natural substrate concentration in vivo.

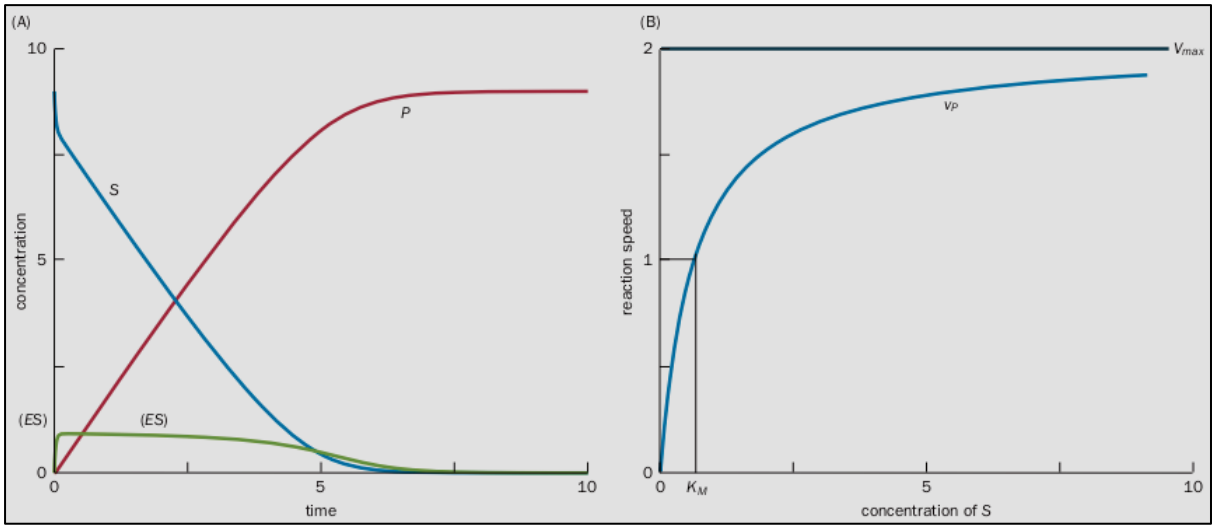


Figure 10.12. Plots of the Michaelis-Menten rate law. (A) Solution of the system of the differential equations in the form of concentrations versus time. (B) Reaction speed versus substrate concentration. Parameters are k_1=4, k_{-1}=0.5, k_2=2, K_M= 0.625, E_{total}=1, V_{max}=2, S(0)=9, (ES)(0)=P(0)=0. (Figure copied from Voit EO: Biological Systems. In: A first course in systems biology. First Edition. New York: Garland Science, Taylor & Francis Group, LLC; 2013).

Figure 10.13 shows an example of a more complex mechanism in which there is not only substrate but also an inhibitor.

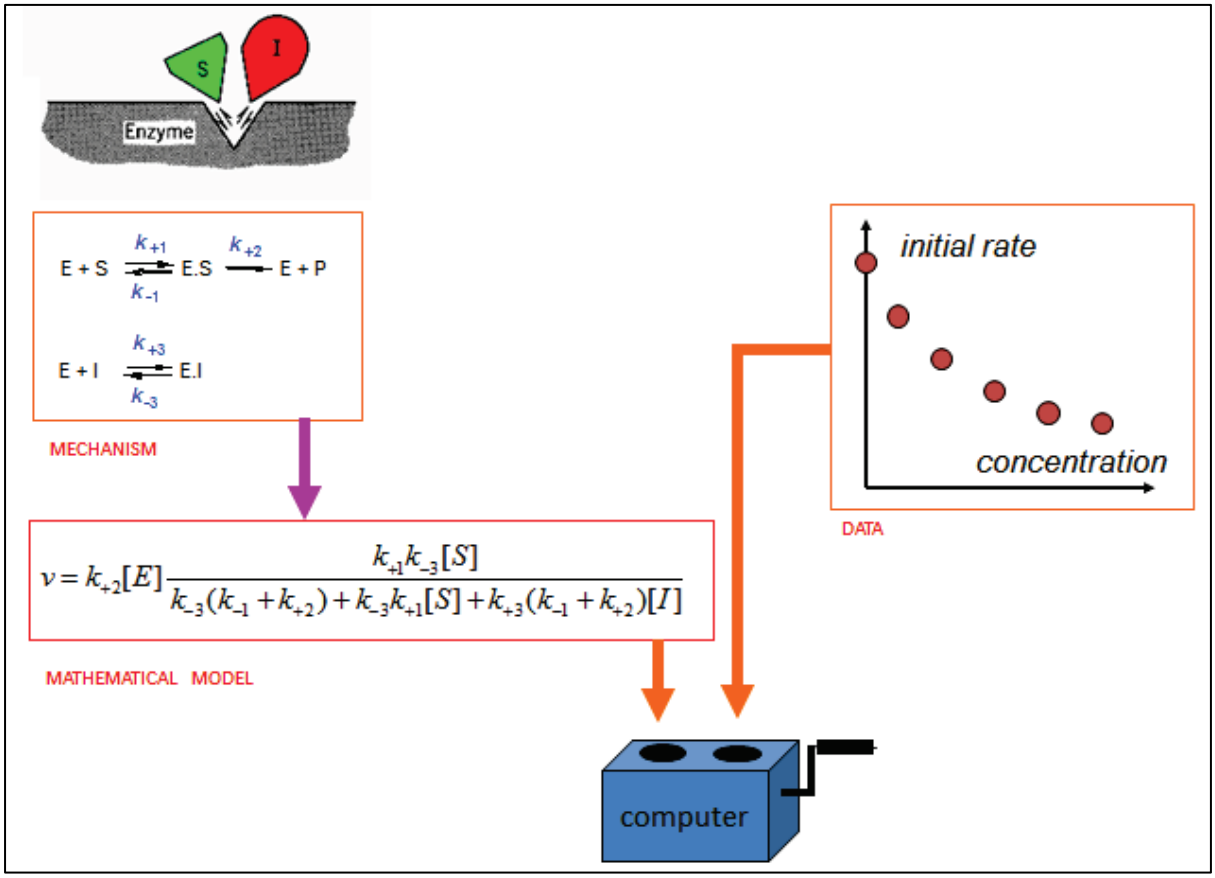


Figure 10.13. Enzyme reaction in presence of an inhibitor (I). This mechanism of inhibition is 'competitive inhibition' and can be described as shown in the 'mechanism' box. This mechanism can also be described by ODEs (not shown), which will give the rate equation for product formation shown in the 'mathematical model' box. This equation is combined with experimental data to estimate the parameters in the equations, in order to predict the rate for e.g., different concentrations of inhibitor.

10.7.3 Types of models

We have already discussed several approaches to model biological systems. Figure 10.16 summarizes these and other models:

- Interaction based models (for example the co-expression network of genes; Figure 10.10);
- Constraint-based models;
- Mechanism-based models (for example the kinetic models of Figure 10.14 and Figure 10.15).

Constrained based models are not discussed in this syllabus. These form an intermediate between the interaction-based models and mechanism-based models.

The interaction-based models are not based on an underlying biological mechanism, stoichiometry, or kinetics and, therefore, do not involve 'biological' parameters. In contrast, the mechanistic models are based on biological models, include stoichiometry of reactions and require knowledge (or measurement) of kinetic parameters. These models are very detailed, but the precise determination of the (kinetic) parameters is a bottleneck in practice.

Constrained-based models, such as flux balance analysis (FBA), allow the calculation of fluxes by defining the stoichiometry of the pathways and imposing additional constraints (e.g., minimization of the overall flux). It is important to realize that stoichiometry is chemistry and not biology! The stoichiometry of a given reaction is preserved across organisms, while the reaction rates may not be preserved (e.g., determined by the environment, regulation, etc).

Interaction-based and constrained-based static models do not describe the system (e.g., pathways) as function of time. They only provide, for example, possible interactions between components such as genes, proteins and metabolites. In contrast, dynamic (kinetic) models describe (and predict) the behavior of systems in time.

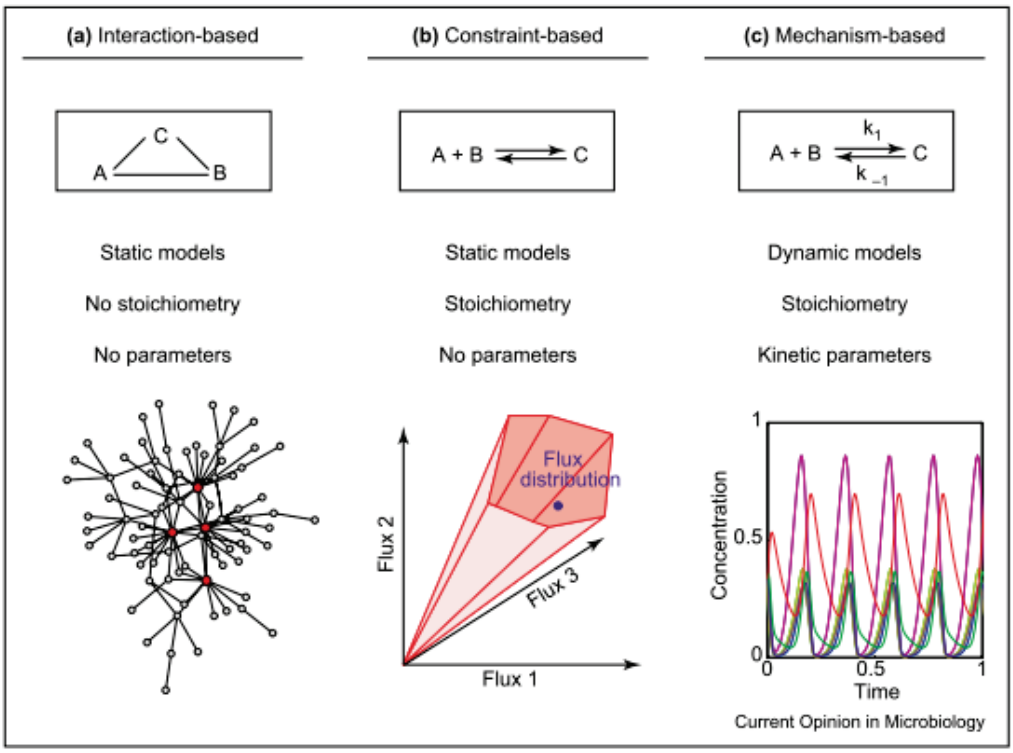


Figure 10.14. Approaches to the mathematical modeling of cellular networks. Mathematical models can start from network representations based on (a) interactions alone, (b) constraints, including network topology, stoichiometries and reaction reversibilities, or (c) (detailed) reaction mechanisms. Schemes in the bottom row illustrate typical analysis results, namely (a) hubs (red circles) in a scale-free interaction network, (b) the cone of admissible flux distributions in a metabolic network constructed from the metabolic pathways, and (c)dynamics in the concentrations of cellular components. (Figure copied from Stelling (2004) Current opinion in microbiology, 7, 513).

10.7.4 Multi-omics (integrative) systems biology

The determination of co-expression networks is one example of the use of omics data in top-down systems biology. However, one may also integrate different types of omics data facilitates systems biology research. This is called multi-omics or integrative systems biology. This is illustrated in

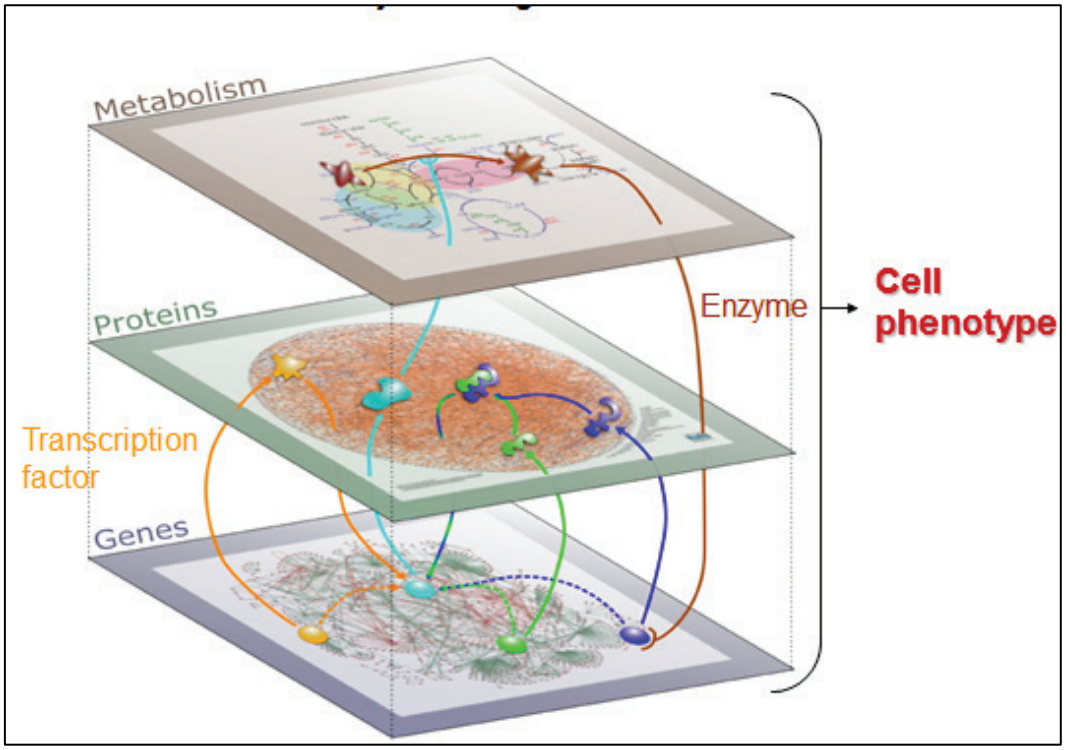


Figure 10.15. Integrative systems biology. Three levels of molecular organization (gene co-expression networks, protein-interaction networks, and metabolic pathways). Different types of interactions exist between these levels. For example, transcription factors (proteins) determine the expression of genes. Expression of genes leads to proteins or enzymes. Metabolites are involved in the regulation of gene expression. At each level omics measurements can be performed. (Figure copied from http://www-dsv.cea.fr/en/institutes/institute-of-biology-and-technology-saclay-ibitec-s/units/integrative-biology-and-molecular-genetics-sbigem/oxidative-stress-and-cancer-lsoc/physiology-and-pathogenicity-of-stress-s.-chedin-j.-labarre)

10.7.5 Multiscale modeling

Multiscale modeling is the field of solving biological problems which have important features at multiple spatial and/or temporal scales (Figure 10.19). In multiscale modeling one aims to integrate multiple temporal and/or spatial levels.

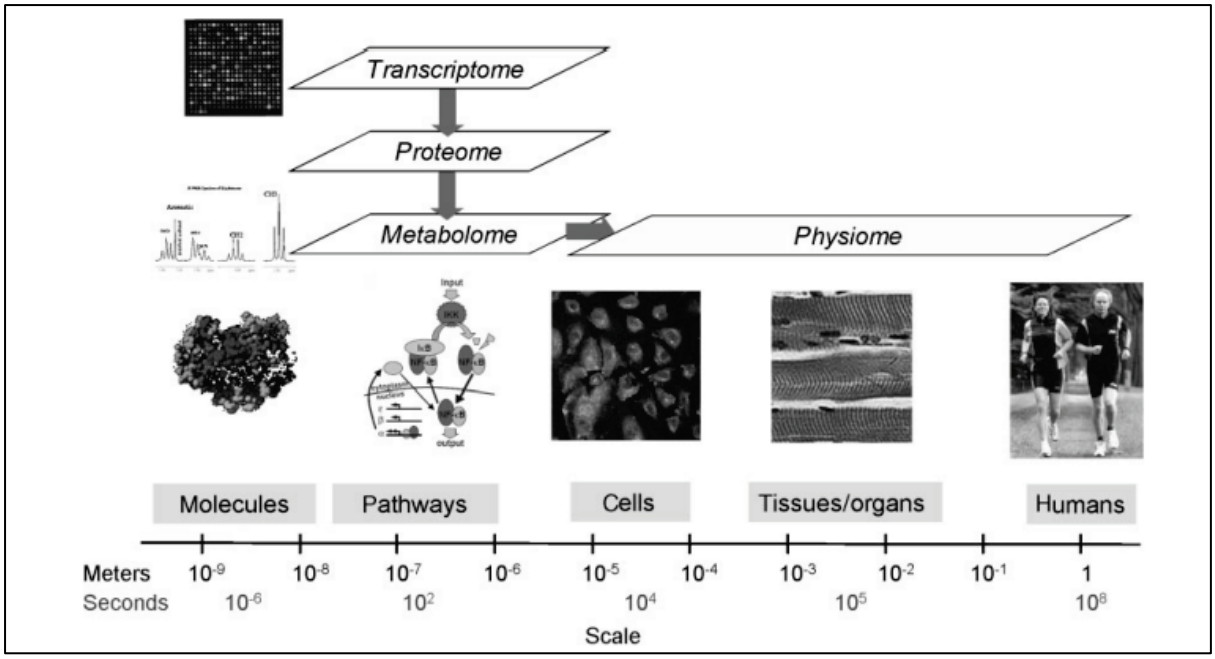


Figure 10.16. Multiscale modeling: integrative physiology and (molecular) systems biology. One could argue that the term 'systems biology' is currently inappropriately limited to the molecular scale and needs to be associated with all spatial and temporal scales. (Figure copied from Van Riel (2006) Briefings in bioinformatics, 7,364).

Figure 10.20 shows one specific example in the context of insulin resistance. The **whole-body energy model** (differential equations; Figure 10.20C) [9] is aimed to determine the fluxes between fat (F), protein (P) and glucose (G) while considering Gluconeogensis (GNG), de novo lipogensis (DNL) and glycerol 3-phosphate synthesis (G3P). Theses fluxes were modeled based on body weight measurements only, which were derived from a study in which macronutrient intakes were controlled, and body weight changes were measured in 32 healthy young men over a period of 1 yr. The subjects participated in a 24-wk semi-starvation period and lost 24% of their body weight. Semi-starvation was followed by a 12-wk controlled refeeding period. Twelve subjects went on to participate in an additional 8-wk *ad libitum* feeding phase (i.e., free access to food and water). Most model parameters were determined from published human data.

At the tissue level, the **insulin-glucose model** (Figure 10.20D) [10] provides a dynamic model (differential equations) for glucose and insulin concentrations and was derived from a glucose tolerance test. This is a medical test in which glucose is given and blood samples taken afterward to determine how quickly it is cleared from the blood. The test is usually used to test for diabetes and insulin resistance.

Finally, the **insulin receptor pathway model** (Figure 10.20E) [11] provides mathematical model (differential equations) of the metabolic insulin signaling pathways at the cellular level.

In multiscale modeling we attempt to connect these models (Figure 10.20A). For example, the glucose levels provided by the whole body energy model can provide input to the insulin-

glucose model. In principle, the insulin concentrations provided by the insulin glucose model can be linked to the pathway model. However, in practice this is not that simple because these models were developed for different temporal and spatial scales (Figure 10.20B). For example, how can we (mathematically) specify the relation between the whole-body energy model that describes changes over a year, and the much faster insulin-glucose model? In other words: what is exactly the mechanism through which these models influence each other? If we cannot specify such mechanisms then it will be hard to develop a true multiscale model. This is the major hurdle in practice.

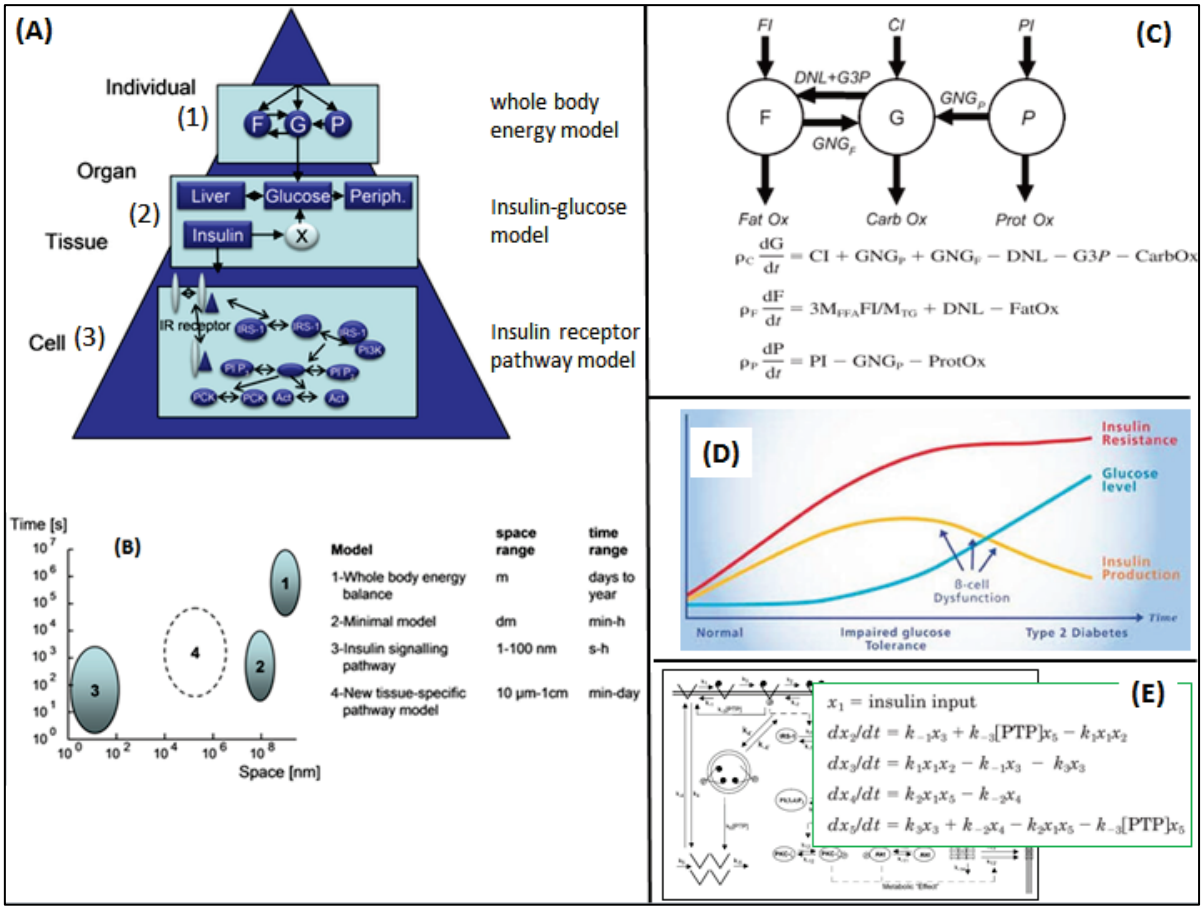


Figure 10.17. Multiscale modeling. Example of a future multiscale model in the area of insulin resistance, built from existing models. (A) schematic overview of the different model layers. (B) Individual model layers plotted along their time–space dimensions. (C) Whole-body energy model. (D) Insulin-Glucose model. (E) Insulin receptor pathway model. (Figures copied from De Graaf AA (2009) PLoS Comput Biol. 5(11), e1000554; Hall KD (2006) American journal of physiology, 291, E23–37; Sedaghat (2002) *American journal of physiology. Endocrinology and metabolism,* **283**, E1084).

10.7.6 Standardization in systems biology (not for examination)

Figure 10.21 shows an overview of different standardization efforts in systems biology. These standards focus on:

- **Minimal requirements**. This is the minimal description of a model, simulation or result to allow reproduction and verification. These requirments are simple documents describing what should be specified.

- **Standard formats**. The standard formats actually provide a technical implementation to specify the minimal information. An 'format' can be seen as a document type such as a word file (.doc). However, in systems biology and bioinformatics one generally prefers XML (eXtende Markup Language) to specify the minimal information. Standard formats such as SBML (systems biology markup language) provide a standard set of 'tags' (e.g., metabolite name) to be provided by the user. The standard formats facilitate integration of e.g., models and the use of these specificiations (models) by software applications.

- **Ontologies**. To ensure that the standard formats use standard terminologies (e.g., for metabolite names), ontologies are specified as a catalog of such terms. This facilitates the comparison between different documents.

