EECS 4425: Introductory Computational Bioinformatics

Suprakash Datta
Email: datta [at] eecs.yorku.ca
Course page: www.cse.yorku.ca/course/4425
Office: LAS 3043

These slides have been taken from the book website

Chapter 4: 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
- discovering new genes or proteins
- discovering variants of genes or proteins
- investigating expressed sequence tags (ESTs)
- exploring protein structure and function

BLASTP search at NCBI: overview of web-based search

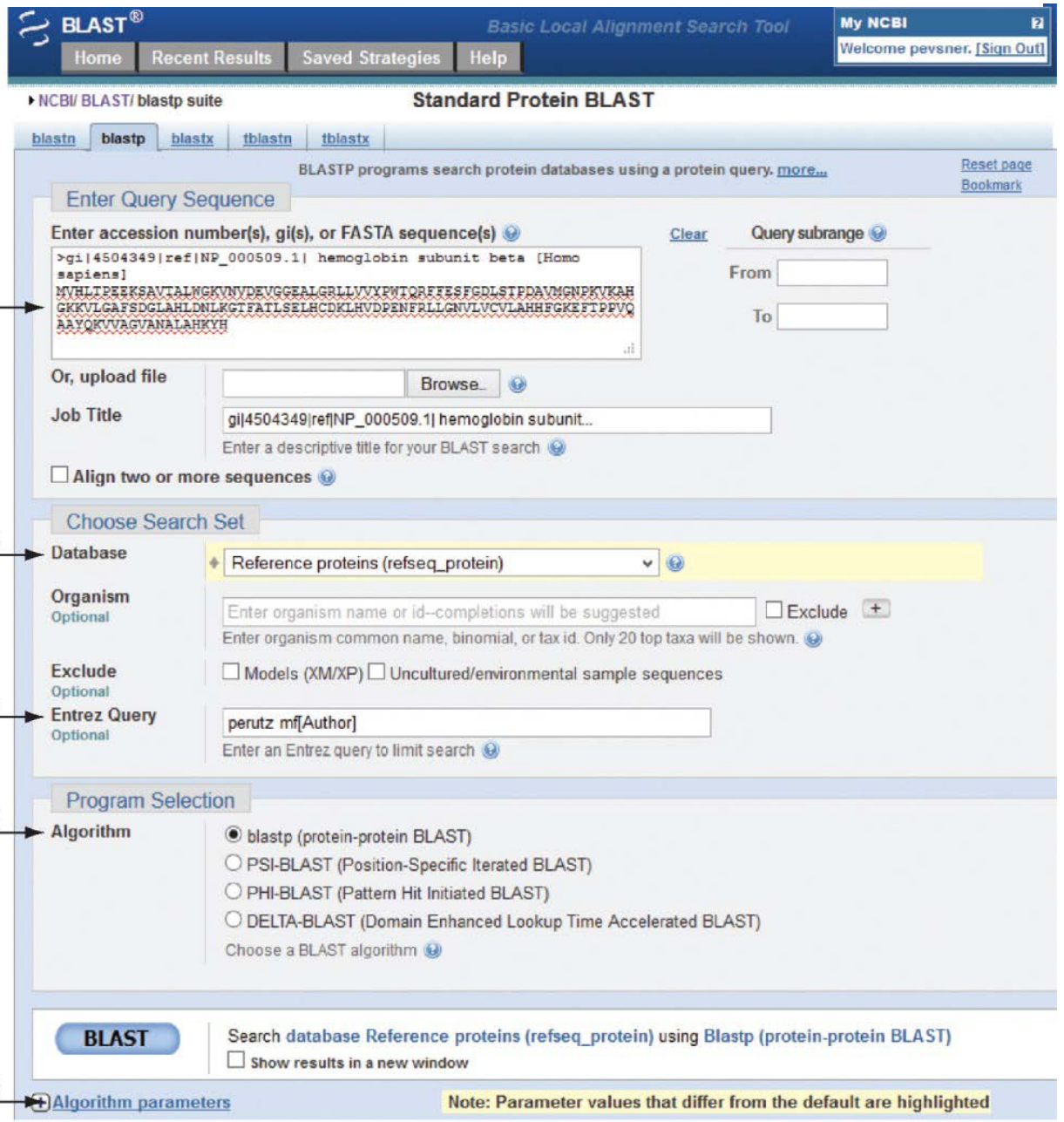
query: FASTA format or accession

database

Entrez query

algorithm

parameters

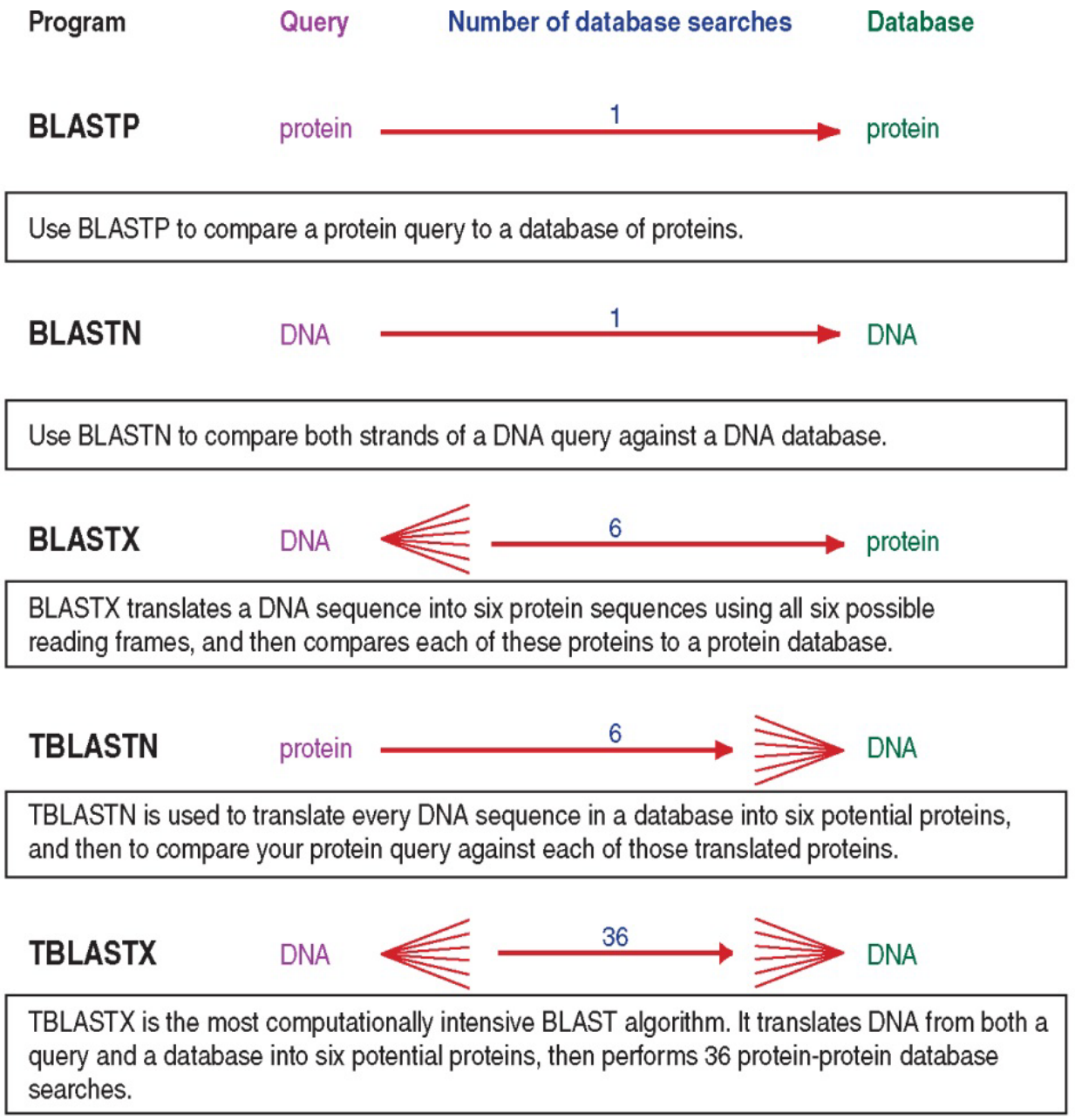


B&FG 3e
Fig. 4-1
Page 123

Step 1: Choose your sequence

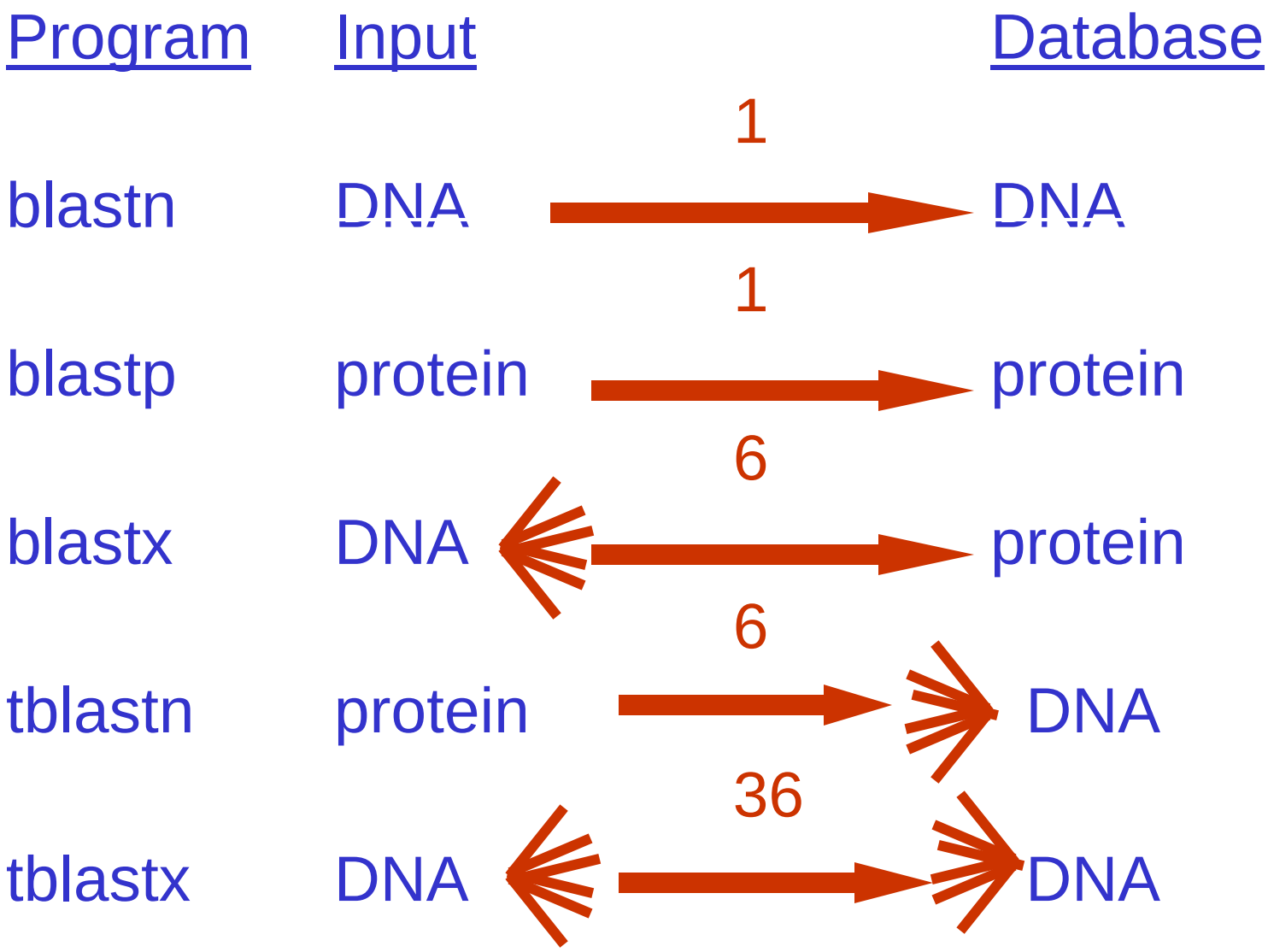
Sequence can be input in FASTA format or as accession number

