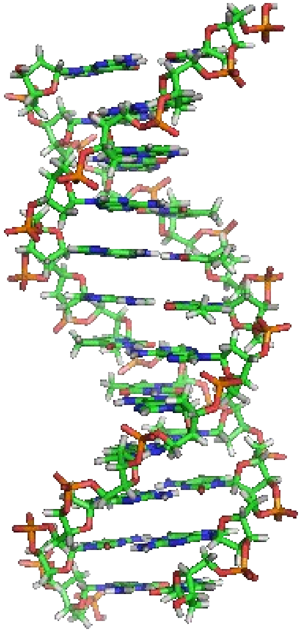
Basic Local Alignment and Search Tool (BLAST)



Database searching

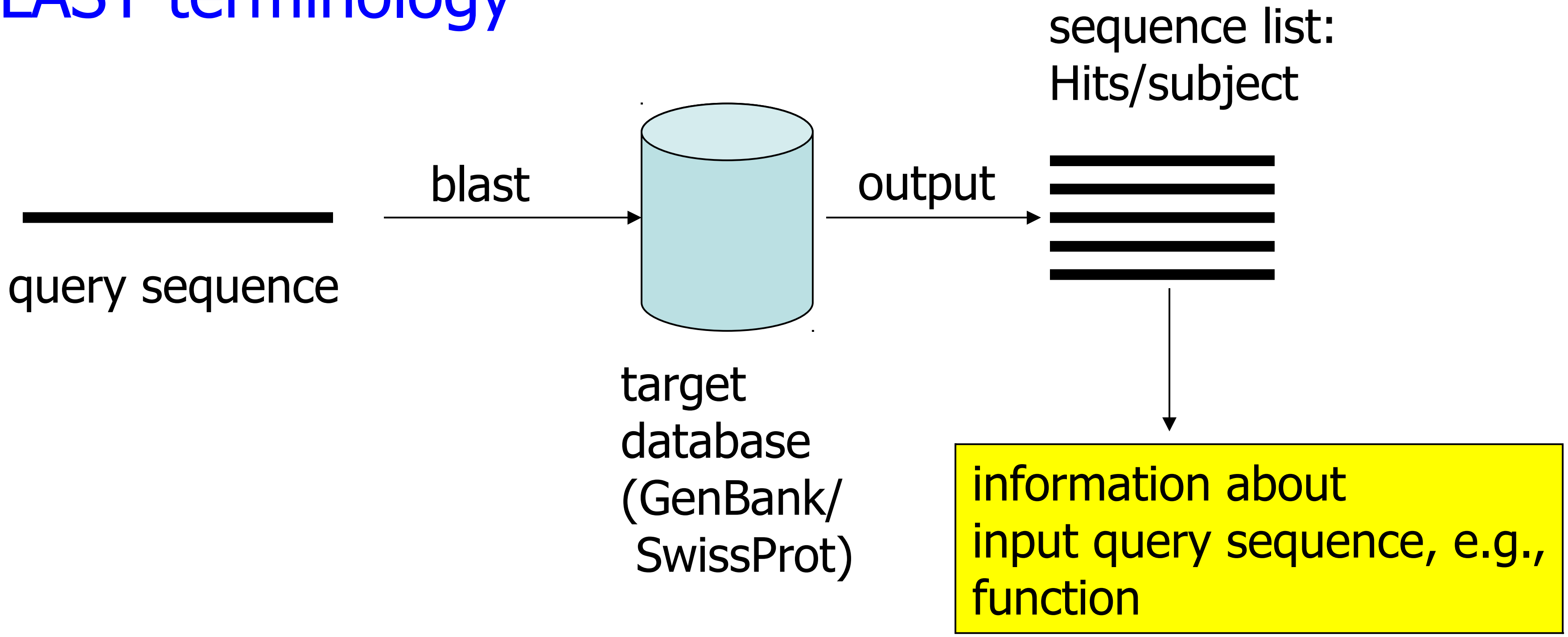
Barbera van Schaik

Why use BLAST?

- Dynamic Programming is not suitable for comparing a query sequence against a database
  - Takes too much time!

- BLAST is a heuristic method to find the highest locally optimal alignments
  - BLAST improved overall speed of searches
  - BLAST maintains good sensitivity

BLAST terminology



sequence list:
Hits/subject

query sequence

blast

target
database
(GenBank/
 SwissProt)

output

information about
input query sequence, e.g.,
function

The aim of a database (blast) search is to discover sequence homology on basis of sequence similarity

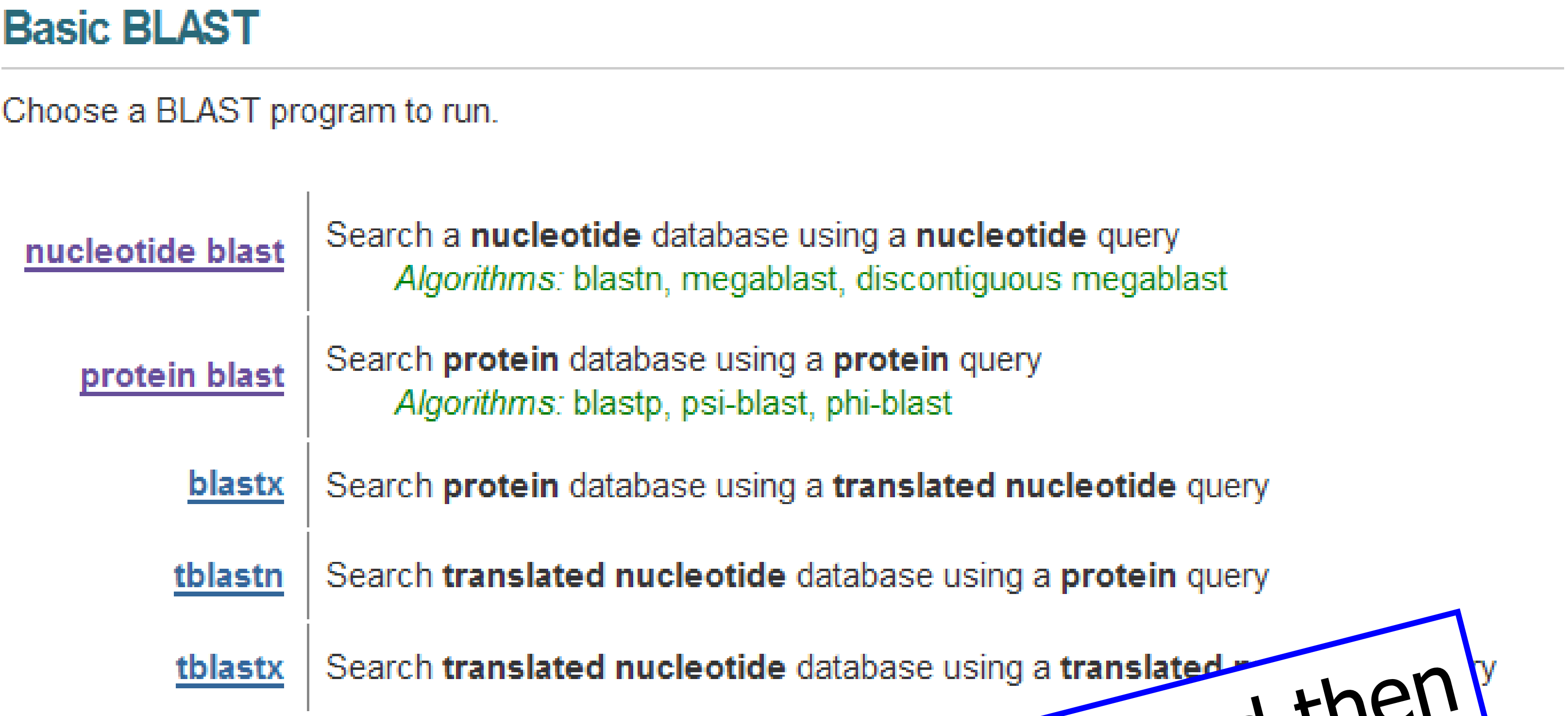
BLAST returns similar sequences, not necessarily biological similar sequences

BLAST variants

Sequence type	**nucleotide database**	**protein database**
nucleotide query	blastn/tblastx	blastx
amino acid query	tblastn	blastp

blastn: finds NT sequences similar to your NT sequence
blastp: finds AA sequences similar to your AA sequence
blastx: finds AA sequences similar to translation of your NT sequence (if you cannot recognize an ORF)
tblastn: translate AA sequence and searches against NT database (for finding pseudogenes)
tblastx: keep computers busy

http://blast.ncbi.nlm.nih.gov/Blast.cgi



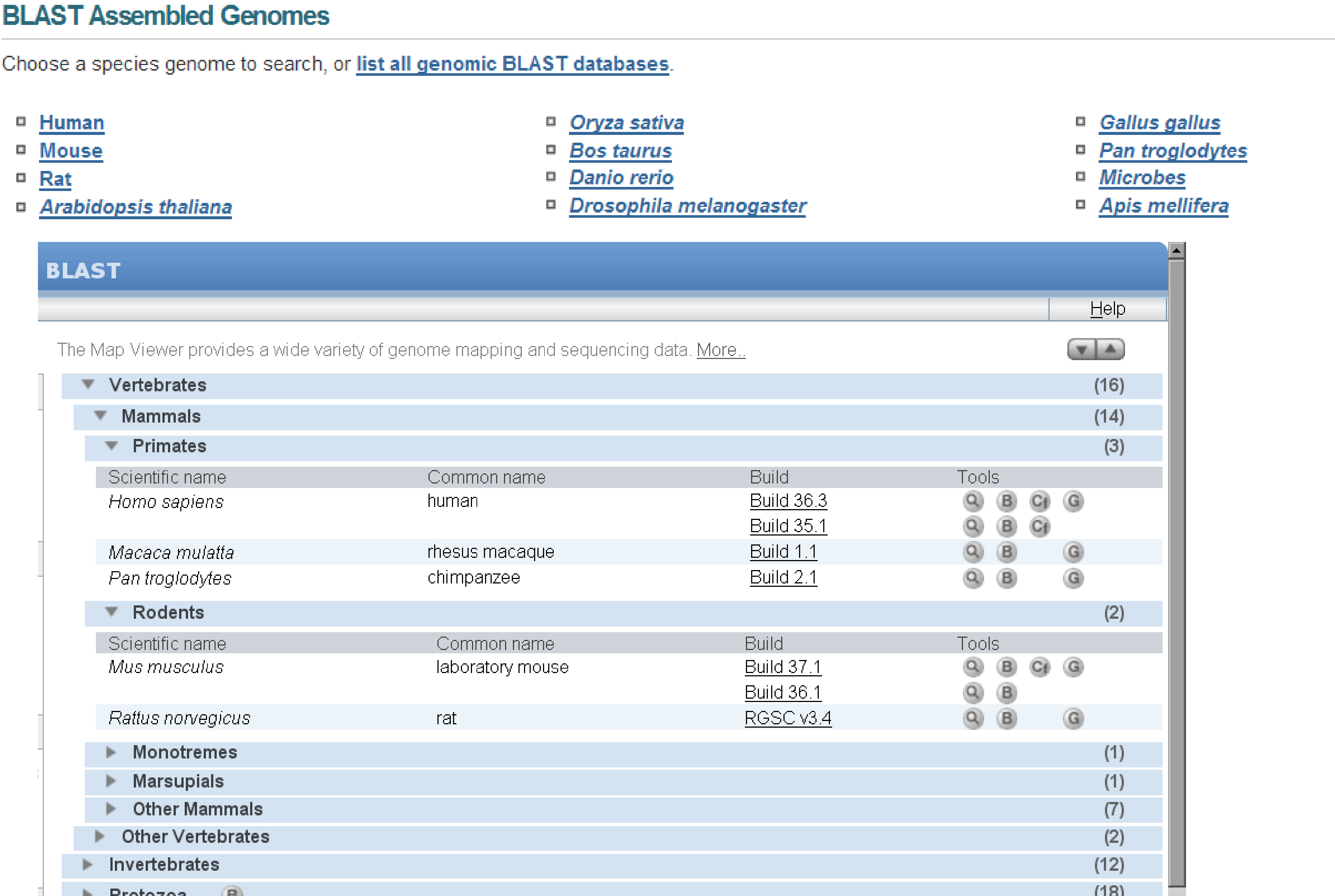
Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated** n ry

Web interface changes now and then

http://blast.ncbi.nlm.nih.gov/Blast.cgi



BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

BLAST

Help

The Map Viewer provides a wide variety of genome mapping and sequencing data. More..

