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/BioinfS11.html

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

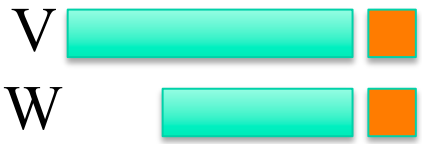
Recurrence Relation

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

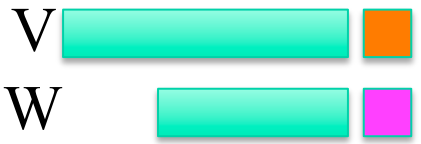
Gap Penalty

What happens with last character(s)?

1. Last characters **MATCH**



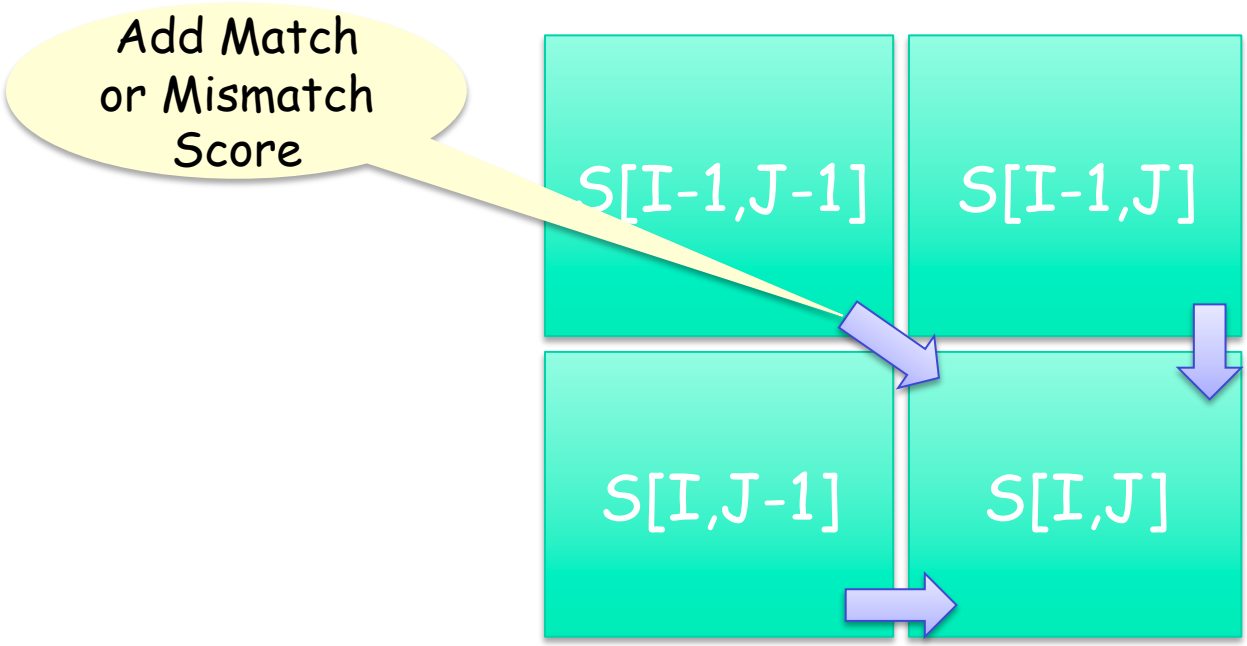
2. Last characters **MISMATCH**



3. Last character of W aligned with GAP

4. Last character of V aligned with GAP

How to fill in the matrix?



