CAP 5510/CGS 5166: Bioinformatics & Bioinformatic Tools

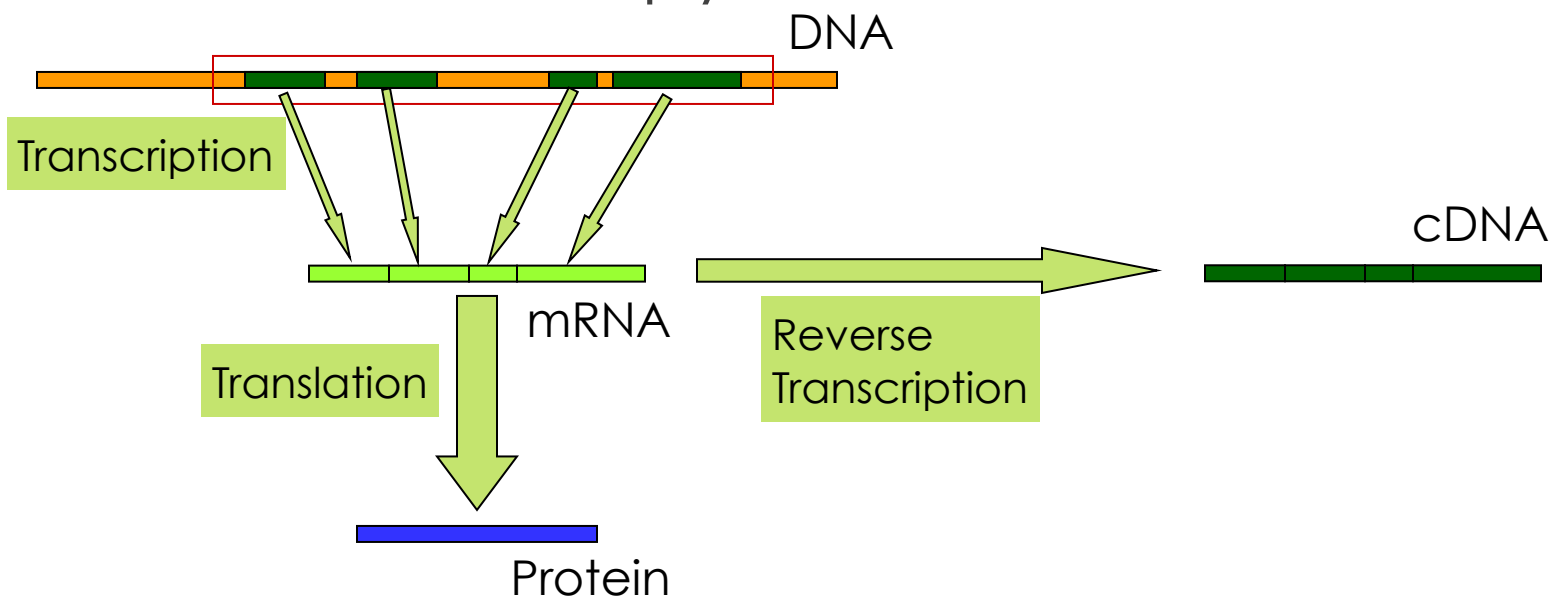
GIRI NARASIMHAN, SCIS, FIU

Sequence Alignment

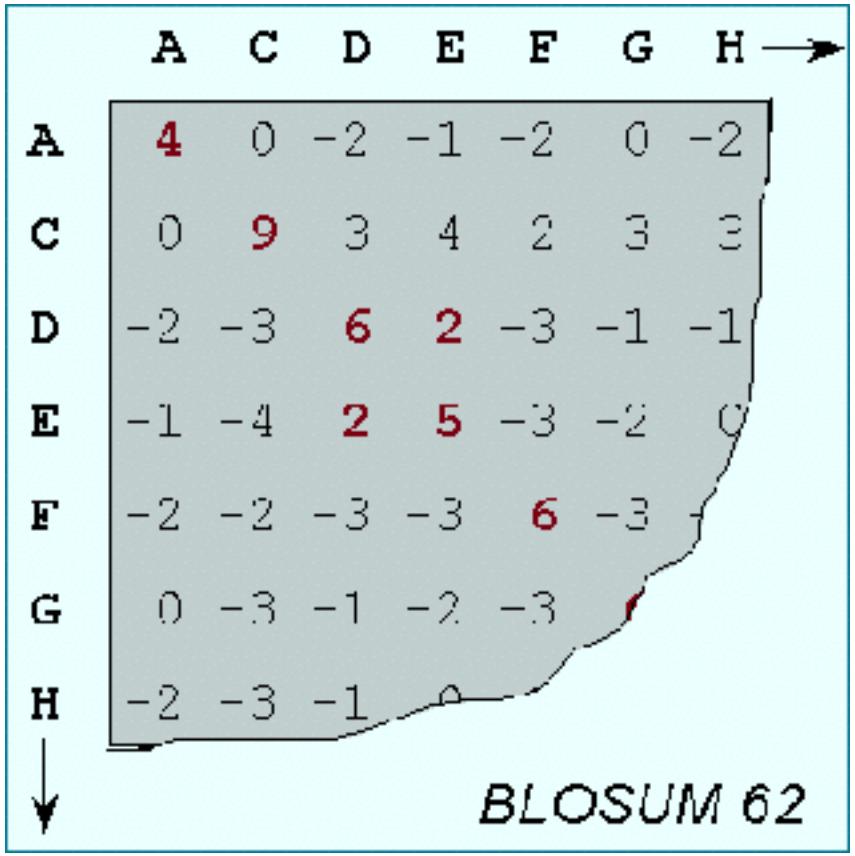
- Global:
  - Needleman-Wunsch-Sellers (1970).
- Local:
  - Smith-Waterman (1981)
  - Useful when commonality is small and global alignment is meaningless. Often unaligned portions "mask" short stretches of aligned portions. Example: comparing long stretches of anonymous DNA; aligning proteins that share only some motifs or domains.
- Dynamic Programming (DP) based.

