Chapter IV: Basic Local Alignment Sequence Tool (BLAST)

Presentations use info from:

Jonathan Pevsner, Ph.D.
http://bioinfbook.org
pevsner@kennedykrieger.org
Bioinformatics and Functional Genomics
(3rd edition, ©2015 John Wiley & Sons, Ltd.)
You may use this PowerPoint for teaching purposes

dr. Stanislav Kolenčík
stanislav.kolencik@famnit.upr.si

What will you learn?

- perform BLAST searches at the NCBI website;
- understand how to vary optional BLAST search parameters;
- explain the three phases of a BLAST search (compile, scan/extend, trace-back);
- define the mathematical relationship between expect values and scores;
- outline strategies for BLAST searching.

Outline

BLAST

BLAST (Basic Local Alignment Search Tool) allows rapid sequence comparison of a query sequence against a database.

The **BLAST** algorithm is fast, accurate, and accessible both via the web and the command line.

Why use BLAST?

BLAST searching is fundamental to understanding the relatedness of any favorite query sequence to other known proteins or DNA sequences.

Applications include

- identifying orthologs and paralogs
- what proteins or genes are present in organism
- discovering new genes or proteins
- discovering variants of genes or proteins
- investigating expressed sequence tags (ESTs)
- exploring protein structure and function

The programs produce high-scoring segment pairs (HSPs) that represent local alignments between your query and database sequences.

BLASTP search at NCBI: overview of web-based search

4 components to performing any web-based BLAST search:

query: FASTA format or accession

database

algorithm

parameters

BLAST® » blastp suite

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

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

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

NP_000509.1

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Databases: Standard databases (nr etc.): New Experimental databases

Try experimental clustered nr database

For more info see What is clustered nr?

Compare: Select to compare standard and experimental database

Standard

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested exclude Add organism

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

Exclude Optional: Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Program Selection

Algorithm:

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

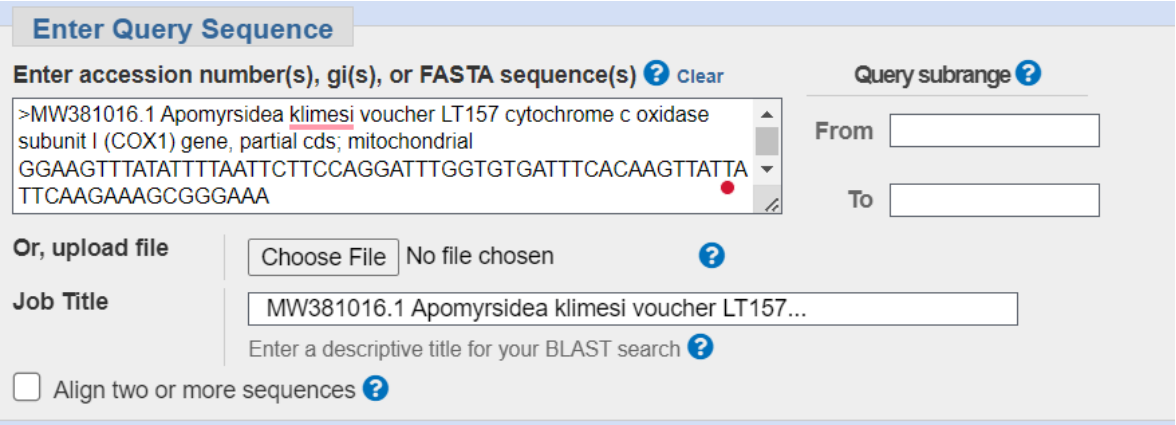
Search database nr using Blastp (protein-protein BLAST)

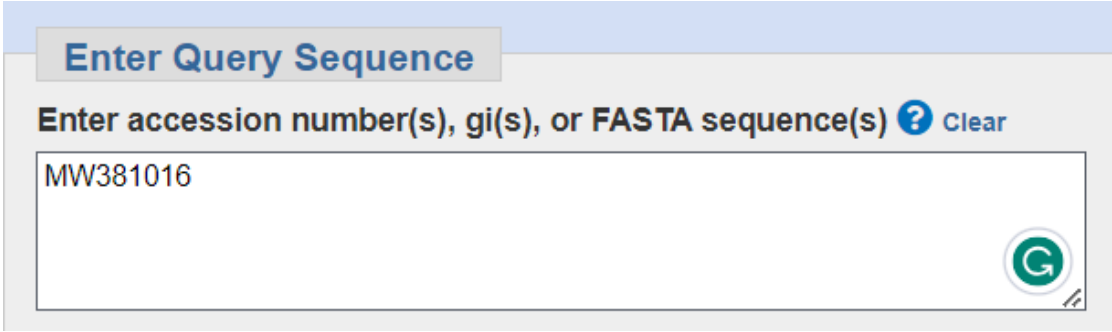
Show results in a new window

+ Algorithm parameters

Step 1: Choose your sequence

Sequence can be input in FASTA format or as accession number





BLAST step 2: Choose Program

Program	Query	Number of database searches	Database
BLASTP	protein	1	protein
BLASTN	DNA	1	DNA
BLASTX	DNA	6	protein
TBLASTN	protein	6	DNA
TBLASTX	DNA	36	DNA

Use BLASTP to compare a protein query to a database of proteins.

Use BLASTN to compare both strands of a DNA query against a DNA database.

BLASTX translates a DNA sequence into six protein sequences using all six possible reading frames, and then compares each of these proteins to a protein database.

TBLASTN is used to translate every DNA sequence in a database into six potential proteins, and then to compare your protein query against each of those translated proteins.

TBLASTX is the most computational[illegible] BLAST algorithm. It translates DNA from both a query and a database into six pot[illegible]n performs 36 protein-protein database searches.

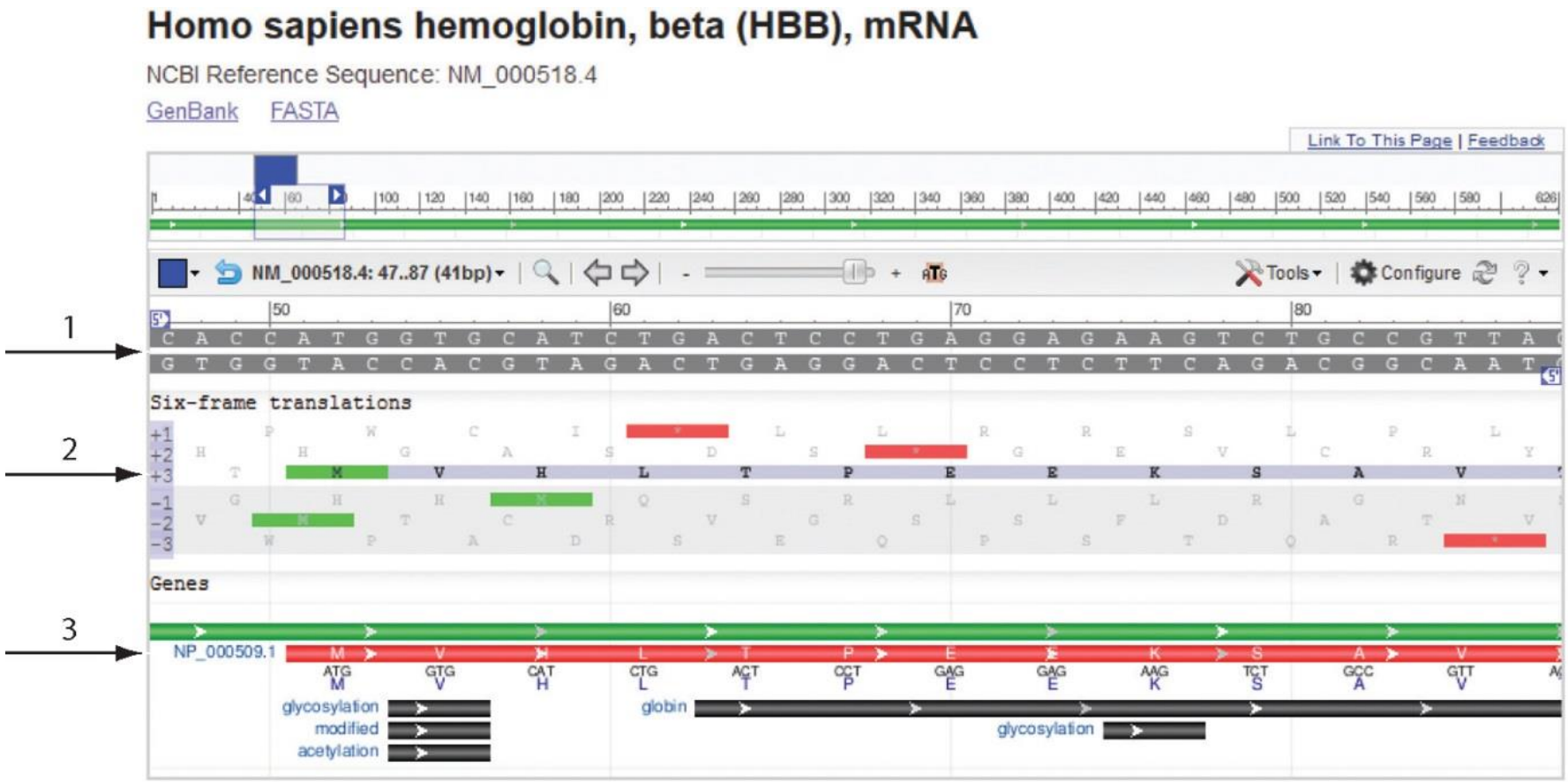
N refers to nucleotide
X refers to a DNA
P refers to protein
T refers to “translating”

Step 2 (choosing the BLAST program): DNA can be translated into six reading frames

DNA

3 forward,
3 reverse frames

protein



This image is from the NCBI Nucleotide entry for HBB

Double-stranded DNA -> Amino Acids -> Codons (3 nucleotides)

- The ribosome starts at a start codon and continues reading until it reaches a stop codon.

- The ribosome can start reading the DNA sequence at any point on the strand.

