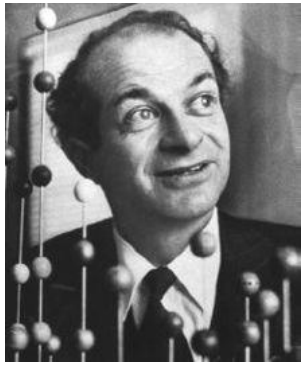
BCB 5200 Introduction to Bioinformatics

BLAST

Bioinformatics and Computational Biology
Saint Louis University

History of sequence analysis/alignment



- Cytochrome sequences to construct phylogenetic tree: Linus Pauling (1960)

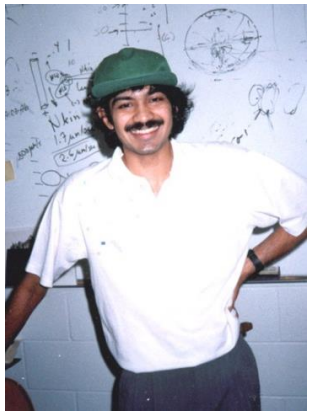
- Homology, sequence alignment, maximum parsimony: Fitch (1967, 1970)

- PAM matrix: Dayhoff (1966-1971)
- Dot-matrix methods (1970)
- Global sequence alignment: Needleman-Wusch (1970)

- Local sequence alignment: Smith-Waterman (1981)

History of sequence analysis/alignment

- Word methods:
  - FASTA (Fasta format): Lipman + Pearson (1985)
  - BLAST: Lipman + Altschul (1990)
- BLOSUM: Henikoff (1992)
- PSI-BLAST: Lipman + Altschul (1997)
- HMMER: Eddy (1998)



- Sequence analysis/classification: Koonin, Bork, Ponting, Aravind, Bateman

Outline

- BLAST algorithm
  - Three steps
  - Threshold
  - Expect value
- BLAST search steps and parameters
- Stand-alone BLAST
- BLAST search strategies
  - How to evaluate the significance of results
  - BLAST searching with multidomain protein: HIV-1 Pol

BLAST

BLAST (Basic Local Alignment Search Tool), is a heuristic method which 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.

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)

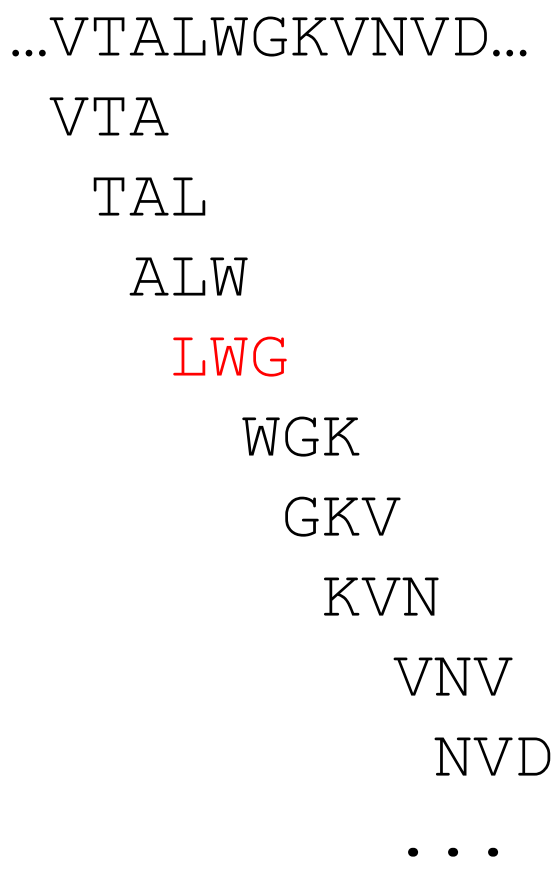
How the original BLAST algorithm works: three phases

1. List
2. Scan
3. Extension

How the original BLAST algorithm works: three phases

1. List: compile a list of word pairs (word length, w =3) above threshold T
   Query sequence: human beta globin (NP_000509)

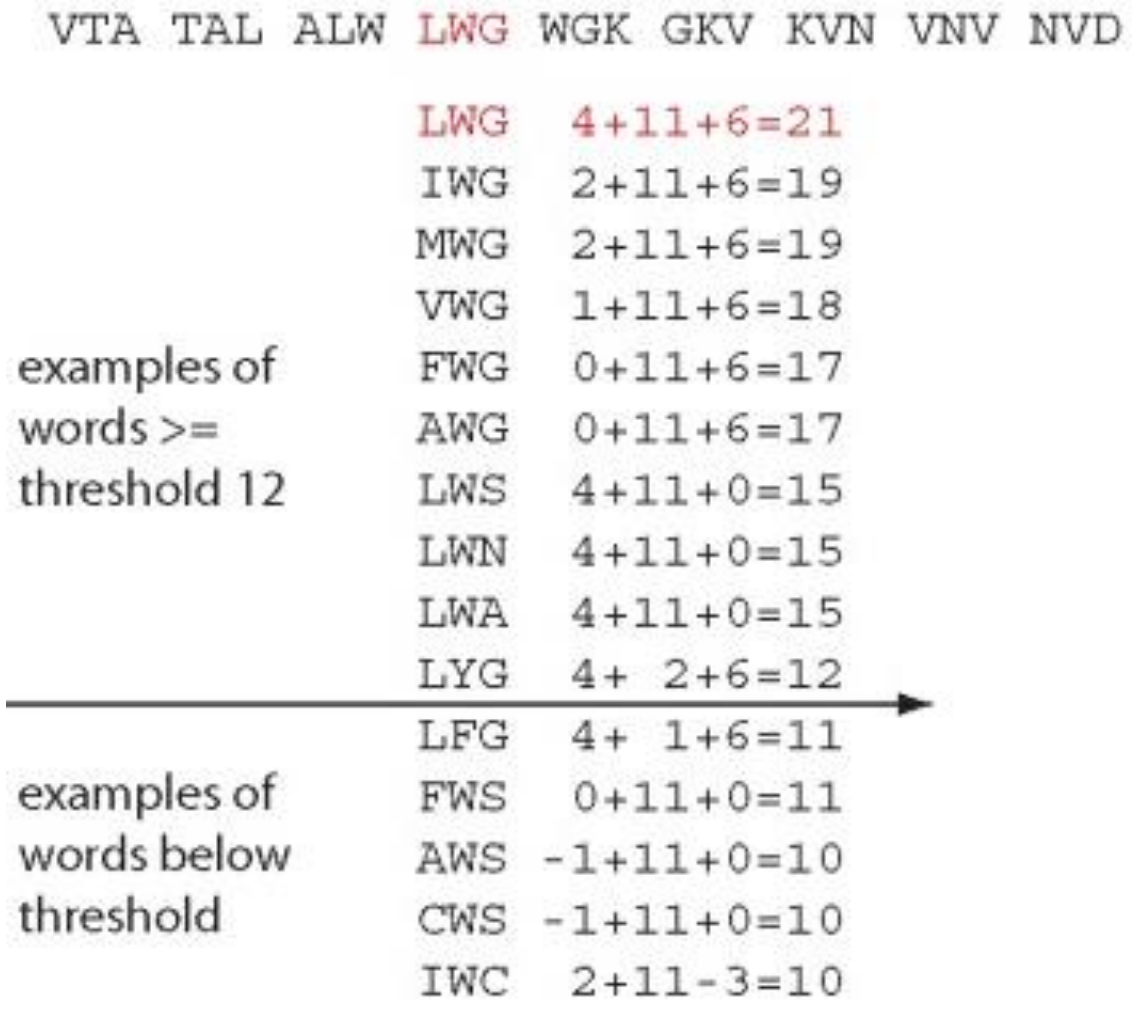
Words are derived from query sequence



1. List : compile a list of neighborhood words

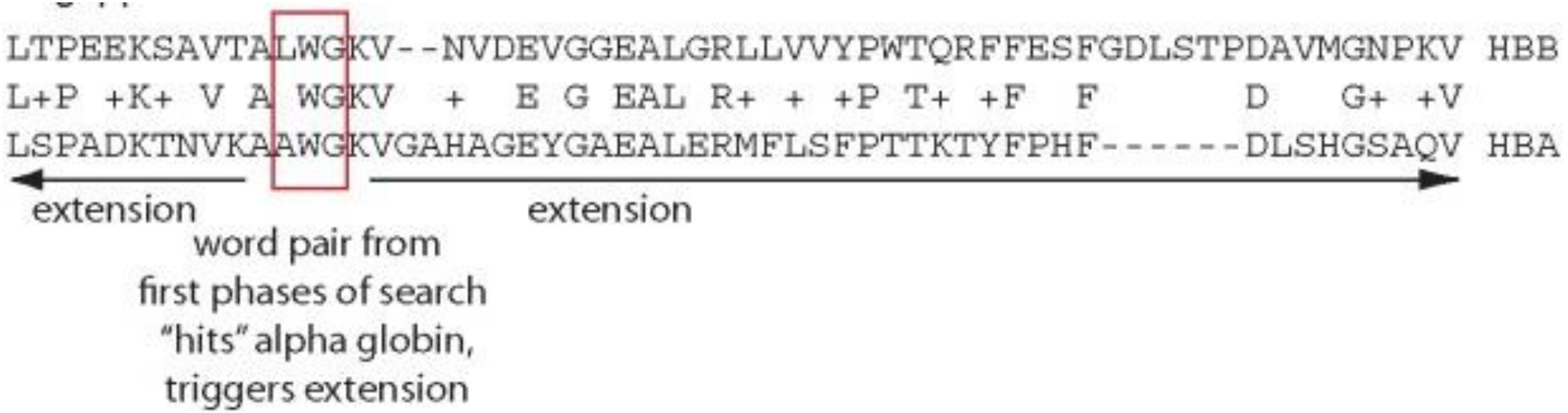
- **For every query word, generate a list of words (neighborhood word hits) matching it, both above and below threshold score (which were derived from BLOSUM62).**
- **We have a collection of neighborhood words based on the query.**

(T=12)



