CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS13.html

Sequence Alignment

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)

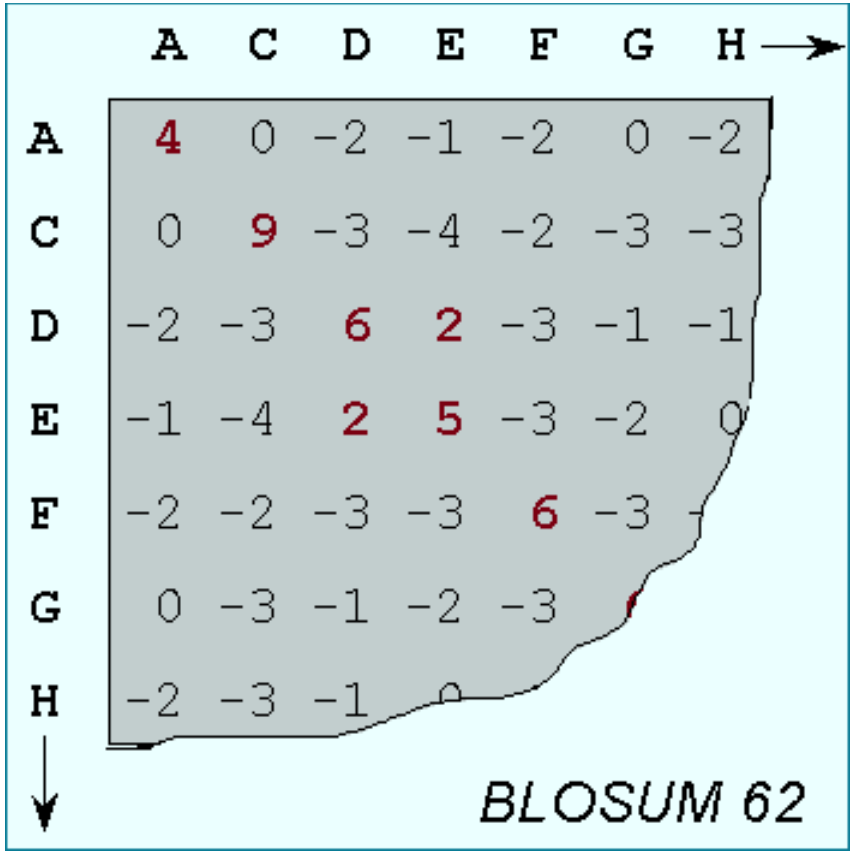
BLAST Cont'd

- **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 Parameters and Output

- Type of sequence, nucleotide/protein
- Word size, w
- Gap penalties, p_1 and p_2
- Neighborhood Threshold Score, T
- Score Threshold, S
- E-value Cutoff, E
- Number of hits to display, H
- Database to search, D
- Scoring Matrix, M
- 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.

How to score mismatches?



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%)

BLAST algorithm: Phase 1

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

Example: for a human RBP query

…FS**GTW**YA…

GTW is a word in this query sequence

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

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**

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 exact hits of similar words list 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

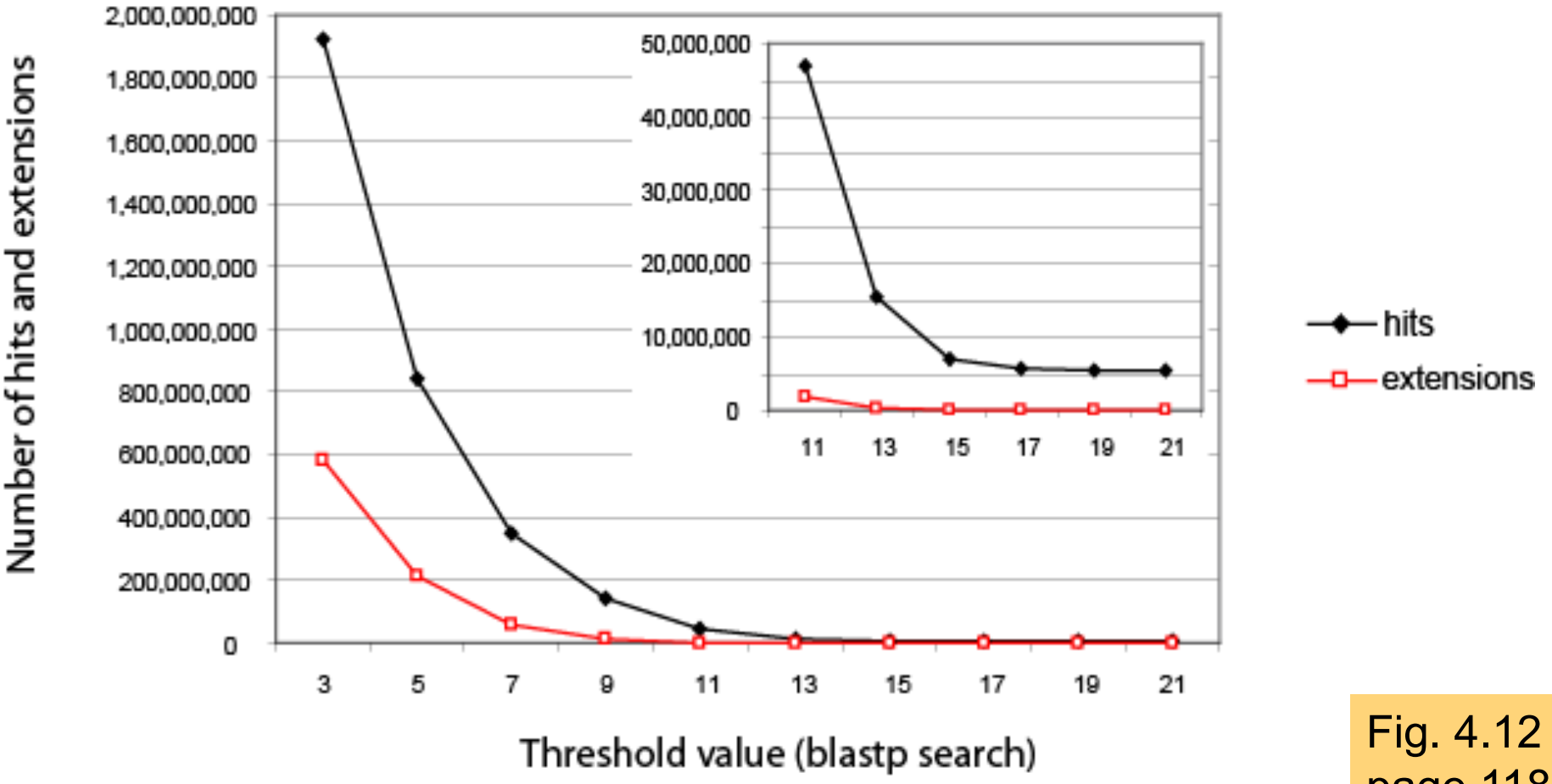
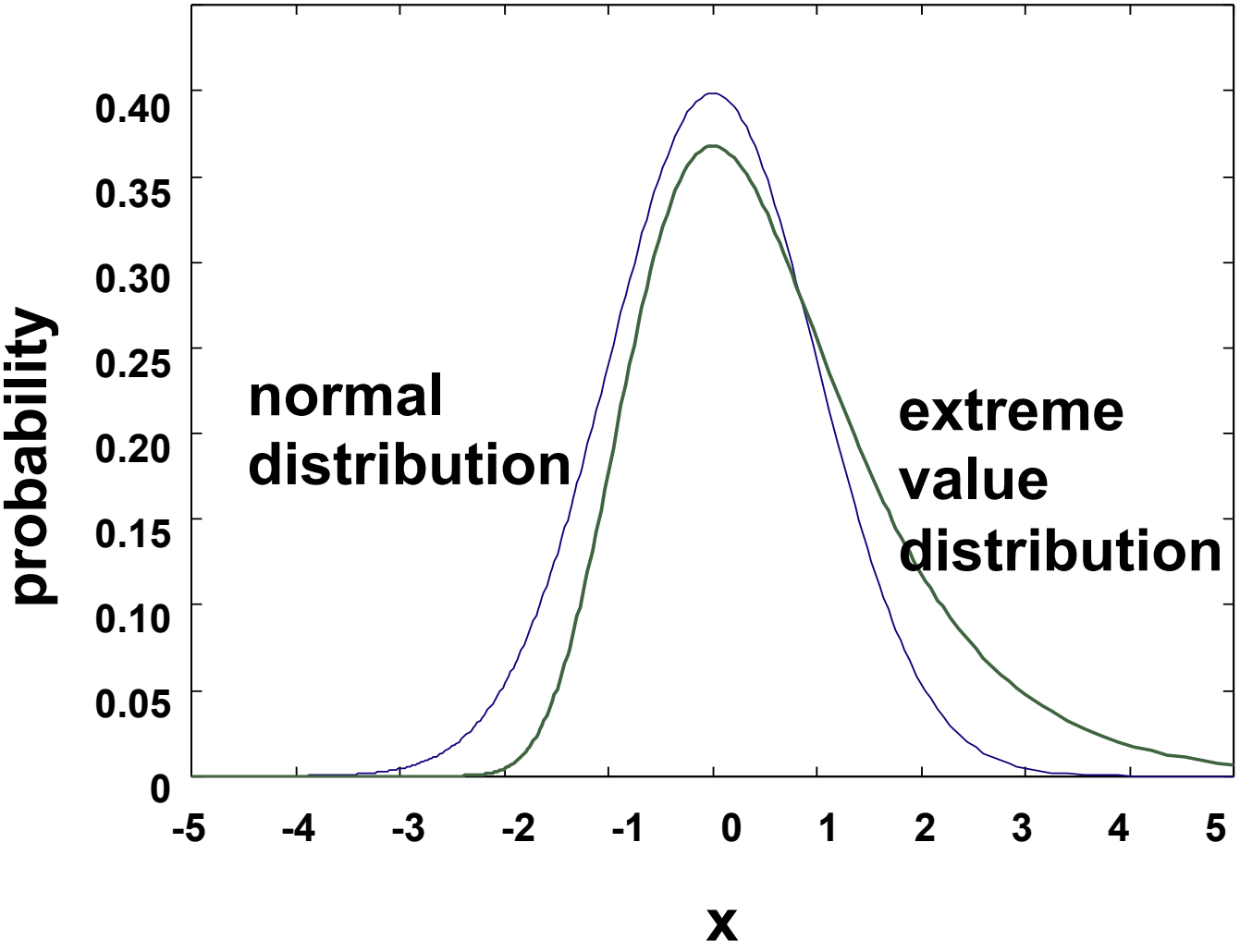