Why gaps?

- Example: Finding the gene site for a given (eukaryotic) cDNA requires "gaps".
- What is cDNA? cDNA = Copy DNA



How to score mismatches?



BLAST & FASTA

- FASTA

  [Lipman, Pearson '85, '88]

- Basic Local Alignment Search Tool

  [Altschul, Gish, Miller, Myers, Lipman '90]

BLAST Overview

- Program(s) to search all sequence databases
- Tremendous Speed/Less Sensitive
- Statistical Significance reported
- WWWBLAST, QBLAST (send now, retrieve results later), Standalone BLAST, BLASTcl3 (Client version, TCP/IP connection to NCBI server), BLAST URLAPI (to access QBLAST, no local client)

BLAST

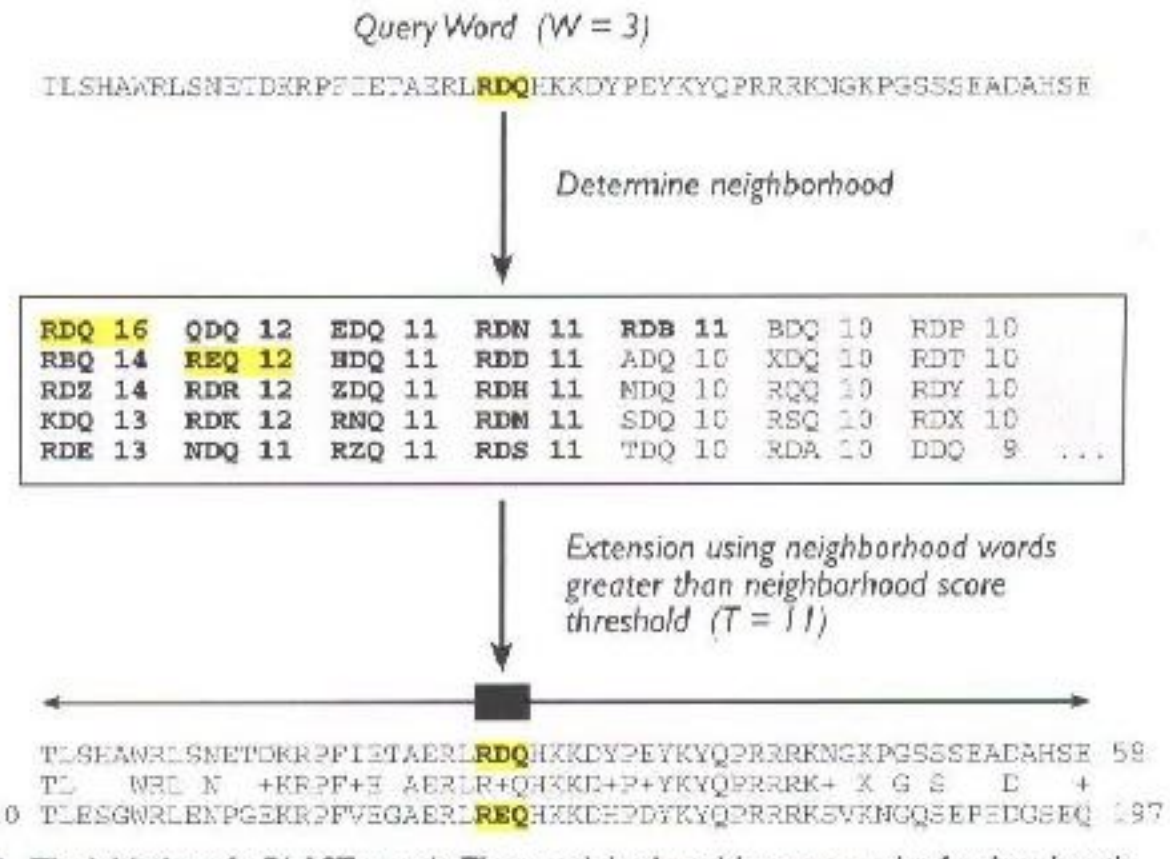


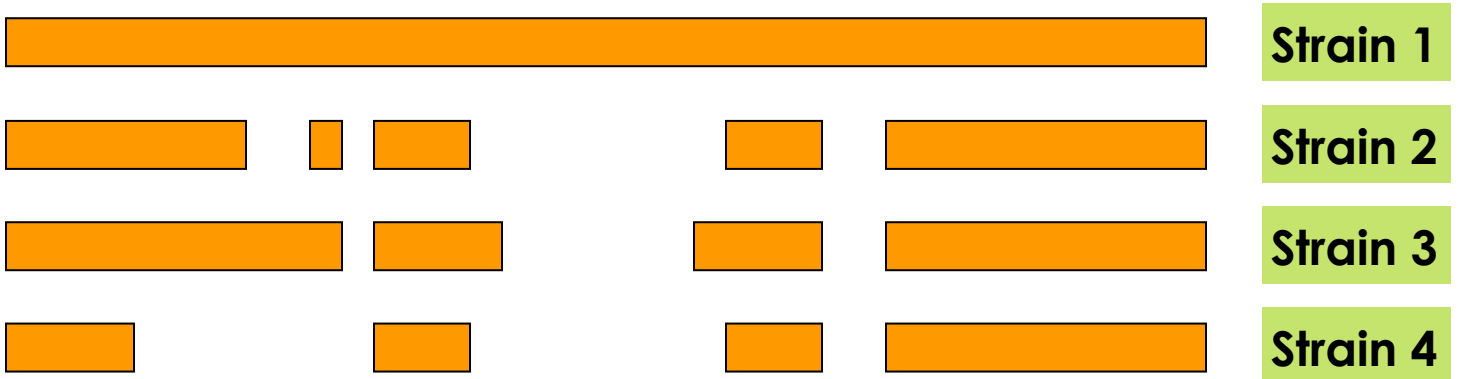
FIGURE 11.7 The initiation of a BLAST search. The search begins with query words of a given length (here, three amino acids) being compared against a scoring matrix to determine additional three-letter words "in the neighborhood" of the original query word. Any occurrences of these neighborhood words in sequences within the target database then are investigated. See text for details.

BLAST Strategy & Improvements

- Lipman et al.: speeded up finding "runs" of "hot spots".
- Eugene Myers '94: "Sublinear algorithm for approximate keyword matching".
- Karlin, Altschul, Dembo '90, '91: "Statistical Significance of Matches"

Why Gaps?

- Example: Aligning HIV sequences.



BLAST Variants

- **Nucleotide BLAST**
  - **Standard blastn**
  - **MEGABLAST** (Compare large sets, Near-exact searches)
  - **Short Sequences** (higher E-value threshold, smaller word size, no low-complexity filtering)
- **Protein BLAST**
  - **Standard blastp**
  - **PSI-BLAST** (Position Specific Iterated BLAST)
  - **PHI-BLAST** (Pattern Hit Initiated BLAST; reg expr. Or Motif search)
  - **Short Sequences** (higher E-value threshold, smaller word size, no low-complexity filtering, PAM-30)
- **Translating BLAST**
  - **Blastx**: Search nucleotide sequence in protein database (6 reading frames)
  - **Tblastn**: Search protein sequence in nucleotide dB
  - **Tblastx**: Search nucleotide seq (6 frames) in nucleotide DB (6 frames)

BLAST Cont'd

- **RPS BLAST**
