



Basic Local Alignment Sequence Tool (BLAST)

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 few results

 - BLAST searching with multidomain protein: HIV-1 Pol

Using BLAST for gene discovery: Find-a-Gene

Learning objectives

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

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 paralog
- 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

The screenshot shows the NCBI BLASTP search interface. The top navigation bar includes 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. The main title is 'Standard Protein BLAST'. The 'Enter Query Sequence' section has a text area containing a FASTA sequence:

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MYHLTPKESAVTALNGKVNDEVGGEALGRLLVYFPIQRFFESFGDLSTPDVACNPKVKAH
GKKVLGAFSDGLAHLNLLKGTFTALSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPEVQ
AAYQKVVGAVANALAHKYH
```

. Below this is a 'Job Title' field with the text 'gi|4504349|ref|NP_000509.1| hemoglobin subunit...'. The 'Choose Search Set' section shows the 'Database' dropdown set to 'Reference proteins (refseq_protein)'. The 'Entrez Query' field contains 'perutz mf[Author]'. The 'Program Selection' section has 'blastp (protein-protein BLAST)' selected. At the bottom, the 'BLAST' button is visible, along with a checkbox for 'Show results in a new window' and a link to 'Algorithm parameters'. Annotations with numbers 1 through 5 point to these specific fields: 1 points to the query sequence, 2 to the database dropdown, 3 to the Entrez query field, 4 to the algorithm selection, and 5 to the algorithm parameters link.

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI Welcome pepsner. [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 [From](#) [To](#)

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

Job Title Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database [Reference proteins \(refseq_protein\)](#) [?](#)

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

Entrez Query Optional Enter an Entrez query to limit search [?](#)

Program Selection

Algorithm ☒ blastp (protein-protein BLAST) ☐ PSI-BLAST (Position-Specific Iterated BLAST) ☐ PHI-BLAST (Pattern Hit Initiated BLAST) ☐ 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

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

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

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.

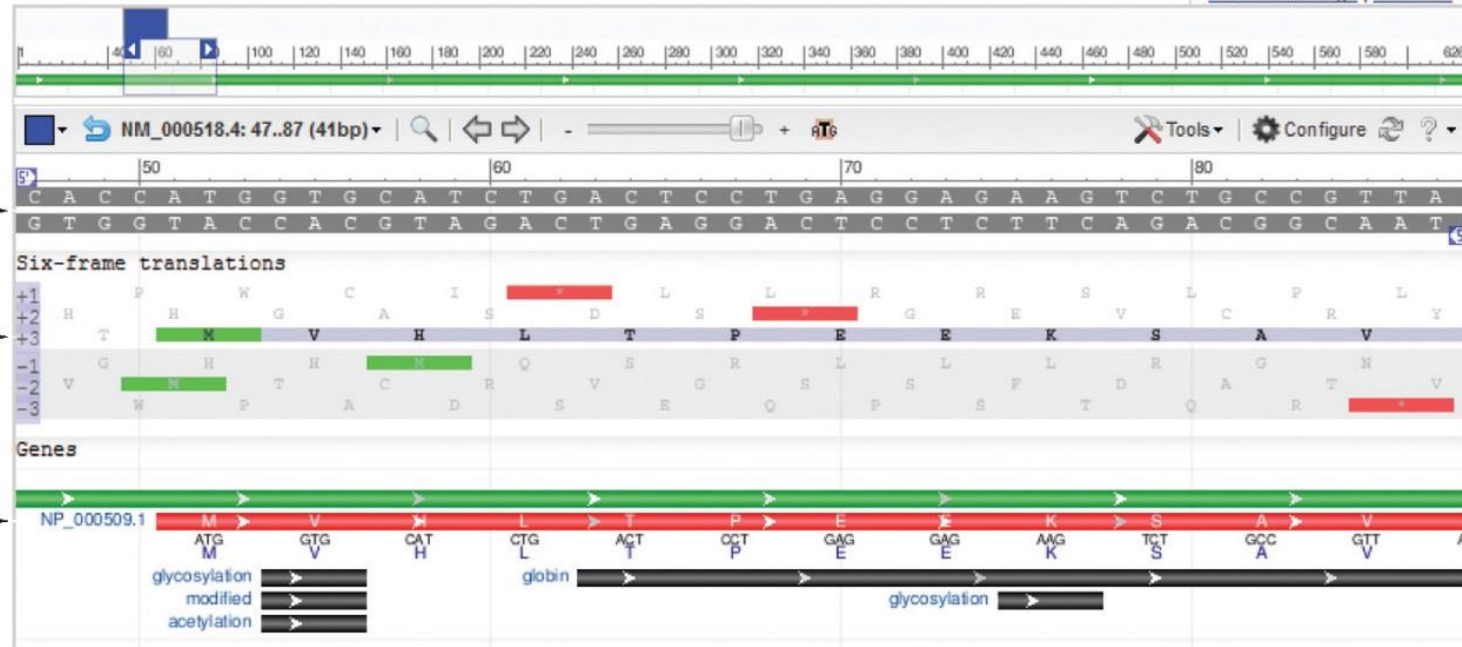
Step 2 (choosing the BLAST program): DNA can be translated into six reading frames

Homo sapiens hemoglobin, beta (HBB), mRNA

NCBI Reference Sequence: NM_000518.4


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

[Link To This Page](#) | [Feedback](#)



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

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

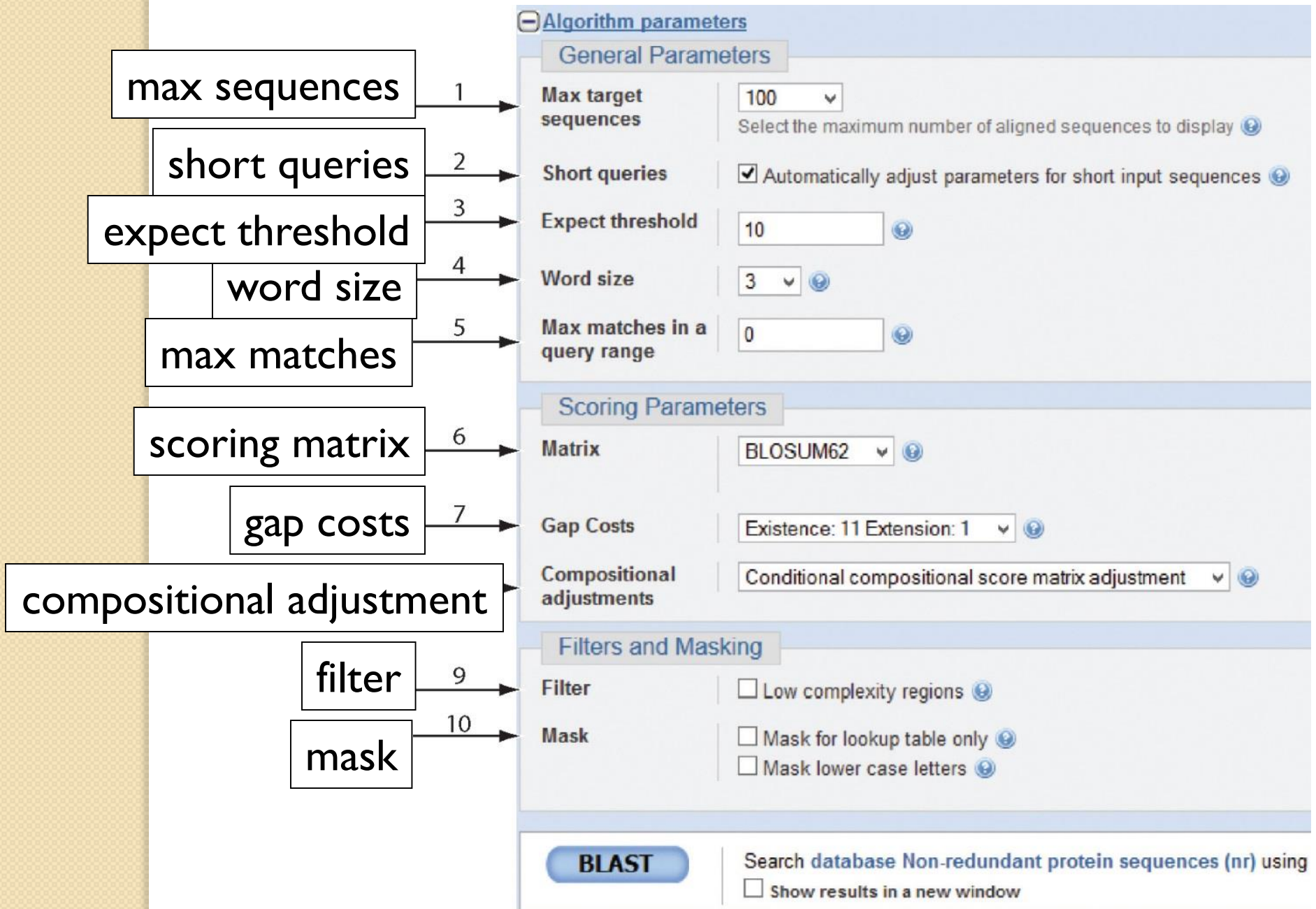
Step 4: optional parameters

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



The diagram illustrates the mapping of search parameters to the BLASTP web interface. On the left, a vertical stack of boxes lists parameters: 'max sequences', 'short queries', 'expect threshold', 'word size', 'max matches', 'scoring matrix', 'gap costs', 'compositional adjustment', 'filter', and 'mask'. Arrows with numbers 1 through 10 point from these boxes to the corresponding fields in the BLASTP interface on the right. The interface is divided into sections: 'Algorithm parameters' (containing 'General Parameters'), 'Scoring Parameters', and 'Filters and Masking'. The 'BLAST' button and database selection are at the bottom.