Fig. 4.12
page 118

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

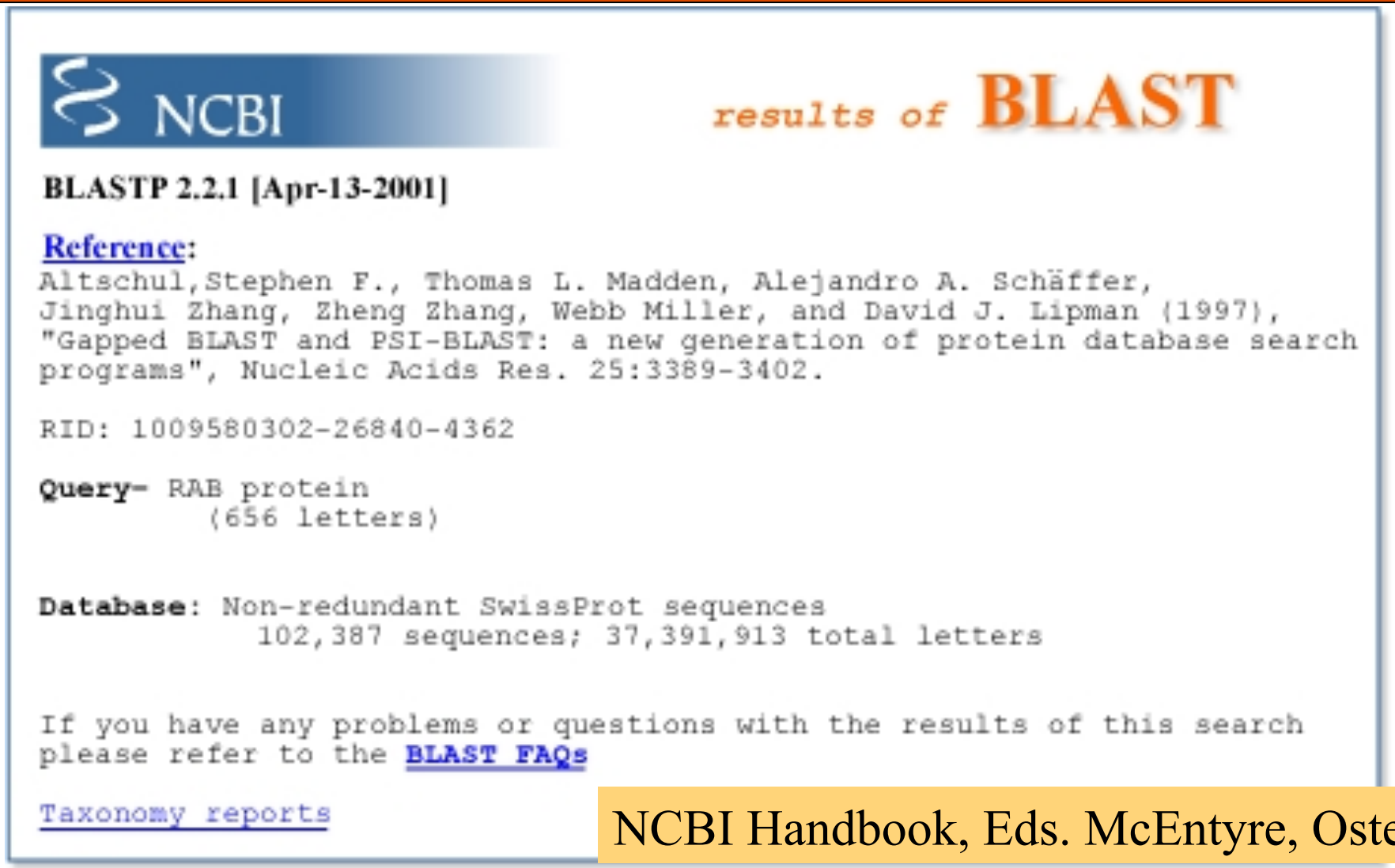
E-value	P-value
10	0.9999546
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258
0.05	0.04877058
0.001	0.00099950
0.0001	0.0001

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)