  - Compare protein sequence against Conserved Domain DB; Helps in predicting rough structure and function
- **Pairwise BLAST**
  - blastp (2 Proteins), blastn (2 nucleotides), tblastn (protein-nucleotide w/ 6 frames), blastx (nucleotide-protein), tblastx (nucleotide w/6 frames-nucleotide w/ 6 frames)
- **Specialized BLAST**
  - Human & Other finished/unfinished genomes
  - *P. falciparum*: Search ESTs, STSs, GSSs, HTGs
  - VecScreen: screen for contamination while sequencing
  - IgBLAST: Immunoglobin sequence database

BLAST Credits

- Stephen Altschul
- Jonathan Epstein
- David Lipman
- Tom Madden
- Scott McGinnis
- Jim Ostell
- Alex Schaffer
- Sergei Shavirin
- Heidi Sofia
- Jinghui Zhang

Databases used by BLAST

- **Protein**
  - nr (everything), swissprot, pdb, alu, individual genomes
- **Nucleotide**
  - nr, dbest, dbsts, htgs (unfinished genomic sequences), gss, pdb, vector, mito, alu, epd
- **Misc**

BLAST Parameters and Output

- Type of sequence, nucleotide/ protein
- Word size, **w**
- Gap penalties, p_1 and p_2
- Neighborhood Threshold Score, **T**
- Database to search, **D**
- Scoring Matrix, **M**
- Score Threshold, **S**
- E-value Cutoff, **E**
- Number of hits to display, **H**
- Score **s** and E-value **e**
  - E-value **e** is the expected number of sequences that would have an alignment score greater than the current score **s**.

Scoring Matrix to Use

- PAM 40 Short alignments with high similarity (70-90%)
- PAM 160 Members of a protein family (50-60%)
- PAM 250 Longer alignments (divergent sequences) (~30%)

- BLOSUM90 Short alignments with high similarity (70-90%)
- BLOSUM80 Members of a protein family (50-60%)
- BLOSUM62 Finding all potential hits (30-40%)
- BLOSUM30 Longer alignments (divergent sequences) (<30%)

Main Ideas in BLAST

- Break sequence into words and look for words in database
- Find hotspots where many words find hits and look more closely
- Instead of looking for **approximate** hits of words …
- … find **exact** hits of nearby words

BLAST algorithm: Phase 1

Phase 1: For each word in query, get words (w=3) within threshold T

Example: for a query sequence

…FS**GTW**YA…

Consider a word **GTW** in the query

Get list of words (w=3) close to GTW:

ATW, GSW, …

Fig. 4.11
page 116

Use BLOSUM to score word hits

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

Phase 1: Find list of similar words

- Find list of words of length w (here w = 3) and distance at least T (here T = 11)
  - GTW 22
  - GSW 18
  - ATW 16
  - NTW 16
  - GTY 13
  - GNW 10
  - GAW 9

BLAST: Phases 2 & 3

- Phase 2: Scan database for hits and find HotSpots
- Phase 3:
  - Extend good 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)
```

BLAST: Threshold vs # Hits & Extensions

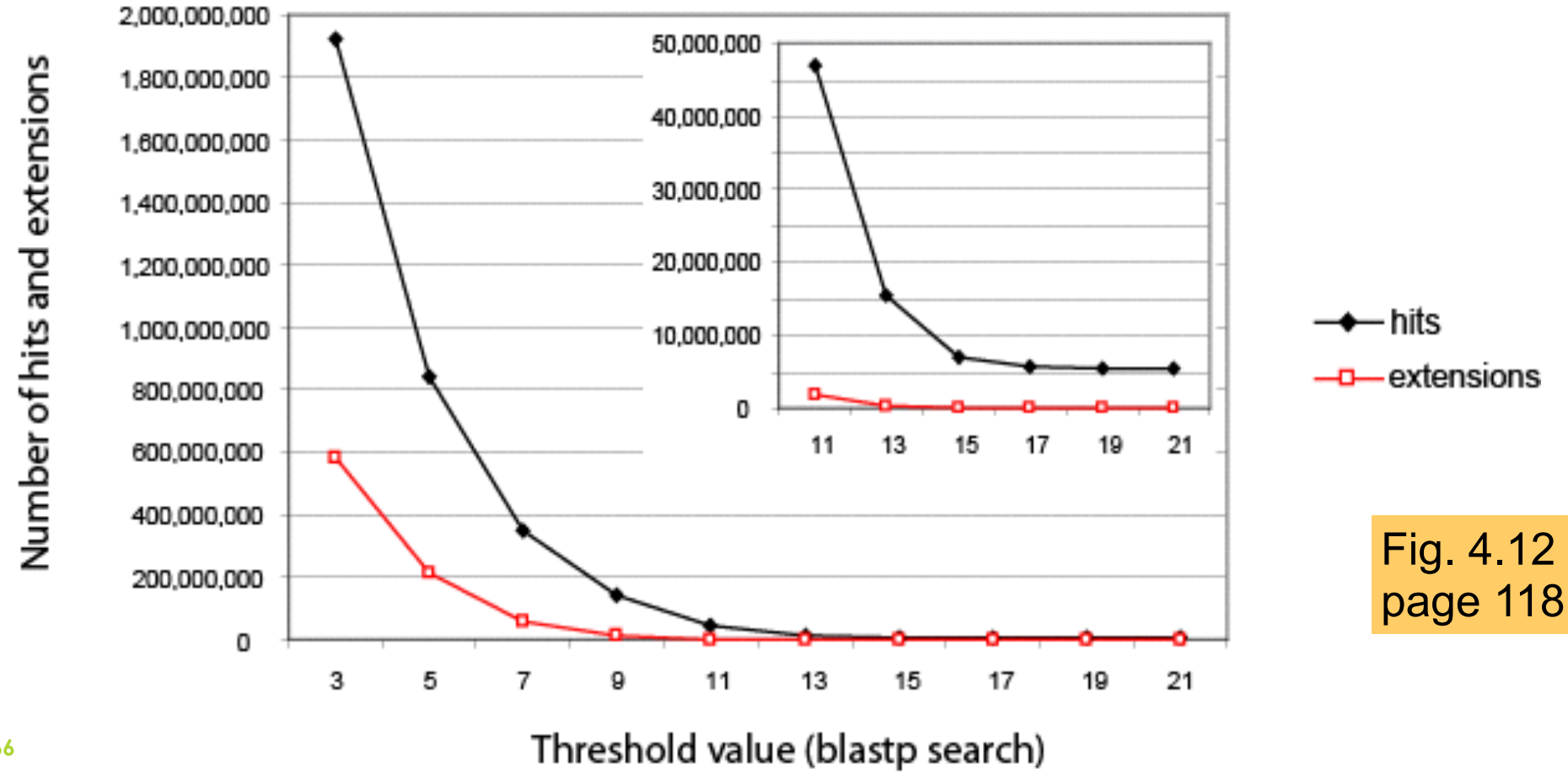
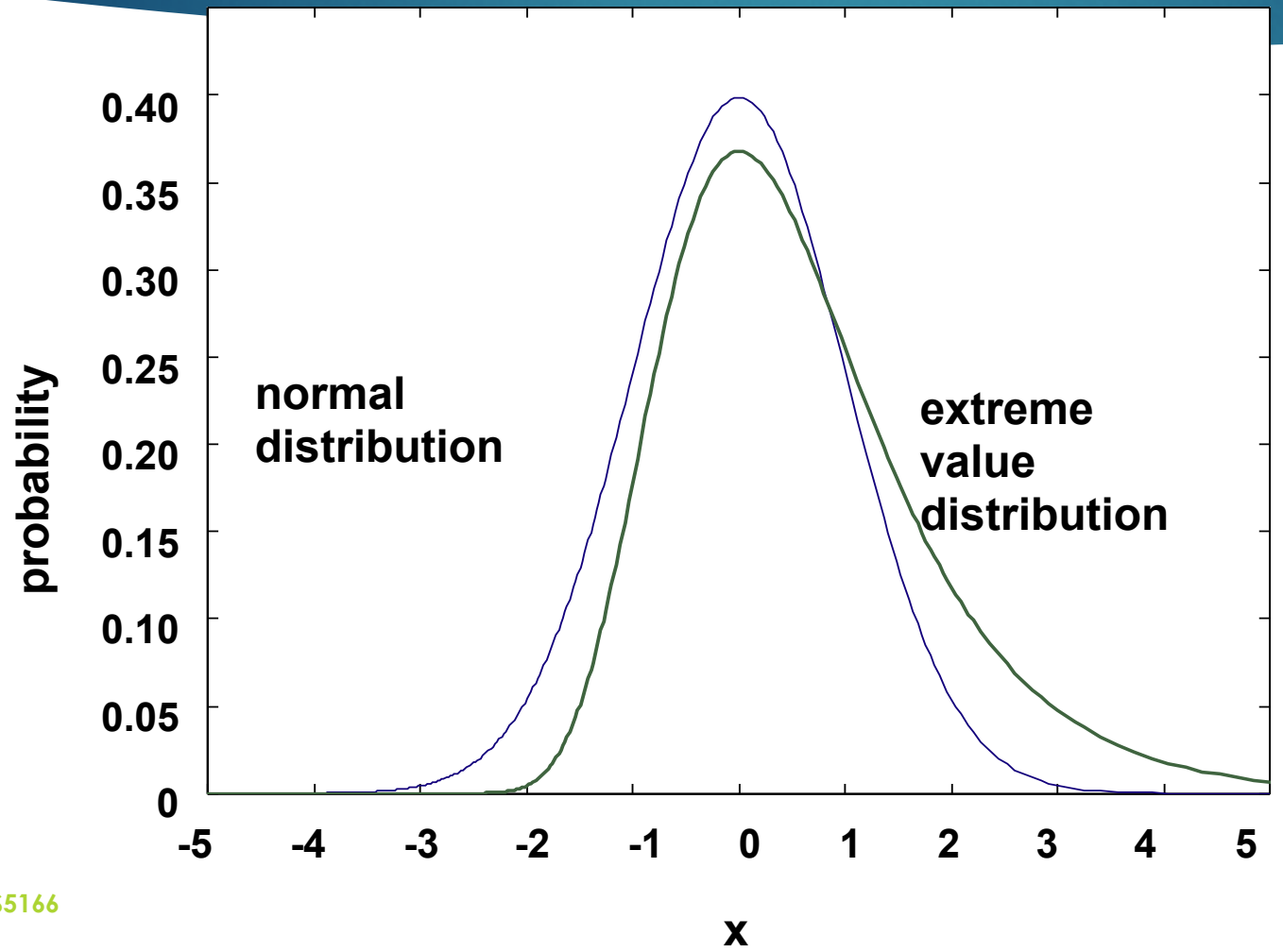


Fig. 4.12
page 118

Word Size

- Blastn: w = 7, 11, or 15.
  - w=15 gives fewer matches and is faster than w=11 or w=7.
- Megablast: w = 28 to 64.
  - Megablast is VERY fast for finding closely related DNA sequences!

Scores: Follow Extreme Value Distribution



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

m,n = seq length
S = Raw Score
K ≈ Search space

$S' = (\lambda S - \ln K) / \ln 2$
S' = Bit Score

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

E-value versus P-value

E-value	P-value
10	0.9999546
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258
0.05	0.04877058
0.001	0.00099950
0.0001	0.0001

E-values are easier to interpret;

If query is short aa sequence, then use very large E-value;

Sometimes even meaningful hits have large E-values.

BLAST: Steps

- Choose your sequence
- Choose your tool
- Choose your database
- Select parameters, if needed
- Interpret your results

BLAST report header

NCBI results of BLAST

BLASTP 2.2.1 [Apr-13-2001]

Reference:
Altschul,Stephen F., Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.

RID: 1009580302-26840-4362