Parameter	BLASTP Field	Value / Option
max sequences	Max target sequences	100
short queries	Short queries	<input checked="" type="checkbox"/> Automatically adjust parameters for short input sequences
expect threshold	Expect threshold	10
word size	Word size	3
max matches	Max matches in a query range	0
scoring matrix	Matrix	BLOSUM62
gap costs	Gap Costs	Existence: 11 Extension: 1
compositional adjustment	Compositional adjustments	Conditional compositional score matrix adjustment
filter	Filter	<input type="checkbox"/> Low complexity regions
mask	Mask	<input type="checkbox"/> Mask for lookup table only <input type="checkbox"/> Mask lower case letters

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

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

(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	LKTRRLRDGVFDECCCLKSCTMDEVLYRC 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	LCGRKLPELTSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQML 87			
Query 88	-----KRGIVEQCCTSICSLYQLENYC 109			
	+ G+ ++CC C++ ++ YC			
Sbjct 88	KTRRLRDGVFDECCCLKSCTMDEVLYRC 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	LCGRKLPELTSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQML 87				
Query 88	-----KRGIVEQCCTSICSLYQLENYC 109				
	+ G+ ++CC C++ ++ YC				
Sbjct 88	KTRRLRDGVFDECCCLKSCTMDEVLYRC 114				

expect = 0.05

Default: conditional
compositional score
matrix adjustment

expect = 0.09

no adjustment

expect = 1e-04

composition-based
statistics

Step 4b: formatting options

The screenshot shows the BLAST Basic Local Alignment Search Tool interface. At the top, there is a navigation bar with links: Home, Recent Results, Saved Strategies, and Help. A user login box on the right says "My NCBI" and "Welcome pevsner. [Sign Out]". Below the navigation bar, the page title is "NCBI/ BLAST/ blastp suite/ Formatting Results - U4X4JS8B014". A message box states: "Your search is limited to records matching entrez query: txid6656 [ORGN].". Below this, there are links: "Edit and Resubmit", "Save Search Strategies", "Formatting options", and "Download". On the right, there are links: "YouTube", "How to read this page", and "Blast report description". The main content area shows the query details: "gi|4504349|ref|NP_000509.1| hemoglobin subunit...". Below this, there are two columns of information. The left column contains: "Query ID" (lc|51620), "Description" (gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]), "Molecule type" (amino acid), and "Query Length" (147). The right column contains: "Database Name" (refseq_protein), "Description" (NCBI Protein Reference Sequences), and "Program" (BLASTP 2.2.28+). At the bottom, there are links for "Other reports": "Search Summary", "Taxonomy reports", "Distance tree of results", and "Multiple alignment". Numbered annotations are present: 1 points to the "Query ID" label, 2 points to the "Query Length" label, 3 points to the "Database Name" label, 4 points to the "Description" label in the right column, 5 points to the "Molecule type" label, and 6 points to the "Query Length" label.

