

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI Welcome pevsner. [Sign Out]

NCBI/BLAST/blastp suite Standard Protein BLAST

blastn blastp blastx tblastn tblastx

BLASTP programs search protein databases using a protein query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) [Query subrange](#)

1
 >gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
 MYHLTEESKSAVTALNGKQNYDEVGGEALGSLVVVXFWTQFFESFGDLSTPDVNGNPFVKAR
 GKQVLGAFSDGLAHLDELKGTATLSELBCRLVDRENPELLGNVLUCYLAHFGKEETIRPQ
 AAYQKVTAGVANAIAKQK

Or, upload file [Browse...](#)

Job Title
 Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

2 Database [...](#)

Organism Optional ☐ Exclude [+](#)
 Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

3 Entrez Query Optional
 Enter an Entrez query to limit search

Program Selection

4 Algorithm ☒ 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 Reference proteins (refseq_protein) using Blastp (protein-protein BLAST)
☐ Show results in a new window

5 [Algorithm parameters](#) **Note: Parameter values that differ from the default are highlighted**

FIGURE 4.1 Main page for a BLASTP search at NCBI. The sequence can be input as an accession number, GI identifier, or FASTA-formatted sequence as shown here (arrow 1). The database must be selected (arrow 2) if the default setting is not selected (as here, in which the database is set to RefSeq proteins); the choice is highlighted in yellow. The search can be restricted to a particular organism or taxonomic group, and Entrez queries can be used to further focus the search (arrow 3); here we limit the search to entries including the author Max Perutz. We discuss the BLASTP algorithm in this chapter (arrow 4), and PSI-BLAST, PHI-BLAST, and DELTA-BLAST in Chapter 5. Many of the search parameters can be modified (arrow 5).

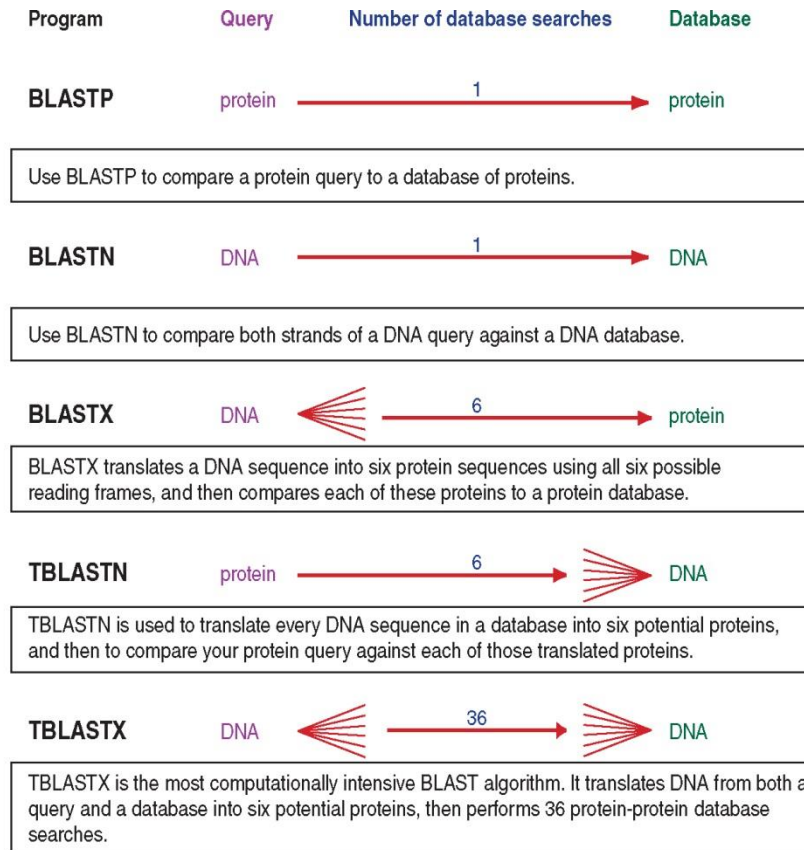


FIGURE 4.2 Overview of the five main BLAST algorithms. Note that the suffix P refers to protein (as in BLASTP), N refers to nucleotide, and X refers to a DNA query that is dynamically translated into six protein sequences. The prefix T refers to “translating,” in which a DNA database is dynamically translated into six proteins.

Homo sapiens hemoglobin, beta (HBB), mRNA

NCBI Reference Sequence: NM_000518.4

[GenBank](#) [FASTA](#)

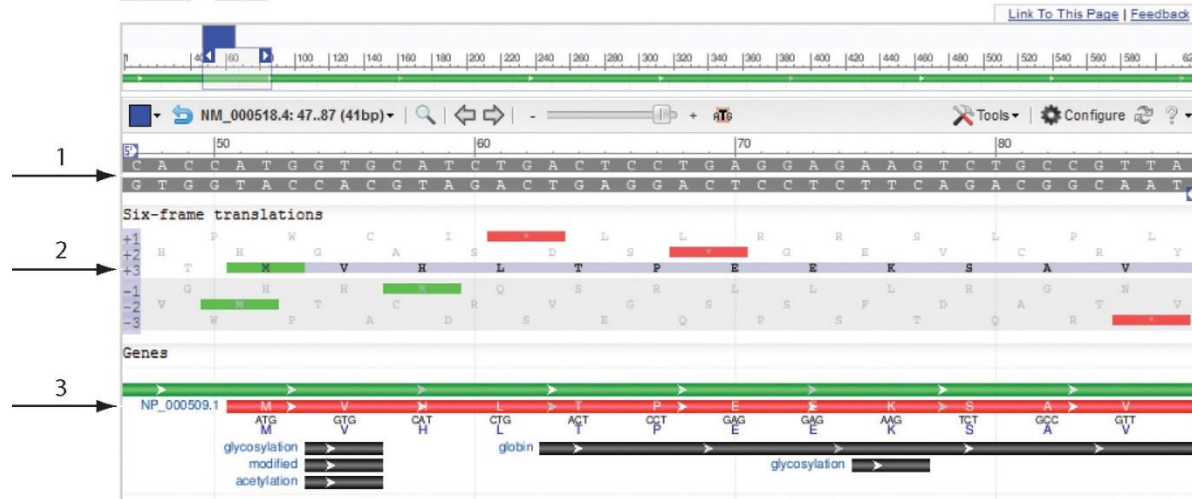


FIGURE 4.3 DNA can potentially encode six different proteins. To demonstrate this, we view the NCBI Nucleotide entry for HBB and select the “graphics” view; The two strands of DNA sequence are shown (arrow 1). In this zoomed view, only a portion of the HBB sequence is displayed. From the top strand, three potential proteins are encoded (frames +1, +2, +3) with the corresponding amino acids indicated in gray using the single-letter amino acid abbreviations. In this case, frame +3 corresponds to the frame used for translation (arrow 2). Note that frames +1 and +2 as well as frame –3 include stop codons (asterisks shaded red). The lower portion of the display includes the amino acid sequence of the corresponding protein (arrow 3) as well as the corresponding nucleotides (matching frame +3); features indicated with black shading represent a site that may be acetylated or glycosylated and a globin domain.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.

© 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.