BLAST step 2: choose program

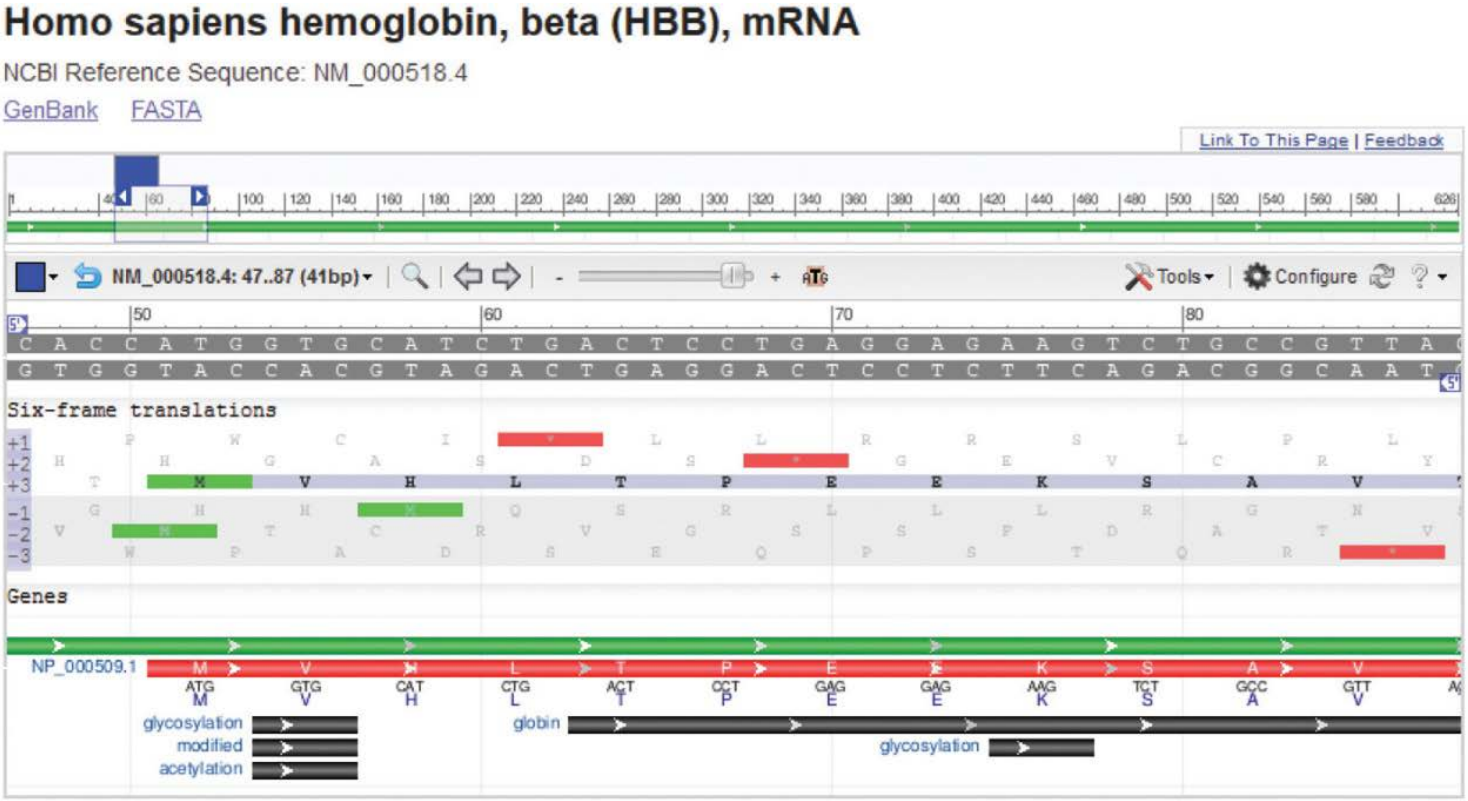


B&FG 3e
Fig. 4-2
Page 124

Step 2. Choose the BLAST program



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



DNA 1

3 forward, 3 reverse frames 2

protein 3

This image is from the 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

B&FG 3e
Table 4-1
Page 126

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

B&FG 3e
Table 4-2
Page 127

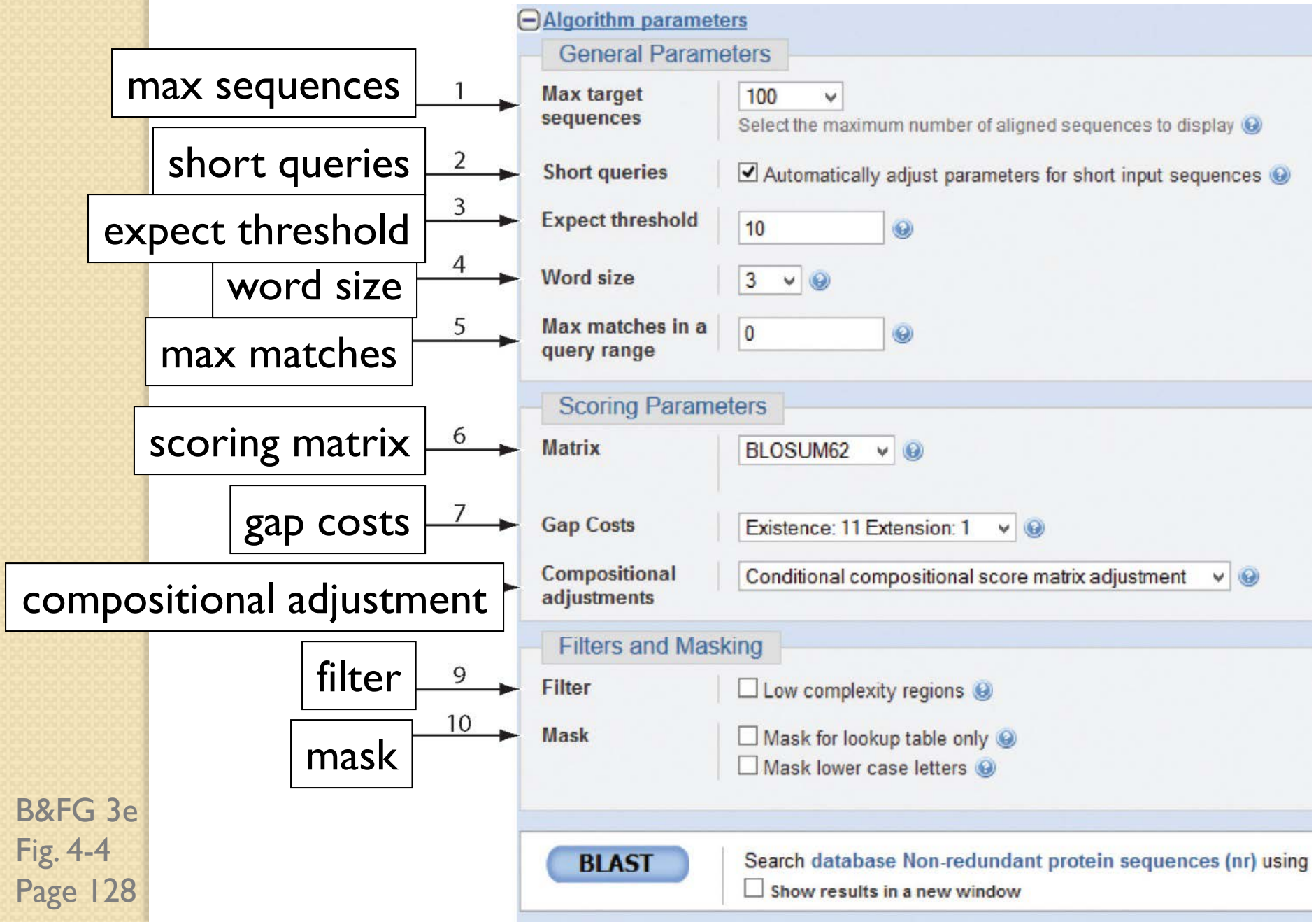
Step 4: optional parameters

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



B&FG 3e
Fig. 4-4
Page 128