- Forward and Reverse x3 = 6 reading frames

DNA

3 forward,
3 reverse frames

protein

NCBI Nucleotide entry for HBB

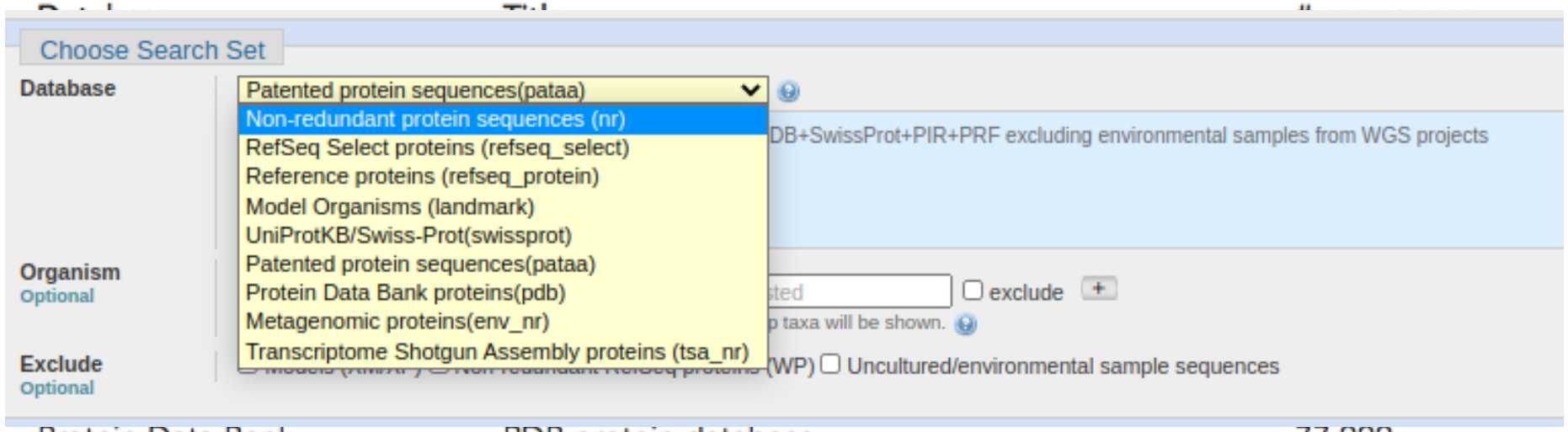
Step 3: choose a database to search (protein databases)

TABLE 4.1 Protein sequence databases that can be searched by BLAST searching at NCBI. PDB, Protein Data Bank. # indicates approximate number of sequences in database. Adapted from BLAST, NCBI, http://blast.ncbi.nlm.nih.gov/.

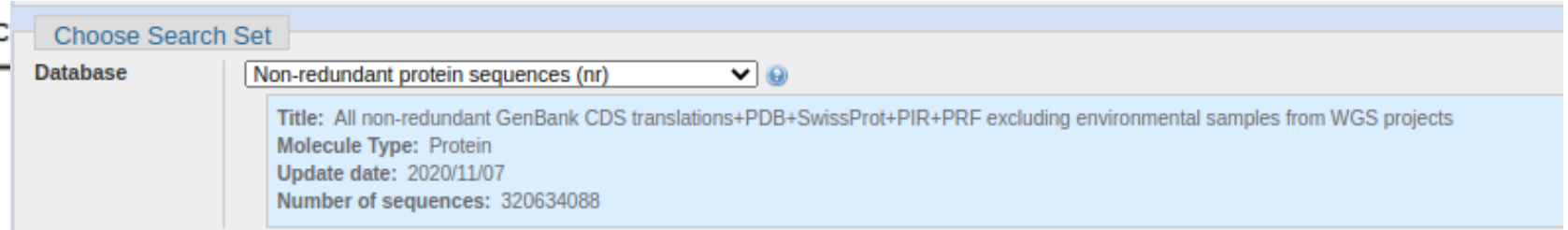
Database	Title	# sequences
nr	All nonredundant GenBank CDS translations + PDB + SwissProt + PIR + PRF excluding environmental samples from WGS projects	65 million
Reference proteins	NCBI protein reference sequences	50 million
UniProtKB/SwissProt	Nonredundant UniProtKB/SwissProt sequences	450,000
Patented protein sequences	Protein sequences derived from the Patent division of GenBank	1.3 million
Protein Data Bank	PDB protein database	77,000
Metagenomic proteins	Proteins from WGS metagenomic projects (env_nr)	6.5 million
Transcriptome	Transcriptome Shotgun Assembly (TSA) sequences	770,000

Step 3: choose a database to search (protein databases)

TABLE 4.1 Protein sequence databases that can be searched by BLAST searching at NCBI. PDB, Protein Data Bank. # indicates approximate number of sequences in database. Adapted from BLAST, NCBI, http://blast.ncbi.nlm.nih.gov/.



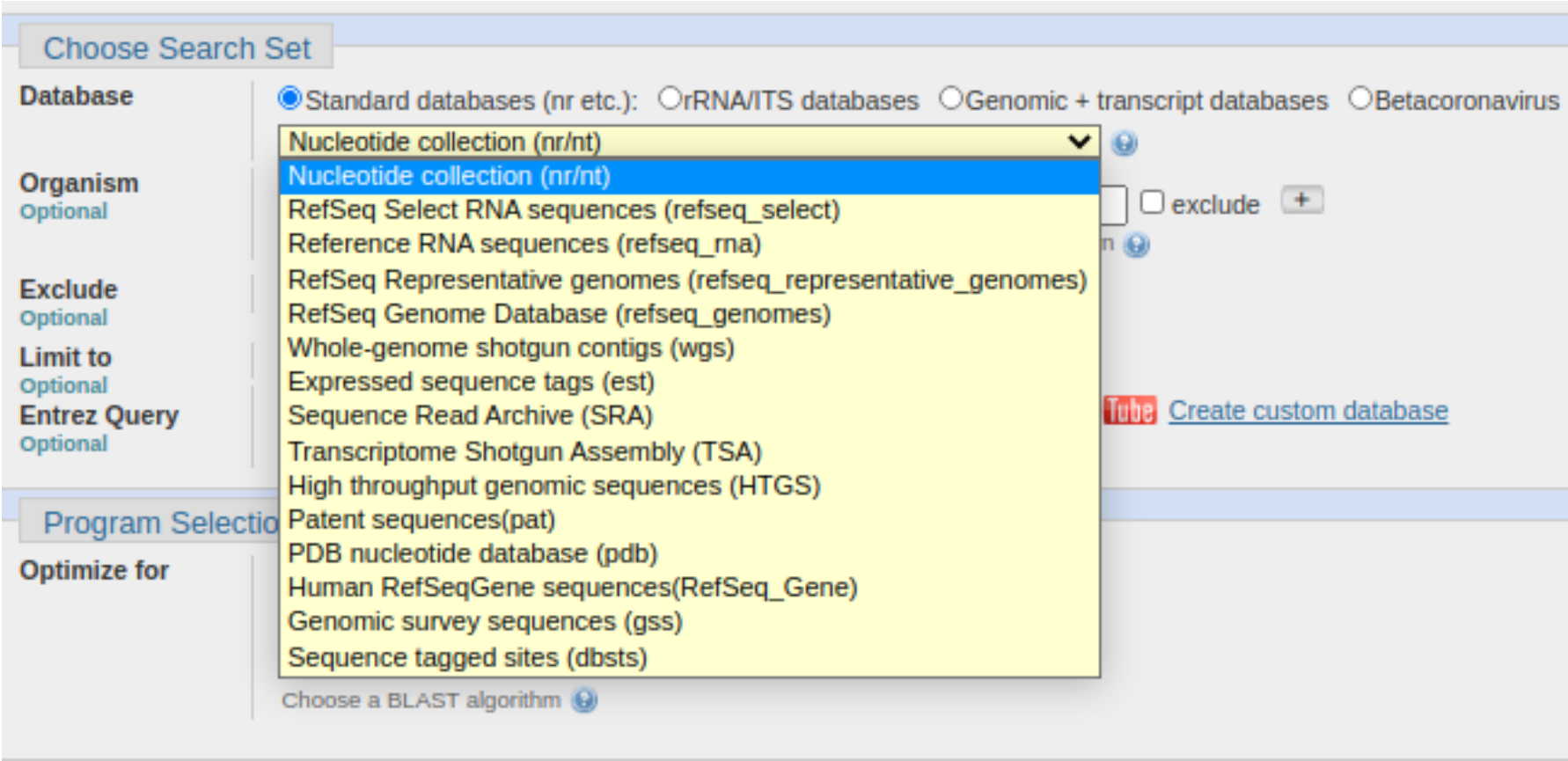
Protein Data Bank	PDB protein database	77,000
Metagenomic proteins	Proteins from WGS metagenomic projects (env_nr)	6.5 million
Transc		



Step 3: choose a database to search (nucleotide)

Database	Title	# sequences
Human Genomic + Transcript	Homo sapiens NCBI Annotation Release 104 RNAs; Homo sapiens all assemblies	55,000
Mouse Genomic + Transcript	Mus musculus NCBI Annotation RNAs; Mus musculus all assemblies	N/A
nr/nt	All GenBank+EMBL+DDBJ+PDB+RefSeq sequences, but excludes EST, STS, GSS, WGS, TSA, patent se[illegible]nces as well as [illegible]se 0, 1, and 2 [illegible]uences	25 million
refseq_rna	[illegible]anscrip[illegible]ences	3.5 million
refseq_genomic	[illegible]ces	2.7 million
NCBI Genomes	[illegible]s	28,000
Expressed sequence tags (EST)	[illegible]MBL+DDBJ [illegible]visions	75 million
Genomic survey sequences (gss)	[illegible] includes [illegible]n-trapped [illegible]ces	36 million
High-throughput genomic sequences (HTGS)	[illegible]ed hig[illegible]omic [illegible]ences; sequ[illegible]es 0,1 and 2	153,000
Patent sequences	[illegible]ucleotide sequenc[illegible]erived from the Patent division of GenBank	21 million
Protein Data Bank	PDB nucleotide database	8000
alu	Human *Alu* repeat elements	325
Sequence tagged sites (STS)	Database of GenBank+EMBL+DDBJ sequences from STS Divisions	1.3 million
Whole-genome shotgun (wgs)	Whole-genome-shotgun contigs	116 million
Transcriptome Shotgun Assembly (TSA)	Transcriptome shotgun assembly (TSA) sequences	15 million
16S ribosomal RNA sequences (Bacteria and Archaea)	16S ribosomal RNA sequences (bacteria and archaea)	7500