Companion Website: www.wiley.com/go/pevsnerbioinformatics

Algorithm parameters

General Parameters

- 1 → **Max target sequences**: 100 (Select the maximum number of aligned sequences to display)
- 2 → **Short queries**: ☒ Automatically adjust parameters for short input sequences
- 3 → **Expect threshold**: 10
- 4 → **Word size**: 3
- 5 → **Max matches in a query range**: 0

Scoring Parameters

- 6 → **Matrix**: BLOSUM62
- 7 → **Gap Costs**: Existence: 11 Extension: 1
- 8 → **Compositional adjustments**: Conditional compositional score matrix adjustment

Filters and Masking

- 9 → **Filter**: ☐ Low complexity regions
- 10 → **Mask**: ☐ Mask for lookup table only
☐ Mask lower case letters

BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
☐ Show results in a new window

FIGURE 4.4 Optional BLASTP parameters. Numbered arrows refer to discussion in the text.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
 © 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
 Companion Website: www.wiley.com/go/pevsnerbioinformatics

(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--GFEDRSLERLLSDSSVQM 86				
Query 88	-----KRGIVEQCCTSIICSLYQLENYC 109				
	+ G+ ++CC C++ ++ YC				
Sbjct 87	LKTRRLRDGVFDECKLSCTMDEVLYC 114				

(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--GFEDRSLERLLSDSSVQML 87			
Query 88	-----KRGIVEQCCTSIICSLYQLENYC 109			
	+ G+ ++CC C++ ++ YC			
Sbjct 88	KTRRLRDGVFDECKLSCTMDEVLYC 114			

(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--GFEDRSLERLLSDSSVQML 87				
Query 88	-----KRGIVEQCCTSIICSLYQLENYC 109				
	+ G+ ++CC C++ ++ YC				
Sbjct 88	KTRRLRDGVFDECKLSCTMDEVLYC 114				

FIGURE 4.5 Pairwise alignments from BLASTP searches illustrating the effects of changing compositional matrices and filtering options. Human insulin (NP_000198.1) was used as a query in a BLASTP search restricted to RefSeq proteins in *Drosophila*. (a) Default settings show a match to a *Drosophila* insulin protein with a score of 31.6 bits and an *E* value of 0.05. Results are shown using (b) no compositional adjustments and (c) composition based statistics. The expect values for these three searches are indicated (red boxes).

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI
Welcome pevsner. [Sign Out]

NCBI/ BLAST/ blastp suite/ Formatting Results - U4X4JS8B014

Your search is limited to records matching entrez query: txid6656 [ORGN].

[Edit and Resubmit](#) [Save Search Strategies](#) [Formatting options](#) [Download](#) [YouTube](#) [How to read this page](#) [Blast report description](#)

gi|4504349|ref|NP_000509.1| hemoglobin subunit...

Query ID	Id 51620	Database Name	refseq_protein
Description	gi 4504349 ref NP_000509.1 hemoglobin subunit beta [Homo sapiens]	Description	NCBI Protein Reference Sequences
Molecule type	amino acid	Program	BLASTP 2.2.28+ Citation
Query Length	147		

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Multiple alignment](#)

FIGURE 4.6 Top portion of a BLAST output describes the search that was performed including the query (arrow 1), the query length (arrow 2), the database that was searched (arrow 3), and the program that was employed (BLASTP 2.2.28 in this case; arrow 4). At the bottom, additional links include a search summary showing details of the search statistics (arrow 5) and taxonomy reports of the results (arrow 6).

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
© 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
Companion Website: www.wiley.com/go/pevsnerbioinformatics

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

FIGURE 4.7 BLAST search summary. The upper portion shows the search parameters (e.g., the program that was used, the expect value (arrow 1), the scoring matrix (arrow 2), any filters that were applied, the threshold (arrow 3)). The middle portion describes the database; in this example it includes about 6.9 billion amino acid residues (arrow 4), and the output has been restricted to txid10090 (i.e., mouse). The bottom portion shows Karlin–Altschul statistics including lambda, K, and H.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
 © 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
 Companion Website: www.wiley.com/go/pevsnerbioinformatics

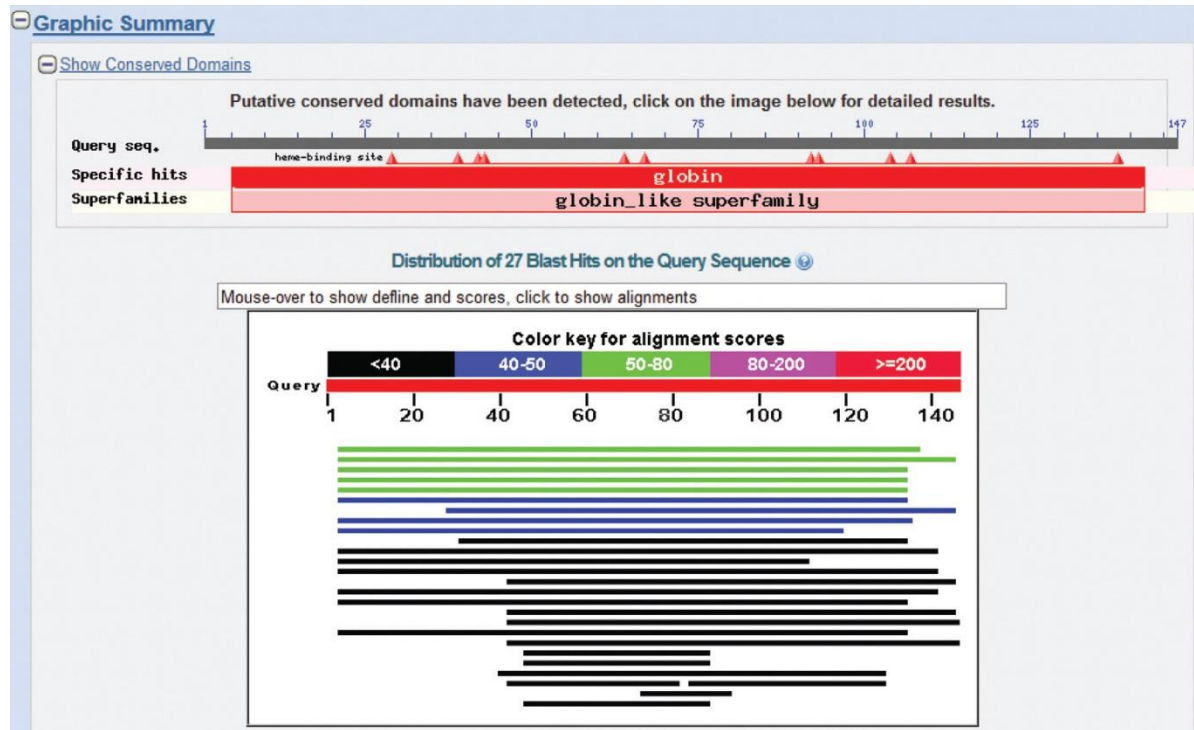


FIGURE 4.8 The graphic summary of BLAST results includes a display of conserved domains (here showing a match to the globin protein family), then a color-coded distribution of hits. Here the x axis corresponds to the length of the query (147 amino acid residues for beta globin), with each database match characterized by a color-coded score (e.g., five matches shaded green have scores of 50–80) and lengths (one of the five green database hits includes an aligned region that extends fully to the carboxy-terminus of the HBB query, while the other four do not). This graphic can be useful to summarize the regions in which database matches align to the query.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
 © 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
 Companion Website: www.wiley.com/go/pevsnerbioinformatics