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

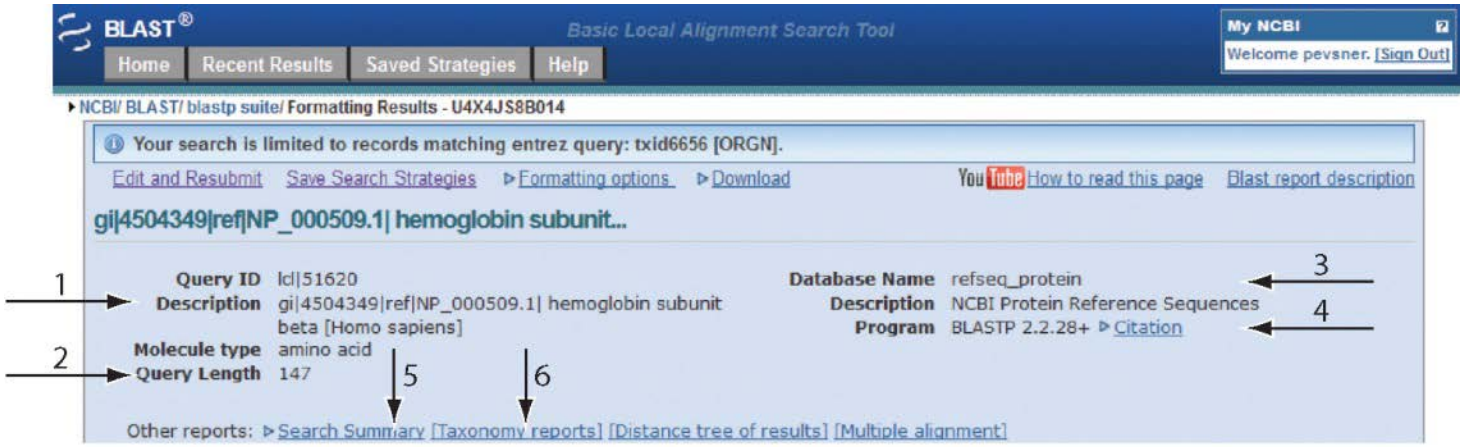
Range 1: 33 to 114 GenPept Graphics

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

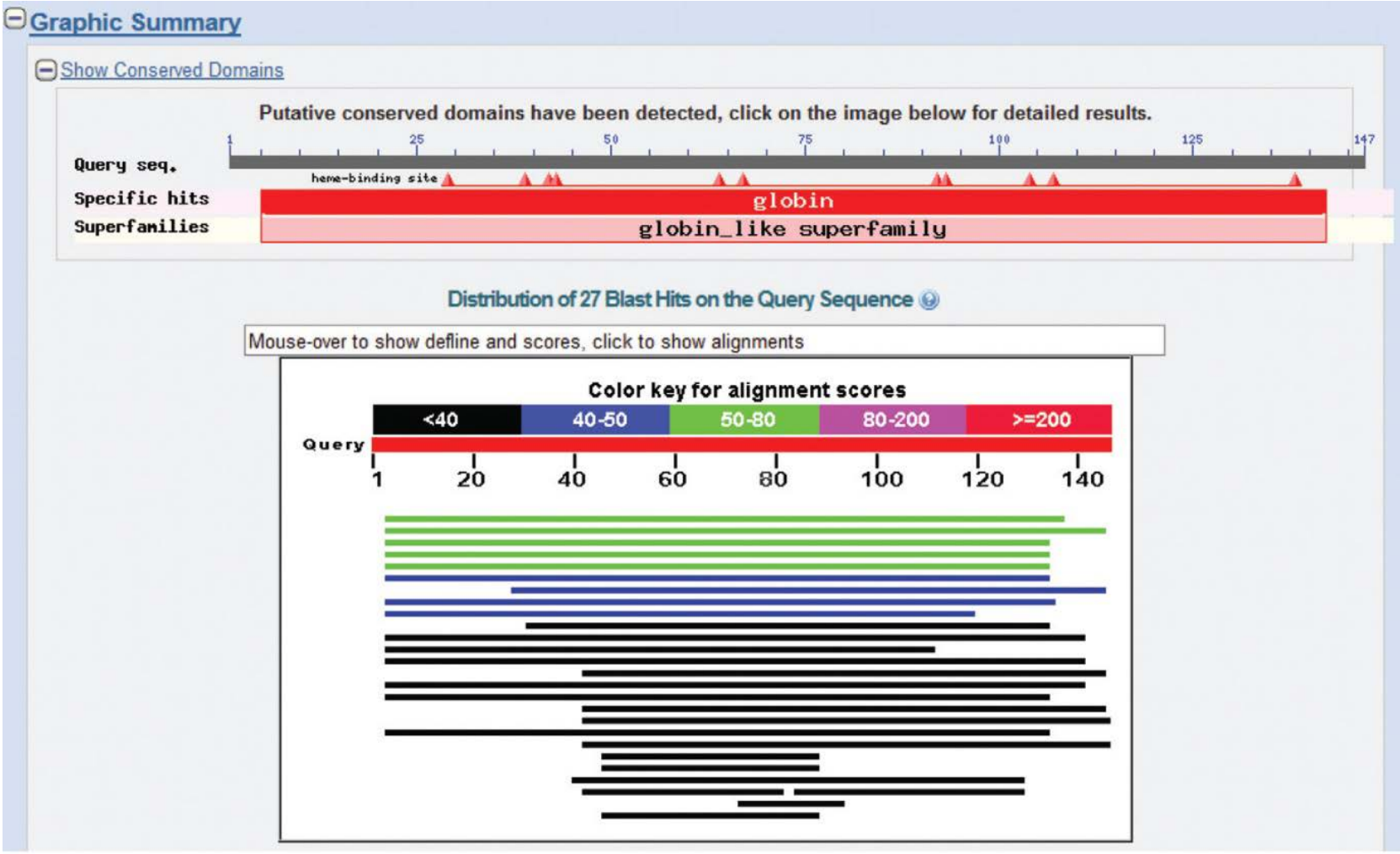
Database	
Posted date	Jun 12, 2013 10:46 AM
Number of letters	6,910,040,539 ← 4
Number of sequences	19,996,853
Entrez query	txid10090 [ORGN]

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

B&FG 3e
Fig. 4-7
Page 133

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)

B&FG 3e
Fig. 4-8
Page 134

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

Sequences producing significant alignments:

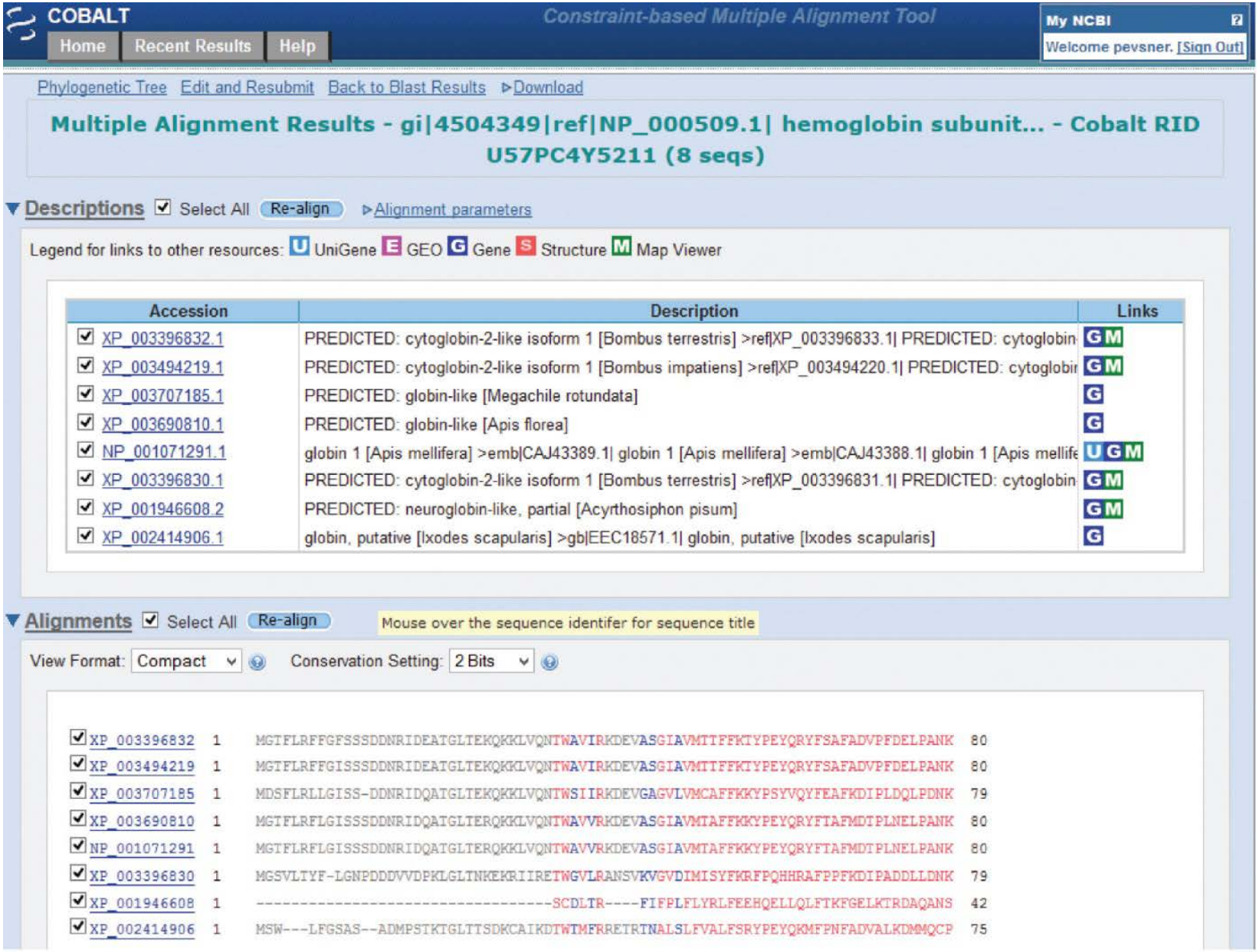
Select: All None Selected:2

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

B&FG 3e
Fig. 4-9
Page 134

BLAST output can be formatted to display multiple alignment



B&FG 3e
Fig. 4-10
Page 135

Outline

Introduction

- BLAST search steps
  - Step 1: Specifying sequence of interest
  - Step 2: Selecting BLAST program
  - Step 3: Selecting a database
  - Step 4: Selecting search parameters and formatting parameters
  - Stand-alone BLAST

BLAST algorithm uses local alignment search strategy

- BLAST algorithm parts: list, scan, extend
- BLAST algorithm: local alignment search statistics and *E* value
- Making sense of raw scores with bit scores
- BLAST algorithm: relation between *E* and *p* values

BLAST search strategies

- General concepts; principles of BLAST searching
- How to evaluate the significance of results
- How to handle too many or too few results
- BLAST searching with multidomain protein: HIV-1 Pol

Using BLAST for gene discovery: Find-a-Gene

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

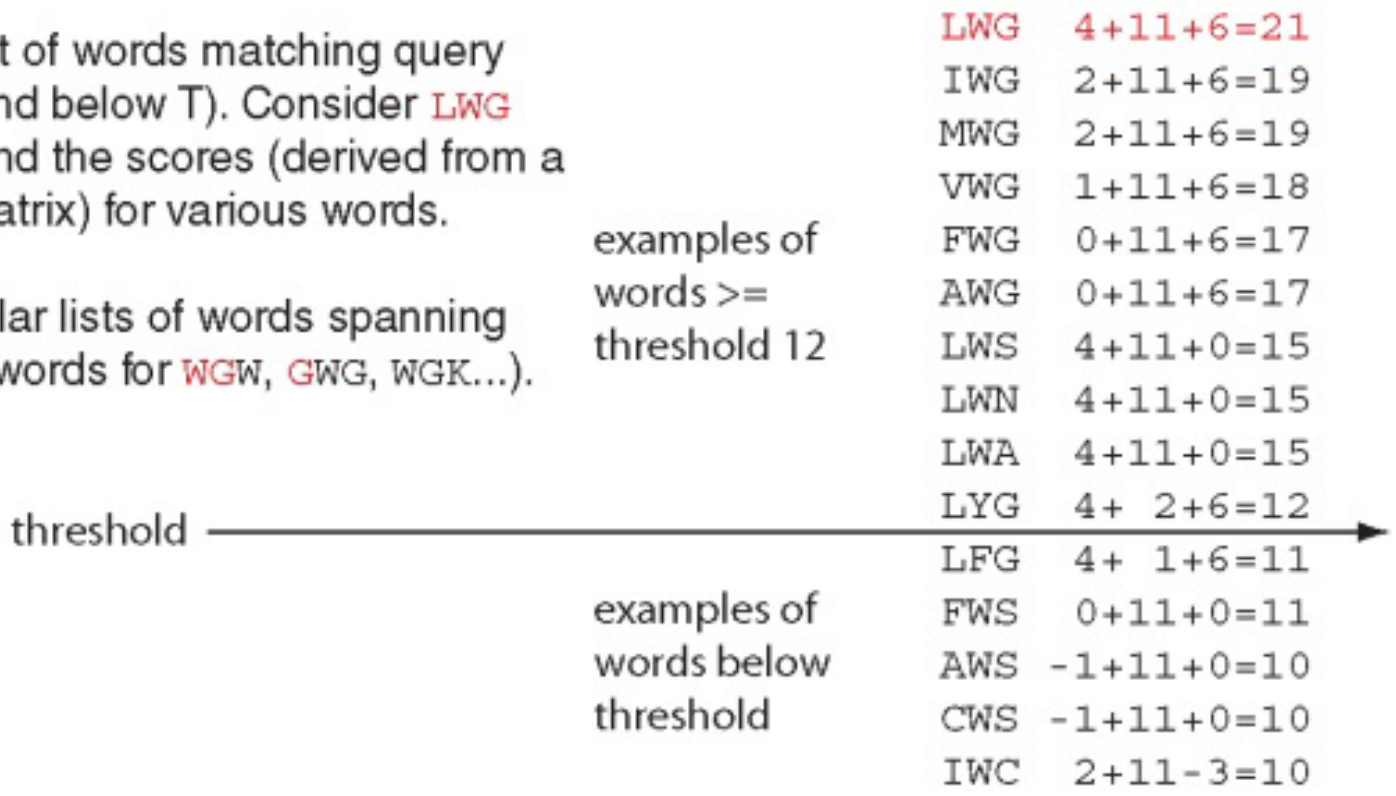
Phase 1: Setup: compile a list of words (w=3) above threshold T

- 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...).