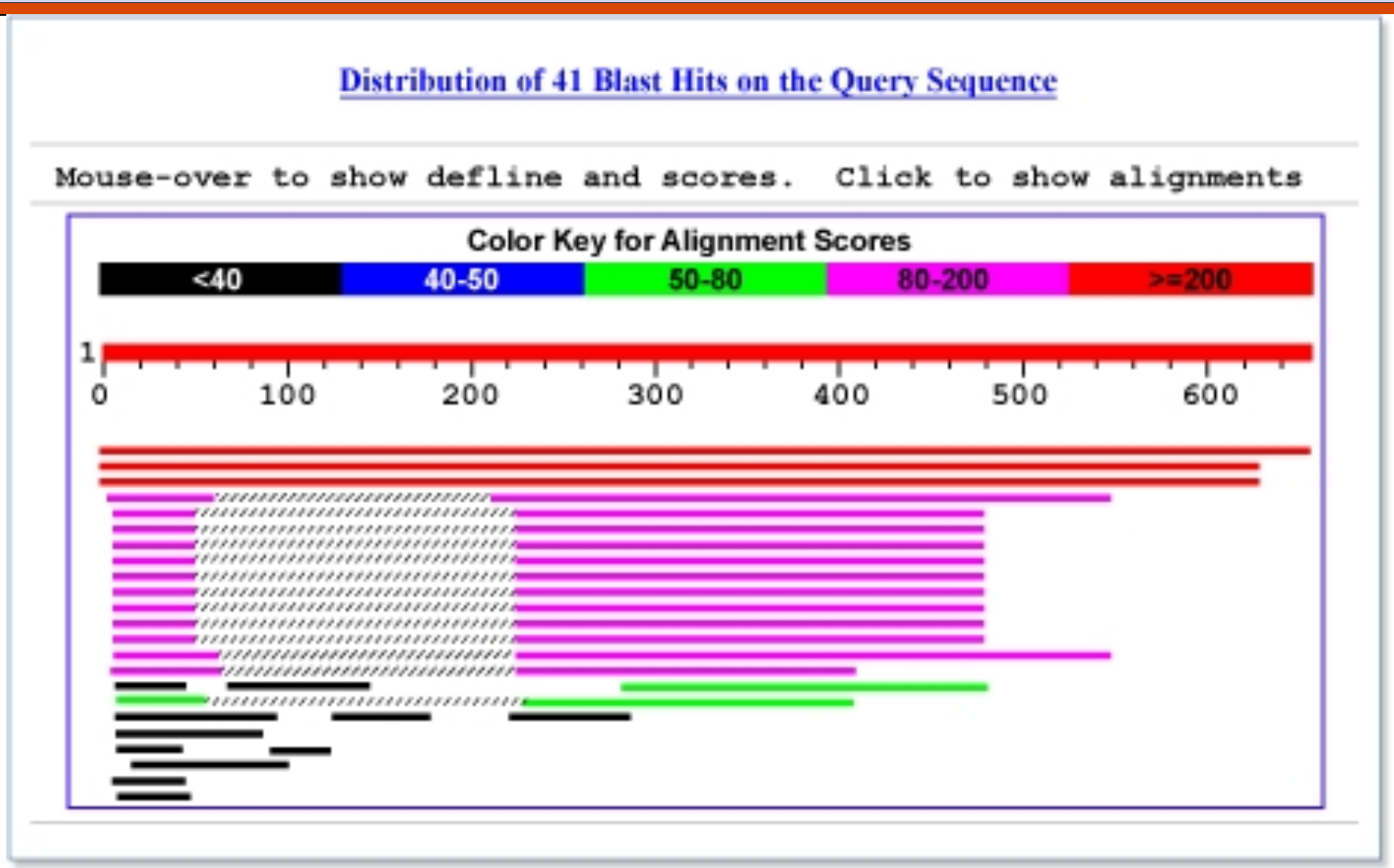
Database: Non-redundant SwissProt sequences
102,387 sequences; 37,391,913 total letters

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|RAE2_HUMAN   Rab proteins geranylgeranyltransferase component A 2 (Rab escort
           protein 2) (REP-2) (Choroideraemia-like protein)
          Length = 656

 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   MADNLPSDFDVIVIGTGLPESIIAAACSRSGQRVLHVDSRSYYGGNWASFSFSGLLSWLK 60

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

Query: 121 KNPSLGVS----NTFTEVLDSALPEESQLSYFNSDEMPAKHTQKSDTEISLEVTDVEESV 176
           KN +   S           S LP   +     S E+PA+ +Q    E S EV D E +
Sbjct: 120 KNHASVTSAQSAEAAEAAETSCLPTAVEPLSMGSCEIPAEQSQCPGPESSPEVNDAEATG 179

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

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

Query: 297 LMKFLTFCLEYEQHPDEYQAFRQCSFSEYLKTKKLTPNLQHFVLHSIAMTSESSCTTIDG 356
           LMKFLTFC+EYE+HPDEY+A+   +FSEYLKT+KLTPNLQ+FVLHSIAMTSE++  T+DG
Sbjct: 292 LMKFLTFCVEYEEHPDEYRAYEGTTFSEYLKTQKLTPNLQYFVLHSIAMTSETTSCTVDG 351

Query: 357 LNATKNFLQCLGRFGNTPFLFPLYGQGEIPQGFCRMCAVFGGIYCLRHKVQCFVVDKESG 416
           L ATK FLQCLGR+GNTPFLFPLYGQGE+PQ FCRMCAVFGGIYCLRH VQC VVDKES
Sbjct: 352 LKATKKFLQCLGRYGNTPFLFPLYGQGELPQCFCRMCAVFGGIYCLRHSVQCLVVDKESR 411

Query: 417 RCKAIIDHFGQRINAKYFIVEDSYLSEETCSNVQYKQISRAVLITDQSILKTDLDQQTSI 476
           +CKA+ID FGQRI +K+FI+EDSYLSE TCS VQY+QISRAVLITD S+LKTD DQQ SI
Sbjct: 412 KCKAVIDQFGQRIISKHFIIEDSYLSENTCSRVQYRQISRAVLITDGSVLKTDADQQVSI 471

Query: 477 LIVPPAEPGACAVRVTELCSSTMTCMKDTYLVHLTCSSSKTAREDLESVVKKLFTPYTET 536
           L VP  EPG+  VRV ELCSSTMTCMK TYLVHLTC SSKTAREDLE VV+KLFTPYTE
Sbjct: 472 LAVPAEEPGSFGVRVIELCSSTMTCMKGTYLVHLTCMSSKTAREDLERVVQKLFTPYTEI 531

Query: 537 EINEEELTKPRLLWALYFNMRDSSGISRSSYNGLPSNVYVCSGPDCGLGNEHAVKQAETL 596
           E   E++ KPRLLWALYFNMRDSS ISR  YN LPSNVYVCSGPD GLGN++AVKQAETL
Sbjct: 532 EAENEQVEKPRLLWALYFNMRDSSDISRDCYNDLPSNVYVCSGPDSGLGNDNAVKQAETL 591

Query: 597 FQXXXXXXXXXXXXXXXXXXXXDGDDKQPEAP 628
           FQ                    DGD  Q E P
Sbjct: 592 FQQICPNEDFCPAPPNPEDIVLDGDSSQQEVP 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|  myoglobin [Homo sapiens]
 ref|NP_976311.1|  myoglobin [Homo sapien
 ref|NP_976312.1|  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     al 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    sum of matches: +60
                  6 4               4
mismatch   -1  1     0  -2 -2  -4  0   sum of mismatches: -13
            -2        0   -3   0
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

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

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

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

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

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.

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.