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
<input checked="" type="checkbox"/>	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396833.1 PREDIC	59.7	59.7	91%	1e-10	29%	XP_003396832.1
<input checked="" type="checkbox"/>	PREDICTED: cytoglobin-2-like isoform 1 [Bombus impatiens] >ref XP_003494220.1 PREDI	58.5	58.5	97%	3e-10	28%	XP_003494219.1
<input type="checkbox"/>	PREDICTED: globin-like [Megachile rotundata]	57.8	57.8	89%	6e-10	29%	XP_003707185.1
<input type="checkbox"/>	PREDICTED: globin-like [Apis florea]	53.9	53.9	89%	1e-08	30%	XP_003690810.1
<input type="checkbox"/>	globin 1 [Apis mellifera]	52.8	52.8	89%	4e-08	30%	NP_001071291.1
<input type="checkbox"/>	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396831.1 PREDIC	45.1	45.1	89%	2e-05	26%	XP_003396830.1
<input type="checkbox"/>	PREDICTED: neuroglobin-like, partial [Acyrtosiphon pisum]	42.4	42.4	80%	2e-04	23%	XP_001946608.2
<input type="checkbox"/>	globin, putative [Ixodes scapularis]	42.7	42.7	90%	2e-04	25%	XP_002414906.1

FIGURE 4.9 A typical BLASTP output includes a list of database sequences that match the query. Links are provided to that database entry (e.g., an NCBI Protein entry) and to the pairwise alignment to the query. The bit score and *E* value for each alignment are also provided. Note that the best matches at the top of the list have large bit scores and small *E* values.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
 © 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
 Companion Website: www.wiley.com/go/pevsnerbioinformatics

COBALT Constraint-based Multiple Alignment Tool My NCBI Welcome pevsner. [Sign Out]

Home Recent Results Help

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
<input checked="" type="checkbox"/> XP_003396832.1	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396833.1 PREDICTED: cytoglobin	GM
<input checked="" type="checkbox"/> XP_003494219.1	PREDICTED: cytoglobin-2-like isoform 1 [Bombus impatiens] >ref XP_003494220.1 PREDICTED: cytoglobin	GM
<input checked="" type="checkbox"/> XP_003707185.1	PREDICTED: globin-like [Megachile rotundata]	G
<input checked="" type="checkbox"/> XP_003690810.1	PREDICTED: globin-like [Apis florea]	G
<input checked="" type="checkbox"/> NP_001071291.1	globin 1 [Apis mellifera] >emb CAJ43389.1 globin 1 [Apis mellifera] >emb CAJ43388.1 globin 1 [Apis mellifera]	U GM
<input checked="" type="checkbox"/> XP_003396830.1	PREDICTED: cytoglobin-2-like isoform 1 [Bombus terrestris] >ref XP_003396831.1 PREDICTED: cytoglobin	GM
<input checked="" type="checkbox"/> XP_001946608.2	PREDICTED: neuroglobin-like, partial [Acyrtosiphon pisum]	GM
<input checked="" type="checkbox"/> XP_002414906.1	globin, putative [Ixodes scapularis] >gb EEC18571.1 globin, putative [Ixodes scapularis]	G

▼ **Alignments** ☒ Select All **Re-align** Mouse over the sequence identifier for sequence title

View Format: **Compact** Conservation Setting: **2 Bits**

<input checked="" type="checkbox"/> XP_003396832	1	MGTFLRFFGSSSDNRIDEATGLTEKQKGLVQNTWAVIRKDEVASGIAMTTFFKTYPEYQRYFSADFVFPDELPA	80
<input checked="" type="checkbox"/> XP_003494219	1	MGTFLRFFGISSSDNRIDEATGLTEKQKGLVQNTWAVIRKDEVASGIAMTTFFKTYPEYQRYFSADFVFPDELPA	80
<input checked="" type="checkbox"/> XP_003707185	1	MDSFLRLGIS--DNRIDQATGLTEKQKGLVQNTWSIIRKDEVGAGVLVMCAFFKKYPSYVQYFEAFKDIPLDQLPDN	79
<input checked="" type="checkbox"/> XP_003690810	1	MGTFLRFLGISSSDNRIDQATGLTEKQKGLVQNTWAVVRKDEVASGIAMTTFKKYPEYQRYFTAFMDTFLNELPA	80
<input checked="" type="checkbox"/> NP_001071291	1	MGTFLRFLGISSSDNRIDQATGLTEKQKGLVQNTWAVVRKDEVASGIAMTTFKKYPEYQRYFTAFMDTFLNELPA	80
<input checked="" type="checkbox"/> XP_003396830	1	MGSVLIYF-LGNPDVVDPKGLTNKRIIRETWGLRANSVKGVDMISYFKRFPQHRAFFPFKDI PADDLLDNK	79
<input checked="" type="checkbox"/> XP_001946608	1	-----SCDLTR-----FIFPLFLYRLFEHQELLQLFTKFGELKTRDAQANS	42
<input checked="" type="checkbox"/> XP_002414906	1	MSW---LFGSAS--ADMPSTKIGLTISDKCAIKDTWIMFRRETRINALSLFVALFSRYPEYQKMFNFAVALKDMNQCP	75

FIGURE 4.10 The lower part of a BLASTP search (or other BLAST family search) consists of a series of pairwise sequence alignments such as those shown in **Figure 4.5**. Using the reformat option, the results can be displayed as a multiple sequence alignment as shown here for a group of globins. Other output format options are available, allowing the user to inspect regions of similarity as well as divergent regions within protein families.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
 © 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
 Companion Website: www.wiley.com/go/pevsnerbioinformatics

Sequence ID: [ref|NM_005330.3|](#) Length: 816 Number of Matches: 1

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

[illegible]

- [Gene](#) - associated gene details
- [UniGene](#) - clustered expressed sequence tags
- [Map Viewer](#) - aligned genomic context
- [GEO Profiles](#) - microarray expression data

FIGURE 4.11 For BLASTN searches, the coding sequence (CDS) option in the reformat page allows the amino acid sequence of the coding regions of the query and the subject (i.e., the database match) to be displayed. Here, human beta globin DNA (NM_000518) was used as a query, and a match to the closely related epsilon 1 globin is shown. The corresponding protein sequences are provided, including mismatches in purple.