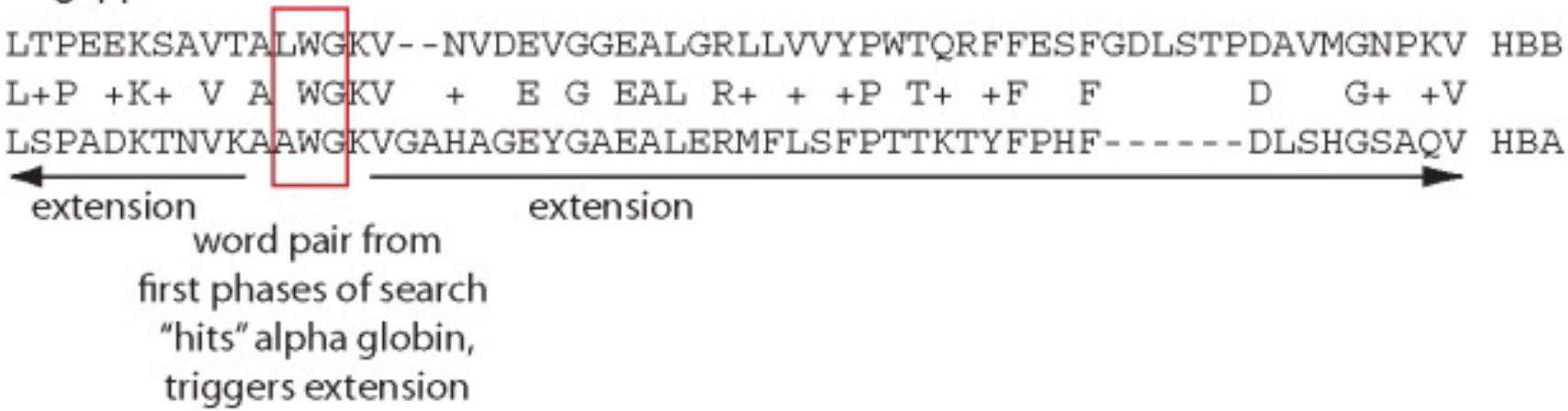


B&FG 3e
Fig. 4-12
Page 139

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



B&FG 3e
Fig. 4-12
Page 139

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

B&FG 3e
Fig. 4-12
Page 139

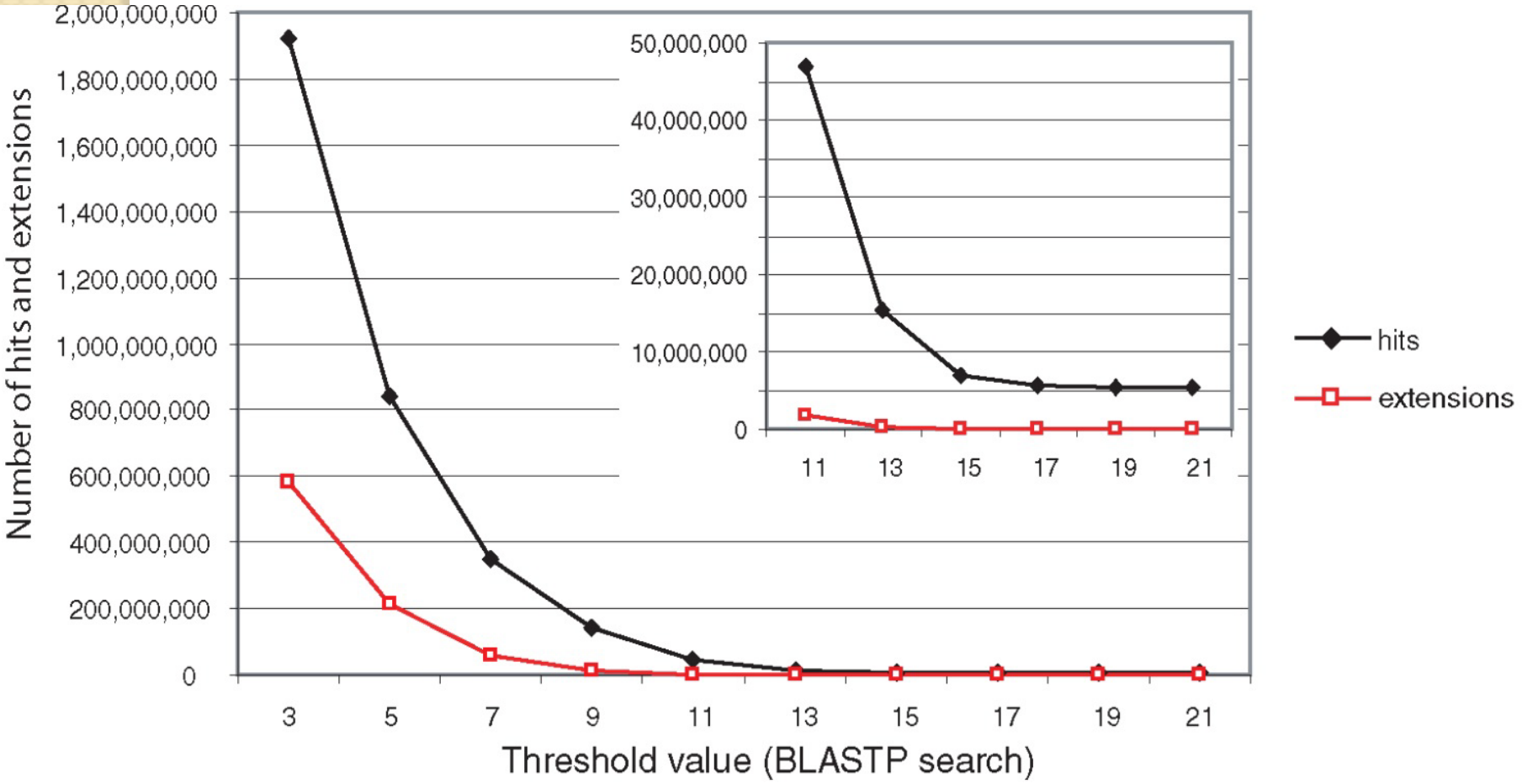
How a BLAST search works: threshold

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+.

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



B&FG 3e
Fig. 4-13
Page 140

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 (see below), 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!

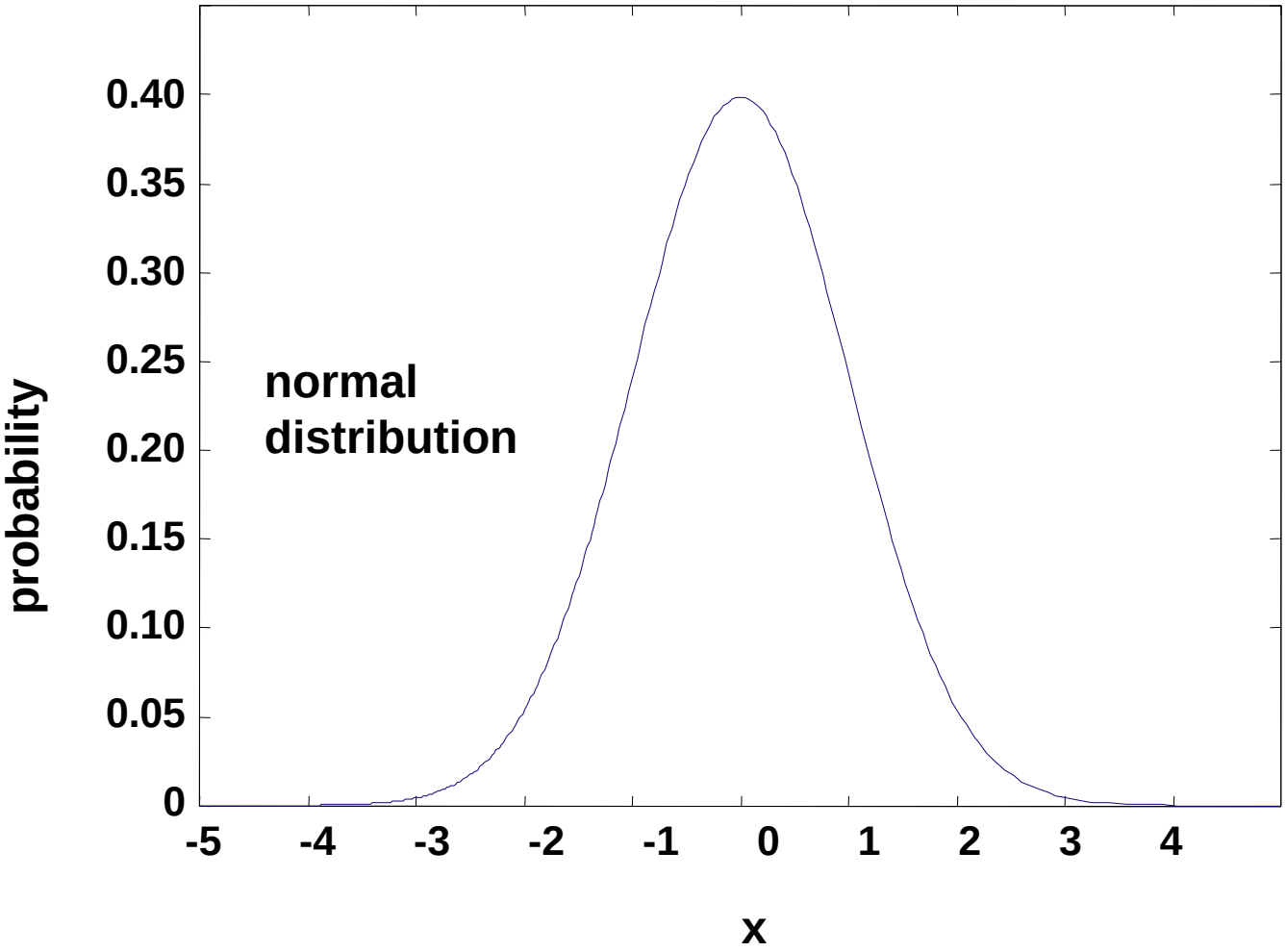
How to interpret a BLAST search: expect value

It is important to assess the statistical significance of search results.