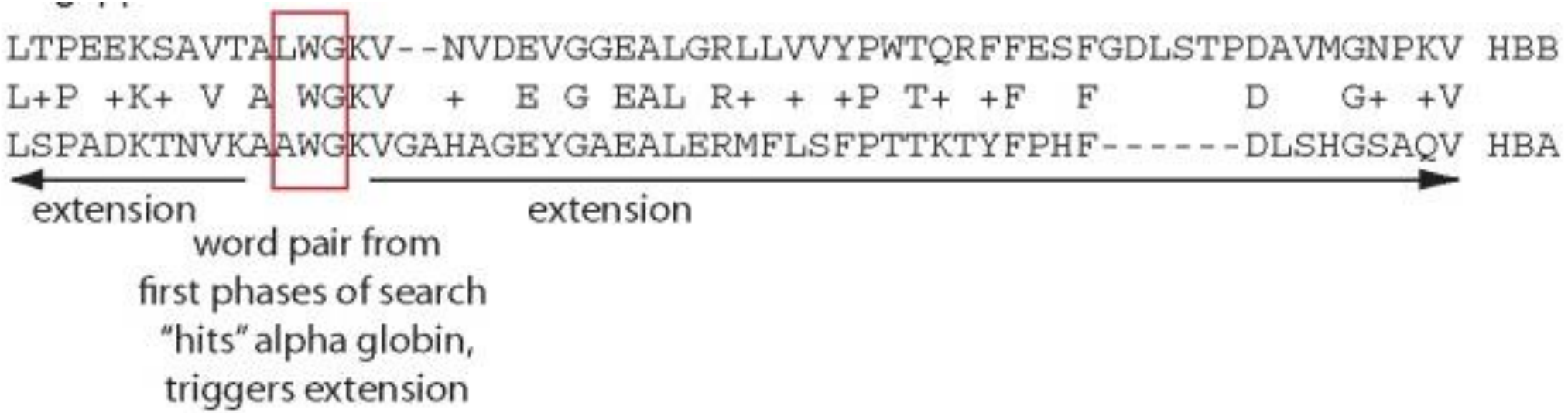
2. Scan the database for matches and extend

- **Select all the neighborhood 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**



3. Extension to generate gapped alignment

- **The database hits are extended in both directions to obtain high-scoring segment pairs (HSPs). If a HSP score exceeds a particular cutoff score S, it is reported in the BLAST output**



How a BLAST search works

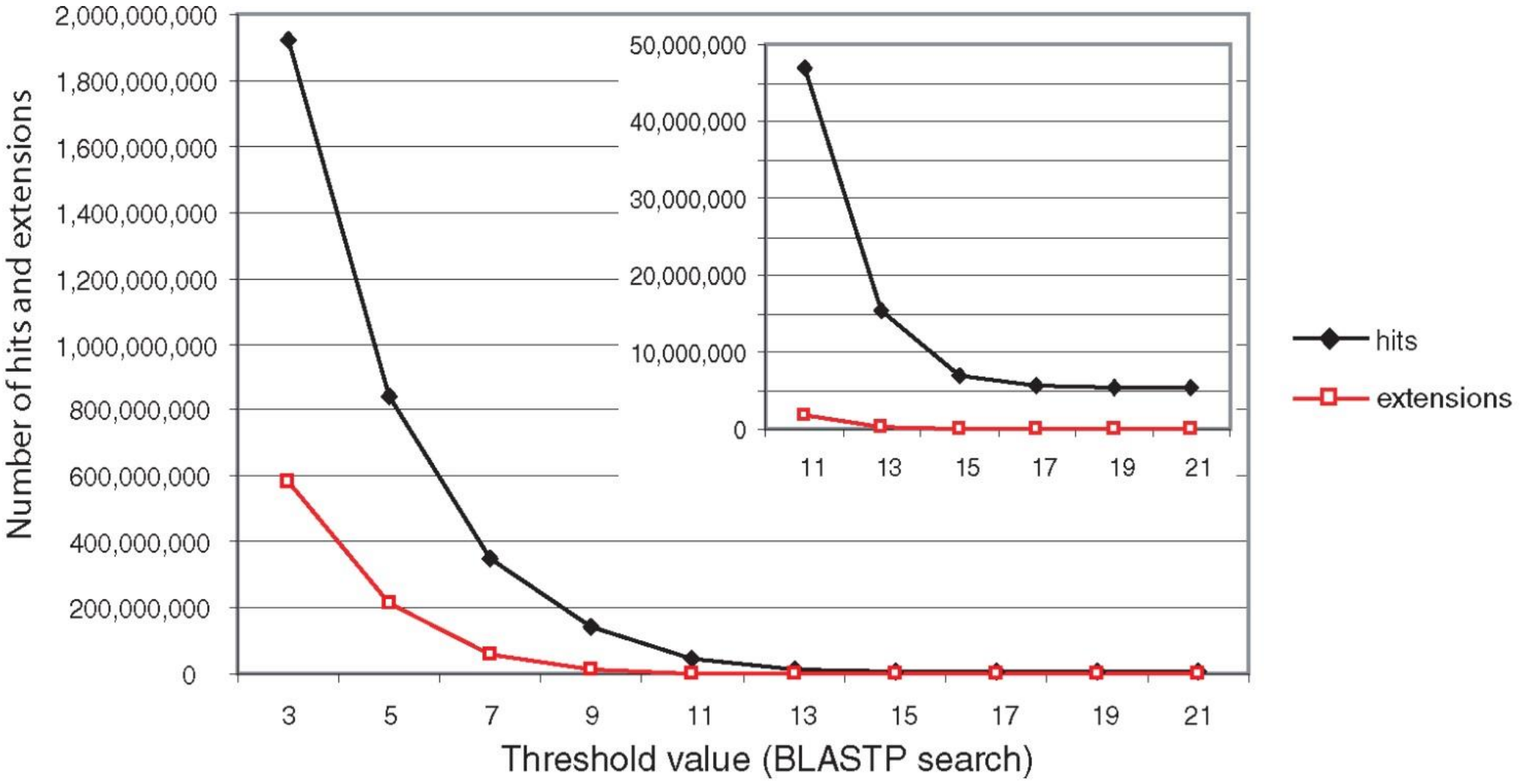
"The central idea of the BLAST algorithm is to confine attention to segment pairs that contain a word pair of length w with a score of at least T."

Altschul et al. (1990)

How a BLAST search works: threshold

- For protein-protein BLAST search, Non-Exact word Matches are taken into account based upon the similarity between words (BLOSUM matrix).
- The word-sizes are typically 2 and 3.
- The default threshold T for protein BLAST search (BLASTP) is 11. But it can be changed.

Effect of changing the threshold T: Lower T yields more database hits (black line) and extensions (red)



Word length for nucleotide BLAST searches

- For nucleotide-nucleotide searches (BLASTN), an Exact Match of the entire word is required before an extension is initiated, so that one normally regulates the sensitivity and speed of the search by increasing or decreasing the word-size.

- For BLASTN, the word size is typically 7, 11, or 15. 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 (see below), the word size is 28 and can be adjusted to 64. MegaBLAST is VERY fast for finding closely related DNA sequences!

How to interpret a BLAST search: expect value (E-value)

- E-value: the number of high-scoring segment pairs (HSPs) expected to occur by chance with a score of at least *S* in a database search
- An indication of statistical significance of a alignment
- Depends on the number of distinct sequences in the database and the length of query sequence
- Lower values signify higher similarity

$$E = kmNe^{-\lambda S}$$

m	# letters in query
N	# letters in database
mN	size of search space
λS	normalized score
k	minor constant

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.

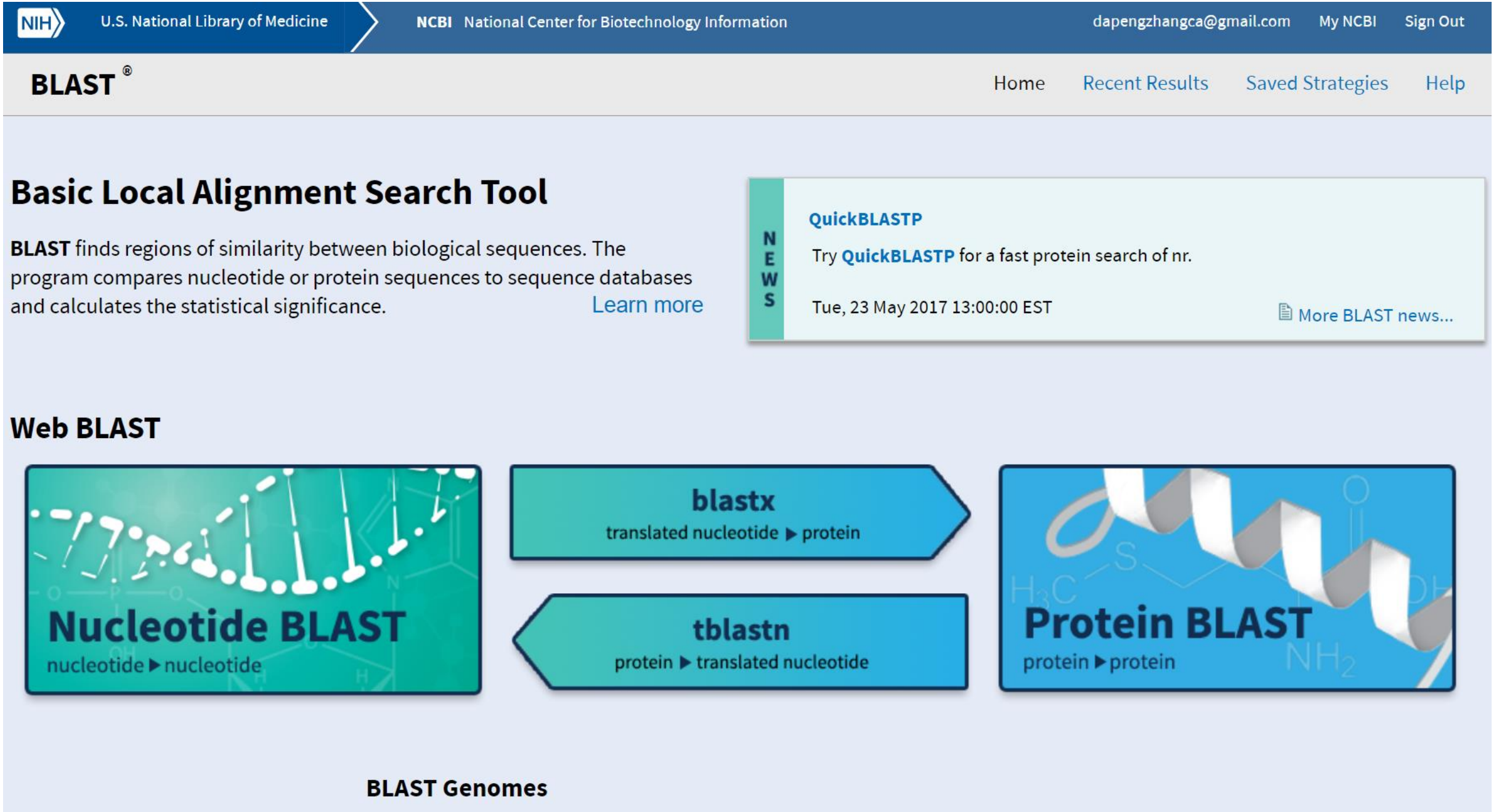
E	*p*
10	0.99995460
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258 (about 0.1)
0.05	0.04877058 (about 0.05)
0.001	0.00099950 (about 0.001)
0.0001	0.00010000