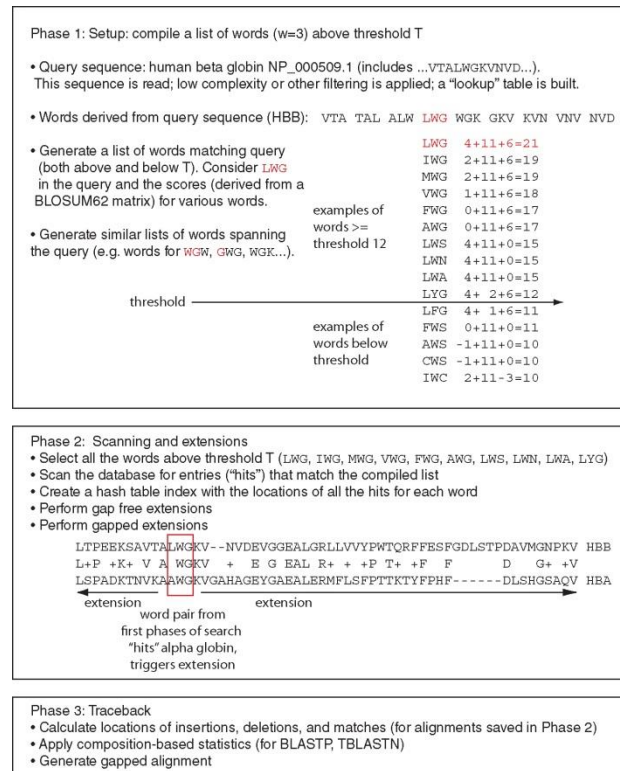


FIGURE 4.12 Schematic of the original BLAST algorithm. In the setup phase a query sequence (such as human beta globin) is analyzed with a given word size (e.g., $w = 3$), and a list of words is compiled having a threshold score (e.g., $T = 11$). Several possible words derived from the query sequence are listed in the figure (from **LWG** to **IWC**); in a BLAST search there are 8000 words compiled for $w = 3$. For a given word, such as the portion of the query sequence consisting of **LWG**, a list of words is compiled with scores greater than or equal to some threshold T (e.g., 12). In this example, 15 words are shown along with their scores from a BLOSUM62 matrix; 10 of these are above the threshold, and 5 are below. In phase 2, a database is scanned to find entries that match the compiled word list. Ungapped and gapped extensions are performed, although (to increase efficiency) positions are not saved. 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. In phase 3, a trace-back is performed and locations of insertions and deletions are recorded. Note that in this particular example the word pair that triggers the extension step is not an exact match (see boxed residues **LWG** aligned to **AWG**). The main idea of the threshold T for protein searches is to also allow both exact and related but nonexact word hits to trigger an extension. For nucleotide BLASTN searches, exact matches are required rather than words above a threshold.

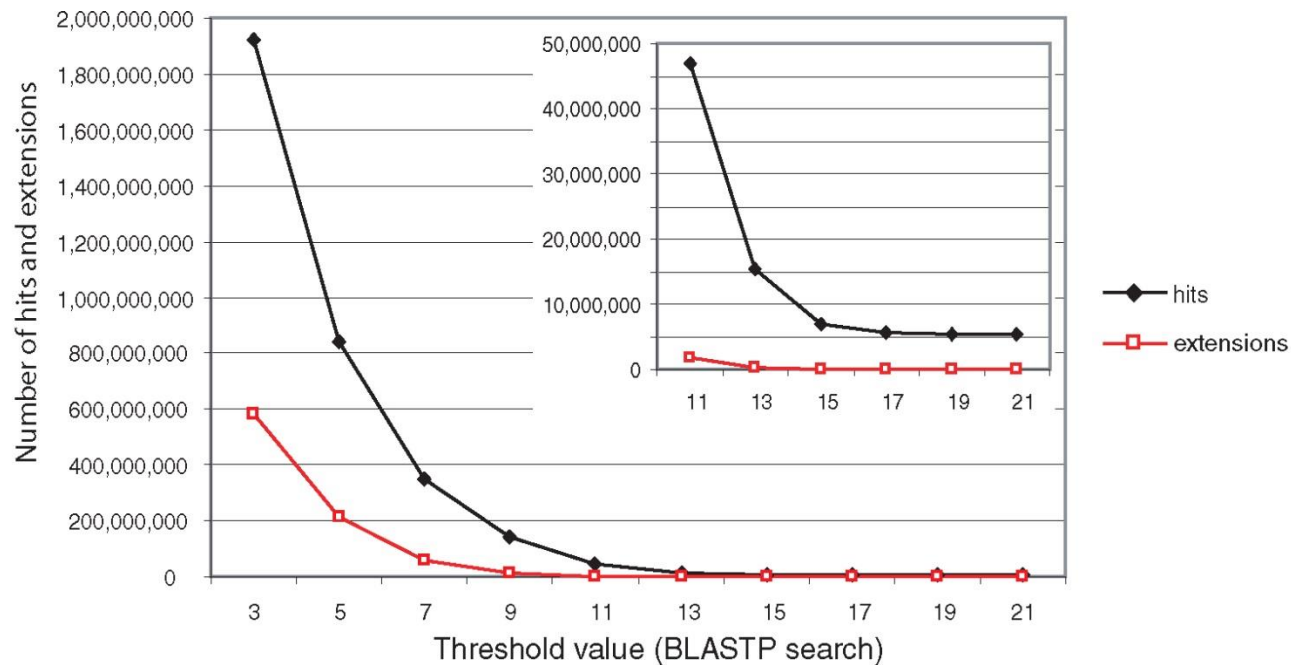


FIGURE 4.13 The effect of varying the threshold (x axis) on the number of database hits (black line) and extensions (red line). BLASTP searches were performed using human beta globin as a query.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
 © 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
 Companion Website: www.wiley.com/go/pevsnerbioinformatics

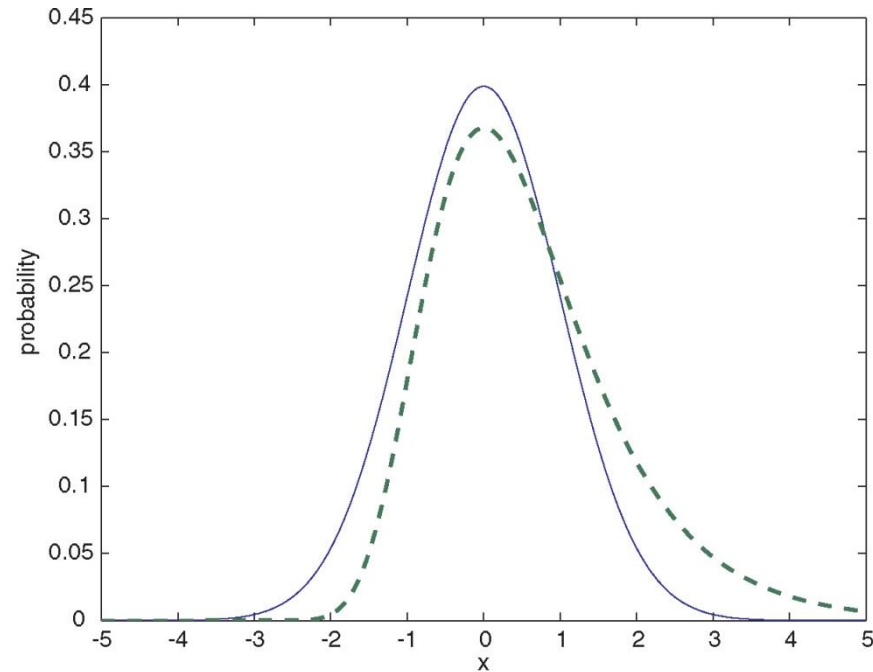


FIGURE 4.14 Normal distribution (solid line) is compared to the extreme value distribution (dotted line). Comparing a query sequence to a set of uniform-length random sequences usually generates scores that fit an extreme-value distribution (rather than a normal distribution). The area under each curve is 1. For the normal distribution, the mean (μ) is centered at zero, and the probability Z of obtaining some score x is given in terms of units of standard deviation (σ) from x to the mean: $Z = (x - \mu)/\sigma$. In contrast to the normal distribution, the extreme value distribution is asymmetric with a skew to the right. It is fit to the equation $f(x) = (e^{-x})(e^{-e^{-x}})$. The shape of the extreme value distribution is determined by the characteristic value u and the decay constant λ ($u = 0$; $\lambda = 1$).