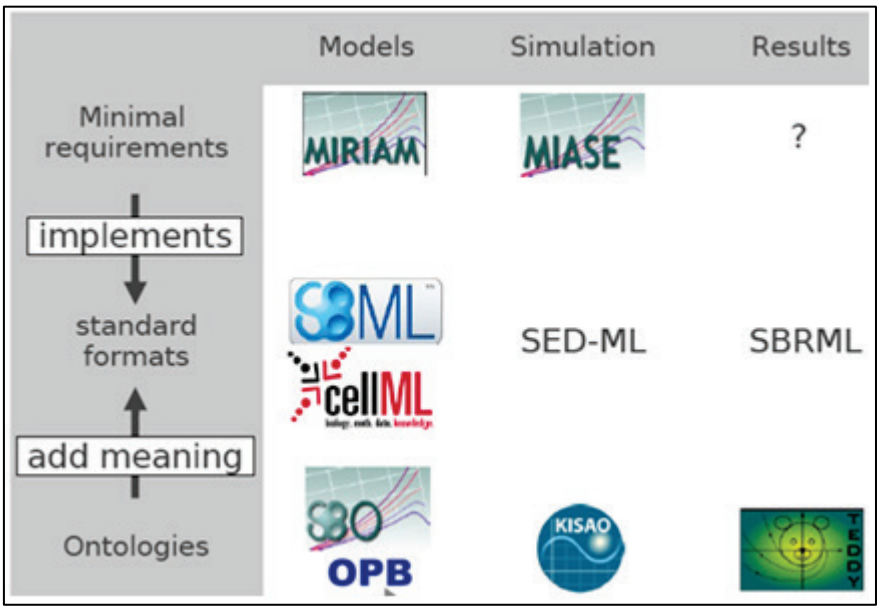


Figure 10.18. Standards in systems biology. (Figure copied from http://www.cellml.org/community/events/workshop/2009/presentations/nicolas_mibbi.pdf)

10.8 References

1. Bernard C. Introduction à l'étude de la médecine expérimentale. Baillière. 1865.
2. Noble D. Claude Bernard, the first systems biologist, and the future of physiology. Experimental physiology. 2008;93(1):16-26. doi: 10.1113/expphysiol.2007.038695. PubMed PMID: 17951329.
3. Bertalanffy Lv. Der organismus als physikalisches system betrachtet. Naturwissenschaften. 1940;33:521-31.
4. Bertalanffy Lv. General System Theory: Foundations, Development, Applications. Braziller. 1969.
5. Mesarovic M. Systems theory and biology - view of a theoretician. In: Mesarovic M, editor.: Springer; 1968. p. 59-87.
6. Rosen R, editor A means toward a new holism: Systems Theory and Biology. Proceedings of the 3rd Systems Symposium; Cleveland, Ohio. New York: Springer-Verlag; 1966.
7. Savageau M. Biochemical System Analysis: A study of function and design in molecular biology: Addison-Wesley; 1976 1976.
8. Hurford A. Mechanistic models: what is the value of understanding? http://theartofmodellingwordpresscom/2012/02/19/mechanistic-models-what-is-the-value-of-understanding/ [Internet]. 2012. Available from: http://theartofmodelling.wordpress.com/2012/02/19/mechanistic-models-what-is-the-value-of-understanding/.
9. Hall KD. Computational model of in vivo human energy metabolism during semistarvation and refeeding. American journal of physiology Endocrinology and metabolism. 2006;291(1):E23-37. doi: 10.1152/ajpendo.00523.2005. PubMed PMID: 16449298; PubMed Central PMCID: PMC2377067.
10. Bergman RN, Phillips LS, Cobelli C. Physiologic evaluation of factors controlling glucose tolerance in man: measurement of insulin sensitivity and beta-cell glucose sensitivity from the response to intravenous glucose. The Journal of clinical investigation. 1981;68(6):1456-67. PubMed PMID: 7033284; PubMed Central PMCID: PMC370948.
11. Sedaghat AR, Sherman A, Quon MJ. A mathematical model of metabolic insulin signaling pathways. American journal of physiology Endocrinology and metabolism. 2002;283(5):E1084-101. doi: 10.1152/ajpendo.00571.2001. PubMed PMID: 12376338.

11 Modelling in Systems Biology

Lecturer: dr Huub Hoefsloot (SILS, UvA)

After reading this chapter you should understand

- Understand the anatomy of a differential equation
- Understand the concept of 'mass action kinetics'
- Understand the concept of 'rate constant', 'reaction order'
- Understand the concept of 'rate law' in enzyme kinetics
- Understand how to set up a model (Ordinary Differential Equations) for a biological system
- Understand how to read a mathematical model
- Understand how to use the equations (models) to understand to investigate the biological system; Understand the kind of questions that can be answered by a model.
- Understand the relevance of mathematical models for biomedical science
- Understand that models may describe various levels (e.g., molecular, population)

Contents

11.1 Mathematical Formulation of Elementary Reactions

We begin with the simple situation in which a chemical compound degrades over time without the involvement of an enzyme or cofactor. This situation is actually not very interesting biochemically, but a good start for model development. It has two prominent applications:

- Radioactive decay, in which radionuclides spontaneously disintegrate and the collective disintegration process is proportional to the presently existing amount.
- A diffusion process, in which the transported amount of material is proportional to the current amount.

The proportionality in both cases gives us a direct hint for how to set up a describing equation: namely, the change at time t is proportional to the amount at time t.

11.2 Ordinary Differential Equations (ODE)

Change is expressed mathematically as the derivative with respect to time, the amount is a function of time, and proportional implies a linear function. Putting the pieces together into an equation yields

$$dX/dt = -kX. \qquad (1)$$

The right-hand side of the **ordinary differential equation** (ODE) contains three items. X is the amount (for example, of the radionuclide), and while it is a function of time and should be written explicitly as X(t), the reference to time is traditionally ignored in ODEs. We often refer to X generically as a pool (of molecules). The second component is the rate constant k, which is always positive (or zero) and constant over time. It quantifies how many units of X are changing per time unit. Finally, the minus sign indicates that the change (X) is in the negative direction. In other words, material disappears rather than accumulates. The formulation in (1) as a description of a chemical process is actually based on considerations of statistical thermodynamics and was proposed more than a hundred years ago by Svante Arrhenius (1859–1927).

Equation (1) is a linear differential equation that describes exponential behaviors. Suppose now that X is converted into Y and that the disappearance of X is captured well by (1). Because all material leaving pool X moves into pool Y, it is easy to see that the change in Y must be equal to the change in X, except that the two have opposite signs. The dynamics of the two variables is therefore easily described as

$$\begin{aligned} dX/dt &= -kX, \qquad (2) \\ dY/dt &= kX. \end{aligned}$$

If we add the two differential equations, we see that

$$dX/dt + dY/dt = 0, \qquad (3)$$

which has the following interpretation. The total change in the system, consisting of the change in X plus the change in Y, equals zero. There is no overall change. This makes sense, because material is just flowing from one pool to the other, while no material is added or lost.

11.3 ODEs and metabolic reactions

Many metabolic reactions involve two substrates and are therefore called bimolecular. Their mathematical description is constructed in analogy to the one-substrate case and leads to a differential equation where the right-hand side involves the product of the two substrates. Specifically, suppose X1 and X2 are the substrates of a bimolecular reaction that generates product X3. Then the increase in the concentration of X3 is given as

$$dX_3/dt = k_3 \; X_1 X_2 \; . \qquad (4)$$

Note that the substrates enter the equation as a product and not a sum, even though one might speak of adding a second substrate or formulate the reaction as $X_1 + X_2 \rightarrow X_3$. The reason for this formulation can be traced back to thermodynamics and to the fact that the two molecules have to come into physical contact within the physical space where the reaction happens, which is a matter of probability and leads to the product. Because X_3 is the recipient of material and does not affect its own synthesis, the right-hand side is positive and does not depend on X_3, and its concentration continues to increase as long as X_1 and X_2 are available. For every molecule of X_3 that is produced, one molecule of X_1 and one molecule of X_2 are used up. Therefore, the loss in either one substrate is

$$dX_1/dt = dX_2/dt = -dX_3/dt = -k_3 \; X_1 X_2 \qquad (5)$$

It is also possible that X_3 is produced from two molecules of type X_1 rather than from X_1 and X_2. In this case, the describing equations are

$$dX_1/dt = -2k_3 \; X_1^2 \qquad (6)$$
$$dX_3/dt = k_3 \; X_1^2 \qquad (6)$$

The product of X_1 and X_2 in (5) becomes X_1^2 in both equations of (6), and one says that the process is of second order with respect to X_1. The first equation in (6) furthermore contains the stoichiometric factor 2, which does not enter the second equation. The reason is that X_1 is used up twice as fast as X_3 is produced, because two molecules of X_1 are needed to generate one molecule of X_3.

The mathematical formulations discussed here are the foundation of **mass action kinetics**. According to this widely used analytical framework, all substrates enter a reaction term as factors in a product, where powers reflect the number of contributing molecules of each type. The term also contains a rate constant, which is positive (or possibly zero) and does not change over time. Mass action formulations implicitly assume that many substrate molecules are available and that they are freely moving about in a homogeneous medium. In many cases, these assumptions provide good approximations to realistic metabolic systems, even if they are not 100% true in reality. They are frequently used because more accurate formulations are incomparably more complicated. For instance, one could more realistically consider the reaction between X_1 and X_2 as a random encounter (stochastic) process. While intuitively plausible, a model of such a process requires heavy-duty mathematics for further analysis.

11.4 Rate Laws

As we discussed earlier, most biochemical reactions in a metabolic system are catalyzed by enzymes. About a century ago, the biochemists Henri, Michaelis, and Menten proposed a mechanism and a mathematical formula describing this action. They postulated that the substrate S and the catalyzing enzyme E reversibly form an intermediate complex (ES), which subsequently breaks apart and irreversibly yields the reaction product P while releasing the enzyme unchanged and ready for recycling. The diagram of the mechanism with typical rate constants for all steps is shown in Figure 11.1.

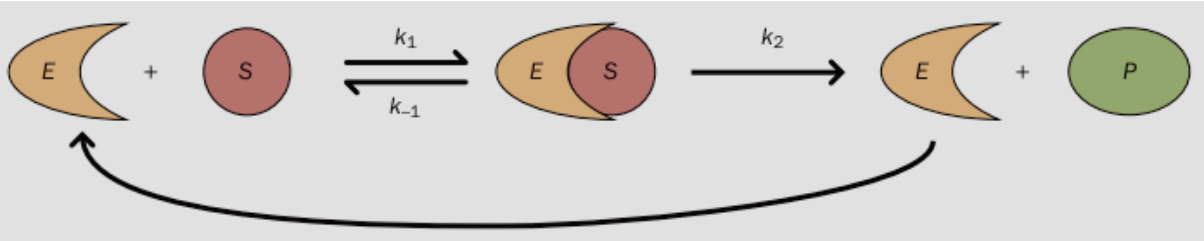


Figure 11.1. Conceptual mechanistic model of an enzyme-catalyzed reaction mechanism, as proposed by Michaelis and Menten. A substrate S and an enzyme E form a complex (ES), which may either return S and E or lead to the product P, thereby releasing free enzyme (Figure copied from Voit EO: Biological Systems. In: A first course in systems biology. First Edition. New York: Garland Science, Taylor & Francis Group, LLC; 2013).

Conceptual mechanistic model of an enzyme-catalyzed reaction mechanism, as proposed by Michaelis and Menten. A substrate S and an enzyme E form a complex (ES), which may either return S and E or lead to the product P, thereby releasing free enzyme.

Modern methods of transient and single-molecule kinetics have shown that this simple scheme is somewhat simplistic and that many reactions in fact consist of entire networks of fast subreactions with different intermediate complexes that form a multidimensional free

energy surface. Nonetheless, the Michaelis–Menten mechanism is a very useful conceptual framework, and an enormous number of studies have measured its characteristic parameters.

Specifically, each process term is formulated by including all variables that are directly involved, as well as the appropriate rate constant.

$$\frac{dS}{dt} = -k_1[S][E] + k_{-1}[ES] \qquad (7)$$
$$\frac{dES}{dt} = k_1[S][E] - k_{-1}[ES] - k_2[\mathrm{ES}]$$
$$\frac{dP}{dt} = k_2[ES]$$

These equations are easy to interpret. For instance, (7) says that the change in substrate is governed by two processes. In the first, $-k_1$ S E, substrate and enzyme enter a bimolecular reaction with rate k_1. The process carries a minus sign, because substrate is lost. The second process, k_{-1}(ES), describes that some of the complex (ES) reverts to S and E, and this happens with rate k_{-1}. The sign here is positive, because the process augments the pool S. An equation for E is not formulated, because the sum of free enzyme E and enzyme bound in the complex (ES) remains constant. Thus, once we know the dynamics of (ES), we can infer the dynamics of E.

11.5 Mitogen-Activated Protein Kinase Cascades

In contrast to the two-component systems (TCS) in bacteria, most higher organisms use a more complicated signal transduction system whose key component is the mitogen-activated protein kinase (**MAPK**) **signaling cascade**. MAPK systems appear to be present in all eukaryotes and also in a few prokaryotes, such as the biofilm bacterium *Myxococcus xanthus*. Intriguingly, MAPK cascades have a highly conserved architecture. The MAPK signaling cascade receives signals from cell surface receptors or cytosolic events, processes them by filtering out noise, usually amplifies them, and ultimately affects various downstream targets, such as cytosolic proteins and nuclear transcription factors. The external signal may consist of a mitogen or inflammatory cytokine, a growth factor, or some physiological stress. This signal is transduced, for example, by a G-protein-coupled receptor (GPCR) that activates the MAPK cascade. For signal result of the MAPK system may be as different as inflammation, differentiation, apoptosis, or initiation of the cell cycle.

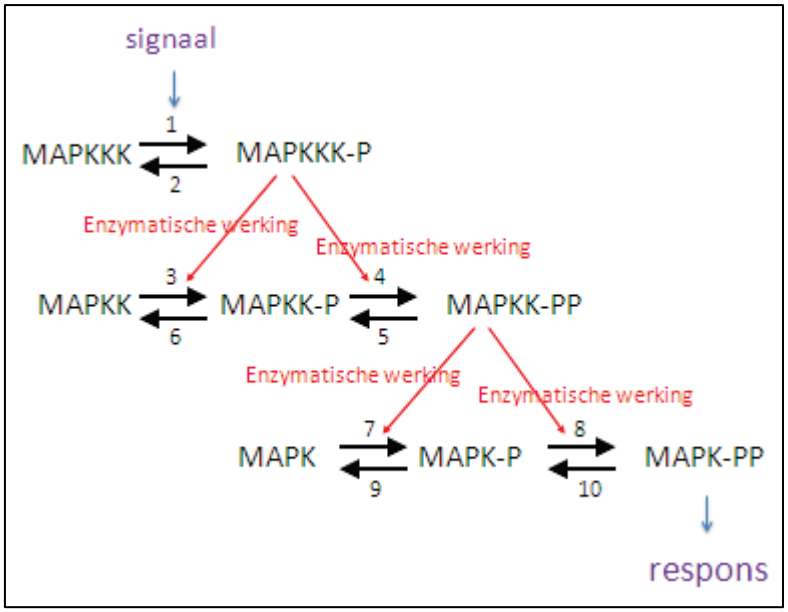


Figure 11.2. Typical MAPK cascade

The typical MAPK cascade is shown in Figure 11.2. It consists of three layers where proteins are phosphorylated or dephosphorylated. The kinase nearest to the signal source is generically called MAP kinase kinase kinase (MAPKKK). If it is activated by a cytosolic or external signal, MAPKKK is phosphorylated to MAPKKK-P. Being a kinase itself, MAPKKK-P activates phosphorylation of the MAP kinase kinase (MAPKK) in the second layer. In fact, full activation of MAPKK requires sequential phosphorylation at two sites: a tyrosine site and a threonine site. The resulting MAPKK-PP in turn phosphorylates and thereby activates the kinase of the third layer, the MAP kinase (MAPK). Again, full activation requires two phosphorylation steps. The activated MAPK-PP can phosphorylate cytosolic targets or can translocate to the nucleus, where it activates specific transcriptional programs. At each layer, a phosphatase can dephosphorylate the active forms into the corresponding inactive forms. Two prominent examples of MAP kinases are the extracellular-signal-regulated kinase (ERK) and the c-Jun N-terminal kinase (JNK), where Jun refers to a family of transcription factors. The responses of the cascades are quite fast: within the first 5 minutes of stimulation, ERK is activated up to 70%, and within 10 minutes, significant amounts of activated ERK are translocated to the nucleus. A mutated form of MAPKKK in the ERK pathway has often been found in malignant melanomas and other cancers.

An obvious question is this: The three-layer cascade has been conserved through evolution, so it is fair to assume that this design has advantages. But what are these, for instance, in comparison with a simple or double phosphorylation at a single layer? We use a model analysis to shed light on this intriguing question.

11.5.1 Key module of the three-layer

The key module of the cascade is the sequential phosphorylation at two sites, which is catalyzed by the same enzyme from the next higher layer. A simplified and detailed diagram of this module is presented in Figure 11.3, where *X* stands for MAPK or MAPKK, *E* is the catalyzing enzyme, *XP* and *XPP* are singly or doubly phosphorylated forms, and *XE* and *XPE* are complexes of *X* and *XP* with the enzyme.

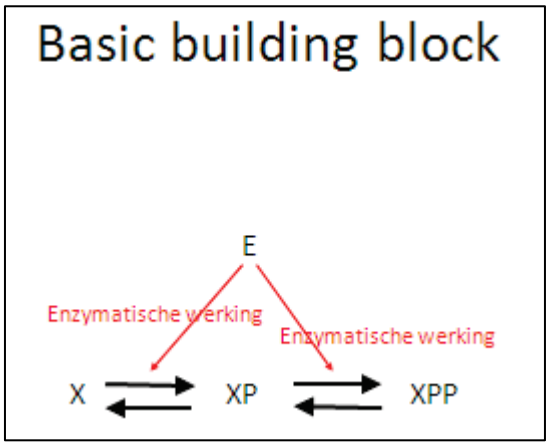
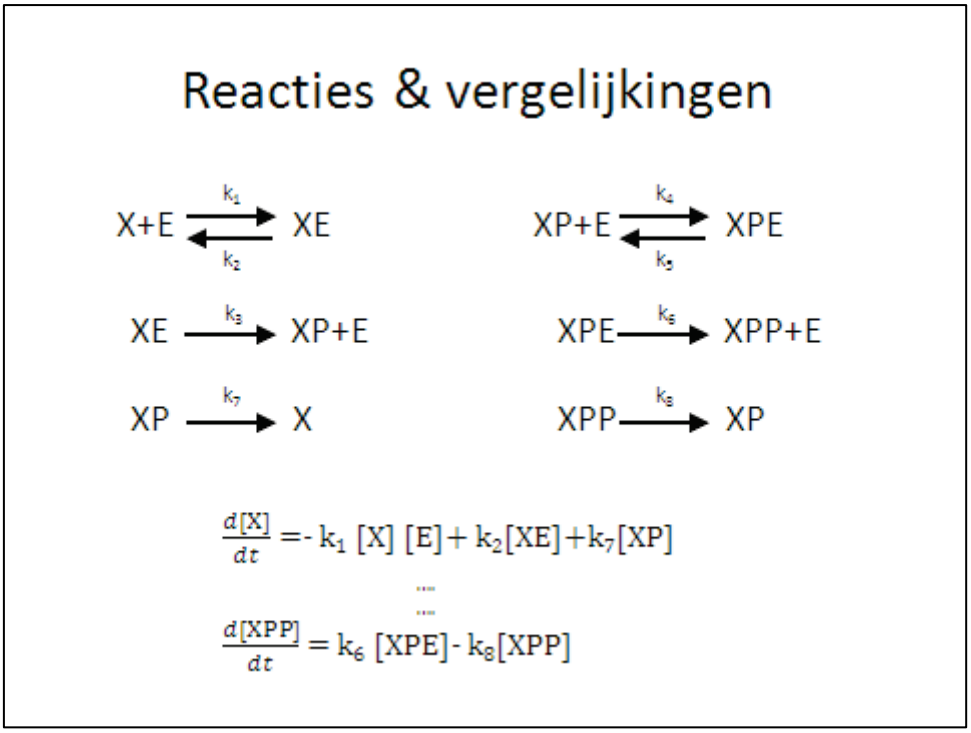


Figure 11.3. Key module of the three-layer cascade.

It is tempting to set up the two phosphorylation steps with Michaelis–Menten rate functions, but such a strategy is not the best option, because (1) the enzyme concentration is not constant, (2) the enzyme concentration is not necessarily smaller than the substrate concentration, and (3) the two reaction steps are competing for the same enzyme. Instead, it is useful to retain the mechanistic ideas of the Michaelis–Menten mechanism, which postulates the formation of a substrate–enzyme complex, and to formulate this mechanism in the basic format of mass-action kinetics. What we do *not* want to do is to rely on the quasi-steady-state assumption that would simplify this system toward the well-known Michaelis–Menten function, because in this format the enzyme concentration is no longer explicit, let alone dynamic. A direct translation of the diagram in Figure 11.3 into mass-action equations is straightforward:



To perform simulations, we specify more or less arbitrarily chosen values for the parameters and initial concentrations in Figure 11.4. It is now easy to study the effects of changes in the enzyme *E* on the variables of the module. At the beginning of the simulation, *E* = 0.01, and the initial conditions for the various forms of *X* are set such that the system is more or less in a steady state. Suppose now that at time *t* = 1 the signal *E* is increased to 10. This is a strong signal, and the system responds with phosphorylation on both sites. After a brief transition, the balance between *X*, *XP*, and *XPP* is entirely switched, with very little unphosphorylated *X* left.

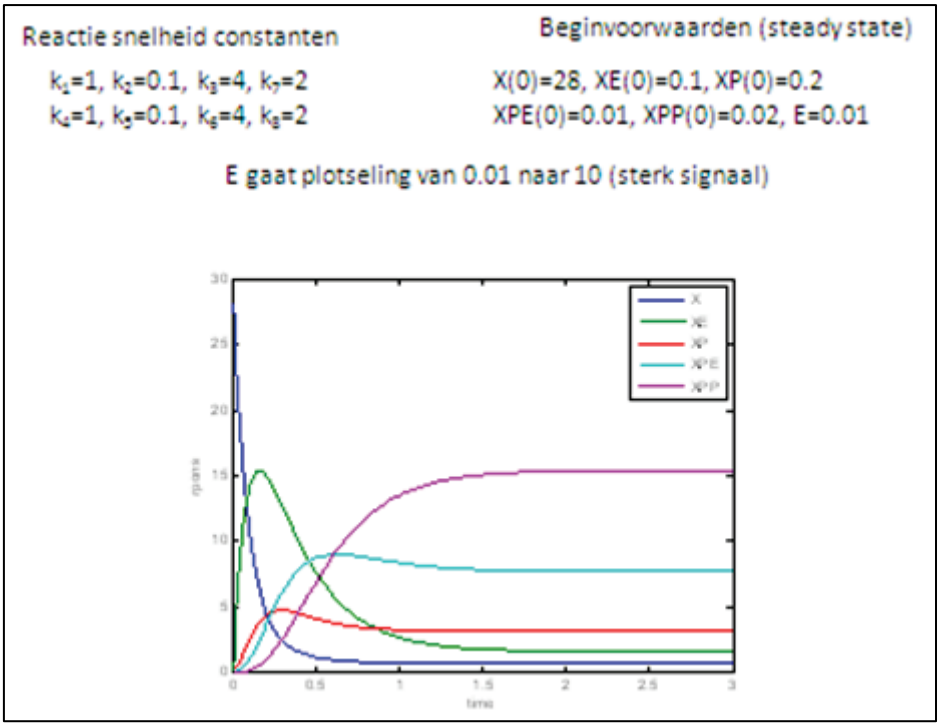


Figure 11.4. Effects of changes in the enzyme E on the variables of the module

Notably, the amount of *XPE* is relatively large, because quite a lot of enzyme is available. If the signal is reset to 0.01, the module returns to its initial state (Figure 11.5)

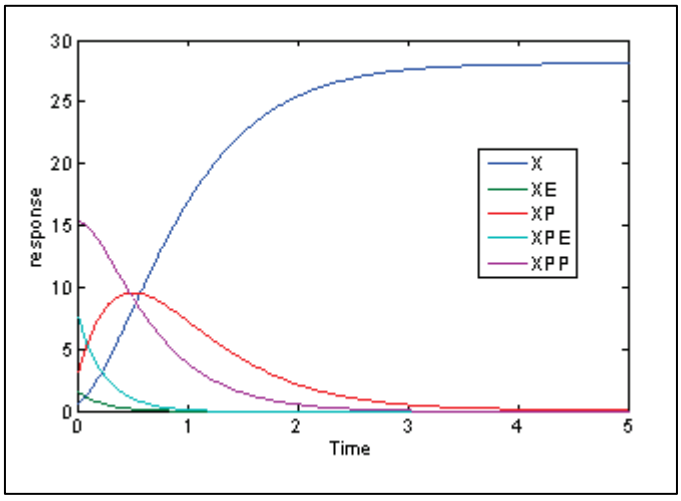


Figure 11.5. The concentration of E is reset from 10 to 0.01. As a result all other concentrations go back to their initial state.

