BLAST
BIFX-550

S. *Ravi*chandran, Ph.D.
Biology,
Hood College, Frederick, MD 21701

Goals

- Explore BLAST from NCBI
- How to carry out BLAST searches?
- Explain the BLAST search in detail.
- BLAST specific inputs/outputs
  - E-values & scores
- Discuss strategies for carrying out BLAST

Drawbacks of Dynamic Programming & lead into Word-based methods

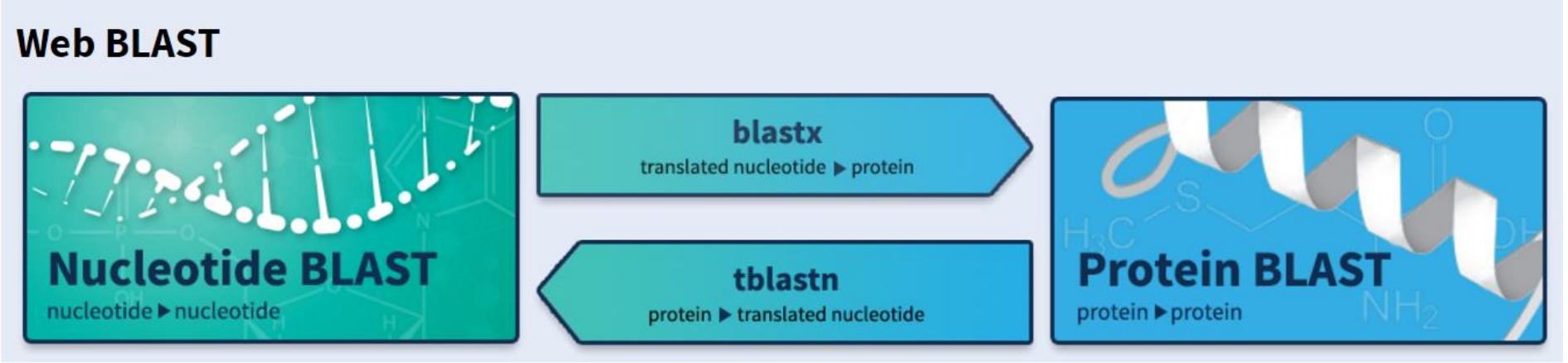
- BLAST, FASTA commonly used today for sequence alignment
  - Heuristic and word-based
    - No guarantee we will find the optimum alignment
    - But, they will find multiple possible solutions quickly

Basic terminology

- Query: your sequence
  - ID, FASTA, Range of residues/nucleotides
- Database: search against your query
  - Typically millions of sequences
- Default method for searching is local alignment
  - Not global alignment
  - Not the local alignment adopted by Smith Waterman
    - BLAST does accomplish using different methodology
    - Original method is computationally too expensive

Basic Local Alignment Search Tool

- What is BLAST?
  - A query tool to retrieve similar (homologous) sequences from a DB
- Flavors
  - Blastp; blastn; Blastx; tblastn; Tblastx
  - BLAST2



What is BLAST used for?

- Identifying homologs for proteins/DNA
  - Orthologs & paralogs
- May have a sequence and would like to identify the identity
  - SMART BLAST
- Discover new genes
  - Genomic BLAST

What is BLAST used for?

- Discovering proteins that have similar domains
- Searching EST database to find out the alternative splicing mechanisms
- Finally functionally relevant amino acid residues either in a protein DB or a PDB database

Summary: Goals of BLAST

- Search for high-scoring pairs (HSPs)
  - Aligned segment score that cannot be extended or trimmed to increase the score
  - Both gapped/ungapped
  - Score must be above score threshold

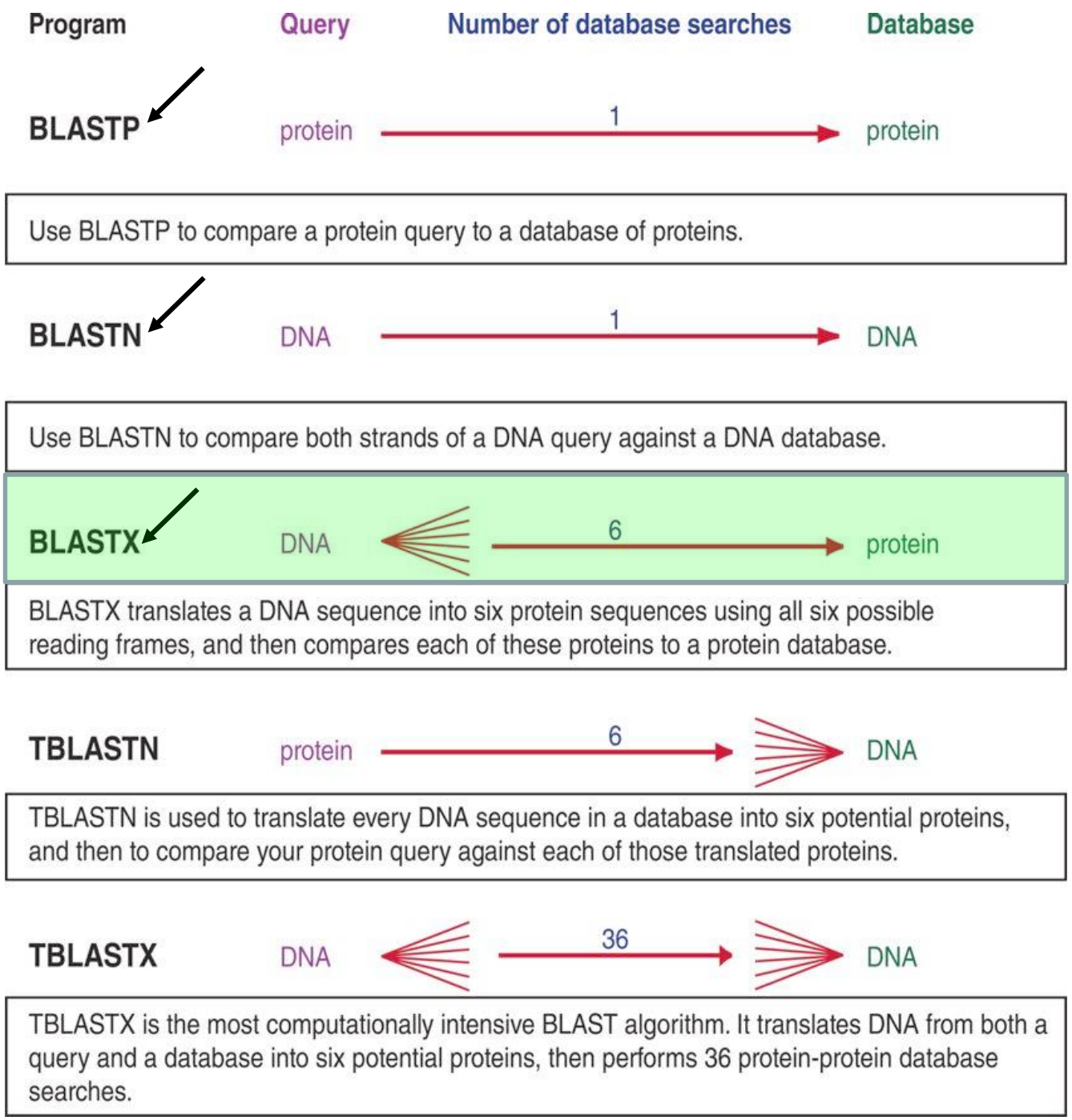
Step1

- Sequence of interest
  - Query
  - DNA/Protein
- Input
  - Sequence or Accession number
  - FASTA
  - Etc.

Step 2

What Flavor of BLAST?

Fig 3.12 from the Pevsner Book III Edition PLEASE DO NOT DISTRIBUTE Copyright figure

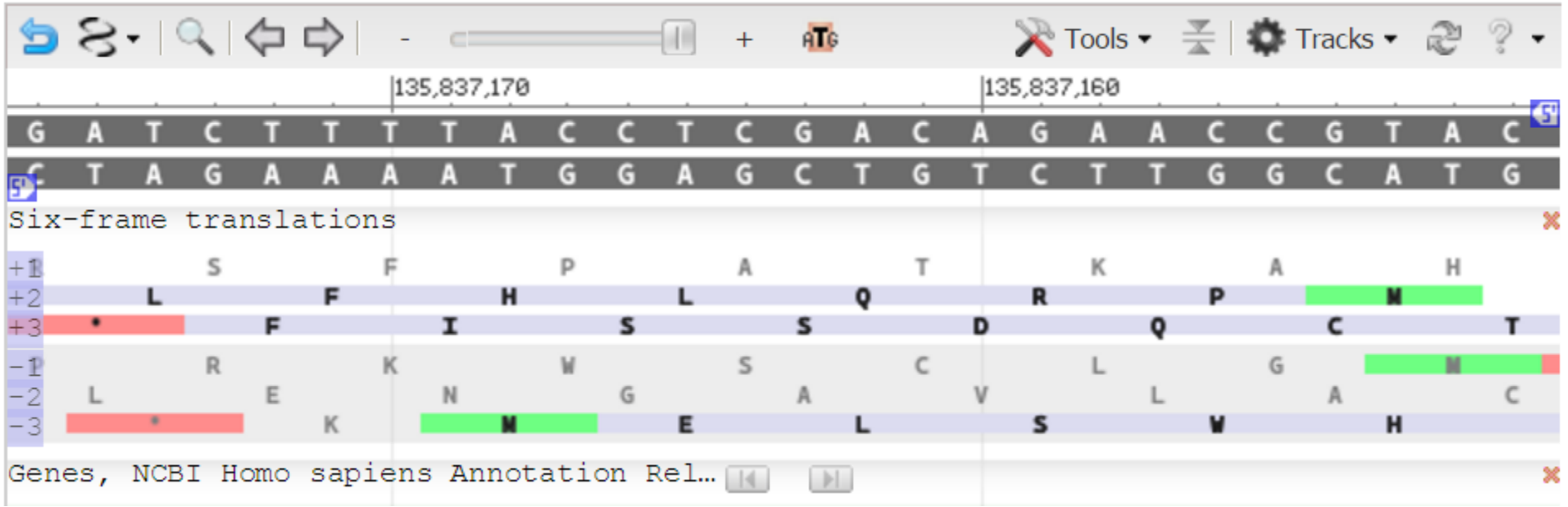


UniGene Uses all nucleotides in its DB to search against all known Protein sequences

BLASTX

Translation of a DNA to a protein and searched against a Protein DB. What is a translation?

LCT Gene → Protein
6 Frames of Translation




```
        10         20         30         40         50
MELSWHVVFI ALLSFSCWGS DWESDRNFIS TAGPLTNDLL HNLSGLLGDQ
```

Step 3: Selecting a DB

- Protein
  - nr database
  - GenBank, PDB, SwissProt, PIR & PRF
  - RefSeq
- DNA
  - BLASTN, TBLASTN,TBLASTX
    - nr/nt (GenBank, EMBL,DDBJ, PDB & RefSeq)
  - Note that nr does not include, EST, STS, WGS, GSS, TSA, Patents or HTGS databases

Table 4.1 from Pevsner III Edition

Database	Title	# sequences	
nr	All nonredundant GenBank CDS translations + PDB + SwissProt + PIR + PRF excluding environmental samples from WGS projects	65 million	~146M
Reference proteins	NCBI protein reference sequences	50 million	~102M
UniProtKB/SwissProt	Nonredundant UniProtKB/SwissProt sequences	450,000	~468K
Patented protein sequences	Protein sequences derived from the Patent division of GenBank	1.3 million	~2.2M
Protein Data Bank	PDB protein database	77,000	~97K
Metagenomic proteins	Proteins from WGS metagenomic projects (env_nr)	6.5 million	~6.9M
Transcriptome	Transcriptome Shotgun Assembly (TSA) sequences	770,000	~2.42M

Accessed Date (2018/02/25)

nr: formed by merging several main protein/DNA DBs
These often contain many identical sequences. Generally only one copy if kept during merging

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

Table 4.2 from Pevsner III Edition

Why is this task difficult?

- Sep 2018
- nrDB
  - ❑ ~171M sequences Protein
- ❑RefSeq
  - ❑~118M Sequences Protein
- Query

Title:All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule Type:Protein
Update date:2018/10/11
Number of sequences:171418145

RID	96CT7C3E014 (Expires on 02-27 05:05 am)
Query ID	NP_000509.1
Description	hemoglobin subunit beta [Homo sapiens]
Molecule type	amino acid
Query Length	147