Step 3: choose a database to search (nucleotide)



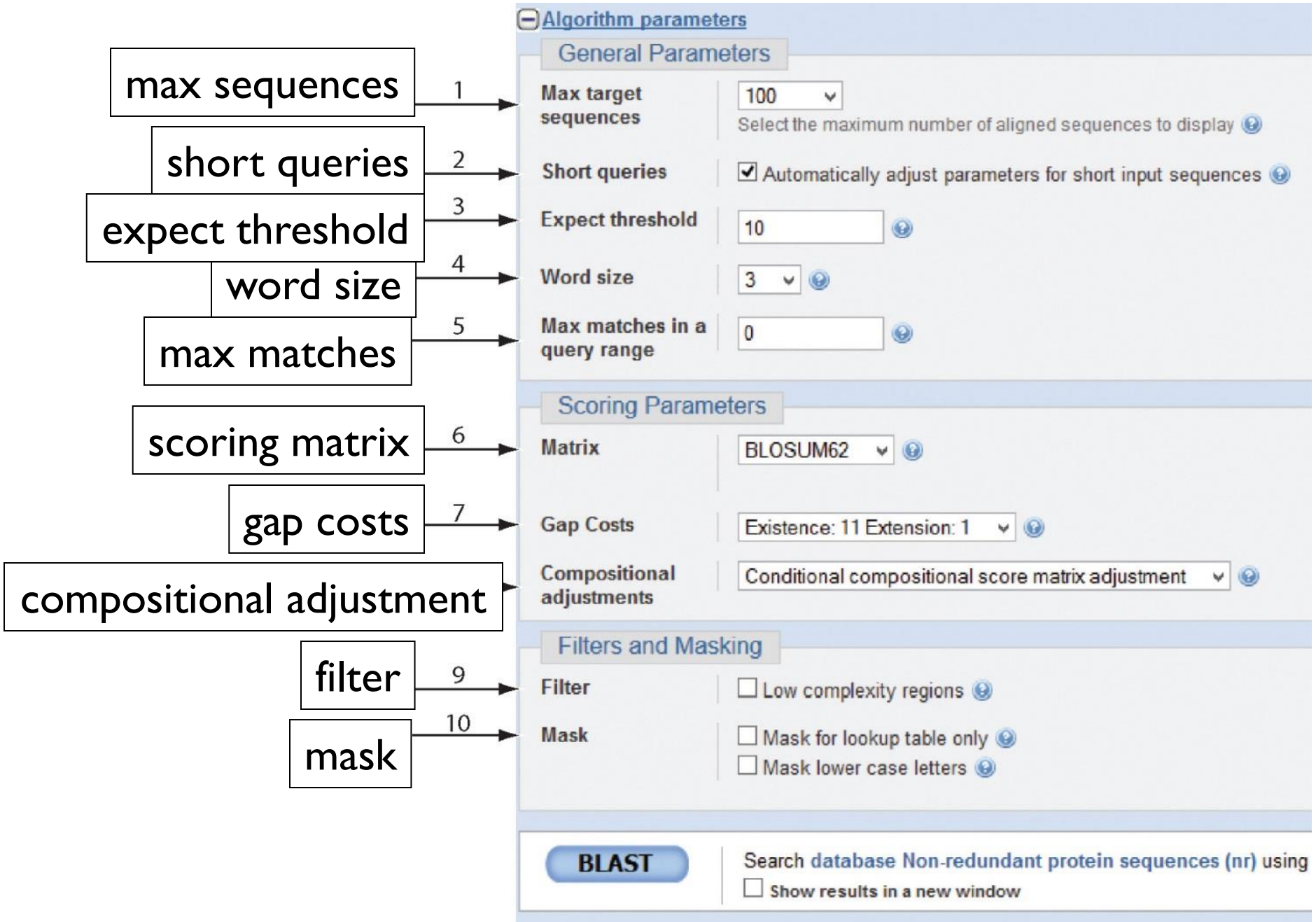
Step 4: optional parameters

You can...

- choose the organism to search
- turn filtering on/off
- change the substitution matrix
- change the expect (e) value
- change the word size

Example: BLASTP human insulin (NP_000198) against a *C. elegans* RefSeq database. Varying some parameters (filtering, compositional adjustments) can greatly affect the alignment itself.

Step 4a: choose optional BLASTP search parameters



Q: What is the Expect (E) value?

The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially as the Score (S) of the match increases.

E value describes the random background noise.
For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance.

Step 4a: compositional adjustment influences score, expect value search results

expect = 0.05

Default: conditional compositional score matrix adjustment

(a) Default: conditional compositional score matrix adjustment

Insulin-like peptide 3 [Drosophila melanogaster]

Sequence ID: ref|NP_648360.2| **Length:** 120 **Number of Matches:** 1

Range 1: 32 to 114 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
31.6 bits(70)	0.050	Compositional matrix adjust.	21/88(24%)	40/88(45%)	12/88(13%)

```
Query  29   HLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQ-  87
             LCG  L E L  +C    + +   T+R  + +   Q++  G      L+ L  + S+Q
Sbjct  32   KLCGRKLPETLSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQM  86

Query  88   ------KRGIVEQCCTSICSLYQLENYC  109
                  + G+ ++CC   C++ ++  YC
Sbjct  87   LKTRRLRDGVFDECCLKSCTMDEVLRYC  114
```

expect = 0.09

no adjustment

(b) No adjustment (by default, filter low complexity regions)

Insulin-like peptide 3 [Drosophila melanogaster]

Sequence ID: ref|NP_648360.2| **Length:** 120 **Number of Matches:** 1

Range 1: 33 to 114 GenPept Graphics

Score	Expect	Identities	Positives	Gaps
33.5 bits(75)	0.009	21/87(24%)	40/87(45%)	12/87(13%)

```
Query  30   LCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQ--  87
            LCG  L E L  +C    + +   T+R  + +   Q++  G      L+ L  + S+Q
Sbjct  33   LCGRKLPETLSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQML  87

Query  88   -----KRGIVEQCCTSICSLYQLENYC  109
                 + G+ ++CC   C++ ++  YC
Sbjct  88   KTRRLRDGVFDECCLKSCTMDEVLRYC  114
```

expect = 1e-04

composition-based statistics

(c) Composition-based statistics

Insulin-like peptide 3 [Drosophila melanogaster]

Sequence ID: ref|NP_648360.2| **Length:** 120 **Number of Matches:** 1

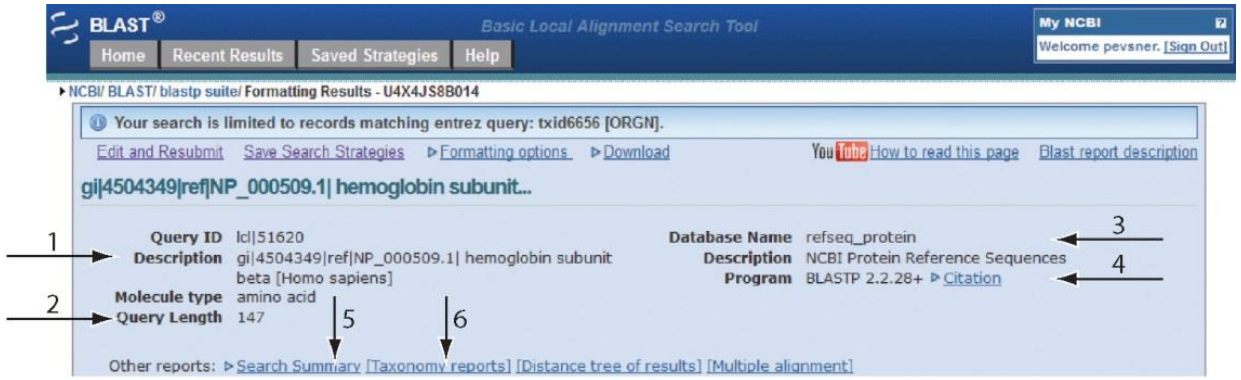
Range 1: 33 to 114 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
30.4 bits(67)	1e-04	Composition-based stats.	21/87(24%)	40/87(45%)	12/87(13%)

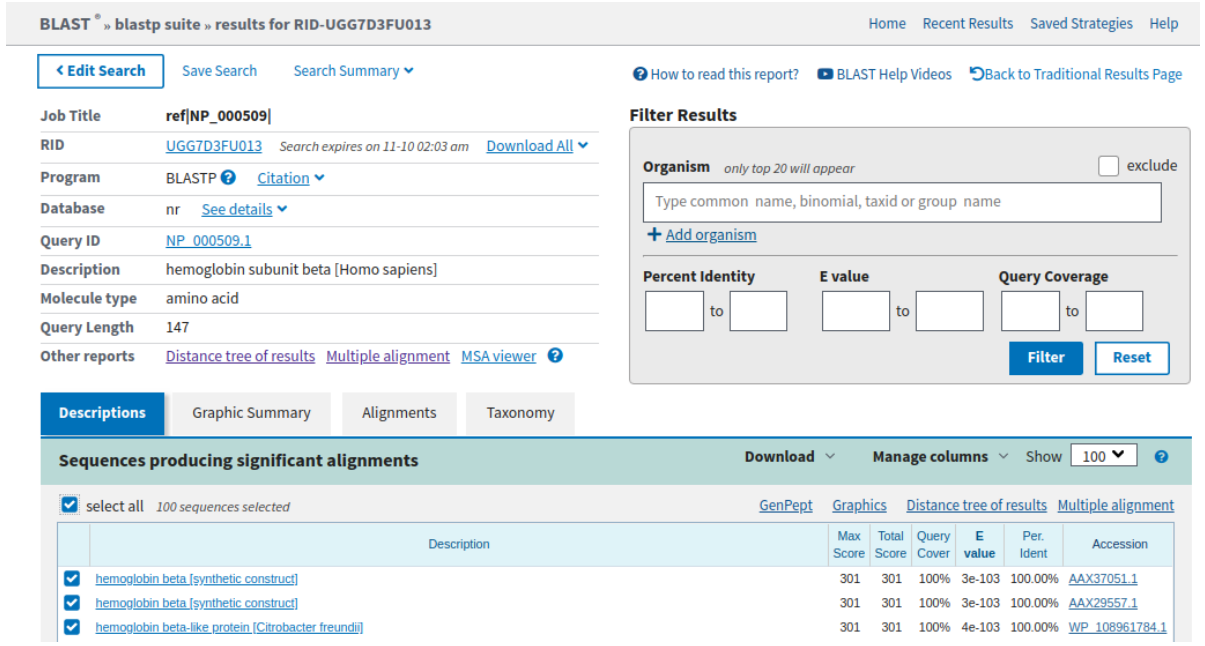
```
Query  30   LCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQ--  87
            LCG  L E L  +C    + +   T+R  + +   Q++  G      L+ L  + S+Q
Sbjct  33   LCGRKLPETLSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQML  87

Query  88   -----KRGIVEQCCTSICSLYQLENYC  109
                 + G+ ++CC   C++ ++  YC
Sbjct  88   KTRRLRDGVFDECCLKSCTMDEVLRYC  114
```

Step 4b: formatting options



The top of the BLAST output summarizes the query, database, and BLAST algorithm. Click to access a summary of the search parameters or a taxonomic report.