It is also easy to study how strong the signal must be to trigger a response. If *E* is set to 3, rather than 10, the response is rather similar to the previous scenario, although the response is not as pronounced (results not shown). If *E* is set to 1 instead, the response is much slower. *X* is reduced only to about 13.6, while *XPP* assumes a value of about 3.2. In other words, no real switch is triggered. For even weaker signals, the response is insignificant.

Now that we have a bit of a feel for the module, we can use it to implement a complete MAPK cascade. Specifically, we set up the module in triplicate and use the same parameter values and initial conditions as before. In addition, we model the single phosphorylation of MAPKKK with a simple Michaelis–Menten mechanism, again formulated in mass-action representation, with $X(0) = 28$ and $XE(0) = XP(0) = 0.1$.

For a relatively weak signal ($S = 0.5$ at $t = 2$; note that S is E in our model), we see the benefit of the cascaded sys-response system. While MAPKKK-*P* only rises to about 5, the ultimate response variable MAPK-*PP* shows a robust response (Figure 11.6).

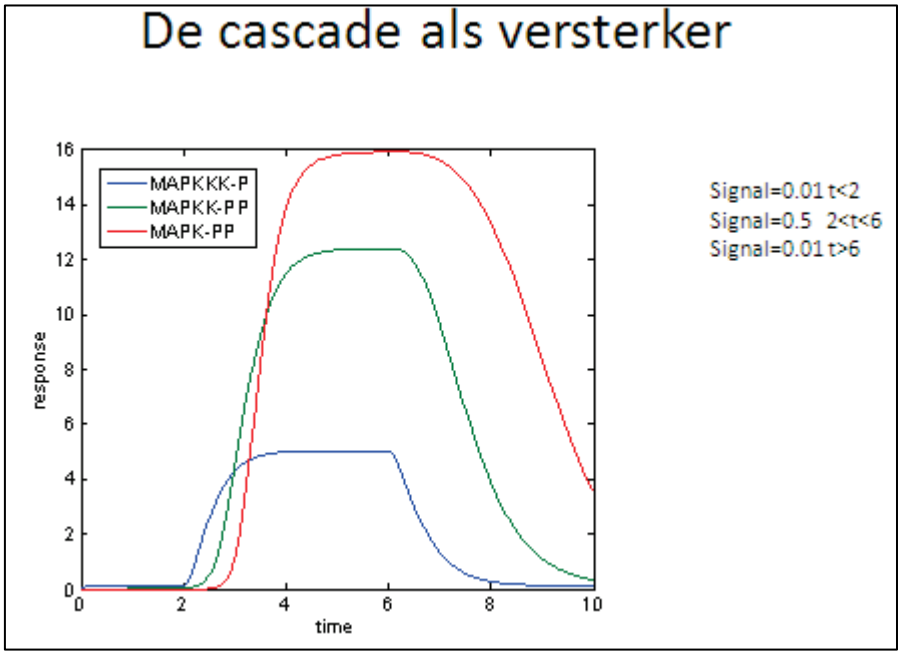


Figure 11.6. The cascade as an amplifier.

In other words, we have solid signal amplification in addition to a steeper response and clear signal transduction. If the signal rises to only two or three times its initial value ($S = 0.02$ or $S = 0.03$ at $t = 2$), the response in MAPK-*PP* is very weak. The results give us a first answer to our earlier question of why there are three layers in the cascade: three layers permit improved signal amplification. At the same time, they effectively reduce noise

It is easy to investigate how the cascade responds to brief repeated signals. We perform a simulation as before, but create an off –on sequence of signals. As an example, if the signal switches between 0.01 and 0.5 every half time unit, beginning with $S = 0.5$ at $t = 2$, we obtain the result shown in **Figure 11.7**.

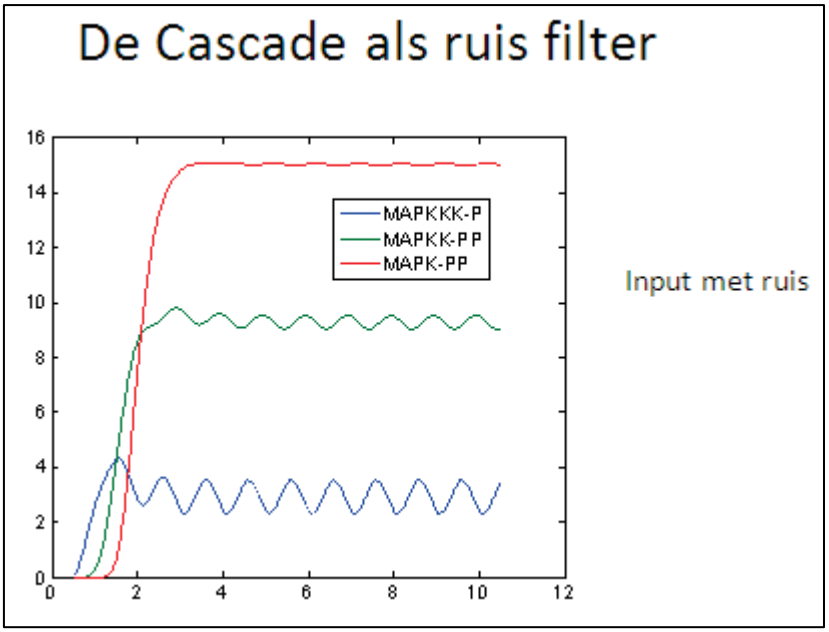


Figure 11.7. Cascade as noise filter.

We can see that the responses at the different layers are quite different. The first layer exhibits fast responses, whereas the ultimate, amplified response in third layer is quite smooth.

11.6 The SIR Model

Following some old ideas of Kermack and McKendrick, we use a similar terminology and keep things as simple as possible; one should, however, note that thousands of variations on this model have been proposed since Kermack and McKendrick's days. We begin by defining just three dependent variables, namely the number of individuals susceptible to the disease (*S*), the number of infected individuals (*I*) and the number of individuals that are "removed" (*R*) from the two pools, because they have acquired immunity. We suppose that all individuals could be in contact with each other, at least in principle, and assume that a certain percentage of the immunized individuals (*R*) lose their immunity and become susceptible again. We also allow for the possibility that individuals are born and that individuals may die while infected. A diagram summarizing this population dynamics is shown in Figure 11.8.

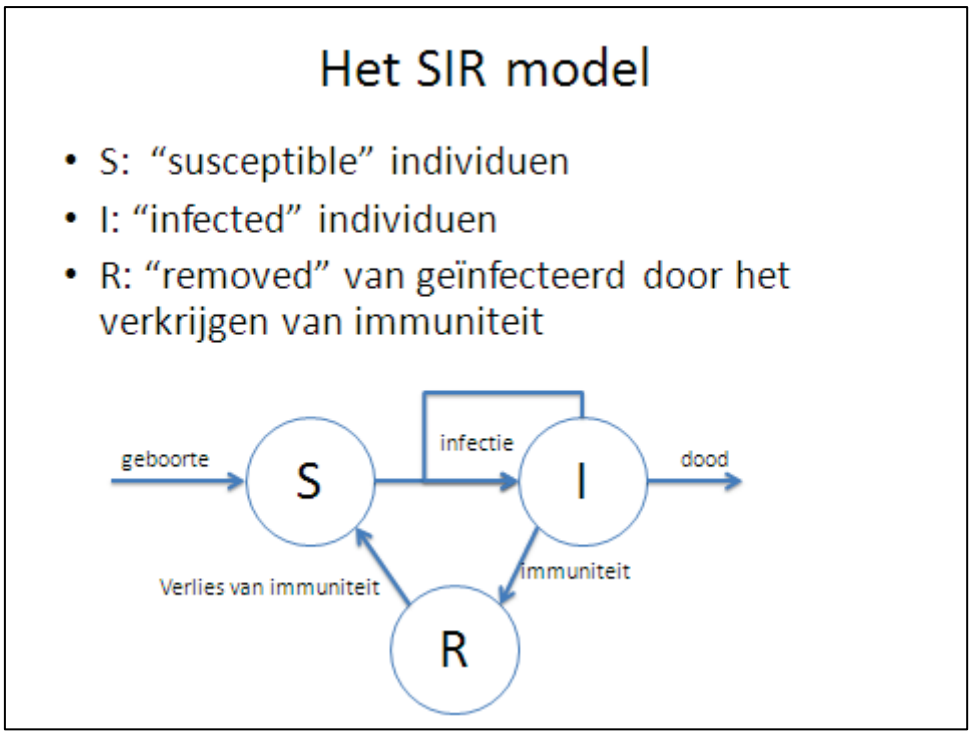


Figure 11.8. The SIR Model.

11.6.1 Model Equations

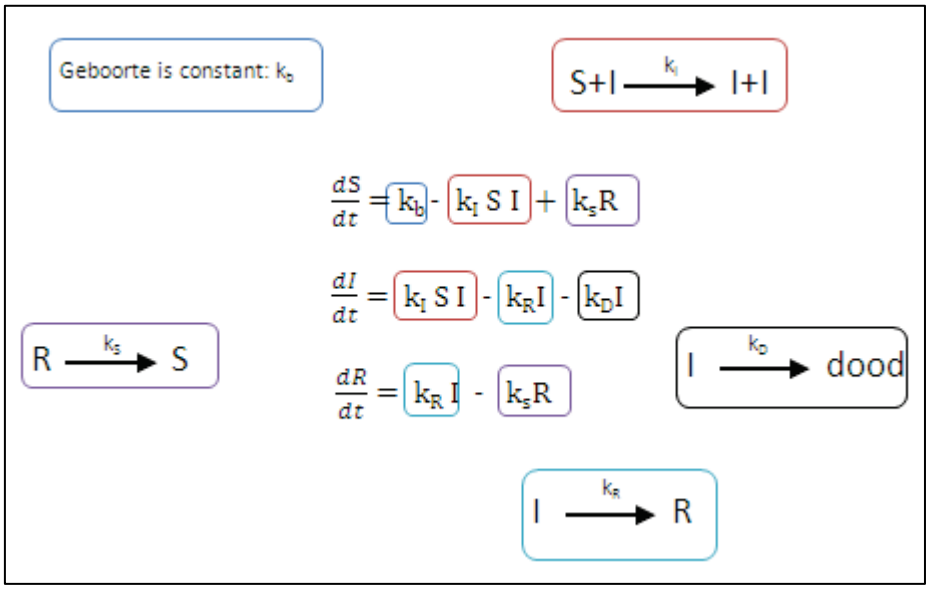
Once we have completed our model diagram and established the lists of components and processes, we need to translate this information into equations that reflect the chosen model structure. At first, these equations are symbolic, which means that no numerical values are yet assigned to the processes that characterize the system.

For the infectious disease problem, we choose a simple differential equation model, called an **SIR** model, which is by its nature moderately explanatory, dynamic, continuous, deterministic, and independent of spatial aspects. More complicated choices could include random processes associated with the infection of susceptible individuals or age distributions within the population with different degrees of susceptibility and recovery.

Like Kermack and McKendrick, we use as a model structure the simple and intuitive process description of mass action kinetics. This structure entails that the terms in each equation are set up either as constants (birth) or as a rate constant multiplied with the dependent variables that are directly involved in the process. This strategy reflects that the number of individuals moving from one pool to another depends on the size of the source pool. The more infected individuals there are, for instance, the more people will acquire immunity.

The specific construction is implemented one differential equation at a time and requires us to introduce new quantities, namely the rates with which the various processes occur. The change in *S* is governed by birth with a constant rate k_b, and replenishment of the susceptible

pool with individuals losing immunity with rate k_s. At the beginning, no immune individuals may be around (R = 0), but the model accounts for the possibility, which may become reality later. The pool S is diminished by the infection process which has a rate k_I and depends on both S and I, because an infection requires that a susceptible and an infected individual come into contact. Thus, the first equation reads:



Here, k_R is the rate of acquiring immunity and k_D is the rate of death. The set-up implies that only infected individuals may die, which may be a matter of debate, or of a later model extension.

We can see that several terms appear twice in the set of equations, once with a positive and once with a negative sign. This is a usual occurrence, because they describe in one case the number of individuals leaving a particular pool and in the other case the same number of individuals entering another pool. The equations thus constructed are symbolic, because we have not yet committed to specific values for the various rates. Therefore, this set of symbolic equations really describes infinitely many models. Many of these will have similar characteristics, but it may also happen that one set of parameter values leads to very different responses than a different set.

Essentially all parameters in our case are rates (which generically describe the number of events per time), so that it is necessary to decide on a time unit. For this example, we use days and set the rates correspondingly. Babies are born at a rate of 3 per day, 2% of the infected individuals actually die per day, and individuals lose immunity at a rate of 1% of R. The other rates are self-explanatory. The initial values say that, at the start of the model period, 99% of the individuals in a population of 1000 are healthy, yet susceptible to the disease (S = 990), that 1% are infected (I =10), for reasons we do not know, and that nobody

is initially immune (R = 0). These settings are entered into the symbolic equations and the resulting parameterized equations are thus:

Stelsel differentiaal vergelijkingen

parameters

k_b=3
k_I=0.0005
k_S=0.01
k_D=0.02
k_R=0.05

Begin waarden

S(0)=990
I(0)=10
R(0)=0

$$\frac{dS}{dt} = k_b - k_I S I + k_s R$$

$$\frac{dI}{dt} = k_I S I - k_R I - k_D I$$

$$\frac{dR}{dt} = k_R I - k_s R$$

11.6.2 MODEL ANALYSIS AND DIAGNOSIS

A typical model analysis contains two phases. The first consists of model diagnostics, which attempts to ensure that nothing is obviously wrong with the model and assesses whether the model has a chance of being useful. For example, a model in which a variable disappears altogether when some physiological feature is changed by just a few percent is not very robust and the actual natural system would not survive in the rough-and-tumble outside world for very long. After we have received a green light from the diagnostics, we enter the second phase of exploring what the model is able or unable to do. Can it oscillate? Can some variable of interest reach a 4 *x* level of 100 units? What would it take to reduce a variable to 10% of its normal 10 value? How long does it take until a system recovers from a **perturbation**?

Because we are dealing with mathematics, we might expect that we could directly compute all properties of a model with calculus or algebra. However, this is not necessarily so, even in apparently simple situations. For our illustration, let's start by solving the equations using a computer simulation with the baseline parameters that we set above. We could use for this purpose software like Mathematica MatLab, or R. Figure 11.9 shows the result, which consists of one time course for each dependent variable.

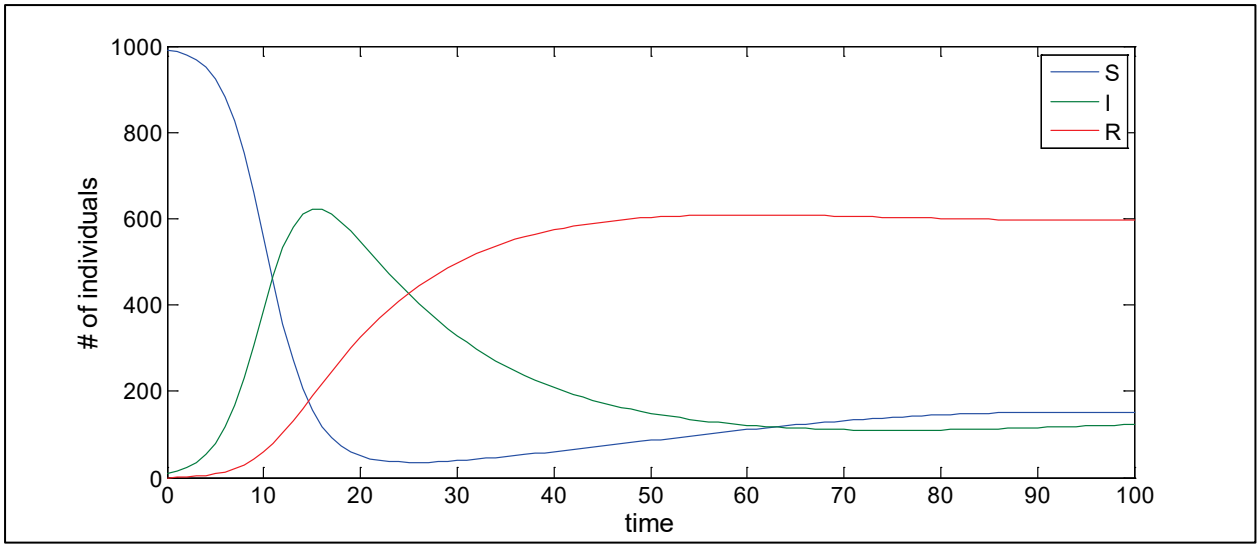


Figure 11.9. Results of a computer simulation of the SIR model.

In our example, the changes within the population are quite dramatic. The subpopulation of susceptible people almost disappears, because almost everyone gets sick, but the number of immune individuals eventually exceeds half the population size. Yet, the infection does not seem to disappear within the first 100 days since the outbreak.

Let's do some computational exploration. The first question is whether the model can reach a **state** where none of the variables changes anymore. The answer is *yes*: we can show such a **steady state** by solving the equations for a sufficient time range (note that the number of immune individuals still climbs after one year). After a long time, the variables eventually reach the values S =140, I = 150, and R = 750, and this steady state is stable, otherwise the model would not approach it (Figure 11.10)

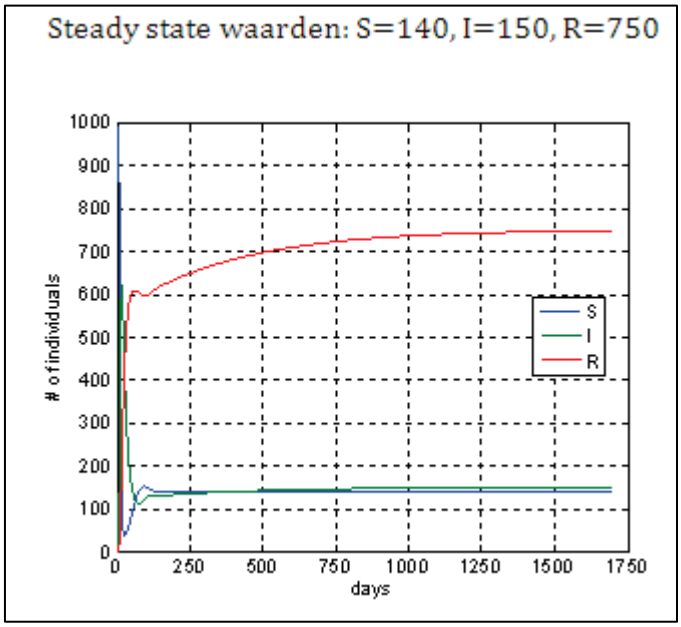


Figure 11.10. Steady state values for SIR model

If we compute the total population size after several years, we note that there are now more people (1040) than at the beginning (1000). Is that reason to worry? No, it just means that the initial values were below the “normal” steady-state size of the population. What happens when we start with *S* = 1500? Because the steady state is stable, the population will shrink and in the end approach 1040 again.

Uitrekenen steady state

$$0 = k_b - k_I S I + k_s R$$

$$0 = k_I S I - k_R I - k_D I$$

$$0 = k_R I - k_s R$$

Het aantal geboorten en het aantal sterfgevallen houdt elkaar in evenwicht

$k_b = k_D I \Rightarrow I = k_b / k_D \Rightarrow I = 150$
$R = k_R / (k_S I) = 750$
$S = (k_b + k_S R) / (k_I I) = 140$

Mogelijke uitbreidingen

- Vaccinatie, quarantaine.
- Asymptomatische personen (mensen met geen of nauwelijks symptomen)
- Sterfgevallen in de S & R groepen
- Rekening houden met persoonlijke karakteristieken
  - Gezondheid
  - Hoeveelheid contacten

12 Analyzing transcriptomics, proteomics and metabolomics data

Lecturer: **Dr. Johan Westerhuis (Biosystems Data Analysis Group)**

After reading this chapter you should understand

- To explain and apply the principles of experimental design (replication, randomisation and blocking).
- To explain and apply multivariate explorative analysis.
- To explain and perform multivariate biomarker selection.
- To describe the reason for data scaling, data transformation and data normalization.

Sources:

- Draghici S. Data Analysis Tools for DNA Microarrays, ISBN 1-58488-3.5-4. 2003. Boca Raton: Chapman & Hall/CRC.
- Gibson G, Spencer VM. A Primer of Genome Science, third edition, ISBN 978-0-87893-236-8. 2009. Sunderland, Massachusetts: Sinauer Associates, Inc Publishers.
- Pevsner J. Bioinformatics and Functional Genomics, ISBN 0-471-21004-8. 2003. Hoboken, New Jersey: John Wiley and Sons.
- Baxevanis AD, Francis Ouellette BF. Bioinformatics, ISBN 0-471-47878-4. 2005. Hoboken, New Jersey: John Wiley and Sons.

Contents

12.1 INTRODUCTION

Transcriptomics, proteomics and metabolomics experiments are performed with different technologies. Each technology therefore has its own specifics with respect to the analysis of the data. Microarray data has different properties than metabolomics data and different data analysis methods will be used in general. This is especially true with respect to data preprocessing and biological interpretation. Each technology will have its own data analysis pipeline. Examples of references dealing with specific data analysis pipelines can e.g. be found in Brown et al. (2005).

Nevertheless, there is also a large overlap between these pipelines on concepts of data analysis. The ways transcriptomics, proteomics and metabolomics experiments are designed are very similar. Also the ways the dimensionality of the data is handled are similar. Basically, all technologies yield many measurements for each sample. Transcriptomics yields many gene expression measurements per sample, proteomics yields many protein expression measurements per sample, and metabolomics yields many metabolite measurements per sample. Thus, each of these technologies yields hundreds or thousands variables per sample. These variables can be different genes, proteins or metabolites. For all these technologies the data can be organized in a matrix. In many cases the samples are organized in the rows (from left to right) and the variables in the columns (top to bottom). In transcriptomics data analysis however, the samples are usually in the columns and the genes in the rows. Nevertheless, the statistical tools to analyze the data matrix are quite similar for transcriptomics, proteomics and metabolomics.

The goal of this text is to present a generalized data analysis pipeline which is based on the common characteristics of the transcriptomics, proteomics and metabolomics pipelines. It should help the student to keep an overview of the procedures used and memorize the relevant components of each pipeline. The generalized data analysis pipeline consists of four components:

1) Experimental design and data collection. Frame a biological question. Identify noise factors and design the experiment. Execute the experiment.

2) Data preprocessing and quality control. Extract the instrumental intensities. Perform data preprocessing and normalization to remove biases introduced by sampling and measurement.

3) Data analysis. Perform exploratory data analysis to find strange phenomena in the data. Perform biomarker selection to find those metabolites that are important in the study.

4) Biological interpretation. Give biological meaning to a group of metabolites that have found to be important in discriminating the patients from the controls.

12.2 EXPERIMENTAL DESIGN AND DATA COLLECTION.

It is very important that the Experimental design step is not ignored, and in many ways it is even the most important step. The starting point of any biological experiment should always be the precise formulation of a biological question that will guide the implementation of an efficient and informative experimental design. Consultation with a statistician on issues such as the levels of replication that will be required to detect an effect or perform a contrast, since the way an experiment is performed can have a large impact on statistical power and hence the conclusions that are reached. A flaw in the experimental design can lead to conclusions that are fundamentally wrong.

12.2.1 Frame a biological question

The first step in the experimental design is to formulate a biological question. The biological question obviously specifies the biological context. However, an important aim of the biological question is that it determines the hypothesis that will be tested (e.g. there is no difference between these groups) and the statistical test that will be executed. In that sense, it also determines the experimental preconditions that should be met for an interpretable and successful outcome. Although the number of biological questions is endless, there are arguably three types of main objectives that each requires a different type of experimental design: 1) Detection of responsive features (genes, proteins or metabolites) under controlled experimental conditions (perturbation study), 2) Detection of biomarkers and 3) Identification of regulatory or mechanistic relationships between variables.

