November 29, 2010

BLAST:

Basic local alignment search tool

Jonathan Pevsner, Ph.D.
Bioinformatics
pevsner@kennedykrieger.org
Johns Hopkins School of Medicine

Copyright notice

Outline of today's lecture

BLAST
- Practical use
- Algorithm
- Strategies

Finding distantly related proteins:
- PSI-BLAST
- Hidden Markov models

BLAST-like tools for genomic DNA
- PatternHunter
- Megablast
- BLAT, BLASTZ

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 web-accessible.

page 101

Why use BLAST?

BLAST searching is fundamental to understanding the relatedness of any favorite query sequence to other known proteins or DNA sequences.

Applications include
- identifying orthologs and paralogs
- discovering new genes or proteins
- discovering variants of genes or proteins
- investigating expressed sequence tags (ESTs)
- exploring protein structure and function

page 102

Four components to a BLAST search

(1) Choose the sequence (query)

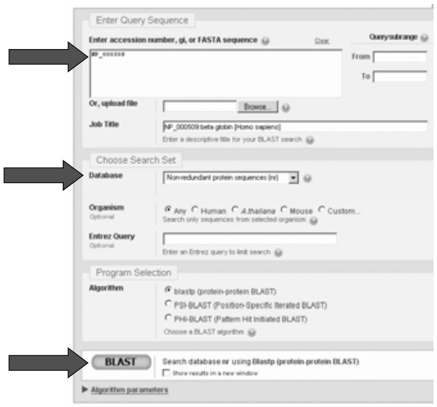
(2) Select the BLAST program

(3) Choose the database to search

(4) Choose optional parameters

Then click "BLAST"

page 102



Enter Query Sequence

Enter accession number, gi, or FASTA sequence

NP_000509

Clear

Query subrange

From

To

Or, upload file [Browse...]

Job Title: NP_000509 beta globin [Homo sapiens]

Enter a descriptive title for your BLAST search

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): ● Any ○ Human ○ A.thaliana ○ Mouse ○ Custom...

Search only sequences from selected organism

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)

Choose a BLAST algorithm

[BLAST] Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

▶ Algorithm parameters

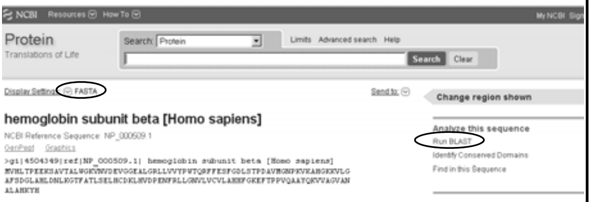
page 103

Step 1: Choose your sequence

Sequence can be input in FASTA format or as accession number

page 103

Example of the FASTA format for a BLAST query



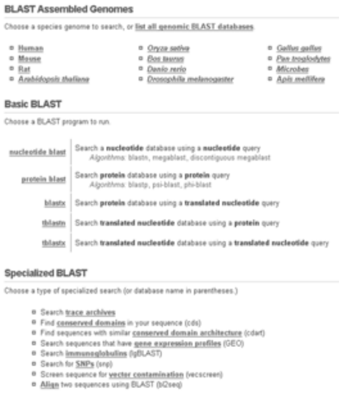
hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

Fig. 2.9
page 32

Step 2: Choose the BLAST program



BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search for SNPs (snp)
- Screen sequence for vector contamination (vecscreen)
- Align two sequences using BLAST (bl2seq)

page 104

Step 2: Choose the BLAST program

blastn (nucleotide BLAST)

blastp (protein BLAST)

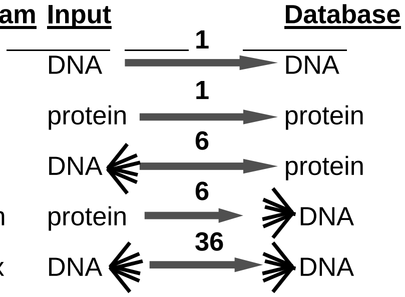
blastx (translated BLAST)

tblastn (translated BLAST)

tblastx (translated BLAST)

page 104

Choose the BLAST program



Program	Input		Database
blastn	DNA	1	DNA
blastp	protein	1	protein
blastx	DNA	6	protein
tblastn	protein	6	DNA
tblastx	DNA	36	DNA

page 104

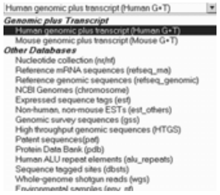
DNA potentially encodes six proteins

```
     5' CAT CAA
      5' ATC AAC
       5' TCA ACT
5' CATCAACTACAACTCCAAAGACACCCTTACACATCAACAAACCTACCCAC 3'
3' GTAGTTGATGTTGAGGTTTCTGTGGGAATGTGTAGTTGTTTGGATGGGTG 5'
                                       5' GTG GGT
                                        5' TGG GTA
                                         5' GGG TAG
```

page 105

Step 3: choose the database

nr = non-redundant (most general database)
dbest = database of expressed sequence tags
dbsts = database of sequence tag sites
gss = genomic survey sequences

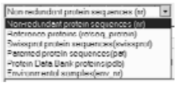


Human genomic plus transcript (Human G+T)

Genomic plus Transcript
- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

Other Databases
- Nucleotide collection (nr/nt)
- Reference mRNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Non-human, non-mouse ESTs (est_others)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun reads (wgs)
- Environmental samples (env_nt)

nucleotide databases

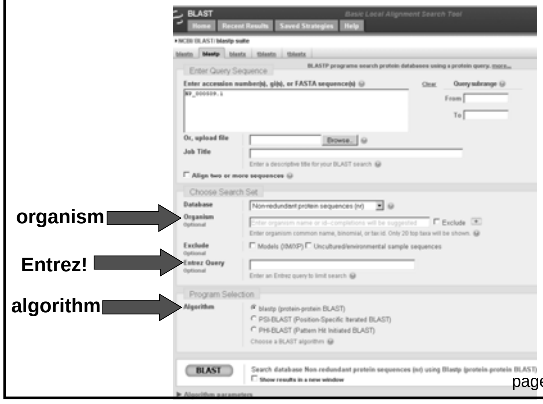


Non-redundant protein sequences (nr)
- Non-redundant protein sequences (nr)
- Reference proteins (refseq_protein)
- Swissprot protein sequences(swissprot)
- Patented protein sequences(pat)
- Protein Data Bank proteins(pdb)
- Environmental samples(env_nr)

protein databases

page 106

Step 4a: Select optional search parameters



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s): NP_000509.1
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: Non-redundant protein sequences (nr)
Organism (Optional): Enter organism name or id--completions will be suggested □ 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)
Choose a BLAST algorithm

BLAST — Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
□ Show results in a new window
▶ Algorithm parameters

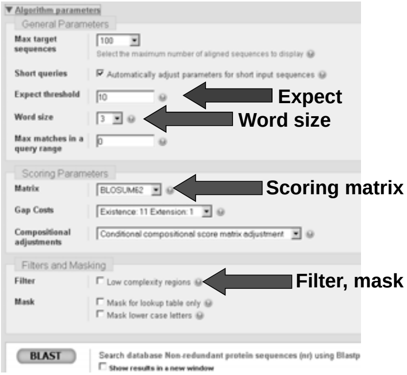
organism

Entrez!

algorithm

page 107

Step 4a: optional blastp search parameters



▼ Algorithm parameters

General Parameters
Max target sequences: 100 — Select the maximum number of aligned sequences to display
Short queries: ☑ Automatically adjust parameters for short input sequences
Expect threshold: 10
Word size: 3
Max matches in a query range: 0

Scoring Parameters
Matrix: BLOSUM62
Gap Costs: Existence: 11 Extension: 1
Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking
Filter: □ Low complexity regions
Mask: □ Mask for lookup table only
□ Mask lower case letters

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

Expect

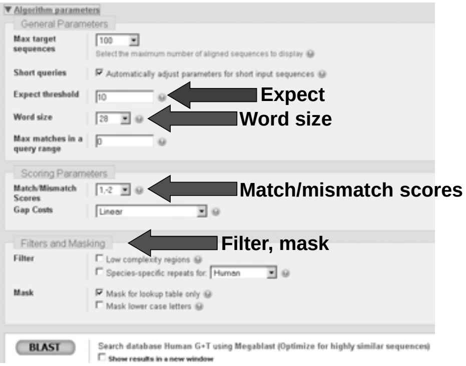
Word size

Scoring matrix

Filter, mask

page 108

Step 4a: optional blastn search parameters



▼ Algorithm parameters

General Parameters
Max target sequences: 100 — Select the maximum number of aligned sequences to display
Short queries: ☑ Automatically adjust parameters for short input sequences
Expect threshold: 10
Word size: 28
Max matches in a query range: 0

Scoring Parameters
Match/Mismatch Scores: 1,-2
Gap Costs: Linear

Filters and Masking
Filter: □ Low complexity regions
□ Species-specific repeats for: Human
Mask: ☑ Mask for lookup table only
□ Mask lower case letters

BLAST — Search database Human G+T using Megablast (Optimize for highly similar sequences)
□ Show results in a new window

Expect

Word size

Match/mismatch scores

Filter, mask

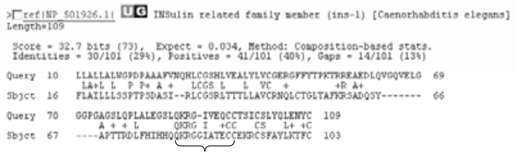
page 108

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

page 106

(a) Query: human insulin NP_000198
Program: blastp
Database: *C. elegans* RefSeq
Default settings:
Unfiltered (“composition-based statistics”)

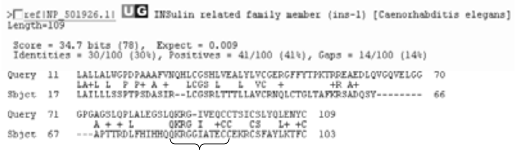


```
>☐ref|NP_501926.1| UG INSulin related family member (ins-1) [Caenorhabditis elegans]
Length=109

 Score = 32.7 bits (73),  Expect = 0.034, Method: Composition-based stats.
 Identities = 30/101 (29%), Positives = 41/101 (40%), Gaps = 14/101 (13%)

Query  10   LLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELG  69
             LA+L L   P P+ A +   LCGS L    L  VC  +          +R A+
Sbjct  16   FLAILLLSSPTPSDASI--RLCGSRLTTTLLAVCRNQLCTGLTAFKRSADQSY-------  66

Query  70   GGPGAGSLQPLALEGSLQKRG-IVEQCCTSICSLYQLENYC  109
                A + + L       QKRG I  +CC   CS   L+ +C
Sbjct  67   ----APTTRDLFHIHHQQKRGGIATECCEKRCSFAYLKTFC  103
```

Our starting point: search human insulin against worm RefSeq proteins by blastp using default parameters

page 109

(b) Query: human insulin NP_000198
Program: blastp
Database: *C. elegans* RefSeq
Option: No compositional adjustment

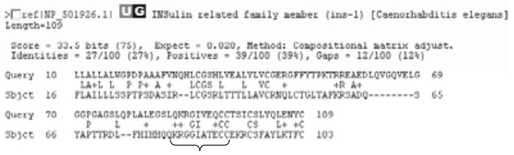


```
>☐ref|NP_501926.1| UG INSulin related family member (ins-1) [Caenorhabditis elegans]
Length=109

 Score = 34.7 bits (78),  Expect = 0.009
 Identities = 30/100 (30%), Positives = 41/100 (41%), Gaps = 14/100 (14%)

Query  11   LALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGG  70
            LA+L L   P P+ A +   LCGS L    L  VC  +          +R A+
Sbjct  17   LAILLLSSPTPSDASIR--LCGSRLTTTLLAVCRNQLCTGLTAFKRSADQSY--------  66

Query  71   GPGAGSLQPLALEGSLQKRG-IVEQCCTSICSLYQLENYC  109
               A + + L       QKRG I  +CC   CS   L+ +C
Sbjct  67   ---APTTRDLFHIHHQQKRGGIATECCEKRCSFAYLKTFC  103
```

Note that the bit score, Expect value, and percent identity all change with the “no compositional adjustment” option

page 109

(c) Query: human insulin NP_000198
Program: blastp
Database: *C. elegans* RefSeq
Option: conditional compositional score matrix adjustment

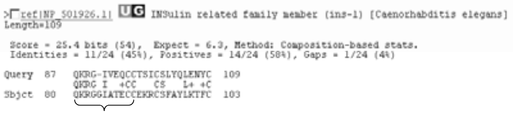


```
>☐ref|NP_501926.1| UG INSulin related family member (ins-1) [Caenorhabditis elegans]
Length=109

 Score = 33.5 bits (75),  Expect = 0.020, Method: Compositional matrix adjust.
 Identities = 27/100 (27%), Positives = 39/100 (39%), Gaps = 12/100 (12%)

Query  10   LLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELG  69
             LA+L L   P P+ A +   LCGS L    L  VC  +          +R A+
Sbjct  16   FLAILLLSSPTPSDASIR--LCGSRLTTTLLAVCRNQLCTGLTAFKRSADQ--------S  65

Query  70   GGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYC  109
              P    L    +    ++ GI  +CC   CS   L+ +C
Sbjct  66   YAPTTRDL--FHIHHQQKRGGIATECCEKRCSFAYLKTFC  103
```

Note that the bit score, Expect value, and percent identity all change with the compositional score matrix adjustment

page 109

(d) Query: human insulin NP_000198
Program: blastp
Database: *C. elegans* RefSeq
Option: Filter low complexity regions

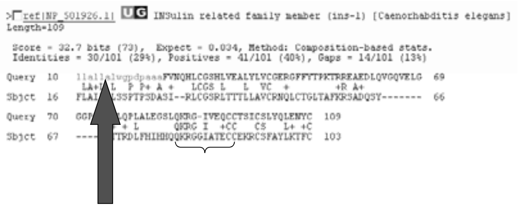


```
>☐ref|NP_501926.1| UG INSulin related family member (ins-1) [Caenorhabditis elegans]
Length=109

 Score = 25.4 bits (54),  Expect = 6.3, Method: Composition-based stats.
 Identities = 11/24 (45%), Positives = 14/24 (58%), Gaps = 1/24 (4%)

Query  87   QKRG-IVEQCCTSICSLYQLENYC  109
            QKRG I  +CC   CS   L+ +C
Sbjct  80   QKRGGIATECCEKRCSFAYLKTFC  103
```

Note that the bit score, Expect value, and percent identity all change with the filter option

page 109

(e) Query: human insulin NP_000198
Program: blastp
Database: *C. elegans* RefSeq
Option: Mask for lookup table only