RefSeq

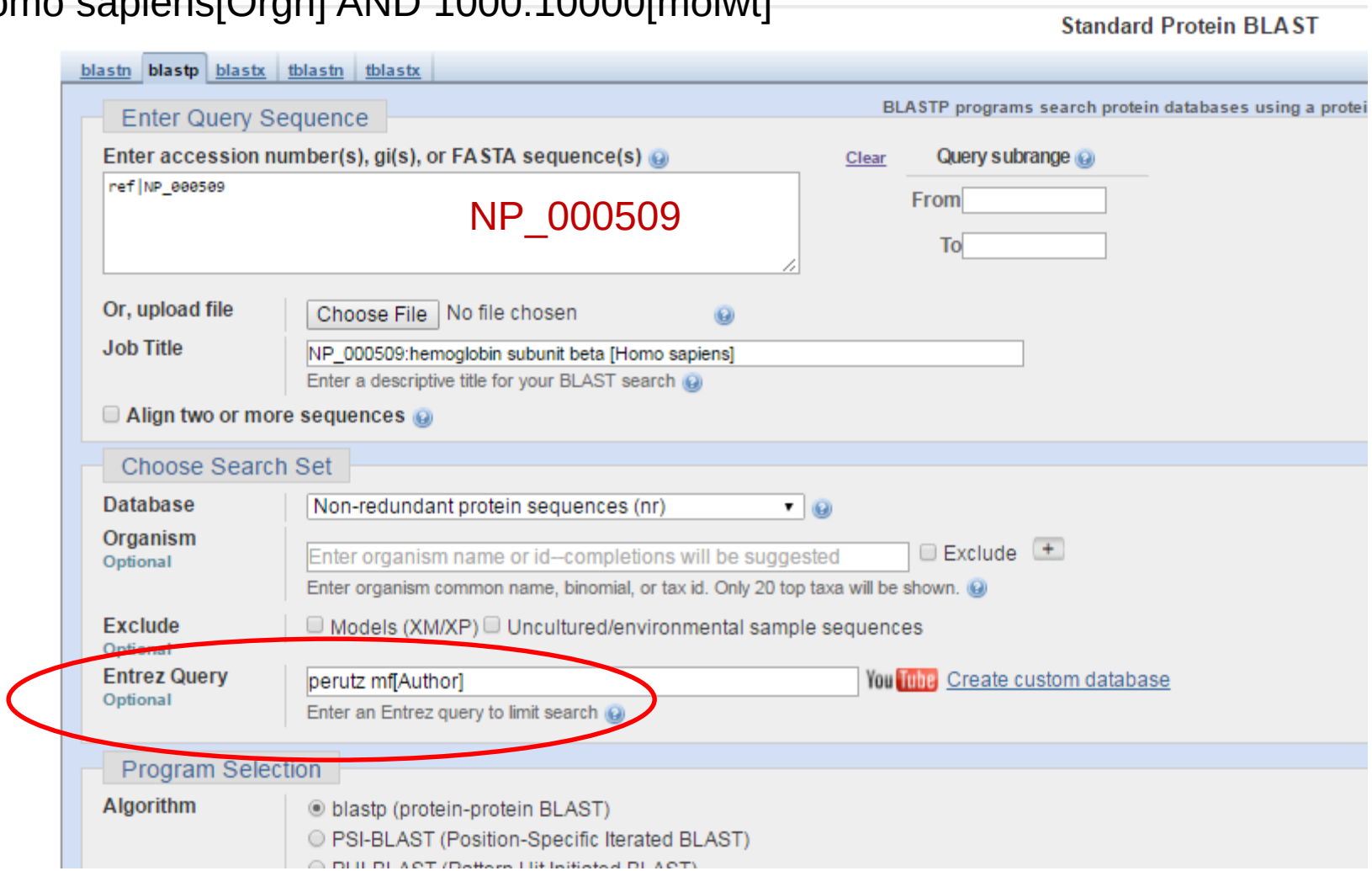
- Difference between RefSeq and nr?
  - RefSeq is also nr (DNA/Protein DB)
  - RefSeq has curation (manual)
    - Genome annotation; gene characterization etc.
  - High quality
  - Maintained by NCBI (derived)

Step 4a: Optional Search Parameters

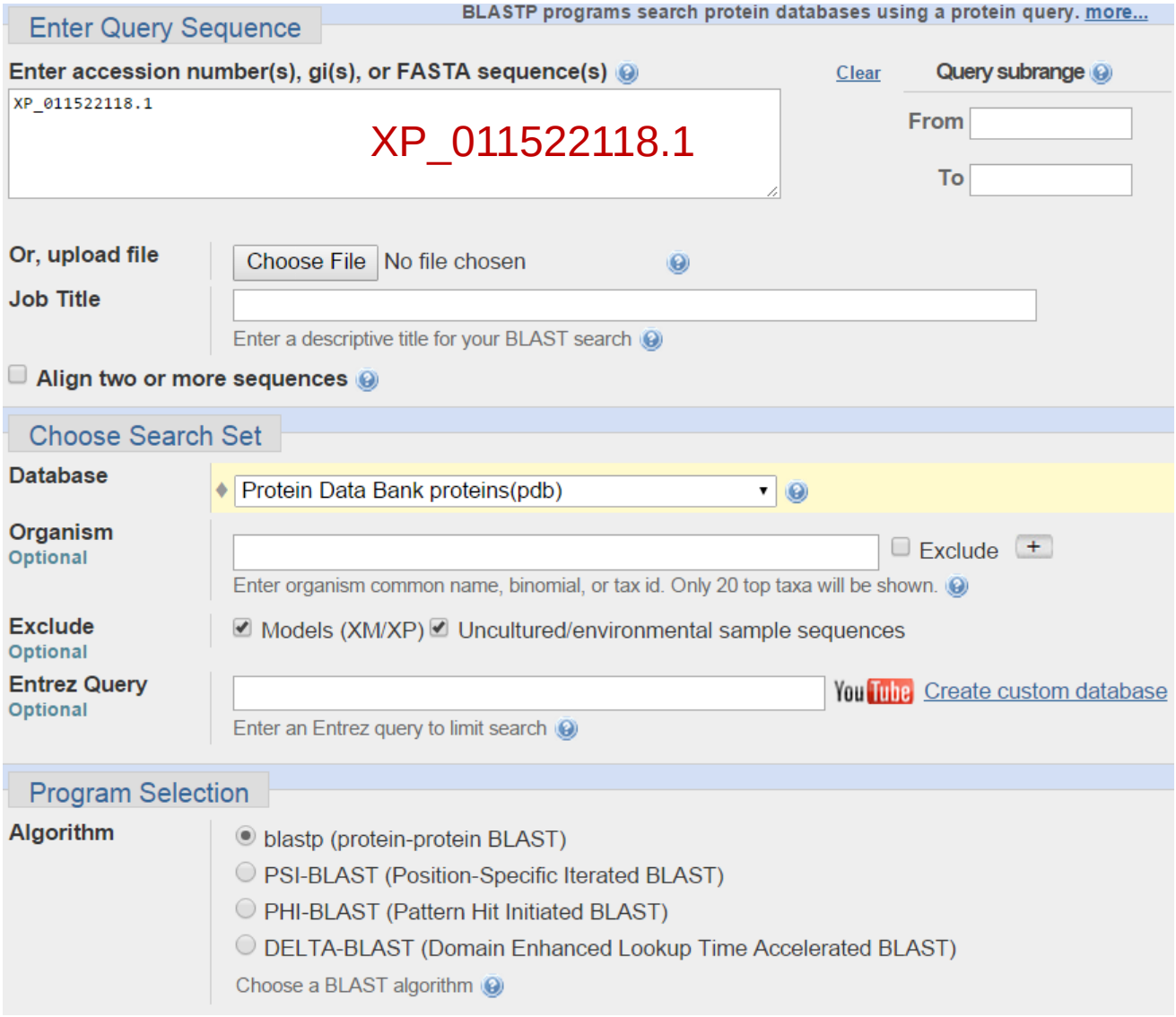
- Limit by Entrez Query
  - "perutz mf[Author]"
  - Check by running through Entrez
- Max Target sequences (default 100)
  - How many do you hits do you want to see?
- Short queries
  - Short sequences, parameters are adjusted for the special case

You can restrict your search. Using the Entrez query. Test your query in Entrez before using it in BLAST

Homo sapiens[Orgn] AND 1000:10000[molwt]



Human retinal guanylyl cyclase 1 isoform X1

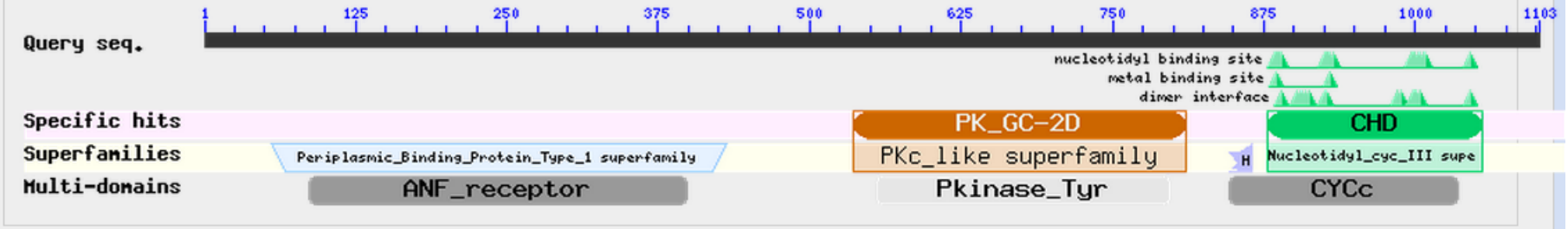


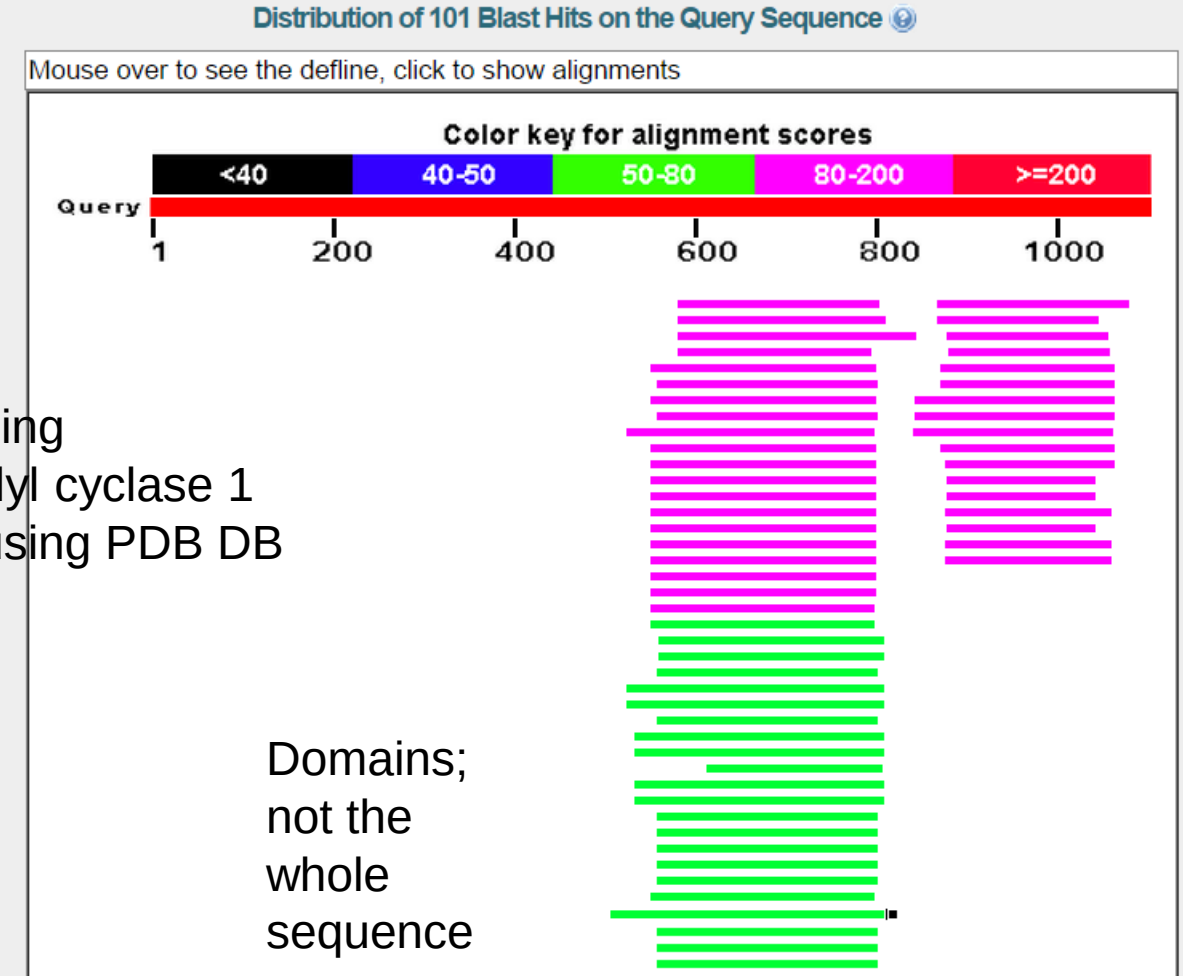
A new BLAST Flavor

- QuickBLASTP is an accelerated version of BLASTP
  - Contains a new pre-processing step
    - Words of length five in the protein-nr (non-redundant) to identify candidate sequences
    - These candidate sequences are aligned and scored with the standard BLASTP algorithm
  - Adv: seconds rather than minutes
  - Limitations: Limited DB and sequence length < 10K

A new BLAST Flavor

- QuickBLASTP is an accelerated version of BLASTP
  - Advantage: seconds rather than minutes
  - Limitations: Limited DB and sequence length < 10K
  - *QuickBLASTP will find around 97% of the database sequences with 70% or more identity to your query and around 98% of the database sequences with 80% or more identity to your query*





Searching using retinal guanylyl cyclase 1 isoform X1 using PDB DB

Domains; not the whole sequence

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

Advanced Options

- Expected Threshold (E-value)
  - "Number of different alignments with scores equal or greater than some score S that are expected to occur in a DB search by chance"
  - Short queries
    - Score is inversely prop to E-value

(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(

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

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

Query human insulin
Hit: insulin-like peptide3 from Drosophila

E value means that to get a score of 31.6 bits or better is expected by chance 1 in 20 times
(for a given DB/choice of parameters)

E-values

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

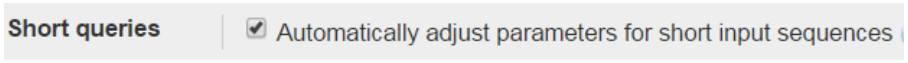
- Default value is 10

  - What this means is, At this E-value, 10 hits with score or better than the alignment score S are expected by chance.
    - Also assumes that you search using a random query with similar length of your actual query

- When you have a small query,

>Query
ELVIS

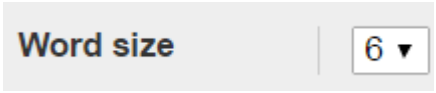
Higher E values (200,000) are set because shorter query cannot get larger scores



Search Parameters	
Program	blastp
Word size	2
Expect value	200000
Hitlist size	100
Gapcosts	9,1
Matrix	PAM30
Filter string	F
Genetic Code	1
Window Size	40
Threshold	11

Advanced Options

- Word size



  - Protein (default, 3) ; DNA (default size 11)
  - What is word size? (more on this later)
    - Smaller subsequence of length equal to the word size
  - Impact of word size on sequences
    - Protein: 3→2 little effect unless a short query
    - DNA: Will have an impact
      - Lower: (11→7) accurate but slow
      - **Higher: Longer sequence search; fast**
        - » Used in MegaBlast