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

Outline

- BLAST algorithm
  - Three steps
  - Threshold
  - Expect value
- BLAST search steps and parameters
- Stand-alone BLAST
- BLAST search strategies
  - How to evaluate the significance of results
  - BLAST searching with multidomain protein: HIV-1 Pol

BLASTP search at NCBI: overview of web-based search



BLASTP search at NCBI: overview of web-based search

query: FASTA format or accession

database

Entrez query

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

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

BLASTN	DNA	1	DNA

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

BLASTX	DNA	6	protein

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.

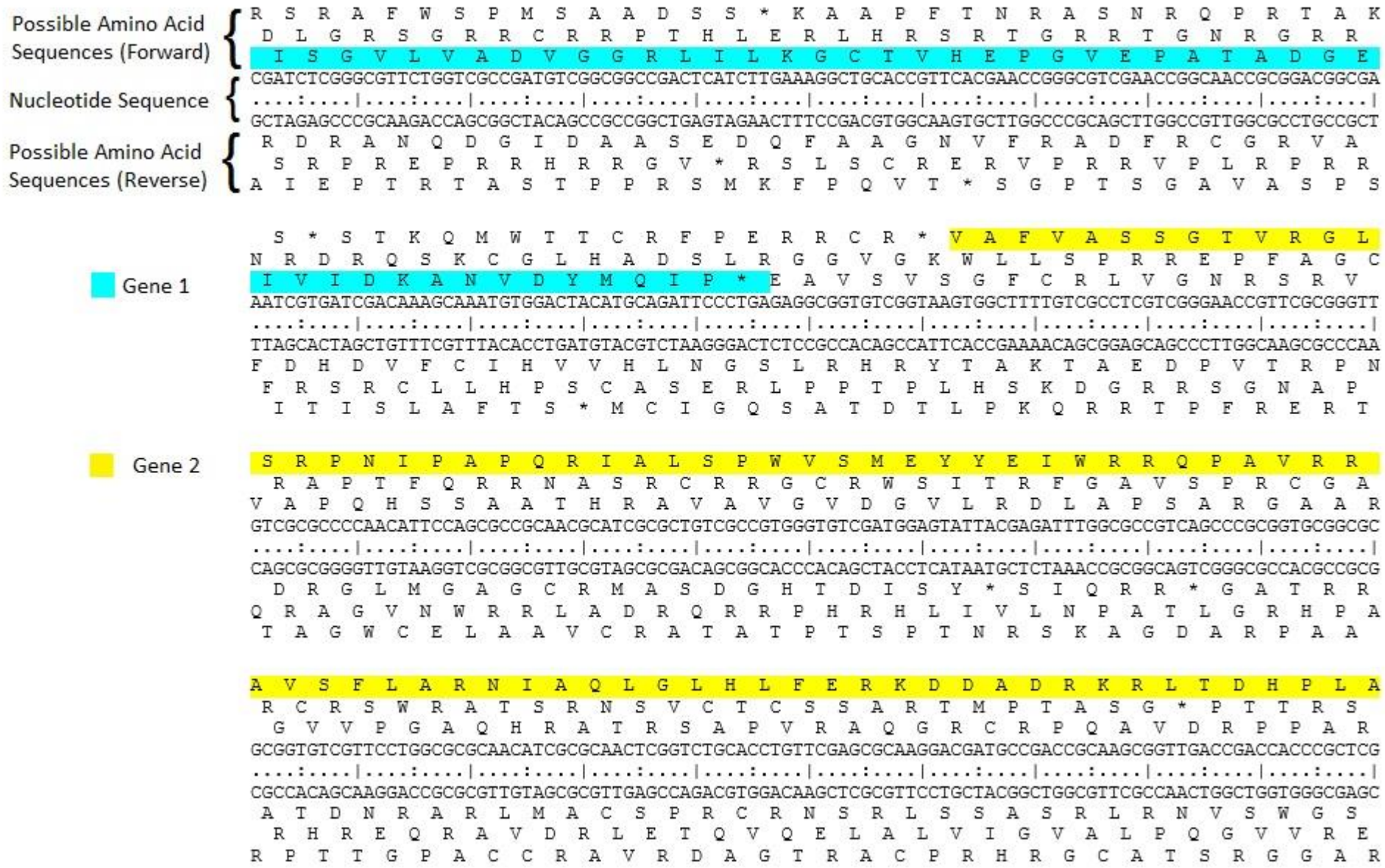
TBLASTN	protein	6	DNA

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.

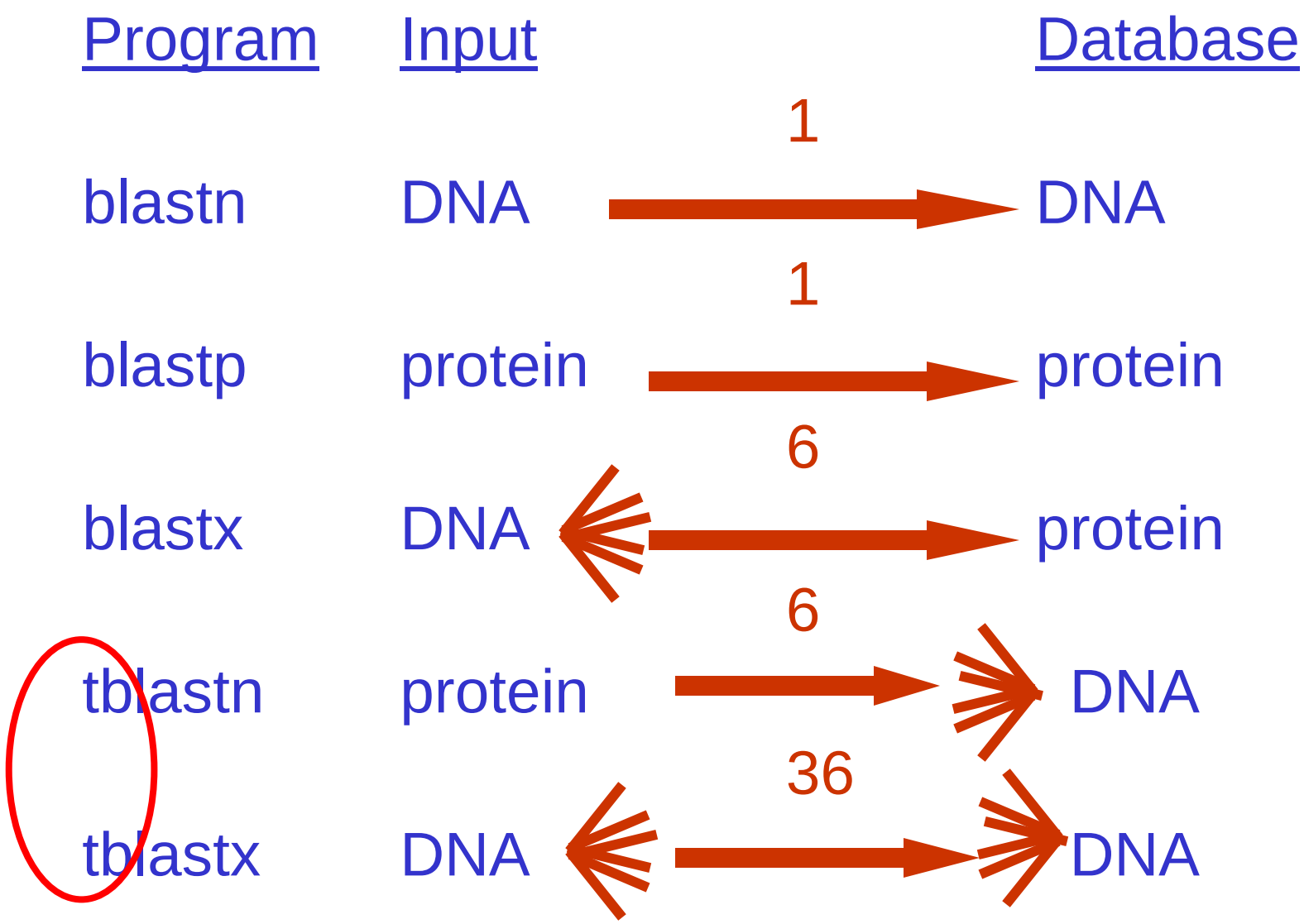
TBLASTX	DNA	36	DNA

TBLASTX is the most computationally intensive BLAST algorithm. It translates DNA from both a query and a database into six potential proteins, then performs 36 protein-protein database searches.