1	Query ID	lc 51620	Database Name	refseq_protein	3
	Description	gi 4504349 ref NP_000509.1 hemoglobin subunit beta [Homo sapiens]	Description	NCBI Protein Reference Sequences	4
	Molecule type	amino acid	Program	BLASTP 2.2.28+ Citation	
2	Query Length	147			

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

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]

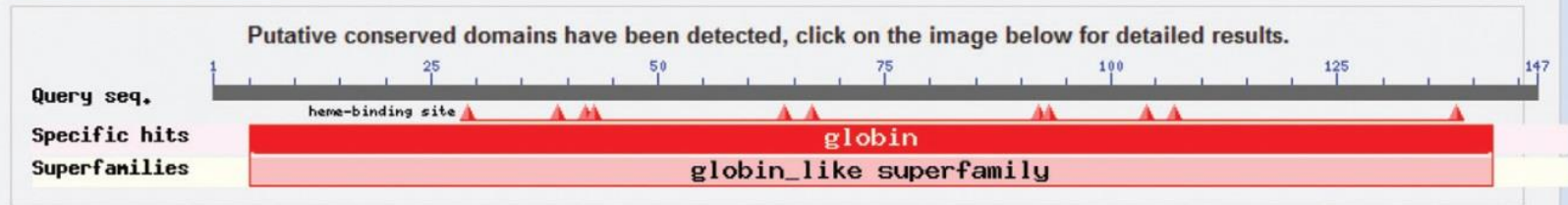
Size of database

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

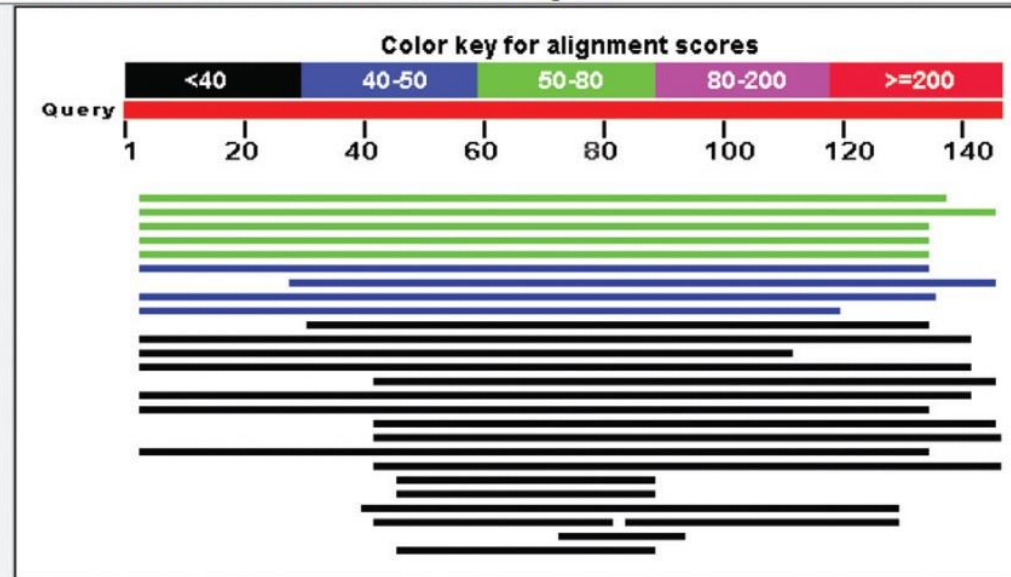
Graphic Summary

Show Conserved Domains



Distribution of 27 Blast Hits on the Query Sequence

Mouse-over to show define and scores, click to show alignments





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
<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

BLAST output can be formatted to display multiple alignment

**COBALT**
Home Recent Results Help

Constraint-based Multiple Alignment Tool

My NCBI
Welcome pevsner. [Sign Out]

Phylogenetic Tree Edit and Resubmit Back to Blast Results Download

Multiple Alignment Results - gi|4504349|ref|NP_000509.1| hemoglobin subunit... - Cobalt RID U57PC4Y5211 (8 seqs)**Descriptions** ☒ Select All [Re-align](#) [Alignment parameters](#)Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer**Alignments** ☒ Select All [Re-align](#) [Mouse over the sequence identifier for sequence title](#)View Format: [Compact](#) [Conservation](#) [Conservation Setting: 2 Bits](#)

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	ATTITGCTTCTGACACAACTGTGTTCACTAGCAACCTCAA---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	CTGACTCCTGAGGAGAAGTCTGCCGTIACCTGCCCTGTGGGCAAGGTGAACGTGGATGAA		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	GTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTTGGACCCAGAGGTICTTT		179
Sbjct	323	GCTGGAGGTGAAGCCTTGGGCAGACTCCTCGTITGTTACCCCTGGACCCAGAGATTTTT		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	GAGTCCITTGGGAATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCT		239
Sbjct	383	GACAGCTTTGGAACCTGTCTCTCCCTCTGCCATCCTGGGCAACCCCAAGGTCAAGGCC		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	CATGGCAAGAAAGTGCTGACTTCTTTGGAGATGCTATTAAAAACATGGACAACCTCAAG		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	TTCAAGCTCCTGGGTAACGTGATGGTGATTATCTGGCTACTCACTTTGGCAAGGAGTTC		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			I P P V Q A A Y Q K V V A G V A N A L A
Query	420	ACCCACCAAGTGCAGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCC		479
Sbjct	623	ACCCCTGAAGTGCAGGCTGCCTGGCAGAAGCTGGTGTCTGCTGTGCCATTGCCCTGGCC		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

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

Command-line BLAST+

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

Three steps:

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

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]
MVHLTPEEKSAVTALWGKVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPKVKAHGKKVLG
AFSDGLAHLDDLKGTFTSLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

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

Do the search:

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

View the results:

```
$ less mysearch1
```

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

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

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 I: compile a list of word pairs ($w=3$)
above threshold T

Example: for a human RBP query
...FS**GTW**YA... (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 I: 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**...).