- Vertebrates (16)
  - Mammals (14)
    - Primates (3)

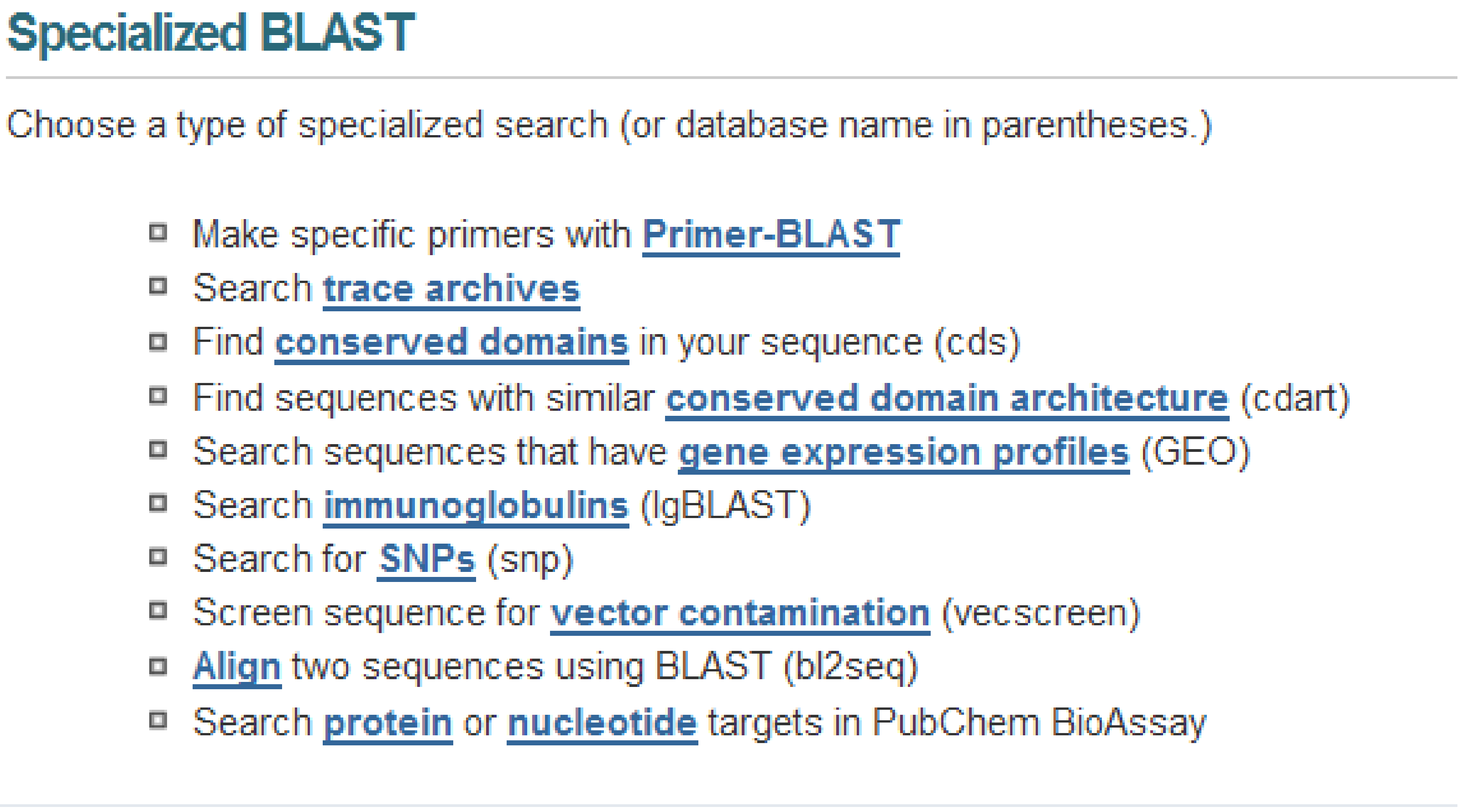
Scientific name	Common name	Build	Tools
Homo sapiens	human	Build 36.3	🔍 B Cf G
		Build 35.1	🔍 B Cf
Macaca mulatta	rhesus macaque	Build 1.1	🔍 B G
Pan troglodytes	chimpanzee	Build 2.1	🔍 B G

- Rodents (2)

Scientific name	Common name	Build	Tools
Mus musculus	laboratory mouse	Build 37.1	🔍 B Cf G
		Build 36.1	🔍 B
Rattus norvegicus	rat	RGSC v3.4	🔍 B G

- Monotremes (1)
- Marsupials (1)
- Other Mammals (7)
- Other Vertebrates (2)
- Invertebrates (12)
- Protozoa (18)

http://blast.ncbi.nlm.nih.gov/Blast.cgi

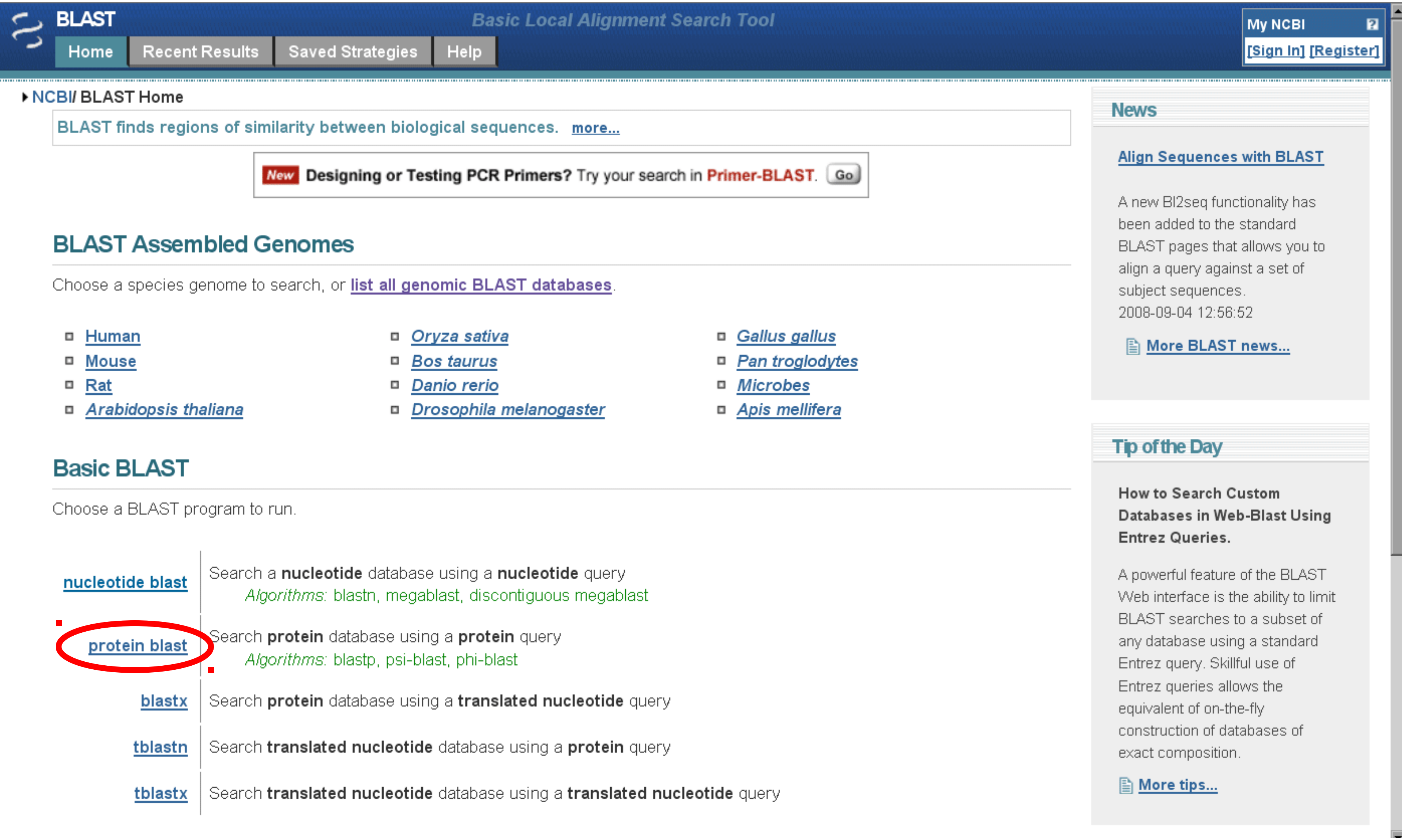


Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with **Primer-BLAST**
- Search **trace archives**
- Find **conserved domains** in your sequence (cds)
- Find sequences with similar **conserved domain architecture** (cdart)
- Search sequences that have **gene expression profiles** (GEO)
- Search **immunoglobulins** (IgBLAST)
- Search for **SNPs** (snp)
- Screen sequence for **vector contamination** (vecscreen)
- **Align** two sequences using BLAST (bl2seq)
- Search **protein** or **nucleotide** targets in PubChem BioAssay

BLASTing a sequence at NCBI – **programs**



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New Designing or Testing PCR Primers? Try your search in Primer-BLAST. Go

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
[✓ protein blast]	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

News

Align Sequences with BLAST

A new Bl2seq functionality has been added to the standard BLAST pages that allows you to align a query against a set of subject sequences.
2008-09-04 12:56:52

More BLAST news...

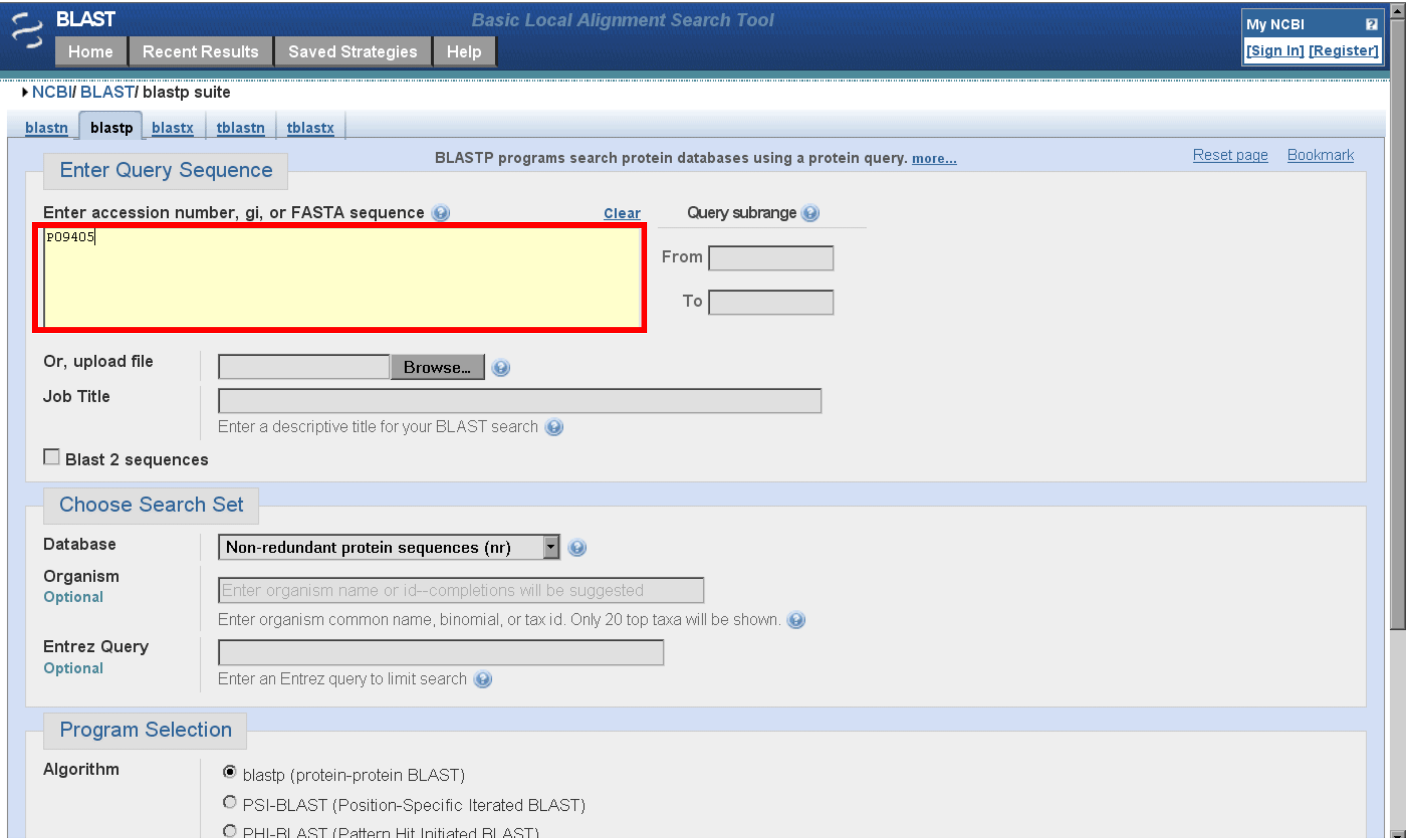
Tip of the Day

How to Search Custom Databases in Web-Blast Using Entrez Queries.

A powerful feature of the BLAST Web interface is the ability to limit BLAST searches to a subset of any database using a standard Entrez query. Skillful use of Entrez queries allows the equivalent of on-the-fly construction of databases of exact composition.

More tips...