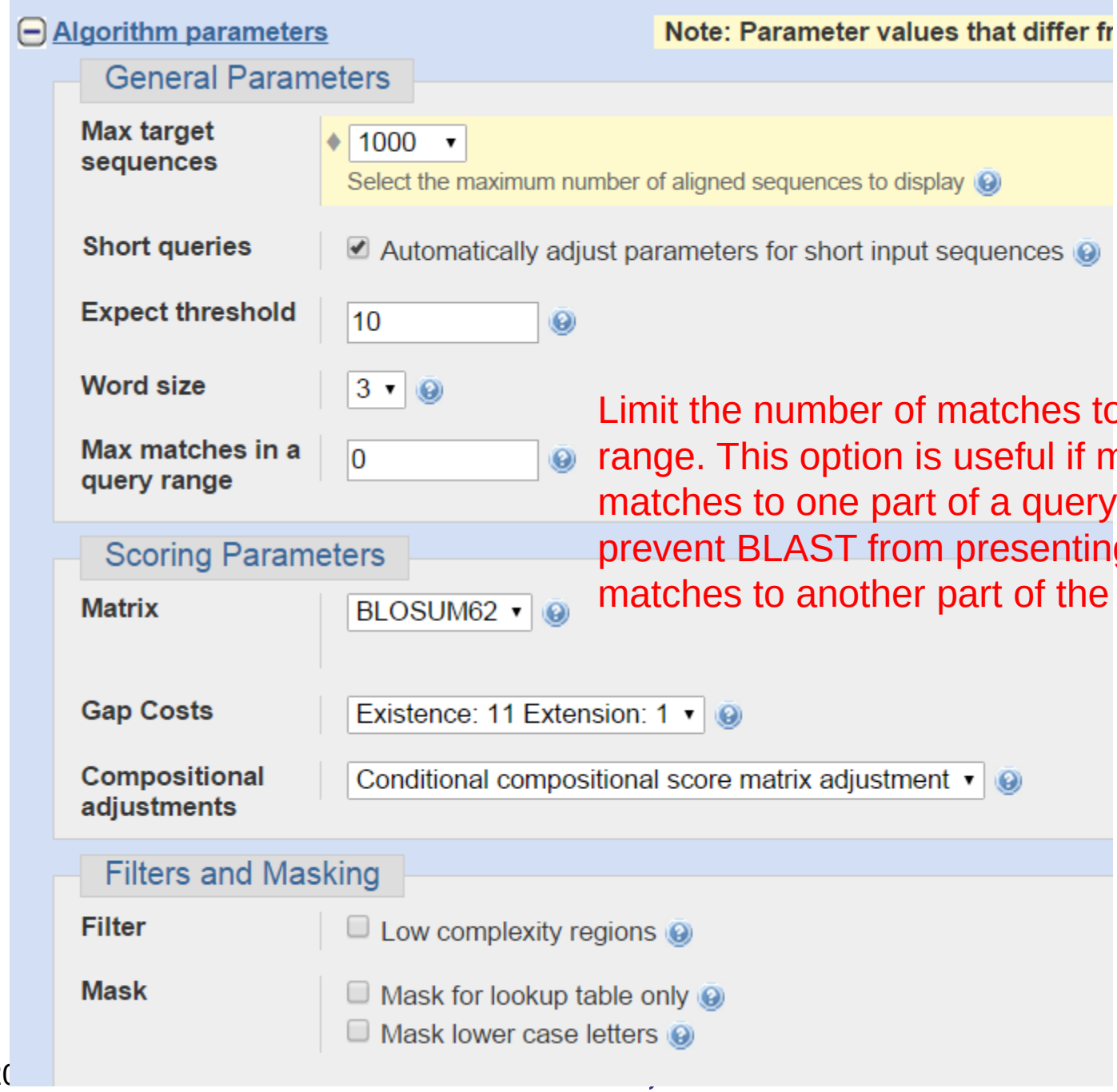
Advanced Options

- Max matches

  - "*Matches to one region of interest can be obscured by frequent matches to a different region of a protein. This feature offers a solution in which redundant database hits are discard*"

- Matrix

  - BLASTP: PAM30/70/250; BLOSUM45/50/62(default)/80/90
  - Switch to different flavors to see how the scoring system vary



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

Advanced Options

- Gap Penalty:
  - ❑ Biology: Mutational events can either cause an insertion or deletion; is an important event
  - ❑ So, to get the alignment right, we have to penalize heavily the insertion and less so for further extension
  - ❑ G + L* n

Affine: penalty for introducing is much greater than extension

  - Affine Gap (bigger constant term + linear term)
  - G = 10-15; L = 1-2

Advanced Options

- Gap Penalty (affine)
  - G + L * n
  - High penalty should help reduce gap openings
  - Low penalty will help us find far diverged match

Advanced Options

- Compositional Adjustments
  - BLAST matrices use an important quantity target frequencies
  - If the target frequencies of the query are different then the matrices and scoring scheme have to be adjusted as well
    - Malaria parasite Plasmodium falciparum is 80% AT (amino acids rich in turn AT-rich codons)

Low-complexity Regions

Rules for identities
>= 25% for proteins
>= 70% for nucleotides

Range 1: 317 to 1004 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
944 bits(2440)	0.0	Compositional matrix adjust.	688/688(100%)	688/688(100%)	0/688(0%)

```
Query  17    LFQPQSVSTAnssssnnnnssTPAALATHsptsnspvsgassasslltaaFGNLFGGSSA  76
            LFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNLFGGSSA
Sbjct  317   LFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNLFGGSSA  376

Query  77    KMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNN  136
            KMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNN
Sbjct  377   KMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNN  436

Query  137   NSIAPANSTPMSNGTNasispgsahssshshqgvspKGSRRVSACSDRSLEAAAADVAGG  196
            NSIAPANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGG
Sbjct  437   NSIAPANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGG  496

Query  197   SPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMqqqq  256
            SPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQ
Sbjct  497   SPPRAASVSSLNGGASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQ  556

Query  257   qqlqekeqlHSKLnnnnnnniaatannnnnttMESINLIDDSEMADIKIKSEPQTAPQPQ  316
            QQLQEKEQLHSKLNNNNNNNIAATANNNNNTTMESINLIDDSEMADIKIKSEPQTAPQPQ
Sbjct  557   QQLQEKEQLHSKLNNNNNNNIAATANNNNNTTMESINLIDDSEMADIKIKSEPQTAPQPQ  616

Query  317   QsphgsshssrsgsgsgshssmasdgslrrkssdsldsHGaqddaqdeedaaPTGQRSES  376
            QSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSDSLDSHGAQDDAQDEEDAAPTGQRSES
Sbjct  617   QSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSDSLDSHGAQDDAQDEEDAAPTGQRSES  676
```

Figure 4.5 from Book

Invoking comp. Statistics we can see that E-value is reduced from 0.05 to 0.0001

(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
```

(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
```