examples of words \geq threshold 12	LWG	$4+11+6=21$
	IWG	$2+11+6=19$
threshold 	MWG	$2+11+6=19$
	VWG	$1+11+6=18$
	FWG	$0+11+6=17$
	AWG	$0+11+6=17$
	LWS	$4+11+0=15$
	LWN	$4+11+0=15$
	LWA	$4+11+0=15$
	LYG	$4+2+6=12$
	LFG	$4+1+6=11$
	FWS	$0+11+0=11$
examples of words below threshold	AWS	$-1+11+0=10$
	CWS	$-1+11+0=10$
	IWC	$2+11-3=10$

Phase 2: scan the database for matches and extend

Phase 2: Scanning and extensions

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

LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPKV HBB
L+P +K+ V A WGKV + E G EAL R+ + +P T+ +F F D G+ +V
LSPADKTNVKAAWGKVGHAHAGEYGAEALERMFSLFPTTKTYFPHF-----DLSHGSAQV HBA

← extension extension →

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

Phase 3: Traceback to generate gapped alignment

Phase 3: Traceback

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

How a BLAST search works: threshold

You can locally install BLAST and modify the threshold parameter.

The default value for BLASTP is 11.

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

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

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

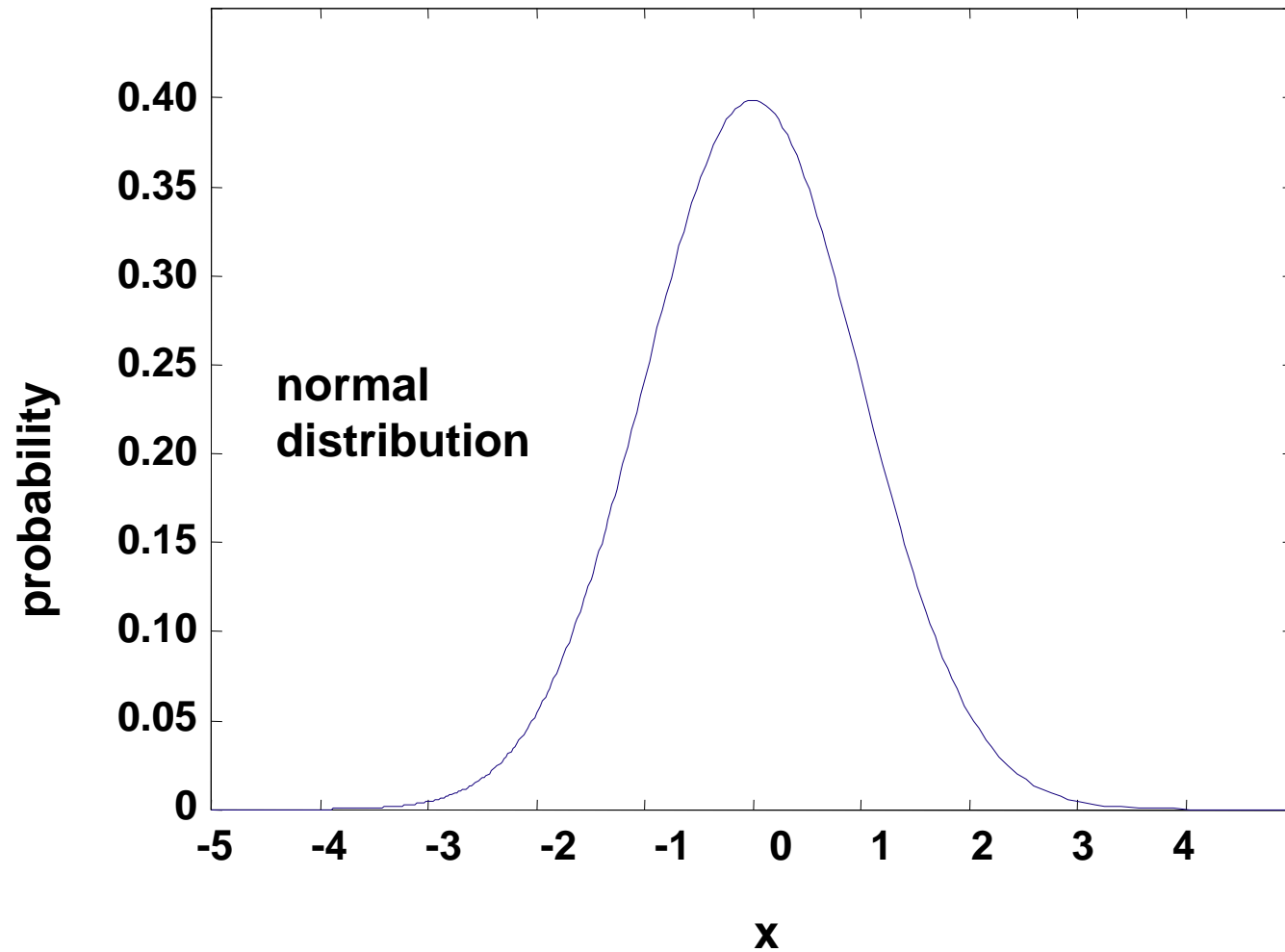
How to interpret a BLAST search: expect value

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

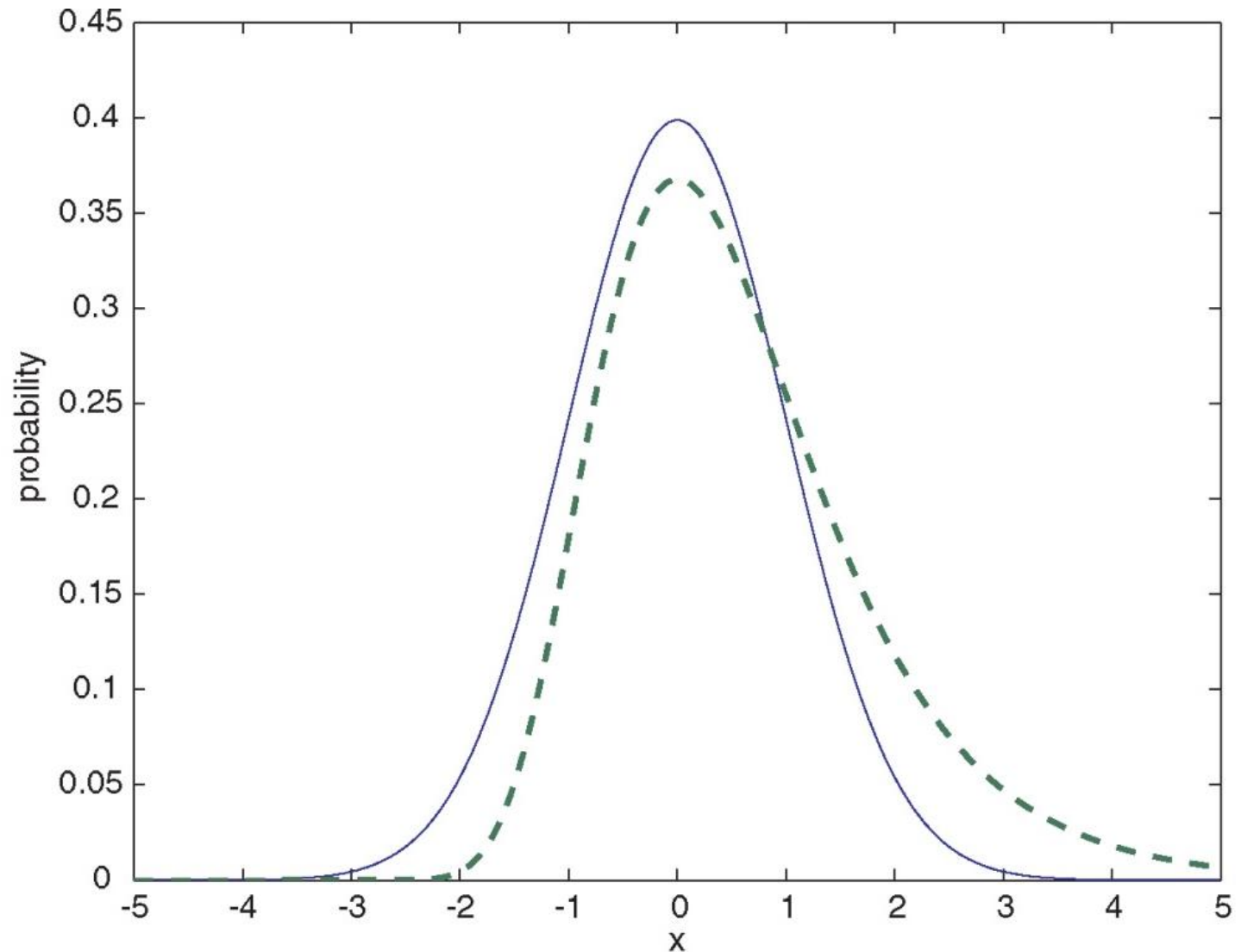
For global alignments, the statistics are poorly understood.

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



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.

<u>E</u>	<u>p</u>