12.2.2 Identify noise factors and design the experiment

Once the biological question is formulated the researcher has to decide on a sampling scheme. To successfully design the experiment the researcher has to overlook the complete process of data acquisition, from the biological experiment up to and including the measurement, to identify possible factors that can disturb a proper measurement. Examples of such factors are temperature, cell batch, climate room, etc. These factors are usually not of experimental interest, but introduce **noise** and possibly **bias**. Such factors are sometimes called “nuisance factors”. There are three basic principles: 1) replication, 2) randomization, 3) blocking, that enable the researcher to deal with these noise factors. The aim of the experimental design is to ensure reliable measurements free from bias.

Replication means to duplicate, repeat, or perform the same measurement more than once. Replication allows the experimenter to obtain an estimate of the experimental error. However, the *type of error* that is estimated depends on how the replication is done. If the purpose is to estimate and control biological variability, different test organisms or batches of cells should be sampled for mRNA or metabolites, and these samples should be processed in exactly the same manner. Replication can be done on different levels. In LCMS (proteomics or metabolomics) only the sample injection can be repeated, the sample workup can be repeated or the whole experiment can be repeated. The replication error in

the first case is usually referred to as repeatability while the other two (with larger errors) are referred to as reproducibility. Sometimes these replicates can also be referred as biological replicates and technical replicates. Biological replication is usually the most important. This is because experiments are usually done to make statements about whole populations, rather than single samples, for instance: "gene X responds to compound Y in mice. The researcher is interested in responses in mice, and not in a single mouse. Hence biological replicates of mice should be taken that are samples from the total population of mice. Technical replication is usually applied to gain statistical power.

Randomization requires the experimenter to use a random choice for every factor that is not of interest but might influence the outcome of the experiment. A simple example is the hybridization of mRNA samples from an experiment comparing a treatment group VS a control group. Hybridization is very sensitive to external conditions and the efficiency can vary during the day. If all control samples are measured first and all treated samples are measured after, then it will be impossible to distinguish between an un-interesting time effect and the interesting treatment effect. These two factors are confounded. If measurement of controls and treated animals are randomly distributed over the day, the bias is eliminated.

Blocking is the arranging of experimental samples in groups (blocks) that are similar to one another. The experimental conditions within a block are expected to be more homogeneous than among blocks. Typically, a blocking factor is a source of variability that is not of primary interest to the experimenter. A simple example is the use of two LC columns in an experiment comparing a treatment group VS a control group. If all control animals are tested using one column and all treated animals are tested using the other column, it will be impossible to distinguish between the un-interesting column effect and the interesting treatment effect. These two factors are confounded. If half of the controls and half of the treated animals are measured with one column and the rest of the animals are analyzed with the second column, the bias is eliminated and the influence of this factor reduced. Blocking is a very powerful tool for experimenters which should always be considered whilst designing an experiment, even more powerful than randomization. The reason is that block effects can be estimated using proper statistical methods, and therefore blocking can increase the statistical power. This is not possible with randomization; randomization can only alleviate the bias. Therefore a simple general rule: "block what you can, randomize what you cannot".

For proteomics and metabolomics, the analytical instruments GCMS and LCMS are not fully stable and may show drift in time. When the number of samples is large and several measurement batches are needed this may lead to problems as instrumental drift causes samples to be measured in the beginning of a series slightly different than when these samples were measured at the end of the series. The order in which the samples are measured is therefore of importance and this is often defined in the *measurement design*.

Samples of healthy controls and patients suffering from metabolic syndrome cannot be measured in separate batches. In that case the observed difference could be solely due to the instrumental drift. Here randomization is crucial.

12.3 DATA PREPROCESSING AND QUALITY CONTROL

Data preprocessing and data pretreatment methods are used to improve the information content of the data. During the process of sampling, sample workup, measurement etc., small or larger disturbances can enter the data. For metabolomics, two types of disturbances can be distinguished:

- Disturbances of a whole sample. Samples may not be comparable due to various reasons. Sometimes the amount of sample measured can be different. The dilution of samples may be different (morning urine is more concentrated than urine obtained later during the day). The sample workup is not exactly equal for all samples. The order of measuring (beginning of day or end of day) has an effect on all measured intensities in the sample.
- Disturbance of a single variable within a sample. The measurement of a single metabolite in a sample may be disturbed e.g. by pH of the sample or by instrumental issues that are different for different variables.

These two different types of disturbances require different preprocessing methods to correct for the disturbances. The correction for the first type of disturbances (affecting the whole sample) is often performed with *normalization methods*. A much used method to correct for measuring a different amount of sample in the *internal standard*. This is a compound that is added to each sample in equal amount. Thus the intensity of this compound has to be the same in all samples. When the intensity of the internal standard in different samples differs, this difference can be used to correct all variables in the sample.

Instrumental drift during the day can be corrected with *quality control (QC) samples*. Often pooled samples are used (a mix of all samples) such that each metabolite is present in this QC sample. After say every 8 samples the QC sample is measured. Thus over the whole day the QC sample is measured many times. The intensity of each metabolite should be the same but due to the instrumental drift it may not the same. However the differences can again be used to correct the study samples that are in between the QC samples.

When dealing with urine samples in proteomics or metabolomics, it is known that morning urine is less diluted than urine obtained later during the day, and concentrations are therefore higher. If we correct this difference in concentration by a certain “concentration measure” we do not have true concentrations anymore, but the samples become better comparable. This correction is also called **normalization.**

Sometimes single variables are distorted during the measurement of the sample. In LCMS and GCMS the chromatographic column may deteriorate due to aging. This may change the retention time of the compounds. This means that the same metabolite appears at different times. *Alignment methods* can be used to align the peaks at different retention times for different samples such that it is clear these peaks belong to the same metabolite or protein. Another problem occurs when the background signal is unequal to zero. Thus even if the concentration for a metabolite is 0, there is still a signal. For such problems *background correction methods* have to be used.

When all necessary data preprocessing has been performed the clean data is usually stored in a data matrix. Thus, the metabolomics data, are going through a process of preprocessing and normalization, and this will result in a **normalized data matrix**. This normalized data matrix is the starting point for data analysis and biological interpretation.

	XVar 1	XVar 2	XVar 3	XVar 4
Object 1	**0.0245**	**2.342**	**0**	**0.034**
Object 2	**0**	**5.112**	**200.1**	**0.032**
Object 3	**0.032**	**6.321**	**800.4**	**0.032**
Object 4	**0.01**	**0**	**732.2**	**0.031**

Figure 12.1. An example of a data matrix

A data matrix contains values for multiple objects and for multiple variables. Each row (from left to right) in this data matrix refers to the data measured for a single object (sample or individual). Each column (from top to bottom) contains the data of a given variable for all objects in the study. In this small example the levels of 4 different variables (XVar 1:XVar 4) are given for 4 different objects.

There is often a relationship between the different objects. Multiple objects can belong to the same group (e.g. patients or healthy), they can belong to the same individual (e.g. before and after treatment. The values in the data matrix (e.g. 0.0245) can be an intensity value or it can also be a concentration. Sometimes the variable can be qualitative (0 vs 1) to represent different groups (patient vs healthy). There is usually a lot of additional information needed to fully understand the data. The 0 can mean not present, or it can be below detection limit, etc. The different variables are often measured in the same units, but this is not always the case. E.g. variable 3 can be measured in mmol/l while the others are measured in µmol/l.

12.4 DATA ANALYSIS.

12.4.1 Exploratory data analysis

After the normalized data matrix has been obtained, the next step is to explore the data. The aim of exploratory data analysis is to summarize the main characteristics in the high dimensional data in easy-to-understand form, often with visual graphs, without using a statistical model or having formulated a hypothesis. Exploratory data analysis enables the researcher visually to examine their data sets, infer the impact of the experimental perturbations and to determine the relevance of nuisance factors. If the outcome of the exploratory data analysis is that a noise factor is influencing the results, then it is wise to go back to "Data processing", to find solutions in preprocessing and normalization to minimize the impact of the noise factor.

Thus, the goal of exploratory data analysis is to see whether strange phenomena, grouping between samples and or variables occurs in the data. As the number of variables is usually so large that this cannot be done for each variable separately we use multivariate explorative data analysis methods to explore the data. The two mainly used methods to do this are principal component analysis (PCA) and clustering analysis. Clustering is also important for the biological interpretation, and will be discussed in the another chapter. Here an introduction in PCA is given below.

Principal component analysis

Principal component analysis (PCA) is a explorative method that summarizes the complex high dimensional data with a few summarizers. These summarizers are chosen in such a way that they all explain different variation in the data, where the first one describes the most variation. Together all summarizers explain the data as well as possible. Now each summarizer can be explored separately or in pairs which is much easier to do than exploring the high dimensional data in one go.

One of the main reasons to use multivariate analysis to study the data is explained in this two variable example below. The difference between the green group and the orange group is not present in either Var 1 or Var 2. Only when both variables are considered simultaneously then the difference between the groups can be observed. Multivariate analysis takes the correlations between the variables into account when describing the data and looking for structures and differences between the samples.

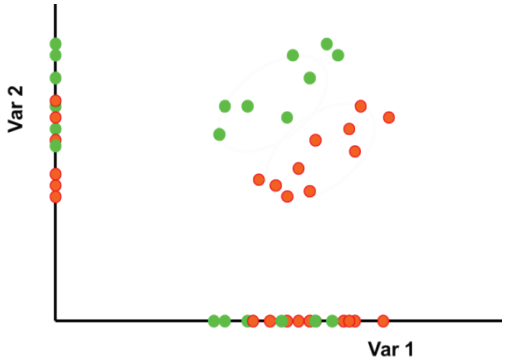


Figure 12.2. Exampling showing the need for multivariate analysis

The PCA model is described as :

$$\mathbf{X} = \mathbf{t_1 p_1^T} + \mathbf{t_2 p_2^T} + \cdots + \mathbf{E}$$

$$\mathbf{X} = \mathbf{T P^T} + \mathbf{E}$$

Here **X** is the high dimensional data of size (*I* objects x *J* variables) described above. $\mathbf{p_1}^T$ is the first loading vector that summarizes the systematic relationship between the variables in the data. E.g. if in the data high values for variable 1 always appear with low values of variable 2 and otherwise, than such relationship will be represented by opposite sign for the values of these two variables in the loading. The corresponding score vector $\mathbf{t}_1$ contains information for each object how much variation they have in the direction of the first loading. Thus if object 1 has very high value for variable 1 and very low for variable 2, it will have a higher score value than object 2 with less high and low values for these variables. $\mathbf{t}_1$ is a column vector of size (*I*x1) and $\mathbf{p}_1^T$ is a row vector of size (1x*J*). Multiplying these vectors as represented in the equation leads to a matrix of size (*IxJ*) with all information described by the first principal component (PC). Similar to the first (PC) the second one describes systematic variation in the data other than described by the first. This can go on until all variation in **X** has been described. When not all PC's are used, then some of the variation in **X** is not described. This is called the residual (**E**) and usually contains nonsystematic noise. The second equation puts all scores in 1 matrix **T** and all loadings in one matrix $\mathbf{P}^T$.

To explore the variation in the data we can now focus on a score plot (values of two scores plotted against each other). Here you can observe how the objects relate to eachother. You may observe object clusters, outlying objects etc. The distance between objects in the score plot represents the distance of the objects in the full data **X**. If there is a logical order in the objects (e.g. time) then the scores can also be plotted vs the logical order.

The loading plot (values of two loadings plotted against each other) describes the relationship between the variables in a similar way as described for the score plot. If there is

a logical order in the variables (e.g. retention time) then the loadings can also be plotted vs the logical order.

In the example paper (Ramadan *et al.*, 2006) PCA is used to explore the NMR data of plasma of healthy men and women. After appropriate pretreatment (centering and scaling) a score plot and a loading plot were made to explore the variation in the NMR data (Figure 12.). In the score plot we see that especially the score of PC2 shows clear separation between the females (red) and males (blue). In the loadings plot we see that variables (0.86, 0.84, 1.24, 1.26) have high loadings for this PC and may be important peaks in the data to distinguish the males from the females. Although the first PC already describes 53.8% of the variation in the data, this variation is between the healthy individuals in general and not specific for male / female differences.

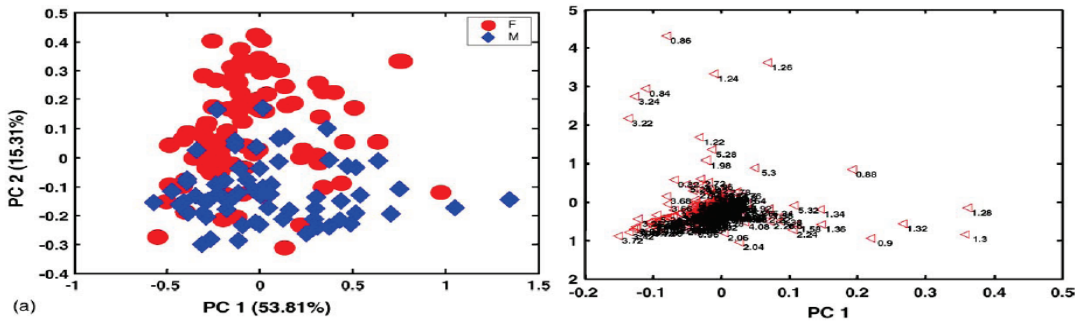


Figure 12.3. A score plot and a loading plot of NMR data

12.4.2 Multivariate hypothesis testing

Principal Component Analysis is a multivariate explorative method that aims to describe as much variation in the data with as few as possible principal components. In Figure 11.3 we saw that this does not mean is focuses on differences between groups, as the groups differences were only found in the 2nd PC.

A slight adjustment of the PCA model is usually used to calculate optimal discrimination models from multivariate data. Such a method can be used to calculate adjusted scores and loadings that are better able to discriminate the cases from the controls. The default approach for this in metabolomics is called partial least squared discriminant analysis (PLSDA). This method combines all measured metabolites simultaneously to make a prediction model that can predict to which class a sample belongs. A PLSDA regression coefficient is calculated for each metabolite ($\mathbf{b_{PLSDA}}$), and multiplying these PLSDA regression coefficients times their corresponding metabolite concentrations gives the prediction of the class.

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{\mathbf{PLSDA}} + \mathbf{f}$$

All metabolite concentrations are in matrix **X** and $\mathbf{b}_{PLSDA}$ is the vector that contains the weights for each metabolite. The vector **y** contains the class label of the group (e.g. 0 or 1), and **f** contains the prediction errors. The important metabolites for class prediction (or biomarkers) are those metabolites that have high (absolute) weights.

The PLSDA model is supervised, which means that the class information is used to calculate the scores. Above in PCA, the loading p was calculated such that the variation described in the scores was maximal. In PLSDA we want to have the variable loadings that provide scores with optimal discrimination between cases and controls. This is achieved by using the covariance between each variable and the vector **y**. The higher this covariance, the more important that variable is for the discrimination and thus the more weight it should have for defining the score.

$\mathbf{t}_{PLS} = \mathbf{X}*(\mathbf{X}^T\mathbf{y})$

Here the $\mathbf{X}^T\mathbf{y}$ is the covariance between each variable of **X** and the response vector **y**. The response **y** can now be predicted from the scores: $\mathbf{y} = \mathbf{t}_{PLS}*q$, where q is just the slope between **y** and $\mathbf{t}_{PLS}$.

Similar to PCA, in PLSDA also multiple components can be calculated. Before a new component is calculated, the information from the 1st PC is removed from the data **X**.

$\mathbf{E} = \mathbf{X} - \mathbf{t}_{PLS}*\mathbf{p}_{PLS}^T$

The second component is then calculated from **E**: $\mathbf{t2}_{PLS} = \mathbf{E}*(\mathbf{E}^T\mathbf{y})$.

New components are also orthogonal to the previous such that each component describes different information from the X-data. This is continued until prediction of **y** does not improve anymore. Finally, the covariances of all components together with the slopes q for each component are combined into the single PLSDA regression vector $\mathbf{b}_{PLSDA}$, which indicates the importance. Details how this is done exactly go beyond the material for this course.

12.4.3 Validation

As PLSDA is a supervised method, meaning that we use class information to calculate the best $\mathbf{b}_{PLSDA}$ regression coefficients its performance is usually too good for the data used to make the model. Validation is a procedure to test whether the PLSDA regression coefficients can also make good predictions for new data that was not used to calculate the $\mathbf{b}_{PLSDA}$ coefficients.

The best approach for testing is to have a separate set of samples (or individuals) called the **test set**. It is most important that this **test set** is never used in the calculation of $\mathbf{b}_{PLSDA}$. Only then it can be tested how well the model is for new data. When a test set is not available the procedure is to separate the samples that we have into two parts, a **training set** and a **test set**. The training set is used to make the model and the test set is used to test it.

When the number of samples is very small, a procedure called cross validation is used. Here the selection of test and training set is repeated until each sample has been in the training set once. Then for each sample a prediction for the class can be given (while it is in the test set) and so an average prediction error for new data can be calculated.

Let's take an example of 30 cases and 30 controls, and 15 individuals have been predicted wrongly. Is 15 misclassifications out of 60 a good result? Is it better than what you can expect by chance? In order to know how well 15 misclassifications actually is we will compare it to situations when there is no difference between the cases and controls and then see how many misclassifications are made. This procedure is done using **permutations.** With **permutations** the class labels are randomly assigned to the samples. This makes that we expect no relation between the measured data and the permuted class labels. Then a model is developed just as was done for the original class labels. The permutation is repeated say 1000 times and then for each time a number of misclassifications is calculated. Finally, our 15 misclassifications can be compared to the 1000 numbers of misclassifications from the permutations. When no more than 5% of the number of misclassifications from the permutations is smaller than the 15, then we can conclude the 15 misclassifications is a significant result.

Similarly to know which of the variables in the data is informative for good classification we can look at the $\mathbf{b}_{PLSDA}$ value. High absolute values are more important than low absolute values, but when is a value high enough? Again we use the information of the permutations. For each permutation a $\mathbf{b}_{PLSDA}$ model is developed. However these values belong to a model from a permuted class label and therefore should be **not important.** Thus the $\mathbf{b}_{PLSDA}$ values of all of the variables from the permutations form a range of values that are nonsignificant. For $\mathbf{b}_{PLSDA}$ values to be significant they need to be outside of this nonsignificant range.

12.5 BIOLOGICAL INTERPRETATION

The next step in the data analysis pipeline is the biological interpretation of the data, or data mining. In general, the aim of the biological interpretation is to infer if expected biological effects are found and if novel discoveries can be made. From the validation approach a list of significant metabolites has been obtained. The goal is now to bring biological context into this list. Are specific pathways over represented in this list? Metabolite Set Enrichment Analysis is a method that can look for over representation of certain groups of metabolites. One approach is to calculate overrepresentation of selected metabolites in a group. As an example consider 4 functional groups of each 13 metabolites, thus 52 in total. In a study, 12 significant were selected. It turned out that 10 of the 12 belong to one of the four groups. What is the chance that 10 out of 12 metabolites would be selected from a single group if there would be no overrepresentation of one of the groups? Thus what is the change that such a result would occur just be chance? We need the **hypergeometrical distribution** to calculate this.

$$Probability = \frac{\binom{13}{10} \times \binom{39}{2}}{\binom{52}{12}}$$

Note $\binom{13}{10} = \left(\frac{13!}{10! * 3!}\right) = \left(\frac{13 * 12 * \ldots * 2 * 1}{10 * 9 * \ldots * 2 * 1 \ * \ 3 * 2 * 1}\right) = \left(\frac{13 * 12 * \ldots * 5 * 4}{10 * 9 * \ldots * 2 * 1}\right)$.

The denominator represents all possible selections of 12 from 52. The first term of the numerator represents the total number of selecting 10 metabolites from a possible 13 in a group. The second term of the numerator represents the possibilities of selecting 2 metabolites out of all metabolites from the other three groups. The result of this calculation is in the order of 1E-6, which is very small. Thus the chance that selecting 10 metabolites from the same group is very rare if there would not be a preference for a group. The 1E-6 can be considered a p-value.

A disadvantage of the overrepresentation procedure is the cutoff, which needs to be set to determine the list of important metabolites. Basically, any cutoff is arbitrary. It may also happen that established significance levels, such as **fdr corrected p-value ≤ 0.05**, no influential metabolites are found. It may then be useful to apply less conservative methods such as enrichment analysis. These tests are however treated in another part of the course.

12.6 REFERENCES

- Brown M et al. A metabolome pipeline: from concept to data to knowledge, Metabolomics,1, 39-51, (2005)
- Ramadan Z, Jacobs D, Grigorov M, Kochhar S. Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms. 2006. Talanta.68(5):1683-91.

13 Information management: public biological databases

Lecturer: Prof. dr. Antoine van Kampen (AMC)

After reading this chapter you should

- Understand the concept of a 'database'
- Have knowledge about few of the major databases (GenBank, UniProt, etc)
- Understand the difference between primary, secondary, and specialized databases
- Understand why we need different ways to access the public databases
- Understand why standardization is important, and know a few examples.
- Understand why we need minimum information standards
- Understand some of the challenges involved with biological databases

Contents

13.1 Introduction

One of the hallmarks of modern genomic research is the generation of enormous amounts of (sequence) data. As the volume of omics data grows, sophisticated computational methodologies are required to manage the data deluge. Thus, the very first challenge in the genomics era is to store and handle the staggering volume of information through the establishment and use of computer databases. The development of databases to handle the vast amount of molecular biological data is thus a fundamental task of bioinformatics, and has become crucial for the work of biologists and researchers from other disciplines. Information management has results in many (hundreds to thousands) public biological databases that provide a wealth of (freely) available information about, for example, nucleotide sequences, proteins and pathways.

"Information management" is the collection and management of information from one or more sources and the distribution of that information to one or more audiences.

In the current context we will use 'information' to denote all (experimental) data, information and knowledge that is available from public databases.

13.2 What is a database?

A **database** is a computerized archive used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria. Databases are composed of computer hardware and software for data management. The main objective of the development of a database is to organize data in a set of **structured records**. Each record, also called an entry, should contain a number of **fields** that hold the actual data items, for example, fields for sequences, organism, sequence type, and sequence features (e.g, exon regions). To retrieve a particular record from the database, a user can specify a particular piece of information, called value, to be found in a particular field and expect the computer to retrieve the whole data record. This process is called making a **query**.

Although data retrieval is the main purpose of all databases, biological databases often have a higher level of requirement, known as knowledge discovery, which refers to the identification of connections between pieces of information that were not known when the information was first entered. For example, databases containing raw sequence information can be used to identify sequence homology or conserved motifs within or across organisms. These features facilitate the discovery of new biological insights from raw data.

13.2.1 Types of databases

Text files

Originally, databases used a flat file format, which is a long text file that contains many individual entries (e.g., sequence records). Within each entry are a number of fields separated by tabs or commas. The text file can be considered a single database table. Thus, to search a

flat file for a particular piece of information, a computer has to read through the entire file, an obviously inefficient process. This is manageable for a small database, but as database size increases or data types become more complex, this database style can become very difficult for information retrieval. (Tab delimited) text files are easily important in Excel if they are not too large.

Database management systems

To facilitate the access and retrieval of data, sophisticated computer software programs for organizing, searching, and accessing data have been developed. They are called database management systems (DBMS). Microsoft Access and MySQL are examples of such systems. It is beyond the score of these lecture notes to explain these systems in more detail. However, they facilitate the organization, integration and querying of data.

13.3 Databases categories

Based on their contents, biological databases can be roughly divided into three categories: primary databases, secondary databases, and specialized databases.

Primary databases contain original experimental data. GenBank is an example of such database (see below).

Secondary databases contain computationally processed or manually curated information, based on original information from primary databases. Translated protein sequence databases containing functional annotation belong to this category. UniProt is an example of such database.

Specialized databases are those that cater to a particular research interest. For example, Flybase, HIV sequence database, and Ribosomal Database Project are databases that specialize, for example, in a particular organism, category of sequences, or other biological entity. Often such databases contain a mixture of data types. For example, FlyBase, a database for drosophila genetics and molecular biology, contains information about sequences, gene expression, phenotype, etc.