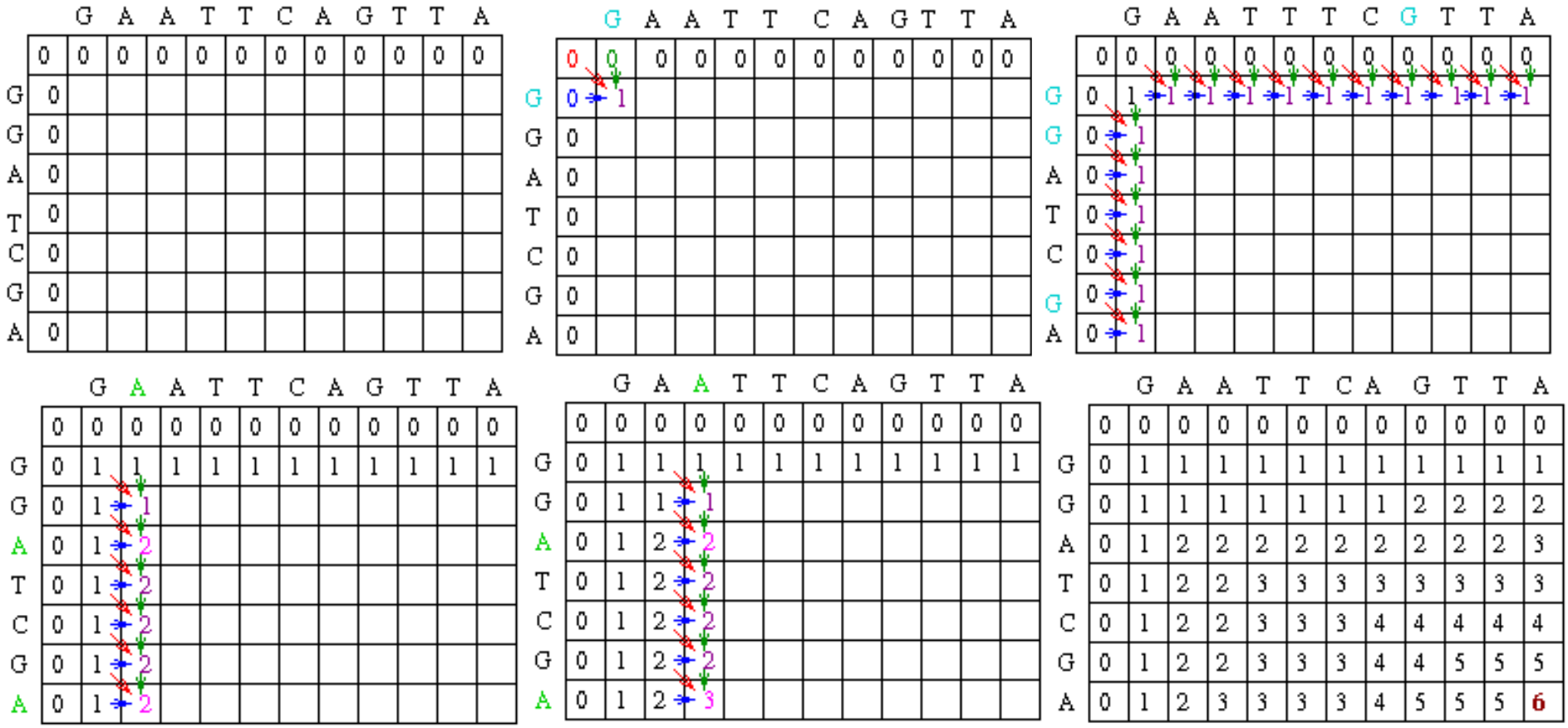
Add gap penalty for gap in seq 1

Add gap penalty for gap in seq 2

Global Alignment: An example

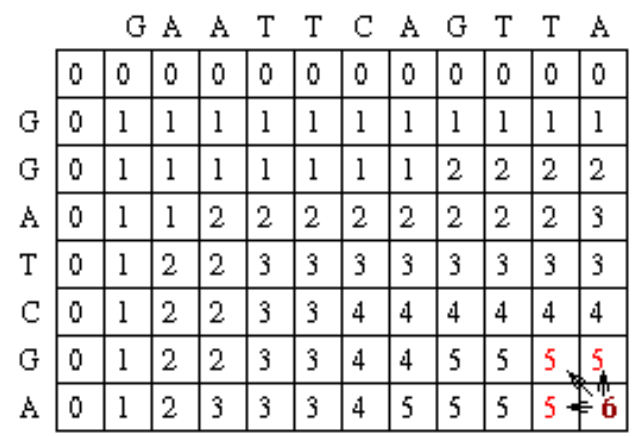
S[I, J] = MAXIMUM {
S[I-1, J-1] + δ(V[I], W[J]),
S[I-1, J] + δ(V[I], —),
S[I, J-1] + δ(— , W[J]) }