BLASTing a sequence at NCBI – **enter accession**



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

P09405

Query subrange

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

□ Blast 2 sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism
Optional

Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query
Optional

Enter an Entrez query to limit search

Program Selection

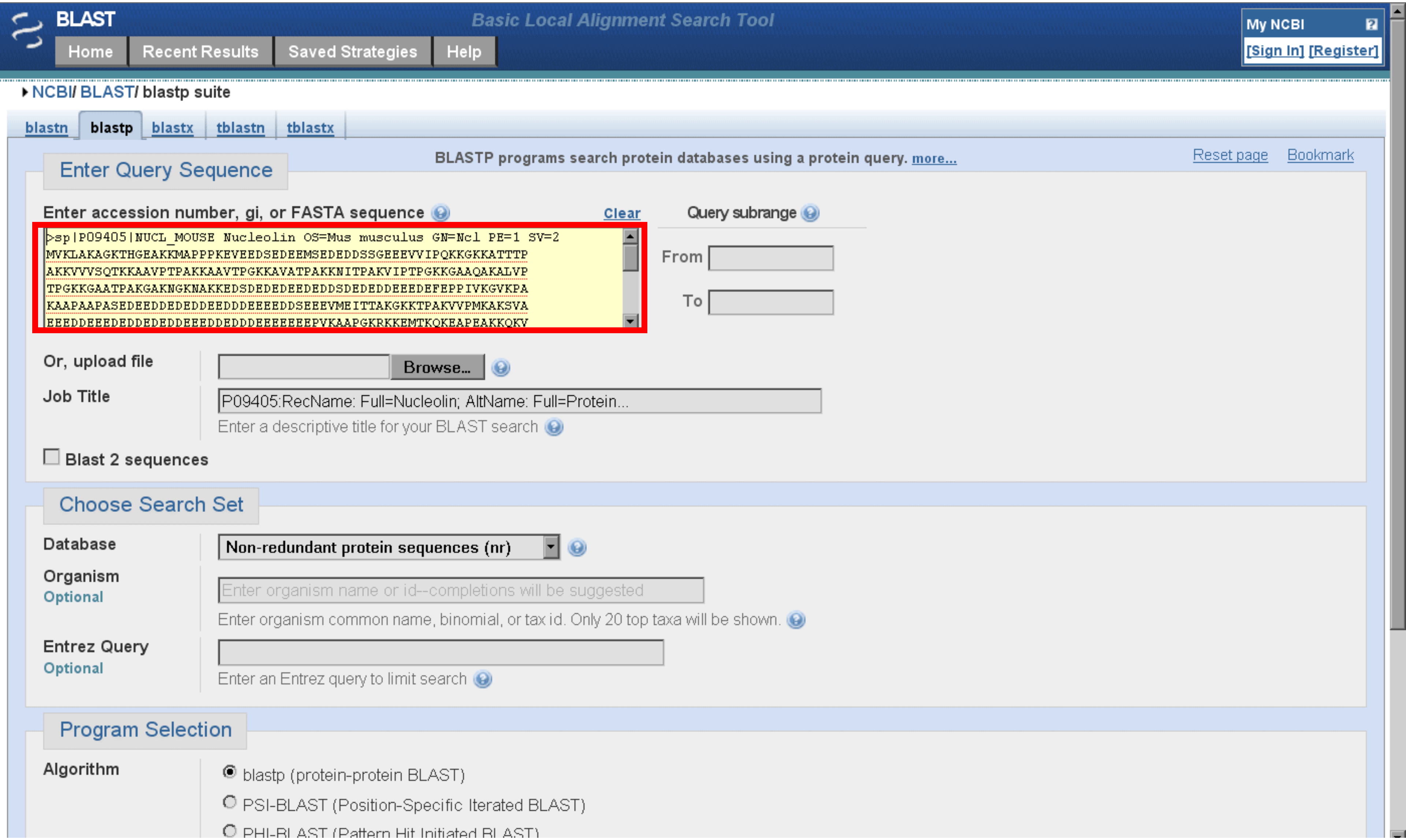
Algorithm

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

BLASTing a sequence at NCBI – **enter sequence**



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

```
>sp|P09405|NUCL_MOUSE Nucleolin OS=Mus musculus GN=Ncl PE=1 SV=2
MVKLAKAGKTHGEAKKMAPPPKEVEEDSEDEEMSEDEDDSSGEEEVVIPQKKGKKATTTP
AKKVVVSQTKKAAVPTPAKKAAVTPGKKAVATPAKKNITPAKVIPTPGKKGAAQAKALVP
TPGKKGAATPAKGAKNGKNAKKEDSDEDEDEEDEDDSDEDEDDEEEDEFEPPIVKGVKPA
KAAPAAPASEDEEDDEDEDDEEDDDEEEEDDSEEEVMEITTAKGKKTPAKVVPMKAKSVA
EEEDDEEEDEDDEDEDDEEEDDEDDDEEEEEEEPVKAAPGKRKKEMTKQKEAPEAKKQKV
```

Query subrange

From

To

Or, upload file

Job Title: P09405:RecName: Full=Nucleolin; AltName: Full=Protein...

Enter a descriptive title for your BLAST search

☐ Blast 2 sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism
Optional

Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query
Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

Database choice

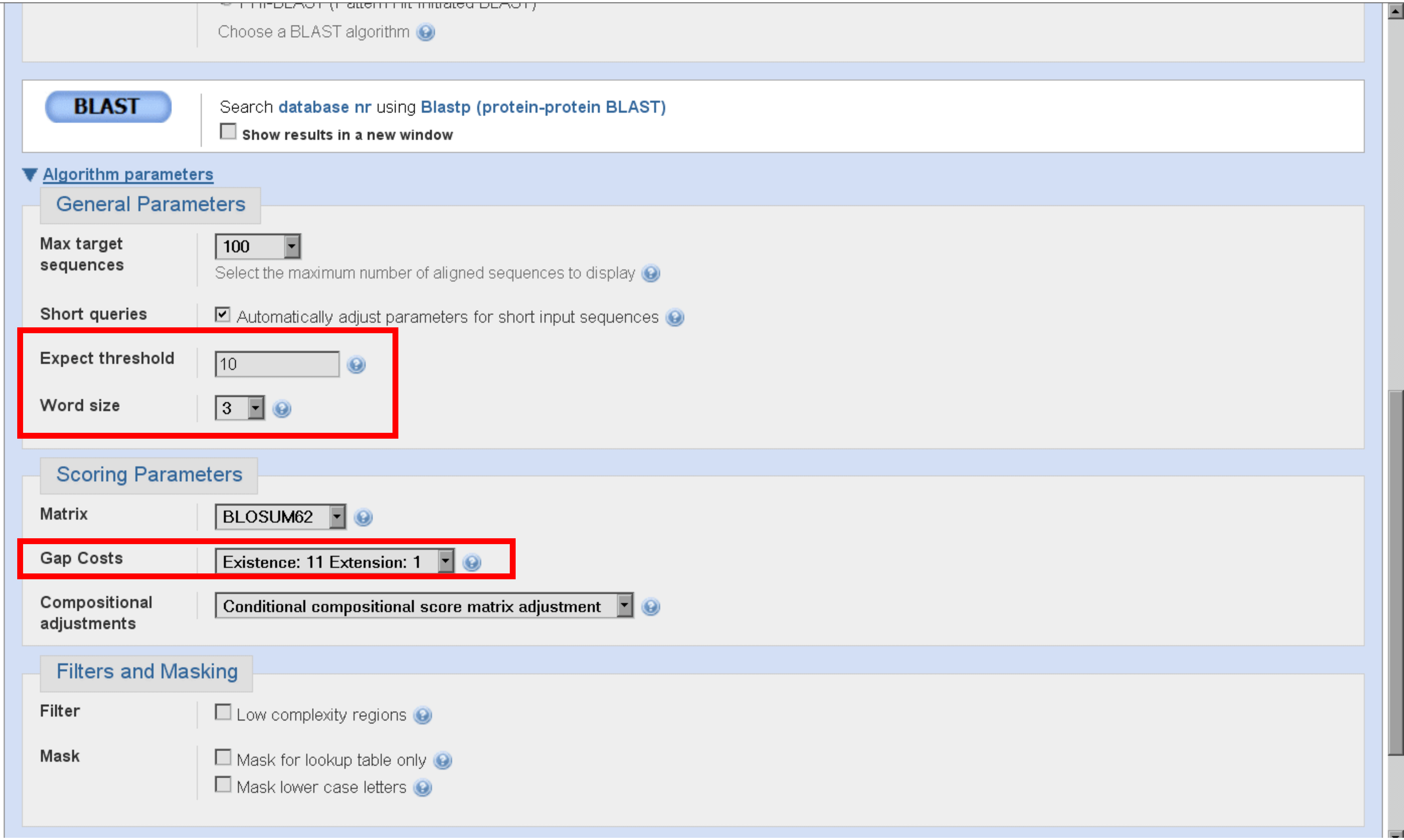
Protein databases

- Good for protein coding nucleotide queries
- Choose a non-redundant database

Nucleotide databases

- Non-redundant database
- Filter on organism / other Entrez query

BLASTing a sequence at NCBI – **parameters**



○ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

▼ Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

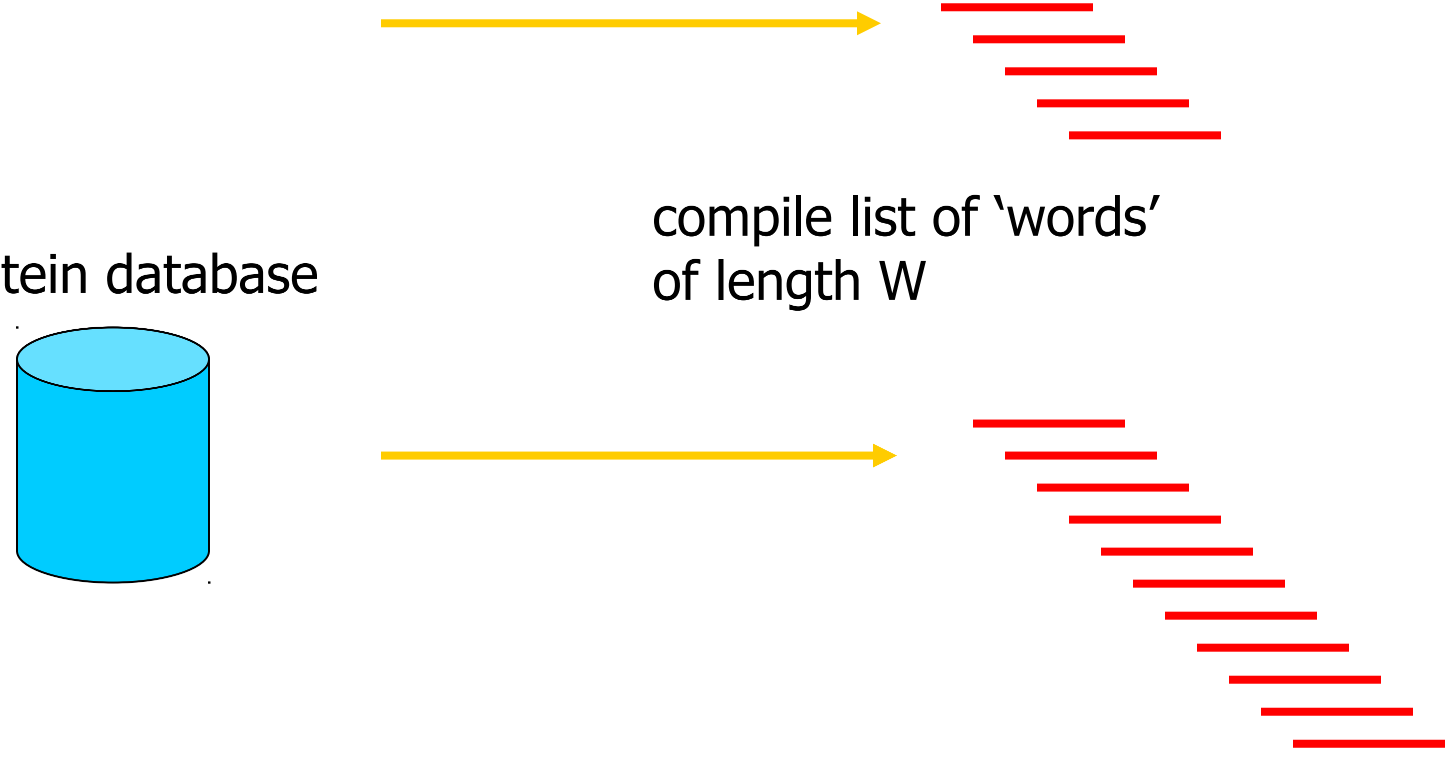
Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: □ Low complexity regions