13.4 Why do we have public biological databases?

In contrast to some other research communities (e.g., chemistry), biologists working in the omics field generally share their data through the use of central public databases. The public data offers various advantages and uses:

- Other researchers are enabled to verify and reproduce results described in scientific publications;
- New data can be compared to public data (e.g., performing a BLAST search of your new sequence against GenBank to determine the origin or function of this new sequence);

- Public data can be retrieved and analyzed in new ways;
- Public data can be retrieved and integrated with your own experimental data to facilitate further analysis and/or interpretation;
- Data from several public databases can be retrieved and integrated for the discovery of new biological knowledge (data mining).

Many journals also require that experimental data described in a paper is made available through one of the public databases. This policy facilitates the collection, organization and dissemination of omics data. In general, when you submit data to a database (e.g. nucleotide sequences from Sanger sequencing to GenBank) you will obtain one or more unique identifiers for the data (experiment) that you submitted. These identifiers are referred to in the publication to facilitate retrieval of your data by other researchers.

13.5 How to find a database?

Many different public databases are available that provide information at various levels of biology (Figure 13.1). If you want to find a specific database then several resources are available that can help you:

- **Bioinformatics Institutes**. Most of the major databases are provided by bioinformatics institutes (e.g., NCBI, EBI, and SIB). You may visit their websites to obtain a list of databases that are provided.
    - NCBI: http://www.ncbi.nlm.nih.gov/
    - EBI: http://www.ebi.ac.uk/
    - SIB: http://www.isb-sib.ch/
- **Nucleotide Acids Research** (NAR; http://nar.oxfordjournals.org). NAR is a scientific journal that publishes the results of research into physical, chemical, biochemical and biological aspects of nucleic acids and proteins involved in nucleic acid metabolism and/or interactions. Each January issue is devoted to the NAR Molecular Biology Database Collection (see below).
- **GeneCards** (http://www.genecards.org). GenCards is a public database that is linked to many other databases and, therefore, may provide a good starting point to your search.
- **Google**

NAR Molecular Biology Database Collection

This database collection comprises a list of databases (not the databases itself) and their descriptions. This list currently (2019) comprises 1613 databases that are categorized as shown in Figure 13.2. The January NAR issue includes a number of scientific papers that describe new databases or about important updates of existing databases. For example, the latest issue comprises 168 papers describing 64 new databases and 92 status updates. Although this collection already comprises 1000+ databases, it is not exhaustive.

The NAR collection, however, provides a good starting point because of the criteria for inclusion. The database should:

- be thoroughly curated (see 13.7.3.1)
- be of interest to a wide variety of biologists
- be comprehensiveness of coverage
- have a degree of added value (e.g., manual curation)
- likely to be maintained for a long period of time

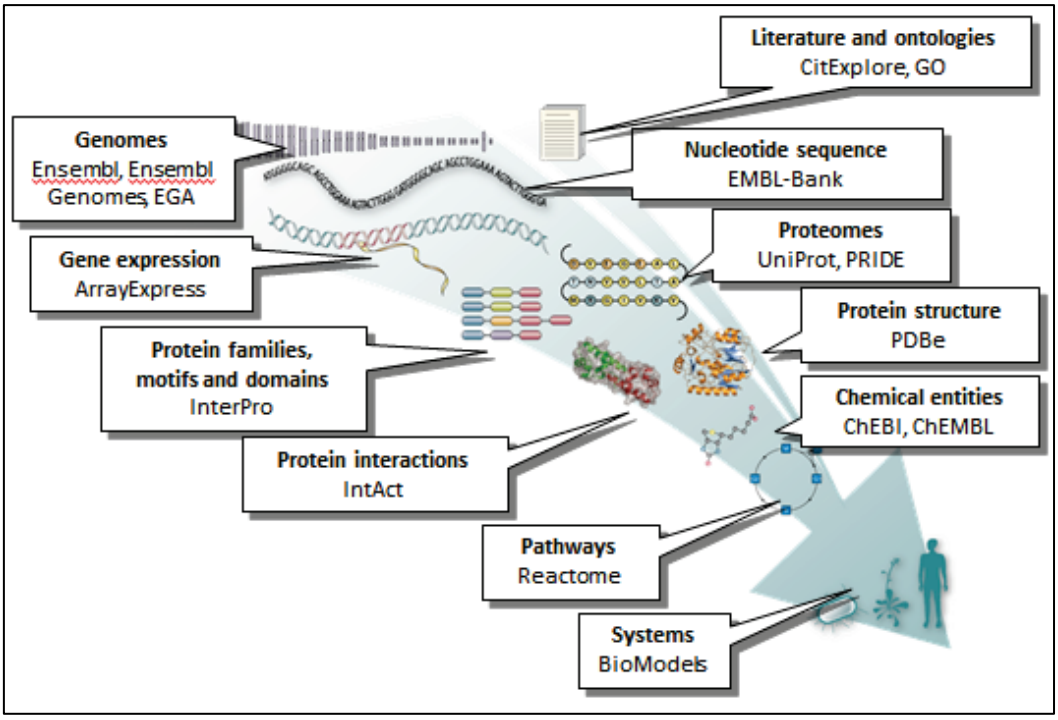


Figure 13.1. Examples of information that is available through public biological databases. Information comes from various levels of biology ranging from DNA sequences to mathematical models of biological systems. The type of information is shown in bold face. Databases names are provided below the type. Figure from the European Bioinformatics Institute.

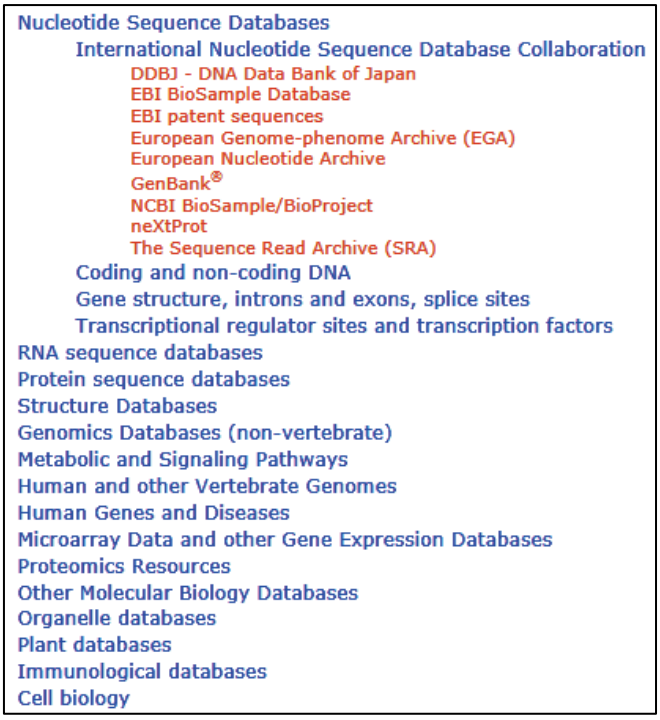


Figure 13.2. Categorization of public databases in the NAR Molecular Biology Database Collection. For example, GenBank is sub-categorized as an 'international nucleotide sequence database collaboration'.

13.6 Examples of public biological databases

Note: you do not have to memorize the names of all these databases. They just provide a few examples of data that you can freely obtain through the public biological databases.

This section will introduce several public biological databases but also to some of the associated (standardization) initiatives. A selection of databases is used to demonstrate the wealth of information that is available. However, we will not thoroughly review the content of each of these databases but focus on specific aspects of these databases that are important to obtain a good understanding of some general principles that underlie these databases. Some of the databases introduced in this section will be used in the computer exercises.

Due to policies of scientific journals and funding agencies, omics data is often made available to the research community via public primary databases. In addition, a wide range of secondary databases have been developed. Biological repositories do not merely archive data and mathematical models but also serve an important role in research.

The main repositories (**primary databases**) are hosted and maintained by the major bioinformatics institutes including EBI, NCBI, and SIB that make a major part of the raw experimental omics data available through a number of primary databases including GenBank[1], GEO[2], PRIDE[3], and Metabolights[4] for sequence, gene expression, MS-based proteomics and MS-based metabolomics data, respectively.

In addition, many **secondary databases** provide information derived from the processing of primary data, for example pathway databases (e.g., Reactome[5], KEGG[6]), protein sequence databases (e.g., UniProtKB[7]), and many others. Pathway databases provide an important resource to construct mathematical models used to study and further refine biological systems[8,9]. Other efforts focus on establishing repositories integrating information from multiple public databases. Joint initiatives of the bioinformatics and systems biology communities resulted in repositories such as BioModels, which contains mathematical models of biochemical and cellular systems [10], and Recon 2 that provides a community-driven, consensus 'metabolic reconstruction' of human metabolism suitable for computational modelling [11]. Another example of a database that may prove to be of value for systems medicine studies is MalaCards, an integrated and annotated compendium of about 17,000 human diseases [12]. MalaCards integrates 44 disease sources into disease cards and establishes gene-disease associations through integration with the well-known GeneCards databases [13,14]. Integration with GeneCards and cross-references within MalaCards enables the construction of networks of related diseases revealing previously unknown interconnections among diseases, which may be used to identify drugs for off-label use.

Databases are used routinely in the analysis, interpretation, and validation of experimental data. For example, the Gene Ontology (GO; [54]) provides a controlled vocabulary of terms for describing gene products, and is often used in gene set analysis to evaluate expression patterns of groups of genes instead of those of individual genes.

13.6.1 Phenotype databases

Phenotype databases are of particular interest to systems medicine. One well-known phenotype repository is the OMIM database, which primarily describes single-gene (Mendelian) disorders [15]. The integration of phenotype repositories with genetic and other molecular information will be a major aim for bioinformatics in the coming decade enabling, for example, the identification of co-morbidities, determination of associations between gene (mutations) and disease, and improvement of disease classifications[16].

13.6.2 Clinical data

To implement and advance systems medicine to the benefit of patients' health, it is crucial to integrate and analyze molecular data together with de-identified individual-level clinical data complementing general phenotype descriptions. Patient clinical data refers to a wide variety of data including basic patient information (e.g., age, sex, and ethnicity), outcomes of physical examinations, patient history, medical diagnoses, treatments, laboratory tests, pathology reports, medical images, and other clinical outcomes. Inclusion of clinical data allows the stratification of patient groups into more homogeneous clinical subgroups. Availability of clinical data will increase the power of downstream data analysis and modelling to elucidate molecular mechanisms, and to identify molecular biomarkers that predict disease onset or

progression, or which guide treatment selection. In biomedical studies clinical information is generally used as part of patient and sample selection, but some omics studies also use clinical data as part of the bioinformatics analysis (e.g., [17,18]). However, in general, clinical data is unavailable from public resources or only provided on an aggregated level. Although good reasons exist for making clinical data available, ethical and legal issues comprising patient and commercial confidentiality, and technical issues are the most immediate challenges [19,20]. This potentially hampers development of systems medicine approaches in a clinical setting since sharing and integration of clinical and nonclinical data is considered a basic requirement [21].

Although clinical data is not yet available on a large scale, the bioinformatics and medical informatics communities have been very active in establishing repositories that provide clinical data. One example is the Database of Genotypes and Phenotypes (dbGaP) [22] developed by the NCBI. Study metadata, summary level (phenotype) data, and documents related to studies are publicly available. Access to de-identified individual-level (clinical) data is only granted after approval by an NIH data access committee. Another example is The Cancer Genome Atlas (TCGA), which also provides individual-level molecular and clinical data through its own portal and the Cancer Genomics Hub (CGHub). Clinical data from TCGA is available without any restrictions but part of the lower-level sequencing and microarray data can only be obtained through a formal request managed by dbGaP.

13.7 A selection of databases

Examples of major databases are:

- Literature databases (**Medline and PubMed;** http://www.ncbi.nlm.nih.gov/pubmed)
- Nucleotide database (**GenBank;** http://www.ncbi.nlm.nih.gov/genbank)
- Hub to other databases (**GeneCards**; http://www.genecards.org)
- Human genome browser (**Ensemble;** http://www.ensembl.org)
- Gene expression (**Gene Expression Omnibus;** http://www.ncbi.nlm.nih.gov/geo)
- Protein sequences (**UniprotKB;** http://www.uniprot.org)
- 3D protein structures (**PDB**; http://www.rcsb.org)
- Small compounds (**ChEBI**; http://www.ebi.ac.uk/chebi)
- Biological pathways (**Reactome;** http://www.reactome.org)
- Protein interactions (**STRING;** http://string-db.org)
- Online Mendelian Inheritance in Man (**OMIM**; http://www.ncbi.nlm.nih.gov/omim)
- Ontologies and vocabularies (**GO;** http://www.geneontology.org)

Visit some of these databases to get a feeling of their content.

A selection of these databases is described below.

13.7.1 Literature databases: Medline and PubMed

MEDLINE is the National Library of Medicine (NLM) journal citation database. Started in the 1960s, it now provides over 24 million references to biomedical and life sciences journal articles back to 1946. MEDLINE includes citations from approximately 5,600 scholarly journals published around the world. The MEDLINE database is directly searchable from NLM as a subset of the PubMed database (http://www.ncbi.nlm.nih.gov/pubmed). PubMed has been available since 1996. Its over 22 million references include the MEDLINE database plus several other types of citations.

Many public biological databases such as GenBank are directly linked to Pubmed, and sometimes literature records are linked to biological resources.

Also keep in mind that the National Center for Biotechnology Information (NCBI) provides a range of online text books (http://www.ncbi.nlm.nih.gov/books)

13.7.2 Nucleotide database: GenBank

GenBank (http://www.ncbi.nlm.nih.gov/genbank) is one of the many database provided by NCBI. GenBank is the NIH genetic sequence database, an annotated collection of all publicly available nucleotide sequences (DNA and RNA). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the EMBL database of the European Bioinformatics Institute (EBI), and GenBank at NCBI. These three organizations exchange data on a daily basis.

Size of GenBank

To get a first indication about the size of this database, consider its statistics from Release 224.0 (February 15 2018):

- 253,630,708,098 bases,
- 207,040,555 sequences
- ~400.000 formally described species
- Required Storage 871 Giga bytes

GenBank grows exponentially and many new sequences (from additional organisms) are submitted to GenBank every day. Note that GenBank does not contain sequence data from NGS technologies, which are contained in the Small Read Archive of NCBI.

GenBank is linked to other databases and software

GenBank is linked to many other (biological) databases that are provided by NCBI. For example, the GenBank database is linked to the protein database and to PubMed. This makes it easy to retrieve lots of information about a gene of interest. GenBank is also linked to online software applications. For example, BLAST allows querying the GenBank database. Because the databases are linked and integrated with various software tools, and because these

databases and applications can be accessed through a web-browser they provide an excellent research tool to the biologist. Because these are available online, you don't have to download the complete database or install the software (although this is possible if you want to).

Searching GenBank

GenBank can be queried in two manners:

1. **Text-based search**. NCBI's Entrez system allows retrieving all records by entering one or more keywords (such as gene names) but you can also construct more complex queries (see http://www.ncbi.nlm.nih.gov/books/NBK44863/).
2. **BLAST**. BLAST allows you to find sequences that are similar to your input sequence.

GenBank content

For an example record (Human HBA1 gene) see powerpoint presentation. Typically, a GenBank record contains the following information (information between the braces provide an example from GenBank record NM_000558):

- **Definition** (Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA);
- **Locus** (NM_000558, 576 bp mRNA);
- **Accession code** (NM_000558). This is a unique identifier for the GenBank sequence, which is also used when citing information from GenBank;
- **Organism** (Homo sapiens; Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo). This includes the full taxonomy;
- **Literature references** (PUBMED 23123858);
- **General comments;**
- **Sequence features** (type, location, description, links to other databases). Such features include, for example, location of exons, polyA signals, etc);
- **Protein sequence.** If possible the nucleotide sequence is translated to the protein sequence and this protein sequence is added to the GenBank record;
- **Nucleotide sequence** (the 576bp mRNA sequence).

Access to GenBank

GenBank can be accessed in several ways:

- **Web-browser**. You can visit GenBank through your web-browser at http://www.ncbi.nlm.nih.gov/genbank. The web-interface allows you to retrieve records from the database and to present/visualize such record in various ways. This is the easiest way to use GenBank;

- **Download specific record (sequence)**. Once you have retrieved a specific record in your web-browser, you can download this specific record to your own computer in any of (currently) 11 different formats! This allows you to copy information directly in your publication, or to further analyse the sequence that you downloaded;

- **File Transfer Protocol (ftp)**. For many research applications it will be necessary to download not a single sequence but, for example, all human sequences or even the complete GenBank database. This is possible since GenBank information is available through a large number of text files each containing many sequences belonging to a specific category (e.g., human sequences). To download ('transfer') such file, you need to use the File Transfer Protocol (ftp). There are two ways of doing this. You can either use your browser to go to the ftp site of GenBank (**ftp**://ftp.ncbi.nih.gov/genbank) and download the required files. However, if you need many files, it is much more convenient to use a dedicated ftp program (e.g., FileZilla; http://filezilla-project.org/).

- **Web-services** (e-utils). A final approach to accessing GenBank data is provided by so-called 'web-services'. Web-services provide a set of software applications (e-utils) that are installed on the computer servers of NCBI, but which you can access and use through your own software. E-utils allows you to 'automate' specific queries or to perform queries that are not possible through the web-interface, without the need to download the complete GenBank database to your own computer. However, the intensity of the queries you perform through e-utils is limited to avoid that computer systems at the NCBI will crash. Figure 13.3 shows a simple example of how e-utils is used.

It is important to realize that data from GenBank (but also from other databases) can be accessed is various ways. For many applications, the standard web-interfaces will not suffice and you will need to use other approaches to access and/or download the data.

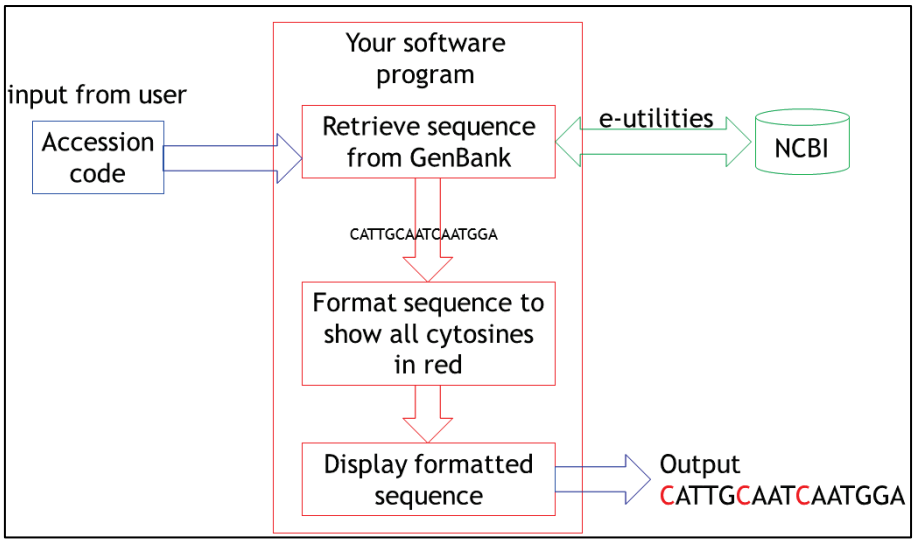


Figure 13.3. Example: the use of e-utils to access GenBank. Suppose you have developed a software program (red boxes) that colors all 'C' nucleotides in a sequence red. Your program will ask to provide an accession code (unique identifier for a nucleotide sequence in GenBank; blue box). Subsequently, your program will use one of the e-util programs to retrieve the corresponding sequence from the GenBank database which is located at the NCBI (USA; green arrow, database). Once the sequence is retrieved your program can format and display the sequence. This way of presenting sequences is not possible through the web-interface of GenBank. Of course,

this is an artificial example. In practice you would probably perform a more complex analysis on the sequence instead of just coloring specific nucleotides.

13.7.3 Protein database: UniProt

Many aspects of GenBank (e.g., data access) are similar for the UniProt protein resource (http://www.uniprot.org). We will not repeat this. Instead, several aspects that are not applicable to GenBank are discussed here.

UniProt is a collaboration between the European Bioinformatics Institute (EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR). Across the three institutes around 90 (!) people are involved through different tasks such as database curation, software development and support.

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. UniProt actually comprises several databases (Figure 13.4):

- UniProt Knowledgebase (UniProtKB)
- UniProt Reference Clusters (UniRef)
- UniProt Archive (UniParc)
- UniProt Metagenomic and Environmental Sequences (UniMES).

The UniRef and UniMES databases are beyond the scope of these lecture notes.

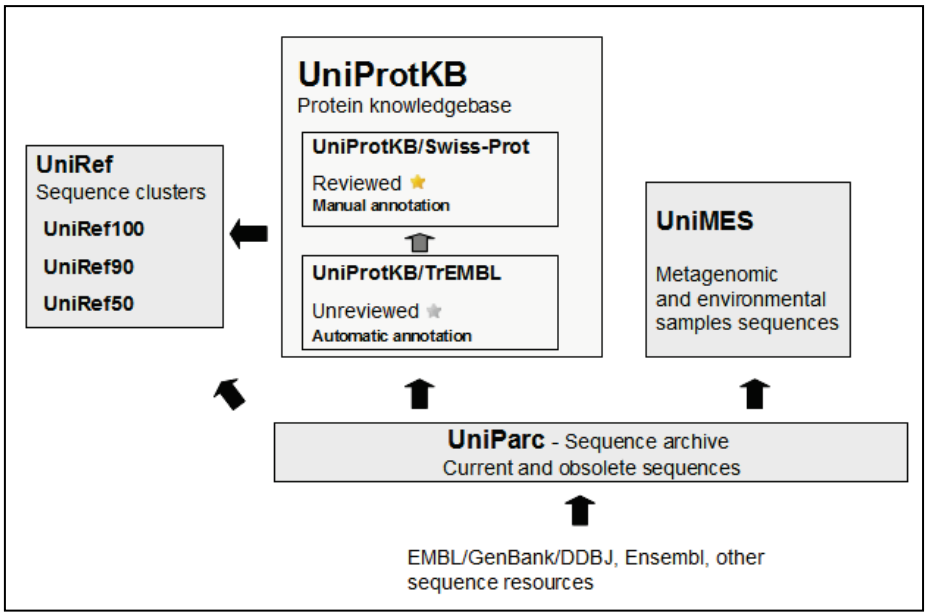


Figure 13.4. The Universal Protein Resource (UniProt) comprises several interconnected database.

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world. For example, UniParc retrieves about 98% of its protein sequences from the GenBank database, which were automatically translated from the mRNA sequences. Proteins may exist in different source databases and in multiple copies

in the same database. UniParc avoids such redundancy by storing each unique sequence only once UniParc contains only protein sequences (no annotation).