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
 © 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
 Companion Website: www.wiley.com/go/pevsnerbioinformatics

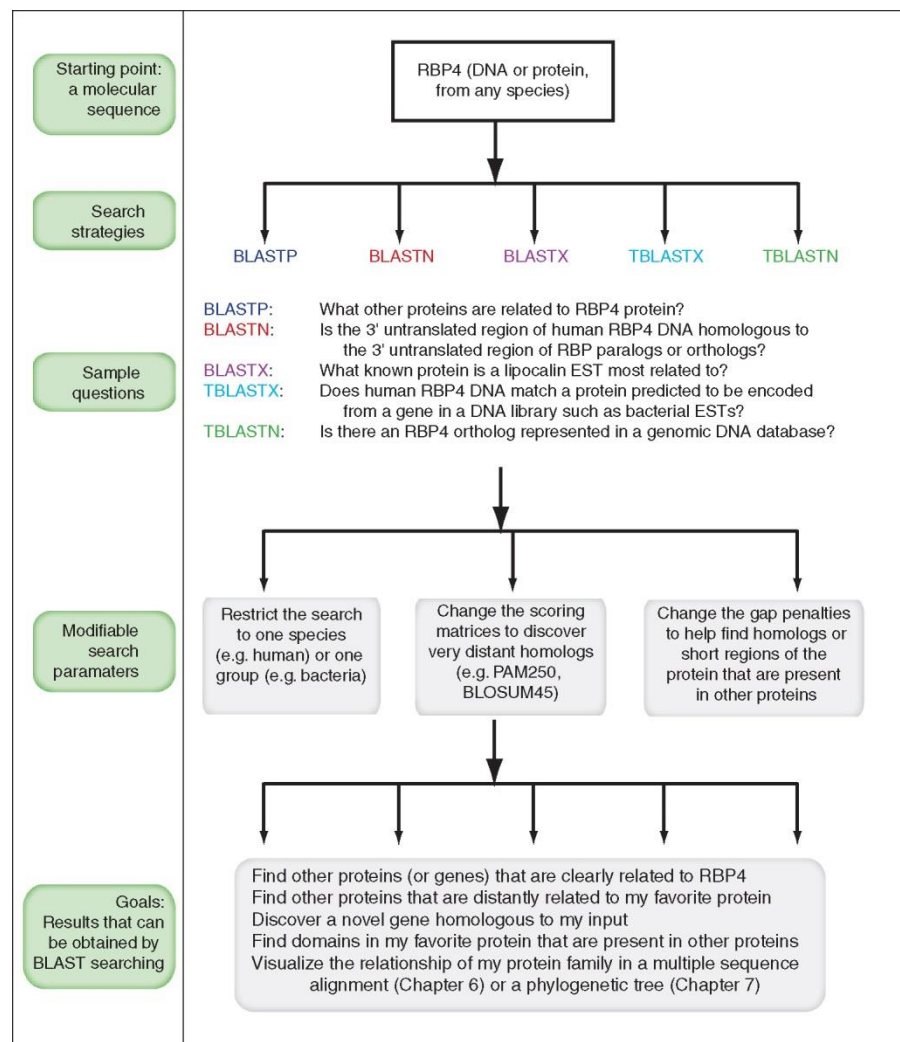
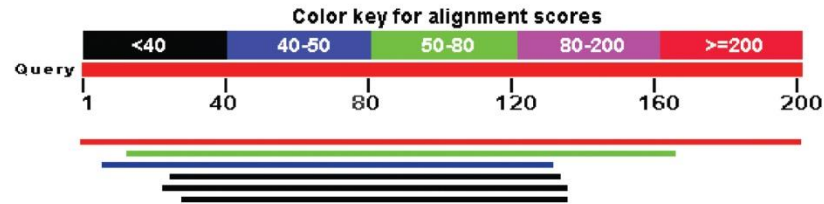


FIGURE 4.15 Overview of BLAST searching strategies. There are many hundreds of questions that can be addressed with BLAST searching, from characterizing the genome of an organism to evaluating the sequence variation in a single gene.

(a) Graphical overview



(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	Accession
<input checked="" type="checkbox"/>	retinol-binding protein 4 precursor [Homo sapiens]	420	420	100%	1e-150	100% NP_006735.2
<input checked="" type="checkbox"/>	apolipoprotein D precursor [Homo sapiens]	55.5	55.5	76%	1e-09	28% NP_001638.1
<input checked="" type="checkbox"/>	glycodelin precursor [Homo sapiens] > ref NP_002562.2 glycodelin precursor [Homo sapiens]	40.0	40.0	62%	5e-04	26% NP_001018059.1
<input checked="" type="checkbox"/>	protein AMBP preproprotein [Homo sapiens]	35.0	35.0	54%	0.034	23% NP_001624.1
<input checked="" type="checkbox"/>	complement component C8 gamma chain precursor [Homo sapiens]	32.3	32.3	56%	0.18	25% NP_000597.2
<input checked="" type="checkbox"/>	lipocalin-15 precursor [Homo sapiens]	28.5	28.5	53%	3.4	23% NP_976222.1

(c) Pairwise alignment of RBP4 and C8G

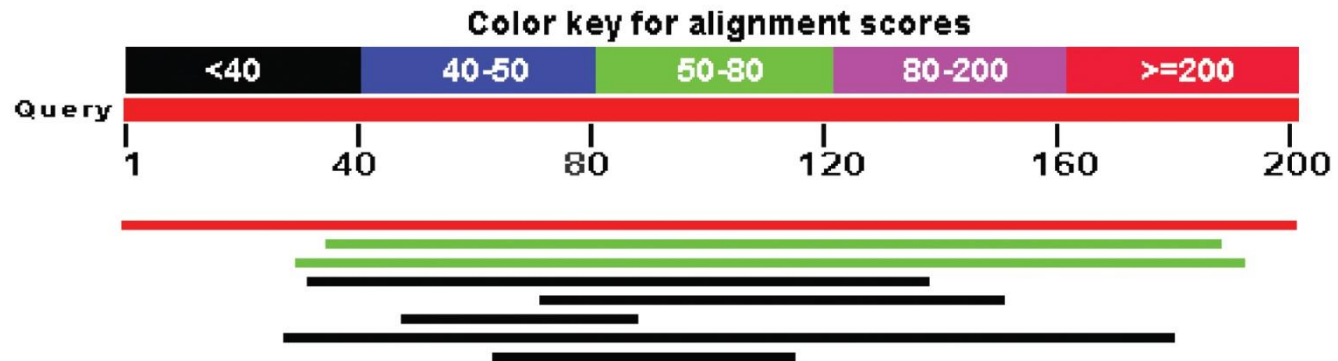
complement component C8 gamma chain precursor [Homo sapiens]

Sequence ID: [ref|NP_000597.2|](#) Length: 202 Number of Matches: 1Range 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	VSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFVSVDETG-QMSATAKGRVRL	82			
	+S+ + K NFD +F+STW +A + AE + Q +A A R L				
Sbjct 33	ISTIQPKANFDAQQFAGTWLLVAVGSACRFLQEQGHRAEATTLHVAPQGTAMAVSTFRKL	92			
Query 83	NNWDVCA DMVGITFDIEDPAKFMKYWGVSFLQKGNDDHWIVDIDYDYAVQY	136			
	+ +C + + DI +F ++ +G + +IDY ++AV Y				
Sbjct 93	DG--ICWQVRQLYGDITGVLRFLQARDA-----RGAVHVVVAETDYQSFAVLY	139			

FIGURE 4.16 Results of a BLASTP nr search using human RBP as a query, restricting the output to human RefSeq proteins. (a) The graphical overview shows that there are 6 hits, only one of which (RBP4 itself) has a high score (bar shaded red) extending across the length of the query. (b) The BLASTP output includes a list of alignments. Inspection of the *E* values suggests that, in addition to RBP itself, several authentic paralogs may have been identified by this search. Is complement component 8 gamma (C8G), having an alignment *E* value of 0.18, likely to be homologous to RBP? (c) Pairwise alignment of RBP4 and C8G, provided as part of the BLASTP output, includes 25% amino acid identity and alignment of a GXW motif (red rectangle) that is consistently conserved among lipocalin carrier proteins such as RBP4.