Related Tools

- Megablast
  - For long, closely-related sequences
  - Uses large w and is very fast
- BLAT
  - UCSC tool
  - DB broken into words; query is searched
- PatternHunter
  - Generalized seeds used instead of words
- BLASTZ, Lagan, SSAHA

Global Alignment: An example

```
V: G A A T T C A G T T A
W: G G A T C G A
```

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0											
G	0											
A	0											
T	0											
C	0											
G	0											
A	0											

Given

$\delta[I, J]$ = Score of Matching
the I^{th} character of sequence V &
the J^{th} character of sequence W

Compute

$S[I, J]$ = Score of Matching
First I characters of sequence V &
First J characters of sequence W

Match/Mismatch score

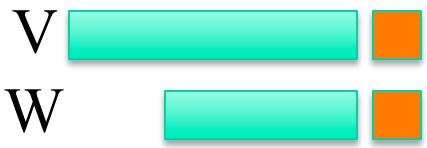
Recurrence Relation

$S[I, J] = MAXIMUM \{$
$S[I-1, J-1] + \delta(V[I], W[J])$,
$S[I-1, J] + \delta(V[I], —)$,
$S[I, J-1] + \delta(— , W[J]) \}$

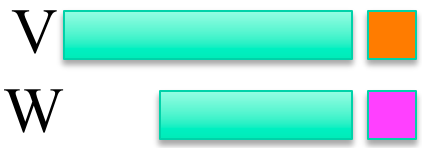
Gap Penalty

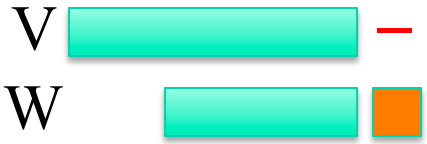
What happens with last character(s)?

1. Last characters **MATCH**

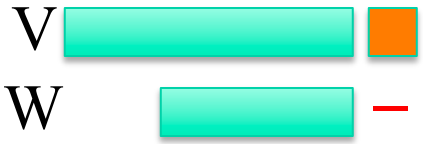


2. Last characters **MISMATCH**



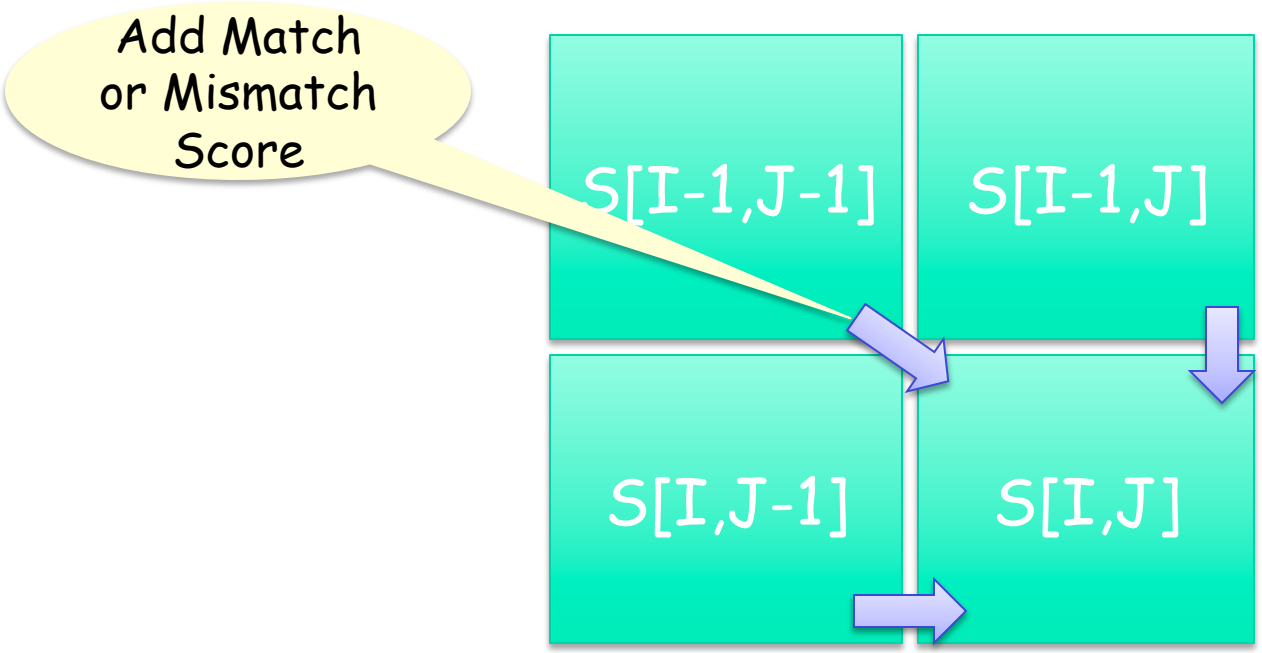


3. Last character of W aligned with GAP



4. Last character of V aligned with GAP

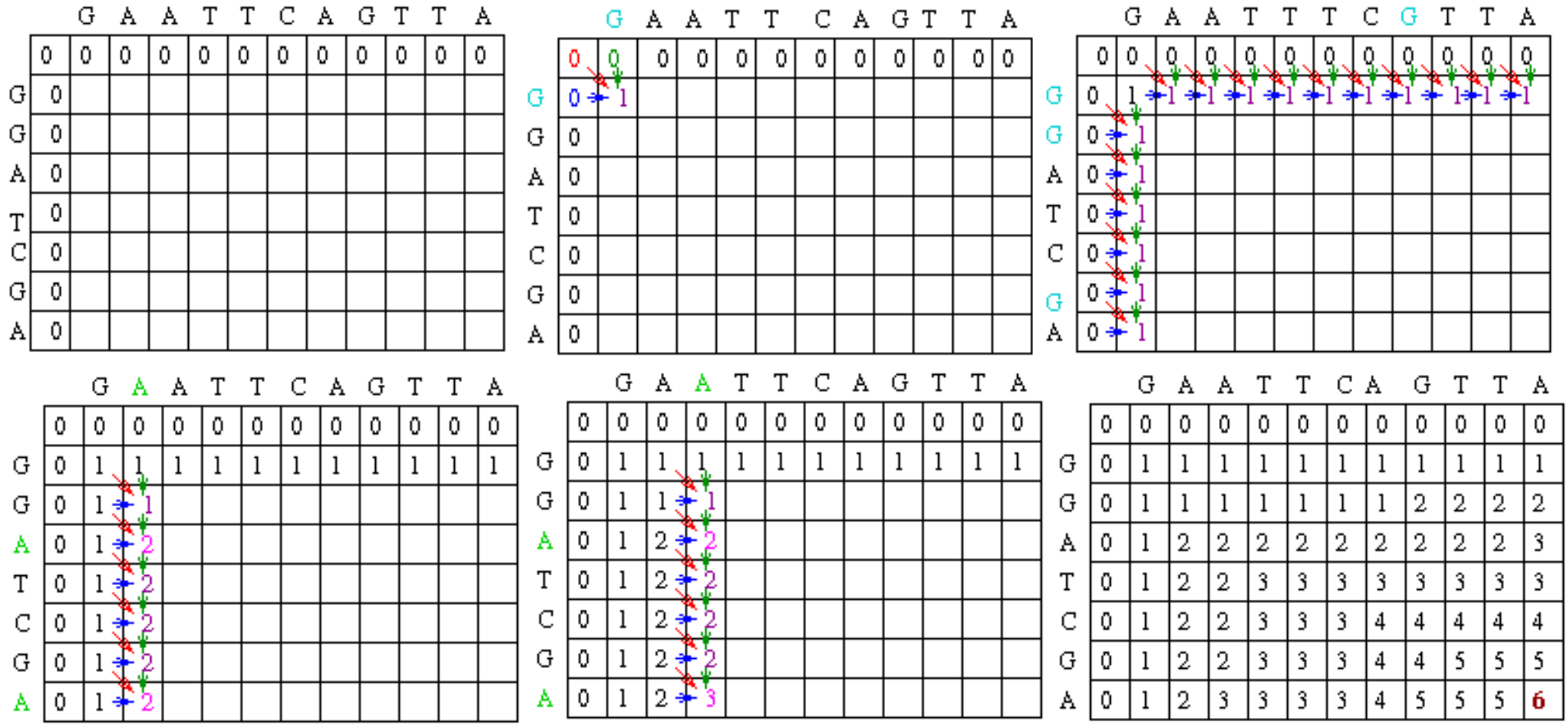
How to fill in the matrix?