UniProtKB/TrEMBL takes protein sequences from UniParc and computationally generates annotations. Thus, these protein sequences are not yet reviews by a human expert. An example of a TrEMBL record for a human globin protein is shown below.

```
ID   Q86YQ4_HUMAN            Unreviewed;        88 AA.
AC   Q86YQ4;
DT   01-JUN-2003, integrated into UniProtKB/TrEMBL.
DT   01-JUN-2003, sequence version 1.
DT   28-NOV-2012, entry version 52.
DE   SubName: Full=Alpha-1 globin;
DE   Flags: Fragment;
GN   Name=HBA1;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE.
RC   TISSUE=Blood;
RA   Elam D., Holley L., Glendenning M., Harbin J.B., Kutlar A., Kutlar F.;
RL   Submitted (NOV-2002) to the EMBL/GenBank/DDBJ databases.
CC   -!- FUNCTION: Involved in oxygen transport from the lung to the
CC       various peripheral tissues (By similarity).
CC   -!- SIMILARITY: Belongs to the globin family.
CC   ---------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   ---------------------------------------------------------------------
DR   EMBL; AY178733; AAO22464.1; -; Genomic_DNA.
DR   IPI; IPI00853068; -.
DR   UniGene; Hs.449630; -.
DR   UniGene; Hs.654744; -.
DR   HSSP; P01922; 1C7C.
DR   ProteinModelPortal; Q86YQ4; -.
DR   STRING; Q86YQ4; -.
DR   PRIDE; Q86YQ4; -.
DR   ArrayExpress; Q86YQ4; -.
DR   GO; GO:0005833; C:hemoglobin complex; IEA:InterPro.
DR   GO; GO:0020037; F:heme binding; IEA:InterPro.
DR   GO; GO:0005506; F:iron ion binding; IEA:InterPro.
DR   GO; GO:0019825; F:oxygen binding; IEA:InterPro.
DR   GO; GO:0005344; F:oxygen transporter activity; IEA:UniProtKB-KW.
DR   Gene3D; 1.10.490.10; Globin_related; 1.
DR   InterPro; IPR000971; Globin.
DR   InterPro; IPR009050; Globin-like.
DR   InterPro; IPR012292; Globin_dom.
DR   InterPro; IPR002338; Haemoglobin_a.
DR   InterPro; IPR018331; Haemoglobin_alpha_chain.
DR   PANTHER; PTHR11442:SF14; PTHR11442:SF14; 1.
DR   Pfam; PF00042; Globin; 1.
DR   PRINTS; PR00612; ALPHAHAEM.
```

```
DR   SUPFAM; SSF46458; Globin_like; 1.
DR   PROSITE; PS01033; GLOBIN; 1.
PE   2: Evidence at transcript level;
KW   Heme; Iron; Metal-binding; Oxygen transport; Transport.
FT   NON_TER       1      1
SQ   SEQUENCE   88 AA;  9482 MW;  383AD3456B489A9C CRC64;
     QVKGHGKKVA DALTNAVAHV DYMPNALSAL SDLHAHKLRV DPVNFKLLSH CLLVTLAAHL
     PAEFTPAVHA SLDKFLASVS TVLTSKYR
//
```

UniProtKB/Swissprot is a high-quality and expert curated and annotated database of protein sequences. A record from this database looks similar but the annotation is often more extensive and precise, while erroneous information has been removed.

13.7.3.1 Curation and annotation

We already mentioned the words 'curation' and 'annotation' several times. These two concepts are of crucial importance for many databases. Note that, in contrast to the UniProtKB database, the GenBank database is not an expertly curated database.

To explain annotation and curation consider the following statement of Amos Bairoch (http://www.youtube.com/watch?v=ERHnBN4wnlw) one of the scientists involved at the Swiss Institute for Bioinformatics:

> *"It's quite depressive to think that we are spending millions in grants for people to perform experiments, produce new knowledge, hide this knowledge in often badly written text and then spend some more millions trying to second guess what the authors really did and found*".

There is a wealth of information available in scientific literature, but this is difficult to use (by computer programs) as part of analysis, to organize, to integrate, etc. Public database provide an extract from literature in electronic format. To do this, information from literature is transferred to database such as UniProtKB to precisely and comprehensively describe the protein sequences.

Annotation

Annotation is the association of (high throughput) data, such as protein sequences, with biological information from literature but also other databases. An annotation is a note, summary, or commentary on some section of data that is intended to explain or illustrate its meaning.

Annotation of protein sequences involves the inclusion of information such as:

- Function(s)
- Enzyme-specific information
- Biologically relevant domains and sites

- Post-translational modifications
- Sub-cellular location(s)
- Tissue specificity
- Developmental specific expression
- Structure
- Interactions
- Splice isoform(s)
- Associated diseases or deficiencies or abnormalities
- Links to other databases
- and much more

Biocuration

Biocuration is also called curation and goes a few steps beyond annotation. Annotation is the process of assigning information to data, while manual curation also involves collection, validation, quality control, use of standards, and communication. Consequently, a biocurator is a professional scientist with a background in life sciences who collects, annotates, and validates data and information that is disseminated by biological databases. His/her sole role encompasses quality control of primary biological research data intended for publication, extracting and organizing data from original scientific literature, and describing the data with standard annotation protocols and vocabularies that enable powerful queries and biological database interoperability. They communicate with researchers to ensure the accuracy of curated information and to foster data exchanges with research laboratories. Accurate and comprehensive representation of biological knowledge, as well as easy access to this data for working scientists and a basis for computational analysis, are primary goals of biocuration. Biocuration is very time consuming and it is virtually impossible to keep up with the flood of new data, which is the reason that other mechanisms are being investigated (such as the use of Wiki's).

UniProt annotation and curation

For example, one of the central activities of the UniProt Consortium is the biocuration of the UniProt Knowledgebase (UniProtKB). In order to respond to the flood of sequencing data, UniProt provides both manual curation by human experts (the curator) and automatic annotation by (statistical) software tools. UniProtKB consists of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. The former contains manually reviewed records with annotation extracted from the literature and curator evaluated computational analysis while the latter contains computationally generated records enhanced by automatic classification and annotation. In general, computational annotation is not free of errors and therefore manual curation is necessary as a second stage. Manual curation consists of a critical review of experimental and predicted data for each protein as well as manual verification of each protein sequence. Curation methods applied to UniProtKB/Swiss-Prot include manual extraction and structuring of information from the literature, manual

verification of results from computational analyses, data mining and integration of large-scale data sets, and continuous updating as new information becomes available.

13.7.4 Pathway databases: Reactome

Reactome (http://www.reactome.org) is an open-source, open access, manually curated and peer-reviewed pathway database. Pathway annotations are authored by expert biologists, in collaboration with Reactome editorial staff and cross-referenced to many public databases.

Examples of biological pathways in Reactome include signaling, innate and acquired immune function, transcriptional regulation, translation, apoptosis and classical intermediary metabolism.

13.7.4.1 Systems Biology Graphical Notation (SBGN)

One unique aspect of Reactome is to convey the rich information in the visual representations of biological pathways familiar from textbooks and articles in a detailed, computationally accessible format. Visualisation of Reactome data is based on the Systems Biology Graphical Notation (SBGN).

A typical figure from a scientific publication is shown in Figure 13.5. Although you will be able to understand (most of) this figure, it is important to realize that it contains many symbols and arrows that are not explicitly defined in the figure. The figure seems to use arrows for several purposes (activation, transport) and the functional differences between the colors and shapes of the boxes of the proteins is not clear. Moreover, nothing of the information in this figure (which is typically in a format such as pdf, jpg, png, gif, etc) is accessible to the computer. It is, for example, for a software application virtually impossible to extract the names of the proteins from this figure for further processing.

SBGN is an attempt to solve these problems. Figure 13.6 is an example of the glycolysis pathway in SBGN format. Each symbol and arrow in this figure is well-defined in the specification of SBGN, and each part of this figure is accessible through software applications. However, these figures seem less-attractive to the biologist than the classical text-book figures.

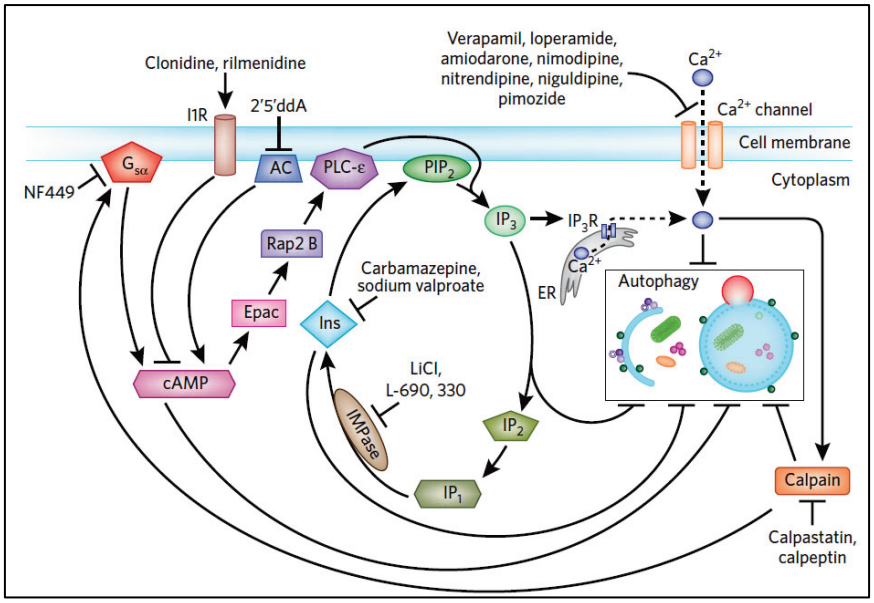


Figure 13.5. Typical figure from scientific publication. The caption for this figure described: "*Chemical screens for new autophagy-inducing agents have identified the cyclical Ca2+–calpain–Gαs and cAMP–Epac–PLC-ε–IP3 pathways as pharmacologically tractable for the modulation of autophagy. Inhibition of various components of these pathways results in autophagy induction. However, the precise mechanism by which levels of cAMP, Ca2+, calpain, inositol or IP3 control autophagy have yet to be elucidated*". (Figure copied from Fleming (2011) Nature Chemical Biology, 7, 9-17).

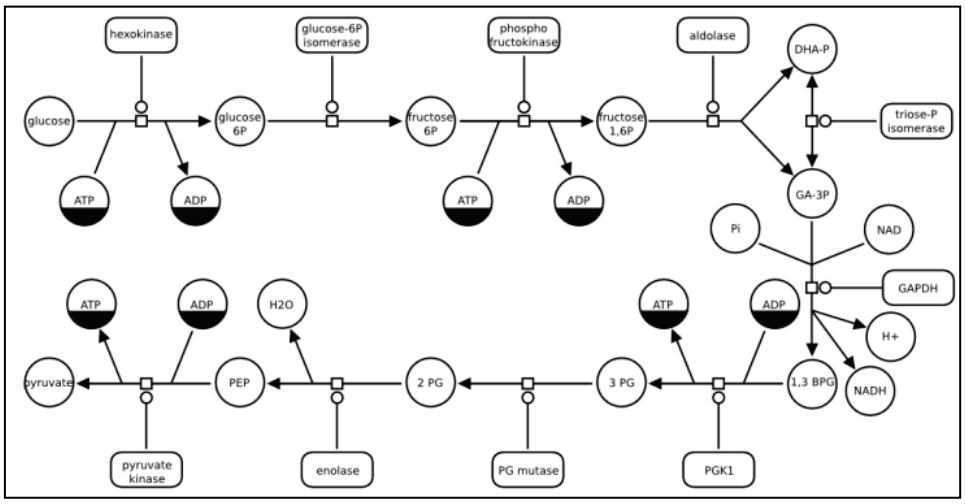


Figure 13.6. Glycolysis according to the Systems Biological Graphical Notation (SBGN). For a precise definition of the symbols go the SBGN website.

13.7.5 Gene Expression Omnibus (GEO)

The Gene Expression Omnibus (GEO) is a public repository that archives and freely distributes gene expression data produced by DNA microarrays, SAGE, next-generation sequencing, etc.

Minimum information guidelines

To **fully understand** the context, data, methods and conclusions of a specific experiment one must have access to a range of background information. The MIBBI project promotes efforts developing minimum information guidelines for the reporting of biological and biomedical science to the wider community. These guidelines also contribute to the **reproducibility of experiments**. MIAME (Minimum Information About a Microarray Experiment) is an example of such minimum information guideline. The MIAME guidelines outline the minimum information that should be included when describing a microarray experiment. Many journals and expression databases require microarray data to comply with MIAME. GEO deposit procedures enable and encourage submitters to supply MIAME compliant data.

The six most critical elements contributing towards MIAME are:

- The raw data for each hybridization;
- The final processed (normalized) data for the set of hybridizations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study);
- The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment);
- The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridizations are technical, which are biological replicates);
- Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number);
- The essential laboratory and data processing protocols (e.g., what normalization method has been used to obtain the final processed data).

13.7.6 The Gene Ontology

The Gene Ontology project provides an ontology of defined terms representing gene product properties. The ontology covers three domains:

- cellular component, the parts of a cell or its extracellular environment;
- molecular function, the elemental activities of a gene product at the molecular level, such as binding or catalysis;
- biological process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

For example, the gene product cytochrome c can be described by the molecular function term oxidoreductase activity, the biological process terms oxidative phosphorylation and induction of cell death, and the cellular component terms mitochondrial matrix and mitochondrial inner membrane.

The GO ontology is structured as a directed acyclic graph, and each term has defined relationships to one or more other terms in the same domain, and sometimes to other domains. The GO vocabulary is designed to be species-neutral, and includes terms applicable to prokaryotes and eukaryotes, single and multicellular organisms. An example of the use of GO is shown in Figure 13.7.

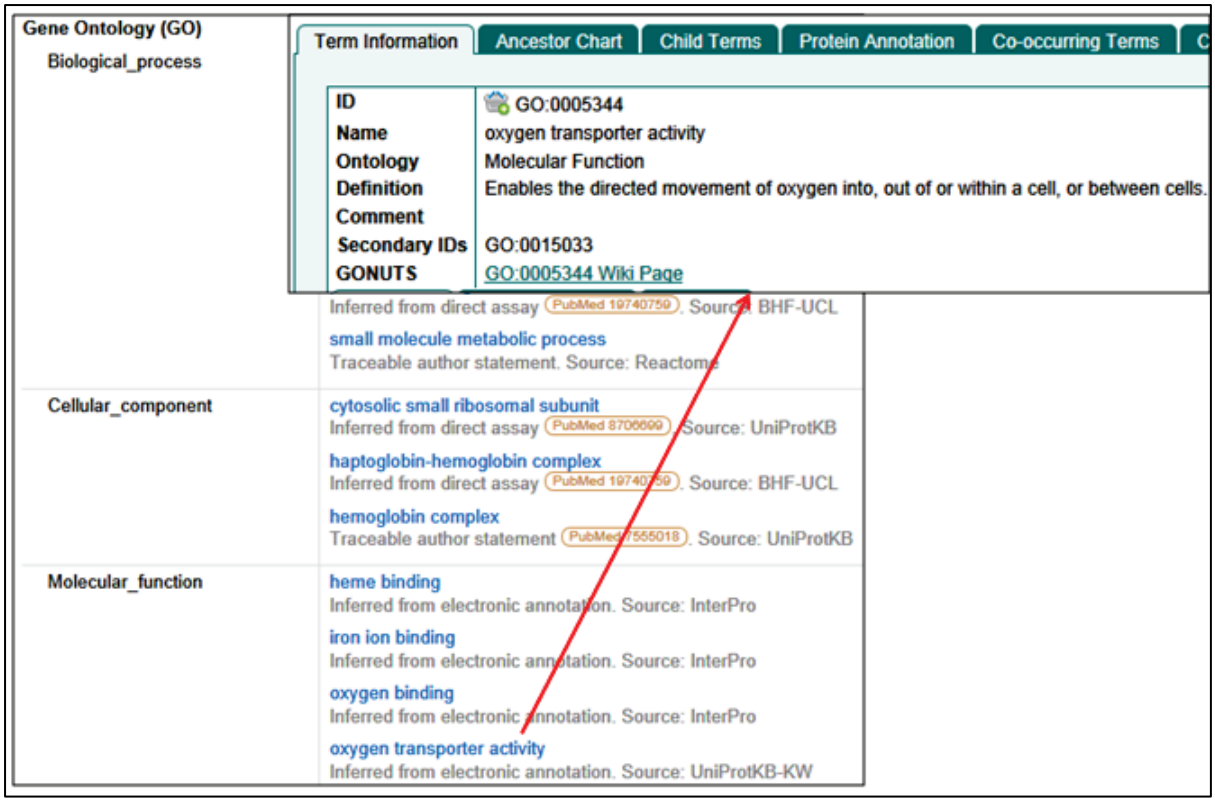


Figure 13.7. Small part of the annotation of the Human Globin protein in UniProtKB. The biological process, cellular component and molecular function are defined by using GO terms. For example, it uses the GO term 'Oxygen transporter activity' (with an unique identifier (GO:0005344) to precisely define one of the functions of this protein. The precise definition of oxygen transporter activity is defined in the Gene Ontology database (see inset). By using GO terms one avoids ambiguities in definitions (across databases). Figure 13.8 shows how this GO term is related to other GO terms in the Gene Ontology.

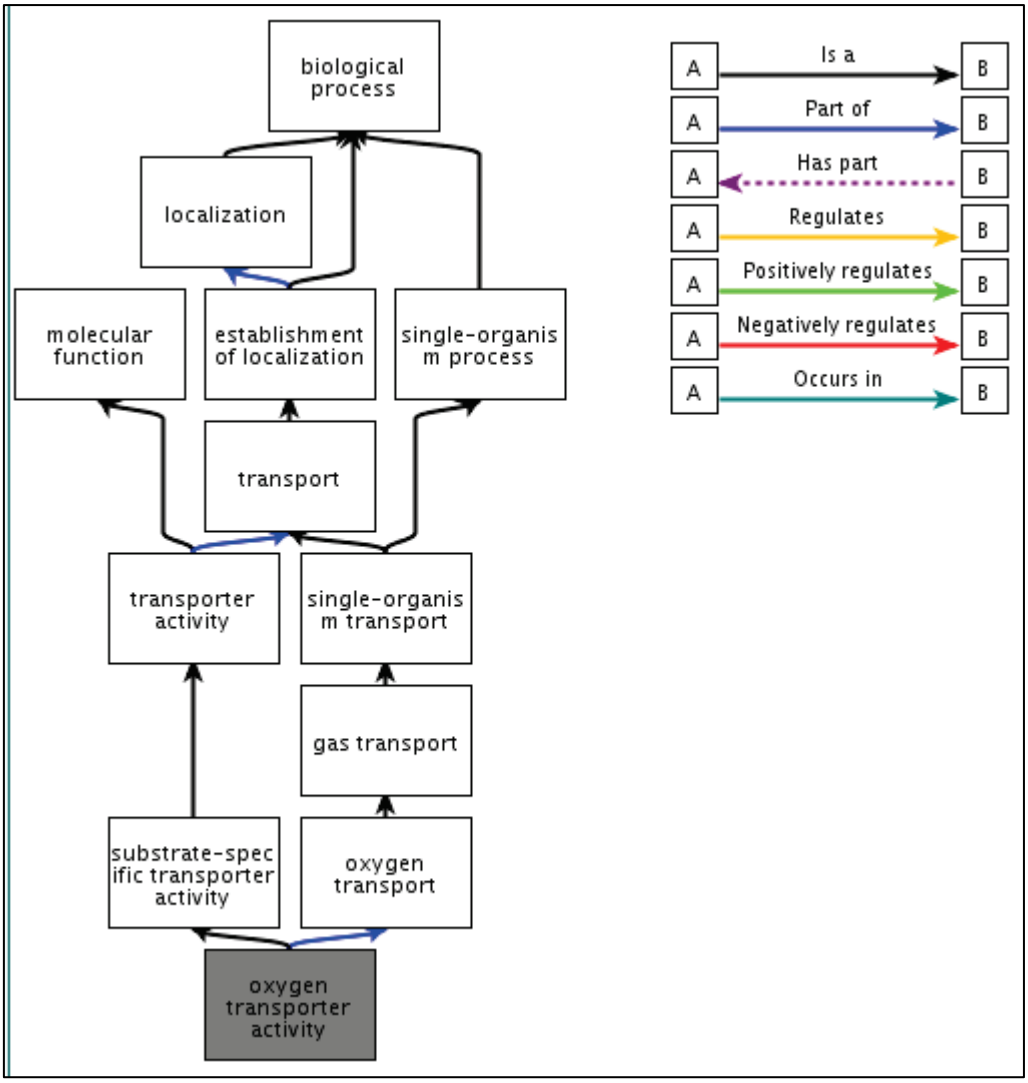


Figure 13.8. This figure shows all parents and relations of 'oxygen transporter activity' in the Gene Ontology.

13.8 Other aspects and challenges of public biological databases

This section discusses several other important aspects related to public biological databases. While reading this section, try to write down how these aspects apply to, for example, exome sequencing.

13.8.1 Access and Open Source

Open access is the practice of providing unrestricted access via the Internet to peer-reviewed scholarly journal articles. OA is also increasingly being provided to theses, scholarly monographs and book chapters. Open Access initiatives have resulted open access journals that provide free access to the full text of scientific papers. Similarly, the Open Data initiative is the practice of providing unrestricted access via the Internet to data such as provided by public biological databases.

In a similar way, software applications are made publicly available as Open-Source software.

Generally, open source refers to a program in which the source code is available to the general public for use and/or modification from its original design. Open source code is typically created as a collaborative effort in which programmers improve upon the code and share the changes within the community. Open source sprouted in the technological community as a response to proprietary software owned by corporations (e.g., Microsoft, IBM).

13.8.2 Fair Data Principles

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the **FAIR Data Principles**.

The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. The FAIR Principles put specific emphasis on enhancing the ability of computers to automatically find and use the data, in addition to supporting its reuse by individuals.

In the FAIR Data approach, data should be:

- Findable – Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets;
- Accessible – Stored for long term such that they can be easily accessed and/or downloaded with well-defined license and access conditions (Open Access when possible), whether at the level of metadata, or at the level of the actual data content;
- Interoperable – Ready to be combined with other datasets by humans as well as computer systems;
- Reusable – Ready to be used for future research and to be processed further using computational methods.

An important step in the FAIR Data approach is to publish existing and new datasets in a semantically interoperable format that can be understood by computer systems. By semantically annotating data items and metadata, we can use computer systems to (semi) automatically combine different data sources, resulting in richer knowledge discovery activities.

13.8.3 Explosive increase in number and size of databases

The number of biological databases is still increasing. We already described the large Molecular Database Collection of NAR with over 1000 databases. However, this raises several questions:

- How do you select the right database?

- How complete is the selected database
- What is the quality of the database
- Is the database easy accessible by the user or software tool?
- Does the database make use of standards?
- What is the redundancy within a database and with alternative databases?

Not all of these questions are always easy to answer.

In addition to a growing number of databases the amount of data is also growing explosively, mainly due to Next Generation Sequencing technologies. Figure 13.9 shows the growth of the data. In the left panel the growth of pre-NGS and NGS data is shown as the number of base pairs that you can sequence for one dollar (e.g., about 10 million base pairs for 1 dollar in 2010). The red line indicates the costs for storage (hard disk) as megabytes that you can store per dollar. The interpretation of this figure is that the costs for sequencing drop faster than the cost for data storage and that, therefore, new approaches such as data compression or the use of cloud need to be used to deal with this problem. Individual research labs often lack the storage, computing power and technical know-how to cope with the current deluge of genomic data. Also data-analysis is complicated by these large datasets.

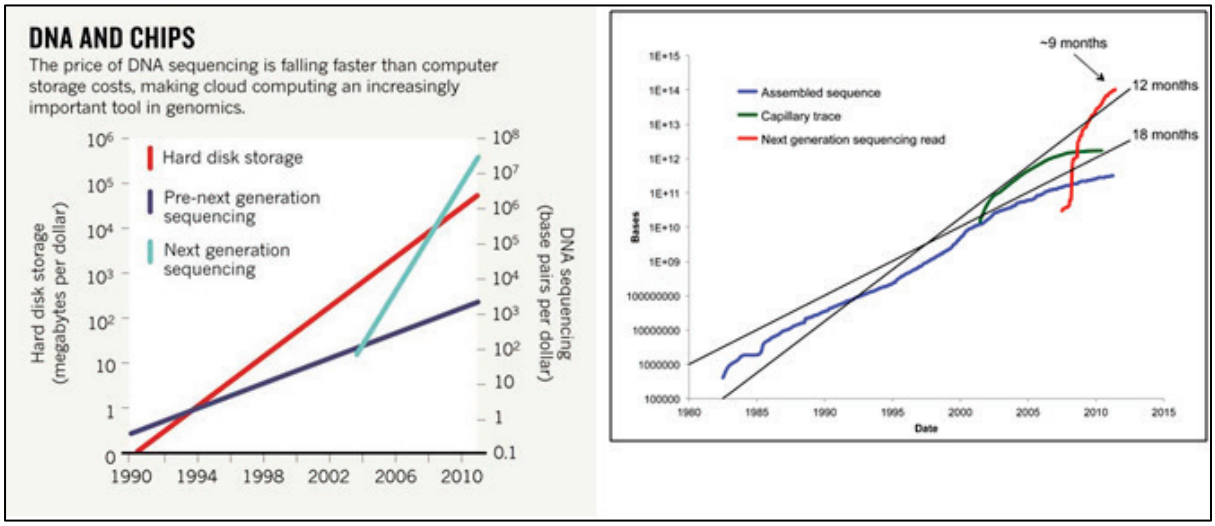


Figure 13.9. Growth of Next Generation Sequencing data. (Left figure copied from Callaway (2011) Nature, 475, 435-437). In the right panel the growth for different categories of sequence data is shown. The blue line shows the amount of assembled sequence data, which doubling time is less than 18 months. The doubling time of capillary Sanger sequencing is between 12 and 18 months. We see a decrease in doubling time since researchers are switching to NGS. The doubling time of NGS data is already 9 months and is expected to further decrease. The relevance of this is shown in the left panel (see the text for an explanation).