Mask: □ Mask for lookup table only
□ Mask lower case letters

Blast algorithm: ***step 1***



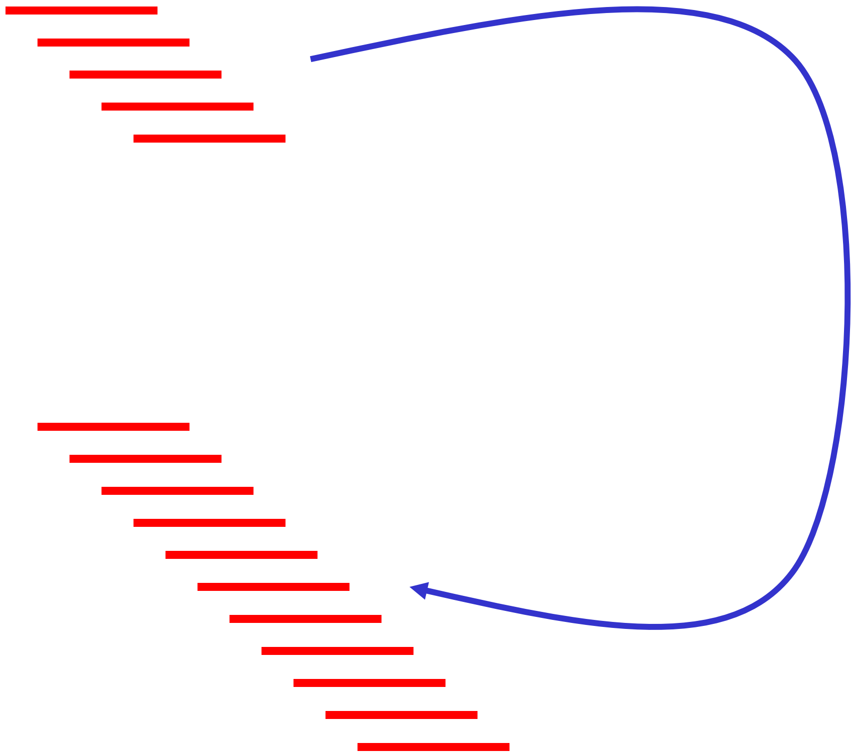
protein query sequence

protein database

compile list of ‘words’
of length W

Blast algorithm: ***step 2***

Initial search



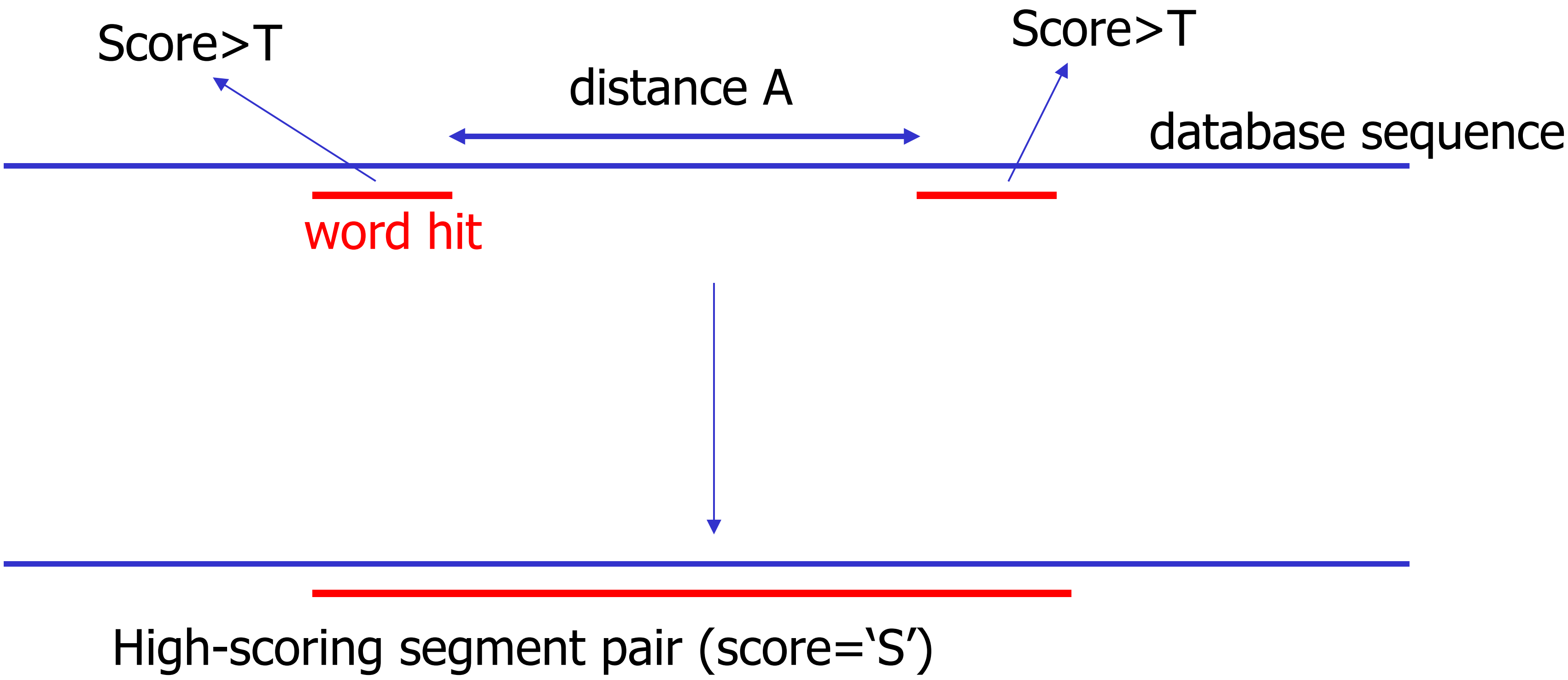
Use PAM/BLOSUM matrix

Find word of length ‘W’ that scores at least ‘T’ (T=11)

Exact matches only

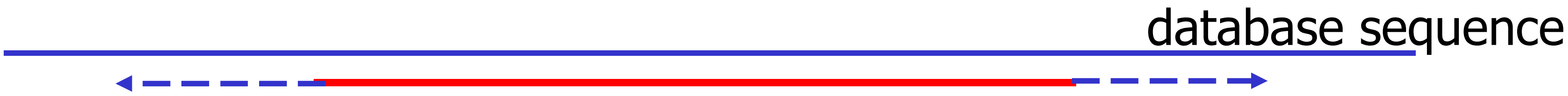
The parameter T dictates the speed and sensitivity
-increasing T increases speed,
 decreases sensitivity

Join words on same diagonal (*ungapped*)



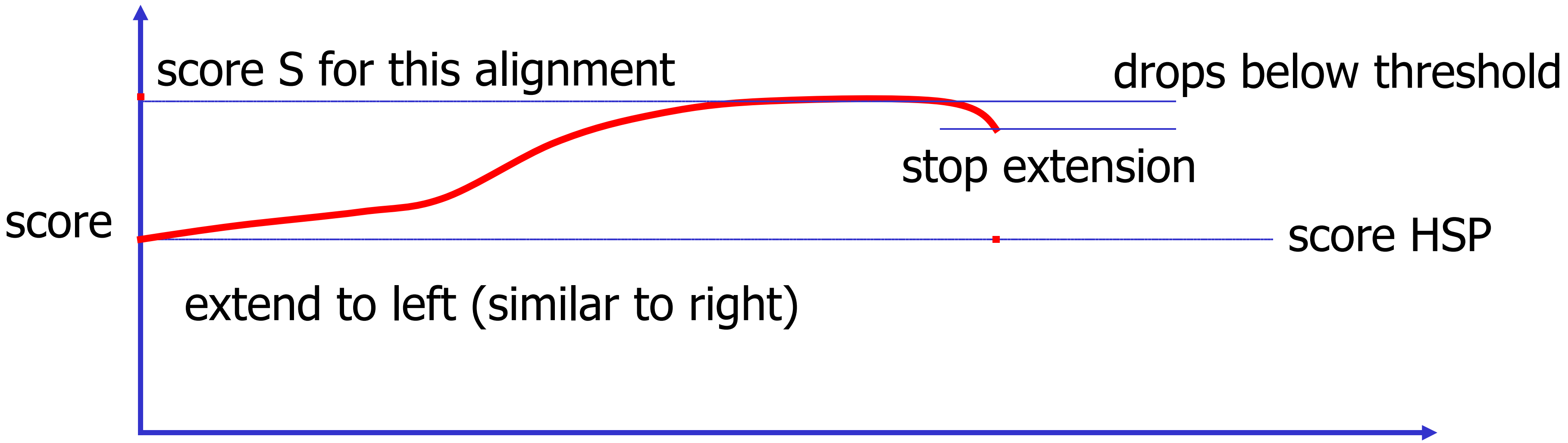
Join words on same diagonal

Extend HSP until score drops small amount below
highest score of shorter alignment



database sequence

High-scoring segment pair (score=‘S’)



score S for this alignment

drops below threshold

stop extension

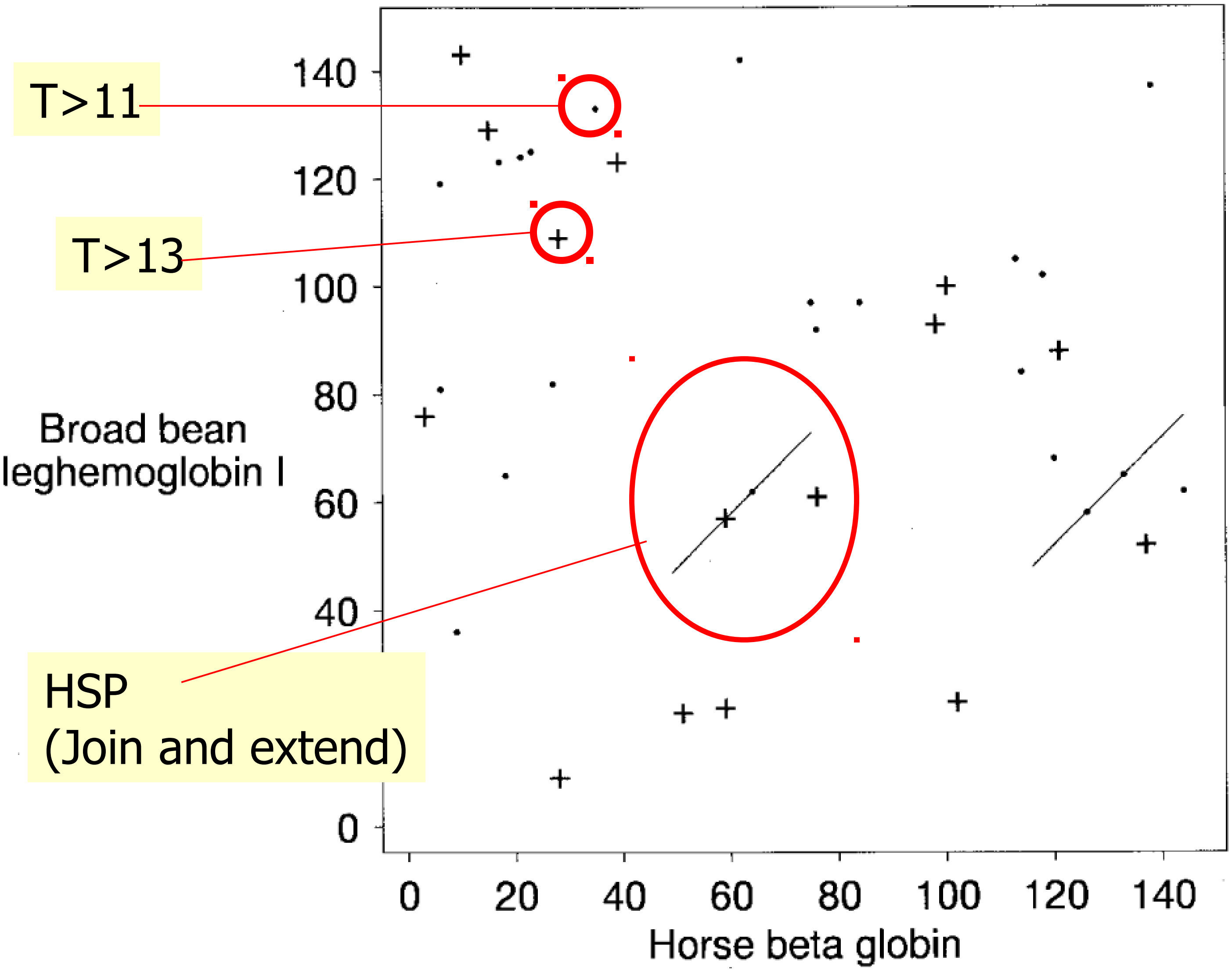
score

score HSP

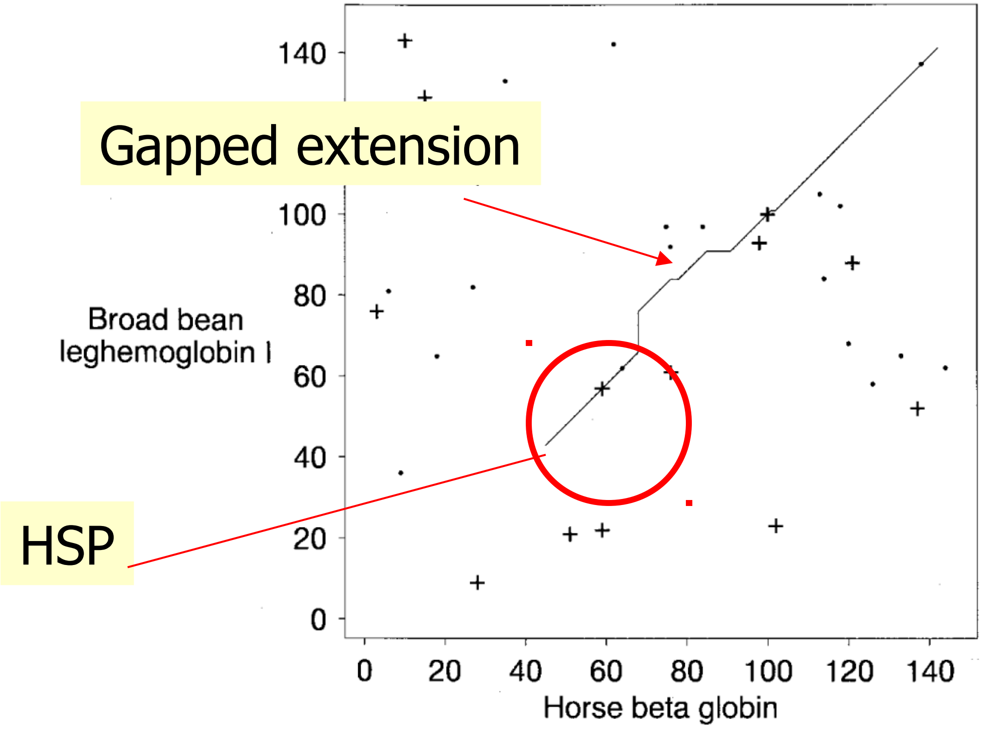
extend to left (similar to right)