Step 4b: formatting options (you can view search parameters)

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

Database	
Posted date	Jun 12, 2013 10:46 AM
Number of letters	6,910,040,539 ← 4
Number of sequences	19,996,853
Entrez query	txid10090 [ORGN]

Karlin-Altschul statistics		
Lambda	0.320339	0.267
K	0.136843	0.041
H	0.422367	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

Expect value

BLOSUM62 matrix

Threshold value T

Size of database

BLAST » blastp suite » results for RID-UGG7D3FU013

Edit Search | Save Search | Search Summary

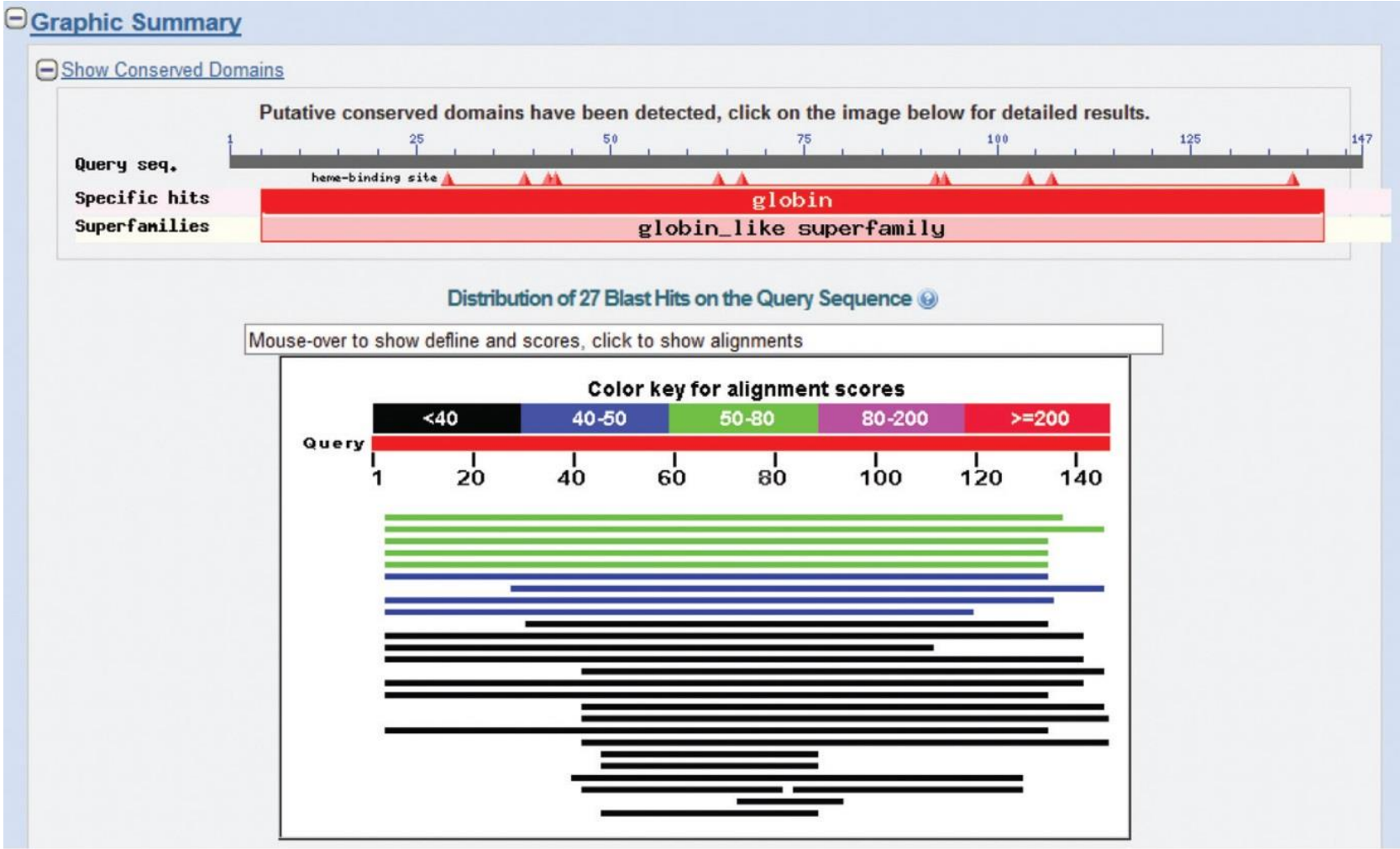
Search Parameters	
Program	blastp
Word size	6
Expect value	0.05
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	21
Composition-based stats	2

Database	
Posted date	Nov 5, 2020 12:42 AM
Number of letters	115,841,673,698
Number of sequences	320,634,088
Entrez query	None

Karlin-Altschul statistics		
Lambda	0.320339	0.267
K	0.136843	0.041
H	0.422367	0.14
Alpha	0.7916	1.9
Alpha_v	4.96466	42.6028
Sigma		43.6362

Results Statistics

Step 4b: formatting options



Graphic summary of the results shows the alignment scores (coded by color) and the length of the alignment (given by the length of the horizontal bars)

BLASTP output includes list of matches; links to the NCBI protein entry; bit score and E value; and download options

Sequences producing significant alignments:

Select: All None Selected:2

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
☑	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref|XP_003396833.1| PREDIC	59.7	59.7	91%	1e-10	29%	XP_003396832.1
☑	PREDICTED: cytoglobin-2-like isoform 1 [Bombus impatiens] >ref|XP_003494220.1| PREDI	58.5	58.5	97%	3e-10	28%	XP_003494219.1
☐	PREDICTED: globin-like [Megachile rotundata]	57.8	57.8	89%	6e-10	29%	XP_003707185.1
☐	PREDICTED: globin-like [Apis florea]	53.9	53.9	89%	1e-08	30%	XP_003690810.1
☐	globin 1 [Apis mellifera]	52.8	52.8	89%	4e-08	30%	NP_001071291.1
☐	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref|XP_003396831.1| PREDIC	45.1	45.1	89%	2e-05	26%	XP_003396830.1
☐	PREDICTED: neuroglobin-like, partial [Acyrthosiphon pisum]	42.4	42.4	80%	2e-04	23%	XP_001946608.2
☐	globin, putative [Ixodes scapularis]	42.7	42.7	90%	2e-04	25%	XP_002414906.1

BLAST output can be formatted to display multiple alignment

COBALT — Constraint-based Multiple Alignment Tool

Home | Recent Results | Help

My NCBI — Welcome pevsner. [Sign Out]

Phylogenetic Tree | Edit and Resubmit | Back to Blast Results | ▷Download

Multiple Alignment Results - gi|4504349|ref|NP_000509.1| hemoglobin subunit... - Cobalt RID U57PC4Y5211 (8 seqs)

▼ Descriptions ☑ Select All Re-align ▷Alignment parameters

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

Accession	Description	Links
☑ XP_003396832.1	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref|XP_003396833.1| PREDICTED: cytoglobin	G M
☑ XP_003494219.1	PREDICTED: cytoglobin-2-like isoform 1 [Bombus impatiens] >ref|XP_003494220.1| PREDICTED: cytoglobir	G M
☑ XP_003707185.1	PREDICTED: globin-like [Megachile rotundata]	G
☑ XP_003690810.1	PREDICTED: globin-like [Apis florea]	G
☑ NP_001071291.1	globin 1 [Apis mellifera] >emb|CAJ43389.1| globin 1 [Apis mellifera] >emb|CAJ43388.1| globin 1 [Apis mellife	U G M
☑ XP_003396830.1	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref|XP_003396831.1| PREDICTED: cytoglobin	G M
☑ XP_001946608.2	PREDICTED: neuroglobin-like, partial [Acyrthosiphon pisum]	G M
☑ XP_002414906.1	globin, putative [Ixodes scapularis] >gb|EEC18571.1| globin, putative [Ixodes scapularis]	G

▼ Alignments ☑ Select All Re-align Mouse over the sequence identifer for sequence title

View Format: Compact Conservation Setting: 2 Bits

```
XP_003396832  1   MGTFLRFFGFSSSDDNRIDEATGLTEKQKKLVQNTWAVIRKDEVASGIAVMTTFFKTYPEYQRYFSAFADVPFDELPANK  80
XP_003494219  1   MGTFLRFFGISSSDDNRIDEATGLTEKQKKLVQNTWAVIRKDEVASGIAVMTTFFKTYPEYQRYFSAFADVPFDELPANK  80
XP_003707185  1   MDSFLRLLGISS-DDNRIDQATGLTEKQKKLVQNTWSIIRKDEVGAGVLVMCAFFKKYPSYVQYFEAFKDIPLDQLPDNK  79
XP_003690810  1   MGTFLRFLGISSSDDNRIDQATGLTERQKKLVQNTWAVVRKDEVASGIAVMTAFFKKYPEYQRYFTAFMDTPLNELPANK  80
NP_001071291  1   MGTFLRFLGISSSDDNRIDQATGLTERQKKLVQNTWAVVRKDEVASGIAVMTAFFKKYPEYQRYFTAFMDTPLNELPANK  80
XP_003396830  1   MGSVLTYF-LGNPDDDVVDPKLGLTNKEKRIIRETWGVLRANSVKVGVDIMISYFKRFPQHHRAFPPFKDIPADDLLDNK  79
XP_001946608  1   ----------------------------------SCDLTR----FIFPLFLYRLFEEHQELLQLFTKFGELKTRDAQANS  42
XP_002414906  1   MSW---LFGSAS--ADMPSTKTGLTTSDKCAIKDTWTMFRRETRTNALSLFVALFSRYPEYQKMFPNFADVALKDMMQCP  75
```