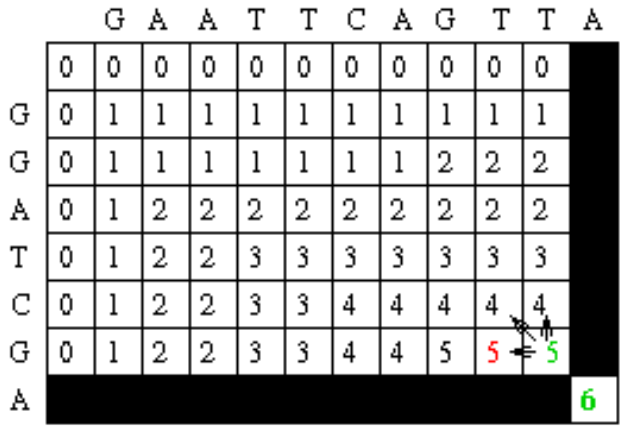
V: G A A T T C A G T T A
W: G G A T C G A

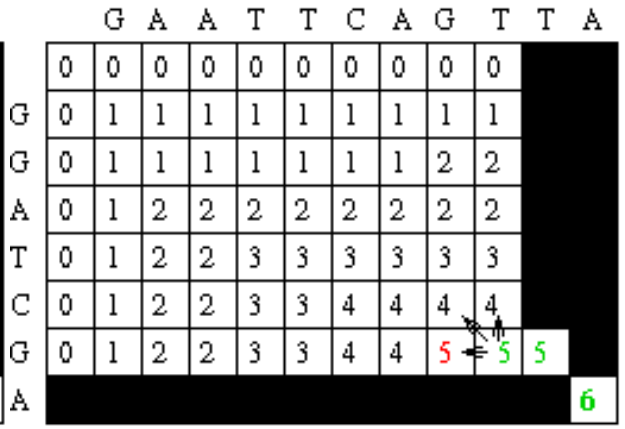


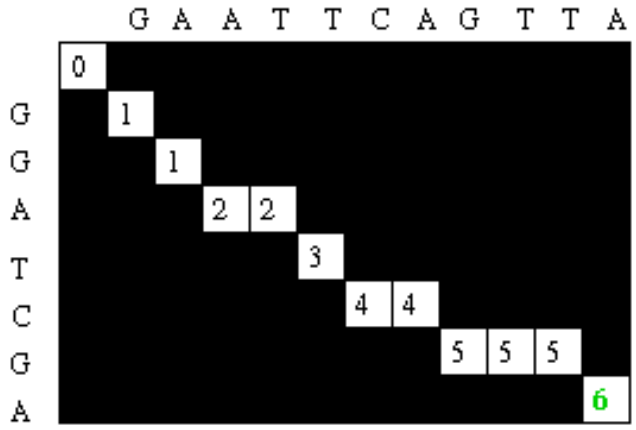
Match score = 1; Mismatch = Gap = -1

Traceback



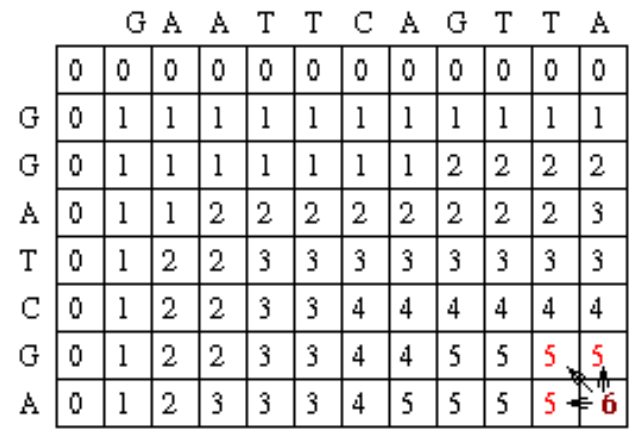


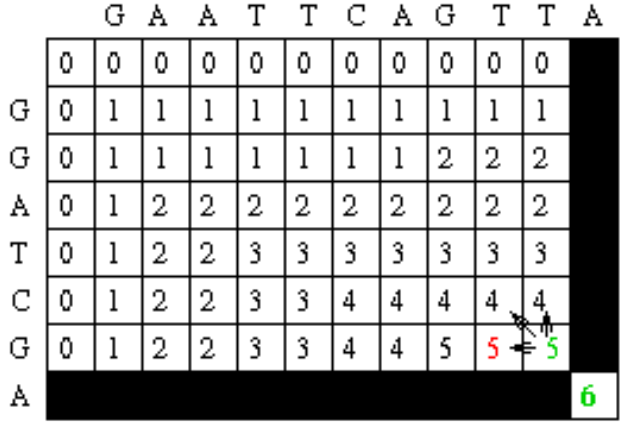


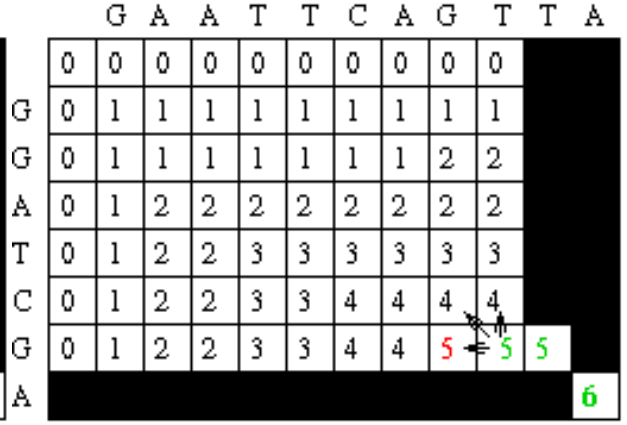


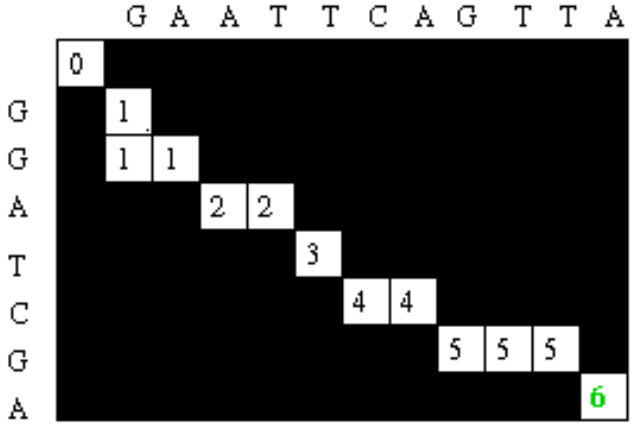

```
V: G A A T T C A G T T A
   |   |   |  |   |     |
W: G G A - T C - G - - A
```

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 = $\Phi(a,b)$
- Weighted Matches = $\Omega(a)$

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

Match +1
Mismatch –1
Gap (-1, -1)