(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
```

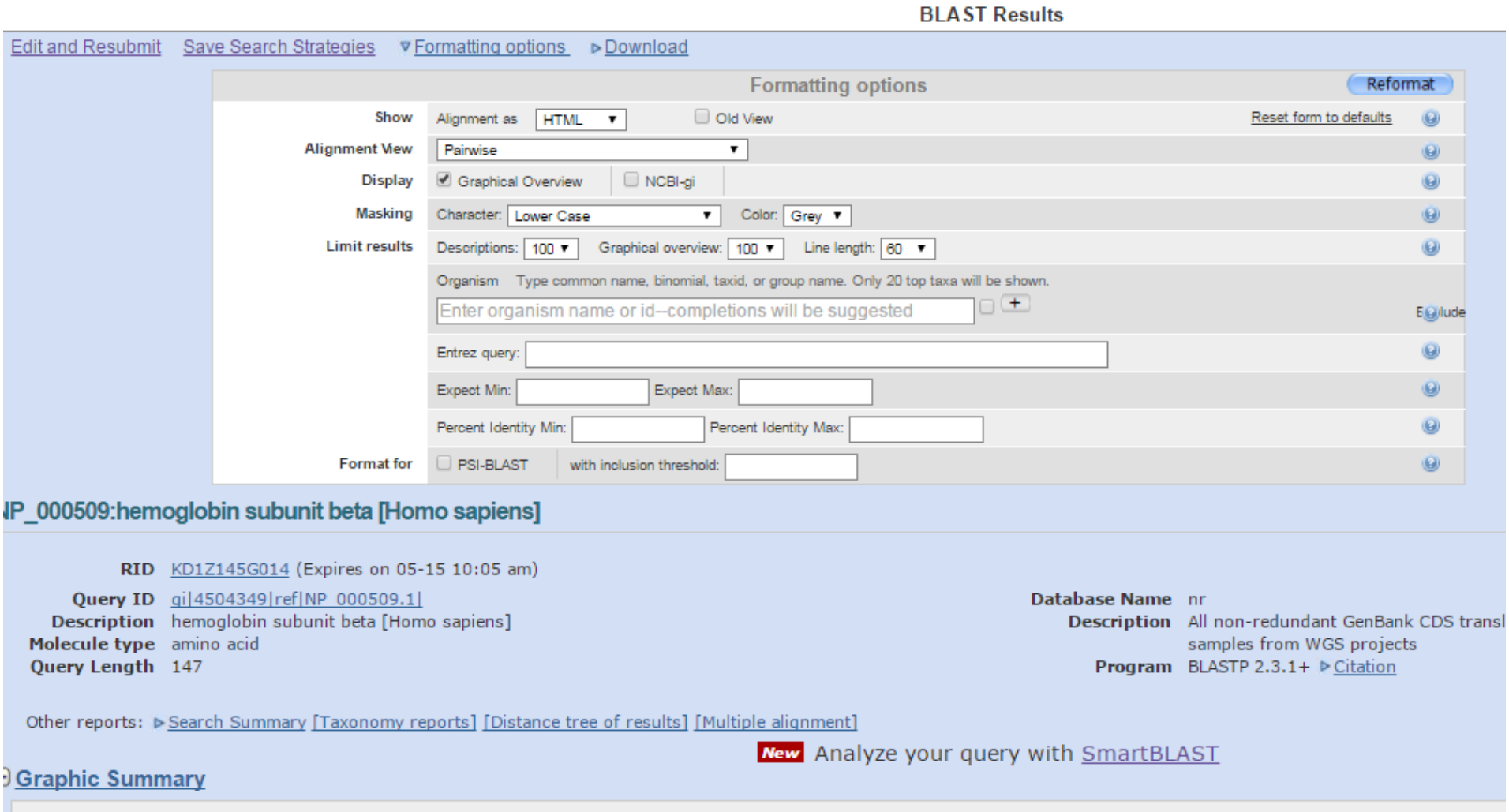
Advanced Options

- Filters
  - What are low-complexity or highly biased compositions
    - Dinucleotide repeats CACACACA
    - Alu repeats or TM region or Hydrophobic amino acid stretches, proline rich regions etc.
  - A mask (only to query) is applied
    - Done to avoid finding another sequence with repeats
  - Proteins: SEG program is used
  - DNA: DUST is employed

Advanced Options

- **Masking**
  - **Mask for lookup table only**
  - BLAST searches consist of two phases, finding hits based upon a lookup table and then extending them. This option masks only for purposes of constructing the lookup table used by BLAST so that no hits are found based upon low-complexity sequence or repeats (if repeat filter is checked). The BLAST extensions are performed without masking and so they can be extended through low-complexity sequence.
  - **Mask Lower Case**
  - With this option selected you can cut and paste a FASTA sequence in upper case characters and denote areas you would like filtered with lower case. This allows you to customize what is filtered from the sequence during the comparison to the BLAST databases.

Step 4b: Formatting Options



Query: ref|NP_000509 with **Entrez query** perutz mf[Author] Others: Default

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.3.1+ ▷Citation

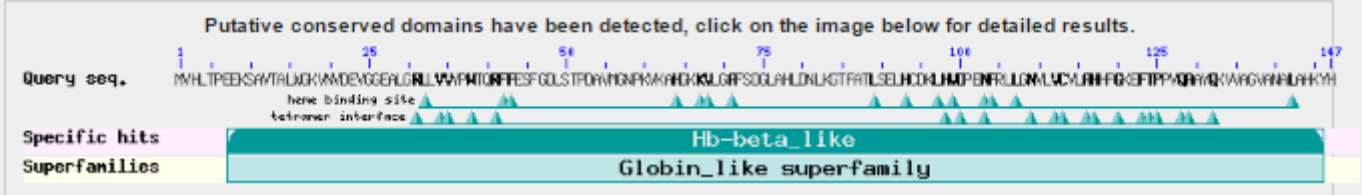
Search Parameters	
Program	blastp
Word size	6
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Filter string	F
Genetic Code	1
Window Size	40
Threshold	21
Composition-based stats	2

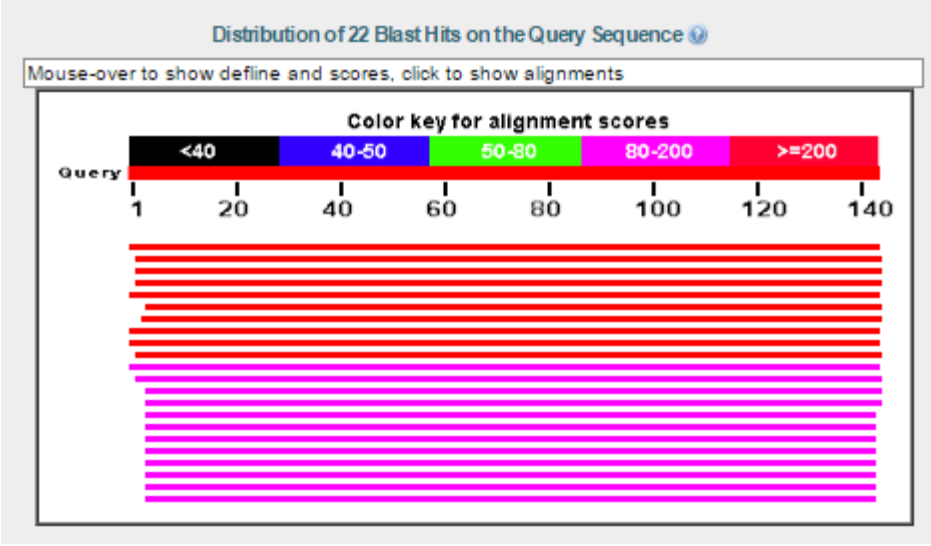
Database	
Posted date	May 12, 2016 5:27 PM
Number of letters	32,001,496,977
Number of sequences	87,229,180
Entrez query	perutz mf[Author]

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

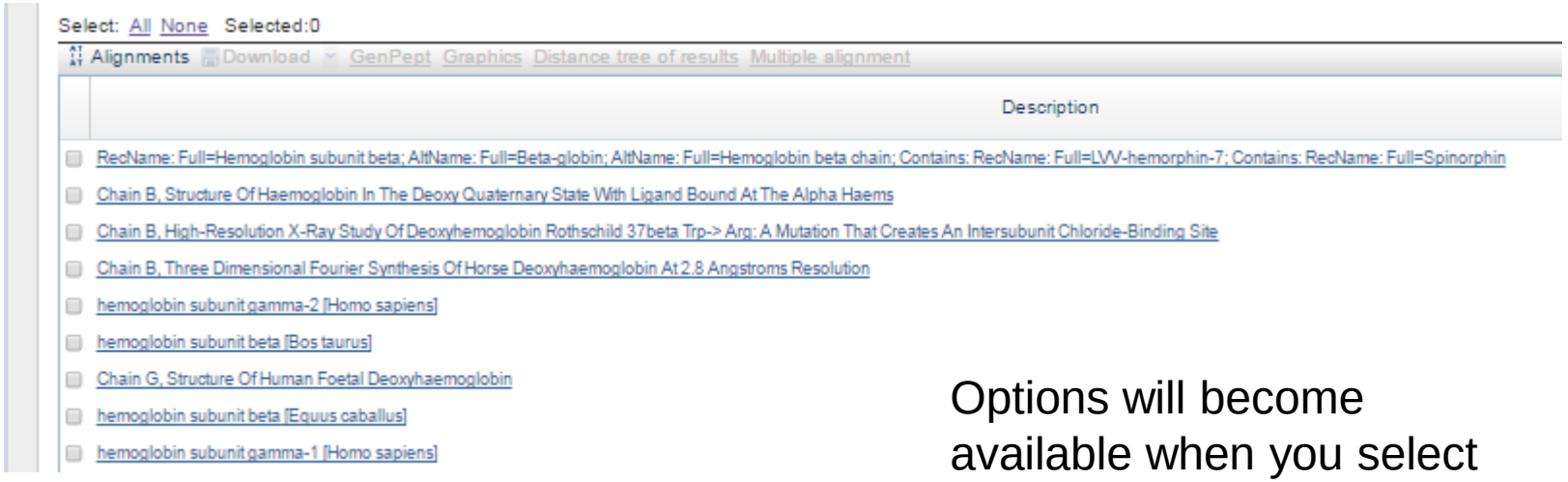
RID KD1Z145G014 (Expires on 05-15 10:05 am)
Query ID gi|4504349|ref|NP_000509.1|
Description hemoglobin subunit beta [Homo sapiens]
Molecule type amino acid
Query Length 147

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.3.1+ ▷Citation

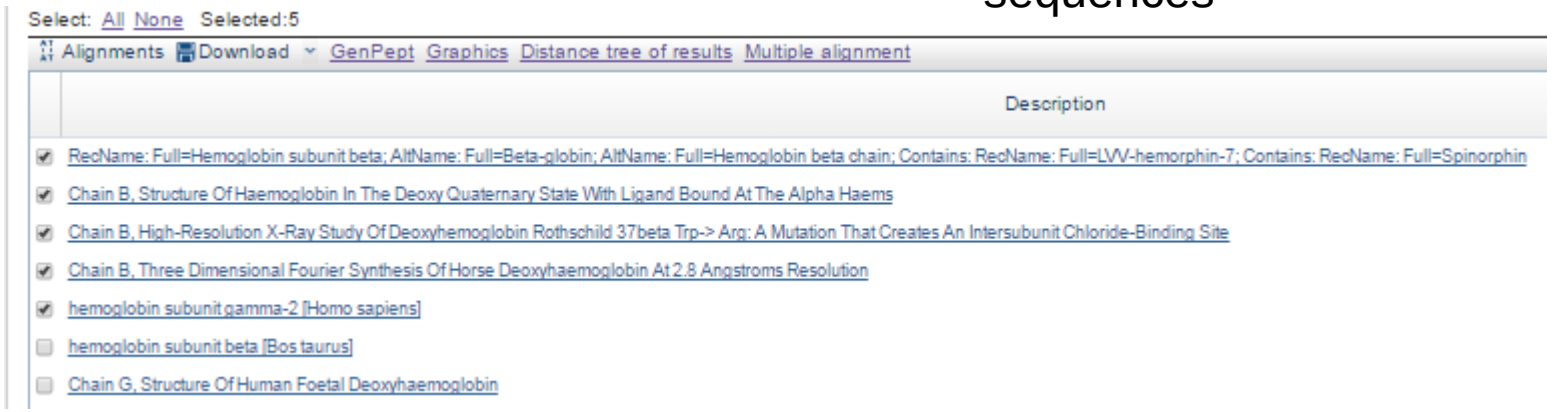




Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=Hemoglobin subunit beta; AltName: Full=Beta-globin; AltName: Full=Hemoglobin	301	301	100%	1e-110	100%	P68871.2
Chain B, Structure Of Haemoglobin In The Deoxy Quaternary State With Ligand Bound At The Alp	298	298	99%	1e-109	100%	1COH_B
Chain B, High-Resolution X-Ray Study Of Deoxyhemoglobin Rothschild 37beta Trp-> Arg: A Mutat	295	295	99%	3e-108	99%	1HBA_B
Chain B, Three Dimensional Fourier Synthesis Of Horse Deoxyhaemoglobin At 2.8 Angstroms Re	247	247	99%	3e-89	82%	2DHB_B
hemoglobin subunit gamma-2 [Homo sapiens]	235	235	100%	2e-84	73%	NP_000175.1
hemoglobin subunit beta [Bos taurus]	234	234	97%	5e-84	85%	NP_776342.1
Chain G, Structure Of Human Foetal Deoxyhaemoglobin	234	234	98%	5e-84	74%	1FDH_G
hemoglobin subunit beta [Equus caballus]	233	233	100%	1e-83	84%	NP_001157490.1
hemoglobin subunit gamma-1 [Homo sapiens]	232	232	100%	2e-83	73%	NP_000550.2



Options will become available when you select sequences



You can select sequences and send them for multiple sequence alignments

View Format: Compact ▾ Conservation Setting: 2 Bits ▾

P68871	1	mVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD	80
1COH_B	1	-VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD	79
1HBA_B	1	-VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPRTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD	79
2DHB_B	1	-VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLD	79
NP_000175	1	mGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPKVKAHGKKVLTSLGDAIKHLD	80

P68871	81	NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH	147
1COH_B	80	NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH	146
1HBA_B	80	NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH	146
2DHB_B	80	NLKGTFAALSELHCDKLHVDPENFRLLGNVLALVVARHFGKDFTPELQASYQKVVAGVANALAHKYH	146
NP_000175	81	DLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMVTGVASALSSRYH	147

View Format: Compact ▾ Conservation Setting: Identity ▾

P68871	1	mVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD	80
1COH_B	1	-VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD	79
1HBA_B	1	-VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPRTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD	79
2DHB_B	1	-VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLD	79
NP_000175	1	mGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPKVKAHGKKVLTSLGDAIKHLD	80

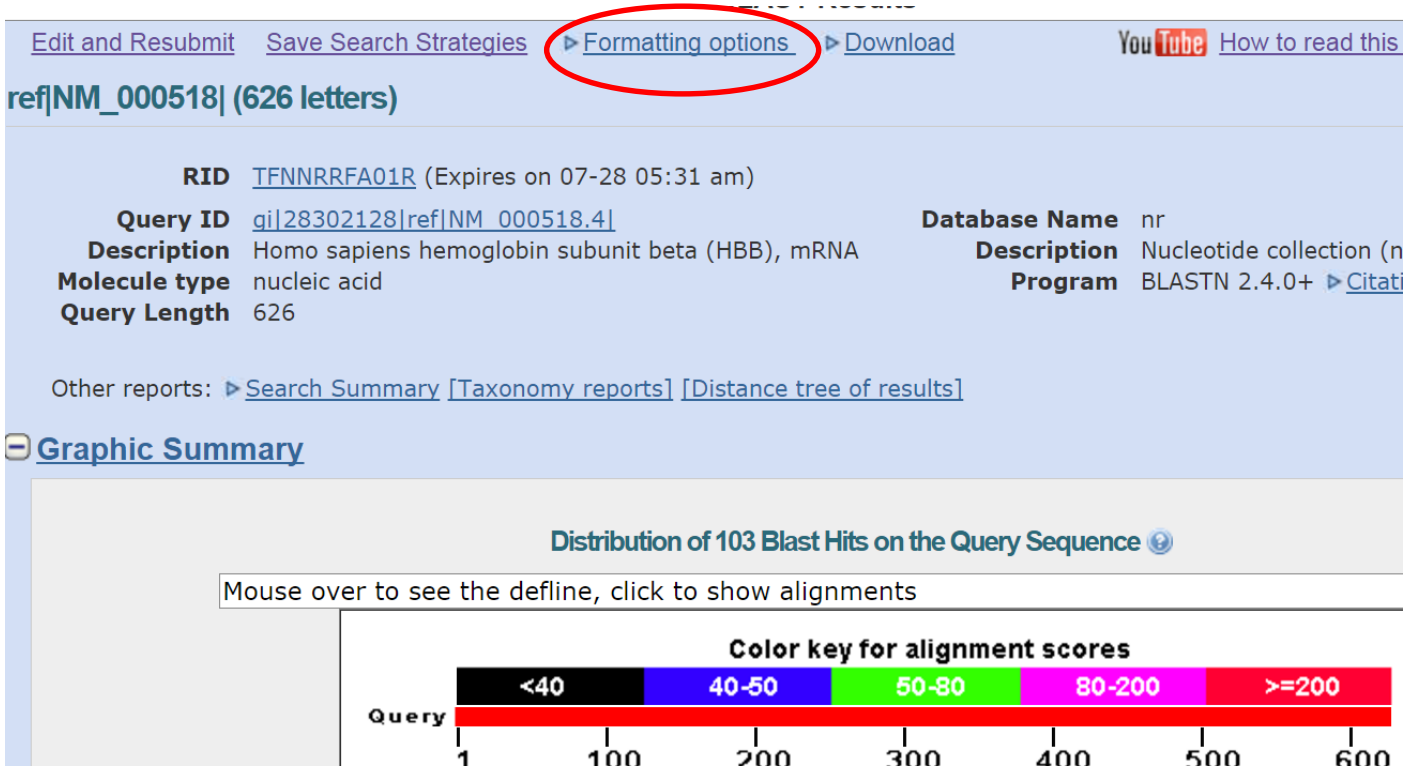
P68871	81	NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH	147
1COH_B	80	NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH	146
1HBA_B	80	NLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH	146
2DHB_B	80	NLKGTFAALSELHCDKLHVDPENFRLLGNVLALVVARHFGKDFTPELQASYQKVVAGVANALAHKYH	146
NP_000175	81	DLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFGKEFTPEVQASWQKMVTGVASALSSRYH	147

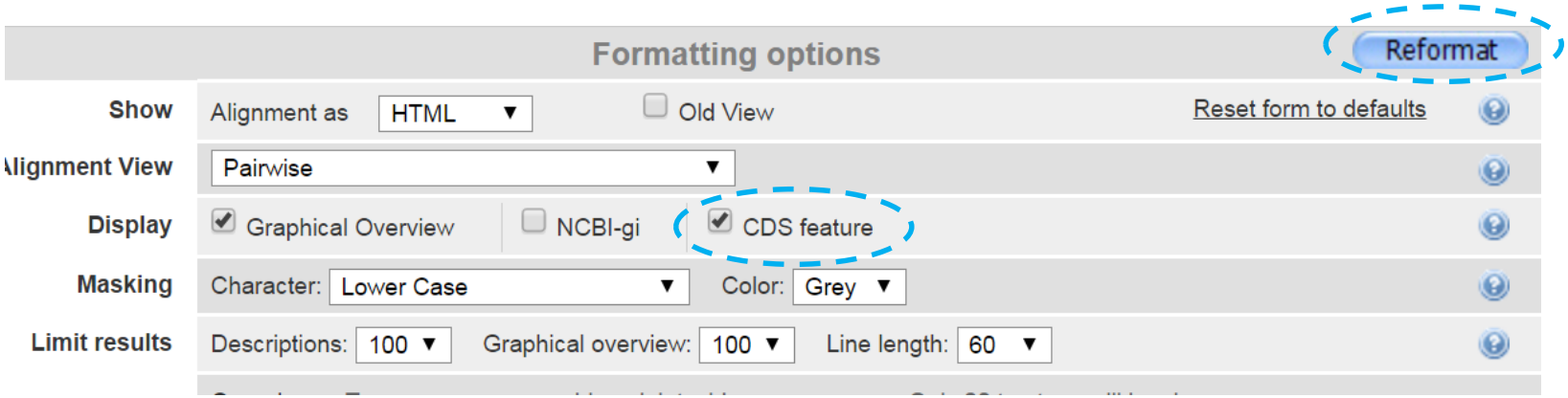
Important Parameters

- E-value cut-off
  - Low (Example,1E-06)
- Word size (WS)
  - Low WS: Higher sensitivity
  - High WS: Higher Specificity
- Match/mismatch Score and Gap Costs
  - Extension scores
- Filters and masking
  - Substitute for RepeatMasker

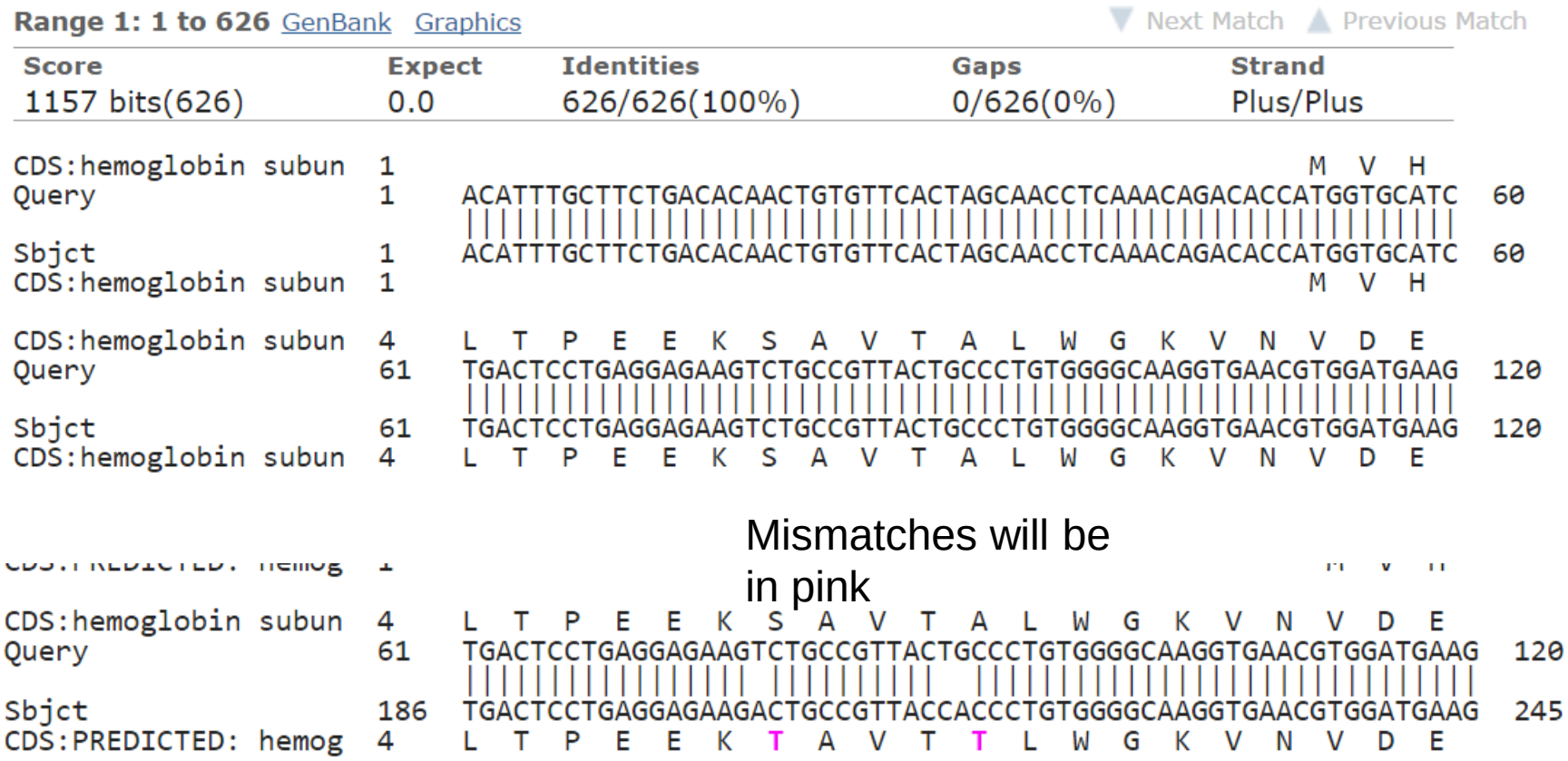
NM_000518.4 HBB (homosapiens) with nr and other default optons

BLASTN search


Edit and Resubmit
Save Search Strategies
Formatting options
Download
How to read this
ref|NM_000518| (626 letters)
RID TFNNRRFA01R (Expires on 07-28 05:31 am)
Query ID gi|28302128|ref|NM_000518.4|
Description Homo sapiens hemoglobin subunit beta (HBB), mRNA
Molecule type nucleic acid
Query Length 626
Database Name nr
Description Nucleotide collection (n
Program BLASTN 2.4.0+
Other reports: Search Summary [Taxonomy reports] [Distance tree of results]
Graphic Summary
Distribution of 103 Blast Hits on the Query Sequence
Mouse over to see the defline, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
100
200
300
400
500
600


Formatting options
Reformat
Show
Alignment as HTML
Old View
Reset form to defaults
Alignment View
Pairwise
Display
Graphical Overview
NCBI-gi
CDS feature
Masking
Character: Lower Case
Color: Grey
Limit results
Descriptions: 100
Graphical overview: 100
Line length: 60

Reformatting options



Standalone BLAST

- Will demo after the lecture

NIH U.S. National Library of Medicine | NCBI National Center for Biotechnology Information

BLAST®

Home Recent Resul

BLAST documentation

Getting Started

- Guide to BLAST home and search pages
- BLAST interface description
- Blast report description

Getting Help

- Email blast-help
- Mailing list
- YouTube BLAST tutorials

About BLAST

- Frequently Asked Questions
- NCBI Handbook: BLAST
- The Statistics of Sequence Similarity Scores
- NAR 2004 Web server issue

Other BLAST information

- Download BLAST Software and Databases
- Developer information
- BLAST Searches at a Cloud Provider

NCBI → BLAST → HELP → Download BLAST Software and Databases

BLAST+ executables

BLAST+ is a suite of command-line tools to run BLAST. For details, please see the BLAST+ user manual, th article in BMC Bioinformatics (PubMed link). BLAST+ is the most current version of BLAST and is the only s

Installers and source code are available from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ .

BLAST ALGORITHM (using BLASTP as an example)

First Phase

- Query