Query= RAB protein
(656 letters)

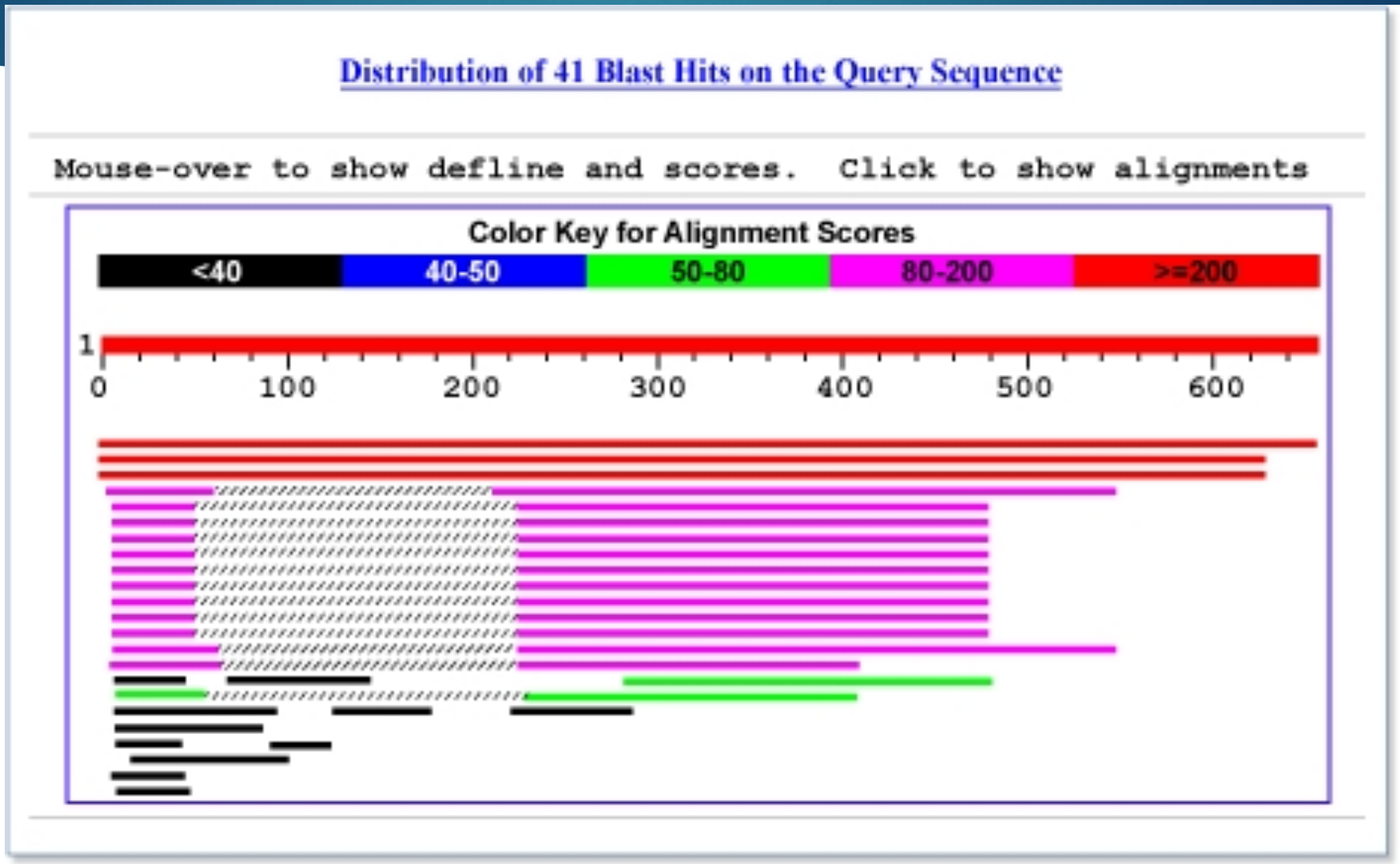
Database: Non-redundant SwissProt sequences
102,387 sequences; 37,391,913 total letters

If you have any problems or questions with the results of this search
please refer to the **BLAST FAQs**

Taxonomy reports

NCBI Handbook, Eds. McEntyre, Ostell

Graphical Overview of BLAST Results



List of hits with one line descriptions

Sequences producing significant alignments:		Score (bits)	E Value
(a)	(b)	(c)	(d)
gi|116365|sp|P26374|RAE2_HUMAN	Rab proteins geranylgeranylt...	1216	0.0
gi|21431807|sp|P24386|RAE1_HUMAN	Rab proteins geranylgerany...	879	0.0
gi|585775|sp|P37727|RAE1_RAT	Rab proteins geranylgeranyltra...	846	0.0
gi|13626886|sp|Q61598|GDIC_MOUSE	RAB GDP dissociation inhib...	127	5e-29
gi|729566|sp|P39958|GDI1_YEAST	SECRETORY PATHWAY GDP DISSOC...	127	5e-29
gi|13626813|sp|O97556|GDIB_CANFA	Rab GDP dissociation inhib...	126	1e-28
gi|13638229|sp|P50397|GDIB_MOUSE	RAB GDP dissociation inhib...	125	3e-28
gi|1707888|sp|P50398|GDIA_RAT	RAB GDP dissociation inhibito...	124	7e-28
gi|121108|sp|P21856|GDIA_BOVIN	Rab GDP dissociation inhibit...	124	7e-28
gi|21903424|sp|P50396|GDIA_MOUSE	Rab GDP dissociation inhib...	124	7e-28
gi|13626812|sp|O97555|GDIA_CANFA	RAB GDP dissociation inhib...	124	8e-28
gi|1707886|sp|P31150|GDIA_HUMAN	Rab GDP dissociation inhibi...	123	9e-28
gi|13638228|sp|P50395|GDIB_HUMAN	Rab GDP dissociation inhib...	122	2e-27
gi|1707891|sp|P50399|GDIB_RAT	RAB GDP DISSOCIATION INHIBITO...	121	5e-27
gi|1723467|sp|Q10305|YD4C_SCHPO	Putative secretory pathway ...	120	8e-27
gi|585776|sp|P32864|RAEP_YEAST	RAB proteins geranylgeranylt...	97	7e-20
gi|10720243|sp|O93831|RAEP_CANAL	RAB proteins geranylgerany...	74	9e-13