For BLASTN, CDS output displays amino acids above DNA sequence of query and subject

Download GenBank Graphics ▼ Nex

Homo sapiens hemoglobin, epsilon 1 (HBE1), mRNA

Sequence ID: ref|NM_005330.3| Length: 816 Number of Matches: 1

Range 1: 203 to 705 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
410 bits(454)	5e-113	393/503(78%)	3/503(0%)	Plus/Plus

```
CDS:hemoglobin subun  1                                                  M  V  H
Query                 3    ATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAA---CAGACACCATGGTGCAT  59
                           || |||||| |||||| |||   |||||||||| |||  |   | | || |||||||||||
Sbjct                 203  ATCTGCTTCCGACACAGCTGCAATCACTAGCAAGCTCTCAGGCCTGGCATCATGGTGCAT  262
CDS:hemoglobin subun  1                                                  M  V  H

CDS:hemoglobin subun  4     L  T  P  E  E  K  S  A  V  T  A  L  W  G  K  V  N  V  D  E
Query                 60   CTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAA  119
                            | ||| |||||||||||| ||||||| |||  ||||||| ||||| |||| ||||| ||
Sbjct                 263  TTTACTGCTGAGGAGAAGGCTGCCGTCACTAGCCTGTGGAGCAAGATGAATGTGGAAGAG  322
CDS:hemoglobin subun  4     F  T  A  E  E  K  A  A  V  T  S  L  W  S  K  M  N  V  E  E

CDS:hemoglobin subun  24    V  G  G  E  A  L  G  R  L  L  V  V  Y  P  W  T  Q  R  F  F
Query                 120  GTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTT  179
                            | ||| ||||| ||| ||||||| || || || || ||||| |||||||||||| || |||
Sbjct                 323  GCTGGAGGTGAAGCCTTGGGCAGACTCCTCGTTGTTTACCCCTGGACCCAGAGATTTTT  382
CDS:hemoglobin subun  24    A  G  G  E  A  L  G  R  L  L  V  V  Y  P  W  T  Q  R  F  F

CDS:hemoglobin subun  44    E  S  F  G  D  L  S  T  P  D  A  V  M  G  N  P  K  V  K  A
Query                 180  GAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCT  239
                            ||   ||||||  | |||||  ||||   |||  |  |||||||||| ||||| |||||
Sbjct                 383  GACAGCTTTGGAAACCTGTCGTCTCCCTCTGCCATCCTGGGCAACCCCAAGGTCAAGGCC  442
CDS:hemoglobin subun  44    D  S  F  G  N  L  S  S  P  S  A  I  L  G  N  P  K  V  K  A

CDS:hemoglobin subun  64    H  G  K  K  V  L  G  A  F  S  D  G  L  A  H  L  D  N  L  K
Query                 240  CATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAG  299
                            |||||||||||| |||||   | ||||| | ||||  |     || |||||||||||||
Sbjct                 443  CATGGCAAGAAGGTGCTGACTTCCTTTGGAGATGCTATTAAAAACATGGACAACCTCAAG  502
CDS:hemoglobin subun  64    H  G  K  K  V  L  T  S  F  G  D  A  I  K  N  M  D  N  L  K

CDS:hemoglobin subun  84    G  T  F  A  T  L  S  E  L  H  C  D  K  L  H  V  D  P  E  N
Query                 300  GGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAAC  359
                             | ||||||| |  |||||||||||||||||||||||||||||| |||||||||||||||
Sbjct                 503  CCCGCCTTTGCTAAGCTGAGTGAGCTGCACTGTGACAAGCTGCATGTGGATCCTGAGAAC  562
CDS:hemoglobin subun  84    P  A  F  A  K  L  S  E  L  H  C  D  K  L  H  V  D  P  E  N

CDS:hemoglobin subun  104   F  R  L  L  G  N  V  L  V  C  V  L  A  H  H  F  G  K  E  F
Query                 360  TTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTC  419
                            |||| ||||||||| |||||| ||||   | | |||||   ||||||||||| || |||
Sbjct                 563  TTCAAGCTCCTGGGTAACGTGATGGTGATTATTCTGGCTACTCACTTTGGCAAGGAGTTC  622
CDS:hemoglobin subun  104   F  K  L  L  G  N  V  M  V  I  I  L  A  T  H  F  G  K  E  F

CDS:hemoglobin subun  124   T  P  P  V  Q  A  A  Y  Q  K  V  V  A  G  V  A  N  A  L  A
Query                 420  ACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCC  479
                            |||||   |||||||||||||  |||||  ||||| ||| ||| || | ||||||||||
Sbjct                 623  ACCCCTGAAGTGCAGGCTGCCTGGCAGAAGCTGGTGTCTGCTGTCGCCATTGCCCTGGCC  682
CDS:hemoglobin subun  124   T  P  E  V  Q  A  A  W  Q  K  L  V  S  A  V  A  I  A  L  A

CDS:hemoglobin subun  144   H  K  Y  H
Query                 480  CACAAGTATCACTAAGCTCGCTT  502
                            || ||||| |||| || || |||
Sbjct                 683  CATAAGTACCACTGAGTTCTCTT  705
CDS:hemoglobin subun  144   H  K  Y  H
```

Command-line BLAST+

Visit the BLAST site at NCBI ("**help**" tab) to find the URL for the BLAST+ download.

Three steps:

(1) Obtain a protein database (we'll use a perl script included in the BLAST+ installation);
(2) Obtain a query protein (we'll use EDirect);
(3) Perform the search

(You will explore command-line workflow during exercises)

Command-line BLAST+ (Step 1: obtain a database)

Visit the BLAST site at NCBI ("help" tab) to find the URL for the BLAST+ download.

```
$ mkdir database # this creates a new directory
$ cd database/ # we navigate into that directory
# Enter the following, without arguments, to see a help document.
$ update_blastdb.pl
# Next get a list of all available databases
$ update_blastdb.pl --showall
$ update_blastdb.pl --showall | less
```

```
$ update_blastdb.pl refseq_protein
```

```
$ tar -zxvf refseq_protein.00.tar.gz
```

You will also need to install the EDirect command-line utility if you wish to look up TAXIDs via the command-line script

(You will explore command-line workflow during exercises)

Command-line BLAST+ (Step 2: obtain a query protein)

Use EDirect to obtain a globin protein.

```
$ esearch -db protein -query "NP_000509" | efetch -format fasta > hbb.
txt
$ cat hbb.txt # cat is the concatenate utility that we use to print the
# file
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

Command-line BLAST+ (Step 3: perform a search!)

Do the search:

```
$ blastp --h # Get help
$ blastp -query hbb.txt -db ./database/refseq_protein -out mysearch1
# Note that we use ./ to specify the directory location of the
# executable which is within the executable directory
```

View the results:

```
$ less mysearch1
```

Try repeating the search,
e.g. changing the database size:

```
$ blastp -query hbb.txt -db ./database/refseq_protein -dbsize 9750000 -out
mysearch2
```

(You will explore command-line workflow during exercises)

How a BLAST search works

"The central idea of the BLAST algorithm is to confine attention to segment pairs that contain a word pair of length w with a score of at least T."

Altschul et al. (1990)

How the original BLAST algorithm works: three phases

Phase 1: compile a list of word pairs (w=3) above threshold T

Example: for a human RBP query
…FSGTWYA… (query word is in green)

A list of words (w=3) is:

```
FSG SGT GTW TWY WYA
YSG TGT ATW SWY WFA
FTG SVT GSW TWF WYS
...
```

Phase 1: compile a list of words (w=3)

neighborhood word hits > threshold

(T=11)

neighborhood word hits < below threshold

GTW	6,5,11	22
GSW	6,1,11	18
ATW	0,5,11	16
NTW	0,5,11	16
GTY	6,5,2	13
GNW		10
GAW		9

A threshold value **T** is established for the score of aligned words.

Phase 1: Setup: compile a list of words (w=3) above threshold T

- Query sequence: human beta globin NP_000509.1 (includes ...VTALWGKVNVD...). This sequence is read; low complexity or other filtering is applied; a "lookup" table is built.

- Words derived from query sequence (HBB): VTA TAL ALW LWG WGK GKV KVN VNV NVD

- Generate a list of words matching query (both above and below T). Consider LWG in the query and the scores (derived from a BLOSUM62 matrix) for various words.

- Generate similar lists of words spanning the query (e.g. words for WGW, GWG, WGK...).

examples of words >= threshold 12

LWG	4+11+6=21
IWG	2+11+6=19
MWG	2+11+6=19
VWG	1+11+6=18
FWG	0+11+6=17
AWG	0+11+6=17
LWS	4+11+0=15
LWN	4+11+0=15
LWA	4+11+0=15
LYG	4+ 2+6=12

threshold →

examples of words below threshold

LFG	4+ 1+6=11
FWS	0+11+0=11
AWS	-1+11+0=10
CWS	-1+11+0=10
IWC	2+11-3=10

Phase 2: scan the database for matches and extend

Phase 2: Scanning and extensions

- Select all the words above threshold T (LWG, IWG, MWG, VWG, FWG, AWG, LWS, LWN, LWA, LYG)
- Scan the database for entries ("hits") that match the compiled list
- Create a hash table index with the locations of all the hits for each word
- Perform gap free extensions
- Perform gapped extensions