If S>threshold (based on random sequences) then keep HSP

Finding HSP's



Trigger gapped extension

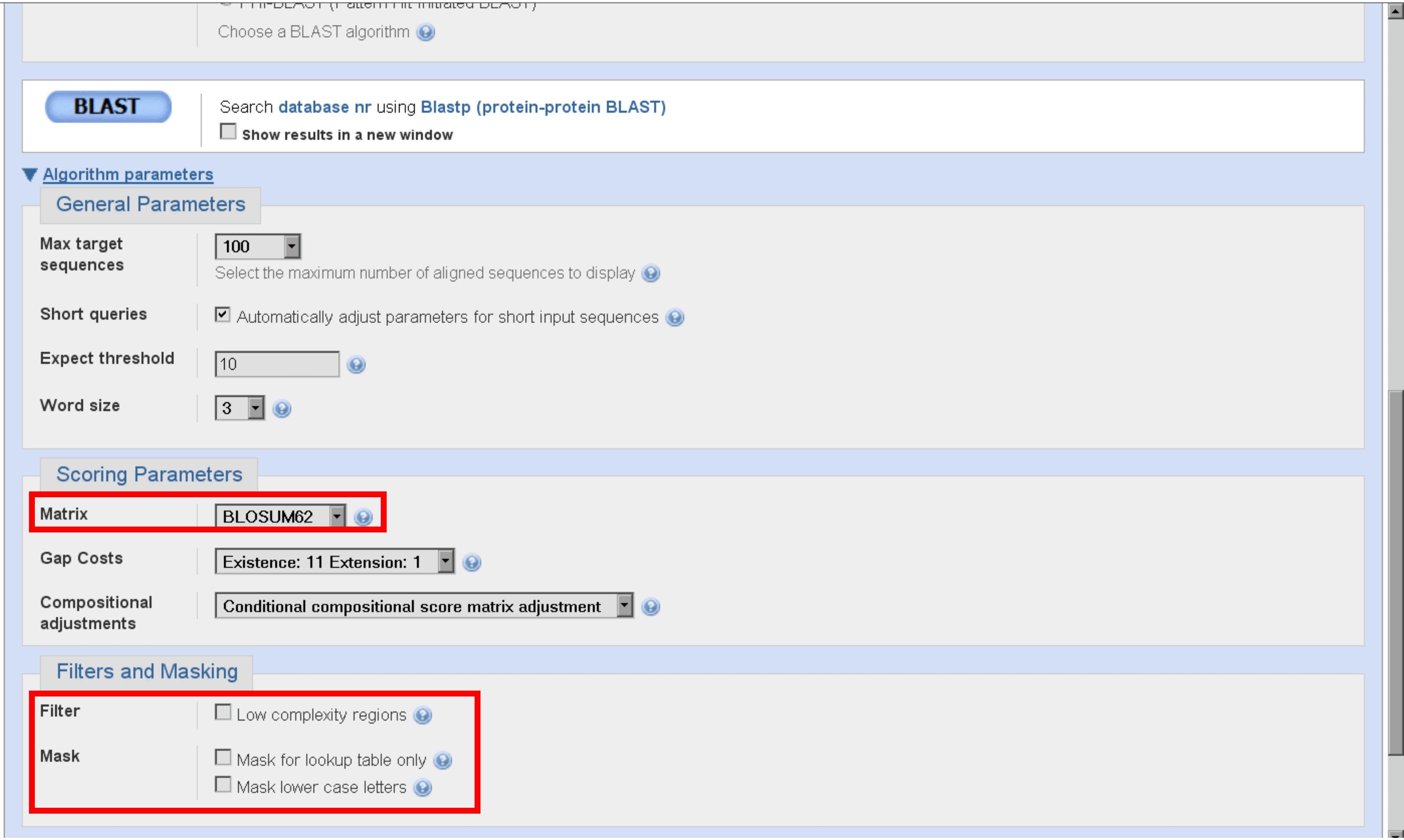


Gapped extension

HSP

```
Leghemoglobin  43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------  90
                  F  L +   V+ +PK+ AH +KV          L + GE V  LD   G+
Beta globin    45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHHLDNLKGTFAALSE  90


Leghemoglobin  91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
                  +H  K  +DP +F ++   L+  +    G  ++ EL A+++    G+A A+
Beta globin    91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

BLASTing a sequence at NCBI – **parameters**



○ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

▼ Algorithm parameters

General Parameters

Max target sequences: 100

Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: □ Low complexity regions

Mask: □ Mask for lookup table only

□ Mask lower case letters

Masking of sequences – **low complexity**

Low complexity repeats in genome

Many amino-acid “stretches” in proteins

BLAST recognizes these regions as similar

but, they are NOT biologically related

Masking of sequences – **highly abundant sequences**

First query sequence against database that contains domains representative of large sequence families

- Alu repeats
- Protein kinase catalytic domains
- Vector sequences

Then mask these domains in the query sequence and continue search

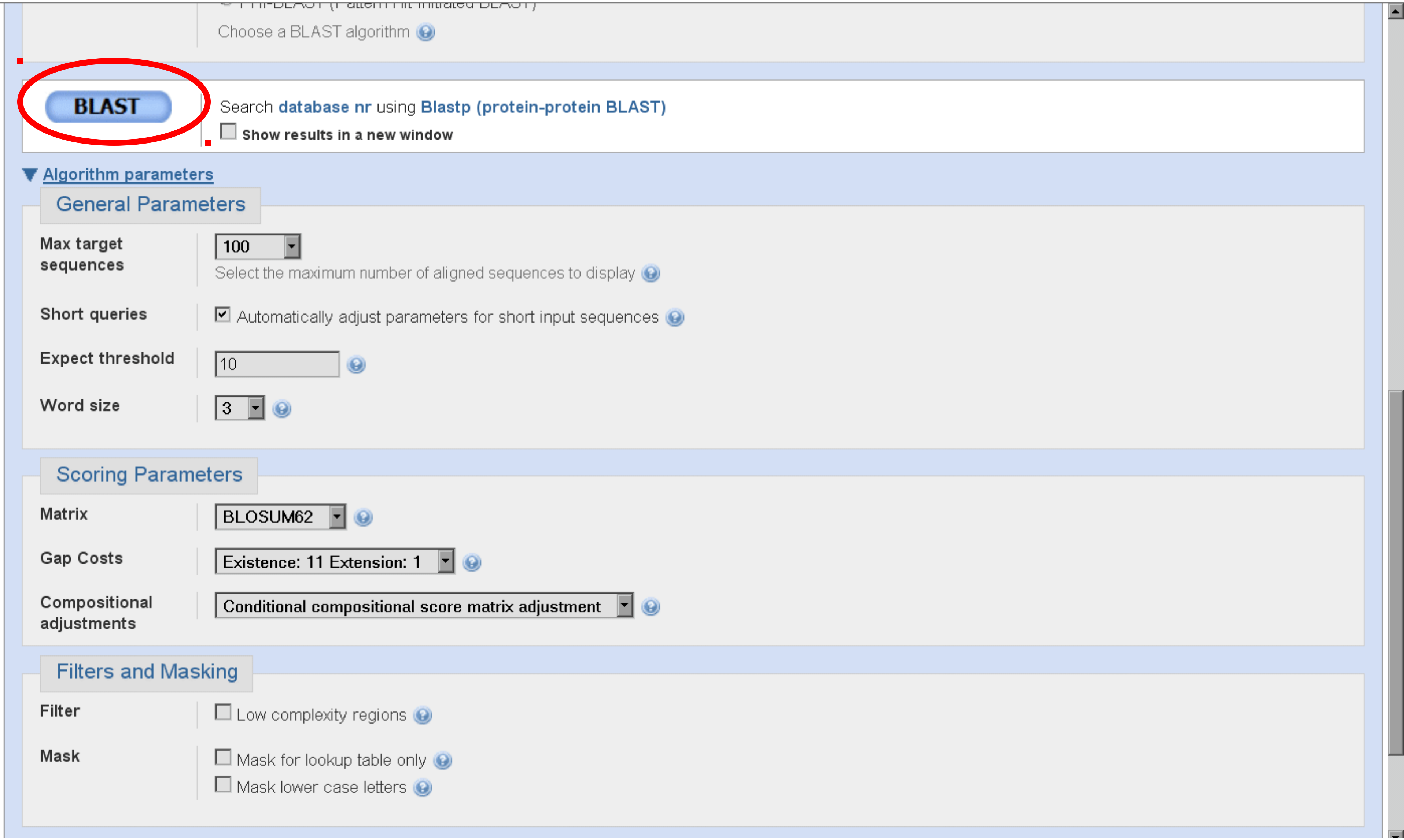
Masking option replaces these regions with XXXXXXX

When do you change the parameters?

Reason	Parameters to change
The sequence you’re interested in contains many identical residues; it has a biased composition	Sequence filter (automatic masking)
BLAST doesn’t report any results	the gap
Your match has a bor	atrix or the gap check the match’s ustness
BLAST reports too	The database you’re searching OR filter the reported entries by keyword OR increase the nr of reported matches OR increase Expect (the evalue threshold) OR reject sequences too similar to the query (those with very low e-values)

Parameters are already optimized

BLASTing a sequence at NCBI – **parameters**



○ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

▼ Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

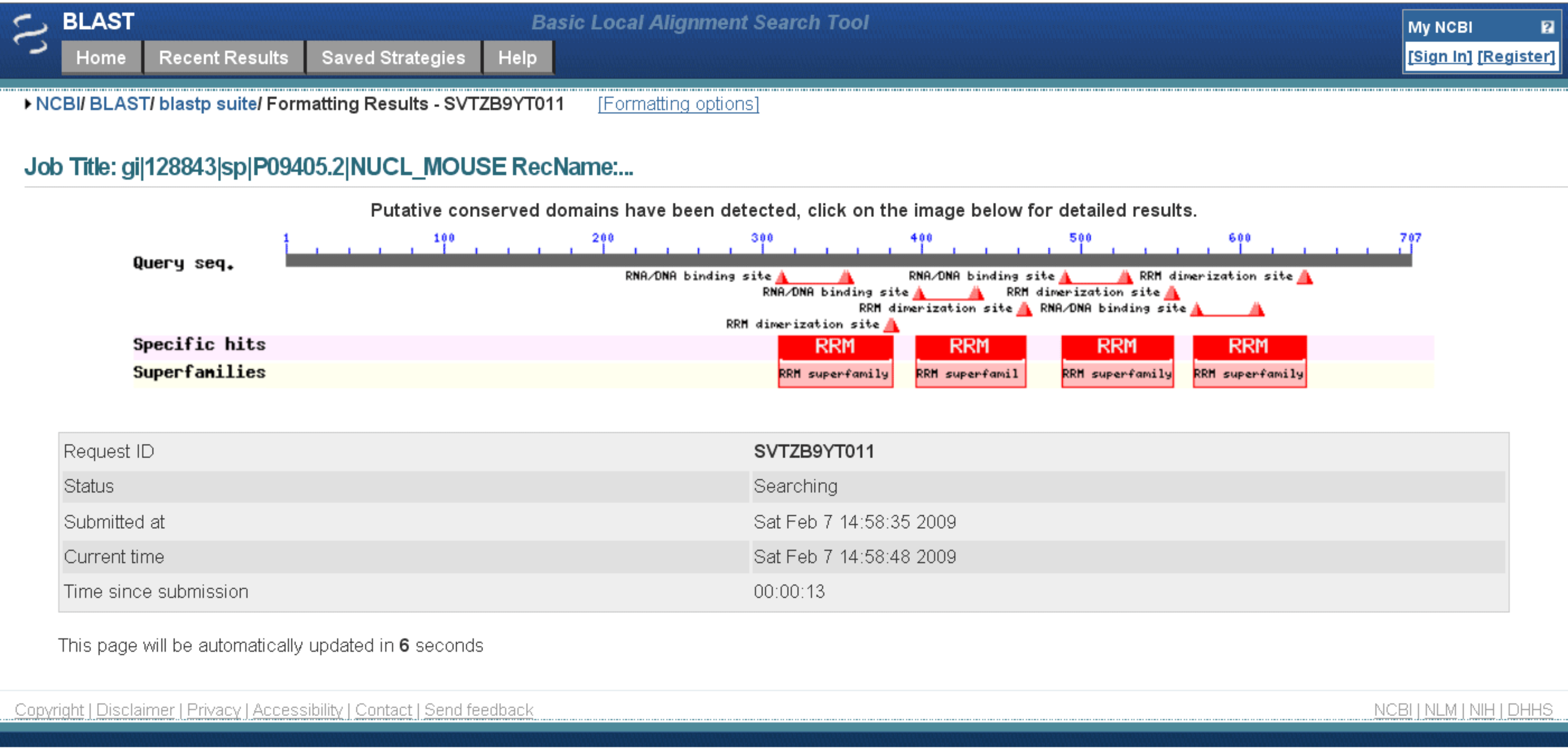
Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

Filter: □ Low complexity regions

Mask: □ Mask for lookup table only
□ Mask lower case letters

BLASTing a sequence at NCBI – **job status**



NCBI/ BLAST/ blastp suite/ Formatting Results - SVTZB9YT011 [Formatting options]