13.8.4 Data and database quality

The biological databases provide a basis for further analysis and interpretation and they provide links to other databases with relevant information. Therefore, database information

should be as accurate as possible to avoid erroneous results and conclusions. Several aspects relate to data quality:

- **Experimental variation and noise**. The experimental data itself should be of high-quality. For example, in a microarray experiment (technical) variation, noise, or other experimental artifacts might be introduced throughout the process of sample preparation, processing of the microarray, and scanning of the resulting image. Even the subsequent, pre-processing steps may introduce artifacts. Similarly, nucleotide sequences may contain sequencing error. Thus, some basic checks should be done before using experimental data from a public database;

- **Database errors**. In addition to low quality experimental data, virtually all databases will contain errors. Such errors may occur in the data itself, in the annotation of the data, or in the links between the databases;

- **High quality annotation.** The annotation (see section 13.7.3.1) of each database record should be complete and of high quality as this increases the utility of the data and confidence in the data. For example, annotation in UniProt is curated by human experts, while in GenBank there is no curation of the data. Minimum Information Standards (see section 13.7.5) will guide researchers to submit complete data annotation;

- **Non-redundancy**. We might prefer a database contains no redundant information. For example, UniProt only contains a single sequence for each protein. In contrast, in GenBank you may find, for example, many sequences related to the human HBA1 gene. Although non-redundant databases are not necessarily of lower quality. In practice is may be convenient to work with a curated non-redundant database because this saves you from the tasks of inspecting multiple entries and resolving conflicts, missing information, etc;

- **Consistency**. The data within a database should be consistent and consistent with other databases; Below is one example of inconsistencies between pathway databases. Inconsistencies may be caused by database errors but may also reflect contradictory views of the underlying biology;

An example: inconsistencies in pathway databases: The TCA cycle

Metabolic pathways in public databases are described through their chemical reactions, and the metabolites, enzyme(s) and gene(s) that are involved (see inset Figure 13.10). Figure 13.10 shows the citric acid cycle as defined in five different pathway databases (Reactome, KEGG, Recon1, EHMN and HumanCyc. The figure includes metabolites, enzymes and genes. The colors indicate in how many database an enzyme (and corresponding metabolites) and genes occur. Red denotes that the enzyme or gene is present in only one database, while green indicates that it is present in all five databases. We see more red boxes than green

boxes, which indicate that is no full agreement between the five databases. That is, they are inconsistent. There are several reasons for these inconsistencies, but this is beyond the scope of these lecture notes.

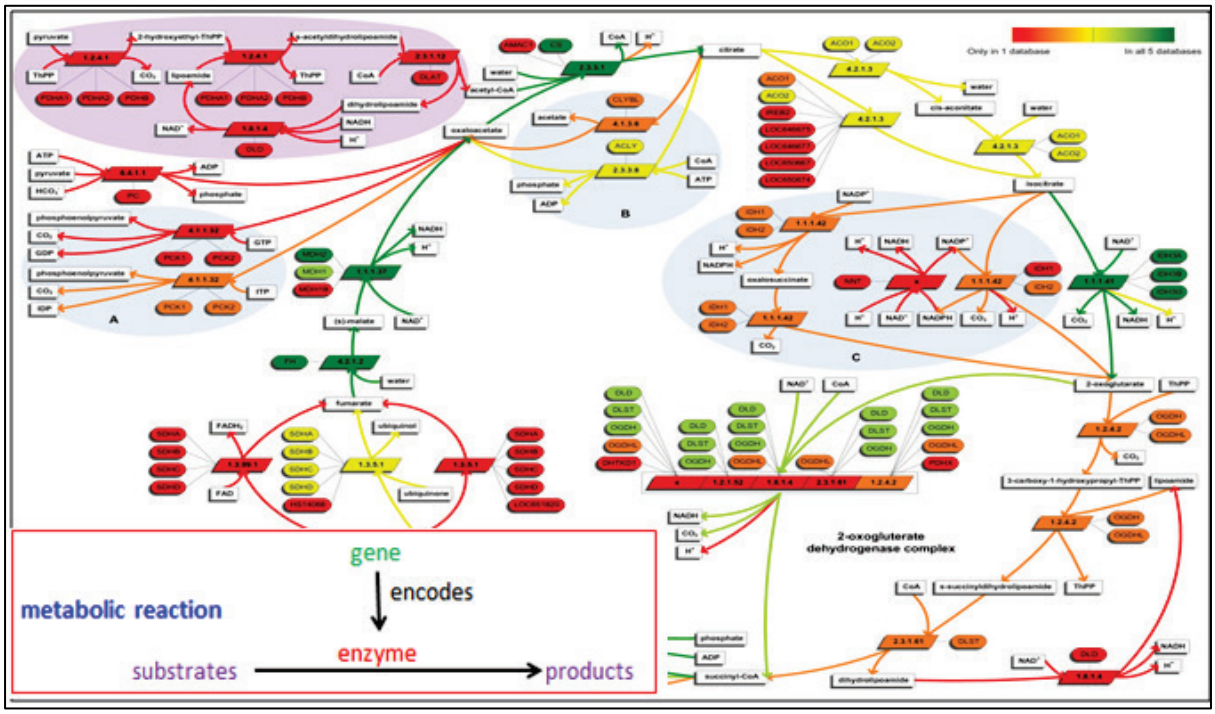


Figure 13.10. The citric acid cycle. Metabolites are shown as white boxes. Enzymes by the parallelograms, and the genes by the rounded rectangles.

Errors in public databases.

Errors occur in virtually every biological database. This was already noted in an editorial in Bioinformatics in 1998 by Peter Karp (Karp (1998) Bioinformatics, 14(9), 753) where he stated that, for the sequence databases. *"…….We have no reliable data regarding either the current rate of errors (incorrect functional annotations) within the public databases, nor on the rate of change of that error rate (we do not even know if it is increasing or decreasing each year)"*. 15 years later, this is still the case although there is some evidence that the errors in GenBank annotation seem to be increasing. Errors also occur in other databases. For example, the differences between the pathway databases w.r.t. the citric acid cycle (Figure 13.10) are partially caused by errors (see powerpoint presentation).

One of the problems with database errors is that they can propagate to other databases (or scientific literature). This might be one of the reasons that the number of errors increases. Figure 13.11 shows one possible mechanism through which database errors might be introduced and propagate. Suppose you have a protein from chicken (protein identifier XP_418136.2) with amino oxidase activity (the protein has an amino oxidase functional domain). Suppose that you find out if this protein has any other functional domains. For this,

you may match this protein sequence against the pfam database which contains proteins sequences for which the domains are annotated. For example, Figure 13.11 shows the LPP3 protein in which three domains (green and pink blocks) are shown. One of these domains is the PAP2 domain, which is a lipid phosphate phosphohydrolase domain. The other two domains are transmembrane regions, which are also present in the amino oxidase protein. Thus both proteins are membrane proteins and the TM regions are the parts that are inside the membrane. These TM regions are not related to the primary function of the proteins. However, if one compares the amino oxidase protein against the LPP3 protein (e.g., by using BLAST) then a match will be found and reported to the researcher. Subsequently, many have made the mistake to assume that, because of this match, the proteins are functionally related in which case one may update the description of the amino oxidase protein by stating that it also has lipid phosphate phosphohydrolase activity. This introduces an error in the protein sequence database, which could easily have been avoided by inspecting the nature of the sequence match more carefully. Once such error is in the protein database it may propagate to other sequences, other databases, or literature if this sequence is returned as a match in future database queries by BLAST and sequence annotation is transferred to the next protein sequence.

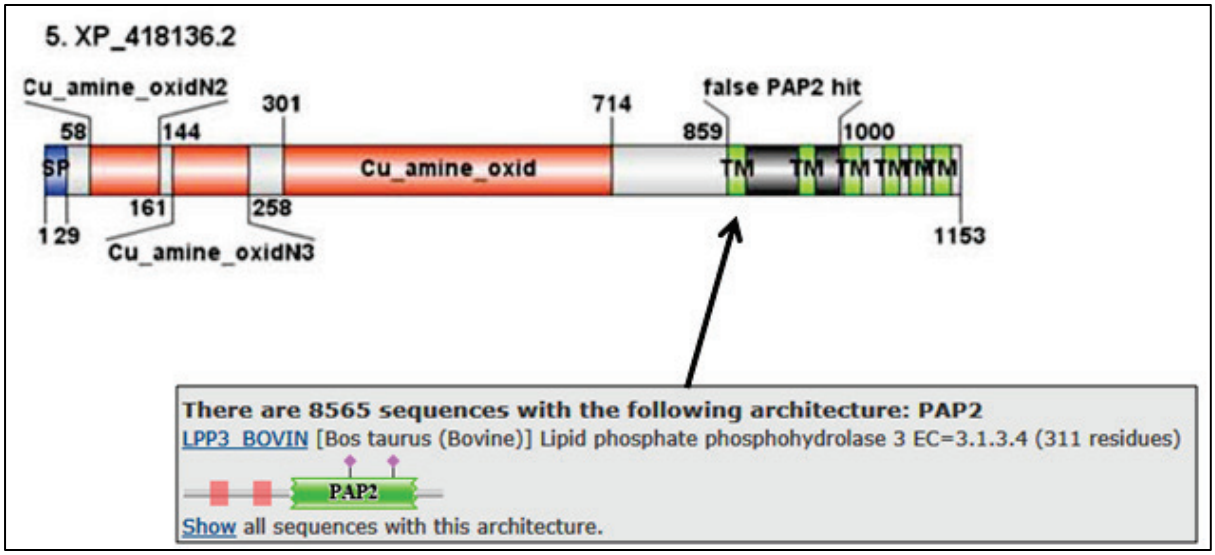


Figure 13.11. Introduction and propagation of database errors.

13.9 References

1. Benson, D.A. *et al.* GenBank. *Nucleic Acids Res* **42**, D32-7 (2014).
2. Barrett, T. & Edgar, R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol* **411**, 352-69 (2006).
3. Vizcaino, J.A. *et al.* The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* **41**, D1063-9 (2013).
4. Haug, K. *et al.* MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res* **41**, D781-6 (2013).
5. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res* **42**, D472-7 (2014).
6. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42**, D199-205 (2014).
7. UniProt, C. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **42**, D191-8 (2014).
8. Buchel, F. *et al.* Path2Models: large-scale generation of computational models from biochemical pathway maps. *BMC Syst Biol* **7**, 116 (2013).
9. Wrzodek, C., Buchel, F., Ruff, M., Drager, A. & Zell, A. Precise generation of systems biology models from KEGG pathways. *BMC Syst Biol* **7**, 15 (2013).
10. Chelliah, V., Laibe, C. & Le Novere, N. BioModels Database: a repository of mathematical models of biological processes. *Methods Mol Biol* **1021**, 189-99 (2013).
11. Thiele, I. *et al.* A community-driven global reconstruction of human metabolism. *Nat Biotechnol* **31**, 419-25 (2013).
12. Rappaport, N. *et al.* MalaCards: an integrated compendium for diseases and their annotation. *Database (Oxford)* **2013**, bat018 (2013).
13. Safran, M. *et al.* GeneCards Version 3: the human gene integrator. *Database (Oxford)* **2010**, baq020 (2010).
14. Stelzer, G. *et al.* In-silico human genomics with GeneCards. *Hum Genomics* **5**, 709-17 (2011).
15. Amberger, J., Bocchini, C. & Hamosh, A. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). *Hum Mutat* **32**, 564-7 (2011).
16. Roque, F.S. *et al.* Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput Biol* **7**, e1002141 (2011).
17. van 't Veer, L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-6 (2002).
18. Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315-22 (2014).
19. Eichler, H.G., Abadie, E., Breckenridge, A., Leufkens, H. & Rasi, G. Open clinical trial data for all? A view from regulators. *PLoS Med* **9**, e1001202 (2012).
20. Rodwin, M.A. & Abramson, J.D. Clinical Trial Data as a Public Good. *Jama-Journal of the American Medical Association* **308**, 871-872 (2012).
21. Wolkenhauer, O., Auffray, C., Jaster, R., Steinhoff, G. & Dammann, O. The road from systems biology to systems medicine. *Pediatr Res* **73**, 502-7 (2013).
22. Tryka, K.A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* **42**, D975-9 (2014).

14 Sequence alignment and BLAST

Lecturer: Prof. dr. Antoine van Kampen (AMC)

After reading this chapter you should understand the concepts of

- Sequence alignment, its evolutionary basis, homology, similarity and identity
  BLAST, variants of BLAST, E-value

This chapter will not be part of the lectures nor the examination. It is provide as background information and to refresh your minds. It is however important to understand the concept of a sequence alignment (e.g., in the context of exome sequencing).

Contents

14.1 Sequence alignment and BLAST

In this section, we will discuss an important topic in bioinformatics: sequence alignment. A good understanding of this topic is necessary since it will be using during the 'exome sequencing' lectures.

A sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Sequence alignment has many applications such as

- Genome comparisons (= comparison of full DNA sequence between organisms)
- Identification of mutations and SNPs in DNA
- Gene and promotor prediction in DNA
- Construction of evolutionary trees (phylogeny)
- Identification of conserved regions (motifs) in DNA or protein sequences
- Database searching (for example with BLAST)

The comparison of two sequences is called pair-wise sequence alignment. The comparison of three or more sequences is called multiple sequence alignment. In this section we will only discuss pair-wise sequence alignment.

An example of a pair-wise DNA sequence alignment of the promotor region of the human and mouse melanocortin 4 receptor (MC4R) gene is shown in Figure 14.1. Heterozygous mutations in the coding region of MC4R are the cause of 1–6% of severe early-onset obesity cases. MC4R is a trans-membrane G-protein–coupled receptor (GPCR) encoded by a single exon gene localized on chromosome 18q22. This alignment shows the differences between the two DNA sequences but also the conserved cis-acting elements.

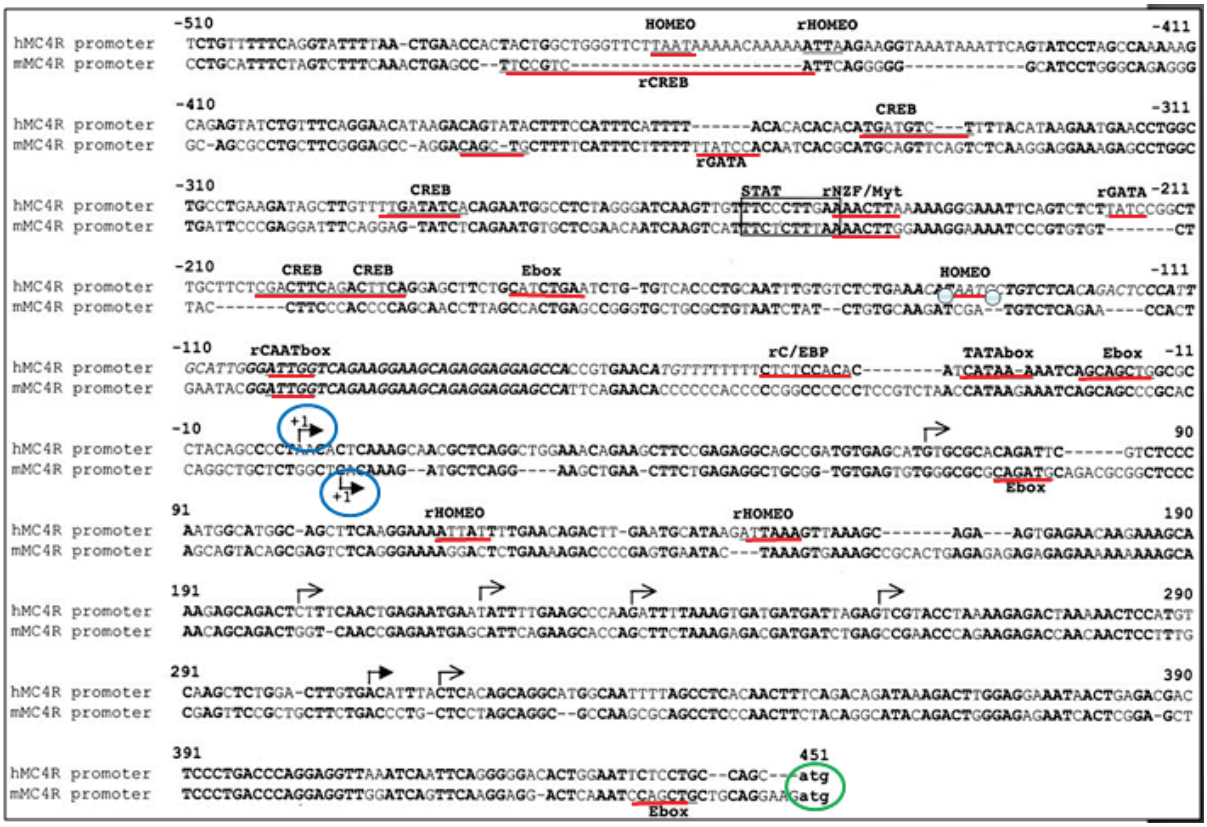


Figure 14.1. DNAsequence alignment of human and mouse MC4R promoter regions. The major transcription start sites are indicated by a "plain" arrow and defined as +1 (blue circle). The translation initiation codon ATG is indicated in lower case (green circle). Conserved regions between human and mouse are indicated in bold. Potential cis-acting elements were identified using the TRANSFAC database and manual searches and are indicated as underlined (red line). (Figure copied from Lubrano-Berthelier (2003) Diabetes, 52, 2996).

14.1.1 Evolutionary basis of sequence alignment

DNA and proteins are products of evolution. The building blocks of these biological macromolecules, nucleotide bases, and amino acids form linear sequences that determine the primary structure of the molecules. These molecules can be considered molecular fossils that encode the history of millions of years of evolution. During this time period, the molecular sequences undergo random changes, some of which are selected during the process of evolution. As the selected sequences gradually accumulate mutations and diverge over time, traces of evolution may still remain in certain portions of the sequences to allow identification of the common ancestry. The presence of evolutionary traces is because some of the residues that perform key functional and structural roles tend to be preserved by natural selection; other residues that may be less crucial for structure and function tend to mutate more frequently. For example, active site residues of an enzyme family tend to be conserved because they are responsible for catalytic functions. Therefore, by comparing sequences through alignment, patterns of conservation and variation can be identified. The degree of sequence conservation in the alignment reveals evolutionary relatedness of

different sequences, whereas the variation between sequences reflects the changes that have occurred during evolution in the form of substitutions, insertions, and deletions.

Identifying the evolutionary relationships between sequences helps to characterize the function of unknown sequences. When a sequence alignment reveals significant similarity among a group of sequences, they can be considered as belonging to the same protein family. If one member within the family has a known structure and function, then that information can be transferred to those that have not yet been experimentally characterized. Therefore, sequence alignment can be used as basis for prediction of structure and function of uncharacterized sequences.

Sequence alignment provides inference for the relatedness of two sequences under study. If the two sequences share significant similarity, it is extremely unlikely that the extensive similarity between the two sequences has been acquired randomly, meaning that the two sequences must have derived from a common evolutionary origin.

14.1.2 Sequence homology versus sequence similarity

An important concept in sequence analysis is sequence homology. When two sequences are descended from a common evolutionary origin, they are said to have a homologous relationship or share homology. A related but different term is sequence similarity, which is the percentage of aligned amino acids or nucleotides.

It is important to distinguish sequence homology from the related term sequence similarity because the two terms are often confused by some researchers who use them interchangeably in scientific literature. To be clear, sequence homology is an inference or a conclusion about a common ancestral relationship drawn from sequence similarity comparison when the two sequences share a high enough degree of similarity. On the other hand, similarity is a direct result of observation from the sequence alignment. Sequence similarity can be quantified using percentages; homology is a qualitative statement. For example, one may say that two sequences share 40% similarity. It is incorrect to say that the two sequences share 40% homology. They are either homologous or non-homologous. Generally, if the sequence similarity level is high enough, a common evolutionary relationship can be inferred.

14.1.3 Sequence similarity versus sequence identity

Another set of related terms for sequence comparison are sequence similarity and sequence identity. Sequence similarity and sequence identity are synonymous for nucleotide sequences. For protein sequences, however, the two concepts are very different. In a protein sequence alignment, sequence identity refers to the percentage of matches of the same

amino acid residues between two aligned sequences. Similarity refers to the percentage of aligned residues that have similar physicochemical characteristics and can be more readily substituted for each other.

14.2 Database similarity searching (BLAST)

A main application of pairwise alignment is retrieving biological sequences in databases (for example GeneBank or Uniprot) based on similarity. This process involves submission of a query sequence and performing a pairwise comparison of the query sequence with all individual sequences in a database. Thus, database similarity searching is pairwise alignment on a large scale. Figure 14.2 shows the principle of BLAST.

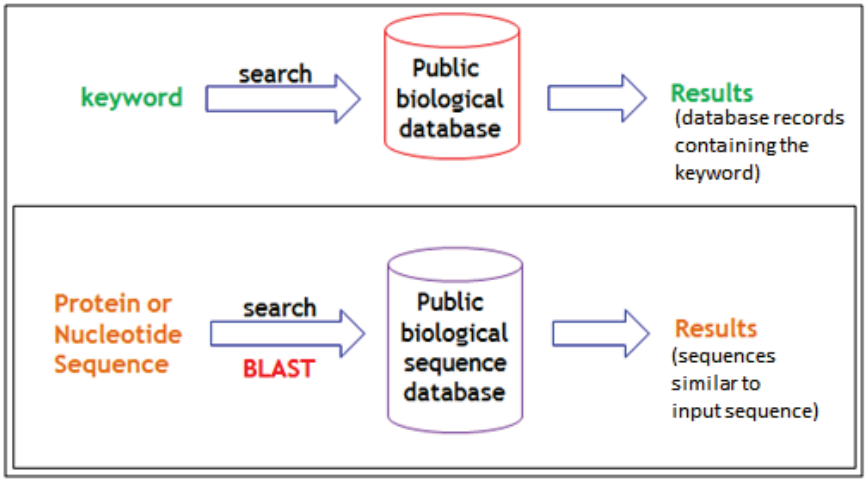


Figure 14.2. The difference between database searching with keywords (top panel) and with sequences (bottom panel).

14.2.1 Unique requirements of database searching

There are unique requirements for implementing algorithms for sequence database searching. The first criterion is sensitivity, which refers to the ability to find as many correct hits as possible. It is measured by the extent of inclusion of correctly identified sequence members of the same family. These correct hits are considered “true positives” in the database searching exercise. The second criterion is selectivity, also called specificity, which refers to the ability to exclude incorrect hits. These incorrect hits are unrelated sequences mistakenly identified in database searching and are considered “false positives.” The third criterion is speed, which is the time it takes to get results from database searches. Depending on the size of the database, speed sometimes can be a primary concern.

Ideally, one wants to have the greatest sensitivity, selectivity, and speed in database searches. However, satisfying all three requirements is difficult in reality. What generally happens is that an increase in sensitivity is associated with decrease in selectivity. A very inclusive search tends to include many false positives. Similarly, an improvement in speed often comes at the

cost of lowered sensitivity and selectivity. A compromise between the three criteria is made by the BLAST program.

One of these heuristic methods for performing database searches is BLAST (Basic Local Alignment Search Tool; www.ncbi.nlm.nih.gov/BLAST) [1,2]. This method is not guaranteed to find the optimal alignment or true homologs. However, it is fast at a moderate expense of sensitivity and specificity of the search.

14.2.2 Variants of BLAST

BLAST is a family of programs that includes BLASTN, BLASTP, BLASTX TBLASTN, and TBLASTX (Figure 14.3). BLASTN queries nucleotide sequences with a nucleotide sequence database. BLASTP uses protein sequences as queries to search against a protein sequence database. BLASTX uses nucleotide sequences as queries and translates them in all six reading frames (3 reading frames for each DNA strand; see also http://www.ncbi.nlm.nih.gov/books/NBK26829/#A1053) to produce translated protein sequences, which are used to query a protein sequence database. TBLASTN queries protein sequences to a nucleotide sequence database with the sequences translated in all six reading frames. TBLASTX uses nucleotide sequences, which are translated in all six frames, to search against a nucleotide sequence database that has all the sequences translated in six frames. In addition, there is also a bl2seq program that performs local alignment of two user-provided input sequences.