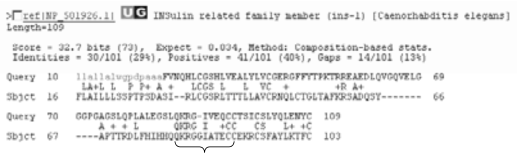


```
>☐ref|NP_501926.1| UG INSulin related family member (ins-1) [Caenorhabditis elegans]
Length=109

 Score = 32.7 bits (73),  Expect = 0.034, Method: Composition-based stats.
 Identities = 30/101 (29%), Positives = 41/101 (40%), Gaps = 14/101 (13%)

Query  10   llallalwgpdpaaaFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELG  69
             LA+  L  P P+ A +   LCGS L    L  VC  +          +R A+
Sbjct  16   FLAI  LSSPTPSDASI--RLCGSRLTTTLLAVCRNQLCTGLTAFKRSADQSY-------  66

Query  70   GGP    LQPLALEGSLQKRG-IVEQCCTSICSLYQLENYC  109
                  + L       QKRG I  +CC   CS   L+ +C
Sbjct  67   ----   TRDLFHIHHQQKRGGIATECCEKRCSFAYLKTFC  103
```

Filtering
(the filtered sequence is the query in lowercase and grayed out)

page 109

(e) Query: human insulin NP_000198
Program: blastp
Database: *C. elegans* RefSeq
Option: Mask for lookup table only

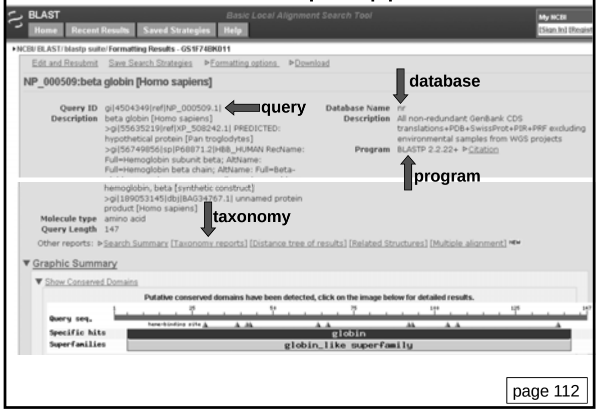


```
>☐ref|NP_501926.1| UG INSulin related family member (ins-1) [Caenorhabditis elegans]
Length=109

 Score = 32.7 bits (73),  Expect = 0.034, Method: Composition-based stats.
 Identities = 30/101 (29%), Positives = 41/101 (40%), Gaps = 14/101 (13%)

Query  10   llallalwgpdpaaaFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELG  69
             LA+L L   P P+ A +   LCGS L    L  VC  +          +R A+
Sbjct  16   FLAILLLSSPTPSDASI--RLCGSRLTTTLLAVCRNQLCTGLTAFKRSADQSY-------  66

Query  70   GGPGAGSLQPLALEGSLQKRG-IVEQCCTSICSLYQLENYC  109
                A + + L       QKRG I  +CC   CS   L+ +C
Sbjct  67   ----APTTRDLFHIHHQQKRGGIATECCEKRCSFAYLKTFC  103
```

Note that the bit score, Expect value, and percent identity could change with the “mask for lookup table only” option

page 109

BLAST search output: top portion



▸NCBI/ BLAST/ blastp suite/ Formatting Results - GS1F74BK011

Edit and Resubmit Save Search Strategies ▸Formatting options ▸Download

NP_000509:beta globin [Homo sapiens]

Query ID gi|4504349|ref|NP_000509.1|
Description beta globin [Homo sapiens]
>gi|55635219|ref|XP_508242.1| PREDICTED: hypothetical protein [Pan troglodytes]
>gi|56749856|sp|P68871.2|HBB_HUMAN RecName: Full=Hemoglobin subunit beta; AltName: Full=Hemoglobin beta chain; AltName: Full=Beta-
hemoglobin, beta [synthetic construct]
>gi|189053145|dbj|BAG34767.1| unnamed protein product [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 2.2.22+ ▸Citation

Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment] NEW

▼ Graphic Summary

▼ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

database

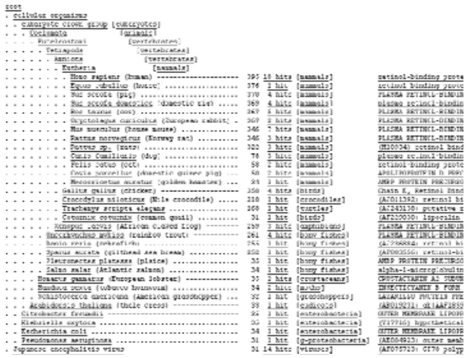
query

program

taxonomy

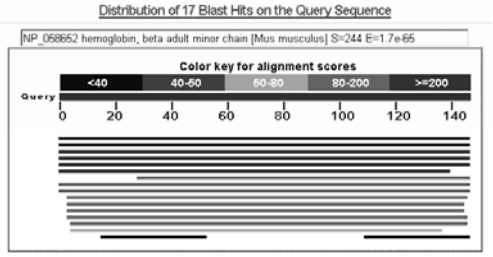
page 112

BLAST search output: taxonomy report summarizes species with matches



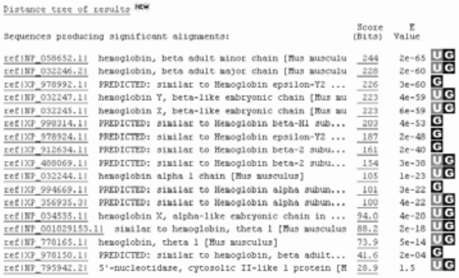
```
root
. cellular organisms
. . Eukaryota crown group [eukaryotes]
. . . Coelomata             [animals]
. . . . Euteleostomi          [vertebrates]
. . . . . Tetrapoda             [vertebrates]
. . . . . . Amniota               [vertebrates]
. . . . . . . Eutheria              [mammals]
. . . . . . . . Homo sapiens (human) -------------------  395 18 hits [mammals]     retinol-binding prote
. . . . . . . . Equus caballus (horse) .................  376  1 hit  [mammals]     retinol binding prote
. . . . . . . . Sus scrofa (pig) .......................  370  4 hits [mammals]     PLASMA RETINOL-BINDIN
. . . . . . . . Sus scrofa domestica (domestic pig) ....  369  4 hits [mammals]     plasma retinol-bindin
. . . . . . . . Bos taurus (cow) .......................  367  6 hits [mammals]     PLASMA RETINOL-BINDIN
. . . . . . . . Oryctolagus cuniculus (European rabbit) .  367  2 hits [mammals]     PLASMA RETINOL-BINDIN
. . . . . . . . Mus musculus (house mouse) .............  346  7 hits [mammals]     PLASMA RETINOL-BINDIN
. . . . . . . . Rattus norvegicus (Norway rat) .........  346  3 hits [mammals]     PLASMA RETINOL-BINDIN
. . . . . . . . Rattus sp. (rats) ......................  322  3 hits [mammals]     (M30034) retinol bind
. . . . . . . . Canis familiaris (dog) .................   78  3 hits [mammals]     plasma retinol-bindin
. . . . . . . . Felis catus (cat) ......................   58  2 hits [mammals]     retinol-binding prote
. . . . . . . . Cavia porcellus (domestic guinea pig) ...   58  2 hits [mammals]     APOLIPOPROTEIN D PREC
. . . . . . . . Mesocricetus auratus (golden hamster) ...   34  1 hit  [mammals]     AMBP PROTEIN PRECURSO
. . . . . . . Gallus gallus (chicken) ------------------  350  4 hits [birds]       Chain E, Retinol Bind
. . . . . . . Crocodylus niloticus (Nile crocodile) .....  310  1 hit  [crocodiles]  (AF011392: retinol bi
. . . . . . . Trachemys scripta elegans ................  308  1 hit  [turtles]     (AF243130: putative r
. . . . . . . Coturnix coturnix (common quail) .........   31  1 hit  [birds]       (AF230030: lipocalin
. . . . . . Xenopus laevis (African clawed frog) -------  299  3 hits [amphibians]  PLASMA RETINOL-BINDIN
. . . . . Oncorhynchus mykiss (rainbow trout) ----------  261  4 hits [bony fishes] PLASMA RETINOL-BINDIN
. . . . . Danio rerio (zebrafish) ......................  255  1 hit  [bony fishes] (AF296884: retinol bi
. . . . . Sparus aurata (gilthead sea bream) ...........  252  1 hit  [bony fishes] (AF003556: retinol-bi
. . . . . Pleuronectes platessa (plaice) ...............   35  1 hit  [bony fishes] AMBP PROTEIN PRECURSO
. . . . . Salmo salar (Atlantic salmon) ................   34  1 hit  [bony fishes] alpha-1-microglobulin
. . . . Homarus gammarus (European lobster) ------------   35  2 hits [crustaceans] CRUSTACYANIN A2 SUBUN
. . . . Manduca sexta (tobacco hornworm) ...............   34  2 hits [moths]       INSECTICYANIN B FORM
. . . . Schistocerca americana (American grasshopper) ...   32  1 hit  [grasshoppers] LAZARILLO PROTEIN PRE
. . . Arabidopsis thaliana (thale cress) ---------------   33  1 hit  [eudicots]    (AB019231: dt|AAF1859
. . Citrobacter freundii -------------------------------   32  1 hit  [enterobacteria] OUTER MEMBRANE LIPOPR
. . Klebsiella oxytoca .................................   32  1 hit  [enterobacteria] (Y17716) hypothetical
. . Escherichia coli ...................................   34  1 hit  [enterobacteria] OUTER MEMBRANE LIPOPR
. . Pseudomonas aeruginosa .............................   31  1 hit  [g-proteobacteria] (AE004923: outer memb
. Japanese encephalitis virus                              31 14 hits [viruses]     (AF075723: GI70 polyp
```

BLAST search output: graphical output



page 112

BLAST search output: tabular output



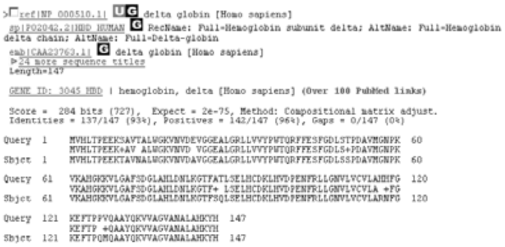
```
Distance tree of results NEW

                                                                     Score     E
Sequences producing significant alignments:                          (Bits)  Value

ref|NP_058652.1|  hemoglobin, beta adult minor chain [Mus musculu     244    2e-65  UG
ref|NP_032246.2|  hemoglobin, beta adult major chain [Mus musculu     228    2e-60  UG
ref|XP_978992.1|  PREDICTED: similar to Hemoglobin epsilon-Y2 ...     226    3e-60  G
ref|NP_032247.1|  hemoglobin Y, beta-like embryonic chain [Mus mu     223    4e-59  UG
ref|NP_032245.1|  hemoglobin Z, beta-like embryonic chain [Mus mu     223    6e-59  UG
ref|XP_998314.1|  PREDICTED: similar to Hemoglobin beta-H1 sub...     203    4e-53  G
ref|XP_978924.1|  PREDICTED: similar to Hemoglobin epsilon-Y2 ...     187    2e-48  G
ref|XP_912634.1|  PREDICTED: similar to Hemoglobin beta-2 subu...     161    2e-40  G
ref|XP_488069.1|  PREDICTED: similar to Hemoglobin beta-2 subu...     154    3e-38  UG
ref|NP_032244.1|  hemoglobin alpha 1 chain [Mus musculus]             105    1e-23  UG
ref|XP_994669.1|  PREDICTED: similar to Hemoglobin alpha subun...     101    3e-22  G
ref|XP_356935.3|  PREDICTED: similar to Hemoglobin alpha subun...     100    4e-22  UG
ref|NP_034535.1|  hemoglobin X, alpha-like embryonic chain in ...    94.0    4e-20  UG
ref|NP_001029153.1|  similar to hemoglobin, theta 1 [Mus musculus    88.2    2e-18  UG
ref|NP_778165.1|  hemoglobin, theta 1 [Mus musculus]                 73.9    5e-14  UG
ref|XP_978150.1|  PREDICTED: similar to hemoglobin, beta adult...    41.6    2e-04  G
ref|NP_795942.2|  5'-nucleotidase, cytosolic II-like 1 protein [M    28.9    1.5    UG
```

High scores low E values

Cut-off: .05? 10^{-10}?

page 113

BLAST search output: alignment output



```
>□ref|NP_000510.1| UG delta globin [Homo sapiens]
 sp|P02042.2|HBD_HUMAN G RecName: Full=Hemoglobin subunit delta; AltName: Full=Hemoglobin
delta chain; AltName: Full=Delta-globin
 emb|CAA23763.1| G delta globin [Homo sapiens]
 ▶24 more sequence titles
 Length=147

 GENE ID: 3045 HBD | hemoglobin, delta [Homo sapiens] (Over 100 PubMed links)

 Score =  284 bits (727),  Expect = 2e-75, Method: Compositional matrix adjust.
 Identities = 137/147 (93%), Positives = 142/147 (96%), Gaps = 0/147 (0%)

Query  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
            MVHLTPEEK+AV ALWGKVNVD VGGEALGRLLVVYPWTQRFFESFGDLS+PDAVMGNPK
Sbjct  1    MVHLTPEEKTAVNALWGKVNVDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPK  60

Query  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
            VKAHGKKVLGAFSDGLAHLDNLKGTF+ LSELHCDKLHVDPENFRLLGNVLVCVLA +FG
Sbjct  61   VKAHGKKVLGAFSDGLAHLDNLKGTFSQLSELHCDKLHVDPENFRLLGNVLVCVLARNFG  120

Query  121  KEFTPPVQAAYQKVVAGVANALAHKYH  147
            KEFTP +QAAYQKVVAGVANALAHKYH
Sbjct  121  KEFTPQMQAAYQKVVAGVANALAHKYH  147
```

Outline of today's lecture

BLAST
- Practical use
- Algorithm
- Strategies

Finding distantly related proteins:
- PSI-BLAST
- Hidden Markov models

BLAST-like tools for genomic DNA
- PatternHunter
- Megablast
- BLAT, BLASTZ

BLAST: background on sequence alignment

There are two main approaches to sequence alignment:

[1] Global alignment (Needleman & Wunsch 1970) using dynamic programming to find optimal alignments between two sequences. (Although the alignments are optimal, the search is not exhaustive.) Gaps are permitted in the alignments, and the total lengths of both sequences are aligned (hence “global”).

page 115

BLAST: background on sequence alignment

[2] The second approach is local sequence alignment (Smith & Waterman, 1980). The alignment may contain just a portion of either sequence, and is appropriate for finding matched domains between sequences.

BLAST is a heuristic approximation to local alignment. It examines only part of the search space.

page 115; 84

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)

(page 115)

How the original BLAST algorithm works: three phases

Phase 1: compile a list of word pairs (w=3) above threshold T

Example: for a human RBP query …FSGTWYA… (query word is in yellow)

A list of words (w=3) is:

```
FSG SGT GTW TWY WYA
YSG TGT ATW SWY WFA
FTG SVT GSW TWF WYS
```

Fig. 4.11
page 116

Phase 1: compile a list of words (w=3)

```
                     GTW 6,5,11  22
neighborhood         GSW 6,1,11 18
word hits            ATW 0,5,11  16
> threshold          NTW 0,5,11  16
                     GTY 6,5,2   13
 (T=11)
                     GNW         10
neighborhood         GAW         9
word hits
< below threshold
```

Fig. 4.11
page 116

Pairwise alignment scores are determined using a scoring matrix such as Blosum62

A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-1	1	1	-2	-1	-3	-2	5								
M	-1	-2	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Page 73

How a BLAST search works: 3 phases

Phase 2:

Scan the database for entries that match the compiled list.