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

(c) Pairwise alignments with nonhomologous proteins

Download ▾ GenPept Graphics					
tenascin-X isoform 1 precursor [Homo sapiens]					
Sequence ID: refINP_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	TTLHVAPOGTAMAVSTFRKLD-GICWQVRQLYGDIGVLGRFLQARDARGAVHVVVAETD 131				
	T L V P+ +AV+ G+ W V Q G FL+Q RDA+G V D				
Sbjct 3255	TLPFVEPRLGELAAVVISDSVGLSWTVAQ-----GPFDSFLVQYRDAQGQPQAVPVSGD 3309				
Query 132	YQSFVAVLYLERAGQLSVKLYA 152				
	++ AV L+ A + L+				
Sbjct 3310	LRVAVAVGLDPARKYKFLFLFG 3330				

Download ▾ GenPept Graphics					
neuroblastoma-amplified sequence [Homo sapiens]					
Sequence ID: refINP_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	GTWLLVAVGSACRFLQEQGHRAEATTLHVAPOGTAMAVSTF 89				
	G W +G R L+E GH AEA +L +A +GT A TF				
Sbjct 2323	GRWDAEELG---RHLREAGHEAEAGSLLLAVRGTHQAFTF 2360				

FIGURE 4.17 Results of a BLASTP search against human proteins via the nonredundant database, using human complement component 8 gamma (C8G) as a query. (a) The graphical overview shows 8 matches, including the query to itself (red bar) and several alignments with low scores (black bars) spanning just short stretches of amino acids. (b) The list of alignments includes RBP4 and other members of the lipocalin family. This “reciprocal” search supports the hypothesis that C8G, identified in a previous RBP4 search, is an authentic homolog. Here three database matches are not homologous (arrows). The *E* values are unconvincingly high, and the proteins are members of protein families other than lipocalins (as can be confirmed by separate BLAST searches). (c) Inspection of pairwise alignments between C8G and two putative nonhomologous proteins shows that these proteins are far larger than typical lipocalins (4242 and 2371 amino acid residues). The tenascin X isoform 1 does not overlap the highly conserved GXW motif. The neuroblastoma-amplified sequence does match the GXW motif, but the region of overlap extends to only 41 residues. These results highlight the need to inspect each pairwise alignment from a BLAST search. The *E* value provides a statistical argument for evaluating possible homology, but it should be complemented by knowledge of the biological properties of the sequences. Here RBP4, C8G, and other lipocalins are soluble, hydrophilic, abundant proteins that probably share similar functions as carrier proteins; they also share similar three-dimensional structures (see Chapter 13).

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
 © 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
 Companion Website: www.wiley.com/go/pevsnerbioinformatics

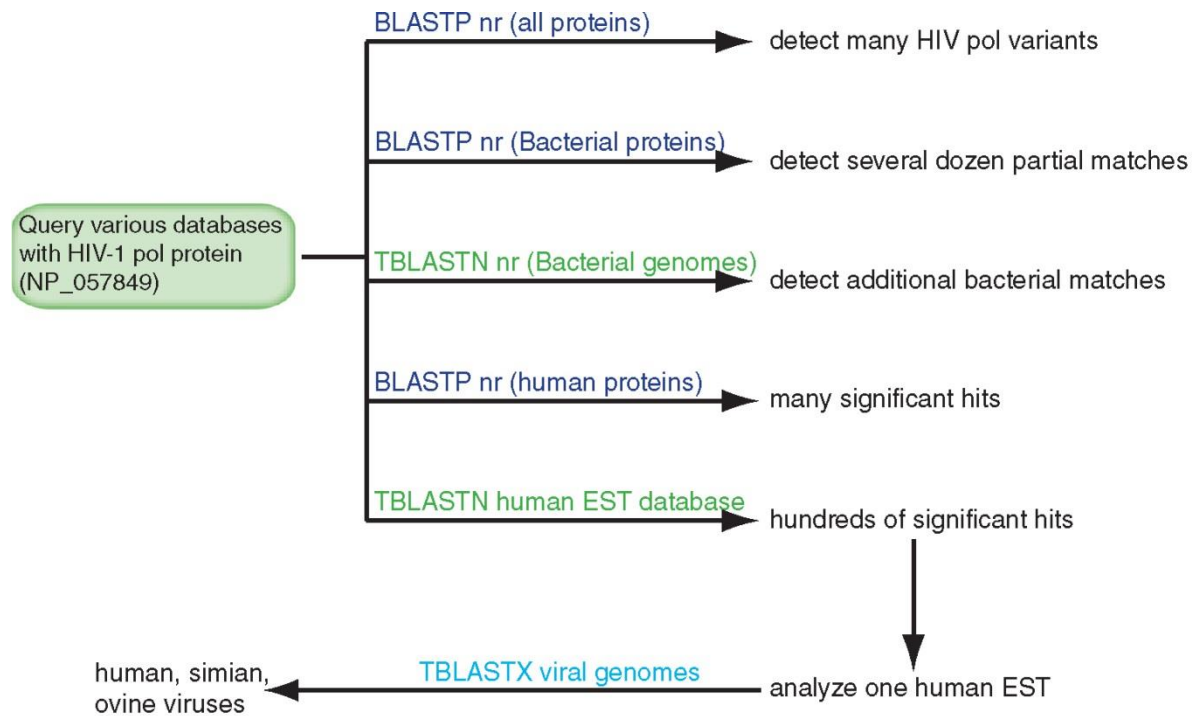
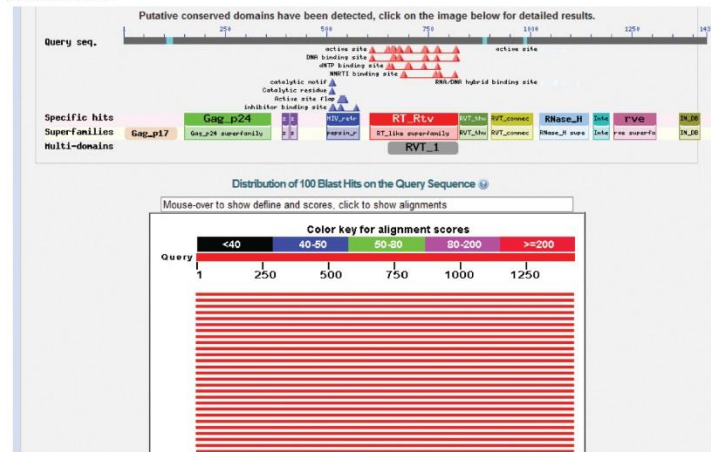


FIGURE 4.18 Overview of BLAST searches beginning with HIV-1 Pol protein. A series of BLAST searches can often be performed to pursue questions about a particular gene, protein, or organism. The number of database matches returned by a BLAST search can vary from none to thousands and depends on the nature of the query, the database, and the search parameters.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
© 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
Companion Website: www.wiley.com/go/pevsnerbioinformatics