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



Step 2. Choose the BLAST program



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 (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 sequences as well as phase 0, 1, and 2 HTGS sequences	25 million
refseq_rna	NCBI transcript reference sequences	3.5 million
refseq_genomic	NCBI genomic reference sequences	2.7 million
NCBI Genomes	NCBI chromosome sequences	28,000
Expressed sequence tags (EST)	Database of GenBank+EMBL+DDBJ sequences from EST Divisions	75 million
Genomic survey sequences (gss)	Genome survey sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences	36 million
High-throughput genomic sequences (HTGS)	Unfinished high-throughput genomic sequences; sequences: phases 0,1 and 2	153,000
Patent sequences	Nucleotide sequences derived 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

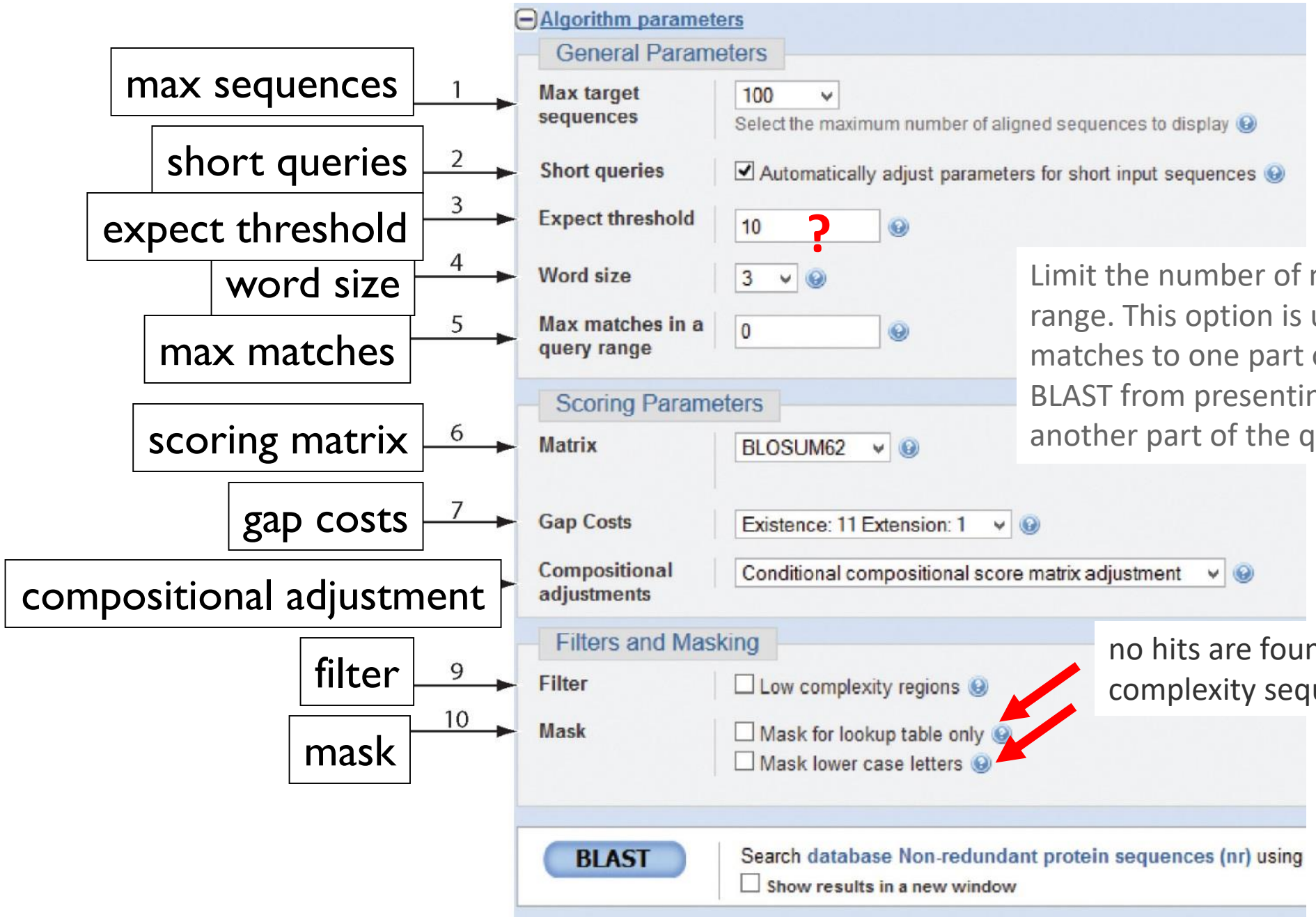
Step 4: optional parameters

You can...

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

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

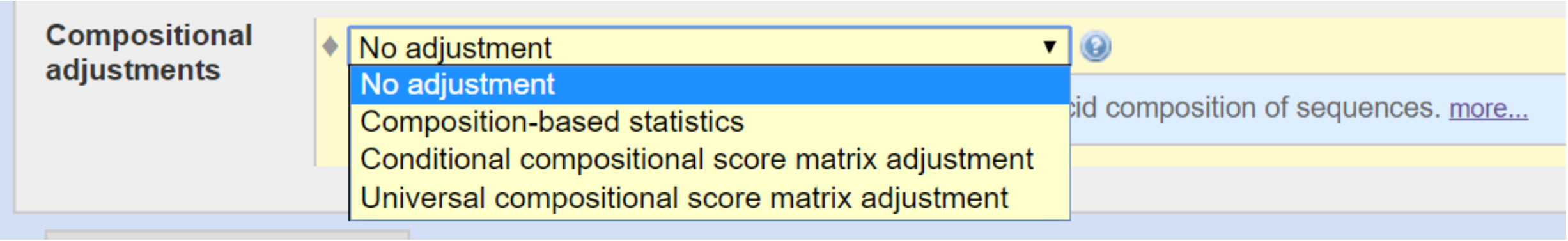
Step 4a: choose optional BLASTP search parameters



Limit the number of matches to a query range. This option is useful if many strong matches to one part of a query may prevent BLAST from presenting weaker matches to another part of the query.

no hits are found based upon low-complexity sequence or repeats

Step 4a: compositional adjustments



- Biology:
  - Some proteins have nonstandard compositions (hydrophobic residues; cysteine-rich regions)
  - Some organisms have the entire gene with a very high G/C or A/T content (Malaria parasite Plasmodium has 80% AT, biasing its proteins towards having residues encoded by AT-rich codons)