gi|2498411|sp|Q49398|GLF_MYCGE	UDP-galactopyranose mutase	35	0.63
gi|11135401|sp|Q9XBQ9|STHA_AZOVI	Soluble pyridine nucleotid...	34	1.0
gi|11135075|sp|O05139|STHA_PSEFL	Soluble pyridine nucleotid...	33	1.3
gi|11135195|sp|P57112|STHA_PSEAE	Soluble pyridine nucleotid...	33	1.8
gi|22257022|sp|Q8TZJ8|RLA0_PYRFU	Acidic ribosomal protein P...	33	2.1
gi|3915516|sp|P94488|YNAJ_BACSU	Hypothetical symporter ynaJ	32	3.4
gi|231788|sp|P30599|CHS2_USTMA	CHITIN SYNTHASE 2 (CHITIN-UD...	32	3.7
gi|2498412|sp|P75499|GLF_MYCPN	UDP-galactopyranose mutase	32	4.2
gi|547891|sp|P36225|MAP4_BOVIN	Microtubule-associated prote...	32	4.2
gi|586602|sp|P37747|GLF_ECOLI	UDP-galactopyranose mutase	32	4.6

List of alignments

>gi|116365|sp|P26374|RAB2_HUMAN Rab proteins geranylgeranyltransferase component A 2 (Rab escort protein 2) (REP-2) (Choroideremia-like protein)
Length = 655

Score = 846 bits (2186), Expect = 0.0
Identities = 432/632 (68%), Positives = 489/632 (77%), Gaps = 13/632 (2%)

Query: 1 MADNLPTEFDVVIIGTGLPESILAAACSRSGQRVLHIDSRSYYGGNWASFSFSGLLSWLK 60
MADNLP++FDV++IGTGLPESI+AAACSRSGQRVLH+DSRSYYGGNWASFSFSGLLSWLK
Sbjct: 1 MADNLPSDFDVIVIGTGLPESILAAACSRSGQRVLHVDSRSYYGGNWASFSFSGLLSWLK 60

Query: 61 EYQQNNDTGRRSTVVWQDLTHFTFRATTLRKKDRTTQHTRAFPYASQDMFDNVFRTGALQ 120
EYQ+NND+ E++ +WQ+ I E EEAI L KD+TIQH E F YASQD+ +VEE GALQ
Sbjct: 61 EYQENNDVVTENS-MWQEQILENEEAIPLSSKDKTIQHVEVFCYASQDLHKDVEEAGALQ 119

Query: 121 KNPSLGVS----NTFTEVLDSALPEESQLSYFNSDEMPAKHTQKSDFEISLEVTDVEEV 176
KN + S S LP + S E+PA+ +Q E S EV D I +
Sbjct: 120 KNHASVTSAQSAEAAEAARTSCLPTAVEPLSMGSCEIPAEQSQCPGPESSPEVNDASATG 179

Query: 177 EKEKYCSDKTSMHTVXXXXXXXXXXXXTVEDKADEPIRNRITYSQIVKEGRRFNIDLVEK 236
+KE + V+D + P +NRITYSQI+KEGRRFNIDLV3+