(a) Graphical overview



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

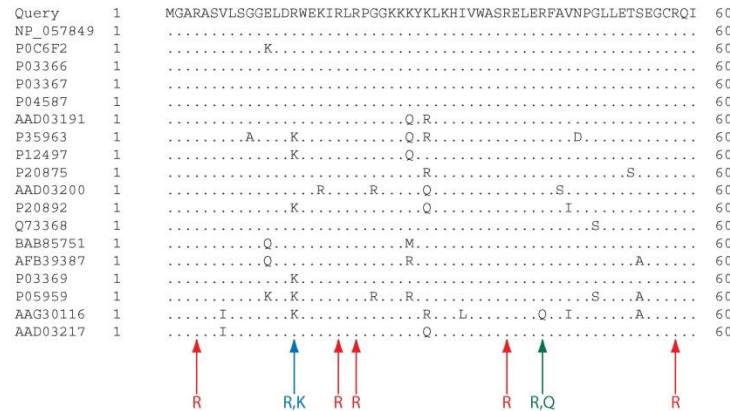


FIGURE 4.19 A BLASTP search with HIV-1 viral Pol (NP_057849). (a) Graphical overview shows conserved domains in the protein. These blocks are clickable and link to the Conserved Domain Database at NCBI (Chapters 5 and 6). The links are to protein domains (Gag_p17, Gag_p24) and abbreviations include rvp, retroviral aspartyl protease; rvt, reverse transcriptase (RNA-dependent DNA polymerase); masH, ribonuclease H; rve, integrase core domain. The red horizontal bars indicate many close matches to viral proteins. (b) The BLAST alignment options include formats such as query-anchored, in which dots correspond to residues in database entries that match the query. This view highlights the occasional sequence differences in viral proteins. Arrows indicate arginine (R) positions in the query that are perfectly conserved, or that are sometimes substituted with lysine (K) or glutamine (Q).

```

Human immunodeficiency virus 1 [viruses] taxid 11676
ref|NP_057849.4| Gag-Pol [Human immunodeficiency virus 1] 2971 0.0
ref|NP_789740.1| Pol [Human immunodeficiency virus 1] 2052 0.0
ref|NP_705927.1| reverse transcriptase [Human immunodeficiency virus 1] 1149 0.0
ref|YP_001856242.1| reverse transcriptase [Human immunodeficiency virus 1] 1149 0.0
ref|NP_789739.1| reverse transcriptase p51 subunit [Human immunodeficiency virus 1] 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 virus 1] 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 virus 1] 204 2e-57
ref|NP_579881.1| nucleocapsid [Human immunodeficiency virus 1] 130 5e-32
ref|NP_787043.1| Gag-Pol Transframe peptide [Human immunodeficiency virus 1] 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 immunodeficiency virus 2] 1675 0.0
ref|NP_056837.1| gag polyprotein [Human immunodeficiency virus 2] 523 3e-167

Simian immunodeficiency virus SIV-mnd 2 [viruses] taxid 159122
ref|NP_758887.1| pol protein [Simian immunodeficiency virus] 1377 0.0
ref|NP_758886.1| gag protein [Simian immunodeficiency virus] 486 2e-153

Feline immunodeficiency virus [viruses] taxid 11673
ref|NP_040973.1| pol polyprotein [Feline immunodeficiency virus] 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 virus] 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 albicans] 90 2e-15
ref|XP_721310.1| hypothetical protein CaO19.6468 [Candida albicans] 86 1e-14

Sus scrofa (wild boar, ...) [even-toed ungulates] taxid 9823
ref|XP_003482346.1| PREDICTED: hypothetical protein LOC100000000 90 2e-15

Tribolium castaneum (rust-red flour beetle) [beetles] taxid 7070
ref|XP_001815322.1| PREDICTED: similar to orf [Tribolium castaneum] 89 5e-15
ref|XP_001808495.1| PREDICTED: similar to orf [Tribolium castaneum] 88 8e-15

Candida dubliniensis CD36 [ascomycetes] taxid 573826
ref|XP_002421195.1| retrovirus-related Pol polyprotein fragment 88 6e-15

Moniliophthora perniciosa PA553 [basidiomycetes] taxid 554373
ref|XP_002387985.1| hypothetical protein MPER_13056 [Moniliophthora perniciosa] 88 7e-15

```

FIGURE 4.20 The taxonomy report for a BLASTP search shows an overview of which species have proteins matching the HIV-1 query. Most matches are viral, but others include rabbit, fungal, pig, and insect sequences. The /// symbols indicate a series of other matches (not shown).

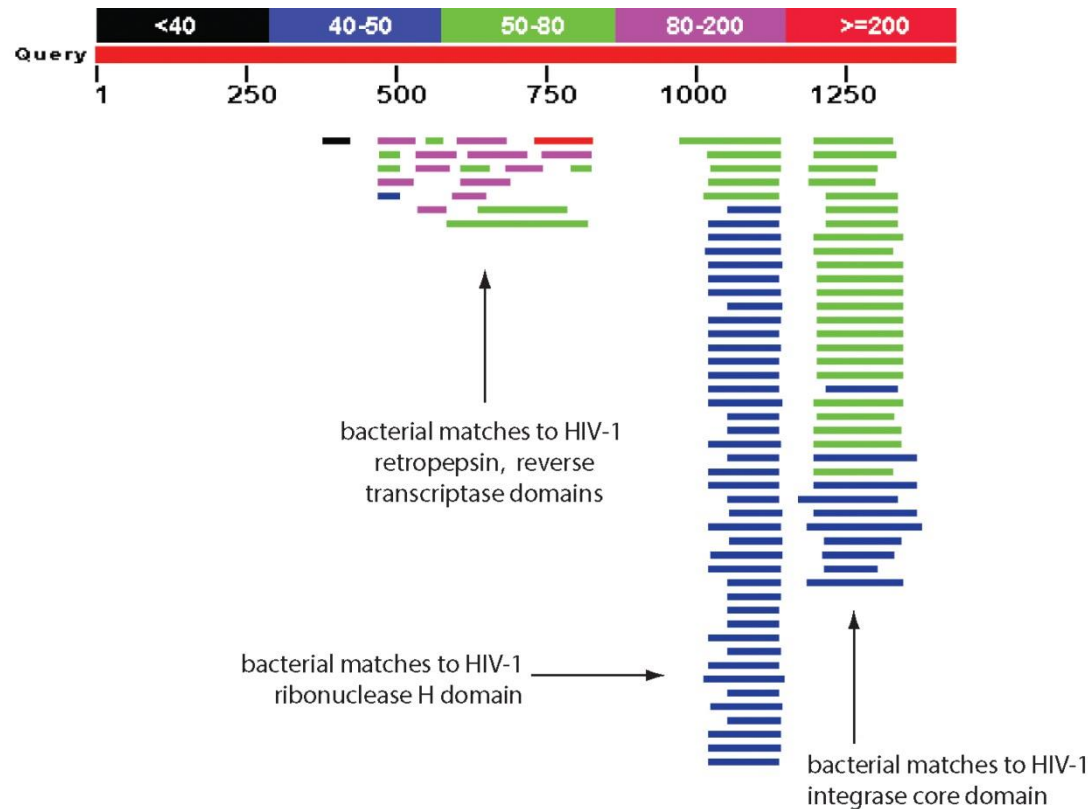


FIGURE 4.21 Result of a BLASTP search with HIV-1 Pol as a query, restricting the output to bacteria. The graphical output of the BLAST search allows identification of the domains within HIV-1 that have bacterial matches. The length of overlap and the number of bacterial sequences are also evident.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
 © 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
 Companion Website: www.wiley.com/go/pevsnerbioinformatics