Global Alignment: An example

$S[I, J] = \text{MAXIMUM}\{$
$S[I-1, J-1] + \delta(V[I], W[J]),$
$S[I-1, J] + \delta(V[I], —),$
$S[I, J-1] + \delta(— , W[J])\}$

```
V: G A A T T C A G T T A
W: G G A T C G A
```



Match score = 1; Mismatch = Gap = -1

Traceback

$$S[I, J] = \text{MAXIMUM}\{ S[I-1, J-1] + \delta(V[I], W[J]),\ S[I-1, J] + \delta(V[I], —),\ S[I, J-1] + \delta(— , W[J]) \}$$

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	1	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5	5
A	0	1	2	3	3	3	4	5	5	5	5	6

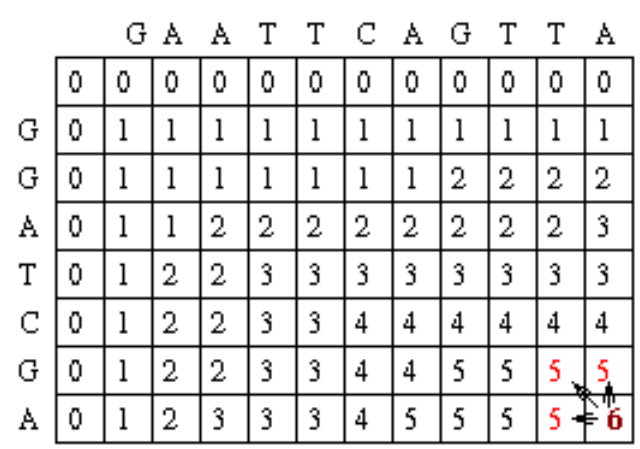
		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	
G	0	1	1	1	1	1	1	1	1	1	1	
G	0	1	1	1	1	1	1	1	2	2	2	
A	0	1	2	2	2	2	2	2	2	2	2	
T	0	1	2	2	3	3	3	3	3	3	3	
C	0	1	2	2	3	3	4	4	4	4	4	
G	0	1	2	2	3	3	4	4	5	5	5	
A												6

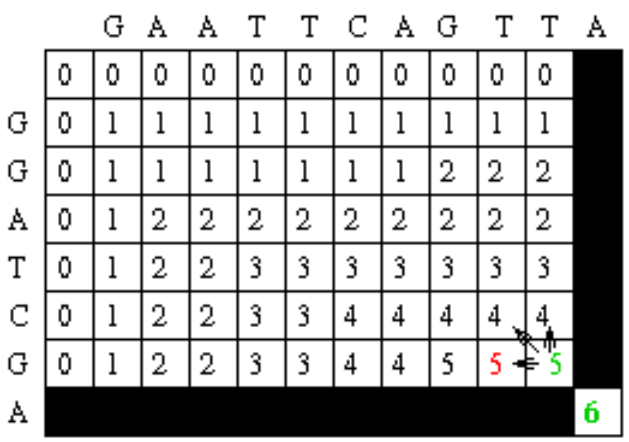
		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	
G	0	1	1	1	1	1	1	1	1	1	1	
G	0	1	1	1	1	1	1	1	2	2	2	
A	0	1	2	2	2	2	2	2	2	2	2	
T	0	1	2	2	3	3	3	3	3	3	3	
C	0	1	2	2	3	3	4	4	4	4	4	
G	0	1	2	2	3	3	4	4	5	5	5	5
A												6

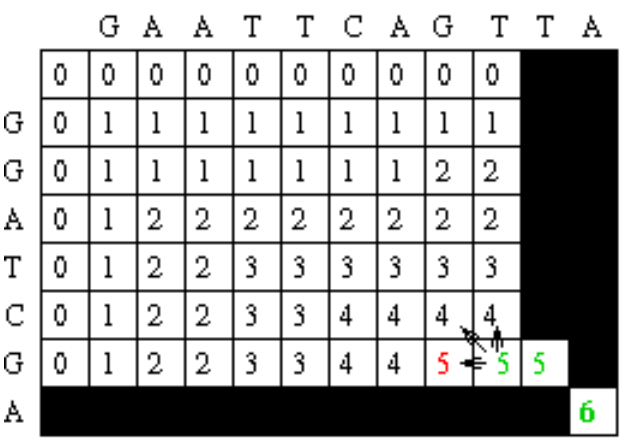
		G	A	A	T	T	C	A	G	T	T	A
	0											
G	1	1										
G		1	2	2								
A					3							
T						4	4					
C								5	5	5		
G											6	
A												

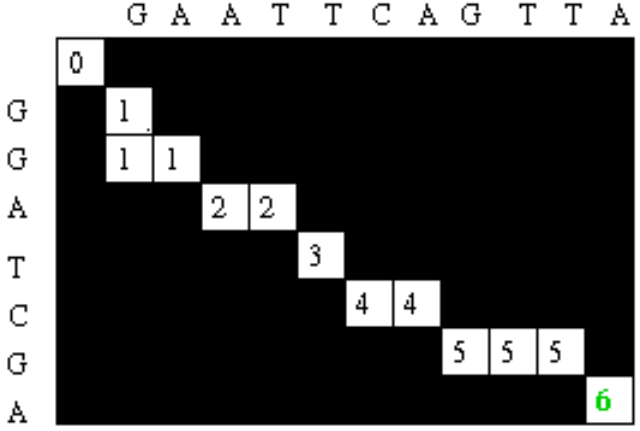
```
V: G A A T T C A G T T A
   |   |   |   |   |
W: G G A - T C - G - - A
```

Alternative Traceback










```
V: G - A A T T C A G T T A
   |   |   | |   |     |
W: G G - A - T C - G - - A
```

```
V: G A A T T C A G T T A
   |   |   | |   |     |
W: G G A - T C - G - - A
```

Previous

Improved Traceback

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	×1	←1	← 1	← 1	← 1	← 1	← 1	×1	← 1	← 1	← 1
G	0	×1	↑1	↑1	↑1	↑1	↑1	↑1	×2	← 2	← 2	← 2
A	0	↑1	↑1	×2	← 2	← 2	← 2	×2	↑2	↑2	↑2	×3
T	0	↑1	← 2	↑2	×3	×3	← 3	← 3	← 3	×3	×3	↑3
C	0	↑1	↑2	↑2	↑3	↑3	×4	← 4	← 4	← 4	← 4	← 4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	×5	← 5	← 5	← 5
A	0	↑1	↑2	×3	↑3	↑3	↑4	×5	↑5	↑5	↑5	×6

Improved Traceback

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	×1	←1	← 1	← 1	← 1	← 1	← 1	×1	← 1	← 1	← 1
G	0	×1	↑1	↑1	↑1	↑1	↑1	↑1	×2	← 2	← 2	← 2
A	0	↑1	↑1	×2	← 2	← 2	← 2	×2	↑2	↑2	↑2	×3
T	0	↑1	← 2	↑2	×3	×3	← 3	← 3	← 3	×3	×3	↑3
C	0	↑1	↑2	↑2	↑3	↑3	×4	← 4	← 4	← 4	← 4	← 4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	×5	← 5	← 5	← 5
A	0	↑1	↑2	×3	↑3	↑3	↑4	×5	↑5	↑5	↑5	×6