```
>QUERY
MESADFYEAEPRPPMSSHLQSPPHAPSSAAFGFPRGAGPAQPPAPPAAPEPLGGICEHET
SIDISAYIDPAAFNDEFLADLFQHSRQQEKAKAAVGPTGGGGGGDFDYPGAPAGPGGAVM
PGGAHGPPPGYGCAAAGYLDGRLEPLYERVGAPALRPLVIKQEPREEDEAKQLALAGLFP
YQPPPPPPPSHPHPHPPPAHLAAPHLQFQIAHCGQTTMHLQPGHPTPPPTPVPSPHPAPA
LGAAGLPGPGSALKGLGAAHPDLRASGGSGAGKAKKSVDKNSNEYRVRRERNNIAVRKSR
DKAKQRNVETQQKVLELTSDNDRLRKRVEQLSRELDTLRGIFRQLPESSLVKAMGNCA
```

- Broken down into **word pairs**
  - NT: 16-256; Proteins: 2-3:
  - Threshold (T=11)
  - Sliding window approach
    - For each word (ex. SAY), the synonyms were formed and high scoring (BLOSUM matrix) words will be chosen.

SAYIDPAAFND

SAY Score
AYI Score
YID Score

IDP Score
DPA Score
PAA Score
AAF Score
AFN Score
FND Score

- Scores using Matrix for word pairs are collected

Word size 3, for 20 aa there can be 203 = 8000 possible words

How does BLAST work?

- SAY Score: 4+4+7 = 15
- AYI Score: 4+7+4 = 15
- YID Score: 7+4+6= 17

Scoring Matrix to calculate Scores

- Now establish a cut-off (say 15)
  - Then only YID and other words that are above the cutoffs are retained

SAYIDPAAFND

SAY Score
AYI Score
YID Score
IDP Score
DPA Score
PAA Score
AAF Score
AFN Score
FND Score

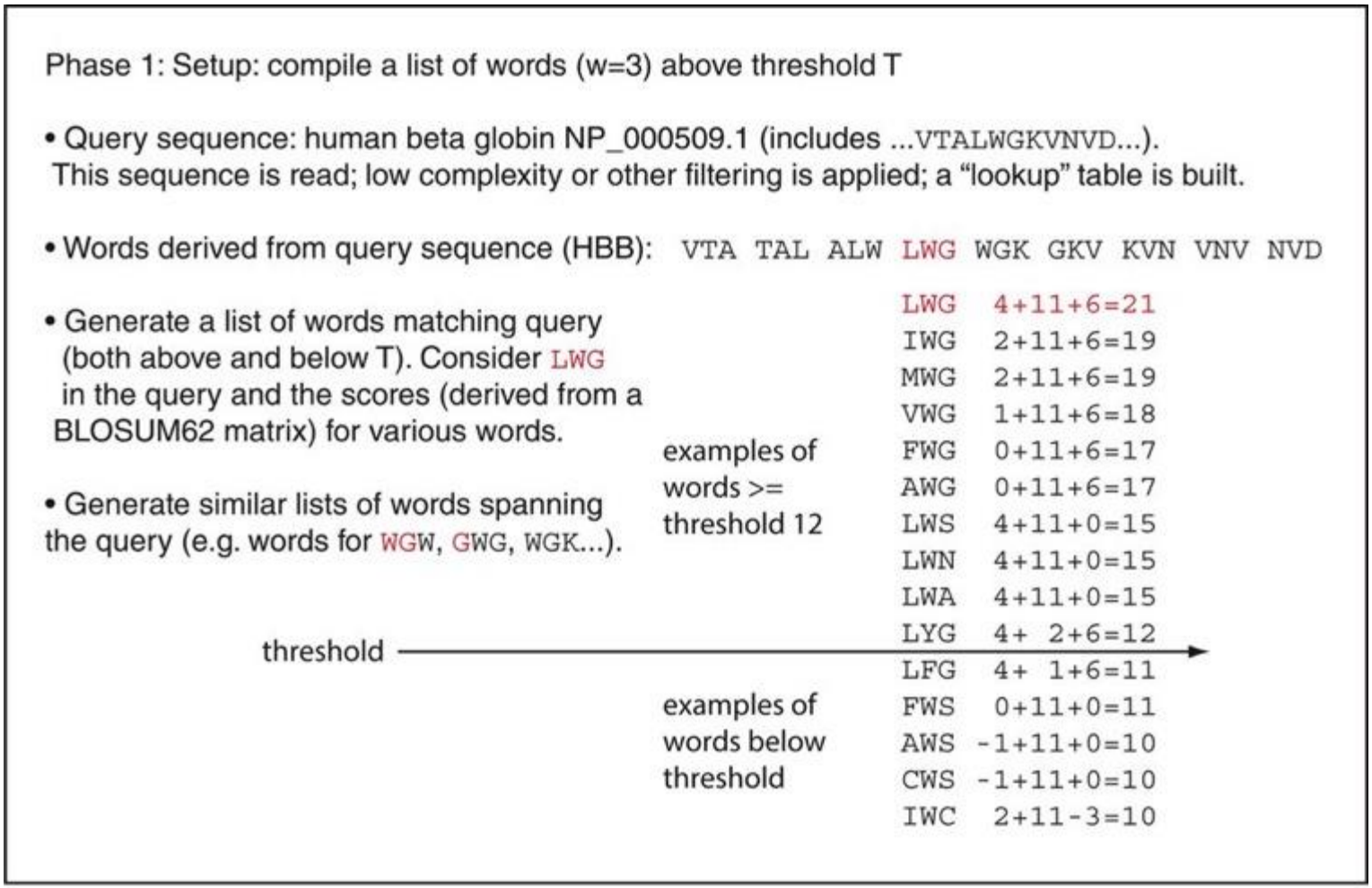
BLASTN

- First phase is slightly different than BLASTP
- Algorithm demands exact matches
  - Default word size is 11 (adjustable by user)
- Choosing a lower word length
  - Slower more accurate

Phase 2

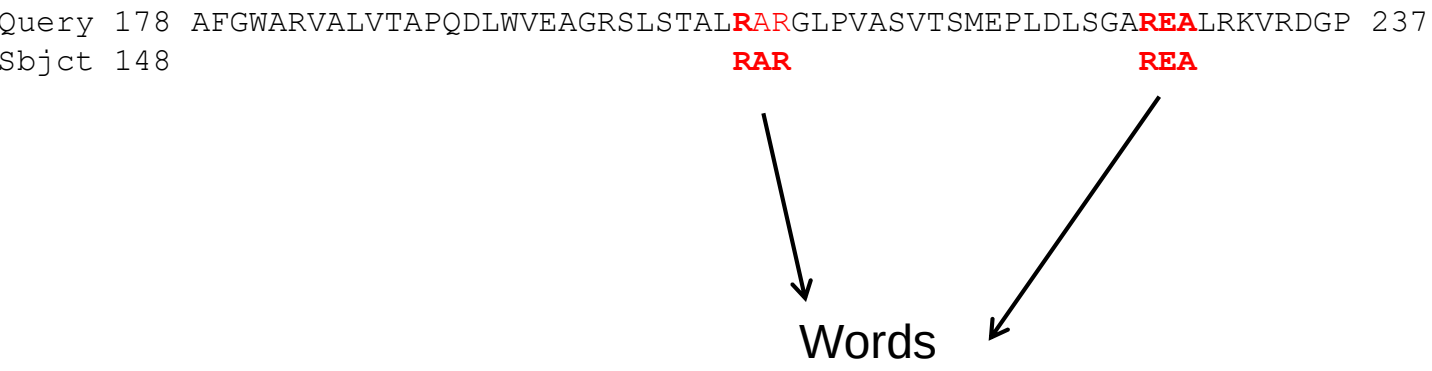
- The selected words are now used to search for sequences that contain these words
- Create a hash table index with the locations of the hits for each word
- Perform two more searches
  - Un-Gapped extensions (first)
  - Followed by gapped extensions
  - Hits above certain score are saved

Fig 4.12 from the Pevsner Book III Edition PLEASE DO NOT DISTRIBUTE Copyright figure



BLAST

- Scans the DB for matches to the words that are present above the Threshold Values
- Requires two hits within the target sequence (the new searched sequence)
- BLAST will set aside sequences with matches above Threshold for further analysis



No need for exact match, but have to be in the list

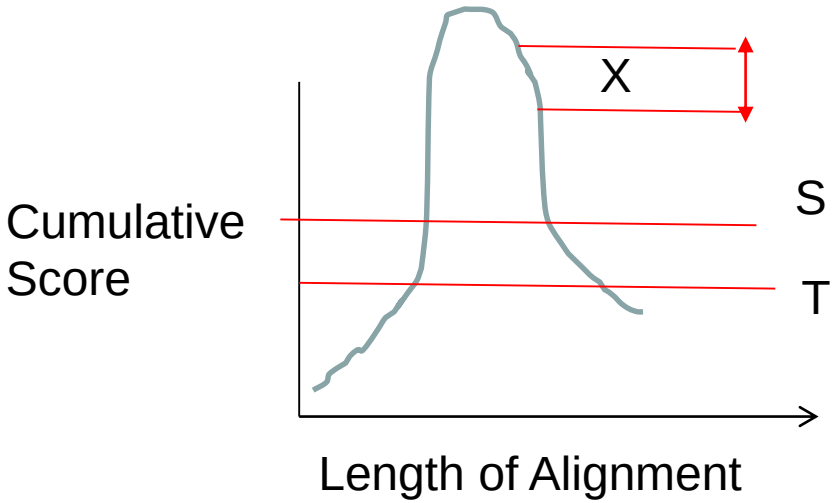
Example:

Sequence = gaacgcctgcgcgatcagcataaaaaataa
word length = W = 7 (there are 24 words possible)

For 'word' = ctgcgcg, a match in the database is found and then a local alignment done until a gap appears.
So...