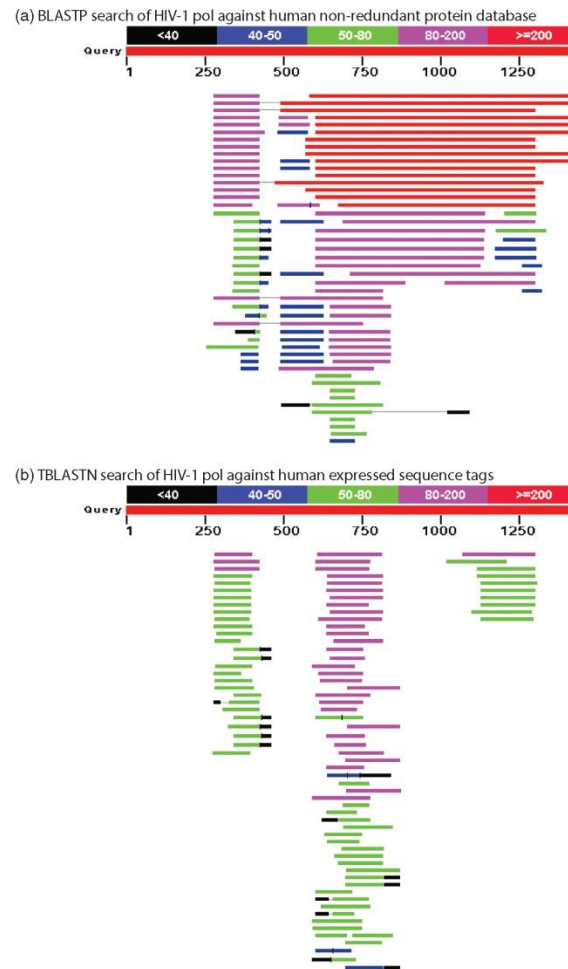


FIGURE 4.22 (a) Graphical output of a BLASTP search using HIV-1 Pol protein to search for matches against human proteins. Note that some human hits have very high scores. (b) Are human transcripts expressed that encode proteins homologous to HIV-1 Pol protein? The results of a TBLASTN search with viral Pol protein against a human EST database are shown. Many human genes are actively transcribed to generate transcripts predicted to make proteins homologous to HIV-1 Pol.

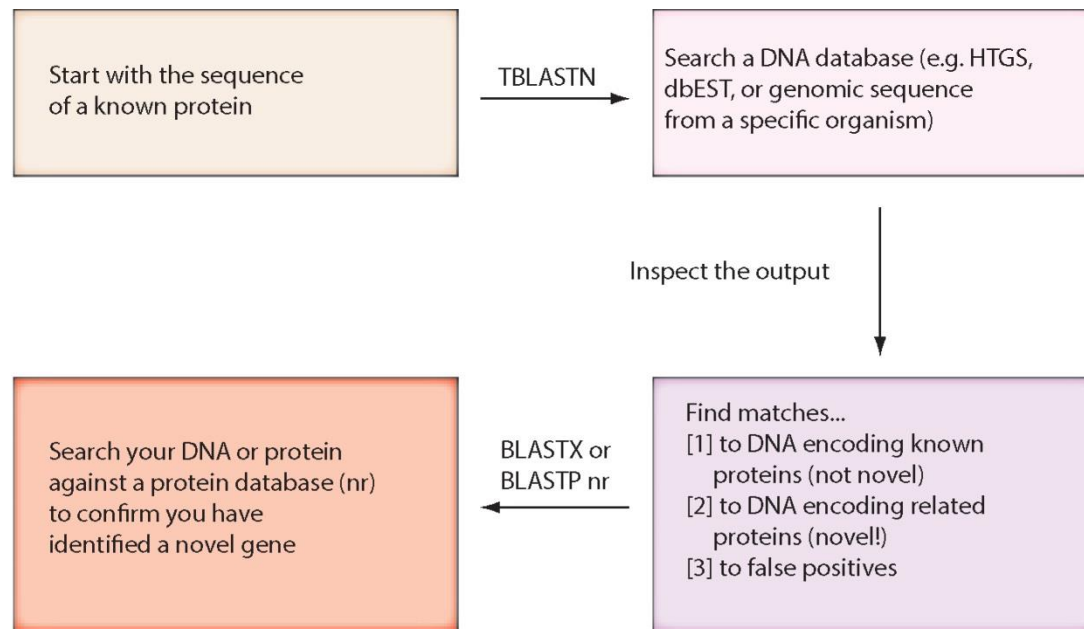


FIGURE 4.23 How to discover a novel gene by BLAST searching. Begin with the sequence of a known protein such as human beta globin. Perform a TBLASTN search of a DNA database. It is unlikely that there are many “novel” genes in the well-characterized genomes of organisms such as human, yeast, or *E. coli*. It may therefore be helpful to search databases of organisms that are poorly characterized or not fully annotated. The TBLASTN search may result in two types of significant matches: (1) matches of your query to known proteins that are already annotated; and (2) homologous proteins that have not yet been annotated (“novel” genes and corresponding novel proteins). (3) The DNA sequence corresponding to the putative novel gene may be searched using the BLASTX algorithm against the nonredundant (nr) database. This may confirm that the DNA does indeed encode a protein that has no perfect match to any described protein.

(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	MVHLTPEEKSAVTALWGKNVDEVGGEALGRLLVVYPTQRFESFGDLSTPDAMGNPK	60				
Sbjct 40	MV T E +A+ +LW K+NV+E+G +A+ RLL+V PWTQR F +FG+LST A+M N K	219				
Query 61	VKAHGKKVLGAFSDGLAHLDNLKGTFAILSELHCDKLHVDPENFRLLGNVLCVLAHFG	120				
Sbjct 220	V H G V+G + ++D++K + LS +H +KLHVD+P+FRLL + +A FG	399				
Query 121	-KEFTPFVQAAYQKVVAGVANALAHKYH	147				
Sbjct 400	EFT VQ A+QK + V +AL +YH	483				

(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|HBB_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	VETIDAEHTAILSLWKKINVEEIGPQAMRLLIVCPWTQRHFANFGNLSTAAAIMNNEKV				222			
Sbjct 1	VENT+E TAI S W KIN+EEIGPQAMRLLIVCPWTQRHFANFGNLSTAAAIMN+KV				60			
Query 223	AKHGTTVMGGLDRAIQNMDDIKNAYRELSVMHSEKLHVDPDNFRLLSEHITLCMAAKFGP				402			
Sbjct 61	AKHGTTVMGGLDRAIQNMDDIKNAYR+LSVMHSEKLHVDPDNFRLL+EHITLCMAAKFGP				120			
Query 403	TEFTADVQEAQKFLMAVTSALGRQYH 483							
Sbjct 121	TEFTADVQEAQKFLMAVTSALGRQYH 147							

FIGURE 4.24 The find-a-gene project was demonstrated using human beta globin (NP_000509) as a query and searching a database of expressed sequence tags (ESTs) restricted to nematodes. (a) The matches included one to an EST from *Anguillicola crassus* (GenBank accession JK511422.1). (b) Using this accession as a query, a BLASTX nr search revealed matches to known beta globins. The best match, shown here, was to a vertebrate globin. However, since there was not a match to an *A. crassus* globin, this suggests that the find-a-gene project resulted in the identification of a DNA sequence that encodes a previously undescribed nematode globin. This novel globin can then be characterized in terms of its full-length sequence, homologs, evolution, structure, and function.

Bioinformatics and Functional Genomics, Third Edition, Jonathan Pevsner.
© 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
Companion Website: www.wiley.com/go/pevsnerbioinformatics