- A standard matrix such as BLOSUM62 is not appropriate

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

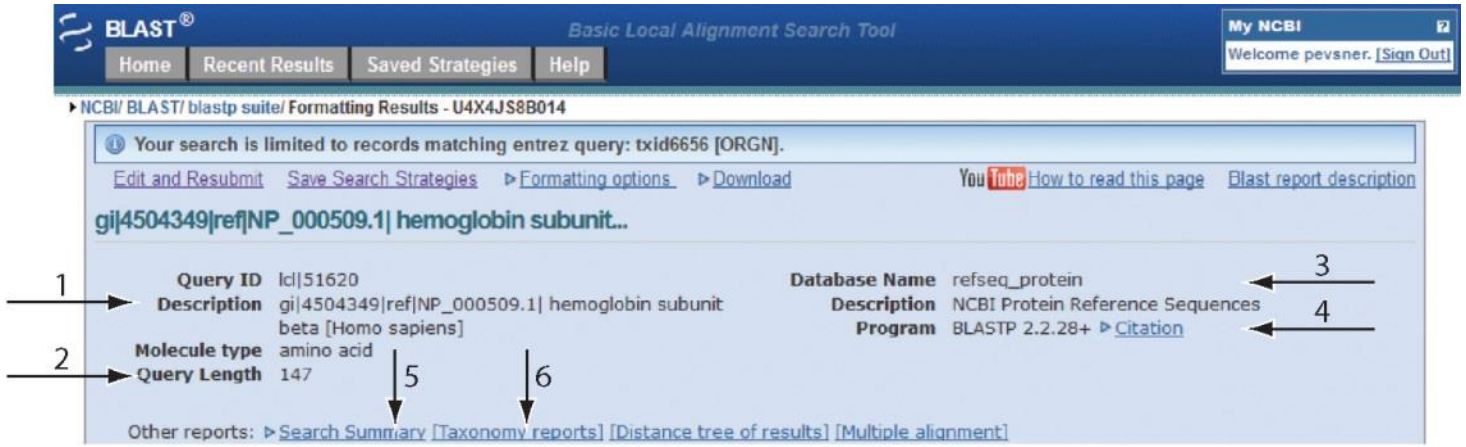
Range 1: 33 to 114 GenPept Graphics

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

Expect value

BLOSUM62 matrix

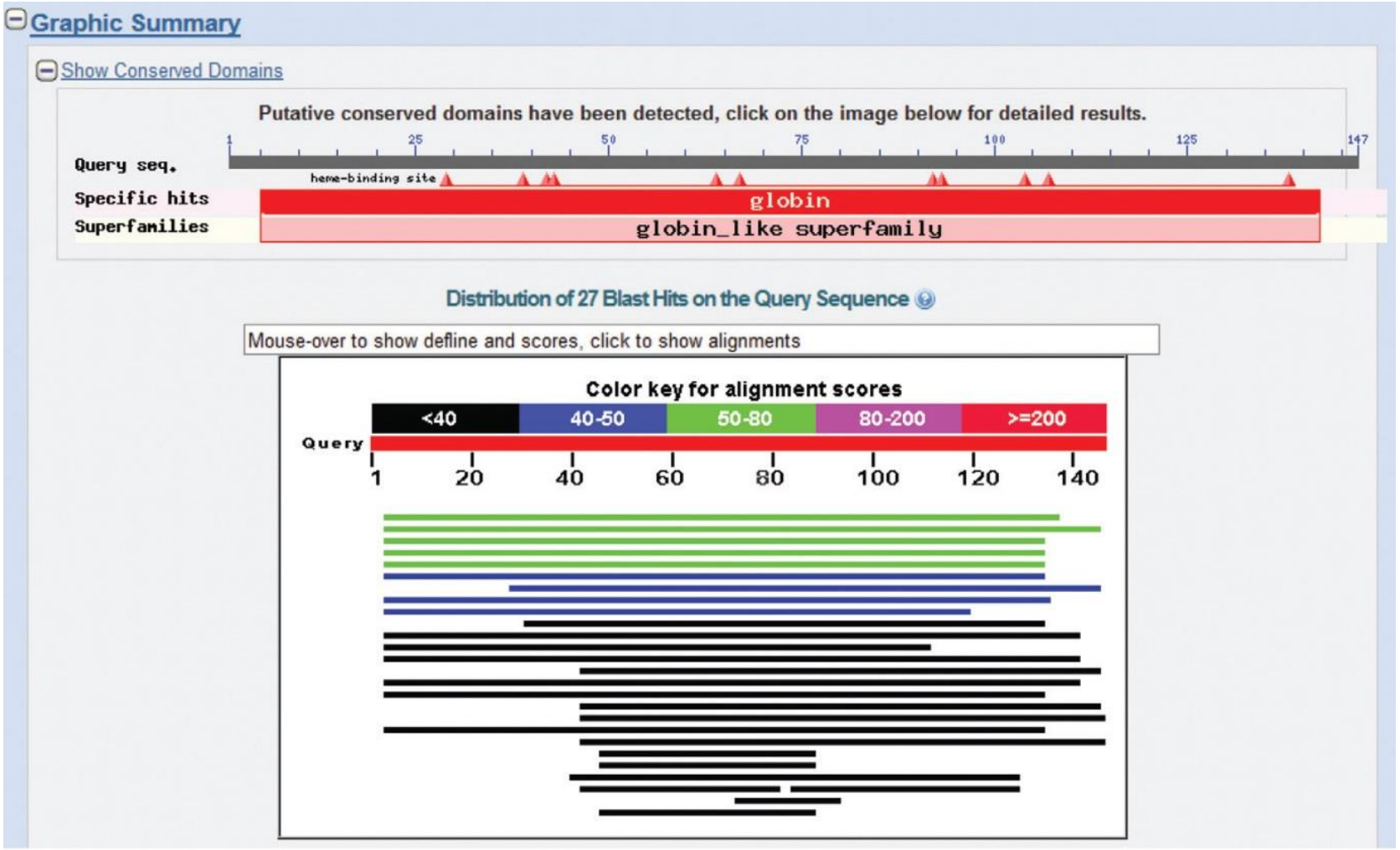
Threshold value T

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]

Size of database

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

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

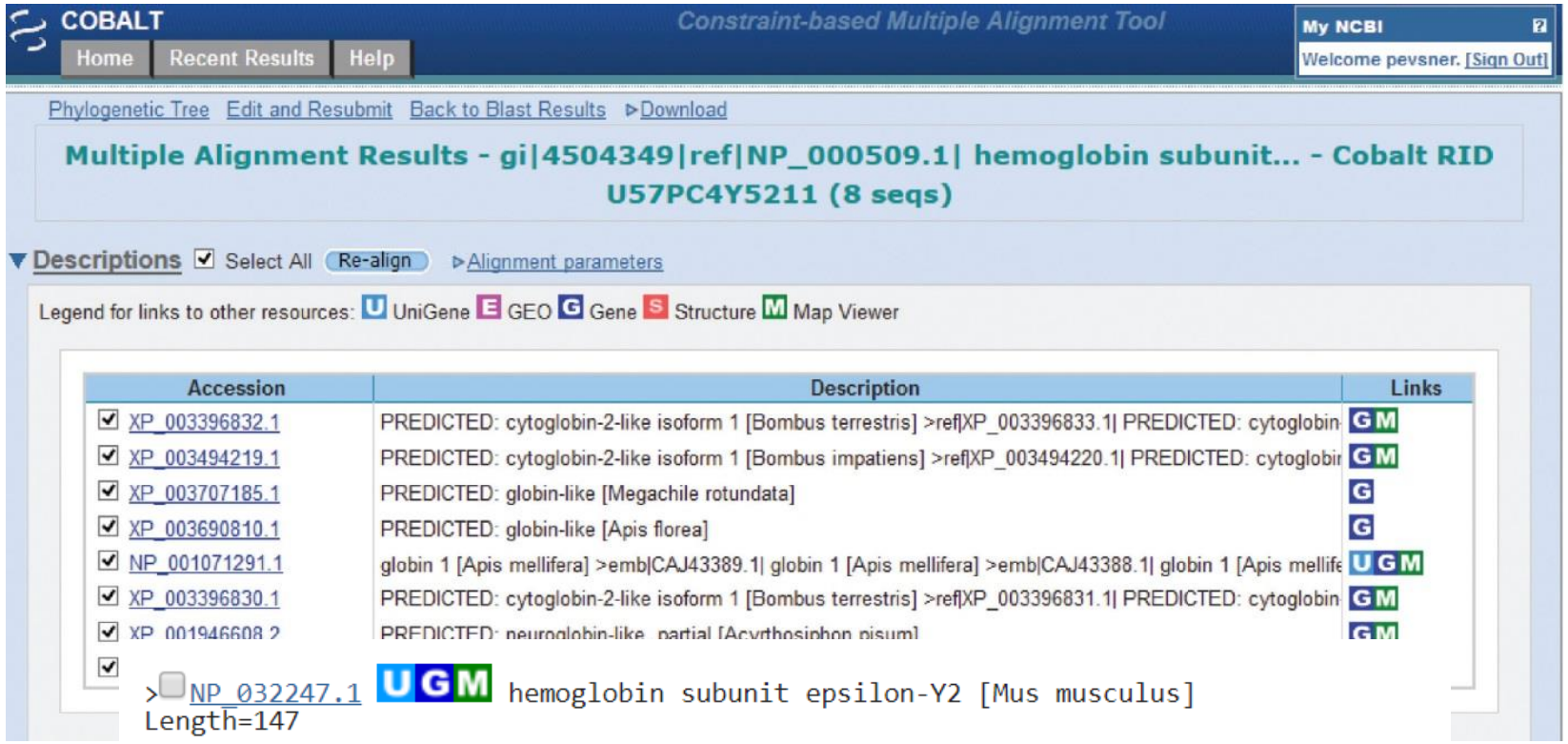
Sequences producing significant alignments:

Select: All None Selected:2

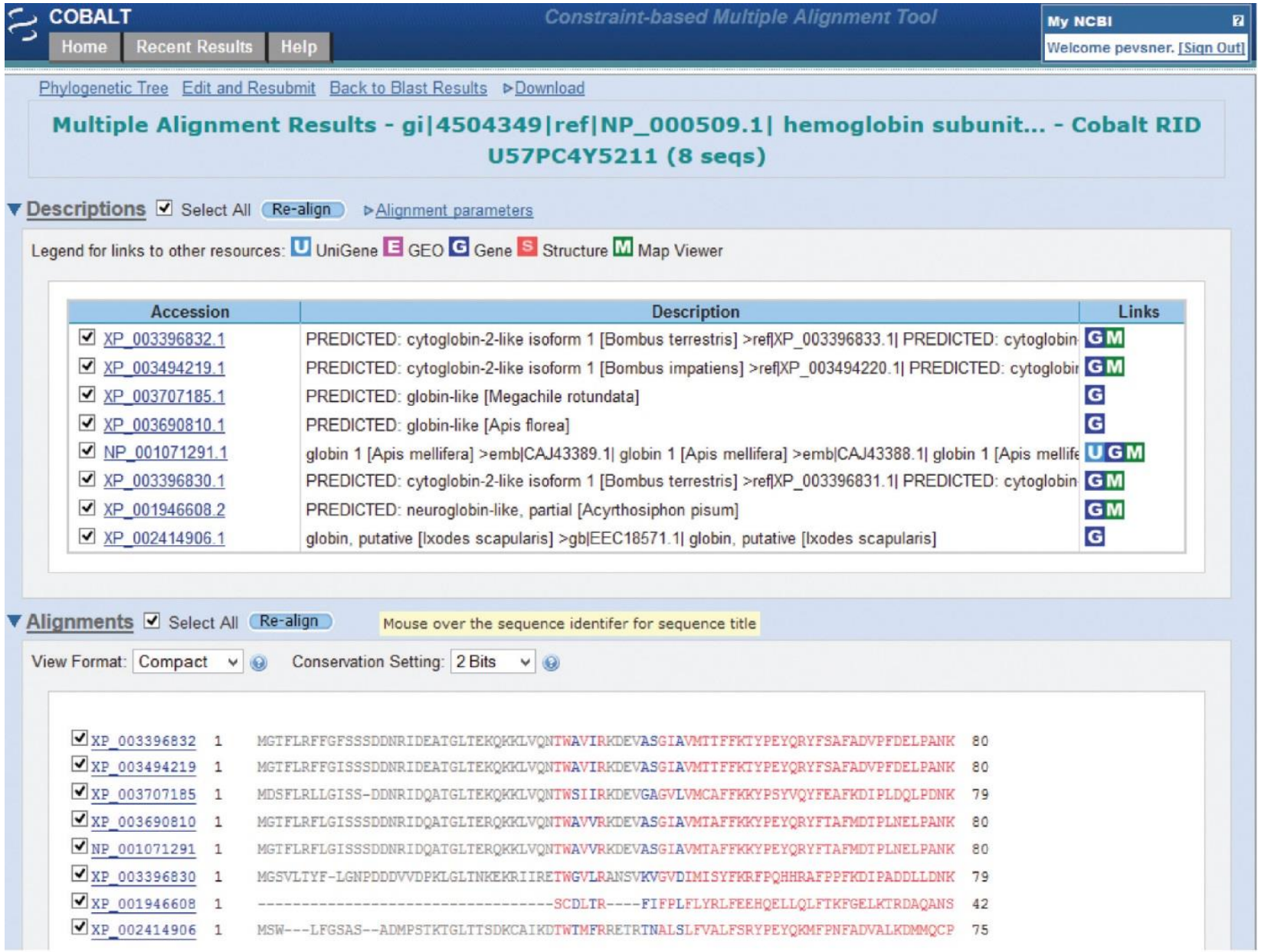
Alignments Download GenPept Graphics Distance tree of results Multiple alignment

	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



BLAST output can be formatted to display multiple alignment



BLAST output: taxonomy report

RID W7UKC5DP015 (Expires on 09-23 00:35 am)

Query ID NP_000509.1
Description hemoglobin subunit beta [Homo sapiens]
Molecule type amino acid
Query Length 147