```
LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV HBB
L+P +K+ V A WGKV  +  E G EAL R+ + +P T+ +F  F      D   G+ +V
LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV HBA
```

extension

extension

word pair from
first phases of search
"hits" alpha globin,
triggers extension

Phase 3: Traceback to generate gapped alignment

Phase 3: Traceback

- Calculate locations of insertions, deletions, and matches (for alignments saved in Phase 2)
- Apply composition-based statistics (for BLASTP, TBLASTN)
- Generate gapped alignment

How a BLAST search works: threshold

You can locally install BLAST and modify the threshold parameter.

The default value for BLASTP is 11.

To change it, enter "-f 16" or "-f 5" in the advanced options of BLAST+.

For **BLASTN**, the word size is typically **7, 11, or 15** (EXACT match). Changing word size is like changing threshold of proteins. w=15 gives fewer matches and is faster than w=11 or w=7.

For **megaBLAST** the word size is **28** and can be adjusted to **64**. What will this do? MegaBLAST is VERY fast for finding closely related DNA sequences!

How to interpret a BLAST search: expect value

The **expect value *E*** is the number of alignments with scores greater than or equal to **score *S*** that are expected to occur by chance in a database search.

An ***E*** value is related to a **probability value *p***.

The key equation describing an *E* value is:

$$E = Kmn\, e^{-lS}$$

$$E = Kmn\, e^{-\lambda S}$$

This equation is derived from a description of the extreme value distribution

S = the score

E = the expect value = the number of high-scoring segment pairs (HSPs) expected to occur with a score of at least *S*

m*, *n = the length of two sequences

l, *K* = Karlin Altschul statistics

The E value depends on the score **and on λ, which is a parameter that scales the scoring system**. E also depends on the length of the query sequence and the length of the database. **The parameter K is a scaling factor for the search space**. The parameters K and λ are described by Karlin and Altschul (1990), and are often referred to as Karlin–Altschul statistics.

From raw scores to bit scores

- There are two kinds of scores: raw scores (calculated from a substitution matrix) and bit scores (normalized scores)

- Bit scores are comparable between different searches because they are normalized to account for the use of different scoring matrices and different database sizes

$S' = \text{bit score} = (\lambda S - \ln K) / \ln 2$

The E value corresponding to a given bit score is:
$E = mn\, 2^{-S'}$

Bit scores allow you to compare results between different database searches, even using different scoring matrices.

How to interpret BLAST: *E* values and *p* values

The **expect value *E*** is the number of alignments with scores greater than or equal to **score *S*** that are expected to occur by chance in a database search. A *p* value is a different way of representing the significance of an alignment.

$$p = 1 - e^{-E}$$

How to interpret BLAST: *E* values and *p* values

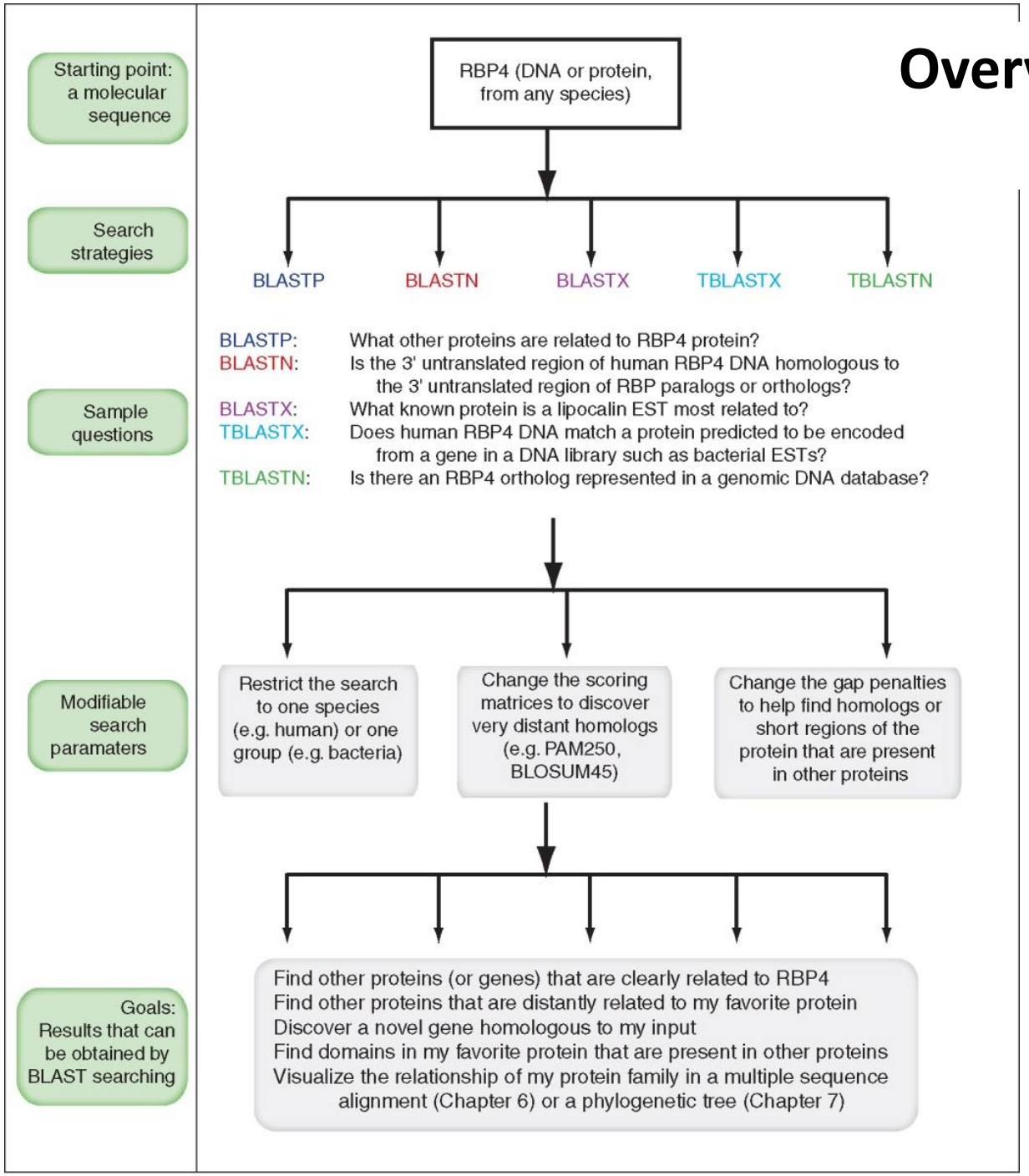
E values of about 1 to 10 are far easier to interpret than corresponding *p* values.

Very small *E* values are very similar to *p* values.

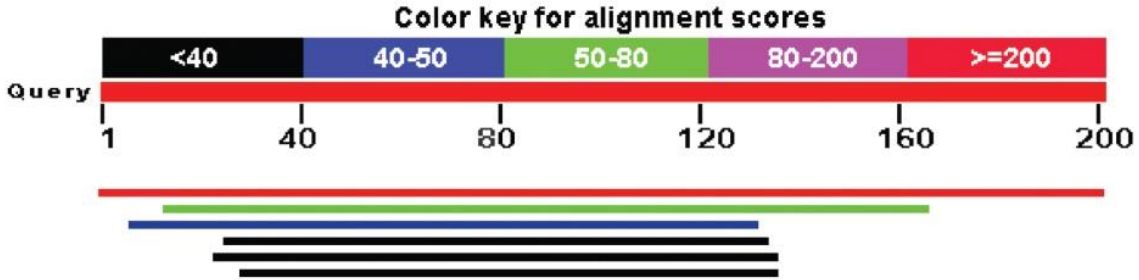
E	*p*
10	0.99995460
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258 (about 0.1)
0.05	0.04877058 (about 0.05)
0.001	0.00099950 (about 0.001)
0.0001	0.00010000

E values are comparable to *p* values and are designed to be more convenient to interpret.

Overview of BLAST search strategies



(a) Graphical overview



BLASTP search: human RBP4 query, human RefSeq database

(b) List of alignments

Sequences producing significant alignments:

Select: All None Selected:6

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Max ident	Accession
retinol-binding protein 4 precursor [Homo sapiens]	420	420	100%	1e-150	100%	NP_006735.2
apolipoprotein D precursor [Homo sapiens]	55.5	55.5	76%	1e-09	28%	NP_001638.1
glycodelin precursor [Homo sapiens] >ref|NP_002562.2| glycodelin precursor [Homo s	40.0	40.0	62%	5e-04	26%	NP_001018059.1
protein AMBP preproprotein [Homo sapiens]	35.0	35.0	54%	0.034	23%	NP_001624.1
complement component C8 gamma chain precursor [Homo sapiens]	32.3	32.3	56%	0.18	25%	NP_000597.2
lipocalin-15 precursor [Homo sapiens]	28.5	28.5	53%	3.4	23%	NP_976222.1

Results include matches (such as CG8) with high E values and limited identity to the query

(c) Pairwise alignment of RBP4 and C8G

complement component C8 gamma chain precursor [Homo sapiens]

Sequence ID: ref|NP_000597.2| Length: 202 Number of Matches: 1

Range 1: 33 to 139 GenPept Graphics ▼ Next Match ▲ Previous Match

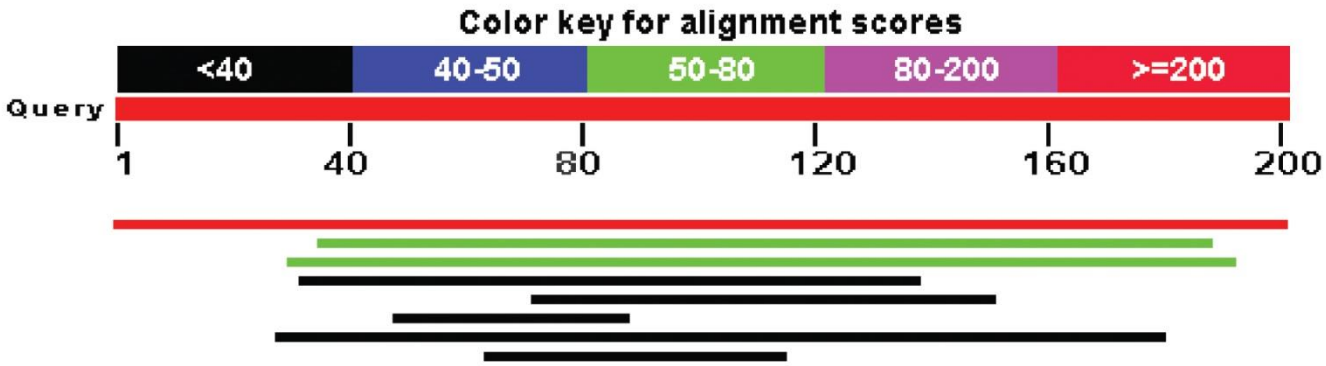
Score	Expect	Method	Identities	Positives	Gaps
32.3 bits(72)	0.18	Compositional matrix adjust.	28/114(25%)	49/114(42%)	8/114(7%)