```
query  = ctgcgcg       7       ctgcgcgatcag 19
match  = ctgcgcg       2598356 ctgcgcgatcag 2598367
```

Peak will tell us what Length we are going to focus



Based on Andy Baxevanis Book and Seminars

High Scoring Segment

Extension

Three letter scores that are greater than T will be carried over to the next step

We keep adding in both directions, more matches than mismatches, so the score keeps going up

Neighborhood threshold (S) Everything above S will be reported in BLAST results

When the score starts decreasing then we go back and pick a Window (X)

BLAST

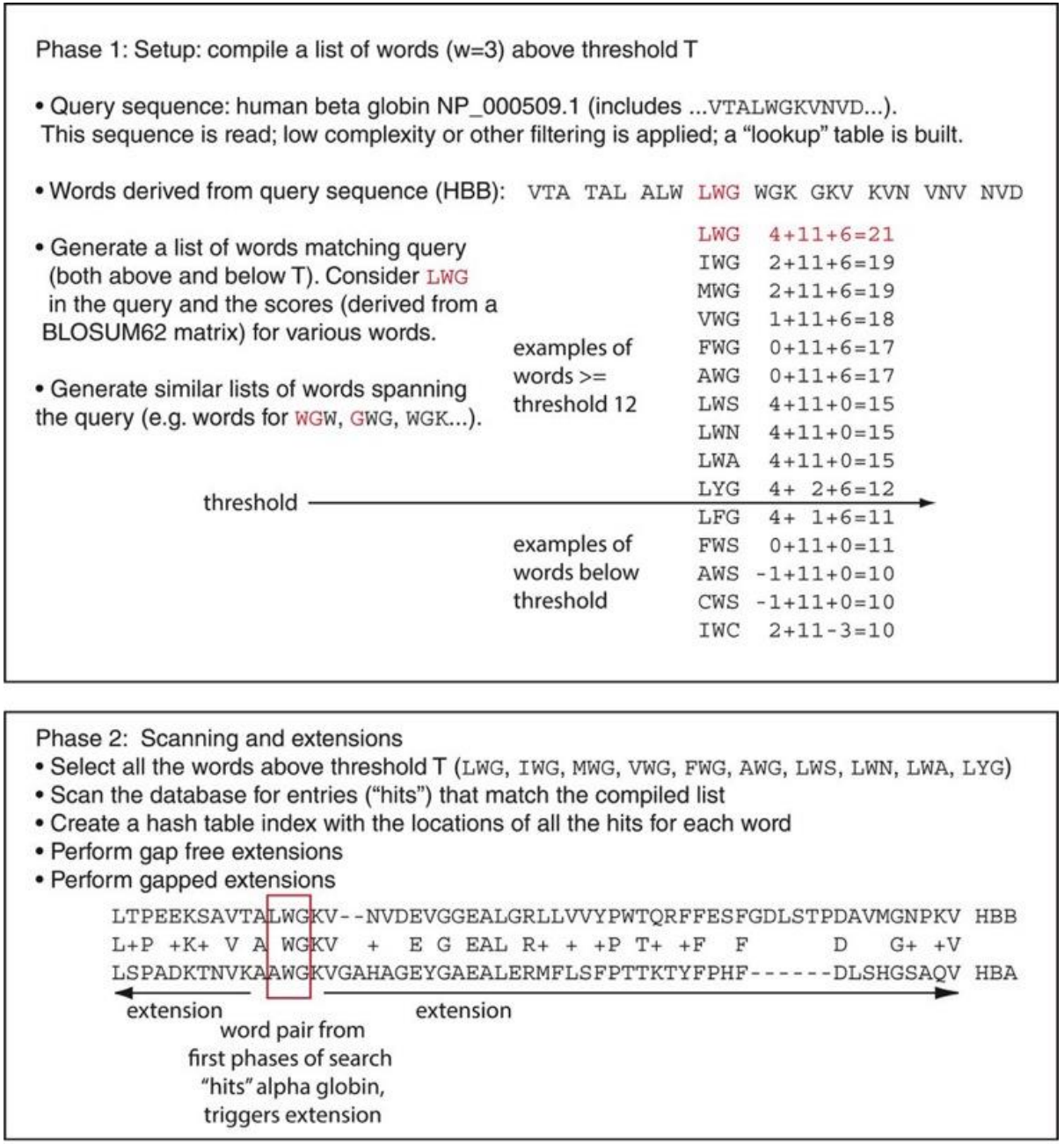
Sequence ID: pdb|3MJP|ALength: 141 Number of Matches: 1

```
Query 178 AFGWARVALVTAPQDLWVEAGRSLSTALRARGLPVASVTSMEPLDLSGAREALRKVRDGP 237
Sbjct 148                                 RAR                 REA
```