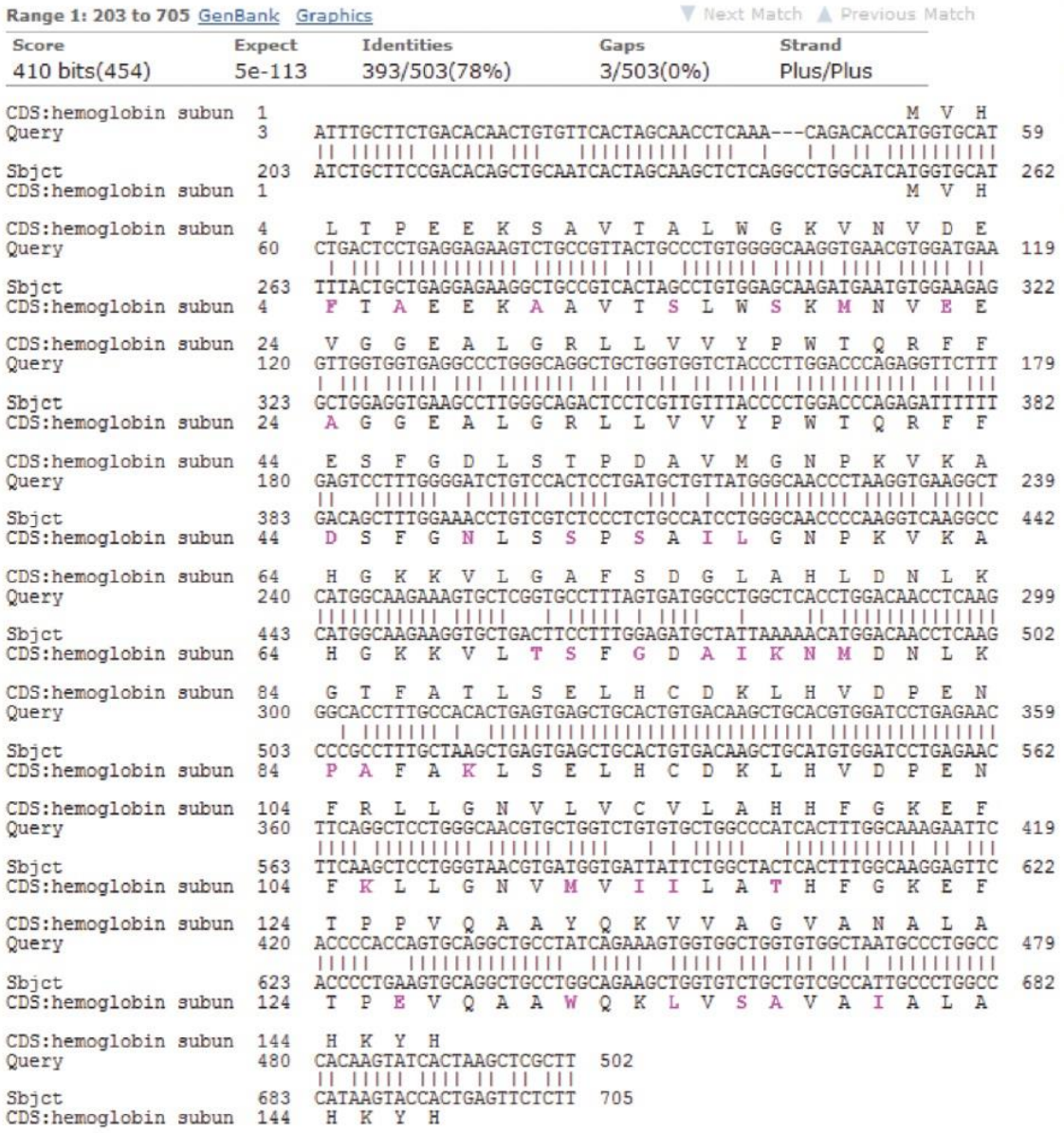
Database Name refseq_protein
Description NCBI Protein Reference Sequer
Program BLASTP 2.7.0+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment] [MSA viewer]

Graphic Summary

.....Panthera pardus	carnivores	251	6	Panthera pardus hits
.....Lipotes vexillifer	whales & dolphins	250	3	Lipotes vexillifer hits
.....Ceratotherium simum simum	odd-toed ungulates	250	4	Ceratotherium simum simum hits
.....Myotis brandtii	bats	250	4	Myotis brandtii hits
.....Camelus ferus	even-toed ungulates	249	3	Camelus ferus hits
.....Vicugna pacos	even-toed ungulates	249	4	Vicugna pacos hits
.....Camelus bactrianus	even-toed ungulates	249	3	Camelus bactrianus hits
.....Camelus dromedarius	even-toed ungulates	249	3	Camelus dromedarius hits
.....Tursiops truncatus	whales & dolphins	248	3	Tursiops truncatus hits
.....Orcinus orca	whales & dolphins	248	4	Orcinus orca hits
.....Equus przewalskii	odd-toed ungulates	248	5	Equus przewalskii hits
.....Condylura cristata	insectivores	247	3	Condylura cristata hits
.....Bos taurus	even-toed ungulates	246	11	Bos taurus hits
.....Capra hircus	even-toed ungulates	246	10	Capra hircus hits
.....Bos mutus	even-toed ungulates	246	8	Bos mutus hits
.....Bison bison bison	even-toed ungulates	246	7	Bison bison bison hits
.....Ovis aries musimon	even-toed ungulates	246	10	Ovis aries musimon hits

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



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  GCTGGAGGTGAAGCCTTGGGCAGACTCCTCGTTGTTTACCCCTGGACCCAGAGATTTTTT  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
```

Outline

- BLAST algorithm
  - Three steps
  - Threshold
  - Expect value
- BLAST search steps and parameters
- Stand-alone BLAST
- BLAST search strategies
  - How to evaluate the significance of results
  - BLAST searching with multidomain protein: HIV-1 Pol

Command-line BLAST+

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

(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

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
```

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
```

Build a custom BLAST database and search

```
$ makeblastdb –n mydb.txt –parse_seqids –dbtype nucl
```

Outline

- BLAST algorithm
  - Three steps
  - Threshold
  - Expect value
- BLAST search steps and parameters
- Stand-alone BLAST
- BLAST search strategies
  - How to evaluate the significance of results
  - BLAST searching with multidomain protein: HIV-1 Pol

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:

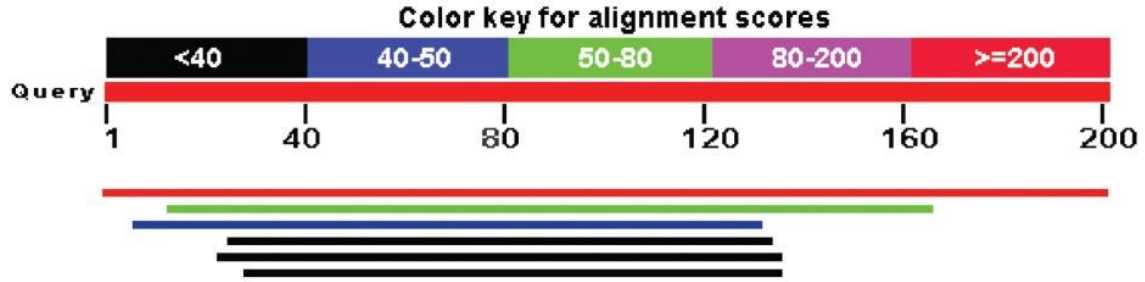
- Identifying orthologs and paralogs
- Determining what proteins or genes are present in a particular organism
- Discovering variants of genes or proteins
- Identifying residues important for protein structure and function
- Identifying distant relationship

Principles of BLAST searching

- How to evaluate the significant results
  - E-value
  - It is also necessary to apply biological criteria to support the inference of homology

  - A common motif or signature?
  - A reasonable multiple sequence alignment?
  - Similar biological function?
  - Structure or fold?
  - Genomic context information?
  - “Reciprical” BLASTP search?

(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

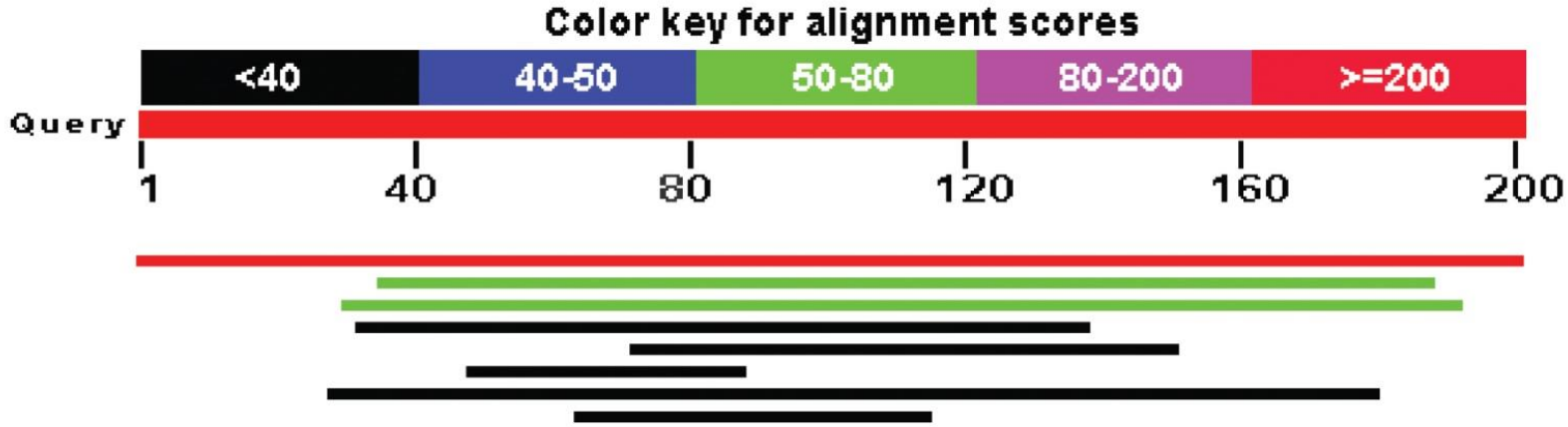
Score	Expect	Method	Identities	Positives	Gaps
32.3 bits(72)	0.18	Compositional matrix adjust.	28/114(25%)	49/114(42%)	8/114(7%)