This is fast and relatively easy.

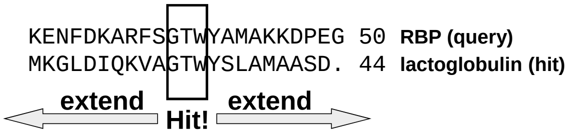
Fig. 4.11
page 116

How a BLAST search works: 3 phases

Phase 3: when you manage to find a hit
(i.e. a match between a “word” and a database entry), extend the hit in either direction.

Keep track of the score (use a scoring matrix)

Stop when the score drops below some cutoff.



```
KENFDKARFSGTWYAMAKKDPEG 50  RBP (query)
MKGLDIQKVAGTWYSLAMAASD. 44  lactoglobulin (hit)
```

page 116

How a BLAST search works: 3 phases

Phase 3:

In the original (1990) implementation of BLAST, hits were extended in either direction.

In a 1997 refinement of BLAST, two independent hits are required. The hits must occur in close proximity to each other. With this modification, only one seventh as many extensions occur, greatly speeding the time required for a search.

page 116

How a BLAST search works: threshold

You can 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+.

(To find BLAST+ go to BLAST → help → download.)

page 117

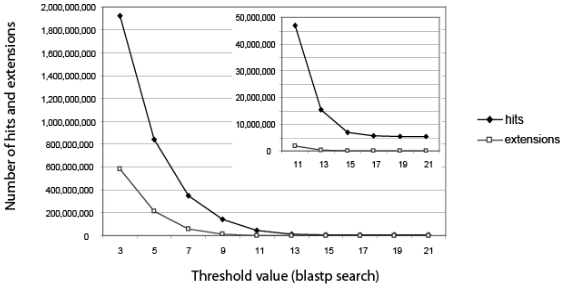


Fig. 4.12
page 118

Phase 1: compile a list of words (w=3)

```
                   GTW 6,5,11  22
neighborhood       GSW 6,1,11 18
word hits          ATW 0,5,11  16
> threshold        NTW 0,5,11  16
                   GTY 6,5,2   13
   (T=11)
                   GNW         10
neighborhood       GAW         9
word hits
< below threshold
```

Fig. 4.11
page 116

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

For megablast (see below), the word size is 28 and can be adjusted to 64. What will this do? Megablast is VERY fast for finding closely related DNA sequences!

How to interpret a BLAST search: expect value

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

For global alignments, the statistics are poorly understood.

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

page 118

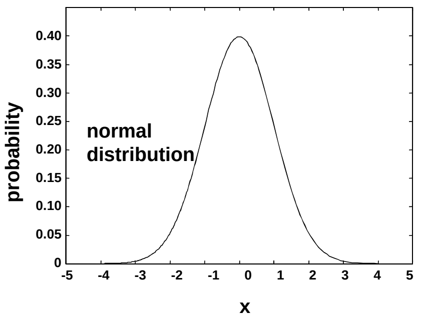


Fig. 4.13
page 119

The probability density function of the extreme value distribution (characteristic value u=0 and decay constant λ=1)

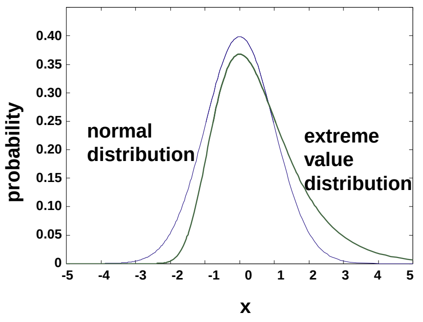


Fig. 4.13
page 119

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:

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

page 120

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

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

page 120

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

page 120

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' = 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.

page 121

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.

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

page 121

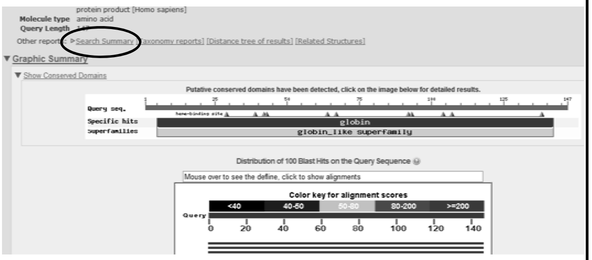
How to interpret BLAST: *E* values and *p* values

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

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

Table 4.3
page 122

How to interpret BLAST: overview



protein product [Homo sapiens]
Molecule type amino acid
Query Length 147
Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures]

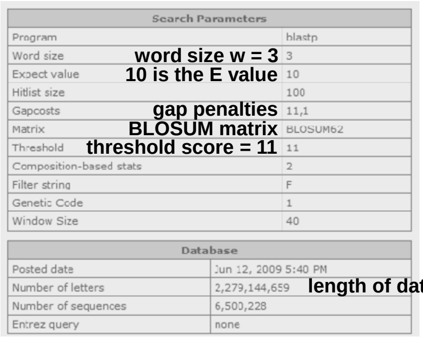
▼ Graphic Summary

▼ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 100 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments

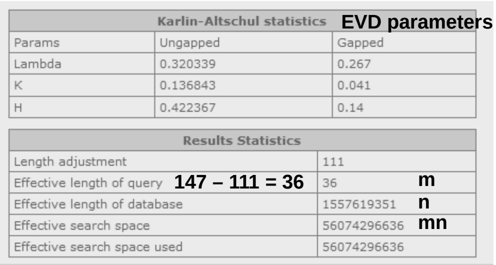


Search Parameters	
Program	blastp
Word size	3
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Threshold	11
Composition-based stats	2
Filter string	F
Genetic Code	1
Window Size	40

Database	
Posted date	Jun 12, 2009 5:40 PM
Number of letters	2,279,144,659
Number of sequences	6,500,228
Entrez query	none

word size w = 3
10 is the E value
gap penalties
BLOSUM matrix
threshold score = 11
length of database

Fig. 4.14
page 122



Karlin-Altschul statistics		
Params	Ungapped	Gapped
Lambda	0.320339	0.267
K	0.136843	0.041
H	0.422367	0.14

Results Statistics	
Length adjustment	111
Effective length of query	36
Effective length of database	1557619351
Effective search space	56074296636
Effective search space used	56074296636

EVD parameters

147 – 111 = 36

m

n

mn

Effective search space
= mn
= length of query x db length

Fig. 4.14
page 122

Why set the E value to 20,000?

Suppose you perform a search with a short query (e.g. 9 amino acids). There are not enough residues to accumulate a big score (or a small E value).

Indeed, a match of 9 out of 9 residues could yield a small score with an E value of 100 or 200. And yet, this result could be "real" and of interest to you.

By setting the E value cutoff to 20,000 you do not change the way the search was done, but you do change which results are reported to you.

Outline of today's lecture

BLAST
- Practical use
- Algorithm
- Strategies

Finding distantly related proteins:
- PSI-BLAST
- Hidden Markov models

BLAST-like tools for genomic DNA
- PatternHunter
- Megablast
- BLAT, BLASTZ

BLAST search strategies

General concepts

How to evaluate the significance of your results

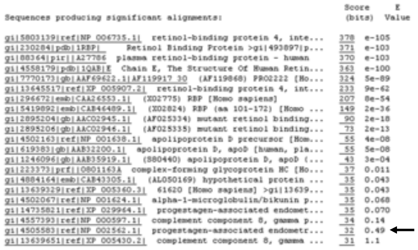
How to handle too many results

How to handle too few results

BLAST searching with HIV-1 pol, a multidomain protein

page 123

Sometimes a real match has an E value > 1



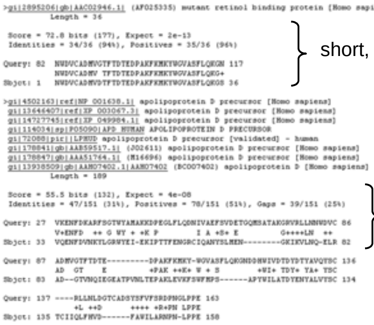
```
                                                               Score    E
Sequences producing significant alignments:                    (bits)  Value

gi|5803139|ref|NP_006735.1|  retinol-binding protein 4, inte...   378  e-105
gi|230284|pdb|1RBP|     Retinol Binding Protein >gi|493897|p...   371  e-103
gi|88364|pir||A27786  plasma retinol-binding protein - human      370  e-103
gi|4558179|pdb|1QAB|E  Chain E, The Structure Of Human Retin...   363  e-100
gi|7770173|gb|AAF69622.1|AF119917_30  (AF119868) PRO2222 [Ho...   324  5e-89
gi|13645517|ref|XP_005907.2|  retinol-binding protein 4, int...   233  9e-62
gi|296672|emb|CAA26553.1|  (X02775) RBP [Homo sapiens]            207  8e-54
gi|5419892|emb|CAB46489.1|  (X02824) RBP (aa 101-172) [Homo ...   149  2e-36
gi|2895204|gb|AAC02945.1|  (AF025334) mutant retinol binding...    90  2e-18
gi|2895206|gb|AAC02946.1|  (AF025335) mutant retinol binding...    73  2e-13
gi|4502163|ref|NP_001638.1|  apolipoprotein D precursor [Hom...    55  4e-08
gi|619383|gb|AAB32200.1|  apolipoprotein D, apoD [human, pla...    55  5e-08
gi|1246096|gb|AAB35919.1|  (S80440) apolipoprotein D, apoD (...    43  3e-04
gi|223373|prf||0801163A  complex-forming glycoprotein HC [Ho...    37  0.011
gi|4884164|emb|CAB43305.1|  (AL050169) hypothetical protein ...    35  0.043
gi|13639329|ref|XP_005360.3|  61620 [Homo sapiens] >gi|13639...    35  0.043
gi|4502067|ref|NP_001624.1|  alpha-1-microglobulin/bikunin p...    35  0.068
gi|14735821|ref|XP_029964.1|  progestagen-associated endomet...    35  0.070
gi|4557393|ref|NP_000597.1|  complement component 8, gamma p...    34  0.14
gi|4505583|ref|NP_002562.1|  progestagen-associated endometr...    32  0.49
gi|13639651|ref|XP_005430.2|  complement component 8, gamma ...    31  1.1
```

real match?

…try a reciprocal BLAST to confirm

Fig. 4.16
page 125

Sometimes a similar *E* value occurs for a short exact match and long less exact match



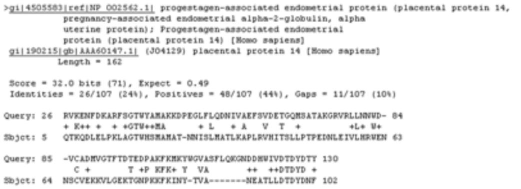
```
>gi|2895206|gb|AAC02946.1| (AF025335) mutant retinol binding protein [Homo sapiens]
          Length = 36

 Score = 72.8 bits (177), Expect = 2e-13
 Identities = 34/36 (94%), Positives = 35/36 (96%)

Query: 82  NWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGN 117
            NWDVCADMV TFTDTEDPAKFKMKYWGVASFLQKG+
Sbjct: 1   NWDVCADMVDTFTDTEDPAKFKMKYWGVASFLQKGS 36

>gi|4502163|ref|NP_001638.1| apolipoprotein D precursor [Homo sapiens]
 gi|13646407|ref|XP_003067.3| apolipoprotein D precursor [Homo sapiens]
 gi|14727745|ref|XP_049984.1| apolipoprotein D precursor [Homo sapiens]
 gi|114034|sp|P05090|APD_HUMAN APOLIPOPROTEIN D PRECURSOR
 gi|72088|pir||LPHUD apolipoprotein D precursor [validated] - human
 gi|178841|gb|AAB59517.1| (J02611) apolipoprotein D precursor [Homo sapiens]
 gi|178847|gb|AAA51764.1| (M16696) apolipoprotein D precursor [Homo sapiens]
 gi|13938509|gb|AAH07402.1|AAH07402 (BC007402) apolipoprotein D [Homo sapiens]
          Length = 189

 Score = 55.5 bits (132), Expect = 4e-08
 Identities = 47/151 (31%), Positives = 78/151 (51%), Gaps = 39/151 (25%)

Query: 27  VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
            V+ENFD  ++ G WY + +K P         I A +S+ E          G++++LN  ++
Sbjct: 33  VQENFDVNKYLGRWYEI-EKIPTTFENGRCIQANYSLMEN--------GKIKVLNQ-ELR 82

Query: 87  ADMVGTFTDTE---------DPAKFKMKY-WGVASFLQKGNDDHWIVDTDYDTYAVQYSC 136
            AD  GT    E         +PAK ++K+ W + S        +WI+ TDY+ YA+ YSC
Sbjct: 83  AD--GTVNQIEGEATPVNLTEPAKLEVKFSWFMPS------APYWILATDYENYALVYSC 134

Query: 137 ----RLLNLDGTCADSYSFVFSRDPNGLPPE 163
                +L ++D      ++++ +R+PN LPPE
Sbjct: 135 TCIIQLFHVD------FAWILARNPN-LPPE 158
```

short, nearly exact

long, only
31% identity,
similar E value

Fig. 4.17
page 125

Assessing whether proteins are homologous



```
>gi|4505583|ref|NP_002562.1| progestagen-associated endometrial protein (placental protein 14,
          pregnancy-associated endometrial alpha-2-globulin, alpha
          uterine protein); Progestagen-associated endometrial
          protein (placental protein 14) [Homo sapiens]
 gi|190215|gb|AAA60147.1| (J04129) placental protein 14 [Homo sapiens]
         Length = 162

 Score = 32.0 bits (71), Expect = 0.49
 Identities = 26/107 (24%), Positives = 48/107 (44%), Gaps = 11/107 (10%)

Query: 26   RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWD- 84
          + K++ +   + +GTW++MA       + L    + A    V   T   +            +L+ W+
Sbjct: 5    QTKQDLELPKLAGTWHSMAMAT-NNISLMATLKAPLRVHITSLLPTPEDNLEIVLHRWEN 63

Query: 85   -VCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTY 130
       C +        T +P KFK+ Y  VA         ++  ++DTDYD +
Sbjct: 64   NSCVEKKVLGEKTGNPKKFKINY-TVA-------NEATLLDTDYDNF 102
```

RBP4 and PAEP:
Low bit score, E value 0.49, 24% identity (“twilight zone”). But they are indeed homologous. Try a BLAST search with PAEP as a query, and find many other lipocalins.