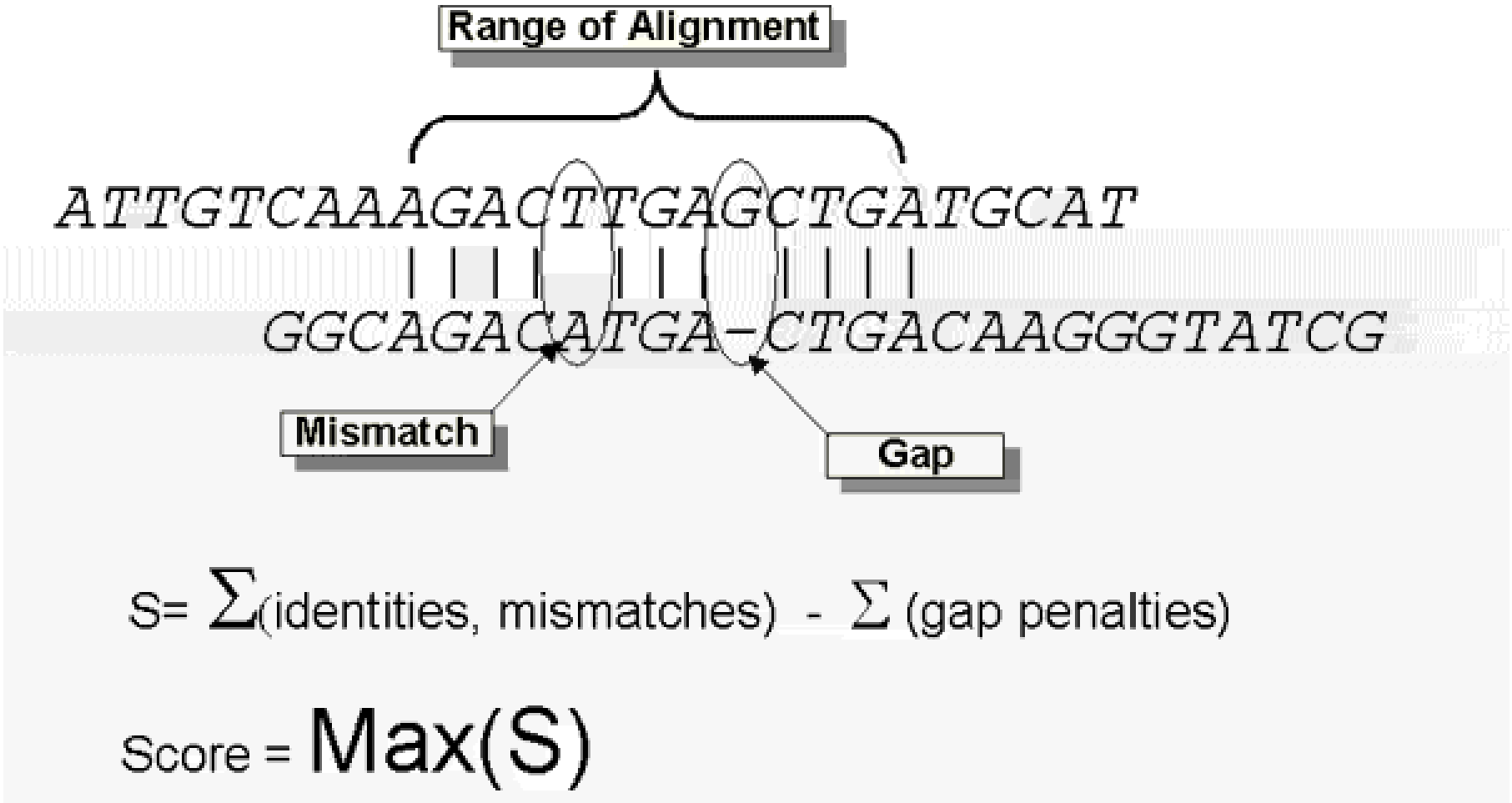
```
V: - G A A T T C A G T T A
     |   | | |   |
W: G G - A T - C - G - - A
```

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

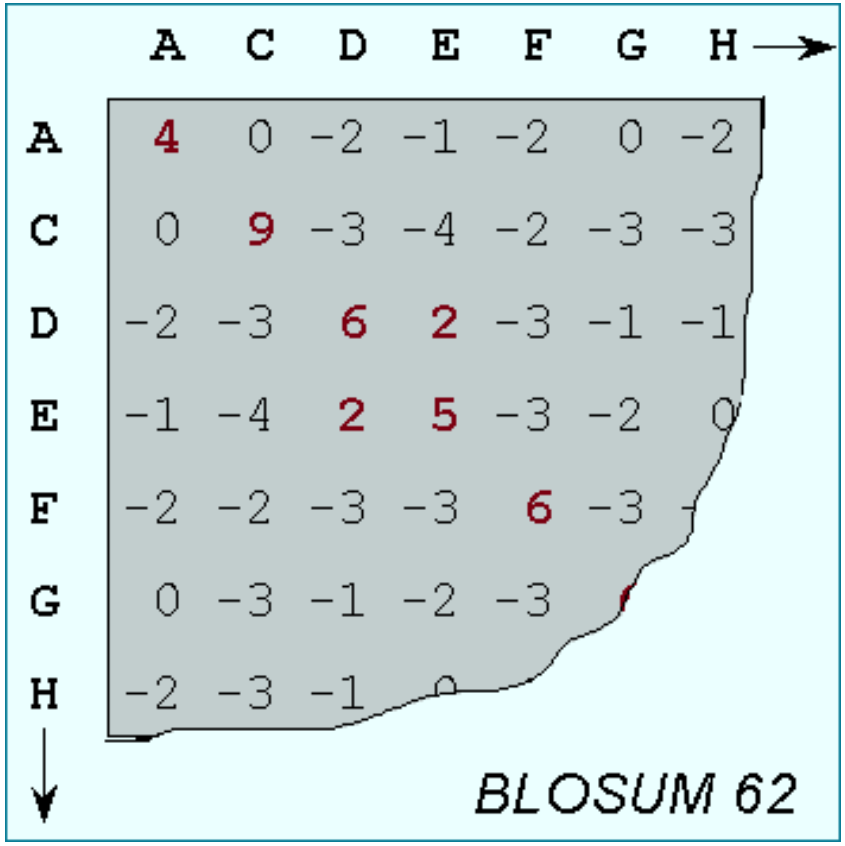
How to score mismatches?