```
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
```

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

(a) Graphical overview



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

(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

Pairwise alignment of CG8 with non-homologous proteins

(c) Pairwise alignments with nonhomologous proteins

- Query and subject are very different lengths
- E values are not significant
- Matches lack GXW motif

Download GenPept Graphics

tenascin-X isoform 1 precursor [Homo sapiens]

Sequence ID: ref|NP_061978.6| Length: 4242 Number of Matches: 1

Range 1: 3255 to 3330 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
30.0 bits(66)	1.5	Compositional matrix adjust.	25/81(31%)	36/81(44%)	6/81(7%)

```
Query  73    ITLHVAPQGTAMAVSTFRKLD-GICWQVRQLYGDTGVLGRFLLQARDARGAVHVVVAETD  131
             I L V P+    +AV+        G+ W V Q      G     FL+Q RDA+G    V    D
Sbjct  3255  IPLPVEPRLGELAVAAVTSDSVGLSWTVAQ-----GPFDSFLVQYRDAQGQPQAVPVSGD  3309

Query  132   YQSFAVLYLERAGQLSVKLYA  152
              ++ AV  L+ A +    L+
Sbjct  3310  LRAVAVSGLDPARKYKFLLFG  3330
```

Download GenPept Graphics

neuroblastoma-amplified sequence [Homo sapiens]

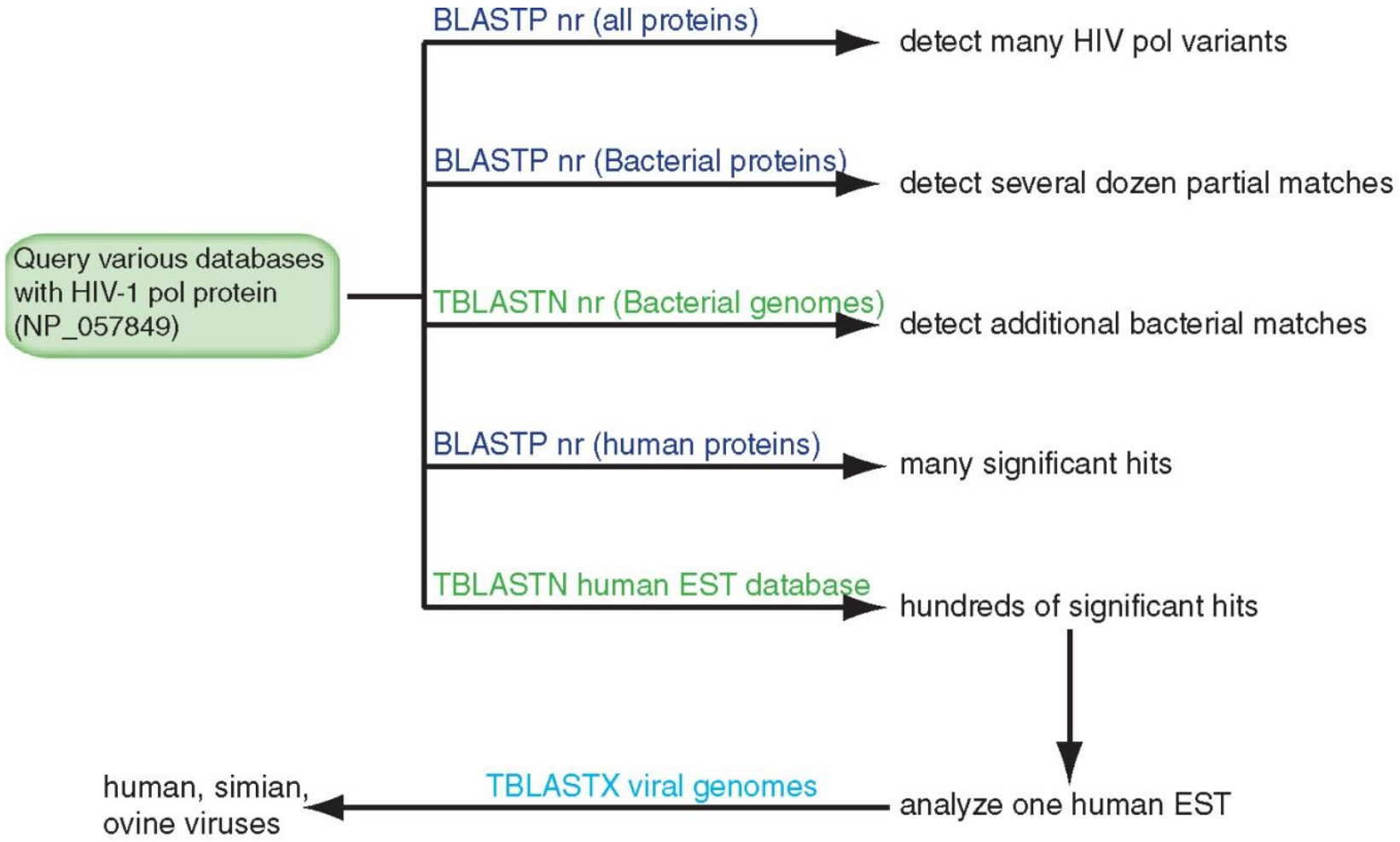
Sequence ID: ref|NP_056993.2| Length: 2371 Number of Matches: 1

Range 1: 2323 to 2360 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
29.6 bits(65)	2.1	Compositional matrix adjust.	18/41(44%)	23/41(56%)	3/41(7%)

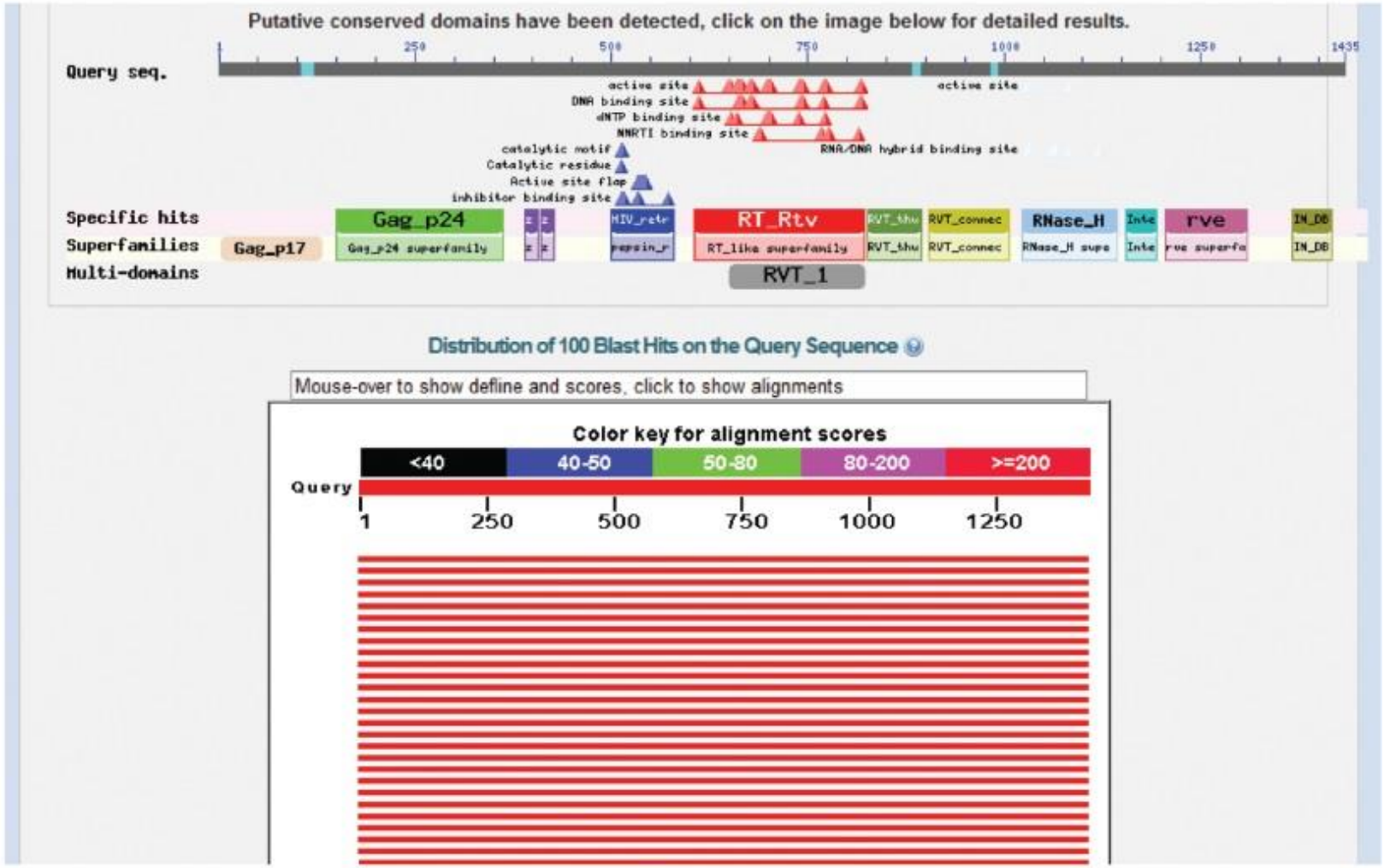
```
Query  49    GTWLLVAVGSACRFLQEQGHRAEATTLHVAPQGTAMAVSTF  89
             G W    +G   R L+E GH AEA +L +A +GT  A  TF
Sbjct  2323  GRWDAEELG---RHLREAGHEAEAGSLLLAVRGTHQAFRTF  2360
```

BLAST searching a multidomain protein: HIV-1 pol



BLAST searching a multidomain protein: HIV-1 pol

(a) Graphical overview



The BLAST output includes a graphic of the various domains in HIV-1 pol

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 (Chapter 5).

Taxonomy report for a BLAST searching HIV-1 pol

Most of the matches are to viruses, but there are also matches to rabbit, fungal, pig, and insect sequences.

Human immunodeficiency virus 1 [viruses] taxid 11676

			0
			0
			0
ref|YP_001856242.1|	reverse transcriptase [Human immunodef...	1149	0.0
ref|NP_789739.1|	reverse transcriptase p51 subunit [Human ...	912	0.0
ref|NP_057850.1|	Pr55(Gag) [Human immunodeficiency virus 1]	908	0.0
ref|NP_705928.1|	integrase [Human immunodeficiency virus 1]	602	0.0
ref|YP_001856243.1|	integrase [Human immunodeficiency viru...	602	0.0
ref|NP_579880.1|	capsid [Human immunodeficiency virus 1]	481	4e-156
ref|NP_579876.2|	matrix [Human immunodeficiency virus 1]	271	7e-81
ref|NP_705926.1|	retropepsin [Human immunodeficiency virus 1]	204	2e-57
ref|YP_001856241.1|	retropepsin [Human immunodeficiency vi...	204	2e-57