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

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

Overview of BLAST search strategies

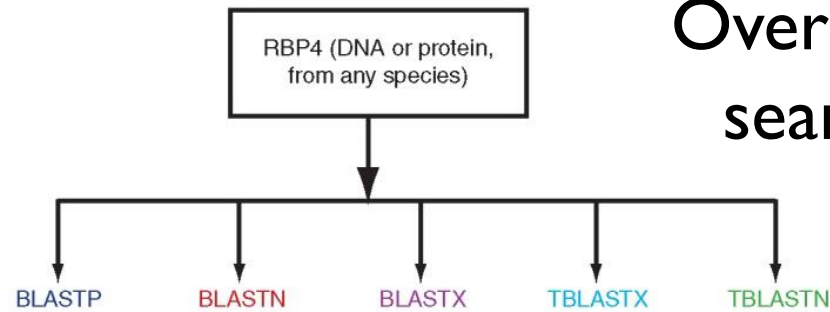
Starting point:
a molecular
sequence

Search
strategies

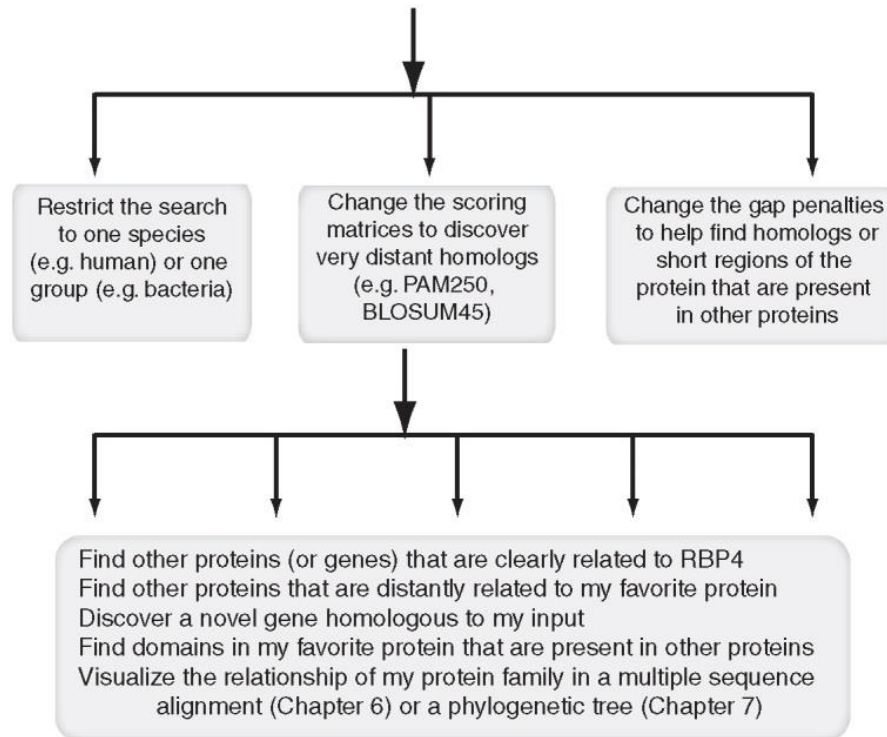
Sample
questions

Modifiable
search
parameters

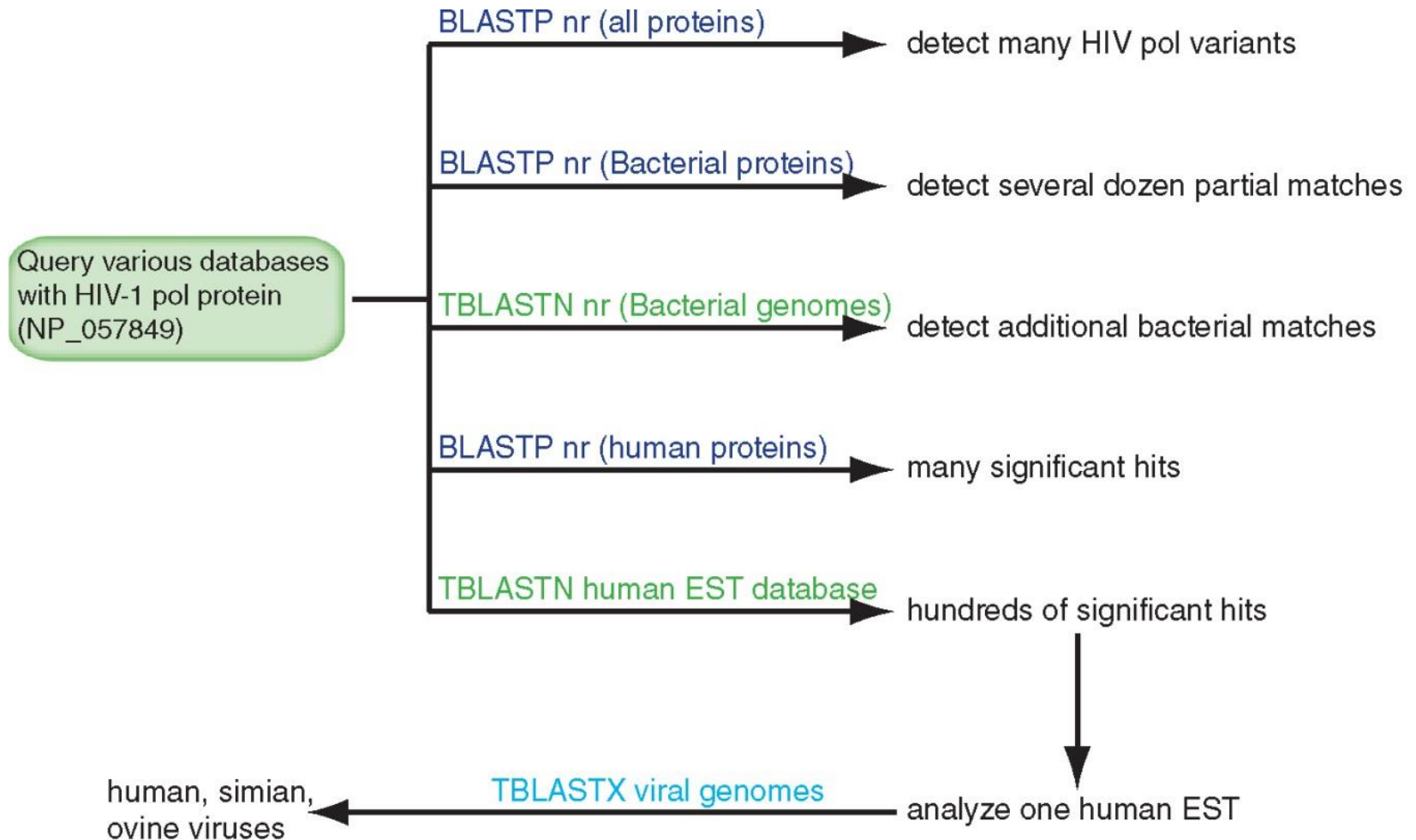
Goals:
Results that can
be obtained by
BLAST searching



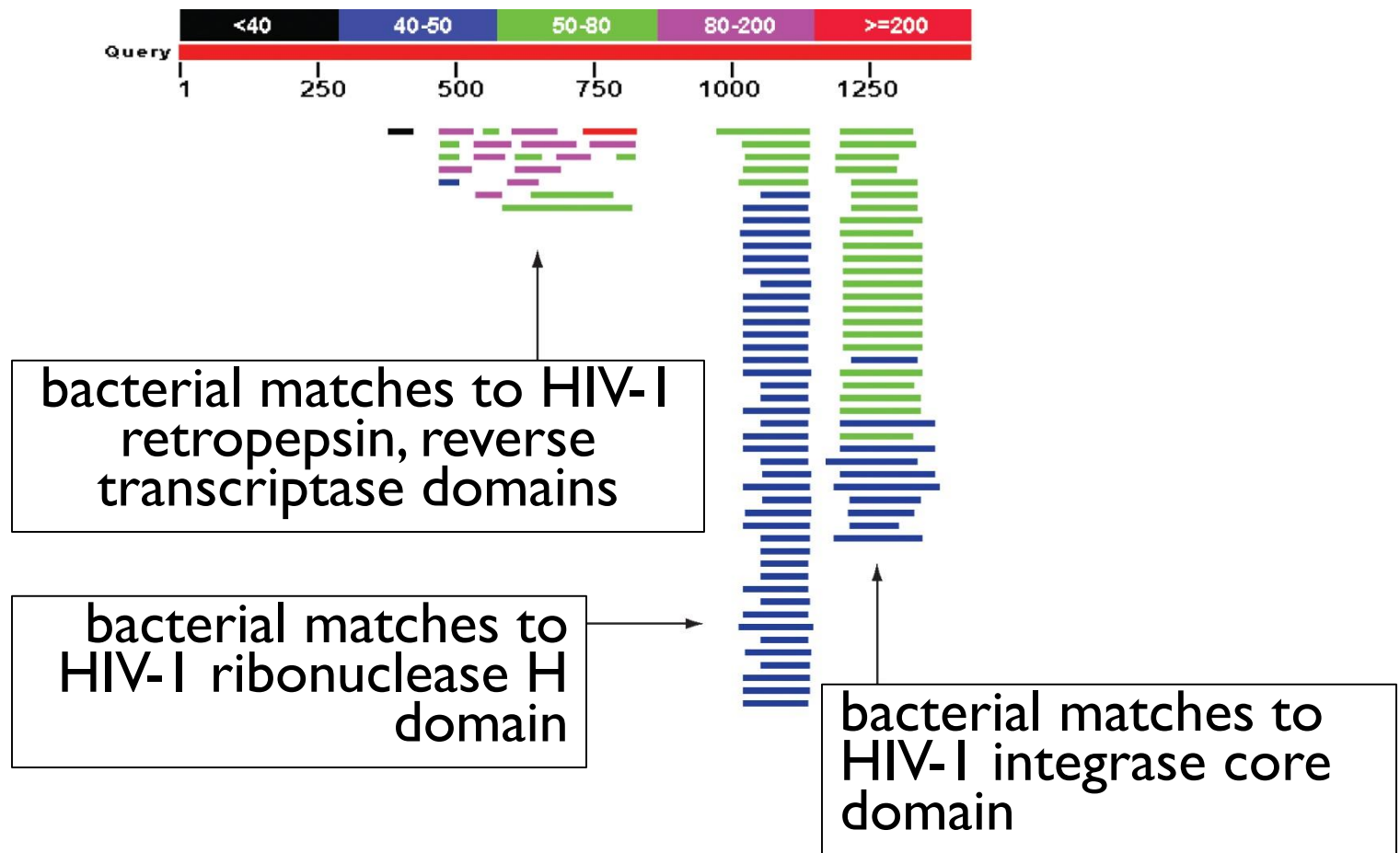
BLASTP: What other proteins are related to RBP4 protein?
BLASTN: Is the 3' untranslated region of human RBP4 DNA homologous to the 3' untranslated region of RBP paralogs or orthologs?
BLASTX: What known protein is a lipocalin EST most related to?