~Fig. 4.18
page 126

The universe of lipocalins (each dot is a protein)

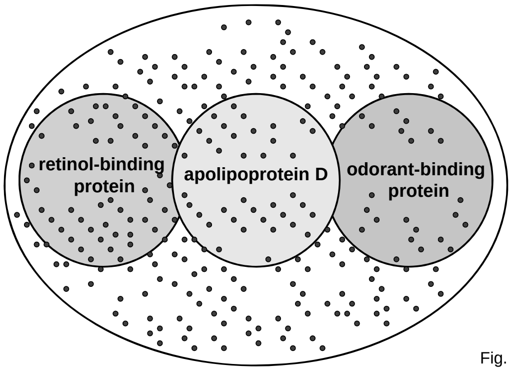
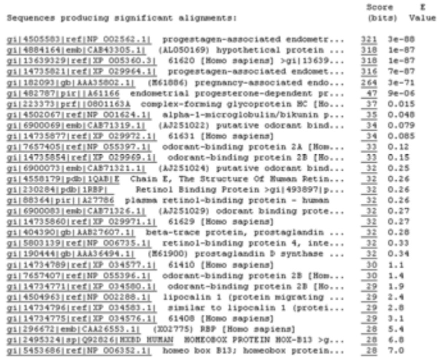


Fig. 5.7
Page 151

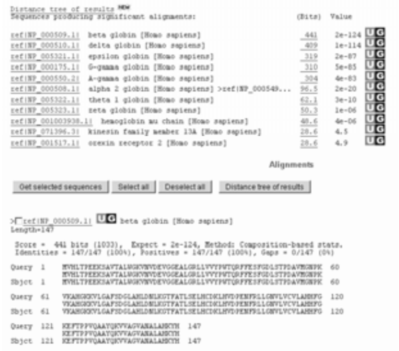
BLAST search with PAEP as a query finds many other lipocalins



```
                                                                 Score    E
Sequences producing significant alignments:                      (bits)  Value

gi|4505583|ref|NP_002562.1|  progestagen-associated endometr...   321   3e-88
gi|4884164|emb|CAB43305.1|  (AL050169) hypothetical protein ...   318   1e-87
gi|13639329|ref|XP_005360.3|  61620 [Homo sapiens] >gi|13639...   318   1e-87
gi|14735821|ref|XP_029964.1|  progestagen-associated endomet...   316   7e-87
gi|182093|gb|AAA35802.1|  (M61886) pregnancy-associated endo...   264   3e-71
gi|482787|pir||A61166  endometrial progesterone-dependent pr...    47   9e-06
gi|223373|prf||0801163A  complex-forming glycoprotein HC [Ho...    37   0.015
gi|4502067|ref|NP_001624.1|  alpha-1-microglobulin/bikunin p...    35   0.048
gi|6900069|emb|CAB71319.1|  (AJ251022) putative odorant bind...    34   0.079
gi|14735877|ref|XP_029972.1|  61631 [Homo sapiens]                 34   0.085
gi|7657405|ref|NP_055397.1|  odorant-binding protein 2A [Hom...    33   0.12
gi|14735854|ref|XP_029969.1|  odorant-binding protein 2B [Ho...    33   0.15
gi|6900073|emb|CAB71321.1|  (AJ251024) putative odorant bind...    32   0.25
gi|4558179|pdb|1QAB|E  Chain E, The Structure Of Human Retin...    32   0.26
gi|230284|pdb|1RBP|     Retinol Binding Protein >gi|493897|p...    32   0.26
gi|88364|pir||A27786  plasma retinol-binding protein - human       32   0.26
gi|6900083|emb|CAB71326.1|  (AJ251029) odorant binding prote...    32   0.27
gi|14735860|ref|XP_029971.1|  61629 [Homo sapiens]                 32   0.27
gi|404390|gb|AAB27607.1|  beta-trace protein, prostaglandin ...    32   0.28
gi|5803139|ref|NP_006735.1|  retinol-binding protein 4, inte...    32   0.33
gi|190444|gb|AAA36494.1|  (M61900) prostaglandin D synthase ...    32   0.34
gi|14734789|ref|XP_034577.1|  61410 [Homo sapiens]                 30   1.1
gi|7657407|ref|NP_055396.1|  odorant-binding protein 2B [Hom...    30   1.4
gi|14734771|ref|XP_034580.1|  odorant-binding protein 2B [Ho...    29   1.9
gi|4504963|ref|NP_002288.1|  lipocalin 1 (protein migrating ...    29   2.4
gi|14734796|ref|XP_034583.1|  similar to lipocalin 1 (protei...    29   2.8
gi|14734775|ref|XP_034576.1|  61408 [Homo sapiens]                 29   3.1
gi|296672|emb|CAA26553.1|  (X02775) RBP [Homo sapiens]             28   5.4
gi|2495324|sp|Q92826|HXBD_HUMAN  HOMEOBOX PROTEIN HOX-B13 >g...    28   6.8
gi|5453686|ref|NP_006352.1|  homeo box B13; homeobox protein...    28   7.0
```

Fig. 4.19
page 127

Using human beta globin as a query, here are the blastp results searching against human RefSeq proteins (PAM30 matrix). Where is myoglobin? It’s absent! We need to use **PSI-BLAST**.



```
Distance tree of results NEW
Sequences producing significant alignments:                        (Bits)  Value

ref|NP_000509.1|  beta globin [Homo sapiens]                          441   2e-124  U G
ref|NP_000510.1|  delta globin [Homo sapiens]                         409   1e-114  U G
ref|NP_005321.1|  epsilon globin [Homo sapiens]                       319   2e-87   U G
ref|NP_000175.1|  G-gamma globin [Homo sapiens]                       310   5e-85   U G
ref|NP_000550.2|  A-gamma globin [Homo sapiens]                       304   4e-83   U G
ref|NP_000508.1|  alpha 2 globin [Homo sapiens] >ref|NP_000549...    96.5   2e-20   U G
ref|NP_005322.1|  theta 1 globin [Homo sapiens]                      62.1   3e-10   U G
ref|NP_005323.1|  zeta globin [Homo sapiens]                         50.3   1e-06   U G
ref|NP_001003938.1|  hemoglobin mu chain [Homo sapiens]              48.6   4e-06   U G
ref|NP_071396.3|  kinesin family member 13A [Homo sapiens]           28.6   4.5     U G
ref|NP_001517.1|  orexin receptor 2 [Homo sapiens]                   28.6   4.9     U G

                                                        Alignments

[Get selected sequences] [Select all] [Deselect all] [Distance tree of results]

>□ref|NP_000509.1| U G beta globin [Homo sapiens]
Length=147

 Score =  441 bits (1033),  Expect = 2e-124, Method: Composition-based stats.
 Identities = 147/147 (100%), Positives = 147/147 (100%), Gaps = 0/147 (0%)

Query  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
            MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
Sbjct  1    MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60

Query  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
            VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
Sbjct  61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120

Query  121  KEFTPPVQAAYQKVVAGVANALAHKYH  147
            KEFTPPVQAAYQKVVAGVANALAHKYH
Sbjct  121  KEFTPPVQAAYQKVVAGVANALAHKYH  147
```

Outline of today’s lecture

BLAST
- Practical use
- Algorithm
- Strategies

Finding distantly related proteins:
- PSI-BLAST
- Hidden Markov models

BLAST-like tools for genomic DNA
- PatternHunter
- Megablast
- BLAT, BLASTZ

Two problems standard BLAST cannot solve

[1] Use human beta globin as a query against human RefSeq proteins, and blastp does not “find” human myoglobin. This is because the two proteins are too distantly related. PSI-BLAST at NCBI as well as hidden Markov models easily solve this problem.

[2] How can we search using 10,000 base pairs as a query, or even millions of base pairs? Many BLAST-like tools for genomic DNA are available such as PatternHunter, Megablast, BLAT, and BLASTZ.

Page 141

Position specific iterated BLAST: PSI-BLAST

The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a scoring matrix that is customized to your query.

Page 146

PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

Page 146

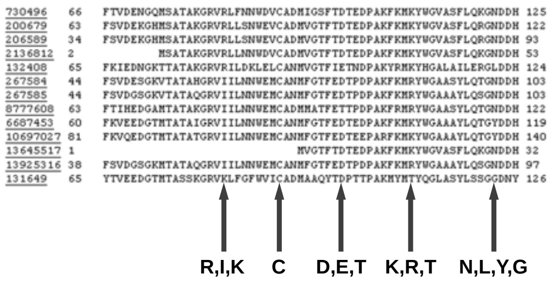
PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a "profile" or specialized position-specific scoring matrix (PSSM)

Page 146

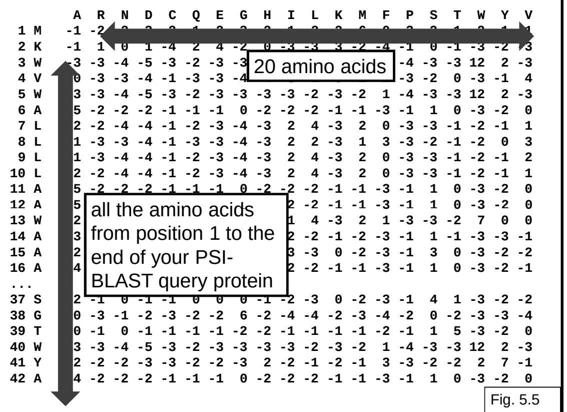
Inspect the blastp output to identify empirical "rules" regarding amino acids tolerated at each position



```
730496   66  FTVDENGQMSATAKGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDH 125
200679   63  FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 122
206589   34  FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTEDPAKFKMKYWGVASFLQRGNDDH 93
2136812  2          MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 53
132408   65  FKIEDNGKTTATAKGRVRILDKLELCANMVGTFIETNDPAKYRMKYHGALAILERGLDDH 124
267584   44  FSVDESGKVTATAHGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAASYLQTGNDDH 103
267585   44  FSVDGSGKVTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 103
8777608  63  FTIHEDGAMTATAKGRVIILNNWEMCADMMATFETTPDPAKFRMRYWGAASYLQTGNDDH 122
6687453  60  FKVEEDGTMTATAIGRVIILNNWEMCANMFGTFEDTEDPAKFKMKYWGAAAYLQTGYDDH 119
10697027 81  FKVQEDGTMTATATGRVIILNNWEMCANMFGTFEDTEEPARFKMKYWGAAAYLQTGYDDH 140
13645517 1                               MVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 32
13925316 38  FSVDGSGKMTATAQGRVIILNNWEMCANMFGTFEDTPDPAKFKMRYWGAAAYLQSGNDDH 97
131649   65  YTVEEDGTMTASSKGRVKLFGFWVICADMAAQYTDPTTPAKMYMTYQGLASYLSSGGDNY 126
```

R,I,K C D,E,T K,R,T N,L,Y,G

Fig. 5.4
Page 147



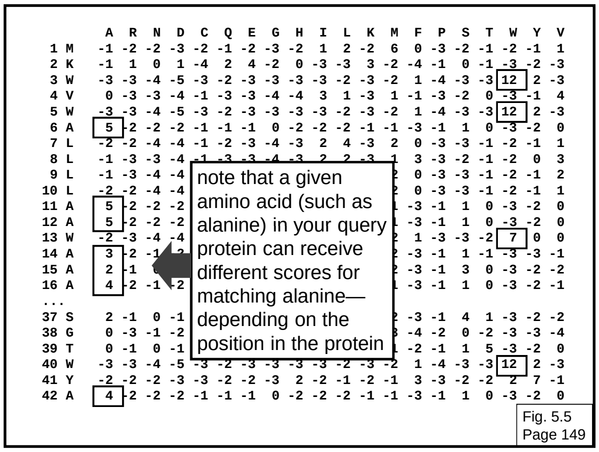
20 amino acids

all the amino acids from position 1 to the end of your PSI-BLAST query protein

Fig. 5.5
Page 149

```
       A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
 1 M  -1 -2 -2 -3 -2 -1 -2 -3 -2  1  2 -2  6  0 -3 -2 -1 -2 -1  1
 2 K  -1  1  0  1 -4  2  4 -2  0 -3 -3  3 -2 -4 -1  0 -1 -3 -2 -3
 3 W  -3 -3 -4 -5 -3 -2 -3 -3 -3 -3 -2 -3 -2  1 -4 -3 -3 12  2 -3
 4 V   0 -3 -3 -4 -1 -3 -3 -4 -4  3  1 -3  1 -1 -3 -2  0 -3 -1  4
 5 W  -3 -3 -4 -5 -3 -2 -3 -3 -3 -3 -2 -3 -2  1 -4 -3 -3 12  2 -3
 6 A   5 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0
 7 L  -2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
 8 L  -1 -3 -3 -4 -1 -3 -3 -4 -3  2  2 -3  1  3 -3 -2 -1 -2  0  3
 9 L  -1 -3 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  2
10 L  -2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
11 A   5 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0
12 A   5 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0
13 W  -2 -3 -4 -4 -2 -2 -3 -4 -3  1  4 -3  2  1 -3 -3 -2  7  0  0
14 A   3 -2 -1 -2 -1 -1 -2  4 -2 -2 -2 -1 -2 -3 -1  1 -1 -3 -3 -1
15 A   2 -1  0 -1 -2  2  0  2 -1 -3 -3  0 -2 -3 -1  3  0 -3 -2 -2
16 A   4 -2 -1 -2 -1 -1 -1  3 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2 -1
...
37 S   2 -1  0 -1 -1  0  0  0 -1 -2 -3  0 -2 -3 -1  4  1 -3 -2 -2
38 G   0 -3 -1 -2 -3 -2 -2  6 -2 -4 -4 -2 -3 -4 -2  0 -2 -3 -3 -4
39 T   0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -3 -2  0
40 W  -3 -3 -4 -5 -3 -2 -3 -3 -3 -3 -2 -3 -2  1 -4 -3 -3 12  2 -3
41 Y  -2 -2 -2 -3 -3 -2 -2 -3  2 -2 -1 -2 -1  3 -3 -2 -2  2  7 -1
42 A   4 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0
```

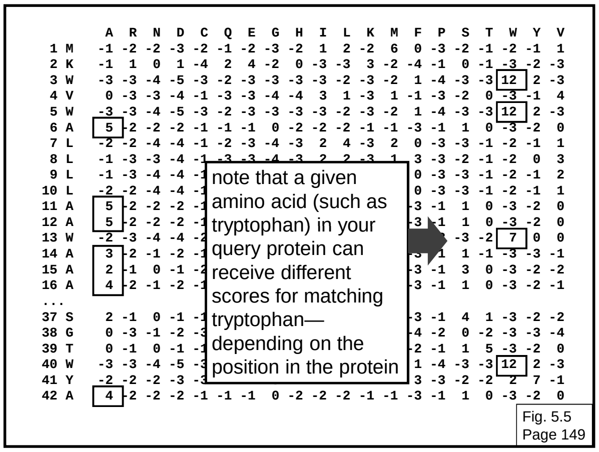
Fig. 5.5
Page 149



```
         A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
  1 M   -1 -2 -2 -3 -2 -1 -2 -3 -2  1  2 -2  6  0 -3 -2 -1 -2 -1  1
  2 K   -1  1  0  1 -4  2  4 -2  0 -3 -3  3 -2 -4 -1  0 -1 -3 -2 -3
  3 W   -3 -3 -4 -5 -3 -2 -3 -3 -3 -3 -2 -3 -2  1 -4 -3 -3 12  2 -3
  4 V    0 -3 -3 -4 -1 -3 -3 -4 -4  3  1 -3  1 -1 -3 -2  0 -3 -1  4
  5 W   -3 -3 -4 -5 -3 -2 -3 -3 -3 -3 -2 -3 -2  1 -4 -3 -3 12  2 -3
  6 A    5 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0
  7 L   -2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
  8 L   -1 -3 -3 -4 -1 -3 -3 -4 -3  2  2 -3  1  3 -3 -2 -1 -2  0  3
  9 L   -1 -3 -4 -4                          2  0 -3 -3 -1 -2 -1  2
 10 L   -2 -2 -4 -4                          2  0 -3 -3 -1 -2 -1  1
 11 A    5 -2 -2 -2                          1 -3 -1  1  0 -3 -2  0
 12 A    5 -2 -2 -2                          1 -3 -1  1  0 -3 -2  0
 13 W   -2 -3 -4 -4                          2  1 -3 -3 -2  7  0  0
 14 A    3 -2 -1                             2 -3 -1  1 -1 -3 -3 -1
 15 A    2 -1  0                             2 -3 -1  3  0 -3 -2 -2
 16 A    4 -2 -1 -2                          1 -3 -1  1  0 -3 -2 -1
 ...
 37 S    2 -1  0 -1                          2 -3 -1  4  1 -3 -2 -2
 38 G    0 -3 -1 -2                          3 -4 -2  0 -2 -3 -3 -4
 39 T    0 -1  0 -1                          1 -2 -1  1  5 -3 -2  0
 40 W   -3 -3 -4 -5 -3 -2 -3 -3 -3 -3 -2 -3 -2  1 -4 -3 -3 12  2 -3
 41 Y   -2 -2 -2 -3 -3 -2 -2 -3  2 -2 -1 -2 -1  3 -3 -2 -2  2  7 -1
 42 A    4 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0
```

note that a given amino acid (such as alanine) in your query protein can receive different scores for matching alanine—depending on the position in the protein