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

How to score mismatches?



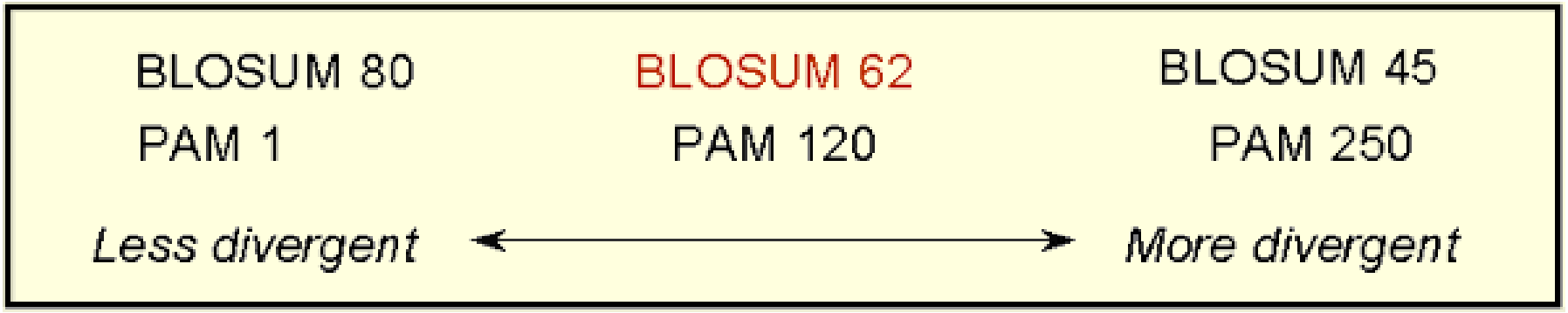
BLOSUM n Substitution Matrices

- 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

BLAST: Steps

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

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

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
- 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.