TBLASTX: Does human RBP4 DNA match a protein predicted to be encoded from a gene in a DNA library such as bacterial ESTs?
TBLASTN: Is there an RBP4 ortholog represented in a genomic DNA database?



BLAST searching a multidomain protein: HIV-1 pol

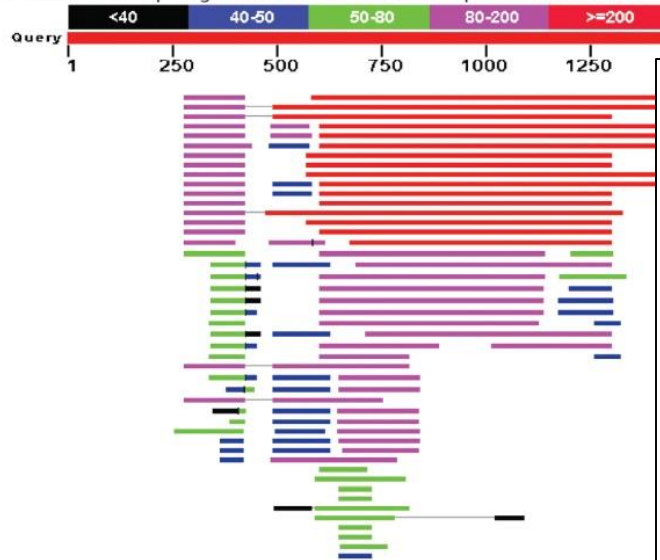


BLASTP searching HIV-1 pol against bacterial proteins



BLAST searching HIV-1 pol against human sequences

(a) BLASTP search of HIV-1 pol against human non-redundant protein database



Question: are there human homologs of HIV-1 pol protein?

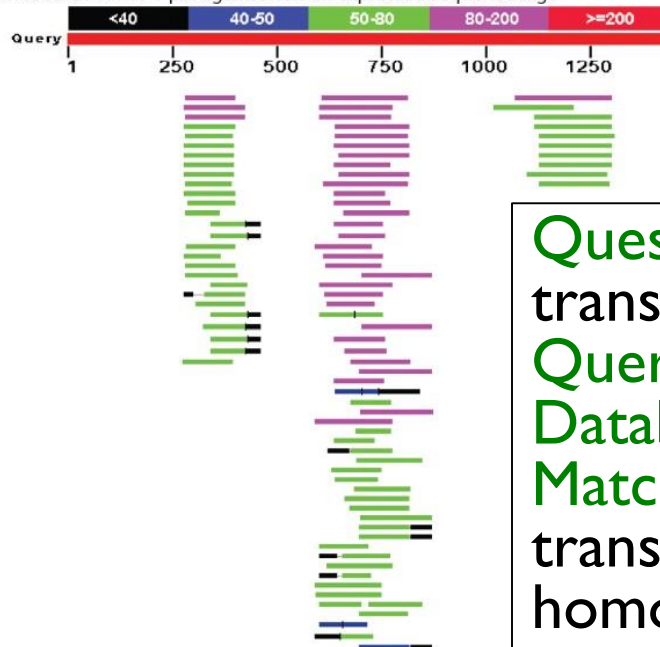
Query: HIV-1 Pol

Program: BLASTP

Database: human nr (nonredundant)

Matches: many human proteins share significant identity.

(b) TBLASTN search of HIV-1 pol against human expressed sequence tags



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.