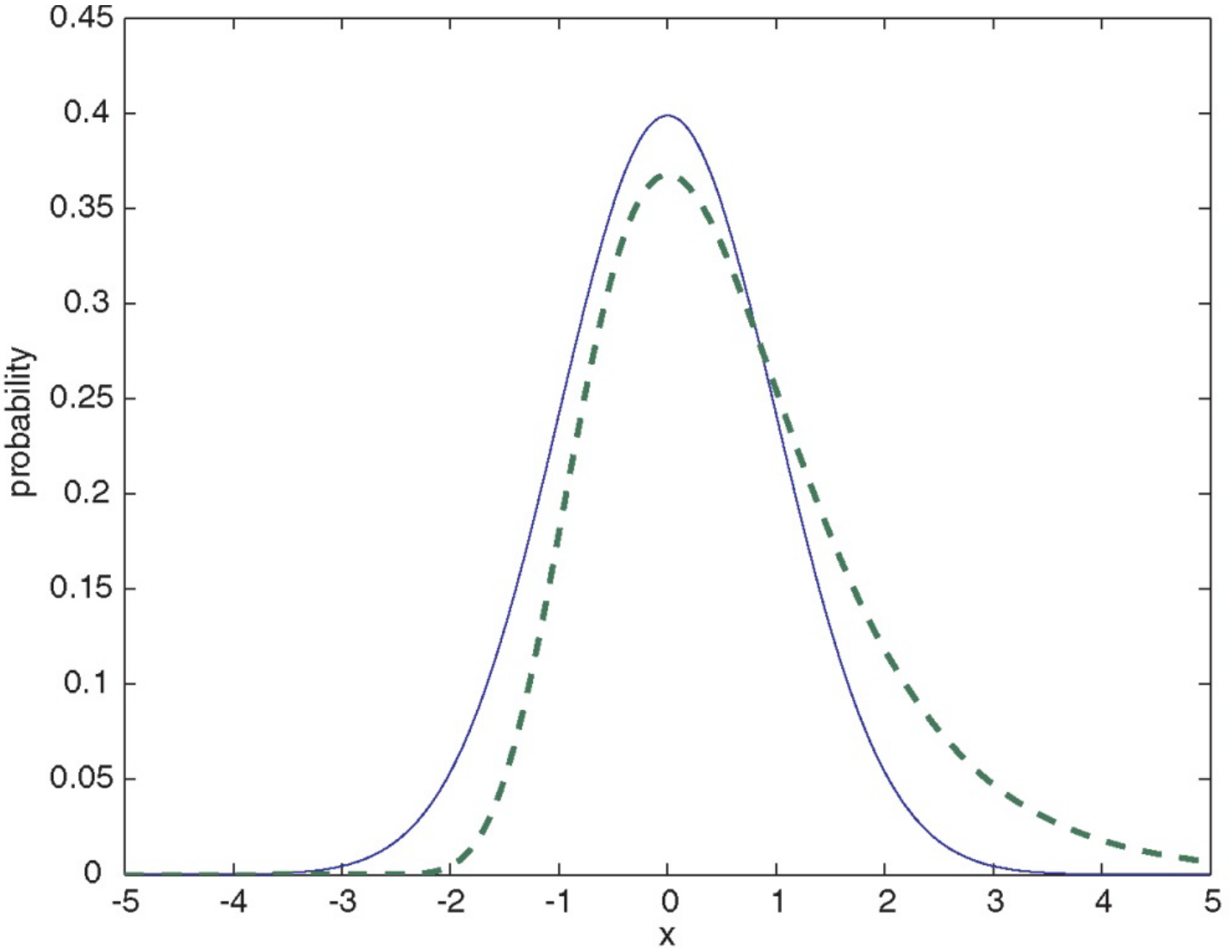
For global alignments, the statistics are poorly understood.

For local alignments (including BLAST search results), the statistics are well understood. The scores follow an extreme value distribution (EVD) rather than a normal distribution.

Normal distribution



Normal distribution (solid line) compared to extreme value distribution (dashed line): note EVD skewing to the right



B&FG 3e
Fig. 4-14
Page 141

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:

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

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

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

λ, K = Karlin Altschul statistics

Some properties of the equation $E = Kmn\ e^{-\lambda S}$

- The value of E decreases exponentially with increasing *S* (higher S values correspond to better alignments). Very high scores correspond to very low *E* values.

- The *E* value for aligning a pair of random sequences must be negative! Otherwise, long random alignments would acquire great scores

- Parameter *K* describes the search space (database).

- For E=1, one match with a similar score is expected to occur by chance. For a very much larger or smaller database, you would expect E to vary accordingly

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

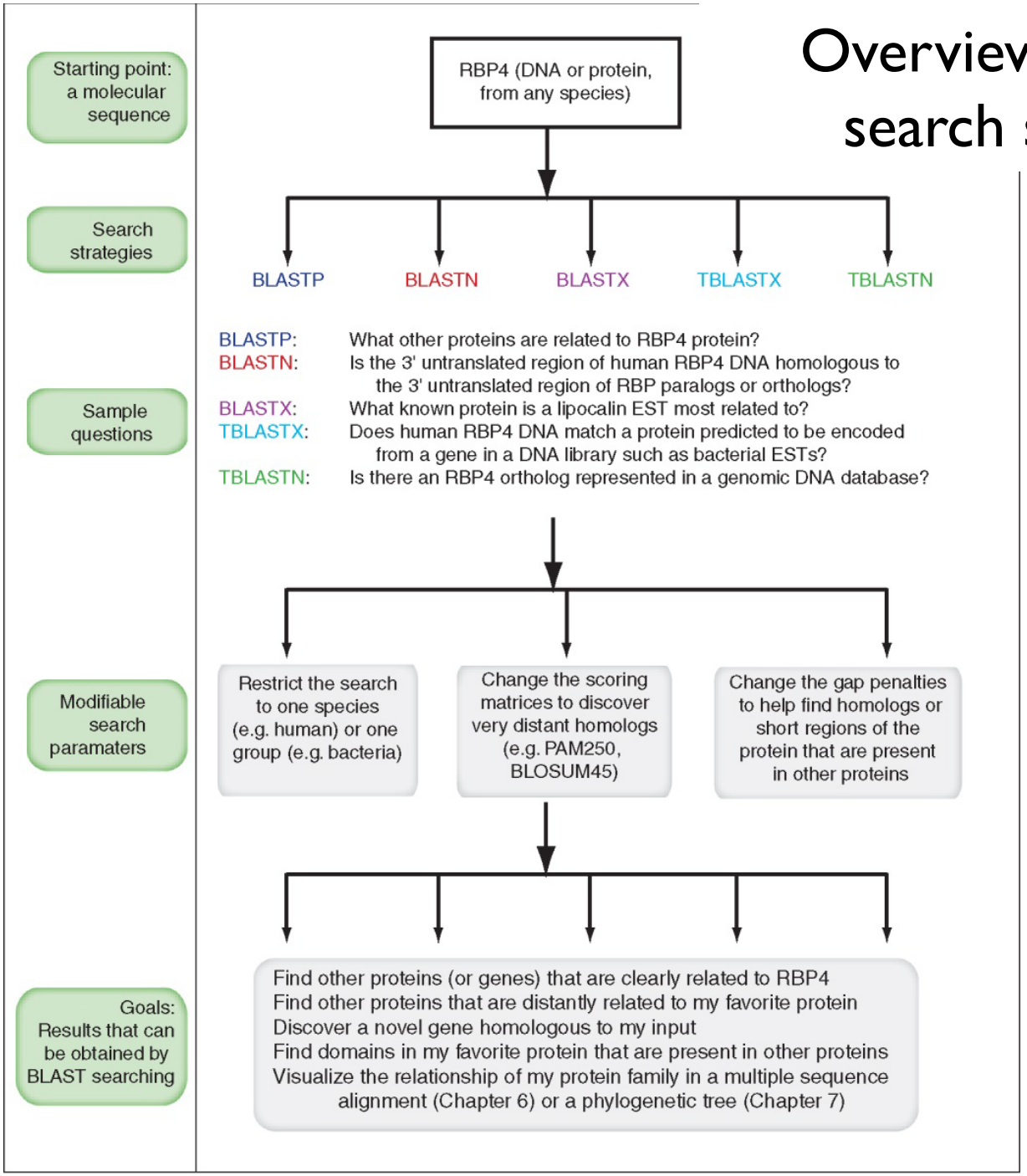
E values of about 1 to 10 are far easier to interpret than corresponding *p* values.

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

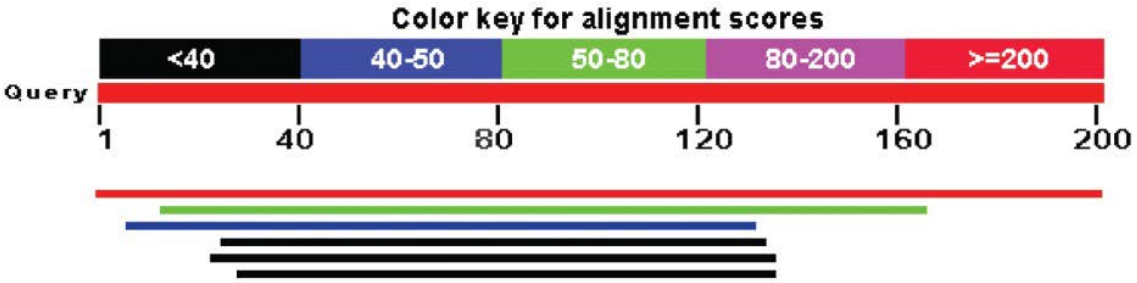
E values are comparable to *p* values, and are designed to be more convenient to interpret.