Job Title: gi|128843|sp|P09405.2|NUCL_MOUSE RecName:...

Putative conserved domains have been detected, click on the image below for detailed results.

Request ID	**SVTZB9YT011**
Status	Searching
Submitted at	Sat Feb 7 14:58:35 2009
Current time	Sat Feb 7 14:58:48 2009
Time since submission	00:00:13

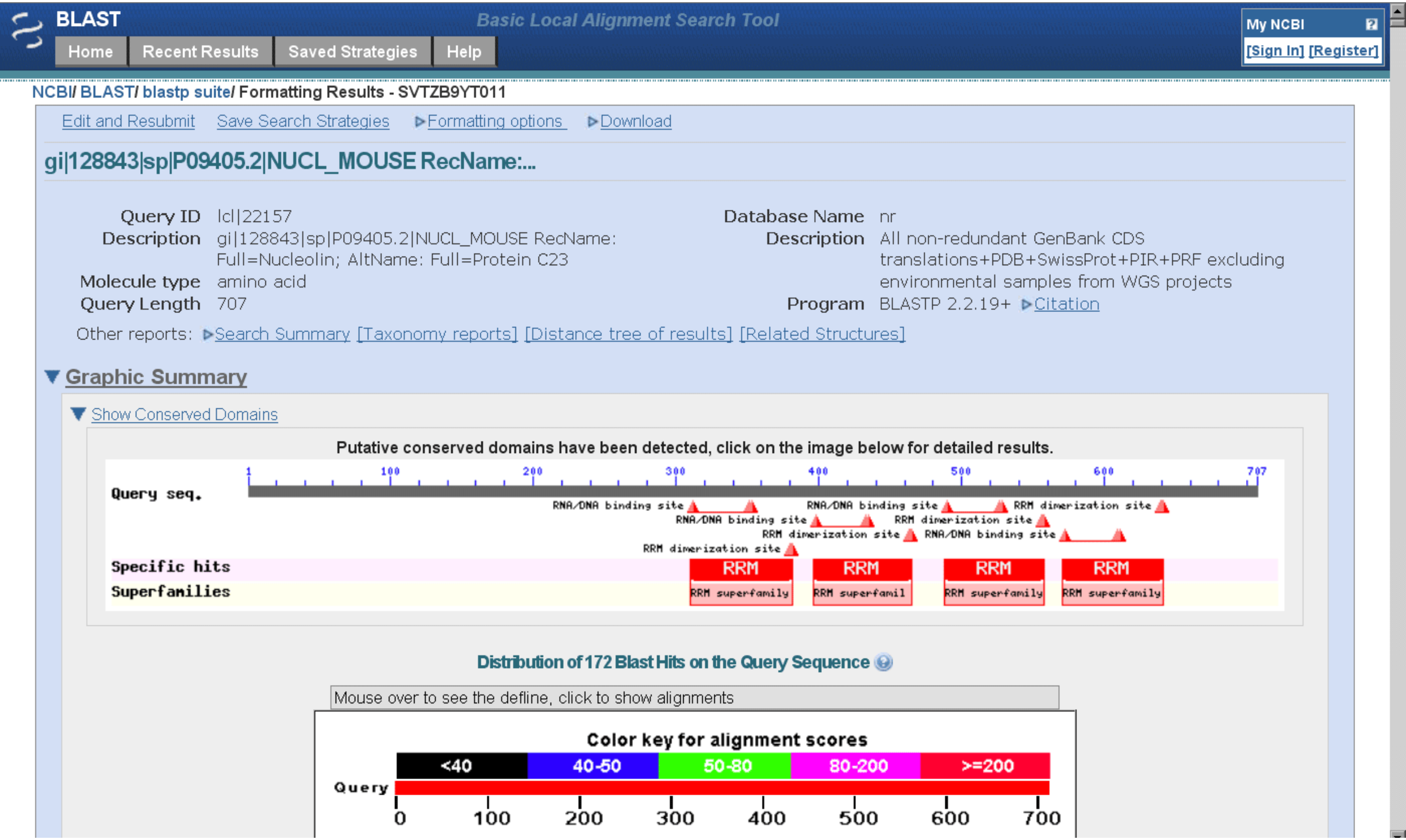
This page will be automatically updated in **6** seconds

If it takes too long: try another BLAST server

Country / continent	Program	URL
USA	BLAST / PSI-BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi
Europe	BLAST	http://www.expasy.org/tools/blast/ bBLAST.html
Europe	BLAST (WU-BLAST)	http://www.ebi.ac.uk/services
Japan	BLAST / PSI-BLAST	http://blast.ddbj.nig.ac.jp/top-e.html

Warning:
different database (versions) !

BLASTing a sequence at NCBI – **blast summary**



BLAST

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI

[Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - SVTZB9YT011

Edit and Resubmit   Save Search Strategies   ▷Formatting options   ▷Download

gi|128843|sp|P09405.2|NUCL_MOUSE RecName:...

Query ID	lcl|22157	Database Name	nr
Description	gi|128843|sp|P09405.2|NUCL_MOUSE RecName: Full=Nucleolin; AltName: Full=Protein C23	Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	amino acid		
Query Length	707	Program	BLASTP 2.2.19+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures]

▼ Graphic Summary

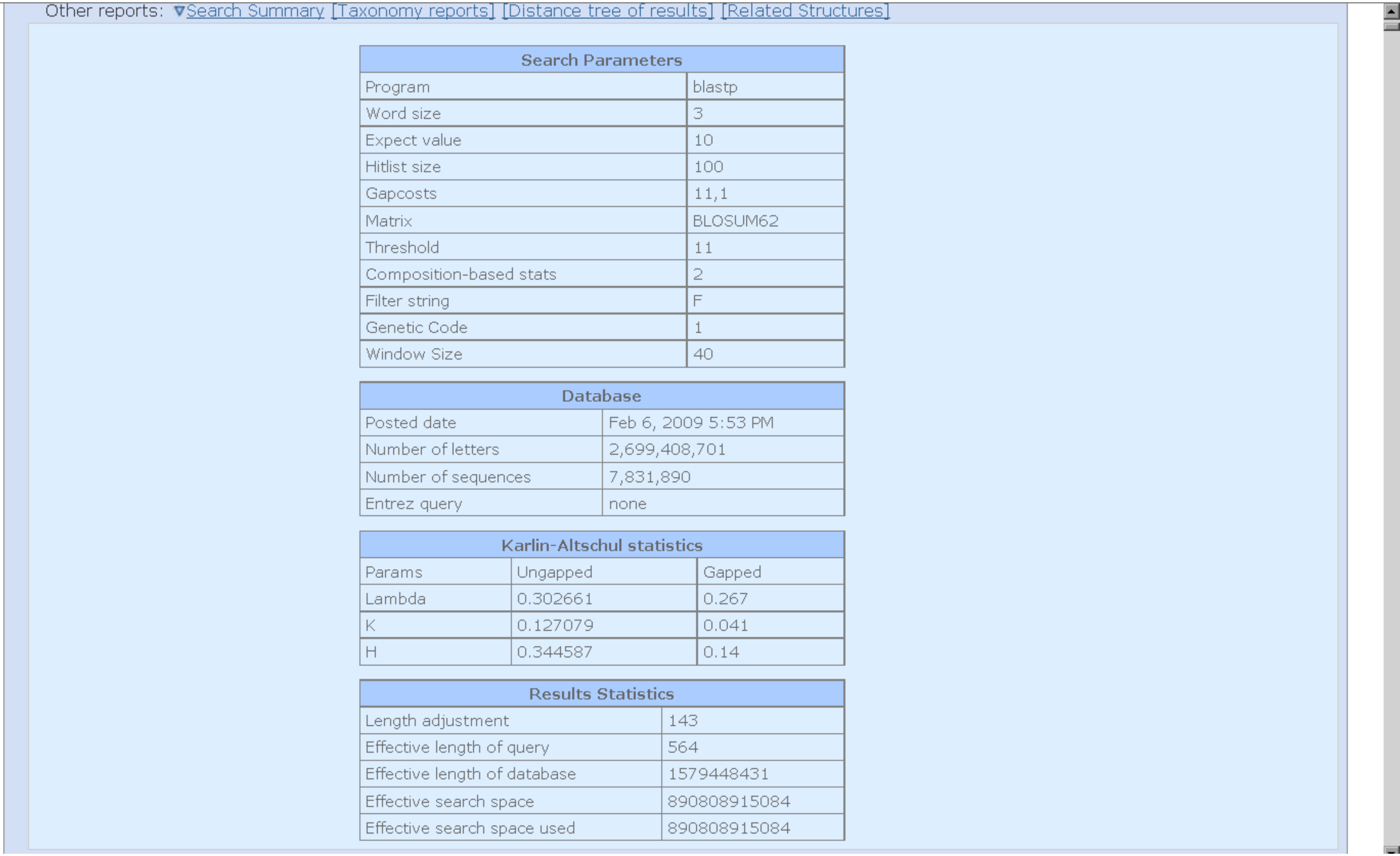
▼ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 172 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

BLASTing a sequence at NCBI – **used parameters**



Other reports: ▼Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures]

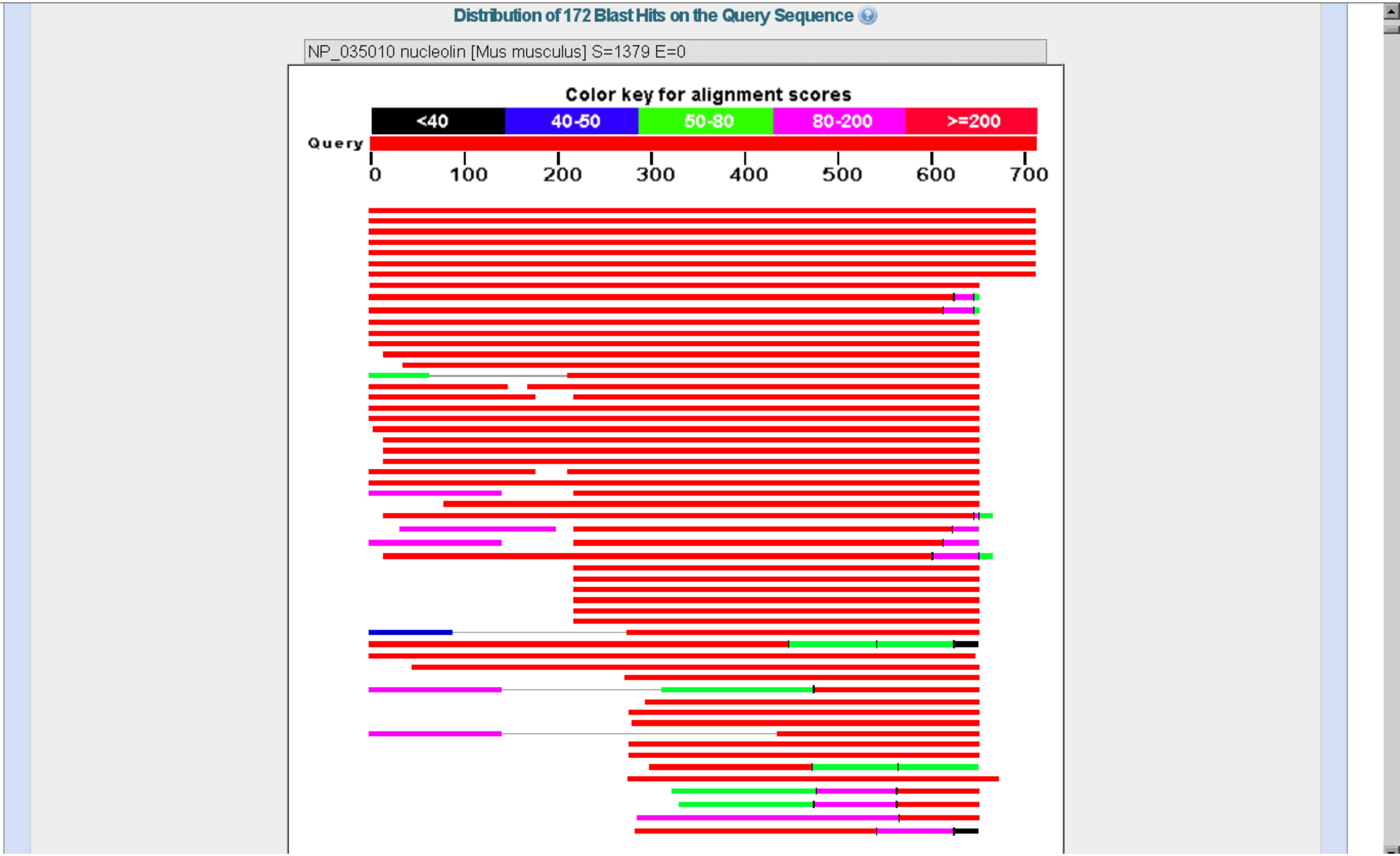
Search Parameters	
Program	blastp
Word size	3
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Threshold	11
Composition-based stats	2
Filter string	F
Genetic Code	1
Window Size	40

Database	
Posted date	Feb 6, 2009 5:53 PM
Number of letters	2,699,408,701
Number of sequences	7,831,890
Entrez query	none

Karlin-Altschul statistics		
Params	Ungapped	Gapped
Lambda	0.302661	0.267
K	0.127079	0.041
H	0.344587	0.14

Results Statistics	
Length adjustment	143
Effective length of query	564
Effective length of database	1579448431
Effective search space	890808915084
Effective search space used	890808915084

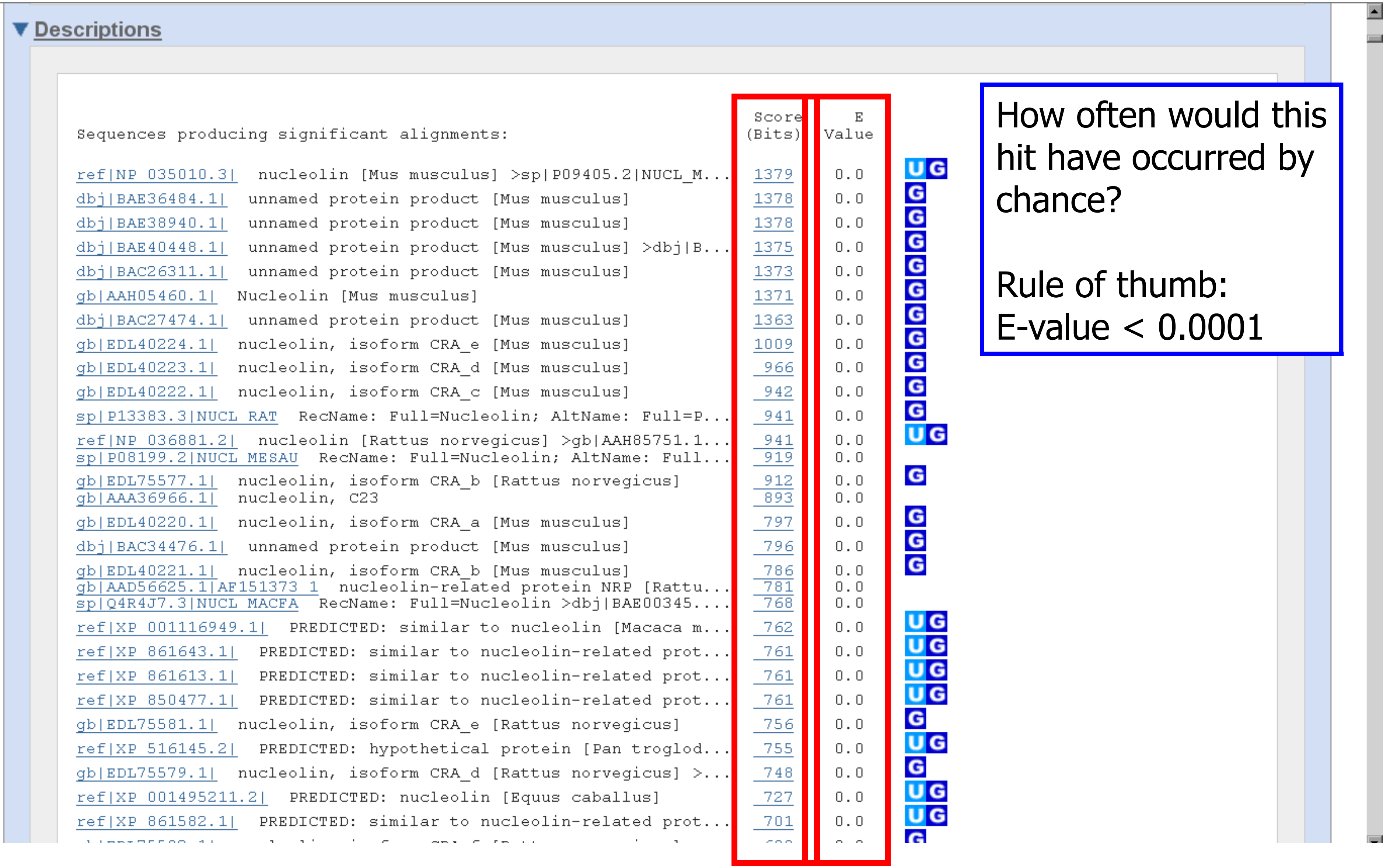
BLASTing a sequence at NCBI – **graphical display**



Distribution of 172 Blast Hits on the Query Sequence

NP_035010 nucleolin [Mus musculus] S=1379 E=0

BLASTing a sequence at NCBI – **hit list**



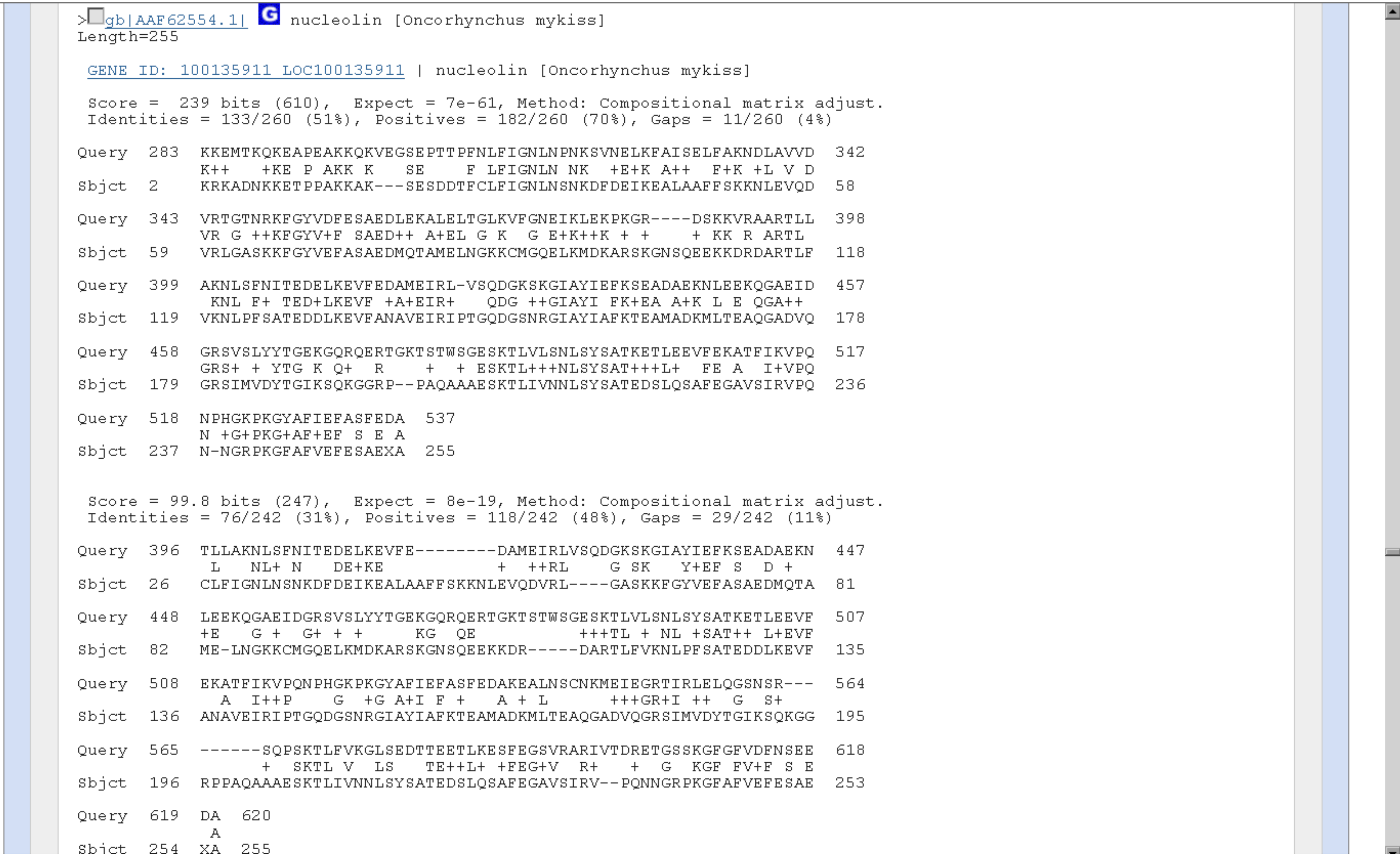
▼ Descriptions