Fig. 5.5
Page 149



```
         A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
  1 M   -1 -2 -2 -3 -2 -1 -2 -3 -2  1  2 -2  6  0 -3 -2 -1 -2 -1  1
  2 K   -1  1  0  1 -4  2  4 -2  0 -3 -3  3 -2 -4 -1  0 -1 -3 -2 -3
  3 W   -3 -3 -4 -5 -3 -2 -3 -3 -3 -3 -2 -3 -2  1 -4 -3 -3 12  2 -3
  4 V    0 -3 -3 -4 -1 -3 -3 -4 -4  3  1 -3  1 -1 -3 -2  0 -3 -1  4
  5 W   -3 -3 -4 -5 -3 -2 -3 -3 -3 -3 -2 -3 -2  1 -4 -3 -3 12  2 -3
  6 A    5 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0
  7 L   -2 -2 -4 -4 -1 -2 -3 -4 -3  2  4 -3  2  0 -3 -3 -1 -2 -1  1
  8 L   -1 -3 -3 -4 -1 -3 -3 -4 -3  2  2 -3  1  3 -3 -2 -1 -2  0  3
  9 L   -1 -3 -4 -4 -1                          0 -3 -3 -1 -2 -1  2
 10 L   -2 -2 -4 -4 -1                          0 -3 -3 -1 -2 -1  1
 11 A    5 -2 -2 -2 -1                         -3 -1  1  0 -3 -2  0
 12 A    5 -2 -2 -2 -1                         -3 -1  1  0 -3 -2  0
 13 W   -2 -3 -4 -4 -2                               -3 -2  7  0  0
 14 A    3 -2 -1 -2 -1                                1 -1 -3 -3 -1
 15 A    2 -1  0 -1 -2                         -3 -1  3  0 -3 -2 -2
 16 A    4 -2 -1 -2 -1                         -3 -1  1  0 -3 -2 -1
 ...
 37 S    2 -1  0 -1 -1                         -3 -1  4  1 -3 -2 -2
 38 G    0 -3 -1 -2 -3                         -4 -2  0 -2 -3 -3 -4
 39 T    0 -1  0 -1 -1                         -2 -1  1  5 -3 -2  0
 40 W   -3 -3 -4 -5 -3                       1 -4 -3 -3 12  2 -3
 41 Y   -2 -2 -2 -3 -3                          3 -3 -2 -2  2  7 -1
 42 A    4 -2 -2 -2 -1 -1 -1  0 -2 -2 -2 -1 -1 -3 -1  1  0 -3 -2  0
```

note that a given amino acid (such as tryptophan) in your query protein can receive different scores for matching tryptophan—depending on the position in the protein

Fig. 5.5
Page 149

PSI-BLAST is performed in five steps

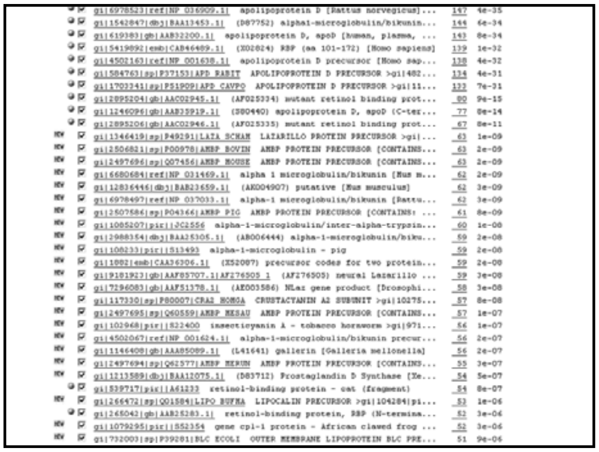
[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)

[3] The PSSM is used as a query against the database

[4] PSI-BLAST estimates statistical significance (E values)

Page 146



```
   ● ☑ gi|6978523|ref|NP_036909.1|  apolipoprotein D [Rattus norvegicus]...   147  4e-35
   ● ☑ gi|1542847|dbj|BAA13453.1|  (D87752) alpha1-microglobulin/bikunin...   144  6e-34
   ● ☑ gi|619383|gb|AAB32200.1|  apolipoprotein D, apoD [human, plasma, ...   143  8e-34
   ● ☑ gi|5419892|emb|CAB46489.1|  (X02824) RBP (aa 101-172) [Homo sapiens]   139  1e-32
   ● ☑ gi|4502163|ref|NP_001638.1|  apolipoprotein D precursor [Homo sap...   138  4e-32
   ● ☑ gi|584763|sp|P37153|APD_RABIT  APOLIPOPROTEIN D PRECURSOR >gi|482...   134  4e-31
   ● ☑ gi|1703341|sp|P51909|APD_CAVPO  APOLIPOPROTEIN D PRECURSOR >gi|11...   133  7e-31
   ● ☑ gi|2895204|gb|AAC02945.1|  (AF025334) mutant retinol binding prot...    80  9e-15
   ● ☑ gi|1246096|gb|AAB35919.1|  (S80440) apolipoprotein D, apoD (C-ter...    77  8e-14
   ● ☑ gi|2895206|gb|AAC02946.1|  (AF025335) mutant retinol binding prot...    67  8e-11
NEW  ☑ gi|1346419|sp|P49291|LAZA_SCHAM  LAZARILLO PROTEIN PRECURSOR >gi|...    63  1e-09
NEW  ☑ gi|2506821|sp|P00978|AMBP_BOVIN  AMBP PROTEIN PRECURSOR [CONTAINS...    63  2e-09
NEW  ☑ gi|2497696|sp|Q07456|AMBP_MOUSE  AMBP PROTEIN PRECURSOR [CONTAINS...    63  2e-09
NEW  ☑ gi|6680684|ref|NP_031469.1|  alpha 1 microglobulin/bikunin [Mus m...    62  2e-09
NEW  ☑ gi|12836446|dbj|BAB23659.1|  (AK004907) putative [Mus musculus]         62  3e-09
NEW  ☑ gi|6978497|ref|NP_037033.1|  alpha-1 microglobulin/bikunin [Rattu...    62  3e-09
NEW  ☑ gi|2507586|sp|P04366|AMBP_PIG  AMBP PROTEIN PRECURSOR [CONTAINS: ...    61  8e-09
NEW  ☑ gi|1085207|pir||JC2556  alpha-1-microglobulin/inter-alpha-trypsin...    60  1e-08
NEW  ☑ gi|2988354|dbj|BAA25305.1|  (AB006444) alpha-1-microglobulin/biku...    59  2e-08
NEW  ☑ gi|108233|pir||S13493  alpha-1-microglobulin - pig                      59  2e-08
NEW  ☑ gi|1882|emb|CAA36306.1|  (X52087) precursor codes for two protein...    59  2e-08
NEW  ☑ gi|9181923|gb|AAF85707.1|AF276505_1  (AF276505) neural Lazarillo ...    59  3e-08
NEW  ☑ gi|7296083|gb|AAF51378.1|  (AE003586) NLaz gene product [Drosophi...    58  3e-08
NEW  ☑ gi|117330|sp|P80007|CRA2_HOMGA  CRUSTACYANIN A2 SUBUNIT >gi|10275...    57  8e-08
NEW  ☑ gi|2497695|sp|Q60559|AMBP_MESAU  AMBP PROTEIN PRECURSOR [CONTAINS...    57  1e-07
NEW  ☑ gi|102968|pir||S22400  insecticyanin A - tobacco hornworm >gi|971...    56  1e-07
NEW  ☑ gi|4502067|ref|NP_001624.1|  alpha-1-microglobulin/bikunin precur...    56  2e-07
NEW  ☑ gi|1146408|gb|AAA85089.1|  (L41641) gallerin [Galleria mellonella]      56  2e-07
NEW  ☑ gi|2497694|sp|Q62577|AMBP_MERUN  AMBP PROTEIN PRECURSOR [CONTAINS...    55  3e-07
NEW  ☑ gi|1213589|dbj|BAA12075.1|  (D83712) Prostaglandin D Synthase [Xe...    54  5e-07
   ● ☑ gi|539717|pir||A61233  retinol-binding protein - cat (fragment)          54  8e-07
NEW  ☑ gi|266472|sp|Q01584|LIPO_BUFMA  LIPOCALIN PRECURSOR >gi|104284|pi...    53  1e-06
   ● ☑ gi|265042|gb|AAB25283.1|  retinol-binding protein, RBP (N-termina...    52  3e-06
NEW  ☑ gi|1079295|pir||S52354  gene cpl-1 protein - African clawed frog ...    52  3e-06
NEW  ☑ gi|732003|sp|P39281|BLC_ECOLI  OUTER MEMBRANE LIPOPROTEIN BLC PRE...    51  9e-06
```

PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)

[3] The PSSM is used as a query against the database

[4] PSI-BLAST estimates statistical significance (E values)

[5] Repeat steps [3] and [4] iteratively, typically 5 times.
At each new search, a new profile is used as the query.

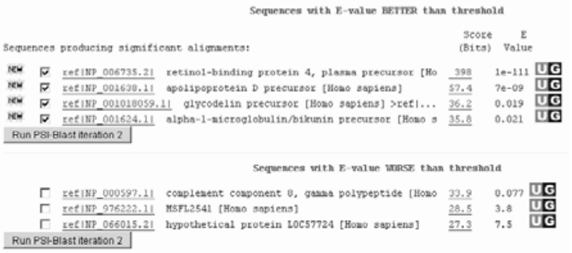
Page 146

Results of a PSI-BLAST search

Iteration	# hits	# hits > threshold
1	104	49
2	173	96
3	236	178
4	301	240
5	344	283
6	342	298
7	378	310
8	382	320

Table 5-2
Page 146

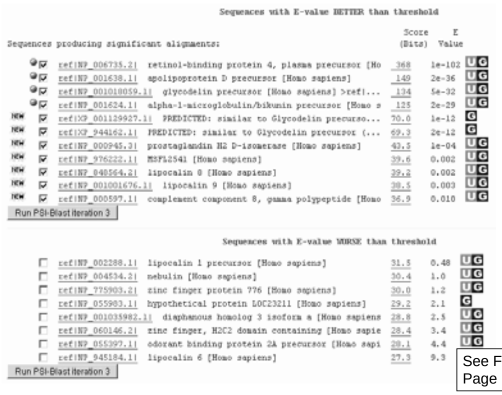
PSI-BLAST search: human RBP versus RefSeq, iteration 1



```
                                    Sequences with E-value BETTER than threshold

                                                                                  Score     E
Sequences producing significant alignments:                                      (Bits)   Value

NEW  ☑  ref|NP_006735.2|  retinol-binding protein 4, plasma precursor [Ho   398    1e-111  UG
NEW  ☑  ref|NP_001638.1|  apolipoprotein D precursor [Homo sapiens]         57.4   7e-09   UG
NEW  ☑  ref|NP_001018059.1|  glycodelin precursor [Homo sapiens] >ref|...   36.2   0.019   UG
NEW  ☑  ref|NP_001624.1|  alpha-1-microglobulin/bikunin precursor [Homo s   35.8   0.021   UG
Run PSI-Blast iteration 2

                                    Sequences with E-value WORSE than threshold

     □  ref|NP_000597.1|  complement component 8, gamma polypeptide [Homo   33.9   0.077   UG
     □  ref|NP_976222.1|  MSFL2541 [Homo sapiens]                           28.5   3.8     UG
     □  ref|NP_066015.2|  hypothetical protein LOC57724 [Homo sapiens]      27.3   7.5     UG
Run PSI-Blast iteration 2
```

See Fig. 5.6
Page 150

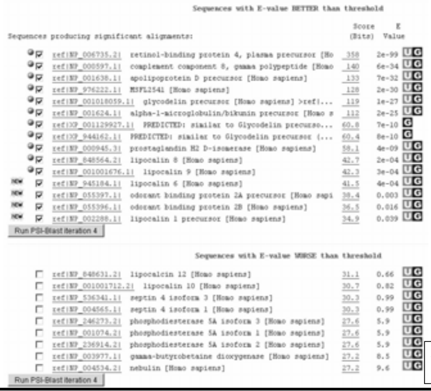
PSI-BLAST search: human RBP versus RefSeq, iteration 2



```
                                  Sequences with E-value BETTER than threshold

                                                                                 Score     E
Sequences producing significant alignments:                                     (Bits)   Value

    ●☑  ref|NP_006735.2|  retinol-binding protein 4, plasma precursor [Ho   368    1e-102  UG
    ●☑  ref|NP_001638.1|  apolipoprotein D precursor [Homo sapiens]         149    2e-36   UG
    ●☑  ref|NP_001018059.1|  glycodelin precursor [Homo sapiens] >ref|...   134    5e-32   UG
    ●☑  ref|NP_001624.1|  alpha-1-microglobulin/bikunin precursor [Homo s   125    2e-29   UG
NEW  ☑  ref|XP_001129927.1|  PREDICTED: similar to Glycodelin precurso...   70.0   1e-12   G
NEW  ☑  ref|XP_944162.1|  PREDICTED: similar to Glycodelin precursor (...   69.3   2e-12   G
NEW  ☑  ref|NP_000945.3|  prostaglandin H2 D-isomerase [Homo sapiens]       43.5   1e-04   UG
NEW  ☑  ref|NP_976222.1|  MSFL2541 [Homo sapiens]                           39.6   0.002   UG
NEW  ☑  ref|NP_848564.2|  lipocalin 8 [Homo sapiens]                        39.2   0.002   UG
NEW  ☑  ref|NP_001001676.1|  lipocalin 9 [Homo sapiens]                     38.5   0.003   UG
NEW  ☑  ref|NP_000597.1|  complement component 8, gamma polypeptide [Homo   36.9   0.010   UG
Run PSI-Blast iteration 3

                                  Sequences with E-value WORSE than threshold

     □  ref|NP_002288.1|  lipocalin 1 precursor [Homo sapiens]              31.5   0.48    UG
     □  ref|NP_004534.2|  nebulin [Homo sapiens]                            30.4   1.0     UG
     □  ref|NP_775903.2|  zinc finger protein 776 [Homo sapiens]            30.0   1.2     UG
     □  ref|NP_055983.1|  hypothetical protein LOC23211 [Homo sapiens]      29.2   2.1     G
     □  ref|NP_001035982.1|  diaphanous homolog 3 isoform a [Homo sapiens   28.8   2.5     UG
     □  ref|NP_060146.2|  zinc finger, H2C2 domain containing [Homo sapie   28.4   3.4     UG
     □  ref|NP_055397.1|  odorant binding protein 2A precursor [Homo sapi   28.1   4.4     UG
     □  ref|NP_945184.1|  lipocalin 6 [Homo sapiens]                        27.3   9.3
Run PSI-Blast iteration 3
```