Improved Traceback

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	×1	←1	← 1	← 1	← 1	← 1	← 1	×1	← 1	← 1	← 1
G	0	×1	↑1	↑1	↑1	↑1	↑1	↑1	×2	← 2	← 2	← 2
A	0	↑1	↑1	×2	← 2	← 2	← 2	×2	↑2	↑2	↑2	×3
T	0	↑1	← 2	↑2	×3	×3	← 3	← 3	← 3	×3	×3	↑3
C	0	↑1	↑2	↑2	↑3	↑3	×4	← 4	← 4	← 4	← 4	← 4
G	0	↑1	↑2	↑2	↑3	↑3	↑4	↑4	×5	← 5	← 5	← 5
A	0	↑1	↑2	×3	↑3	↑3	↑4	×5	↑5	↑5	↑5	×6

```
V: G A - A T T C A G T T A
   |   |   |   |   |     |
W: G - G A - T C - G - - A
```

Subproblems

- Optimally align V[1..I] and W[1..J] for every possible values of I and J.
- Having optimally aligned
  - V[1..I-1] and W[1..J-1]
  - V[1..I] and W[1..J-1]
  - V[1..I-1] and W[1, J]

  it is possible to optimally align V[1..I] and W[1..J]

- O(mn),
  where m = length of V,
  and n = length of W.

Generalizations of Similarity Function

- Mismatch Penalty = α
- Spaces (Insertions/Deletions, InDels) = β
- Affine Gap Penalties:

  (Gap open, Gap extension) = (γ,δ)
- Weighted Mismatch = Φ(a,b)
- Weighted Matches = Ω(a)

Alternative Scoring Schemes

		G	A	A	T	T	C	A	G	T	T	A
	0	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
G	-2	× 1	← -1	← -2	← -3	← -4	← -5	← -6	← -7	← -8	← -9	← -10
G	-3	↑-1	× -1	← -3	← -4	← -5	← -6	← -7	× -5	← -7	← -8	← -9
A	-4	↑-2	× 0	× 0	← -2	← -3	← -4	← -5	← -6	← -7	← -8	× -7
T	-5	↑-3	↑ -2	↑-2	× 1	← -1	← -2	← -3	← -4	← -5	← -6	← -7
C	-6	↑-4	↑ -3	↑-3	↑-1	× -1	× 0	← -2	← -3	← -4	← -5	← -6
G	-7	↑-5	↑-4	↑-4	↑-2	↑-3	↑-2	× -2	× -1	← -3	← -4	← -5
A	-8	↑-6	↑-5	↑-5	↑-3	↑-4	↑-3	× -1	↑-3	× -3	× -5	× -3

Match +1
Mismatch –2
Gap (-2, -1)

```
V: G A A T T C A G T T A
   |   | |   |   |     |
W: G G A T - C - G - - A
```

Local Sequence Alignment

- Example: comparing long stretches of anonymous DNA; aligning proteins that share only some motifs or domains.
- Smith-Waterman Algorithm

Recurrence Relations (Global vs Local Alignments)

- S[I, J] = MAXIMUM {
  S[I-1, J-1] + δ(V[I], W[J]),
  S[I-1, J] + δ(V[I], —),
  S[I, J-1] + δ(— , W[J]) }

Global Alignment

- S[I, J] = MAXIMUM { 0,
  S[I-1, J-1] + δ(V[I], W[J]),
  S[I-1, J] + δ(V[I], —),
  S[I, J-1] + δ(— , W[J]) }

Local Alignment

Local Alignment: Example

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	× 1	0	0	0	0	0	0	0	0	0	0
G	0	× 1	← 0	0	0	0	0	0	× 1	0	0	0
A	0	0	× 2	× 1	0	0	0	× 1	0	0	0	× 1
T	0	0	↑ 0	× 1	× 2	← 1	0	0	0	× 1	× 1	0
C	0	0	0	0	↑ 0	× 0	× 2	0	0	0	0	0
G	0	0	0	0	0	0	0	0	× 1	0	0	0
A	0	0	× 1	× 1	0	0	0	× 1	0	0	0	× 1

```
V: - G A A T T C A G T T A
     |   |   |     |
W: G G - A T - C - G - - A
```

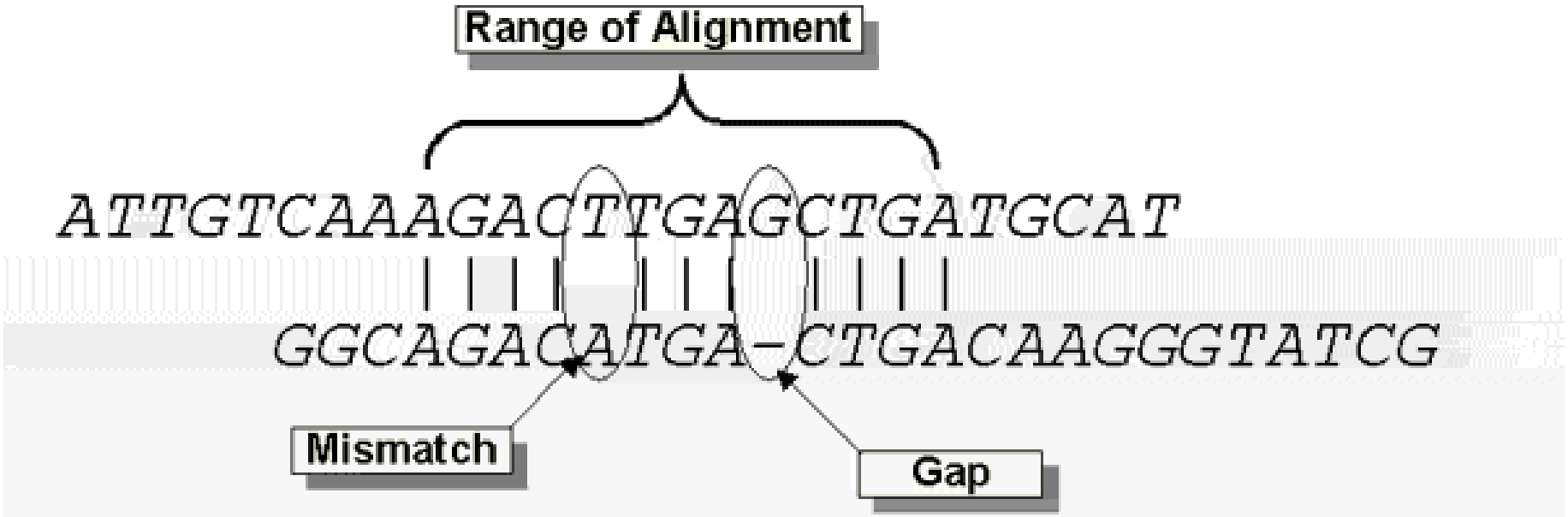
Match +1
Mismatch –1
Gap (-1, -1)

Properties of Smith-Waterman Algorithm

- How to find all regions of "high similarity"?
  - Find all entries above a threshold score and traceback.
- What if: Matches = 1 & Mismatches/spaces = 0?
  - Longest Common Subsequence Problem
- What if: Matches = 1 & Mismatches/spaces = $-\infty$?
  - Longest Common Substring Problem
- What if the average entry is positive?
  - Global Alignment

Slide: Courtesy J. Pevsner

Calculation of an alignment score



$$S = \sum(\text{identities, mismatches}) - \sum(\text{gap penalties})$$

$$\text{Score} = \text{Max}(S)$$

Source: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Alignment_Scores2.html

How to score mismatches?