Sbjct: 180 KKENSDAKSS--------TEEPSENVPKVQDNTETPKKNRITYSQIIKEGRRFNIDLVSQ 231

Query: 237 LLVSQGLLIDLLIKSDVSRYVEFKNVTRILAFREGKVEQVPCSRADVFNSKELTMVEKRM 296
LLYS+GLLIDLLIKS+VSRY EFKN+TRILAFREG VEQVPCSRADVFNSK+LTMVEKRM
Sbjct: 232 LLYSRGLLIDLLIKSNVSRYAEFKNITRILAFREGTVEQVPCSRADVFNSKQLTMVEKRM 291

Query: 297 LMKFLTFCLEYEQHPDEYQAFTQCSPSEYLKTKKLTPNLQHFVLHSIAMTSEESCTTIDC 356
LMKFLTFC+EYE+HPDEY+A+ +PSEYLKT+KLTPNLQ+FVLHSIAMTSE++ T+DG
Sbjct: 292 LMKFLTFCVEYEEHPDEYRAYEGTTPSEYLKTQKLTPNLQYFVLHSIAMTSEITSCTVDG 351

Query: 357 LNATKNFLQCLGRFGNTPFLFPLYGQGELPQGFCRMCAVFGGIYCLRHKVQCFVVDKEDG 416
L ATK FLQCLGR+GNTPFLFPLYGQGELPQ FCRMCAVFGGIYCLRH VQC VVDKED
Sbjct: 352 LKATKKFLQCLGRYGNTPFLFPLYGQGELPQCFCRMCAVFGGIYCLRHSVQCLVVDKEDR 411

Query: 417 RCKAIIDMFGQRINAKYFIVEDGYLSEEICSNVQYKQISRAVLIIDQSILKTDLDQQTSI 476
+CKA+ID FGQRI +K+FI+EDGYLSE TCS VQY+QISRAVLIID SILKTD DQQ SI
Sbjct: 412 KCKAVIDQFGQRIISKHFIIEDGYLSENTCSRVQYRQISRAVLIIDGSVLKTDADQQVSI 471

Query: 477 LIVPAEEGALAVRVIRLCSSTMTCMKGTYLVHLICMSSKTARRDLERVVQKLFTPYTST 536
L VD EDG+ VDV BLCSSTMTCMK TYLVHLIC SSKTARRDLE VV+KLFTDVTS
Sbjct: 472 LAVPAEEPGSFGVRVIRLCSSTMTCMKGTYLVHLICMSSKTARRDLERVVQKLFTPYTSI 531

Query: 537 EINEEELTKPRLLWALYFNMRDSGISRSSYNGLPSNVYVCSGPDCSLGNEHAVKQAETL 596
E E++ KPRLLWALYFNMRDSS ISR YN LPSNVYVCSGPD CLGN++AVKQAETL
Sbjct: 532 EAENEQVEKPRLLWALYFNMRDSSDISRQCYNDLPSNVYVCSGPDSSLGNDNAVKQAETL 591

Query: 597 FQKXXXXXXXXXXXXXXXXXXXUGDMQFEAP 628
FQ DGD Q E P
Sbjct: 592 FQQTCPVRDFCPAPPNPFNTVTAGDSSQQRVP 623

Pairwise alignment result of human beta globin and myoglobin

Myoglobin RefSeq

Information about this alignment: score, expect value, identities, positives, gaps…

Middle row displays identities; + sign for similar matches