See Fig. 5.6
Page 150

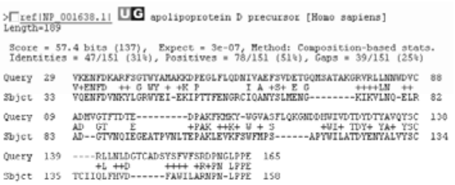
PSI-BLAST search: human RBP versus RefSeq, iteration 3



```
                                  Sequences with E-value BETTER than threshold

                                                                                 Score     E
Sequences producing significant alignments:                                     (Bits)   Value

    ●☑  ref|NP_006735.2|  retinol-binding protein 4, plasma precursor [Ho   358    2e-99   UG
    ●☑  ref|NP_000597.1|  complement component 8, gamma polypeptide [Homo   140    6e-34   UG
    ●☑  ref|NP_001638.1|  apolipoprotein D precursor [Homo sapiens]         133    7e-32   UG
    ●☑  ref|NP_976222.1|  MSFL2541 [Homo sapiens]                           128    2e-30   UG
    ●☑  ref|NP_001018059.1|  glycodelin precursor [Homo sapiens] >ref|...   119    1e-27   UG
    ●☑  ref|NP_001624.1|  alpha-1-microglobulin/bikunin precursor [Homo s   112    2e-25   UG
    ●☑  ref|XP_001129927.1|  PREDICTED: similar to Glycodelin precurso...   60.8   7e-10   G
    ●☑  ref|XP_944162.1|  PREDICTED: similar to Glycodelin precursor (...   60.4   8e-10   G
    ●☑  ref|NP_000945.3|  prostaglandin H2 D-isomerase [Homo sapiens]       58.1   4e-09   UG
    ●☑  ref|NP_848564.2|  lipocalin 8 [Homo sapiens]                        42.7   2e-04   UG
    ●☑  ref|NP_001001676.1|  lipocalin 9 [Homo sapiens]                     42.3   3e-04   UG
NEW  ☑  ref|NP_945184.1|  lipocalin 6 [Homo sapiens]                        41.5   4e-04   UG
NEW  ☑  ref|NP_055397.1|  odorant binding protein 2A precursor [Homo sapi   38.4   0.003   UG
NEW  ☑  ref|NP_055396.1|  odorant binding protein 2B [Homo sapiens]         36.5   0.016   UG
NEW  ☑  ref|NP_002288.1|  lipocalin 1 precursor [Homo sapiens]              34.9   0.039   UG
Run PSI-Blast iteration 4

                                  Sequences with E-value WORSE than threshold

     □  ref|NP_848631.2|  lipocalcin 12 [Homo sapiens]                      31.1   0.66    UG
     □  ref|NP_001001712.2|  lipocalin 10 [Homo sapiens]                    30.7   0.82    UG
     □  ref|NP_536341.1|  septin 4 isoform 3 [Homo sapiens]                 30.3   0.99    UG
     □  ref|NP_004565.1|  septin 4 isoform 1 [Homo sapiens]                 30.3   0.99    UG
     □  ref|NP_246273.2|  phosphodiesterase 5A isoform 3 [Homo sapiens]     27.6   5.9     UG
     □  ref|NP_001074.2|  phosphodiesterase 5A isoform 1 [Homo sapiens]     27.6   5.9     UG
     □  ref|NP_236914.2|  phosphodiesterase 5A isoform 2 [Homo sapiens]     27.6   5.9     UG
     □  ref|NP_003977.1|  gamma-butyrobetaine dioxygenase [Homo sapiens]    27.2   8.5     UG
     □  ref|NP_004534.2|  nebulin [Homo sapiens]                            27.2   9.6     UG
Run PSI-Blast iteration 4
```

See Fig. 5.6
Page 150

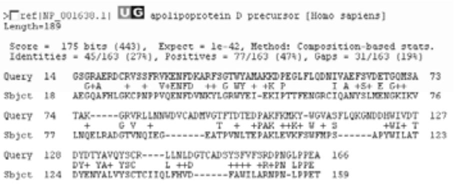
RBP4 match to ApoD, PSI-BLAST iteration 1
E value 3e-07



```
>□ref|NP_001638.1| UG apolipoprotein D precursor [Homo sapiens]
Length=189

 Score = 57.4 bits (137),  Expect = 3e-07, Method: Composition-based stats.
 Identities = 47/151 (31%), Positives = 78/151 (51%), Gaps = 39/151 (25%)

Query  29   VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC  88
            V+ENFD  ++ G WY + +K P         I A +S+ E G        ++++LN  ++
Sbjct  33   VQENFDVNKYLGRWYEI-EKIPTTFENGRCIQANYSLMENG--------KIKVLNQ-ELR  82

Query  89   ADMVGTFTDTE---------DPAKFKMKY-WGVASFLQKGNDDHWIVDTDYDTYAVQYSC  138
            AD  GT    E         +PAK ++K+ W + S         +WI+ TDY+ YA+ YSC
Sbjct  83   AD--GTVNQIEGEATPVNLTEPAKLEVKFSWFMPS------APYWILATDYENYALVYSC  134

Query  139  ----RLLNLDGTCADSYSFVFSRDPNGLPPE  165
                +L ++D      ++++ +R+PN LPPE
Sbjct  135  TCIIQLFHVD------FAWILARNPN-LPPE  158
```

Fig. 5.6
Page 150

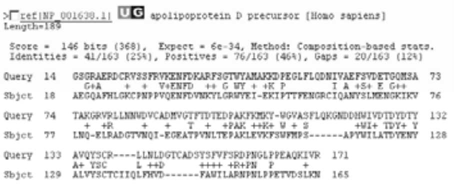
RBP4 match to ApoD, PSI-BLAST iteration 2
E value 1e-42
Note that PSI-BLAST E values can improve dramatically!



```
>□ref|NP_001638.1| UG apolipoprotein D precursor [Homo sapiens]
Length=189

 Score =  175 bits (443),  Expect = 1e-42, Method: Composition-based stats.
 Identities = 45/163 (27%), Positives = 77/163 (47%), Gaps = 31/163 (19%)

Query  14   GSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSA  73
              G+A    +  + V+ENFD  ++ G WY + +K P         I A +S+ E G++
Sbjct  18   AEGQAFHLGKCPNPPVQENFDVNKYLGRWYEI-EKIPTTFENGRCIQANYSLMENGKIKV  76

Query  74   TAK-----GRVRLLNNWDVCADMVGTFTDTEDPAKFKMKY-WGVASFLQKGNDDHWIVDT  127
              +     G V  +               T  +  +PAK ++K+ W + S         +WI+ T
Sbjct  77   LNQELRADGTVNQIEG-------EATPVNLTEPAKLEVKFSWFMPS------APYWILAT  123

Query  128  DYDTYAVQYSCR----LLNLDGTCADSYSFVFSRDPNGLPPEA  166
            DY+ YA+ YSC     L ++D      ++++ +R+PN LPPE
Sbjct  124  DYENYALVYSCTCIIQLFHVD------FAWILARNPN-LPPET  159
```

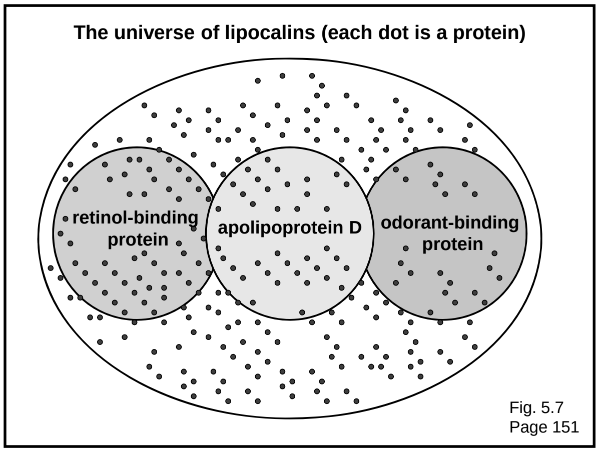
Fig. 5.6
Page 150

RBP4 match to ApoD, PSI-BLAST iteration 3
E value 6e-34



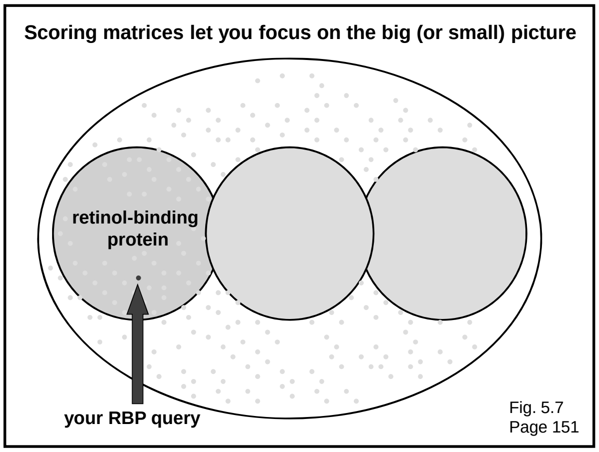
```
>□ref|NP_001638.1| UG apolipoprotein D precursor [Homo sapiens]
Length=189

 Score =  146 bits (368),  Expect = 6e-34, Method: Composition-based stats.
 Identities = 41/163 (25%), Positives = 76/163 (46%), Gaps = 20/163 (12%)

Query  14   GSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSA  73
              G+A    +  + V+ENFD  ++ G WY + +K P         I A +S+ E G++
Sbjct  18   AEGQAFHLGKCPNPPVQENFDVNKYLGRWYEI-EKIPTTFENGRCIQANYSLMENGKIKV  76

Query  74   TAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKY-WGVASFLQKGNDDHWIVDTDYDTY  132
              +  +R      +   +   T  +  +PAK ++K+ W + S         +WI+ TDY+ Y
Sbjct  77   LNQ-ELRADGTVNQI-EGEATPVNLTEPAKLEVKFSWFMPS------APYWILATDYENY  128

Query  133  AVQYSCR----LLNLDGTCADSYSFVFSRDPNGLPPEAQKIVR  171
            A+ YSC     L ++D      ++++ +R+PN  P    +
Sbjct  129  ALVYSCTCIIQLFHVD------FAWILARNPNLPPETVDSLKN  165
```

Fig. 5.6
Page 150



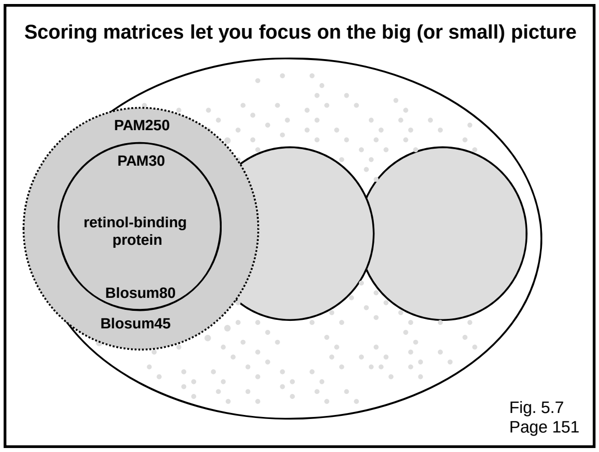
The universe of lipocalins (each dot is a protein)

Fig. 5.7
Page 151



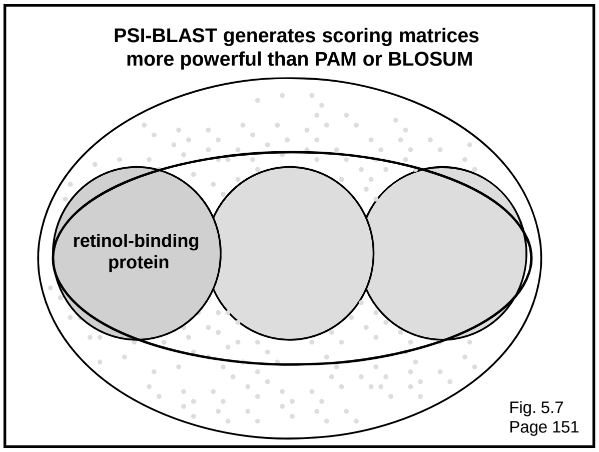
Scoring matrices let you focus on the big (or small) picture

Fig. 5.7
Page 151



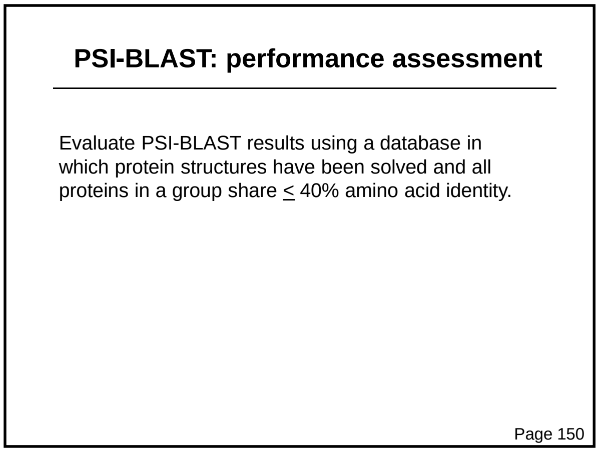
Scoring matrices let you focus on the big (or small) picture

Fig. 5.7
Page 151



PSI-BLAST generates scoring matrices
more powerful than PAM or BLOSUM

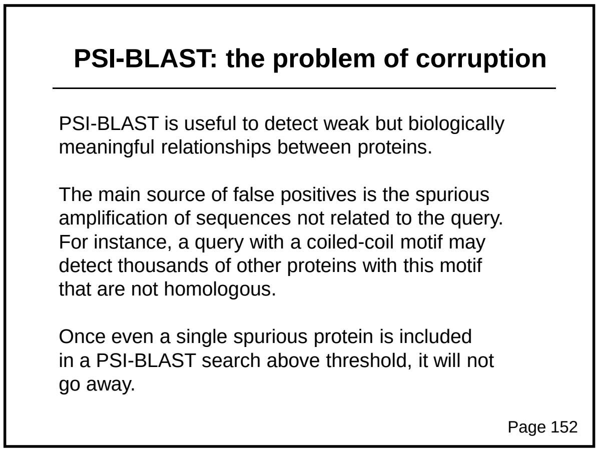
Fig. 5.7
Page 151



PSI-BLAST: performance assessment

Evaluate PSI-BLAST results using a database in which protein structures have been solved and all proteins in a group share ≤ 40% amino acid identity.