Blosum62 scoring matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
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

Slide: Courtesy J. Pevsner

BLOSUM n Substitution Matrices

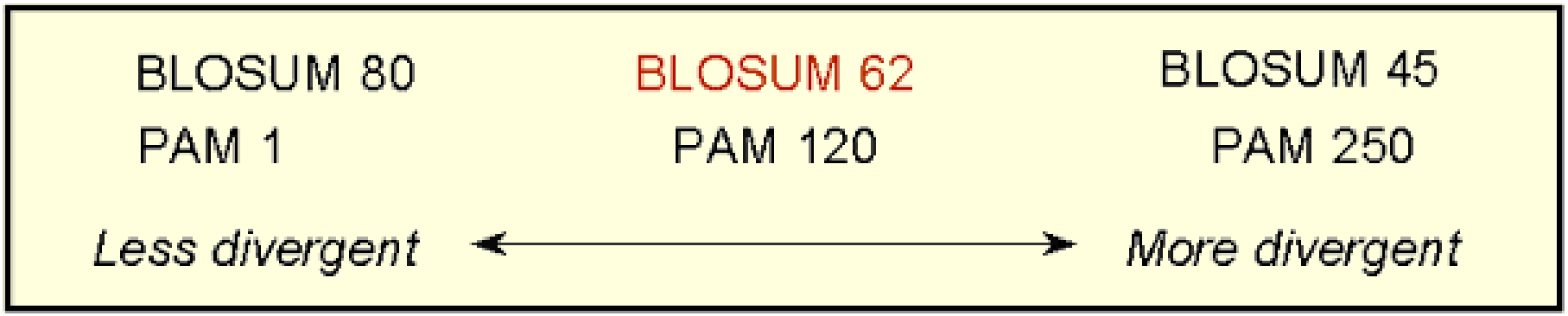
- For each amino acid pair a, b
  - For each BLOCK
    - Align all proteins in the BLOCK
    - Eliminate proteins that are more than n% identical
    - Count F(a), F(b), F(a,b)
    - Compute Log-odds Ratio

$$\log\left(\frac{F(a,b)}{F(a)F(b)}\right)$$

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%)



Slide: Courtesy J. Pevsner

Local/Standalone BLAST

- Go to: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/
- Right click on a desired archive and select "Save link as..." from the popup menu
- In the prompt, switch to a desired directory (folder) and click the "Save" button to save the archive to a desired location on the local disk
- Installation details are at:
  - Windows: http://www.ncbi.nlm.nih.gov/books/NBK52637/
  - Unix: http://www.ncbi.nlm.nih.gov/books/NBK52640/
  - Help: http://www.ncbi.nlm.nih.gov/books/NBK1762/
- With the help of this installation, you can run BLAST with preformatted databases or format your own database before you run BLAST queries.

Multiple Sequence Alignment

Multiple Alignments

- Global
  - ClustalW, ClustalX
  - MSA
  - T-Coffee
- Local
  - BLOCKS
  - eMOTIF
  - GIBBS
  - HMMER
  - MACAW
  - MEME
- Other
  - Profile Analysis from msa (UCSD)
  - SAM HMM (from msa)

MSA of glyceraldehyde 3-phosphate dehydrogenases: example of high conservation

```
fly        GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
human      GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS CTTNCLAPLA
plant      GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
bacterium  GAKKVVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
yeast      GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
archaeon   GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS CTTNSITPVA

fly        KVINDNFEIV EGLMTTVHAT TATQKTVDGP SGKLWRDGRG AAQNIIPAST
human      KVIHDNFGIV EGLMTTVHAI TATQKTVDGP SGKLWRDGRG ALQNIIPAST
plant      KVVHEEFGIL EGLMTTVHAT TATQKTVDGP SMKDWRGGRG ASQNIIPSST
bacterium  KVINDNFGII EGLMTTVHAT TATQKTVDGP SHKDWRGGRG ASQNIIPSST
yeast      KVINDAFGIE EGLMTTVHSL TATQKTVDGP SHKDWRGGRT ASGNIIPSST
archaeon   KVLDEEFGIN AGQLTTVHAY TGSQNLMDGP NGKP.RRRRA AAENIIPTST

fly        GAAKAVGKVI PALNGKLTGM AFRVPTPNVS VVDLTVRLGK GASYDEIKAK
human      GAAKAVGKVI PELNGKLTGM AFRVPTANVS VVDLTCRLEK PAKYDDIKKV
plant      GAAKAVGKVL PELNGKLTGM AFRVPTSNVS VVDLTCRLEK GASYEDVKAA
bacterium  GAAKAVGKVL PELNGKLTGM AFRVPTPNVS VVDLTVRLEK AATYEQIKAA
yeast      GAAKAVGKVL PELQGKLTGM AFRVPTVDVS VVDLTVKLNK ETTYDEIKKV
archaeon   GAAQAATEVL PELEGKLDGM AIRVPVPNGS ITEFVVDLDD DVTESDVNAA
```


Slide: Courtesy J. Pevsner

Multiple Alignments: CLUSTALW

* identical

: conserved substitutions

. semi-conserved substitutions

```
gi|2213819      CDN-ELKSEAIIEHLCASEFALR-------------MKIKEVKKENGDKK 223
gi|12656123     ----ELKSEAIIEHLCASEFALR-------------MKIKEVKKENGD-   31
gi|7512442      CKNKNDDDNDIMETLCKNDFALK-------------IKVKEITYINRDTK 211
gi|1344282      QDECKFDYVEVYETSSSGAFSLLGRFCGAEPPPHLVSSHHELAVLFRTDH 400
                    : .   : *  . . *:*               . :*:
```

Red: AVFPMLW (Small & hydrophobic)

Blue: DE (Acidic)

Magenta: RHK (Basic)

Green: STYHCNGQ (Hydroxyl, Amine, Basic)

Gray: Others

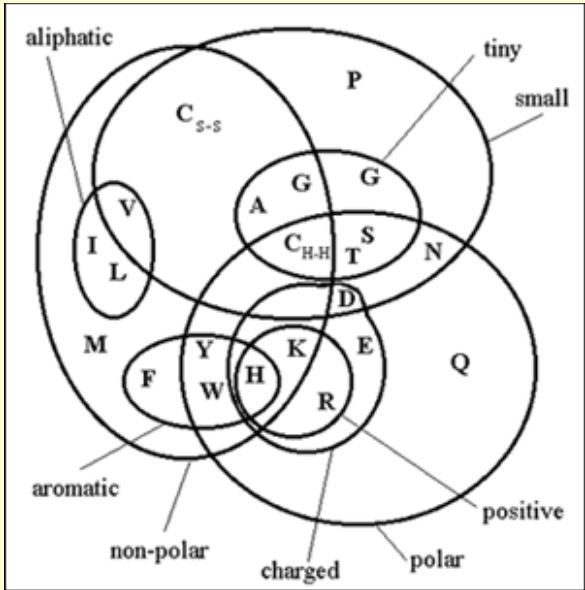
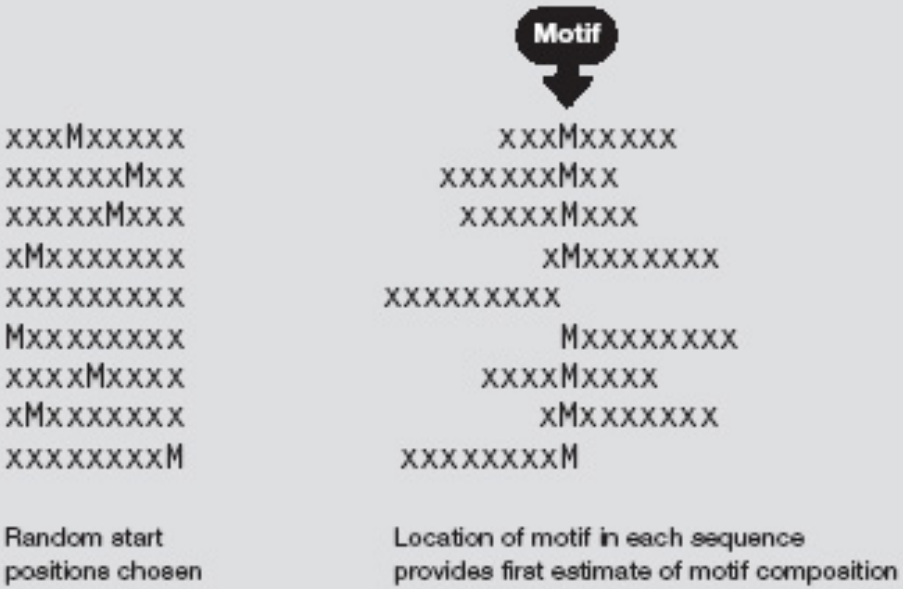