```
Query  24   VSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETG-QMSATAKGRVRLL  82
            +S+ + K NFD  +F+GTW  +A         +    AE +      Q +A A    R L
Sbjct  33   ISTIQPKANFDAQQFAGTWLLVAVGSACRFLQEQGHRAEATTLHVAPQGTAMAVSTFRKL  92

Query  83   NNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQY  136
            +   +C  +   + DT    +F ++         +G     + +TDY ++AV Y
Sbjct  93   DG--ICWQVRQLYGDTGVLGRFLLQARDA-----RGAVHVVVAETDYQSFAVLY  139
```

“Reciprocal” BLASTP search with CG8 as query includes RBP4 and other lipocalins

(a) Graphical overview



(b) List of alignments

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Max ident	Accession
complement component C8 gamma chain precursor [Homo sapiens]	412	412	100%	3e-147	100%	NP_000597.2
lipocalin-15 precursor [Homo sapiens]	69.7	69.7	76%	1e-14	34%	NP_976222.1
protein AMBP preproprotein [Homo sapiens]	68.9	68.9	80%	1e-13	25%	NP_001624.1
retinol-binding protein 4 precursor [Homo sapiens]	33.1	33.1	52%	0.12	25%	NP_006735.2
tenascin-X isoform 1 precursor [Homo sapiens] ← Not homologous	30.0	30.0	39%	1.5	31%	NP_061978.6
neuroblastoma-amplified sequence [Homo sapiens] ← Not homologous	29.6	29.6	20%	2.1	44%	NP_056993.2
neutrophil gelatinase-associated lipocalin precursor [Homo sapiens]	28.9	28.9	75%	2.9	21%	NP_005555.2
HBS1-like protein isoform 1 [Homo sapiens] ← Not homologous	28.5	28.5	25%	5.4	33%	NP_006611.1

This confirms that the finding of CG8 using RBP4 as a query was a true positive

BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 3 2008, pages 319–324
doi:10.1093/bioinformatics/btm585

Sequence analysis

Choosing BLAST options for better detection of orthologs as reciprocal best hits

Gabriel Moreno-Hagelsieb* and Kristen Latimer
Department of Biology, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON, Canada, N2L 3C5

Received on August 29, 2007; revised on October 21, 2007; accepted on November 19, 2007
Advance Access publication November 26, 2007
Associate Editor: John Quackenbush

ABSTRACT

Motivation: The analyses of the increasing number of genome sequences requires shortcuts for the detection of orthologs, such as Reciprocal Best Hits (RBH), where orthologs are assumed if two genes each in a different genome find each other as the best hit in the other genome. Two BLAST options seem to affect alignment scores the most, and thus the choice of a best hit: the filtering of low information sequence segments and the algorithm used to produce the final alignment. Thus, we decided to test whether such options would help better detect orthologs.

Results: Using *Escherichia coli* K12 as an example, we compared the number and quality of orthologs detected as RBH. We tested four different conditions derived from two options: filtering of low-information segments, hard (default) versus soft; and alignment algorithm, default (based on matching words) versus Smith–Waterman. All options resulted in significant differences in the number of orthologs detected, with the highest numbers obtained with the combination of soft filtering with Smith–Waterman alignments. We compared these results with those of Reciprocal Shortest Distances (RSD), supposed to be superior to RBH because it uses an evolutionary measure of distance, rather than BLAST statistics, to rank homologs and thus detect orthologs. RSD barely increased the number of orthologs detected over those found with RBH. Error estimates, based on analyses of conservation of gene order, found small differences in the quality of orthologs detected using RBH. However, RSD showed the highest error rates. Thus, RSD have no advantages over RBH.

Availability: Orthologs detected as Reciprocal Best Hits using soft masking and Smith–Waterman alignments can be downloaded from http://popolvuh.wlu.ca/Orthologs.

Contact: gmoreno@wlu.ca

computational biologists often need their own orthologous sets for variable reasons such as: (1) a newly sequenced genome that needs annotation; (2) a need for updated mappings not available in published ortholog databases; (3) the lack of agreement about the genome annotations to use, for instance, those provided by the authors of a genome, corrections such as those within the RefSeq database (Maglott *et al.*, 2000; Pruitt *et al.*, 2005), the Genome Reviews (http://www.ebi.ac.uk/GenomeReviews/), the HAMAP project (Boeckmann *et al.*, 2003; Gattiker *et al.*, 2003) or even those re-annotations produced by particular research groups (Besemer *et al.*, 2001).

Orthologs are defined as genes that have diverged after a speciation event (Fitch, 2000). Another way to define them might be as the 'same genes' in different organisms. This evolutionary relationship implies that products of orthologous genes should tend to keep their original functions. Paralogs, on the other hand, are defined as genes that have diverged after a duplication event (Fitch, 2000). These have been proposed as a source of functional innovation (Francino, 2005; Ohno, 1970) and are less expected to have similar functions. It is therefore very important to be able to differentiate between orthologs and extra-paralogs, paralogous genes residing in different organisms (Janga and Moreno-Hagelsieb, 2004).

The definitions above are based on the event separating the histories of the homologous genes in question. In practice, one has to rely on sequence similarity and suitable statistics to detect homologs. Once putative homologs have been detected, evolutionary models such as phylogenetic trees, would be too computationally intensive to run for orthology detection, especially given the growth rate of genome sequence databases. Thus, most research in comparative genomics relies on some sort of shortcut, or working definition, to detect orthology.

Essentially, if gene a in genome A finds gene b as its best, highest scoring, match in genome B; and gene b finds gene a as its best match in genome A, they are RBH and thus inferred to be orthologs.

Sequence analysis

Progress in quickly finding orthologs as reciprocal best hits

Julie Hernández-Salmerón and Gabriel Moreno-Hagelsieb*

Department of Biology, Wilfrid Laurier University, 75 University Ave W, Waterloo, Ontario, Canada N2L 3C5.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: While the authors of software for the quick comparison of protein sequences evaluate the speed of their software and compare their results against the most usual software for the task, it is not common for them to evaluate their software for more particular uses, such as finding orthologs as reciprocal best hits (RBH). Here we focused on comparing RBH results using software that runs faster than blastp. Namely, lastal, diamond, and MMseqs2.

Results: Of the programs tested, lastal required the least time to produce results. However, it produced fewer results than any other program when comparing evolutionarily distant genomes. The program producing the most similar number of RBH as blastp was MMseqs2. This program was also resulted in the lowest error estimates compared to any of the programs used. The results with diamond were very close to those obtained with MMseqs2, with diamond running much faster. Our results suggest that the best of the programs tested was diamond, ran with the "sensitive" option, which took 7% of the time as blastp to run, and produced results with lower error rates than blastp.

Availability: A program to obtain reciprocal best hits using each sequence comparison program is maintained at https://github.com/Computational-conSequences/SequenceTools

Contact: gmorenohagelsieb@wlu.ca

Supplementary information: There is no Supplementary data associated with this manuscript.

…two somewhat recently developed programs for sequence comparison include **diamond (Buchfink et al., 2015)** and **MMseqs2 (Steinegger and Söding, 2017)** (from now on mmseqs). Here we use lastal as a reference to the previous report (Ward and Moreno-Hagelsieb, 2014), where **lastal** was the program producing the most-similar-to-blastp results, and test these two new programs in terms of the proportion of RBH found and their quality compared to blastp.

Standard workflow:

DIAMOND or MMseqs2 (recommended, although BLAST+ can be used instead)

The MCL graph clustering algorithm

FastME (The appropriate version for your system, e.g. 'fastme-2.1.5-linux64', should be renamed `fastme', see instructions below.)

OrthoFinder: phylogenetic orthology inference for comparative genomics

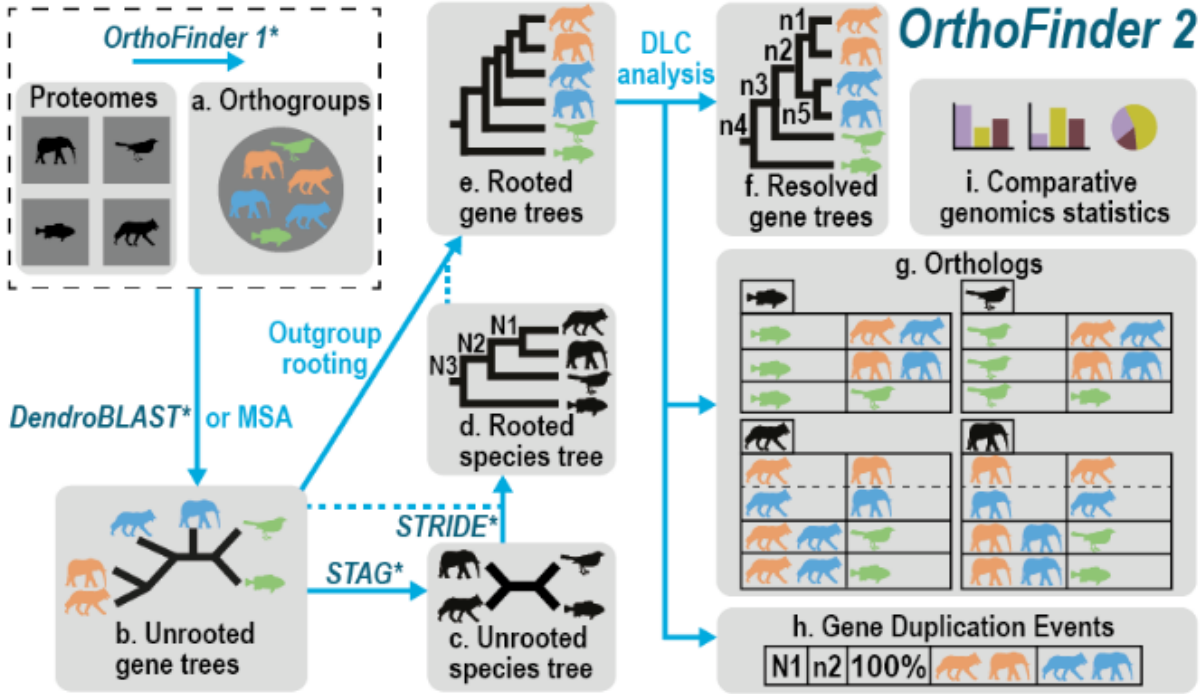


Figure 1: Automatic OrthoFinder analysis

What does OrthoFinder do?

OrthoFinder is a fast, accurate and comprehensive platform for comparative genomics. It finds **orthogroups** and **orthologs**, infers **rooted gene trees** for all orthogroups and identifies all of the **gene duplication events** in those gene trees. It also infers a **rooted species tree** for the species being analysed and maps the gene duplication events from the gene trees to branches in the species tree. OrthoFinder also provides **comprehensive statistics** for comparative genomic analyses. OrthoFinder is simple to use and all you need to run it is a set of protein sequence files (one per species) in FASTA format.

https://github.com/davidemms/OrthoFinder

BLAST searching a multidomain protein: HIV-1 pol

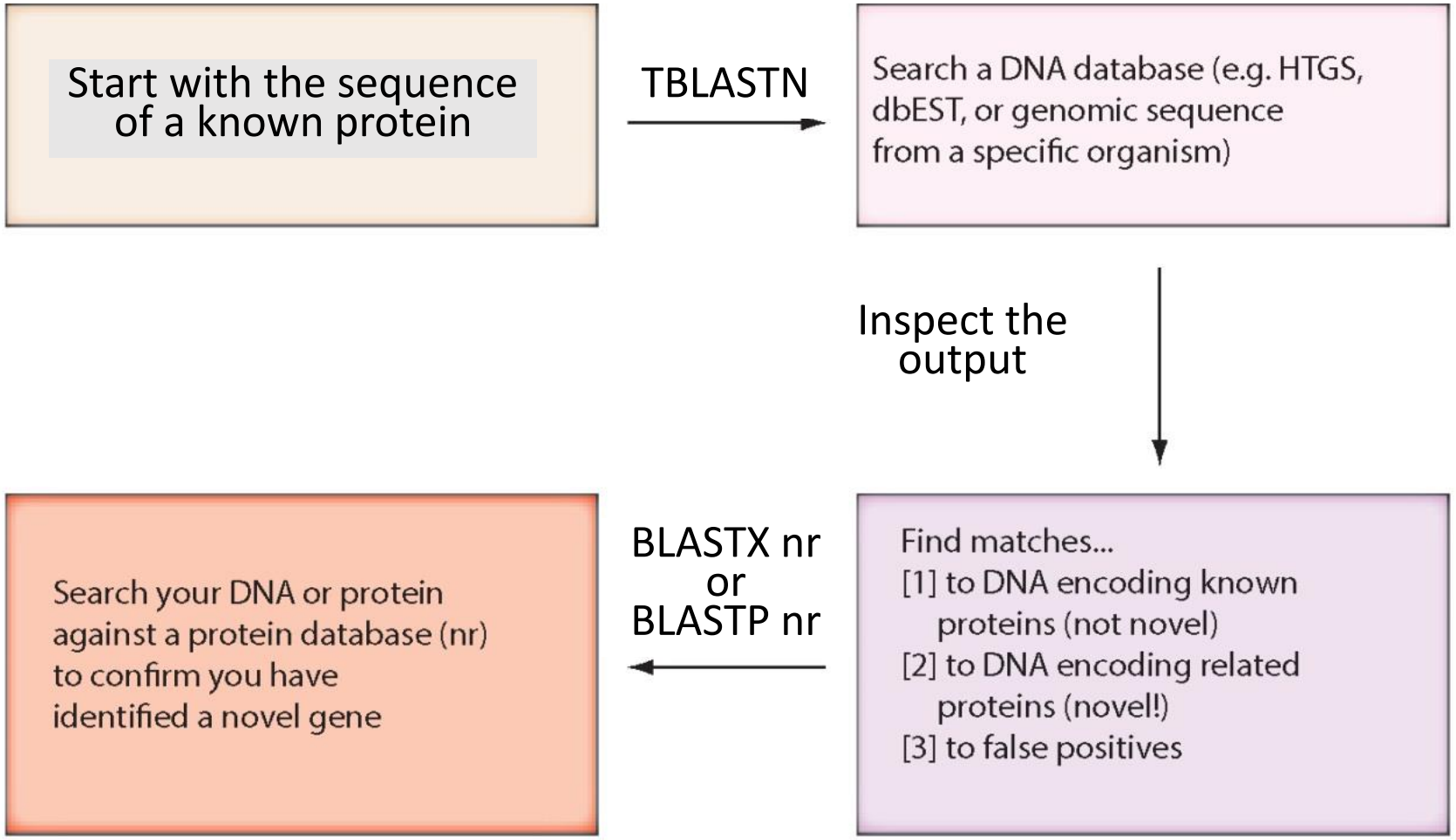
(b) List of alignments (query-anchored with dots for identities)