Page 150



PSI-BLAST: the problem of corruption

PSI-BLAST is useful to detect weak but biologically meaningful relationships between proteins.

The main source of false positives is the spurious amplification of sequences not related to the query. For instance, a query with a coiled-coil motif may detect thousands of other proteins with this motif that are not homologous.

Once even a single spurious protein is included in a PSI-BLAST search above threshold, it will not go away.

Page 152

PSI-BLAST: the problem of corruption

Corruption is defined as the presence of at least one false positive alignment with an E value < 10^{-4} after five iterations.

Three approaches to stopping corruption:

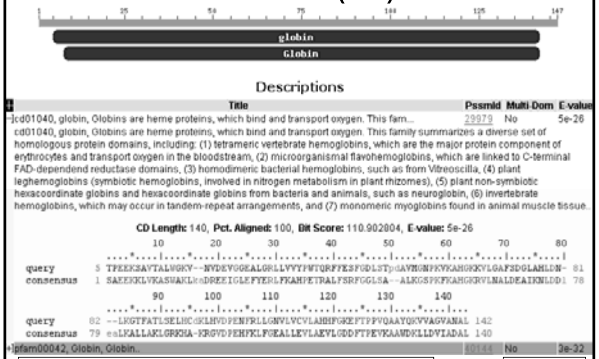
[1] Apply filtering of biased composition regions

[2] Adjust E value from 0.001 (default) to a lower value such as E = 0.0001.

[3] Visually inspect the output from each iteration. Remove suspicious hits by unchecking the box.

Page 152

Conserved domain database (CDD) uses RPS-BLAST



Descriptions

Title	Pssmld	Multi-Dom	E-value
cd01040, globin, Globins are heme proteins, which bind and transport oxygen. This fam...	29979	No	5e-26

cd01040, globin, Globins are heme proteins, which bind and transport oxygen. This family summarizes a diverse set of homologous protein domains, including: (1) tetrameric vertebrate hemoglobins, which are the major protein component of erythrocytes and transport oxygen in the bloodstream, (2) microorganismal flavohemoglobins, which are linked to C-terminal FAD-dependend reductase domains, (3) homodimeric bacterial hemoglobins, such as from Vitreoscilla, (4) plant leghemoglobins (symbiotic hemoglobins, involved in nitrogen metabolism in plant rhizomes), (5) plant non-symbiotic hexacoordinate globins and hexacoordinate globins from bacteria and animals, such as neuroglobin, (6) invertebrate hemoglobins, which may occur in tandem-repeat arrangements, and (7) monomeric myoglobins found in animal muscle tissue...

CD Length: 140, **Pct. Aligned:** 100, **Bit Score:** 110.902804, **E-value:** 5e-26

```
                   10        20        30        40        50        60        70        80
           ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
query      5 TPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTpdAVMGNPKVKAHGKKVLGAFSDGLAHLDN- 81
consensus  1 SAEEKKLVKASWAKLkaDREEIGLEFYERLFKAHPETRALFSRFGGLSA--ALKGSPKFKAHGKRVLNALDEAIKNLDDl 78
                   90       100       110       120       130       140
           ....*....|....*....|....*....|....*....|....*....|....*....|...
query     82 --LKGTFATLSELHCdKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANAL 142
consensus 79 eaLKALLAKLGRKHA-KRGVDPEHFKLFGEALLEVLAEVLGDDFTPEVKAAWDKLLDVIADAL 140
```

pfam00042, Globin, Globin..	40144	No	3e-32

Main idea: you can search a query protein against a database of position-specific scoring matrices

Fig. 5.8
Page 153

Outline of today's lecture

BLAST
- Practical use
- Algorithm
- Strategies

Finding distantly related proteins:
- PSI-BLAST
- Hidden Markov models

BLAST-like tools for genomic DNA
- PatternHunter
- Megablast
- BLAT, BLASTZ

Multiple sequence alignment to profile HMMs

- in the 1990's people began to see that aligning sequences to profiles gave much more information than pairwise alignment alone.

- Hidden Markov models (HMMs) are "states" that describe the probability of having a particular amino acid residue at arranged in a column of a multiple sequence alignment

- HMMs are probabilistic models

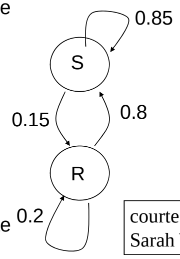
- Like a hammer is more refined than a blast, an HMM gives more sensitive alignments than traditional techniques such as progressive alignments

Page 155

Simple Markov Model

Rain = dog may not want to go outside

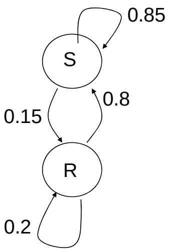
Sun = dog will probably go outside



Markov condition = no dependency on anything but nearest previous state ("memoryless")

courtesy of Sarah Wheelan

Simple Hidden Markov Model



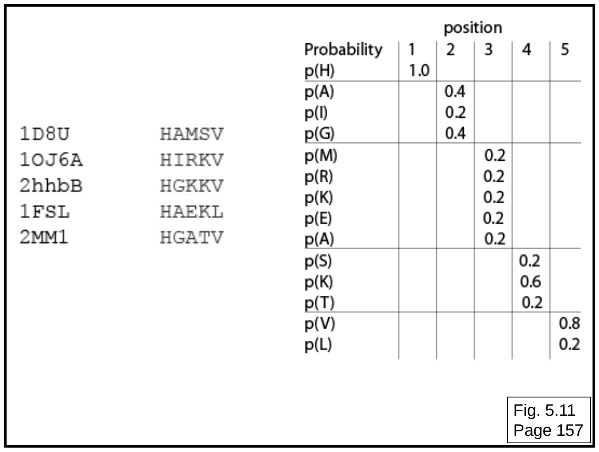
P(dog goes out in rain) = 0.1

P(dog goes out in sun) = 0.85

Observation: YNNNYYNNNYN

(Y=goes out, N=doesn't go out)

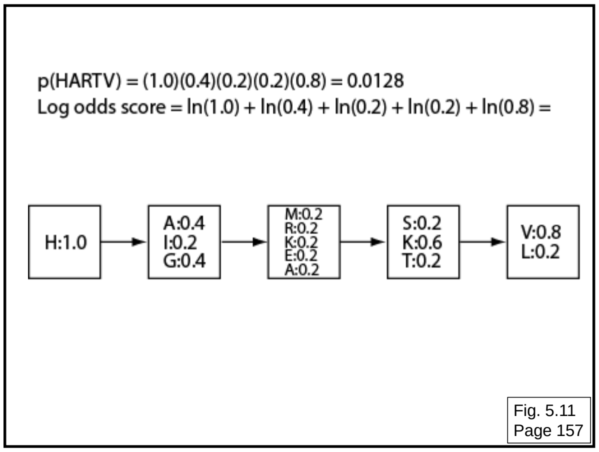
What is underlying reality (the hidden state chain)?



1D8U	HAMSV
1OJ6A	HIRKV
2hhbB	HGKKV
1FSL	HAEKL
2MM1	HGATV

	position				
Probability	1	2	3	4	5
p(H)	1.0				
p(A)		0.4			
p(I)		0.2			
p(G)		0.4			
p(M)			0.2		
p(R)			0.2		
p(K)			0.2		
p(E)			0.2		
p(A)			0.2		
p(S)				0.2	
p(K)				0.6	
p(T)				0.2	
p(V)					0.8
p(L)					0.2

Fig. 5.11
Page 157



p(HARTV) = (1.0)(0.4)(0.2)(0.2)(0.8) = 0.0128

Log odds score = ln(1.0) + ln(0.4) + ln(0.2) + ln(0.2) + ln(0.8) =

Fig. 5.11
Page 157

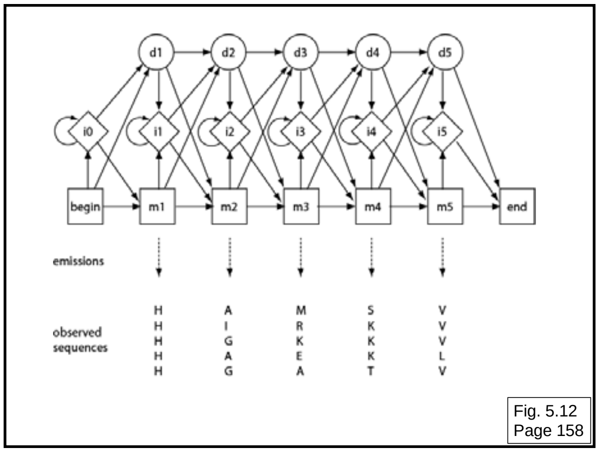


Fig. 5.12
Page 158

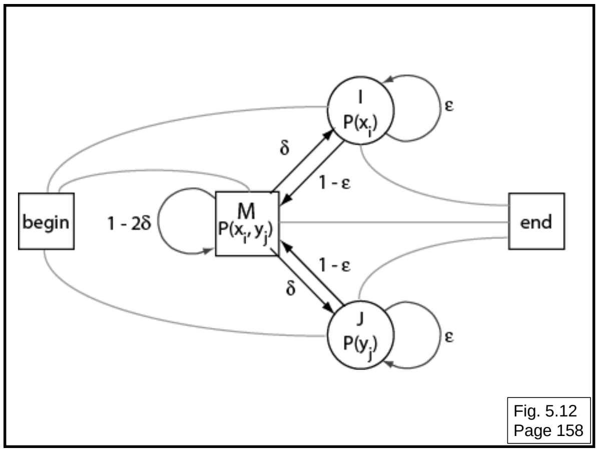


Fig. 5.12
Page 158

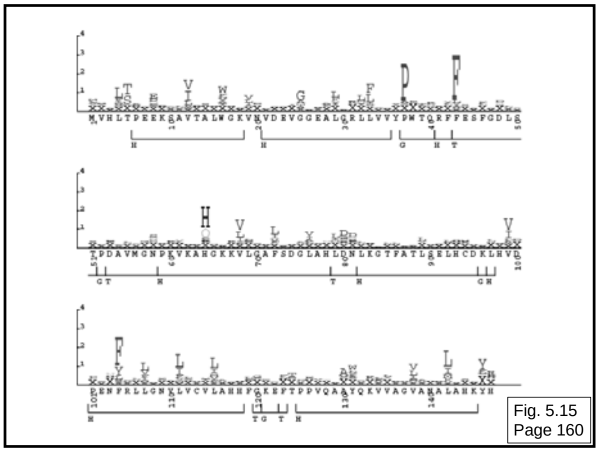
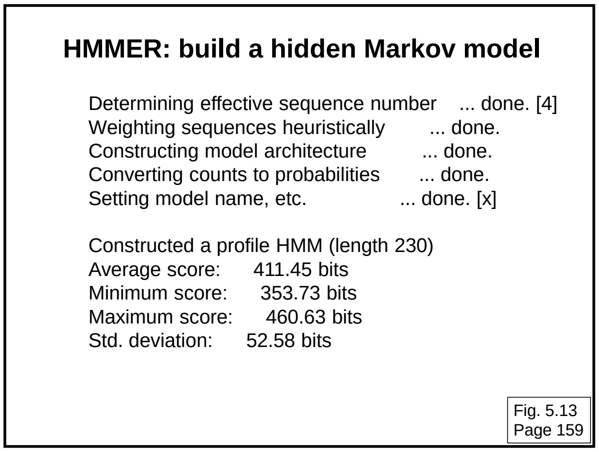


Fig. 5.15
Page 160



HMMER: build a hidden Markov model

```
Determining effective sequence number    ... done. [4]
Weighting sequences heuristically        ... done.
Constructing model architecture          ... done.
Converting counts to probabilities       ... done.
Setting model name, etc.                 ... done. [x]

Constructed a profile HMM (length 230)
Average score:      411.45 bits
Minimum score:      353.73 bits
Maximum score:      460.63 bits
Std. deviation:     52.58 bits
```

Fig. 5.13
Page 159

HMMER: calibrate a hidden Markov model

```
HMM file:                lipocalins.hmm
Length distribution mean: 325
Length distribution s.d.: 200
Number of samples:        5000
random seed:              1034351005
histogram(s) saved to:    [not saved]
POSIX threads:            2
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

HMM    : x
mu     : -123.894508
lambda :    0.179608
max    :  -79.334000
```

Fig. 5.13
Page 159

HMMER: search an HMM against GenBank

```
Scores for complete sequences (score includes all domains):
Sequence                              Description       Score    E-value  N
--------                              -----------       -----    ------- ---
gi|20888903|ref|XP_129259.1|          (XM_129259) ret   461.1   1.9e-133   1
gi|132407|sp|P04916|RETB_RAT          Plasma retinol-   458.0   1.7e-132   1
gi|20548126|ref|XP_005907.5|          (XM_005907) sim   454.9   1.4e-131   1
gi|5803139|ref|NP_006735.1|           (NM_006744) ret   454.6   1.7e-131   1
gi|20141667|sp|P02753|RETB_HUMAN      Plasma retinol-   451.1   1.9e-130   1
.
.
gi|16767588|ref|NP_463203.1|          (NC_003197) out   318.2    1.9e-90   1



gi|5803139|ref|NP_006735.1|: domain 1 of 1, from 1 to 195: score 454.6, E = 1.7e-131
                   *->mkwVMkLLLLaALagvfgaAErdAfsvgkCrvpsPPRGfrVkeNFDv
                      mkwV++LLLLaA  +  +aAErd      Crv+s    frVkeNFD+
  gi|5803139     1    MKWVWALLLLAA--W--AAAERD------CRVSS----FRVKENFDK 33

                   erylGtWYeIaKkDprFErGLllqdkItAeySleEhGsMsataeGrirVL
                   +r++GtWY++aKkDp  E GL+lqd+I+Ae+S++E+G+Msata+Gr+r+L
  gi|5803139    34 ARFSGTWYAMAKKDP--E-GLFLQDNIVAEFSVDETGQMSATAKGRVRLL 80

                   eNkelcADkvGTvtqiEGeasevfLtadPaklklKyaGvaSflqpGfddy
                   +N+++cAD+vGT+t++E          dPak+k+Ky+GvaSflq+G+dd+
  gi|5803139    81 NNWDVCADMVGTFTDTE----------DPAKFKMKYWGVASFLQKGNDDH 120
```

Fig. 5.13
Page 159

PFAM is a database of HMMs and an essential resource for protein families http://pfam.sanger.ac.uk/

Pfam 24.0 (October 2009, 11912 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUERY PFAM BY KEYWORD

Search for keywords in text data in the Pfam database.

You can also use the keyword search box at the top of every page.

Outline of today’s lecture

BLAST
- Practical use
- Algorithm
- Strategies

Finding distantly related proteins:
- PSI-BLAST
- Hidden Markov models

BLAST-like tools for genomic DNA
- PatternHunter
- Megablast
- BLAT, BLASTZ

BLAST-related tools for genomic DNA

The analysis of genomic DNA presents special challenges:
- There are exons (protein-coding sequence) and introns (intervening sequences).
- There may be sequencing errors or polymorphisms
- The comparison may between be related species (e.g. human and mouse)

Page 161

BLAST-related tools for genomic DNA

Recently developed tools include:

- MegaBLAST at NCBI.