Overview of BLAST search strategies



B&FG 3e
Fig. 4-15
Page 145

(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

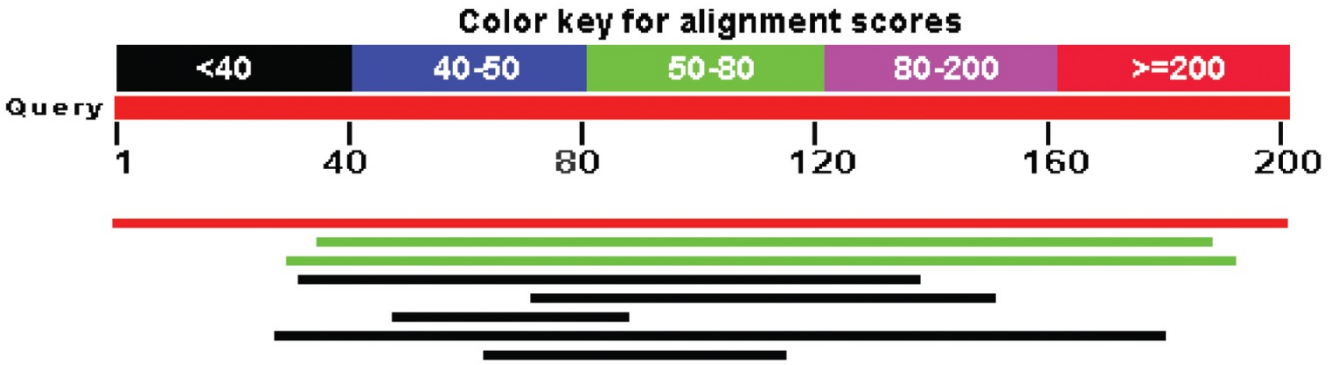
Score	Expect	Method	Identities	Positives	Gaps
32.3 bits(72)	0.18	Compositional matrix adjust.	28/114(25%)	49/114(42%)	8/114(7%)

```
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
```

"Recipricol" 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

Pairwise alignment of CG8 with non-homologous proteins

(c) P

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    TTLHVAPQGTAMAVSTFRKLD-GICWQVRQLYGDTGVLGRFLLQARDARGAVHVVVAETD  131
             T L V P+    +AV+       G+ W V Q      G    FL+Q RDA+G    V    D
Sbjct  3255  TPLPVEPRLGELAVAAVTSDSVGLSWTVAQ-----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

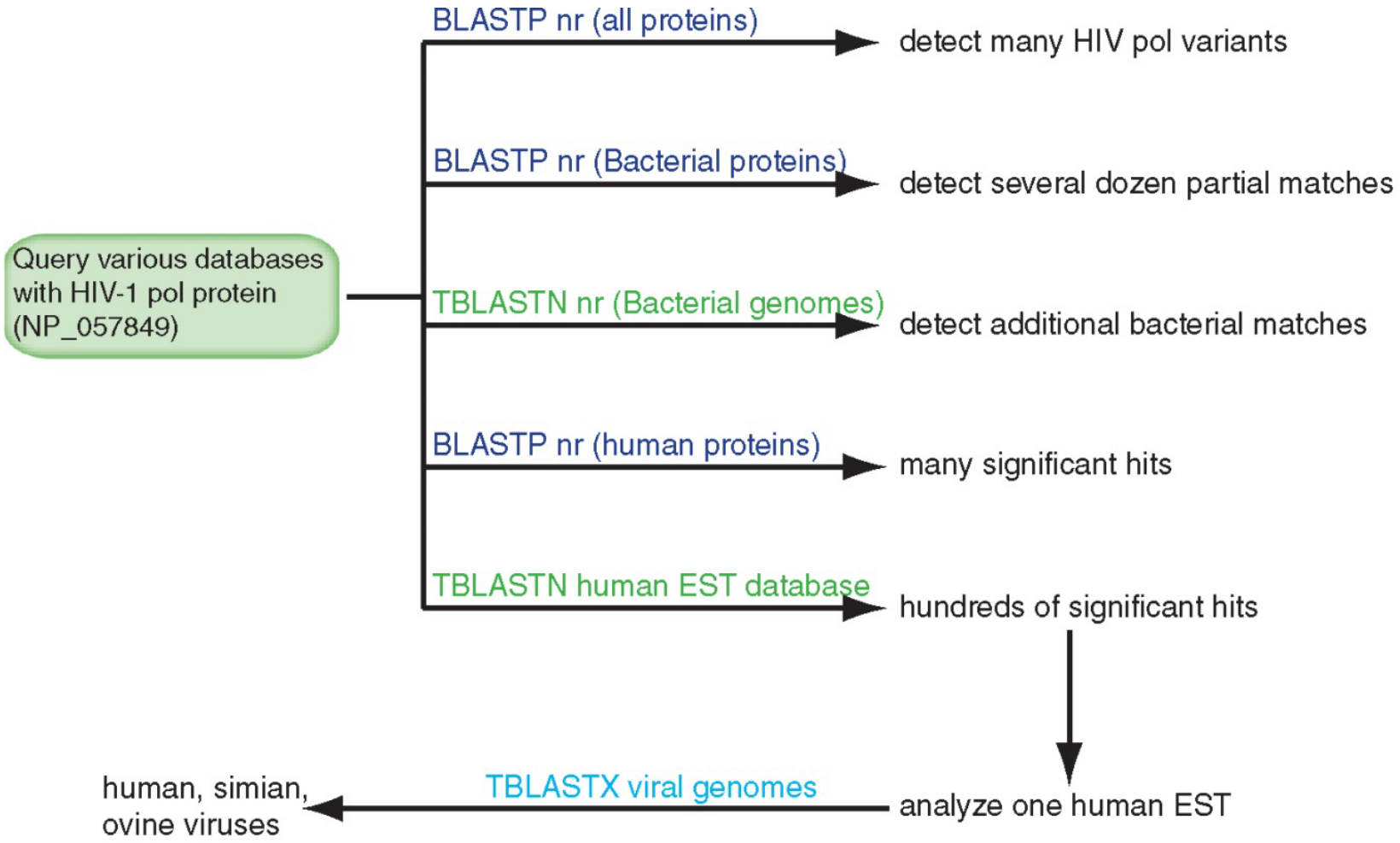
Range 1: 2323 to 2360 GenPept Graphics Next Match Previous Match