```
                                                                        Score     E
Sequences producing significant alignments:                            (Bits)  Value

ref|NP_035010.3|  nucleolin [Mus musculus] >sp|P09405.2|NUCL_M...      1379    0.0     U G
dbj|BAE36484.1|  unnamed protein product [Mus musculus]                 1378    0.0     G
dbj|BAE38940.1|  unnamed protein product [Mus musculus]                 1378    0.0     G
dbj|BAE40448.1|  unnamed protein product [Mus musculus] >dbj|B...      1375    0.0     G
dbj|BAC26311.1|  unnamed protein product [Mus musculus]                 1373    0.0     G
gb|AAH05460.1|  Nucleolin [Mus musculus]                                1371    0.0     G
dbj|BAC27474.1|  unnamed protein product [Mus musculus]                 1363    0.0     G
gb|EDL40224.1|  nucleolin, isoform CRA_e [Mus musculus]                 1009    0.0     G
gb|EDL40223.1|  nucleolin, isoform CRA_d [Mus musculus]                  966    0.0     G
gb|EDL40222.1|  nucleolin, isoform CRA_c [Mus musculus]                  942    0.0     G
sp|P13383.3|NUCL_RAT  RecName: Full=Nucleolin; AltName: Full=P...       941    0.0     G
ref|NP_036881.2|  nucleolin [Rattus norvegicus] >gb|AAH85751.1...       941    0.0     U G
sp|P08199.2|NUCL_MESAU  RecName: Full=Nucleolin; AltName: Full...       919    0.0
gb|EDL75577.1|  nucleolin, isoform CRA_b [Rattus norvegicus]            912    0.0     G
gb|AAA36966.1|  nucleolin, C23                                          893    0.0
gb|EDL40220.1|  nucleolin, isoform CRA_a [Mus musculus]                 797    0.0     G
dbj|BAC34476.1|  unnamed protein product [Mus musculus]                 796    0.0     G
gb|EDL40221.1|  nucleolin, isoform CRA_b [Mus musculus]                 786    0.0     G
gb|AAD56625.1|AF151373_1  nucleolin-related protein NRP [Rattu...       781    0.0
sp|Q4R4J7.3|NUCL_MACFA  RecName: Full=Nucleolin >dbj|BAE00345....       768    0.0
ref|XP_001116949.1|  PREDICTED: similar to nucleolin [Macaca m...       762    0.0     U G
ref|XP_861643.1|  PREDICTED: similar to nucleolin-related prot...       761    0.0     U G
ref|XP_861613.1|  PREDICTED: similar to nucleolin-related prot...       761    0.0     U G
ref|XP_850477.1|  PREDICTED: similar to nucleolin-related prot...       761    0.0     U G
gb|EDL75581.1|  nucleolin, isoform CRA_e [Rattus norvegicus]            756    0.0     G
ref|XP_516145.2|  PREDICTED: hypothetical protein [Pan troglod...       755    0.0     U G
gb|EDL75579.1|  nucleolin, isoform CRA_d [Rattus norvegicus] >...       748    0.0     G
ref|XP_001495211.2|  PREDICTED: nucleolin [Equus caballus]              727    0.0     U G
ref|XP_861582.1|  PREDICTED: similar to nucleolin-related prot...       701    0.0     U G
```

How often would this hit have occurred by chance?

Rule of thumb:
E-value < 0.0001

BLASTing a sequence at NCBI



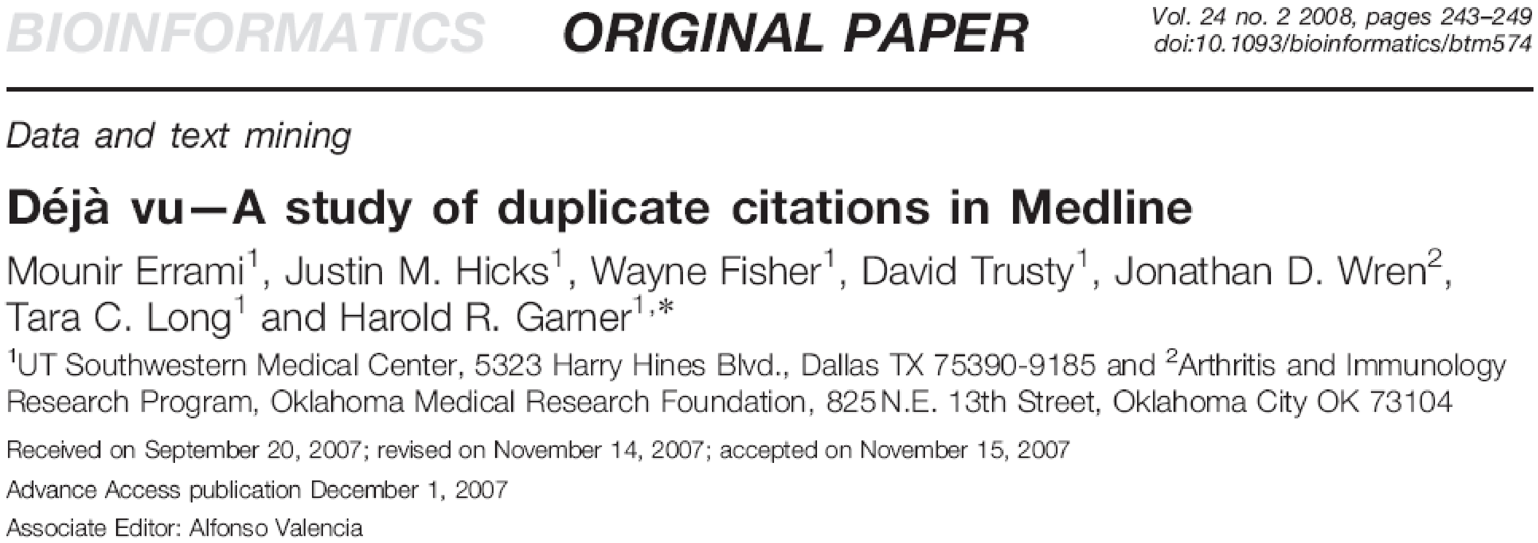
```
>□gb|AAF62554.1| G nucleolin [Oncorhynchus mykiss]
Length=255

 GENE ID: 100135911 LOC100135911 | nucleolin [Oncorhynchus mykiss]

 Score =  239 bits (610),  Expect = 7e-61, Method: Compositional matrix adjust.
 Identities = 133/260 (51%), Positives = 182/260 (70%), Gaps = 11/260 (4%)

Query  283  KKEMTKQKEAPEAKKQKVEGSEPTTPFNLFIGNLNPNKSVNELKFAISELFAKNDLAVVD  342
            K++   +KE P AKK K   SE    F LFIGNLN NK  +E+K A++  F+K +L V D
Sbjct  2    KRKADNKKETPPAKKAK---SESDDTFCLFIGNLNSNKDFDEIKEALAAFFSKKNLEVQD  58