- BLAT (BLAST-like alignment tool). BLAT parses an entire genomic DNA database into words (11mers), then searches them against a query. Thus it is a mirror image of the BLAST strategy. See http://genome.ucsc.edu

- SSAHA at Ensembl uses a similar strategy as BLAT. See http://www.ensembl.org

Page 162

PatternHunter

```
11111111111
ATGGTGCATCT (example of a seed)(extended)
ATTGTGCATCT (example of a mismatch)(not extended)

110100110010101111
ATGGTGCATCTGACTCCT (example of a seed)(extended)
ATTGTGCATCTGACTCCT (example of an acceptable match)(extended)
```

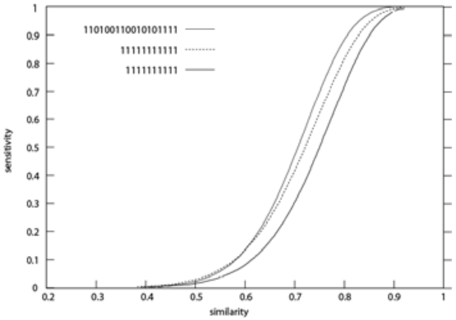


Fig. 5.16
Page 163

MegaBLAST at NCBI

--very fast
--uses very large word sizes (e.g. W=28)
--use it to align long, closely related sequences

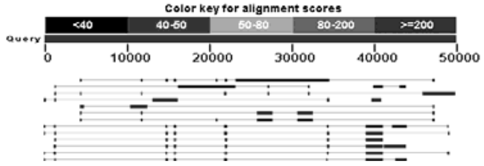
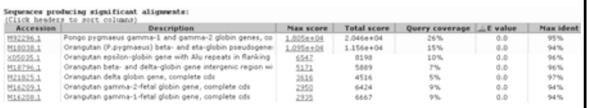


Fig. 5.19
Page 167

MegaBLAST output

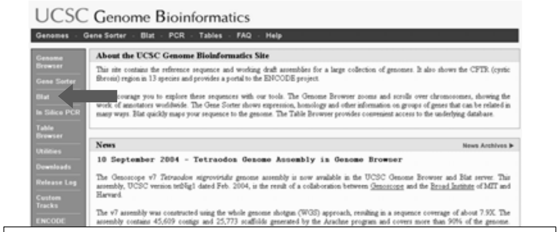


Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
M92296.1	Pongo pygmaeus gamma-1 and gamma-2 globin genes, co	1.805e+04	2.046e+04	26%	0.0	95%
M18038.1	Orangutan (P.pygmaeus) beta- and eta-globin pseudogene:	1.095e+04	1.156e+04	15%	0.0	94%
X05035.1	Orangutan epsilon-globin gene with Alu repeats in flanking	6547	8198	10%	0.0	96%
M18796.1	Orangutan beta- and delta-globin gene intergenic region wi	5171	5889	7%	0.0	96%
M21825.1	Orangutan delta globin gene, complete cds	3616	4516	5%	0.0	97%
M16209.1	Orangutan gamma-2-fetal globin gene, complete cds	2950	6424	9%	0.0	94%
M16208.1	Orangutan gamma-1-fetal globin gene, complete cds	2935	6667	9%	0.0	94%

Fig. 5.19
Page 167

To access BLAT, visit http://genome.ucsc.edu



About the UCSC Genome Bioinformatics Site

This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also shows the CFTR (cystic fibrosis) region in 13 species and provides a portal to the ENCODE project.

...courage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database.

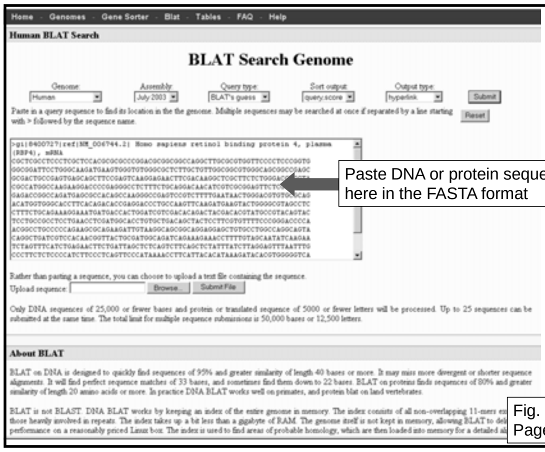
News

10 September 2004 - Tetraodon Genome Assembly in Genome Browser

The Genoscope v7 *Tetraodon nigroviridis* genome assembly is now available in the UCSC Genome Browser and Blat server. This assembly, UCSC version tetNig1 dated Feb. 2004, is the result of a collaboration between Genoscope and the Broad Institute of MIT and Harvard.

The v7 assembly was constructed using the whole genome shotgun (WGS) approach, resulting in a sequence coverage of about 7.9X. The assembly contains 45,609 contigs and 25,773 scaffolds generated by the Arachne program and covers more than 90% of the genome.

"BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 20 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein blat on land vertebrates." --BLAT website



Human BLAT Search

BLAT Search Genome

Genome: Human
Assembly: July 2003
Query type: BLAT's guess
Sort output: query,score
Output type: hyperlink

Paste in a query sequence to find its location in the the genome. Multiple sequences may be searched at once if separated by a line starting with > followed by the sequence name.

```
>gi|8400727|ref|NM_006744.2| Homo sapiens retinol binding protein 4, plasma
(RBP4), mRNA
CGCTCGCCTCCCTCGCTCCACGCGCGCCCGGACGCGGCGGCCAGGCTTGCGCGTGGTTCCCCTCCCGGTG
GGCGGATTCCTGGGCAAGATGAAGTGGGTGTGGGCGCTCTTGCTGTTGGCGGCGTGGGCAGCGGCCGAGC
GCGACTGCCGAGTGAGCAGCTTCCGAGTCAAGGAGAACTTCGACAAGGCTCGCTTCTCTGGGAC
CGCCATGGCCAAGAAGGACCCCGAGGGCCTCTTTCTGCAGGACAACATCGTCGCGGAGTTCTC
GAGACCGGCCAGATGAGCGCCACAGCCAAGGGCCGAGTCCGTCTTTTGAATAACTGGGACGTGTG
ACATGGTGGGCACCTTCACAGACACCGAGGACCCTGCCAAGTTCAAGATGAAGTACTGGGGCGTAGCCTC
CTTTCTGCAGAAAGGAAATGATGACCACTGGATCGTCGACACAGACTACGACACGTATGCCGTACAGTAC
TCCTGCCGCCTCCTGAACCTCGATGGCACCTGTGCTGACAGCTACTCCTTCGTGTTTTCCCGGGACCCCA
ACGGCCTGCCCCCAGAAGCGCAGAAGATTGTAAGGCAGCGGCAGGAGGAGCTGTGCCTGGCCAGGCAGTA
CAGGCTGATCGTCCACAACGGTTACTGCGATGGCAGATCAGAAAGAAACCTTTTGTAGCAATATCAAGAA
TCTAGTTTCATCTGAGAACTTCTGATTAGCTCTCAGTCTTCAGCTCTATTTATCTTAGGAGTTTAATTTG
CCCTTCTCTCCCCATCTTCCCTCAGTTCCCATAAAACCTTCATTACACATAAAGATACACGTGGGGGTCA
```

Paste DNA or protein sequence here in the FASTA format

Rather than pasting a sequence, you can choose to upload a text file containing the sequence.
Upload sequence:

Only DNA sequences of 25,000 or fewer bases and protein or translated sequence of 5000 or fewer letters will be processed. Up to 25 sequences can be submitted at the same time. The total limit for multiple sequence submissions is 50,000 bases or 12,500 letters.

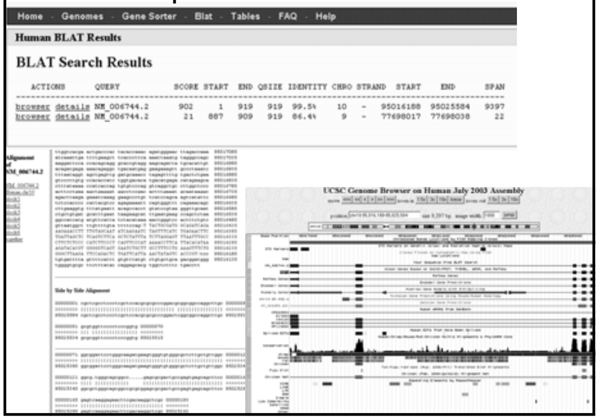
About BLAT

BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 22 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein blat on land vertebrates.

BLAT is not BLAST. DNA BLAT works by keeping an index of the entire genome in memory. The index consists of all non-overlapping 11-mers ex those heavily involved in repeats. The index takes up a bit less than a gigabyte of RAM. The genome itself is not kept in memory, allowing BLAT to del performance on a reasonably priced Linux box. The index is used to find areas of probable homology, which are then loaded into memory for a detailed al

Fig. 5.20
Page 167

BLAT output includes browser and other formats

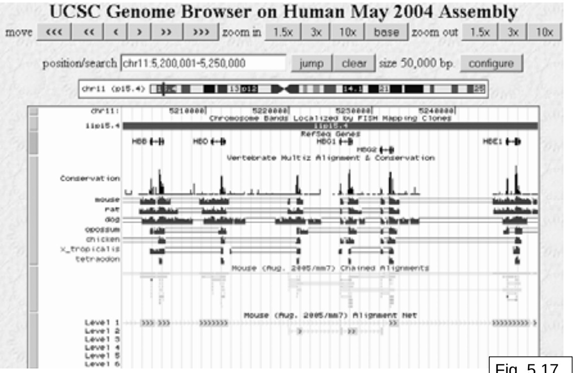

Home - Genomes - Gene Sorter - Blat - Tables - FAQ - Help
Alignment of NM_006744.2
Side by Side Alignment
UCSC Genome Browser on Human July 2003 Assembly

Human BLAT Results

BLAT Search Results

```
   ACTIONS      QUERY           SCORE START  END QSIZE IDENTITY CHRO STRAND  START     END      SPAN
---------------------------------------------------------------------------------------------------
browser details NM_006744.2       902     1  919   919  99.5%    10   -    95016188 95025584   9397
browser details NM_006744.2        21   887  909   919  86.4%     9   -    77698017 77698038     22
```

Blastz

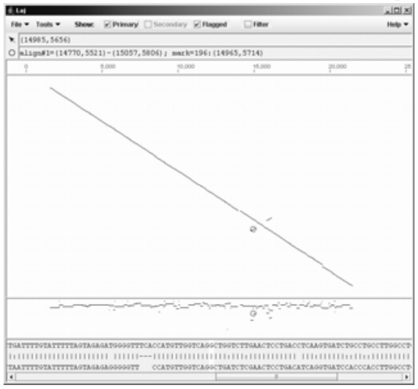


UCSC Genome Browser on Human May 2004 Assembly

position/search chr11:5,200,001-5,250,000 size 50,000 bp.

Fig. 5.17
Page 165

Blastz (laj software): human versus rhesus duplication



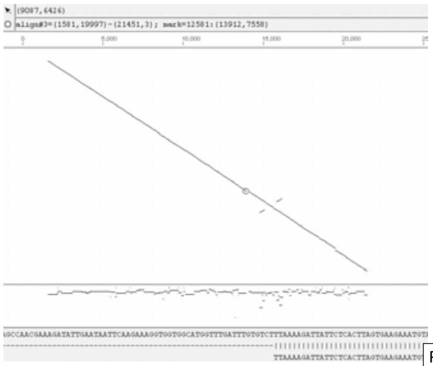
(14985,5656)

align#1=(14770,5521)-(15057,5806); mark=196:(14965,5714)

```
TGATTTTGTATTTTTAGTAGAGATGGGGTTTCACCATGTTGGTCAGGCTGGTCTTGAACTCCTGACCTCAAGTGATCTGCCTGCCTTGGCCT
TAATTTTGTATTTTTAGTAGAGAGGGGGTT   CCATGTTGGTCAGGCTGATCTCGAACTCCTGACATCAGGTGATCCACCCACCTTGGCCT
```

Fig. 5.18
Page 166

Blastz (laj software): human versus rhesus gap



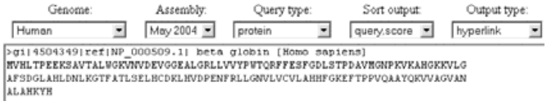
(9087,6426)

align#3=(1581,19997)-(21451,3); mark=12581:(13912,7558)

```
AGCCAACGAAAGATATTGAATAATTCAAGAAAGGTGGTGGCATGGTTTGATTTGTGTCTTTAAAAGATTATTCTCACTTAGTGAAGAAATGTA
                                                              TTAAAAGATTATTCTCACTTAGTGAAGAAATGT
```

Fig. 5.18
Page 166

BLAT

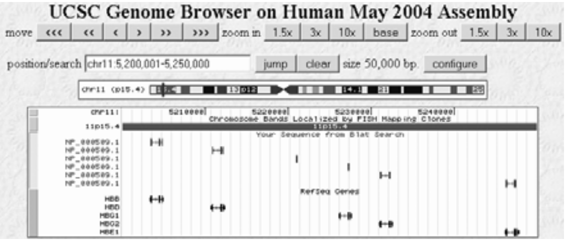


Genome: Human
Assembly: May 2004
Query type: protein
Sort output: query.score
Output type: hyperlink

```
>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

Fig. 5.20
Page 167

BLAT



UCSC Genome Browser on Human May 2004 Assembly

position/search chr11:5,200,001-5,250,000 size 50,000 bp.

Fig. 5.20
Page 167

LAGAN

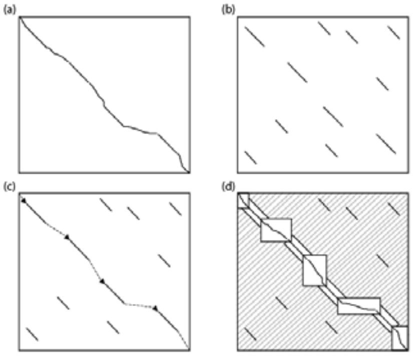
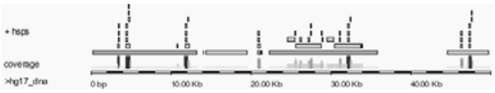


Fig. 5.21
Page 168

SSAHA



Outline of today's lecture

BLAST
- Practical use
- Algorithm
- Strategies

Finding distantly related proteins:
- PSI-BLAST
- Hidden Markov models

BLAST-like tools for genomic DNA
- PatternHunter
- Megablast
- BLAT, BLASTZ

Where we are in the course

--We started with “access to information” (Chapter 2)

--We next covered pairwise alignment (Chapter 3), then BLAST in which one sequence is compared to a database (Chapters 4, 5)

--Next we'll describe multiple sequence alignment (Chapter 6)

--We'll then visualize multiple sequence alignments as phylogenetic trees (Chapter 7). That topic spans molecular evolution.