ref|NP_579881.1|	nucleocapsid [Human immunodeficiency viru...	130	5e-32
ref|NP_787043.1|	Gag-Pol Transframe peptide [Human immunod...	119	4e-28

Simian immunodeficiency virus [viruses] taxid 11723

ref|NP_687035.1|	Gag-Pol [Simian immunodeficiency virus]	1687	0.0
ref|NP_054369.1|	gag protein [Simian immunodeficiency virus]	502	1e-159

Human immunodeficiency virus 2 [viruses] taxid 11709

ref|NP_663784.1|	gag-pol fusion polyprotein [Human immunod...	1675	0.0
ref|NP_056837.1|	gag polyprotein [Human immunodeficiency v...	523	3e-167

Simian immunodeficiency virus SIV-mnd 2 [viruses] taxid 159122

ref|NP_758887.1|	pol protein [Simian immunodeficiency viru...	1377	0.0
ref|NP_758886.1|	gag protein [Simian immunodeficiency viru...	486	2e-153

Feline immunodeficiency virus [viruses] taxid 11673

ref|NP_040973.1|	pol polyprotein [Feline immunodeficiency ...	489	2e-148
ref|NP_040972.1|	gag protein [Feline immunodeficiency virus]	158	8e-38

Equine infectious anemia virus [viruses] taxid 11665

ref|NP_056902.1|	pol polyprotein [Equine infectious anemia...	424	1e-123
ref|NP_056901.1|	gag protein [Equine infectious anemia virus]	154	2e-36

///

Candida albicans SC5314 [ascomycetes] taxid 237561

ref|XP_888860.1|	hypothetical protein CaO19_6468 [Candida ...	90	2e-15
ref|XP_721310.1|	hypothetical protein CaO19.6468 [Candida ...	86	1e-14

Sus scrofa (wild boar, ...) [even-toed ungulates] taxid 9823

ref|XP_003482346.1|	PREDICTED: hypothetical protein LOC100...	90	2e-15

Tribolium castaneum (rust-red flour beetle) [beetles] taxid 7070

ref|XP_001815322.1|	PREDICTED: similar to orf [Tribolium c...	89	5e-15
ref|XP_001808495.1|	PREDICTED: similar to orf [Tribolium c...	88	8e-15

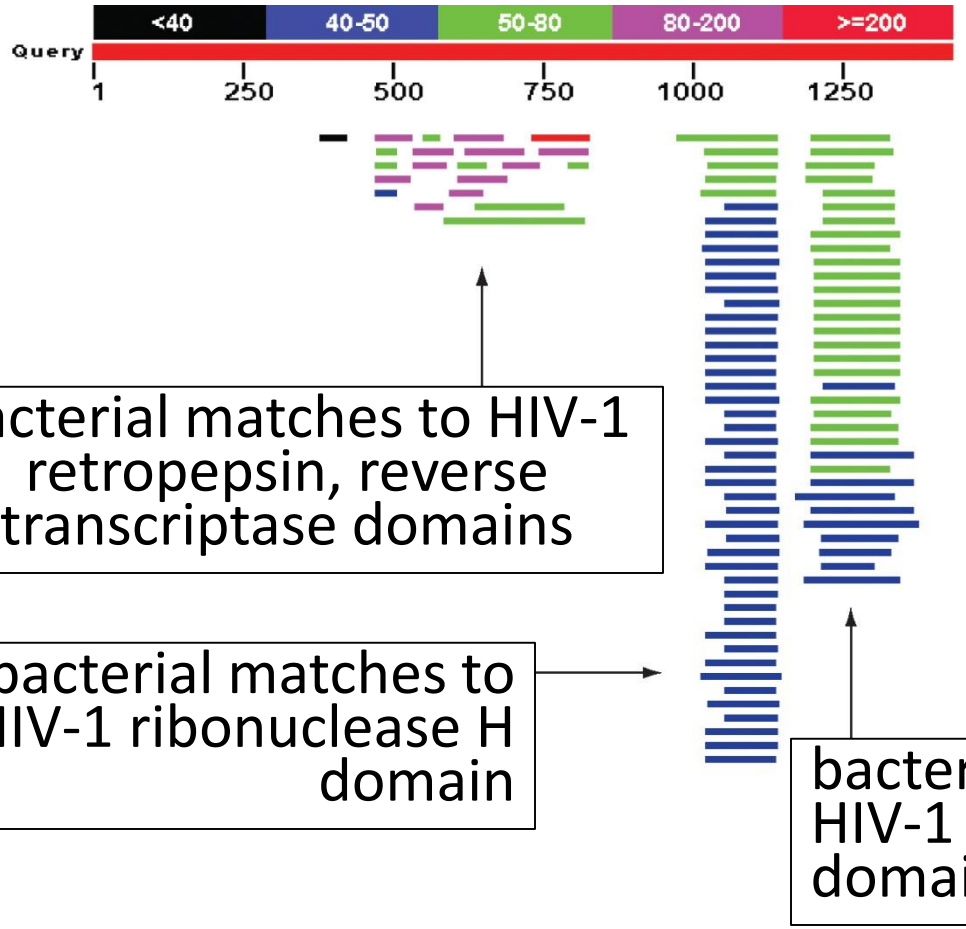
Candida dubliniensis CD36 [ascomycetes] taxid 573826

ref|XP_002421195.1|	retrovirus-related Pol polyprotein fro...	88	6e-15

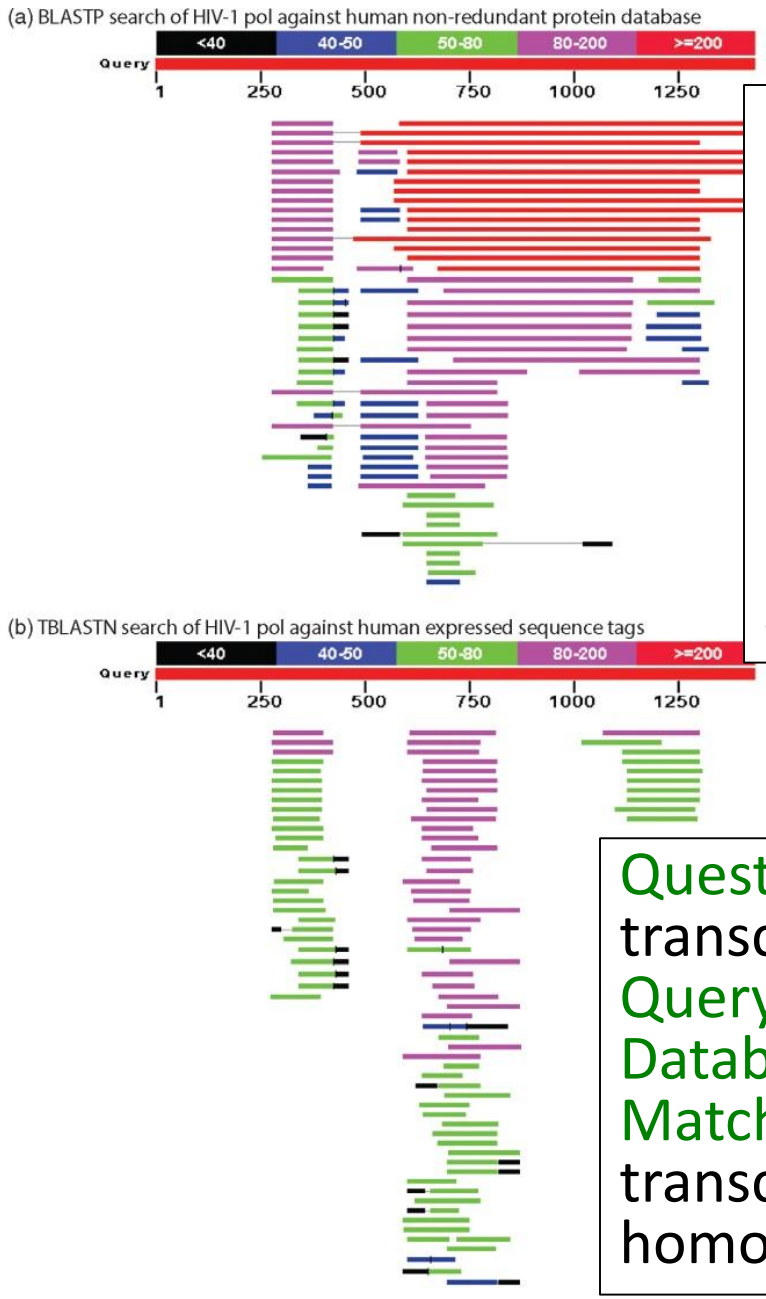
Moniliophthora perniciosa FA553 [basidiomycetes] taxid 554373

ref|XP_002387985.1|	hypothetical protein MPER_13056 [Monil...	88	7e-15

BLASTP searching HIV-1 pol against bacterial proteins



BLAST searching HIV-1 pol against human sequences



Question: are there human homologs of HIV-1 pol protein?
Query: HIV-1 Pol
Program: BLASTP
Database: human nr (nonredundant)
Matches: many human proteins share significant identity.

Question: are there human RNA transcripts corresponding to HIV-1 pol?
Query: HIV-1 Pol Program: TBLASTN
Database: human ESTs
Matches: many human genes are actively transcribed to generate transcripts homologous to HIV-1 pol.

Learning objectives

- The three phases of a BLAST search
  List, scan, extention
- Perform BLAST searches locally and at the NCBI website
- Understand how to vary optional BLAST search parameters;
- Outline strategies for BLAST searching.