Query  343  VRTGTNRKFGYVDFESAEDLEKALELTGLKVFGNEIKLEKPKGR----DSKKVRAARTLL  398
            VR G ++KFGYV+F SAED++ A+EL G K  G E+K++K + +    + KK R ARTL
Sbjct  59   VRLGASKKFGYVEFASAEDMQTAMELNGKKCMGQELKMDKARSKGNSQEEKKDRDARTLF  118

Query  399  AKNLSFNITEDELKEVFEDAMEIRL-VSQDGKSKGIAYIEFKSEADAEKNLEEKQGAEID  457
             KNL F+ TED+LKEVF +A+EIR+   QDG ++GIAYI FK+EA A+K L E QGA++
Sbjct  119  VKNLPFSATEDDLKEVFANAVEIRIPTGQDGSNRGIAYIAFKTEAMADKMLTEAQGADVQ  178

Query  458  GRSVSLYYTGEKGQRQERTGKTSTWSGESKTLVLSNLSYSATKETLEEVFEKATFIKVPQ  517
            GRS+ + YTG K Q+  R    +  + ESKTL+++NLSYSAT+++L+  FE A  I+VPQ
Sbjct  179  GRSIMVDYTGIKSQKGGRP--PAQAAAESKTLIVNNLSYSATEDSLQSAFEGAVSIRVPQ  236

Query  518  NPHGKPKGYAFIEFASFEDA  537
            N +G+PKG+AF+EF S E A
Sbjct  237  N-NGRPKGFAFVEFESAEXA  255


 Score = 99.8 bits (247),  Expect = 8e-19, Method: Compositional matrix adjust.
 Identities = 76/242 (31%), Positives = 118/242 (48%), Gaps = 29/242 (11%)

Query  396  TLLAKNLSFNITEDELKEVFE--------DAMEIRLVSQDGKSKGIAYIEFKSEADAEKN  447
             L   NL+ N   DE+KE           +  ++RL    G SK   Y+EF S  D +
Sbjct  26   CLFIGNLNSNKDFDEIKEALAAFFSKKNLEVQDVRL----GASKKFGYVEFASAEDMQTA  81

Query  448  LEEKQGAEIDGRSVSLYYTGEKGQRQERTGKTSTWSGESKTLVLSNLSYSATKETLEEVF  507
            +E   G +  G+ + +     KG  QE          +++TL + NL +SAT++ L+EVF
Sbjct  82   ME-LNGKKCMGQELKMDKARSKGNSQEEKKDR-----DARTLFVKNLPFSATEDDLKEVF  135

Query  508  EKATFIKVPQNPHGKPKGYAFIEFASFEDAKEALNSCNKMEIEGRTIRLELQGSNSR---  564
              A  I++P    G  +G A+I F +   A + L      +++GR+I ++  G  S+
Sbjct  136  ANAVEIRIPTGQDGSNRGIAYIAFKTEAMADKMLTEAQGADVQGRSIMVDYTGIKSQKGG  195

Query  565  ------SQPSKTLFVKGLSEDTTEETLKESFEGSVRARIVTDRETGSSKGFGFVDFNSEE  618
                  +  SKTL V  LS    TE++L+ +FEG+V  R+    +  G  KGF FV+F S E
Sbjct  196  RPPAQAAAESKTLIVNNLSYSATEDSLQSAFEGAVSIRV--PQNNGRPKGFAFVEFESAE  253

Query  619  DA  620
             A
Sbjct  254  XA  255
```

Alternatives for homology searches

Country / continent	**Program**	**Address**
USA	FASTA	http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml
Europe	FASTA	http://www.ebi.ac.uk/Tools/sss/fasta/
Europe	SSEARCH	http://www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=fasta&program=ssearch&context=protein
USA	BLAT	http://genome.ucsc.edu/

Alternative use
Of alignment algorithm

Pairwise comparison of Medline abstracts



BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 2 2008, pages 243–249
doi:10.1093/bioinformatics/btm574

Data and text mining

Déjà vu—A study of duplicate citations in Medline

Mounir Errami[1], Justin M. Hicks[1], Wayne Fisher[1], David Trusty[1], Jonathan D. Wren[2], Tara C. Long[1] and Harold R. Garner[1,*]

[1]UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas TX 75390-9185 and [2]Arthritis and Immunology Research Program, Oklahoma Medical Research Foundation, 825 N.E. 13th Street, Oklahoma City OK 73104

Received on September 20, 2007; revised on November 14, 2007; accepted on November 15, 2007

Advance Access publication December 1, 2007

Associate Editor: Alfonso Valencia

Pairwise comparison of Medline abstracts

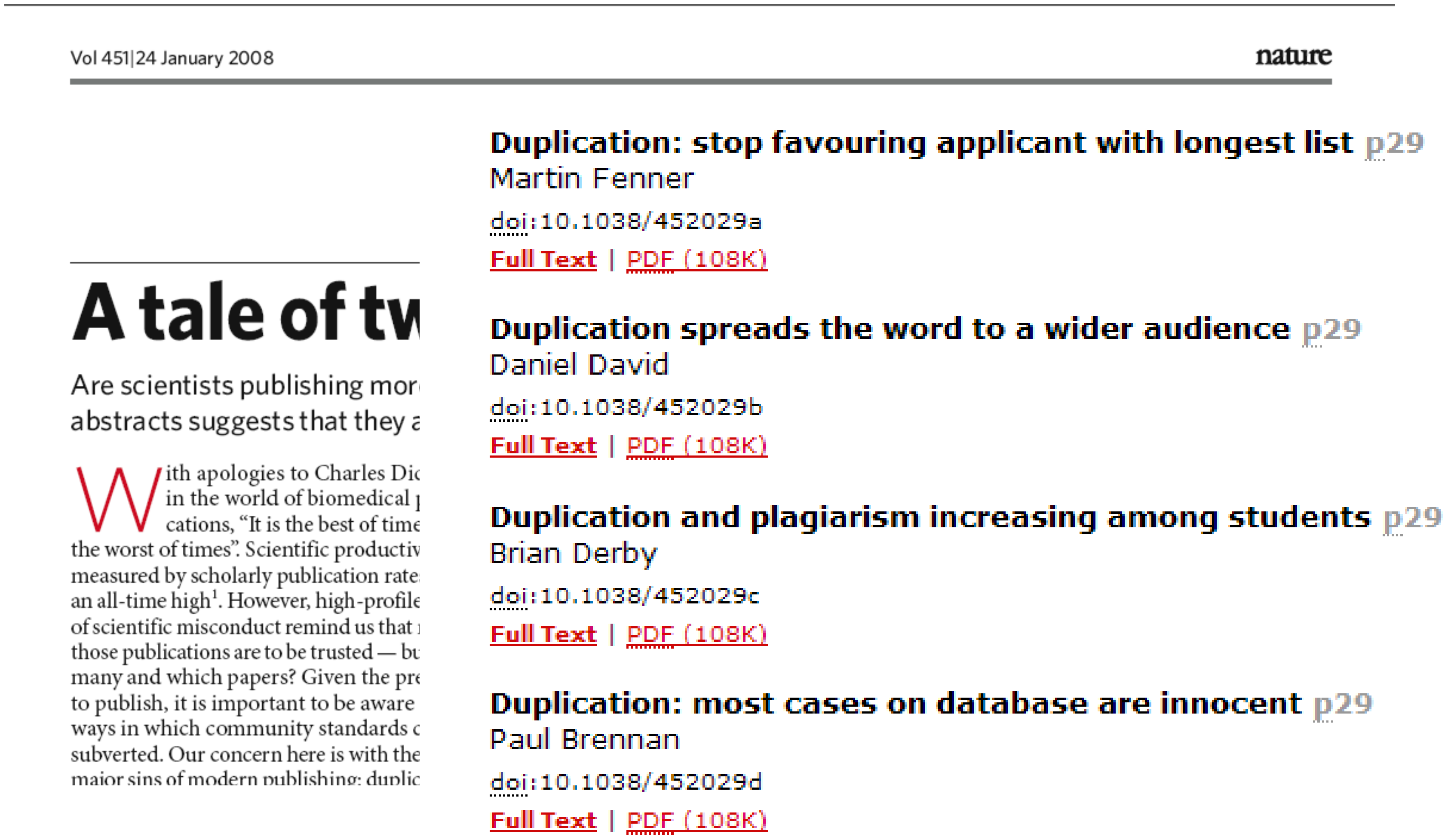
eTBLAST implementation

A sample of 62 213 Medline citations

1.35% with shared authors were sufficiently similar

0.04% of the citations with no shared authors were highly similar (potential plagiarism)

Pairwise comparison of Medline abstracts



Vol 451|24 January 2008 nature

A tale of tw

Are scientists publishing mor
abstracts suggests that they a

With apologies to Charles Dic
in the world of biomedical p
cations, "It is the best of time
the worst of times". Scientific productiv
measured by scholarly publication rate
an all-time high[1]. However, high-profile
of scientific misconduct remind us that
those publications are to be trusted — bu
many and which papers? Given the pre
to publish, it is important to be aware
ways in which community standards c
subverted. Our concern here is with the
major sins of modern publishing: duplic

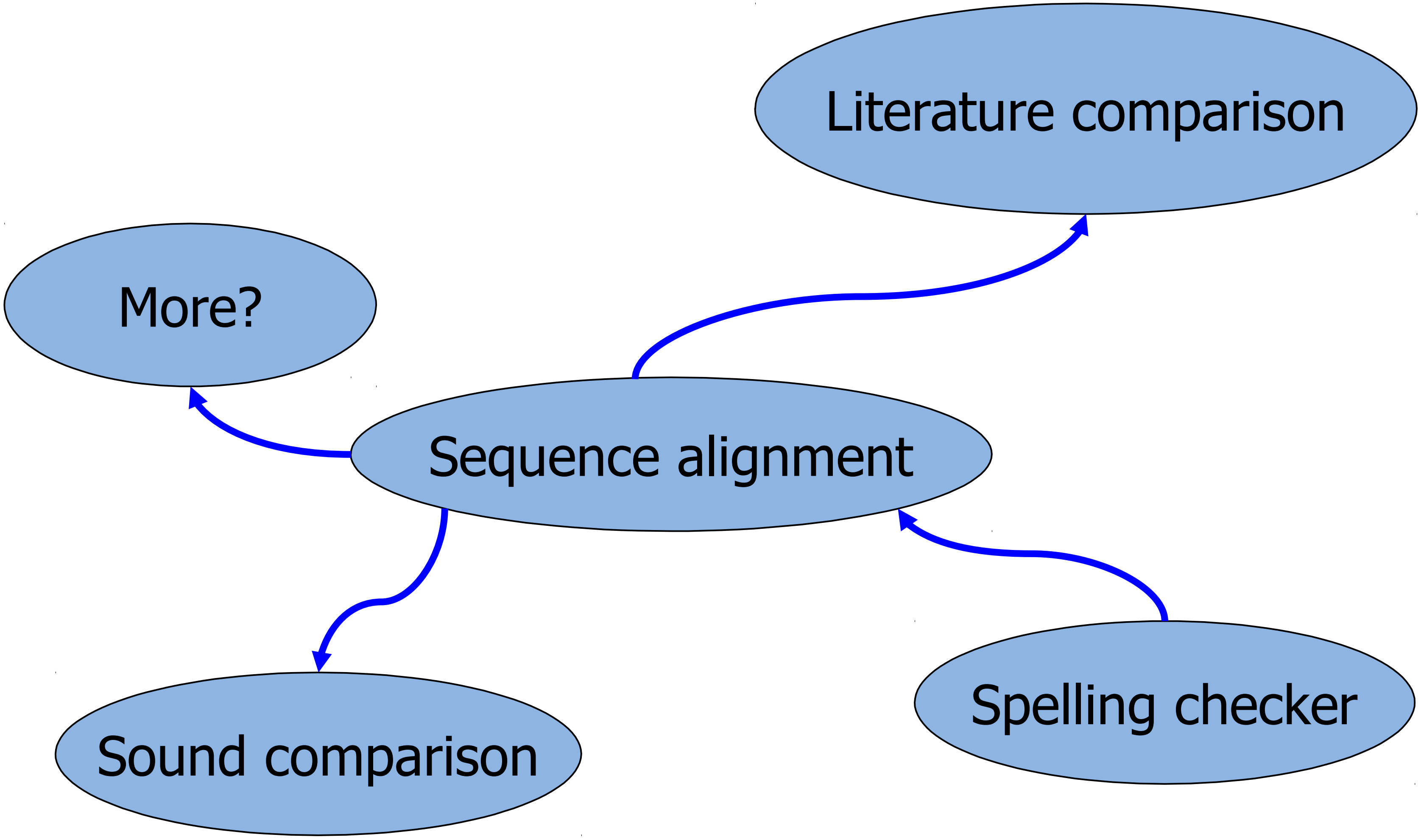
Duplication: stop favouring applicant with longest list p29
Martin Fenner
doi:10.1038/452029a
Full Text | PDF (108K)

Duplication spreads the word to a wider audience p29
Daniel David
doi:10.1038/452029b
Full Text | PDF (108K)

Duplication and plagiarism increasing among students p29
Brian Derby
doi:10.1038/452029c
Full Text | PDF (108K)

Duplication: most cases on database are innocent p29
Paul Brennan
doi:10.1038/452029d
Full Text | PDF (108K)

Other applications



Literature comparison

More?

Sequence alignment

Spelling checker

Sound comparison