```
Query 178 AFGWARVALVTAPQDLWVEAGRSLSTALRARGLPVASVTSMEPLDLSGAREALRKVRDGP 237
            G AR               GR      RARGLPVA VTSMEP DLSGAREA      GP
Sbjct 148 --GAAR--------------GR-----WRARGLPVALVTSMEPSDLSGAREA--SASAGP 184
```

Extension until the score drops

Fig 4.12 from the Pevsner Book III Edition PLEASE DO NOT DISTRIBUTE Copyright figure



Phase 3

- Traceback
  - Identify the locations of INDELS and matches from phase 2
  - If applicable, use composition-based statistics (BLASTP, TBLASN)
  - Generate final gapped alignment

Summary

- BLAST
  - Heuristic algorithm (optimized for speed/sensitivity)
  - Threshold is increased
    - Speed in increased/fewer hits
      - Distantly related are missed
  - Threshold is lowered
    - Speed is lowered/large number of matches
    - Sensitivity is increased

Threshold size
T = 10 (BLASTP)
will compile words
>= 10

Impact of Threshold Score(T)

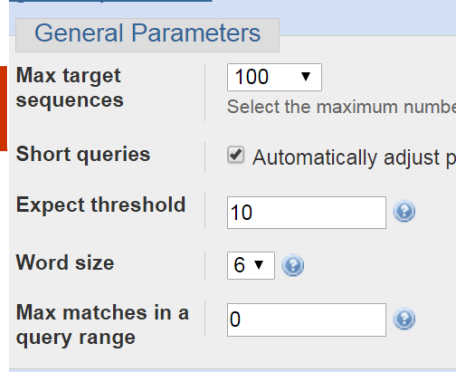
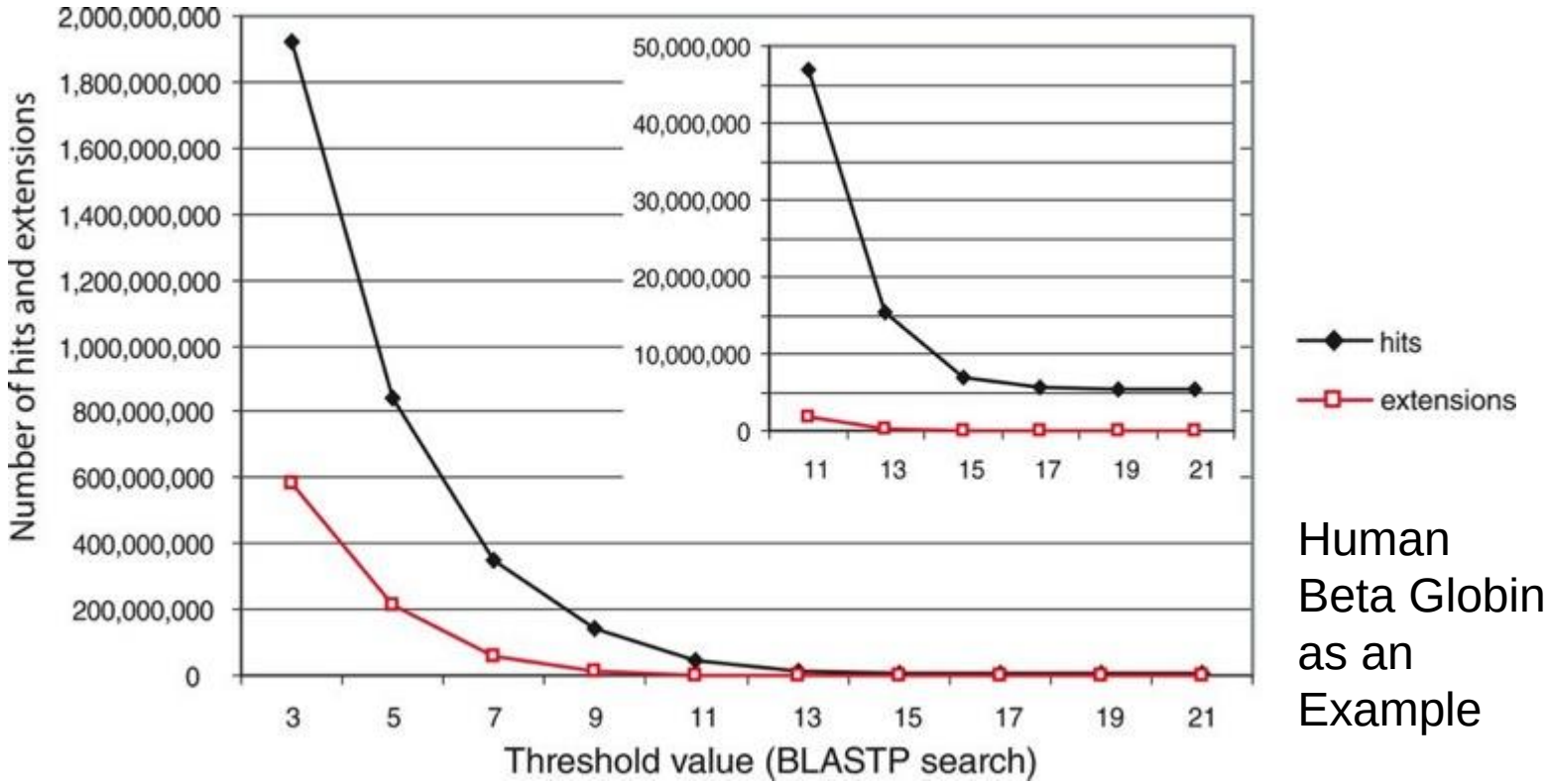


Fig 4.13 from the Pevsner Book III Edition PLEASE DO NOT DISTRIBUTE Copyright figure



Human Beta Globin as an Example

What matrix to pick?

BLOSUM	Suitable For; Based on experience	% similarity
90	Short alignments; highly similar	70-90
80	Best for identifying family members	50-60
62	**MOST EFFECTIVE for identifying all potential similarities (default in NCBI)**	**30-40**
30	Longer/weaker local alignments	<30

Is One matrix is enough?

- David Altschul prescribes a “Triple Strategy”
- Pick the default and a higher/lower BLOSUMn
- Analyze and pick the appropriate matrix

Statistics of Alignments

Let us begin with a simple diagram that explains global alignment

Global Alignment

- Distribution behavior of global alignments is not known (not Gaussian/normal)
  - Usually approximated using simulations

Local Alignment

- Statistics/distributions are known
  - Altschul many papers
- Start with Ungapped alignments
  - Random
  - Gapped alignments
  - Proteins

Local Alignment

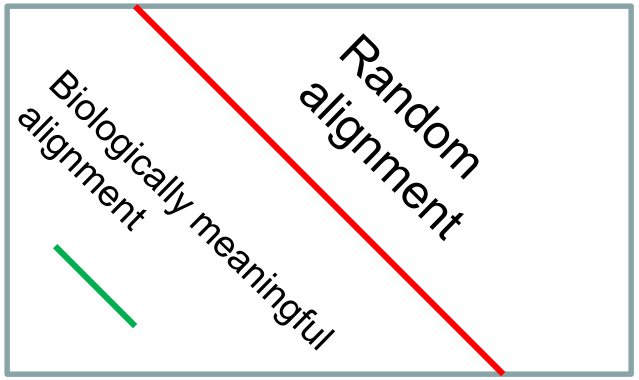
- To access how high a score can occur by chance, we need random sequences and their scores
  - Conditions
    - **Expected score** for aligning random pair of amino acids has to be NEGATIVE

Expected Score matrix constraints

- For alignment algorithms that seek to capture the local alignments of variable length should have a negative expected score (a necessary condition)

$$\sum_{i,j} p_i p_j s_{i,j} < 0$$

For local alignments of random sequences, Negative Expected Score



Condition II: At least one of the score has to be positive (s_i,j)

Log-odds Scores

- "*with the previous two assumptions, the scores of any substitution matrix (with a negative expected value and at least one positive score) can be written in the form*" *(Karlin & Altschul, PNAS, 872264(1990)*

$$s_{i,j} = \frac{\left(\ln \frac{q_{i,j}}{p_i p_j}\right)}{\lambda} = \log\left(\frac{q_{i,j}}{p_i p_j}\right)$$

Change of base

λ: Scaling Parameter

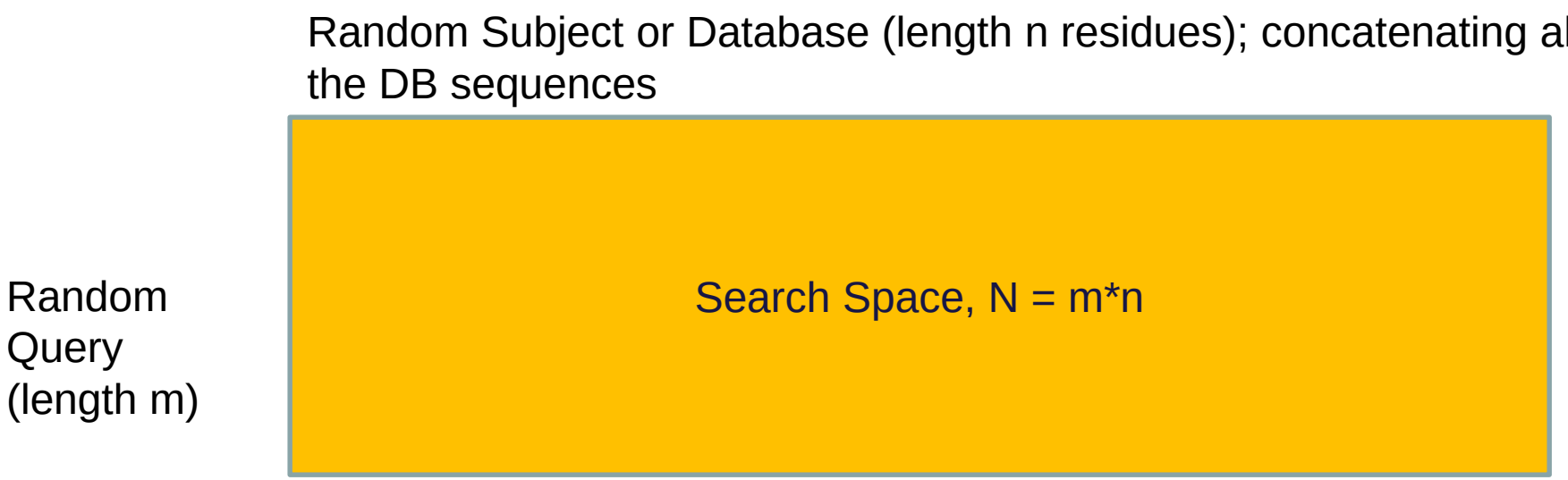
x, a, b *are all positive*

$a \neq 1; b \neq 1$

$$\log_8 x = \frac{\ln x}{\ln 8} = \frac{1}{\ln 8} \ln x$$

What is a search space

- "*Given a scoring system, how many **distinct local alignments** with score >= S (S is some number) can one **find by chance** by comparing **two random sequences** of length m and n*" S. Altschul

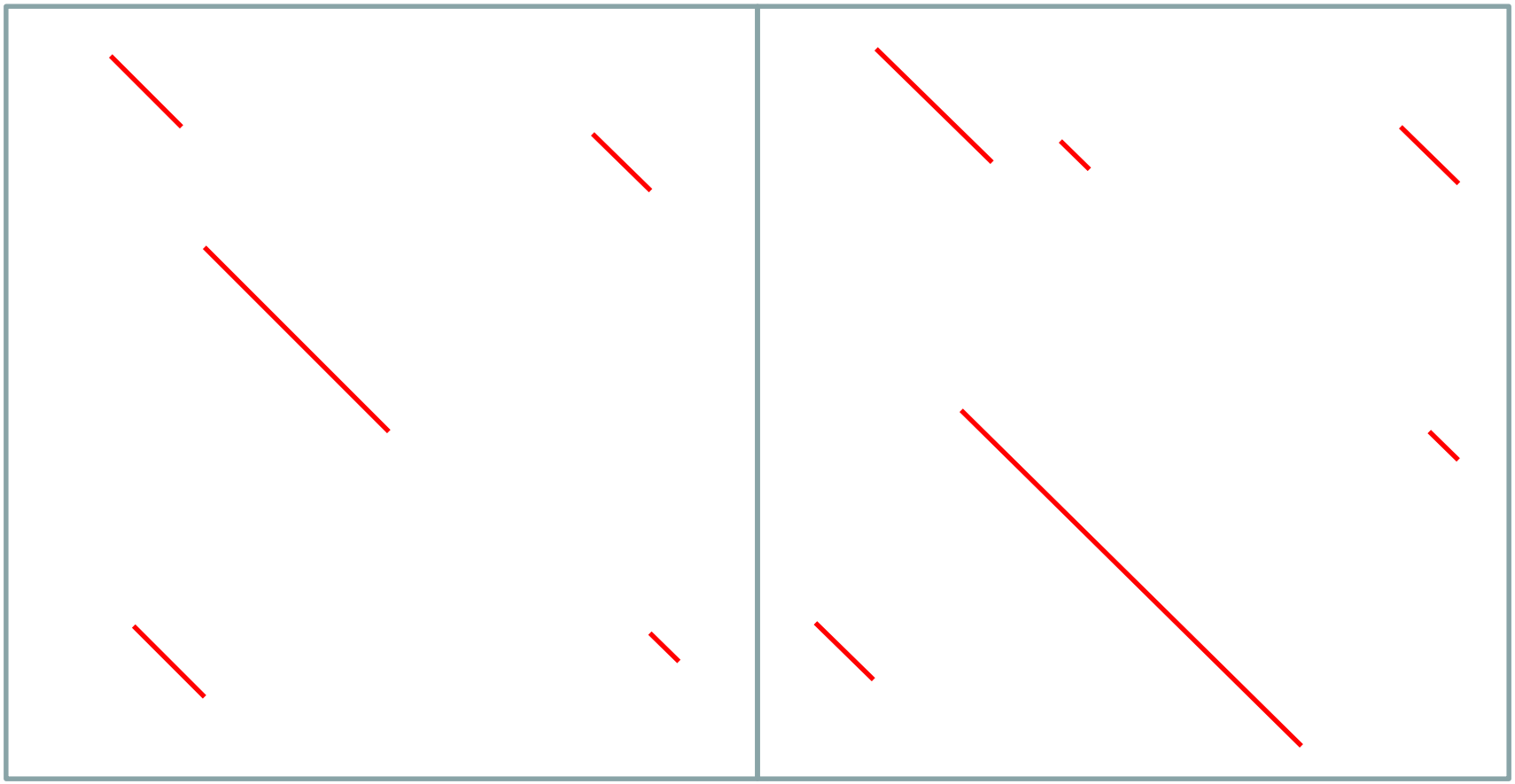


- Answer: E(S,m,n), E = expected score will depend on S, m and n

Not a proof

Number of Random High-Scoring Alignments ~ Search Space Size

$$E(S,m,n)\ \alpha\ mn$$ Asymptotic result



Double the size then HAS will roughly double.

The # of random alignments with Score >= S should decrease Exponentially with S

Any scoring system, the probability that the optimal local alignment that starts at a particular position has a score ≥ S decreases exponentially with S

$$E(S,m,n) \ \alpha \ e^{-\alpha S}$$

Alpha turns out to the same as λ

$$E(S,m,n) \ \alpha \ e^{-\alpha S}$$

$$E(S,m,n) \ \alpha \ mn$$

$$\longrightarrow$$

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

E: "Number of different alignments with scores equal or greater than some score S that are expected to occur in a DB search by chance"

Poisson

- Prob of finding 0 alignments (or none) with score ≥ S is e^{-E}

$$P(\text{k events in interval}) = \frac{e^{-L} L^k}{k!}$$

The average number of events in an interval is designated as L.

- Prob. of finding at least one alignment with score ≥ S is $p = 1 - e^{-E}$
- This is called "p-value" associated with S.
- When E ≤ 0.1, p ~ E

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

Next page for derivation

Derivation of Normalized Scores

- To calculate E-value associated with a S, we need to know λ and K.
  - Refer you to Altschul papers on derivation

$$S' = \frac{(\lambda S - \ln K)}{\ln 2}$$

- But, these values can be wrapped into a reduced form as shown above, then S' can be easily connected to E
- Refer back, N = Search Space

$$E = \frac{N}{2^{S'}}$$

Number of alignments with ≥S

$$E = (kmn)e^{-\lambda S}$$

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

$$E = Ne^{-\ln_e k}e^{-\lambda S}$$

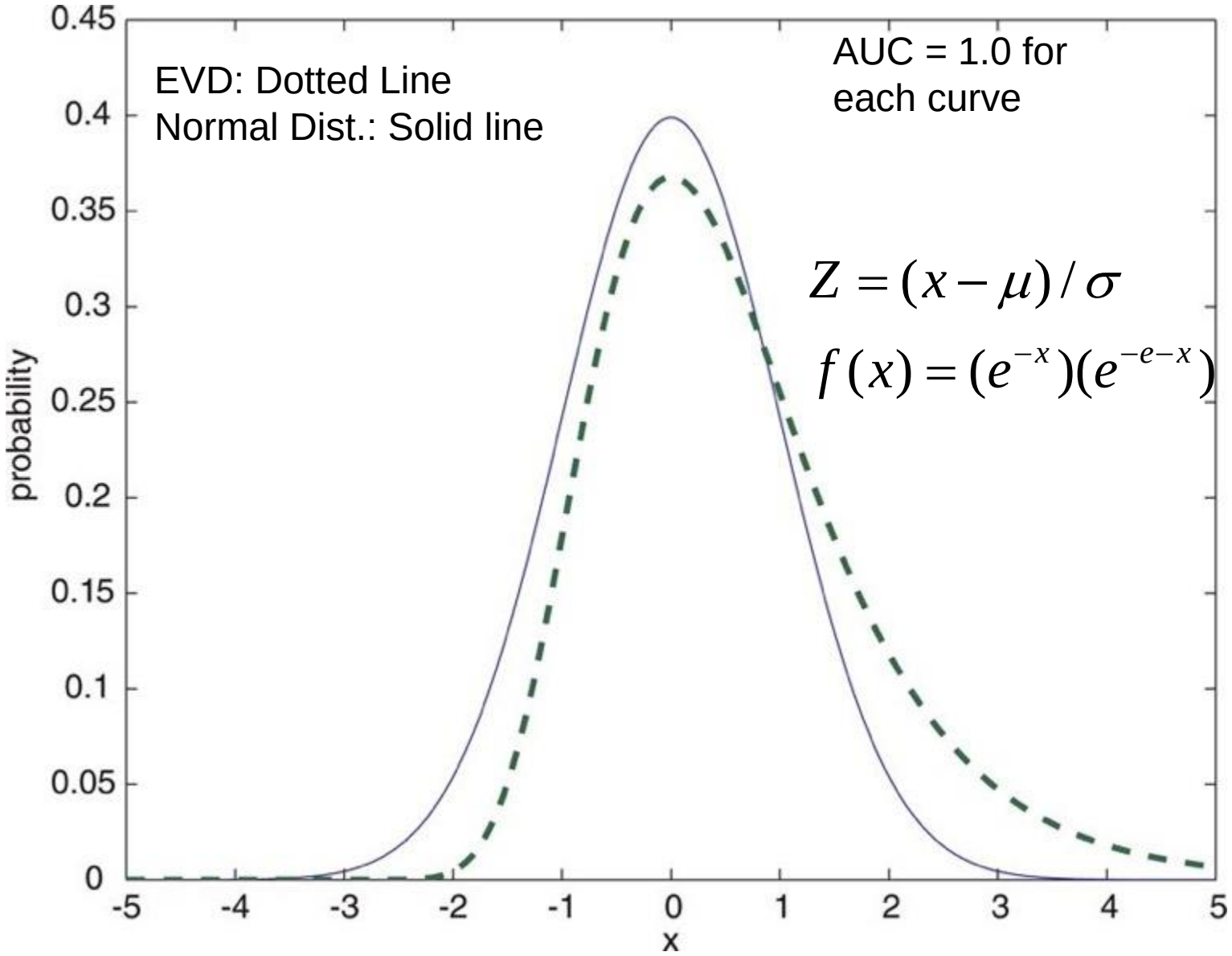
$$E = Ne^{-\left[\frac{(\lambda S - \ln_e k)}{\ln_e 2}\right]\ln_e 2}$$

$$E = N2^{-S'}$$

$$E = \frac{N}{2^{S'}}$$

$$S' = \frac{(\lambda S - \ln K)}{\ln 2}$$