```
>ref|NP_005359.1|  G  myoglobin [Homo sapiens]
 ref|NP_976311.1|  UG  myoglobin [Homo sapien
 ref|NP_976312.1|  G  myoglobin [Homo sapien
 ▶11 more sequence titles
 Length=154

 GENE ID: 4151 MB | myoglobin [Homo s ens] (Over 10 PubMed

 Score = 47.4 bits (144),  Expect = 8e-11, Method: Compos  l matrix adjust.
 Identities = 37/145 (25%), Positives = 57/145 (39%), Ga  2/145 (1%)

Query  4    LTPEEKSAVTALWGKVNVDEVG--GEALGRLLVVYPWTQRF ESFGDLSTPDAVMGNPKV  61
            L+  E   V  +WGKV  D  G   E L RL   +P T   F+ F  L + D +  +  +
Sbjct  3    LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL  62

Query  62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
            K HG  VL A    L    + +     L++ H  K  +  +    +   ++ VL
Sbjct  63   KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG  122

Query  122  EFTPPVQAAYQKVVAGVANALAHKY  146
            +F    Q A  K +      +A  Y
Sbjct  123  DFGADAQGAMNKALELFRKDMASNY  147
```

Slide: Courtesy J. Pevsner

Query = HBB; Subject = MB

Pairwise alignment result of human beta globin and myoglobin: the score is a sum of match, mismatch, gap creation, and gap extension scores


```
 Score = 18.1 bits (35),  Expect = 0.015, Method: Composition-based stats.
 Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query  12  VTALWGKVNVD--EVGGEALGRLL  33
           V  +WGKV  D    G E L RL
Sbjct  11  VLNVWGKVEADIPGHGQEVLIRLF  34
```

match	4 11 5 6 6 5 4 5 / 6 4 4	sum of matches: +60
mismatch	-1 1 0 -2 -2 -4 0 / 2 0 3 0	sum of mismatches: -13
gap open	11	sum of gap penalties: 12
gap extend	1	
		total raw score: 60 13 12 = 35

Slide: Courtesy J. Pevsner

Pairwise alignment result of human beta globin and myoglobin: the score is a sum of match, mismatch, gap creation, and gap extension scores


```
 Score = 18.1 bits (35),  Expect = 0.015, Method: Composition-based stats.
 Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query  12  VTALWGKVNVD--EVGGEALGRLL  33
           V  +WGKV  D    G E L RL
Sbjct  11  VLNVWGKVEADIPGHGQEVLIRLF  34
```

match	4 11 5 6 6 5 4 5 6 4 4	sum of matches: +60
mismatch	-1 1 0 -2 -2 -4 0 2 0 3 0	sum of mismatches: -13
gap open	11	sum of gap penalties: 12
gap extend	1	
		total raw score: 60 13 12 = 35

V matching V earns +4
T matching L earns -1

These scores come from a "scoring matrix"!

Slide: Courtesy J. Pevsner

Bit Score

- If S is the (raw) score for a local alignment, the **normalized** score S' (in bits) is given by

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

The parameters K and λ depend on the scoring system.

Expect value or E-value

- E-value is **not a probability**, but describes strength of random background noise.
- E-value describes number of hits one can "expect" to see by chance when searching a database of a particular size.
- It decreases exponentially with the score (S).
- **E-value = 1** means "in a database of current size, one might expect to see one match with a similar score simply by chance. Lower E-value mean more "significant" match.
- WARNING: Short sequences can be virtually identical and have relatively high E-values.
  - Calculation of E-value takes into account length of query sequence. Since shorter sequences have a high probability of occurring in the database purely by chance, E-values can be high.

BLAST Tutorial

- http://www.ncbi.nlm.nih.gov/books/NBK21097/#A614

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.

Difficulties with BLAST

- 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.
- 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.

Rules of Thumb

- Most sequences with significant similarity over their entire lengths are homologous.
- Matches that are > 50% identical in a 20-40 aa region occur frequently by chance.
- Distantly related homologs may lack significant similarity. Homologous sequences may have few absolutely conserved residues.
- A homologous to B & B to C $\Rightarrow$ A homologous to C.
- Low complexity regions, transmembrane regions and coiled-coil regions frequently display significant similarity without homology.
- Greater evolutionary distance implies that length of a local alignment required to achieve a statistically significant score also increases.

Rules of Thumb

- Results of searches using different scoring systems may be compared directly using normalized scores.
- If S is the (raw) score for a local alignment, the normalized score S' (in bits) is given by

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

The parameters depend on the scoring system.

- Statistically significant normalized score,

$$S > \log\left(\frac{N}{E}\right)$$

where E-value = E, and N = size of search space.