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
```

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

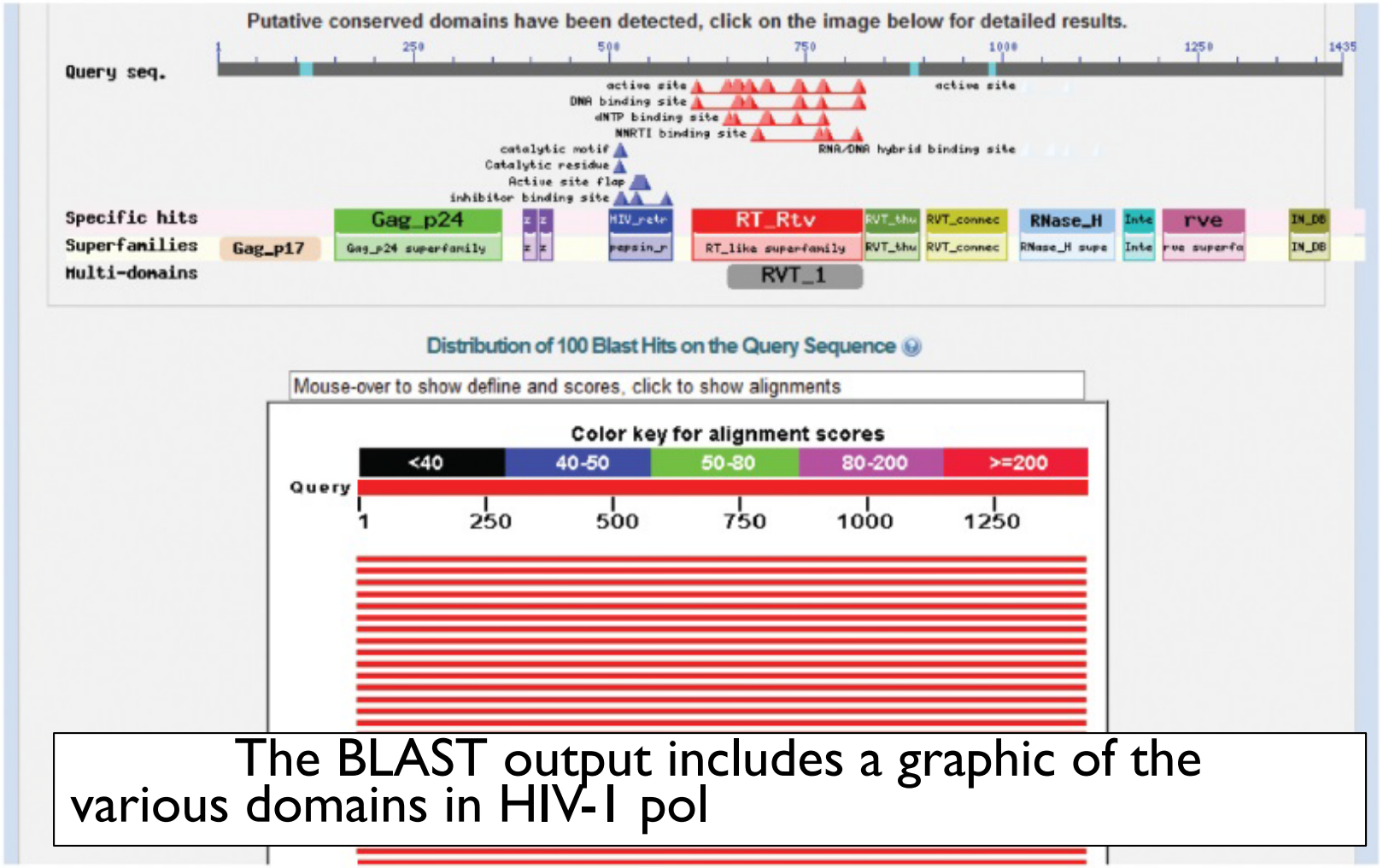
BLAST searching a multidomain protein: HIV-1 pol



B&FG 3e
Fig. 4-18
Page 151

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).

B&FG 3e
Fig. 4-19
Page 153

Taxonomy report for a BLAST searching HIV-1 pol

Human immunodeficiency virus 1 [viruses] taxid 11676

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

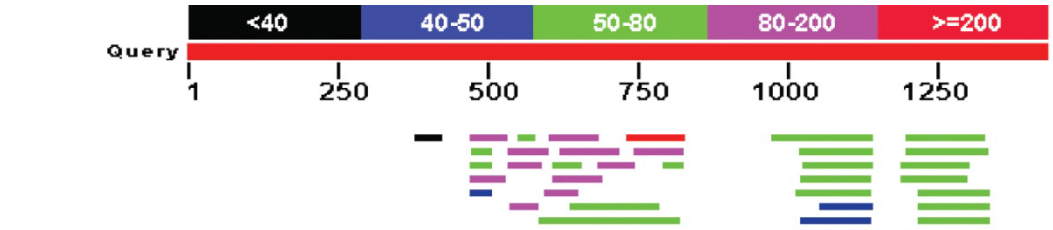
ref|XP_002421195.1|	retrovirus-related Pol polyprotein fro...	88	6e-15

Moniliophthora perniciosa FA553 [basidiomycetes] taxid 554373

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

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

BLASTP searching HIV-1 pol against bacterial proteins



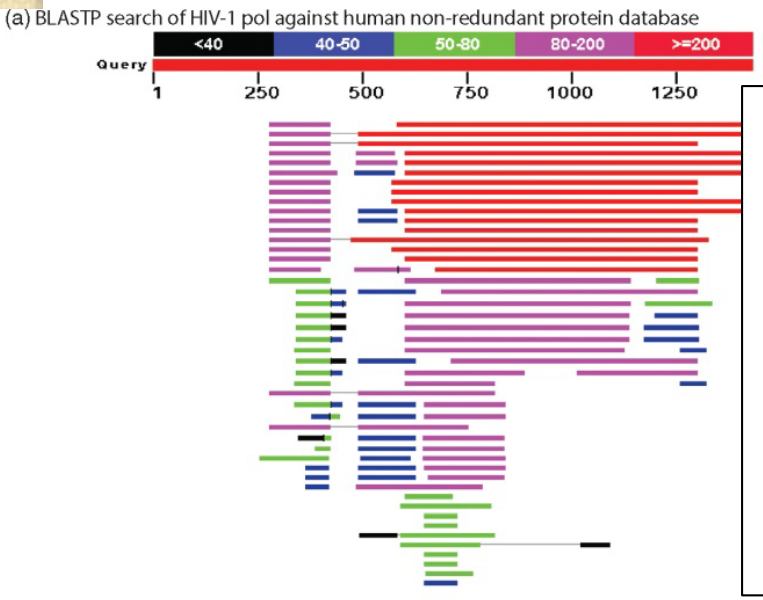
bacterial matches to HIV-1 retropepsin, reverse transcriptase domains

bacterial matches to HIV-1 ribonuclease H domain

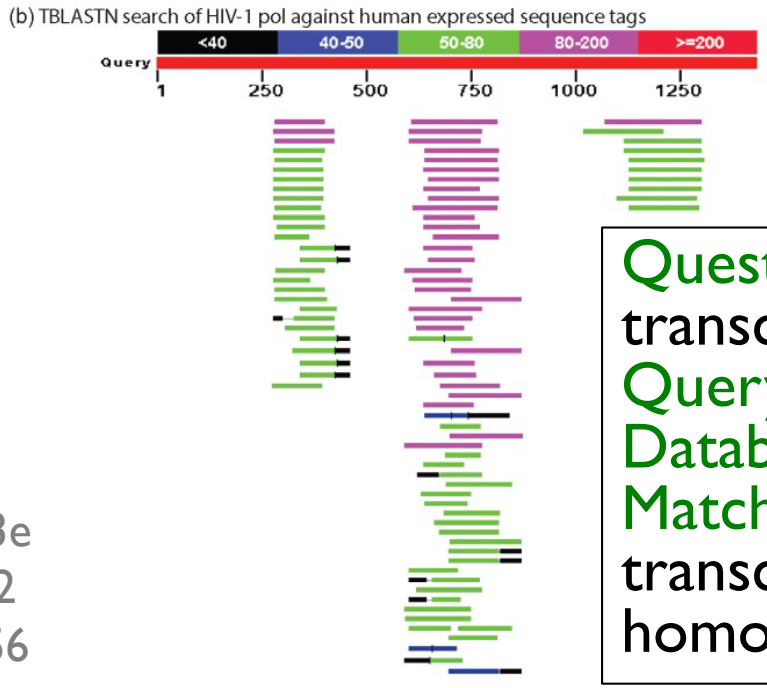
bacterial matches to HIV-1 integrase core domain

B&FG 3e
Fig. 4-21
Page 155

BLAST searching HIV-1 pol against human sequences



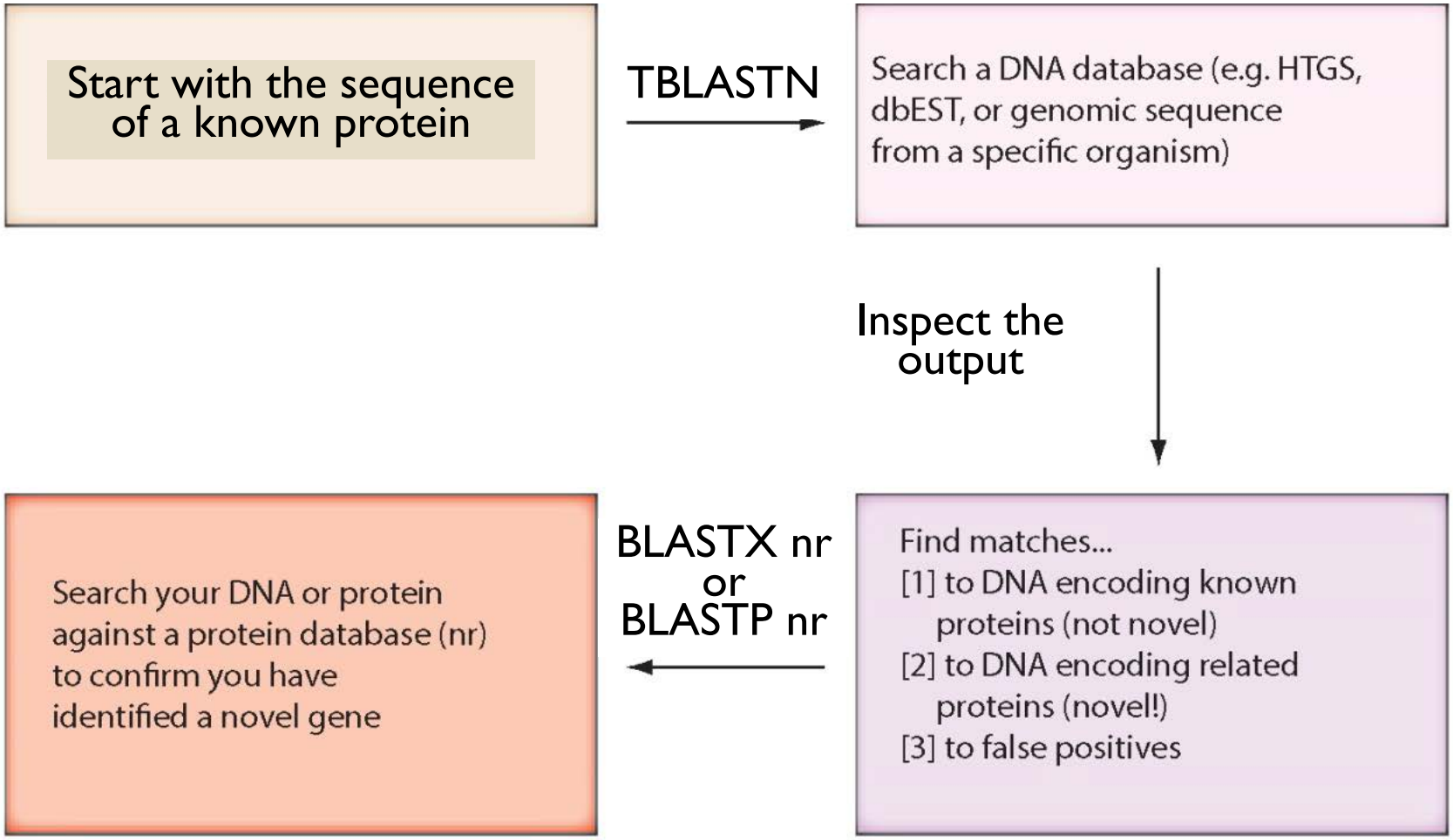
Question: are there human homologs of HIV-1 pol protein?
Query: HIV-1 Pol
Program: BLASTP
Database: human nr (nonredundant)
Matches: many human proteins share significant identity.



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.

B&FG 3e
Fig. 4-22
Page 156

"Find-a-gene project" to practice BLAST



B&FG 3e
Fig. 4-23
Page 157

"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 (Chapter 6), phylogeny (Chapter 7), and predict its protein structure (Chapter 13) and its function (Chapter 14).

Perspective

BLAST searching has emerged as an indispensable tool to analyze the relation of a DNA or protein sequence to millions or even trillions of sequences in public databases. All database search tools confront the issues of sensitivity (i.e., the ability to minimize false negative results), selectivity (i.e., the ability to minimize false positive results), and time. As the size of the public databases has grown exponentially in recent years, the BLAST tools have evolved to provide a rapid, reliable way to screen the databases. For protein searches we have focused on BLASTP. However, for most biologists performing even routine searches with a protein query, the DELTA-BLAST or HMMER programs described in Chapter 5 are strongly preferred. This is because of their more optimally constructed scoring matrices.

B&FG 3e
Fig. 2-3
Page 22