Query compared to a set of random seq of same length as query. The alignment scores will take a extreme value distribution (EVD)



Ungapped → Gapped

- Everything discussed up to this apply to gapped alignment (as long gap scores are negative enough; not close to zero)
- Not proved up until now!
- According to Altschul, gapped alignments there is no way to statistical theory (?) to calculate the statistical parameters, K and λ but we can estimate them.

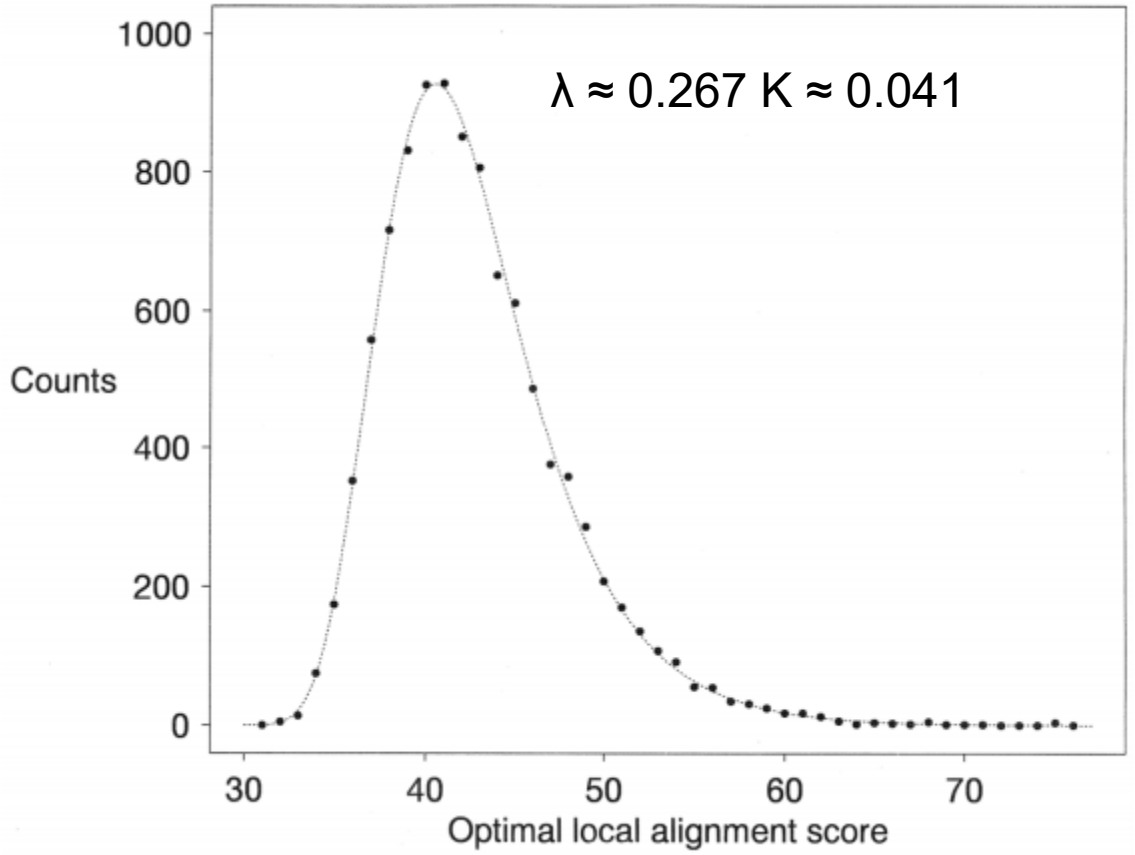
$$S' = \frac{(\lambda S - \ln K)}{\ln 2}$$

Local Alignment with Gaps

Simulation:

10,000 pairs of random protein sequence, each of length 1000 were compared using BLOSUM-62 substitution score, gap score of -11-k for a gap of length k

After several simulations, Altschul has plotted a histogram of how many times, he saw the scores. He fitted them to EVD and estimated Lambda and K



Random Sequence, Ungapped to Real proteins

- The theory still holds except for the following cases
  - Low-complexity filtering
  - Mask the sequence segments by giving a negative score
- Let us start with the following equation
  - $E = kmNe^{-\lambda S}$
  - E gives the # of HSPs found purely by chance

$$E = kmne^{-\lambda S'}$$

$$S' = \frac{(\lambda S - \ln K)}{\ln 2}$$

- Raw score (S)
  - Sub. Matrix (gap penalty etc)
- Bit Score (S', scaled value)
  - Bit scores can be compared even using different scoring matrices
- E values are derived from Bit Scores (S')
- Prob of chance alignment occurring with the score or better

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

Why BLAST doesn't report P?
It is easier to think of the number of HSAs rather than the probability values; High-Scoring Alignment (HAS)

E	*p*
10	0.99995460
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258
0.05	0.04877058
0.001	0.00099950
0.0001	0.0001000

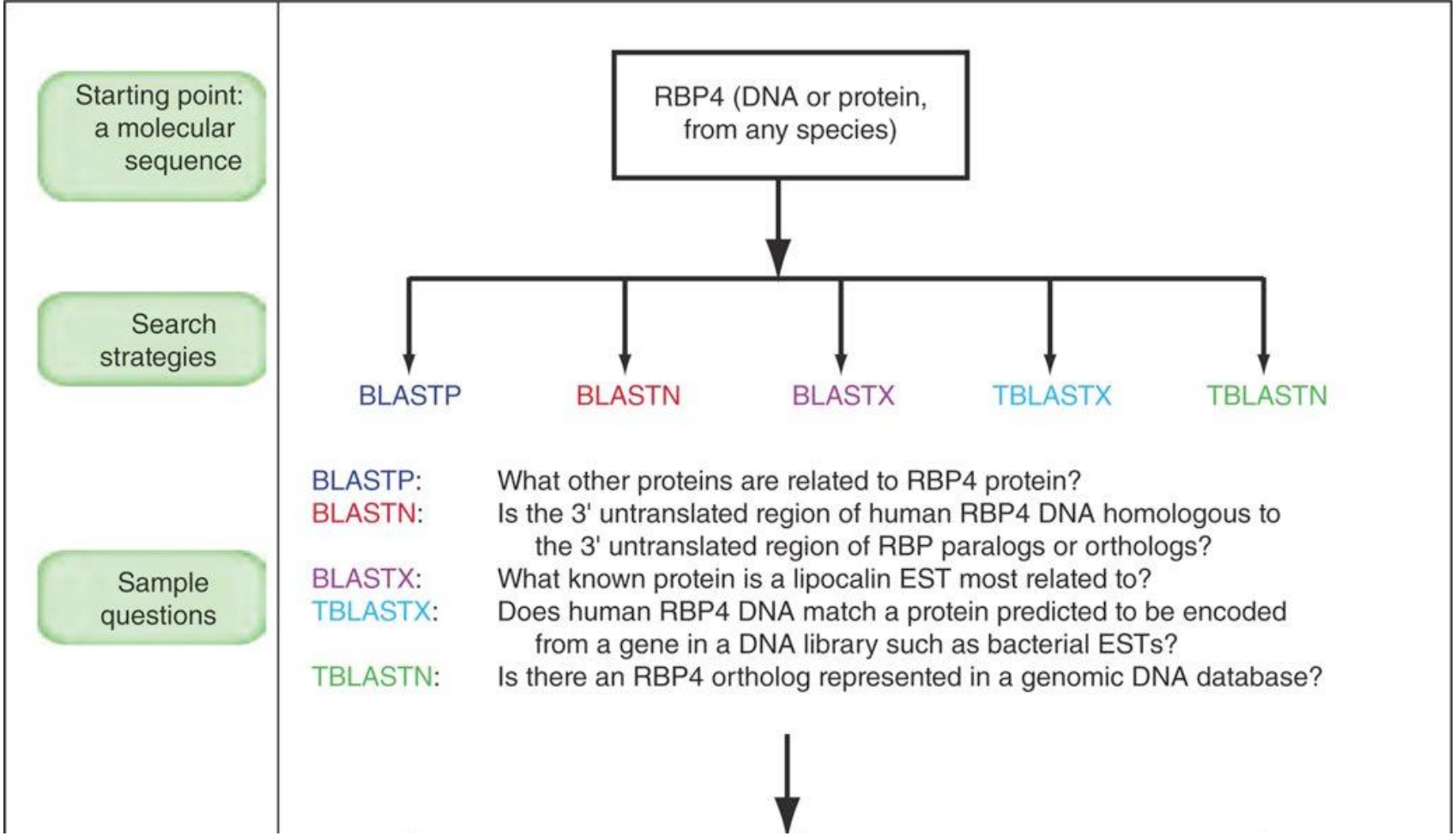
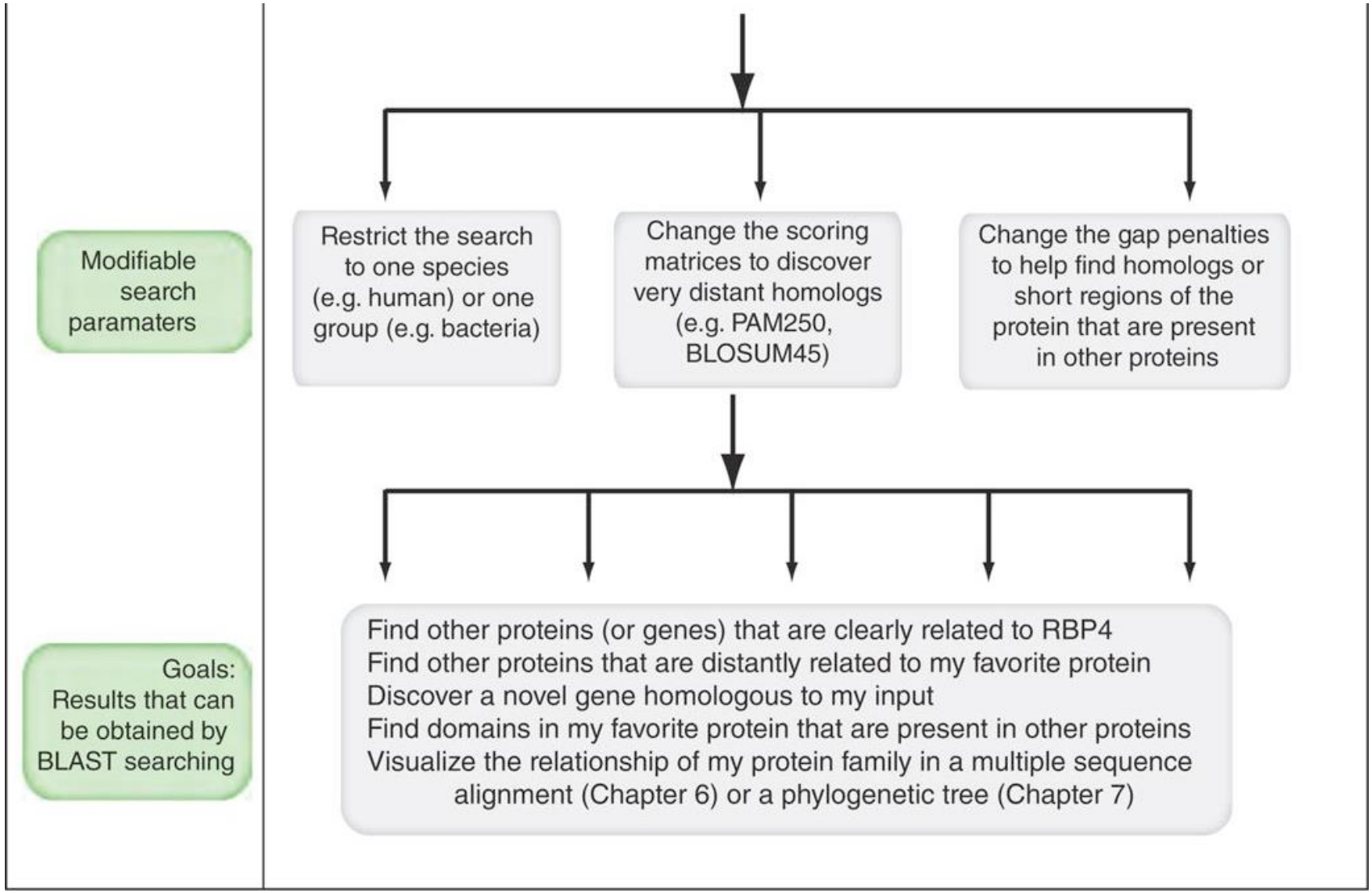


Fig 4.15 from Pevsner III edition

Fig 4.15 from Pevnsner III edition



Principles of DB searching using Calcin family as an example

- Lipocalcin family proteins in this example share very limited sequence similarity
  - RBP4, NP_006735.2
  - Odorant-binding protein (OBP)
- BLAST
  - DB: nr; organism: Homo sapiens; others: def
  - Restricting the output only to Human RefSeq proteins (how can we do this?)

Too many hits?

- Refseq
  - Nr database
  - Restrict to only specific organisms
  - Restrict to the domain of interest
    - How to find this?
    - UniProt
  - Adjust
    - Scoring matrix
    - Expect value

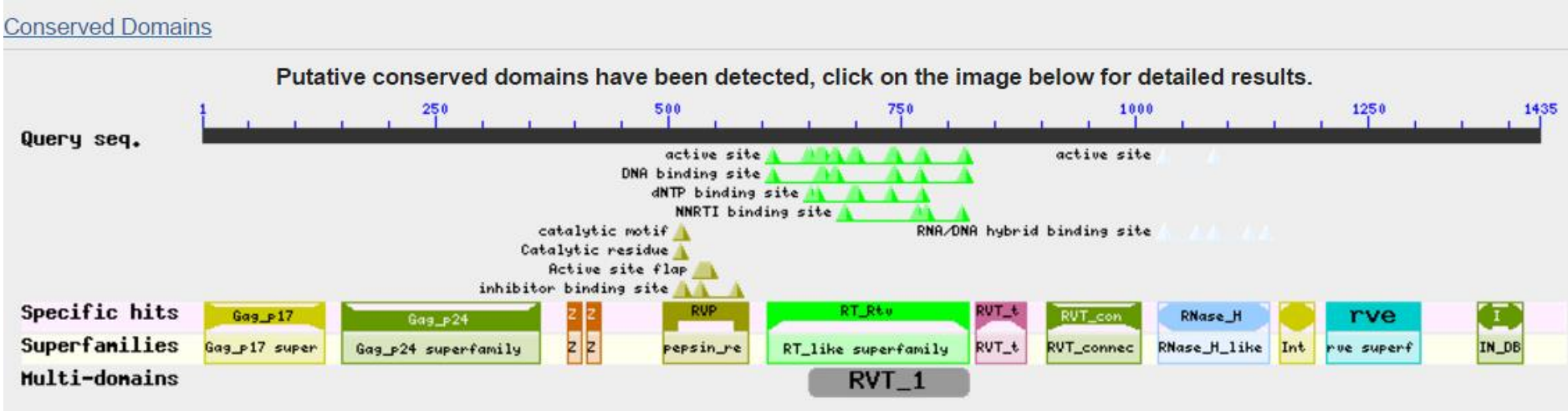
Too Small hits?

- How can this happen?
  - Exploring microbial/viral genomes
    - Only few are sequenced
  - Reset the BLAST page (to remove prev limits)
  - Matrices
    - High PAM or lower BLOSUM
    - Include all DBs (HTGS/GSS)
    - Search to include model sequences
    - Finally, use HMM based searches (PSI-BLAST etc)

HIV-1 Pol: Second Example

- Mutiple domain protein
- http://www.ncbi.nlm.nih.gov/gene/155348
- http://www.uniprot.org/uniprot/P04585
- 1435 aa
- What will happen when do a blastp search for this query?

NP_057849.4


Conserved Domains
Putative conserved domains have been detected, click on the image below for detailed results.
Query seq.
1
250
500
750
1000
1250
1435
active site
DNA binding site
dNTP binding site
NNRTI binding site
catalytic motif
Catalytic residue
Active site flap
inhibitor binding site
active site
RNA/DNA hybrid binding site
Specific hits
Superfamilies
Multi-domains
Gag_p17
Gag_p17 super
Gag_p24
Gag_p24 superfamily
RVP
pepsin_re
RT_RTv
RT_like superfamily
RVT_t
RVT_t
RVT_con
RVT_connec
RNase_H
RNase_H_like
Int
rve
rve superf
I
IN_DB
RVT_1

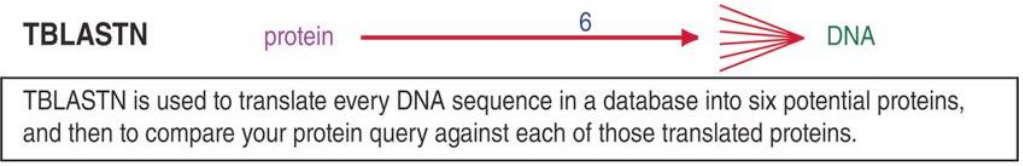
"New Gene" == discovery of some DNA sequence in a DB that has not been annotated yet

Find-a-Gene Tips

Start with the sequence of a known protein

TBLASTN

Search a DNA database (e.g. HTGS, dbEST, or genomic sequence from a specific organism)



Inspect the output

Find matches...
[1] to DNA encoding known proteins (not novel)
[2] to DNA encoding related proteins (novel!)
[3] to false positives

BLASTX or BLASTP nr

Search your DNA or protein against a protein database (nr) to confirm you have identified a novel gene

Things to Report

- Query sequence
- TBLASTN
  - What DB? What Matrix; what non-optional parameters?
  - Hits (follow the font and other details as Dr. Pevsner has suggested)

Things to Report

- Use additional BLASTX/BLASTP to confirm that the protein that you had identified is novel
  - (*follow the suggestions of Prof. Pevnser on what is novel; page 159 of the book*)
  - Again list DB, matrix; hits (top 10)
  - Name your protein, example “Anguilicola Globin”
    - Because of the organism and family it belongs to

Please refer to the word document (available on Blackboard) for details on the final project

Things to Report

- Carry out Multiple sequence alignment
  - Your novel protein + 5 or 10 (max 30) from the novel protein speculated family
- Create a phylogenetic tree
- Secondary/tertiary structure of your novel protein
- Provide whether the gene is under positive/negative selection (optional)
- Significance of the novel gene

Computer Lab

- Try the following exercises:
  - 4-1, 4-2, 4-3, 4-5 and 4-9

Thanks

ravichandran@hood.edu