Sequence type	nucleotide database	protein database
nucleotide query	blastn/tblastx	blastx
amino acid query	tblastn	blastp

Figure 14.3. Family of BLAST programs.

The choice of the type of sequences also influences the sensitivity of the search. Generally speaking, there is a clear advantage of using protein sequences in detecting homologs. This is because DNA sequences only comprise four nucleotides, whereas protein sequences contain twenty amino acids. Amino acid substitution matrices (which are used in BLAST and other alignment methods) incorporate subtle differences in physicochemical properties between amino acids, meaning that protein sequences are far more informative and sensitive in detection of homologs. For that reason, if the input sequence is a protein-encoding DNA sequence, it is preferable to use BLASTX, which translates it in six open reading frames before sequence comparisons are carried out.

14.2.3 Statistical Significance

The BLAST output provides a list of pairwise sequence matches ranked by statistical significance. The significance scores help to distinguish evolutionarily related sequences from unrelated ones. Generally, only hits above a certain threshold are displayed. In BLAST searches, this statistical indicator is known as the E-value (expectation value), and it indicates the probability that the resulting alignments from a database search are caused by random chance. The E-value is related to the statistical p-value used to assess significance of single pairwise alignment.

BLAST output format

The BLAST output (Figure 14.4) includes a graphical overview box, a matching list and a text description of the alignment. The graphical overview box contains colored horizontal bars that allow quick identification of the number of database hits and the degrees of similarity of the hits. The color coding of the horizontal bars corresponds to the ranking of similarities of the sequence hits (red: most related; green and blue: moderately related; black: unrelated). The length of the bars represents the spans of sequence alignments relative to the query sequence. Below the graphical box is a list of matching hits ranked by the E-values in ascending order. Each hit includes the accession number, title (usually partial) of the database record, bit score, and E-value.

This list is followed by the text description, which may be divided into three sections: the header, statistics, and alignment. The header section contains the gene index number or the reference number of the database hit plus a one-line description of the database sequence. This is followed by the summary of the statistics of the search output, which includes the bit score, E-value, percentages of identity, similarity ("Positives"), and gaps. In the actual alignment section, the query sequence is on the top of the pair and the database sequence is at the bottom of the pair labeled as Subject. In between the two sequences, matching identical residues are written out at their corresponding positions, whereas non-identical but similar residues are labeled with "+".

Go to the BLAST website and try it yourself!

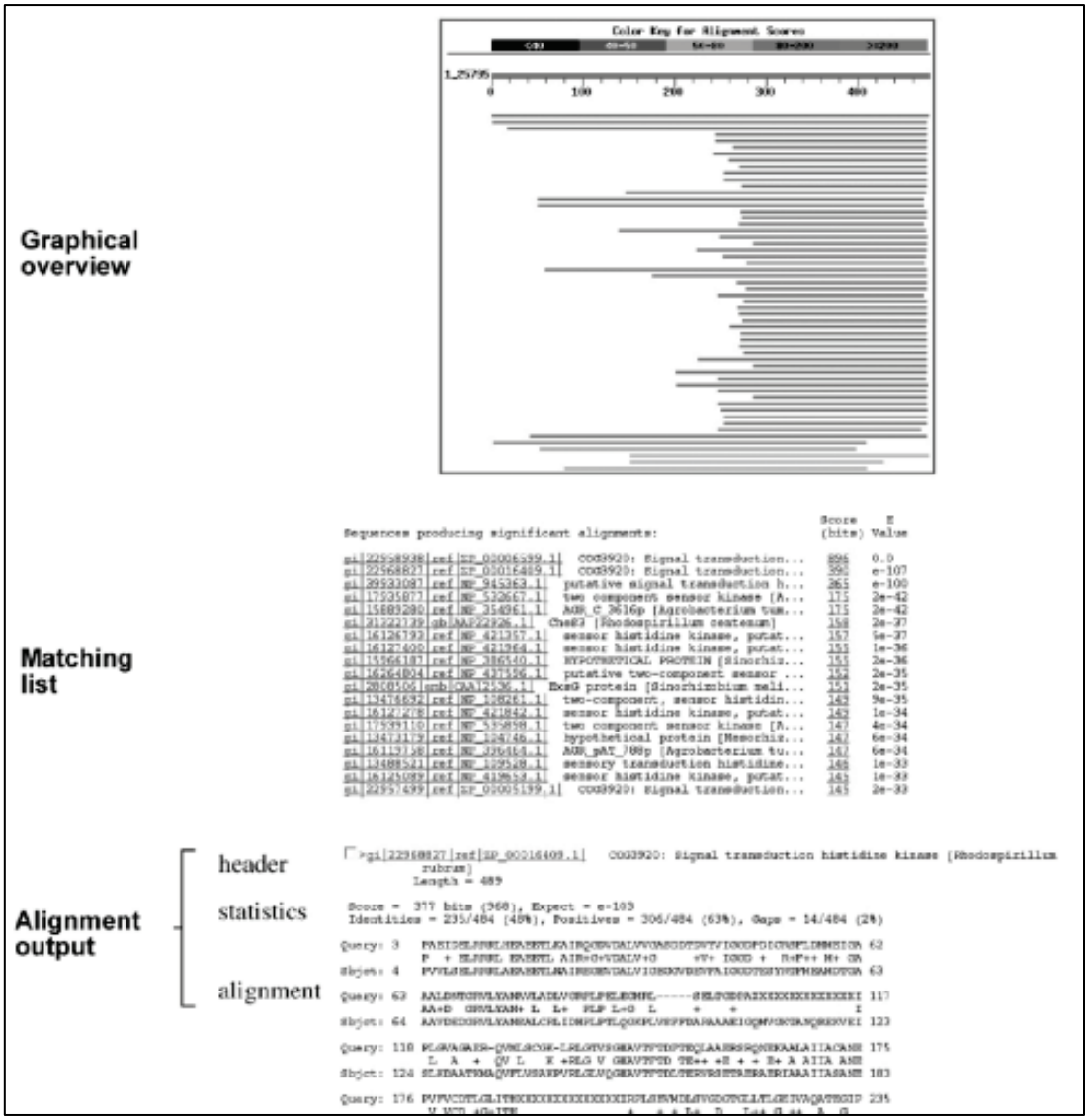


Figure 14.4. An example of a BLAST output showing three portions: the graphical overview box, the list of matching hits, and the text portion containing header, statistics, and the actual alignment. (Figure copied from Jin Xiong (2006) Essential Bioinformatics, Cambridge University Press, Cambridge).

14.3 References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of molecular biology 215: 403-410.
2. Karlin S, Altschul SF (1990) Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. Proceedings of the National Academy of Sciences of the United States of America 87: 2264-2268.

15 Brief introduction to Unix

Lecturer: Prof. dr. Antoine van Kampen (AMC)

After reading this chapter you should understand

- Operating systems
- Unix and Linux
- Basic Unix/Linux commands

This chapter is not for the examination but is provided to give you basic background about Unix, which will help you to make the exome sequencing practicum.

Contents

15.1 Introduction

Much software used in bioinformatics is developed for the Unix or Linux operating system and does not work on a Windows or Apple computer. Therefore, some basic knowledge about the Unix/Linux operating system is useful. We will also make use of Linux during the exome sequence practicum.

15.2 Computer operating systems

An operating system (OS; Figure 15.1) is a collection of software that manages computer hardware resources and provides common services for computer programs (such as Microsoft Word). The operating system is an essential component of the system software in a computer system. Typical functionality of an OS includes:

- Process management
- Memory management
- File system
- Device drivers (e.g., for your printer)
- Networking
- Security (Process/Memory protection)
- Input and output (etc, keyboard, monitor)

Operating systems can be found on almost any device that contains a computer (from cellular phones and video game consoles to supercomputers and web servers).

Examples of popular modern operating systems include

- Android
- iOS
- Linux
- Microsoft Windows
- Windows Phone

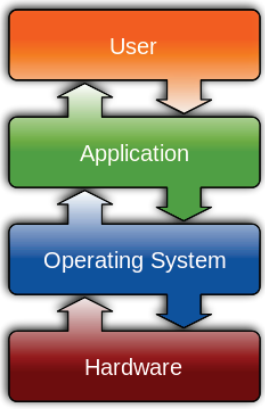


Figure 15.1. The operating system of a computer functions as the interface between a computer application (such as Microsoft Word) and the hardware (e.g., your laptop). (Figure copied from Wikipedia; http://en.wikipedia.org/wiki/Operating_system)

15.3 Unix and Linux

[The text below is mainly copied from Wikipedia: http://en.wikipedia.org/wiki/Unix and http://en.wikipedia.org/wiki/Linux]

Unix (officially trademarked as UNIX) is a multitasking, multi-user computer operating system that exists in many variants. From the power user's or programmer's perspective, Unix systems are characterized by a modular design that is sometimes called the "Unix philosophy," meaning the OS provides a set of simple tools that each perform a limited, well-defined function, with a unified filesystem as the main means of communication and a shell scripting and command language to combine the tools to perform complex workflows.

During the late 1970s and 1980s, Unix developed into a **standard operating system for academia**. AT&T tried to commercialize it by licensing the OS to third-party vendors, leading to a variety of both academic (e.g., BSD) and commercial variants of Unix (such as Xenix) and eventually to the "Unix wars" between groups of vendors.

The Open Group, an industry standards consortium, now owns the UNIX trademark and allows its use for certified operating systems compliant with its standard. Other operating systems that emulate Unix to some extent may be called Unix-like. The term Unix is also often used informally to denote any operating system that closely resembles the trademarked system. The most common version of Unix (bearing certification) is Apple's OS X, while **Linux** is the most popular non-certified workalike

Linux

Linux is a Unix-like computer operating system assembled under the model of free and open source software development and distribution. Linux was originally developed as a free operating system for Intel x86-based personal computers. It has since been ported to more computer hardware platforms than any other operating system. It is a leading operating system on servers and other big iron systems such as mainframe computers and supercomputers: as of June 2013, more than 95% of the world's 500 fastest supercomputers run some variant of Linux. Linux also runs on embedded systems (devices where the operating system is typically built into the firmware and highly tailored to the system) such as mobile phones, tablet computers, network routers, building automation controls, televisions and video game consoles; the Android system in wide use on mobile devices is built on the Linux kernel.

The development of Linux is one of the most prominent examples of free and open source software collaboration: the underlying source code may be used, modified, and distributed—commercially or non-commercially—by anyone under licenses such as the GNU General Public License. Typically, Linux is packaged in a format known as a Linux distribution for desktop and server use. Some popular mainstream Linux distributions include Debian (and its

derivatives such as Ubuntu and Linux Mint) and Fedora (and its derivatives such as the commercial Red Hat Enterprise Linux and its open equivalent CentOS).

A distribution oriented toward desktop use will typically include X11 as the windowing system, and an accompanying desktop environment such as GNOME or the KDE Software Compilation.

In the following some basic constructs of Linux commands are given. This will help you to recognize and understand them during the computer practicum.

15.4 Basic Linux (Unix) principles

Although modern Linux distributions allow you to work with software applications in a way similar to Microsoft Windows, much of the scientific (bioinformatics) software is executed from the Linux prompt. For example, the command 'cd', which is an abbreviation for 'change directory' allows you to go to another directory (which is called a 'folder' in Microsoft Windows). For example, the command

```
> cd /home/vankampen/gene
```

will take you to the directory /home/vankampen/gene such that you can inspect and/or use the files in this folder. Note that the '>' represents the Linux prompt in this example and should NOT BE TYPED. Also note that this directory may not exist on your computer system, in which case you will get an error.

Once we are in this directory, we can ask for its content by using the command 'ls' (list):

```
> ls
```

In general, a Linux command takes the form:

```
> command [-arguments]
```

where the arguments may be optional. For example, 'ls' can take many arguments. One of this is the argument 'l':

```
> ls -l
```

This will also show the content of the current directory, but provides more details about the files (this is comparable to the 'detailed' view in Microsoft windows).

If you need help with a certain Linux command, then you can consult the Linux manual pages. For example, if you want to know which arguments are available for the 'ls' command then you can simply type:

```
> man ls
```

15.5 Redirection

A commonly occurring situation is that you want to redirect the output of an application or Linux command to a file instead of having the output printed to your screen. This is done by the '>' or '>>' sign. The difference between these two is that '>' will overwrite the file, while '>>' will append the output to the content that is already in the file.

Let us suppose that you want to store the output of 'ls –a' in a file. This is very straightforward:

```
> ls -l > output.txt
```

Here output.txt is the name of the file. This file can now be printed or further manipulated. For example, we can use the Unix command 'grep' to show all lines that contain have the word 'sequence' in their file name:

```
> grep sequence output.txt
```

15.6 Piping

Another powerful mechanism in Unix is 'piping'. This allows the output from one application to serve as the input of a next application. This is done by using the character '|'. Continuing the previous example we could also have done the following:

```
> ls -l | grep sequence
```

15.7 Commonly used Linux commands

man – Display manual page. Most Unix systems have a very comprehensive set of documents. To know how to use a command, use man command name. To search for a command based on a keyword, use man -k keyword.

ls – LiSt the directory content. Filenames that start with a dot are normally hidden for this command. To see all files including "hidden" files, the option -a must be used. The option -l causes a "long" listing including file ownership, permission etc.

cp – CoPy files. This command requires at least two arguments. The last argument is the destination and all other arguments are source files. If the destination is a directory, the source file(s) are copied with the same name into the destination directory. If only two

arguments are supplied and the destination is a file or doesn't exist yet, the source file is copied as the file with the destination's name.

mv – MoVe files. This is analogous to the cp command, but files are either moved to a new directory or renamed.

rm – ReMove files. This is used to remove files (not directories). A Unix OS doesn't make a habit of asking if you are sure you want to do something, so use this command with care! If the option -r is used with a directory as an argument, it will recursively remove the complete directory tree. Very dangerous!

pwd – Show the name of the Working Directory.

cd – Change Directory. Change the current directory to the argument or to the users home directory if no argument is given.

mkdir – MaKe DIRectory.

rmdir – ReMove DIRectory. This only works if the directory is empty. To remove a directory including all its content, use rm -r.

cat – CATenate file(s). Without argument, this will copy stdin to stdout. When filenames are passed as arguments the contents of all these files are copied successively to stdout.

head – Print the first few lines of either stdin or the file(s) in the arguments to stdout. With the option -n the number of lines to print can be specified.
tail – Analogous to head, but print the last few lines.

wc – Word Count. Count the number of characters, lines and/or words from stdin or files in the arguments.
sort – sort a file or stdin based on its lines.

uniq – remove duplicated lines. With the proper options this can also print only unique or non-unique lines. uniq expects duplicated lines the input to be consecutive.

tar – tape archive. Create an (uncompressed) archive of a set of files to stdout, a file or a (tape) device.

compress/uncompress – (Un)compress a (single) file or stdin. Because of patents on the LZW algorithm that compress uses, the GNU project developed a patent-free compressor named gzip. This gained much popularity and is a “de facto” standard today.

grep – Search for lines with specified substrings

more – Print the input to the screen, one page at a time

Definitions, links and further information

1000 Genomes Project

http://www.l000genomes.org

Acyclic graph

In mathematics and computer science, a directed acyclic graph is a directed graph with no directed cycles. That is, it is formed by a collection of vertices and directed edges, each edge connecting one vertex to another, such that there is no way to start at some vertex v and follow a sequence of edges that eventually loops back to v again (Source Wikipedia).

Allele

One of two or more forms of a gene or a genetic locus

allogeneic haematopoietic progenitor cell transplant

Allogeneic: not from a genetically identical donor of the same species. Haematopoietic; pertaining to the formation of blood or blood cells. Progenitor cell transplant: The transplantation of parent cells which may give rise to progeny (daughter cells) which serve more specialized functions. Transplants may be from the peripheral blood or bone marrow.

Annotation

Annotation is the association of (high throughput) data, such as protein sequences, with biological information from literature but also other databases. An annotation is a note, summary, or commentary on some section of data that is intended to explain or illustrate its meaning.

Biotynilation

In biochemistry, biotinylation is the process of covalently attaching biotin to a protein, nucleic acid or other molecule. Biotinylation is rapid, specific and is unlikely to perturb the natural function of the molecule due to the small size of biotin. Biotin binds to streptavidin with an extremely high affinity, fast on-rate, and high specificity, and these interactions are exploited in many areas of biotechnology to isolate biotinylated molecules of interest. (source Wikipedia).

Candidate gene re-sequencing

The sequencing of part of an individual's genome in order to detect sequence differences between the individual and the standard genome of the species

Complex traits

Disease that are influenced by more than one factor. The factors can be genetic or environmental. This is in contrast to simple genetic traits, whose variations are controlled by variations in single genes

Consensus coding sequence (CCDS) project:

The Consensus CDS (CCDS) project is a collaborative effort to identify a core set of human and mouse protein coding regions that are consistently annotated and of high quality. http://www.ncbi.nlm.nih.gov/CCDS/

Copy-number variants.

A form of structural variation. These are alterations of the DNA of a genome that results in the cell having an abnormal number of copies of one or more sections of the DNA. CNVs correspond to relatively large regions of the

genome that have been deleted (fewer than the normal number) or duplicated (more than the normal number) on certain chromosomes. (source: Wikipedia).

Chromatin immunoprecipitation.

Is a type of immunoprecipitation experimental technique used to investigate the interaction between proteins and DNA in the cell. It aims to determine whether specific proteins are associated with specific genomic regions, such as transcription factors on promoters or other DNA binding sites,

Curation

Involves information collection, validation, quality control, use of standards, and communication to provide annotation for public biological databases.

dbSNP

http://www.ncbi.nlm.nih.gov/snp

EBI

European Bioinformatics Institute (http://www.ebi.ac.uk/)

Endomembrane

The endomembrane system is composed of the different membranes that are suspended in the cytoplasm within a eukaryotic cell. These membranes divide the cell into functional and structural compartments, or organelles. In eukaryotes the organelles of the endomembrane system include: the nuclear envelope, the endoplasmic reticulum, the Golgi apparatus, lysosomes, vacuoles, vesicles, endosomes and the cell membrane. The system is defined more accurately as the set of membranes that form a single functional and developmental unit, either being connected directly, or exchanging material through vesicle transport

Epigenetics

Changes in the expression of a gene or set of genes that occur without changing the DNA sequence (source: Sadava et al (2013) Life, the science of biology. W.H. Freeman and company, Ninth Edition)

Exome

The subset of a genome that is protein coding. In addition to the exome, commercially available capture probes target non-coding exons, sequences flanking exons and microRNAs.

Gaussian Kernel.

Can be thought of as a histogram smoothed by a Gaussian function.

G protein coupled receptors (GPCRs)

A large protein family of receptors that sense molecules outside the cell and activate inside signal transduction pathways and, ultimately, cellular responses. (source: Wikipedia)

GWAS

Genome wide association study. A GWAS is an examination of many common genetic variants in different individuals to see if any variant is associated with a trait. GWAS typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major diseases. (source:wikipedia)

Haplotype

A combination of alleles on a single chromosome.

Heritability

The proportion of the total phenotypic variation in a given characteristic that can be attributed to additive genetic effects.

Human Variome Project

http://www.humanvariomeproject.org

Inflection points

An inflection point is a point on a curve at which the curvature changes sign from plus to minus or from minus to plus. The curve changes from positive curvature to negative curvature, or vice versa. If one imagines driving a vehicle along a winding road, inflection is the point at which the steering-wheel is momentarily "straight" when being turned from left to right or vice versa. (source: Wikipedia)

Innate immune system

The innate immune system, also known as non-specific immune system and first line of defense,[1] comprises the cells and mechanisms that defend the host from infection by other organisms in a non-specific manner. This means that the cells of the innate system recognize and respond to pathogens in a generic way, but unlike the adaptive immune system, it does not confer long-lasting or protective immunity to the host.[2] Innate immune systems provide immediate defense against infection, and are found in all classes of plant and animal life.

Inversions

An inversion is a chromosome rearrangement in which a segment of a chromosome is reversed end to end. An inversion occurs when a single chromosome undergoes breakage and rearrangement within itself. (source: Wikipedia).

Isotopes

Isotopes are variants of a particular chemical element: while all isotopes of a given element share the same number of protons, each isotope differs from the others in its number of neutrons (soure: Wikipedia)

Linkage mapping

Genetic linkage is the tendency of genes that are located proximal to each other on a chromosome to be inherited together during meiosis. Genes whose loci are nearer to each other are less likely to be separated onto different chromatids during chromosomal crossover, and are therefore said to be genetically linked. A linkage map is a genetic map of a species that shows the position of its known genes or genetic markers relative to each other in terms of recombination frequency, rather than a specific physical distance along each chromosome. Linkage mapping is critical for identifying the location of genes that cause genetic diseases.

Locus heterogeneity

The appearance of phenotypically similar characteristics resulting from mutations at different genetic loci. Differences in effect size or in replication between studies and samples are often ascribed to different loci leading to the same disease.

Major histocompatibility complex (MHC)

MHC a cell surface molecule encoded by a large gene family in all vertebrates. MHC molecules mediate interactions of leukocytes, also called white blood cells, which are immune cells, with other leukocytes or body cells. MHC determines compatibility of donors for organ transplant as well as one's

susceptibility to an autoimmune disease. In humans, MHC is also called human leukocyte antigen (HLA). (source: Wikipedia)

Mendelian disorders

Phenotypes caused by a mutation (or mutations) in a single gene and inherited in a dominant, recessive or X-linked pattern.

Michaelis–Menten equation

In biochemistry, Michaelis–Menten kinetics is one of the simplest and best-known models of enzyme kinetics. The model takes the form of an equation describing the rate of enzymatic reactions (v), by relating reaction rate to the concentration of a substrate S (v=Vmax*[S]/(Km+[S])). (source: Wikipedia).

Minor allele frequency (MAF)

The frequency at which the less common allele occurs in a given population

NCBI

National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov)

Next-generation DNA sequencing

Highly parallelized DNA-sequencing technologies that produce many hundreds of thousands or millions of short reads (25-500 bp) for a low cost and in a short time.

Nonlinear

In mathematics, a nonlinear system is one that does not satisfy the superposition principle, or one whose output is not directly proportional to its input.

OMIM

http://www.ncbi.nlm.nih.gov/omim

Open-source

Generally, open source refers to a program in which the source code is available to the general public for use and/or modification from its original design. Open source code is typically created as a collaborative effort in which programmers improve upon the code and share the changes within the community. Open source sprouted in the technological community as a response to proprietary software owned by corporations

Ordinary Differential Equation (ODE)

An ordinary differential equation is an equality involving a function and its derivatives. (source: http://mathworld.wolfram.com)

Penetrance

The proportion of individuals with a specific phenotype among carriers of a particular genotype.

Polyphen:

Polymorphism Phenotyping, http://genetics. bwh.harvard.edu/pph2

Positional cloning studies

Positional cloning is a method of gene identification in which a gene for a specific phenotype is identified only by its approximate chromosomal location; this is known as the candidate region. Initially, the candidate region can be defined using techniques such as linkage analysis (see linkage mapping), and

positional cloning is then used to narrow the candidate region until the gene and its mutations are found. Positional cloning typically involves the isolation of partially overlapping DNA segments from genomic libraries to progress along the chromosome toward a specific gene. During the course of positional cloning, one needs to determine whether the DNA segment currently under consideration is part of the gene.

Processed pseudogenes

Copies of the coding sequences of genes that lack promoters and introns, contain poly(A) tails.

RefSeq:

http://www.ncbi.nlm.nih.gov/RefSeq

Sample indexing

Sequencing more than one sample in a single sequencing lane.

Scale free (interaction) network

A scale-free network is a network whose degree distribution follows a power law. Such network often comprises highly connected hubs of, for example, genes or proteins.

SIB

Swiss Institute for Bioinformatics (http://www.isb-sib.ch)

SIFT

http://sift.jcvi.org

Single nucleotide variants (SNV) / single nucleotide polymorphisms (SNPs)

The most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called a nucleotide. For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA. SNPs occur normally throughout a person's DNA. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome. Most commonly, these variations are found in the DNA between genes. They can act as biological markers, helping scientists locate genes that are associated with disease. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene's function.

Translocations

A chromosome translocation is a chromosome abnormality caused by rearrangement of parts between nonhomologous chromosomes. (source: Wikipedia).