```
Query       1    MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWASRELERFAVNPGLLETSEGCRQI  60
NP_057849   1    ............................................................  60
P0C6F2      1    ...........K................................................  60
P03366      1    ............................................................  60
P03367      1    ............................................................  60
P04587      1    ............................................................  60
AAD03191    1    ............................Q.R.............................  60
P35963      1    .........A....K............Q.R................D.............  60
P12497      1    ..............K............Q...............................  60
P20875      1    ...............................R......................S.......  60
AAD03200    1    ................R.....R.....Q..............S...............  60
P20892      1    ..............K..............Q................I..............  60
Q73368      1    ...............................................S...........  60
BAB85751    1    ...........Q................M...............................  60
AFB39387    1    ...........Q................R.........................A......  60
P03369      1    ..............K.............................................  60
P05959      1    ...........K..K........R...R....................S....A......  60
AAG30116    1    ......I.......K..............R...L........Q..I.......A......  60
AAD03217    1    ......I......................Q...............................  60
```

R R,K R R R R,Q R

This output shows identical residues as a dot “.”. Note that the column positions that contain an arginine (R) can sometimes also contain a lysine (K) or glutamine (Q) in a position-specific pattern. This is a preview of the concept of position-specific scoring matrices.

“Find-a-gene project” to practice BLAST



“Find-a-gene project” example: novel globin

(a) Result of TBLASTN against nematode ESTs using human beta globin as a query

Ac_EH1r_01A07_M13 Adult Anguillicola crassus Anguillicola crassus cDNA clone Ac_EH1r_01A07

Sequence ID: gb|JK511422.1| Length: 559 Number of Matches: 1

Range 1: 40 to 483 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
149 bits(375)	6e-44	Compositional matrix adjust.	69/148(47%)	97/148(65%)	1/148(0%)	+1

```
Query  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
            MV  T  E +A+ +LW K+NV+E+G +A+ RLL+V PWTQR F +FG+LST  A+M N K
Sbjct  40   MVEWTDAEHTAILSLWKKINVEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNEK  219

Query  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
            V  HG  V+G     + ++D++K  +  LS +H +KLHVDP+NFRLL   +   +A  FG
Sbjct  220  VAKHGTTVMGGLDRAIQNMDDIKNAYRELSVMHSEKLHVDPDNFRLLSEHITLCMAAKFG  399

Query  121  -KEFTPPVQAAYQKVVAGVANALAHKYH  147
              EFT  VQ A+QK +  V +AL  +YH
Sbjct  400  PTEFTADVQEAWQKFLMAVTSALGRQYH  483
```

Query: NP_000509
Program: TBLASTN
Database: EST (nematodes)
Match: novel globin

(b) BLASTX result with a nematode EST showing its closest known protein match is in a vertebrate

RecName: Full=Hemoglobin anodic subunit beta; AltName: Full=Hemoglobin anodic beta chain

Sequence ID: sp|P80946.1|HBBA_ANGAN Length: 147 Number of Matches: 1

Range 1: 1 to 147 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
290 bits(742)	2e-97	Compositional matrix adjust.	136/147(93%)	141/147(95%)	0/147(0%)	+1

```
Query  43   VEWTDAEHTAILSLWKKINVEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNEKV  222
            VEWT+ E TAI S W KIN+EEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNN+KV
Sbjct  1    VEWTEDERTAIKSKWLKINIEEIGPQAMRRLLIVCPWTQRHFANFGNLSTAAAIMNNDKV  60

Query  223  AKHGTTVMGGLDRAIQNMDDIKNAYRELSVMHSEKLHVDPDNFRLLSEHITLCMAAKFGP  402
            AKHGTTVMGGLDRAIQNMDDIKNAYR+LSVMHSEKLHVDPDNFRLL+EHITLCMAAKFGP
Sbjct  61   AKHGTTVMGGLDRAIQNMDDIKNAYRQLSVMHSEKLHVDPDNFRLLAEHITLCMAAKFGP  120

Query  403  TEFTADVQEAWQKFLMAVTSALGRQYH  483
            TEFTADVQEAWQKFLMAVTSAL RQYH
Sbjct  121  TEFTADVQEAWQKFLMAVTSALARQYH  147
```

Confirmation
Query: nematode EST
Program: BLASTX
Best match: a globin, but *not* a previously annotated globin

“Find-a-gene project”

- The find-a-gene project is meant to be a very focused, specific project to help you understand how to use various BLAST tools (e.g. TBLASTN, BLASTX, BLASTP) and various databases.
- You can start with (almost) any protein, from the organism of your choice, and discover a “novel” gene in another organism that is homologous but has never been annotated before as related to your query. Therefore, you are discovering a new gene.
- You can take your new gene/protein, name it, then search it against databases to confirm it has not been described before.
- You can further perform multiple sequence alignment, phylogeny, and predict its protein structure and its function.