Figure 1. A Venn diagram showing the relationship of the 20 naturally occurring amino acids to a selection of physio-chemical properties thought to be important in the determination of protein structure.

Multiple Alignment

A. Estimate the amino acid frequencies in the motif columns of all but one sequence. Also obtain background.



How to Score Multiple Alignments?

- Sum of Pairs Score (SP)
  - Optimal alignment: $O(d^N)$ [Dynamic Prog]
  - Approximate Algorithm: Approx Ratio 2
    - Locate Center: $O(d^2N^2)$
    - Locate Consensus: $O(d^2N^2)$

Consensus char: char with min distance sum

Consensus string: string of consensus char

Center: input string with min distance sum

Multiple Alignment Methods

- Phylogenetic Tree Alignment (NP-Complete)
  - Given tree, task is to label leaves with strings
- Iterative Method(s)
  - Build a MST using the distance function
- Clustering Methods
  - Hierarchical Clustering
  - K-Means Clustering

Multiple Alignment Methods (Cont'd)

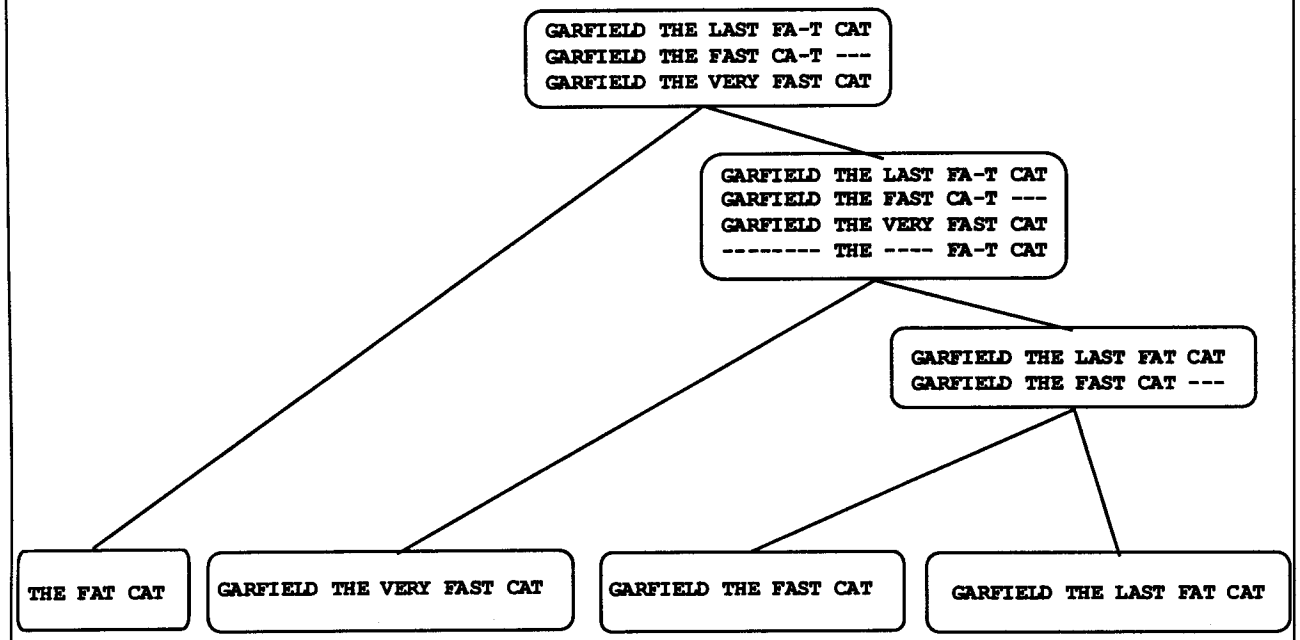
- Gibbs Sampling Method
  - Lawrence, Altschul, Boguski, Liu, Neuwald, Winton, *Science*, 1993
- Hidden Markov Model
  - Krogh, Brown, Mian, Sjolander, Haussler, *JMB*, 1994

Multiple Sequence Alignments (MSA)

- Choice of Scoring Function
  - Global vs local
  - Gap penalties
  - Substitution matrices
  - Incorporating other information
  - Statistical Significance
- Computational Issues
  - Exact/heuristic/approximate algorithms for optimal MSA
  - Progressive/Iterative/DP
  - Iterative: Stochastic/Non-stochastic/Consistency-based
- Evaluating MSAs
  - Choice of good test sets or benchmarks (BAliBASE)
  - How to decide thresholds for good/bad alignments

Progressive MSA: CLUSTALW

Figure 1. Limits of the progressive strategy.



This example shows how a progressive alignment strategy can be misled. In the initial alignment of sequences 1 and 2, ClustalW has a choice between aligning CAT with CAT and making an internal gap or making a mismatch between C and F and having a terminal gap. Since terminal gaps are much cheaper than internals, the ClustalW scoring schemes prefers the former. In the next stage, when the extra sequence is added, it turns out that properly aligning the two CATs in the previous stage would have led to a better scori ng sums-of-pairs multiple alignment.

C. Notredame, *Pharmacogenomics*, **3**(1), 2002.

Software for MSA

Table 1. Some recent and less recent available methods for MSAs.

Name	Algorithm	URL	Ref.
MSA	Exact	http://www.ibc.wustl.edu/ibc/msa.html	[28]
OMA	Iterative DCA	http://bibiserv.techfak.uni-biefield.de/oma	[61]
MultAlin	Progressive	http://www.toulouse.inra.fr/multalin.html	[41]
ComAlign	Consistency-based	http://www.daimi.au.df/~ ocaprani	[75]
Praline	Iterative/progressive	jhering@nimr.mrc.ac.uk	[48]
Prrp	Iterative/Stochastic	ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/	[47]
HMMER	Iterative/Stochastic/HMM	http://hmmer.wustl.edu/	[68]
GA	Iterative/Stochastic/GA	czhang@watnow.uwaterloo.ca	[52]

C. Notredame, *Pharmacogenomics*, **3**(1), 2002.

MSA: Conclusions

- Very important
  - Phylogenetic analyses
  - Identify members of a family
  - Protein structure prediction
- No perfect methods
- Popular
  - Progressive methods: CLUSTALW
  - Recent interesting ones: Prrp, SAGA, DiAlign, T-Coffee
- Review of Methods [C. Notredame, *Pharmacogenomics,* **3**(1), 2002]
  - CLUSTALW works reasonably well, in general
  - DiAlign is better for sequences with long insertions & deletions (indels)
  - T-Coffee is best available method