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

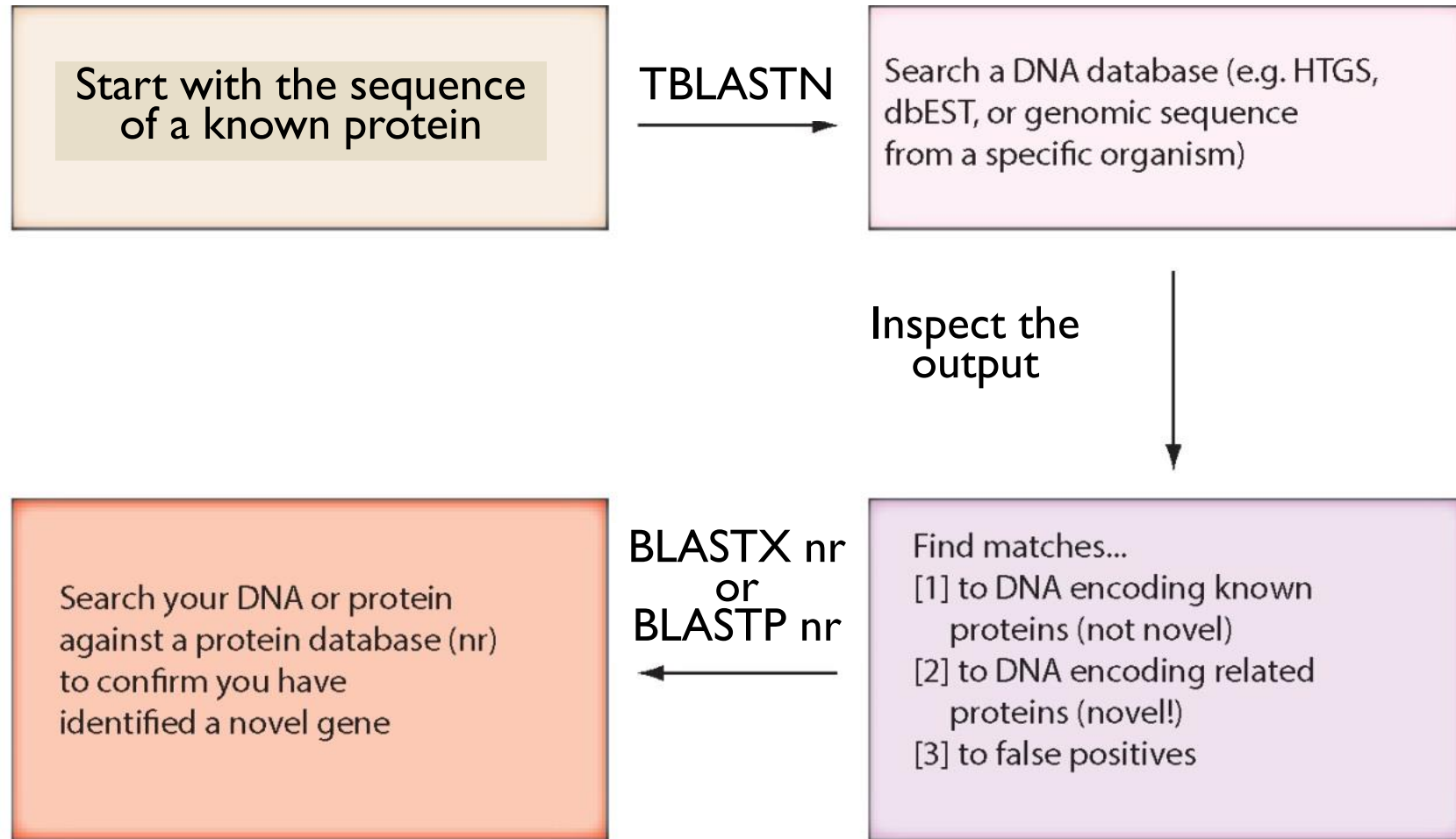
How to evaluate the significance of results

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

“Find-a-gene project” to practice BLAST



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

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

Ac_EH1r_01A07_M13 Adult Anguillicola crassus Anguillicola crassus cDNA clone Ac_EH1r_01A07

Sequence ID: [gb|JK511422.1](#) Length: 559 Number of Matches: 1

Range 1: 40 to 483			GenBank	Graphics	▼ Next Match	▲ Previous Match
Score	Expect	Method	Identities	Positives	Gaps	Frame
149 bits(375)	6e-44	Compositional matrix adjust.	69/148(47%)	97/148(65%)	1/148(0%)	+1
Query 1	MVHLTPEEKSAVTALWGKVNVEVGGEALGRLLVVPWTQRFFESFGDLSTPDAVMGNPK				60	
	MV T E +A+ +LW K+NV+E+G +A+ RLL+V PWTQR F +FG+LST A+M N K					
Sbjct 40	MVEWIDAEHTAILSLWKKINVEEIGPQAMRRLIVCPWTQRHFANFGNLSTAAAIMNNEK				219	
Query 61	VKAHGKKVLGAFSDGLAHLNKLGTIFATLSSELHCDKLHVDPENFRLLGNVLVCVLAHHFG				120	
	V HG V+G + ++D++K + LS +H +KLHVDPE+FRLL + +A FG					
Sbjct 220	VAKHGTTVMGGGLDRAIQNMDDIKNAYRELSVMHSEKLVDPDNFRLLSEHITLCMAAKFG				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	VEWTDAEHTAILS	SLWKKINVEEIGPQAMRRLIVCPWTQRHFANFGNLSTAAAIMNNEKV			222	
	VEWT+ E TAI S W KIN+EEIGPQAMRRLIVCPWTQRHFANFGNLSTAAAIMNN+KV					
Sbjct 1	VEWTEDETAISKWLKINIEEIGPQAMRRLIVCPWTQRHFANFGNLSTAAAIMNNDKV				60	
Query 223	AKHGTTVMGGGLDRAIQNMDDIKNAYRELSVMHSEKLVDPDNFRLLSEHITLCMAAKFGP				402	
	AKHGTTVMGGGLDRAIQNMDDIKNAYR+LSVMHSEKLVDPDNFRLL+EHITLCMAAKFGP					
Sbjct 61	AKHGTTVMGGGLDRAIQNMDDIKNAYRQLSVMHSEKLVDPDNFRLLAEHITLCMAAKFGP				120	
Query 403	TEFTADVQEAWQKFLMAVTSALGRQYH	483				
	TEFTADVQEAWQKFLMAVTSAL RQYH					
Sbjct 121	TEFTADVQEAWQKFLMAVTSALARQYH	147				

Conf

Quer

Progr

Best

not a

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.