BLAST

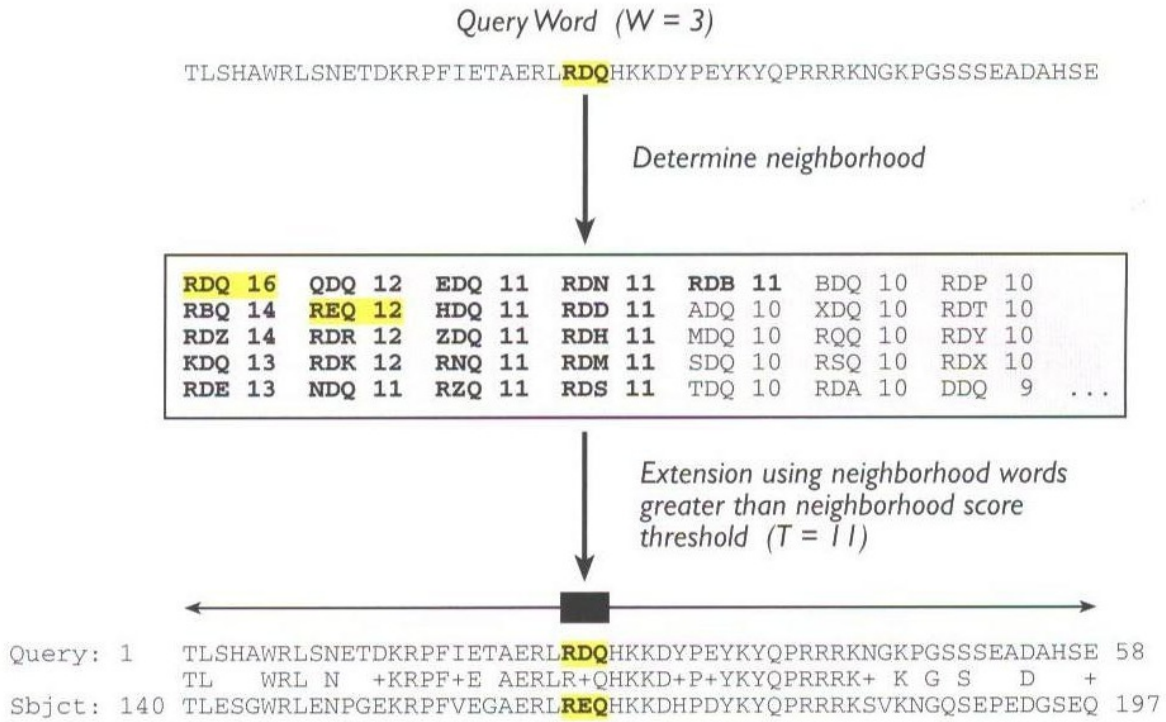
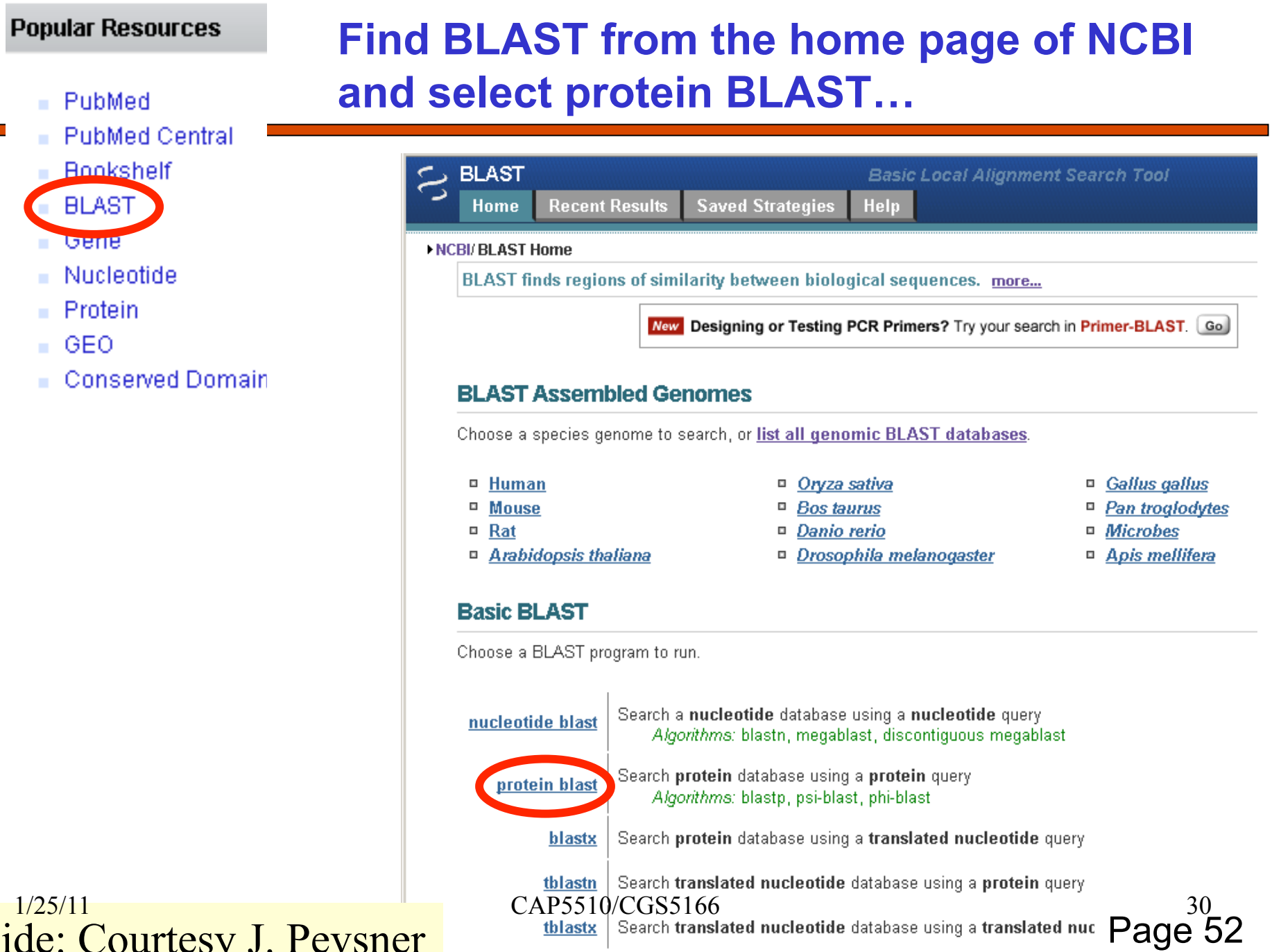
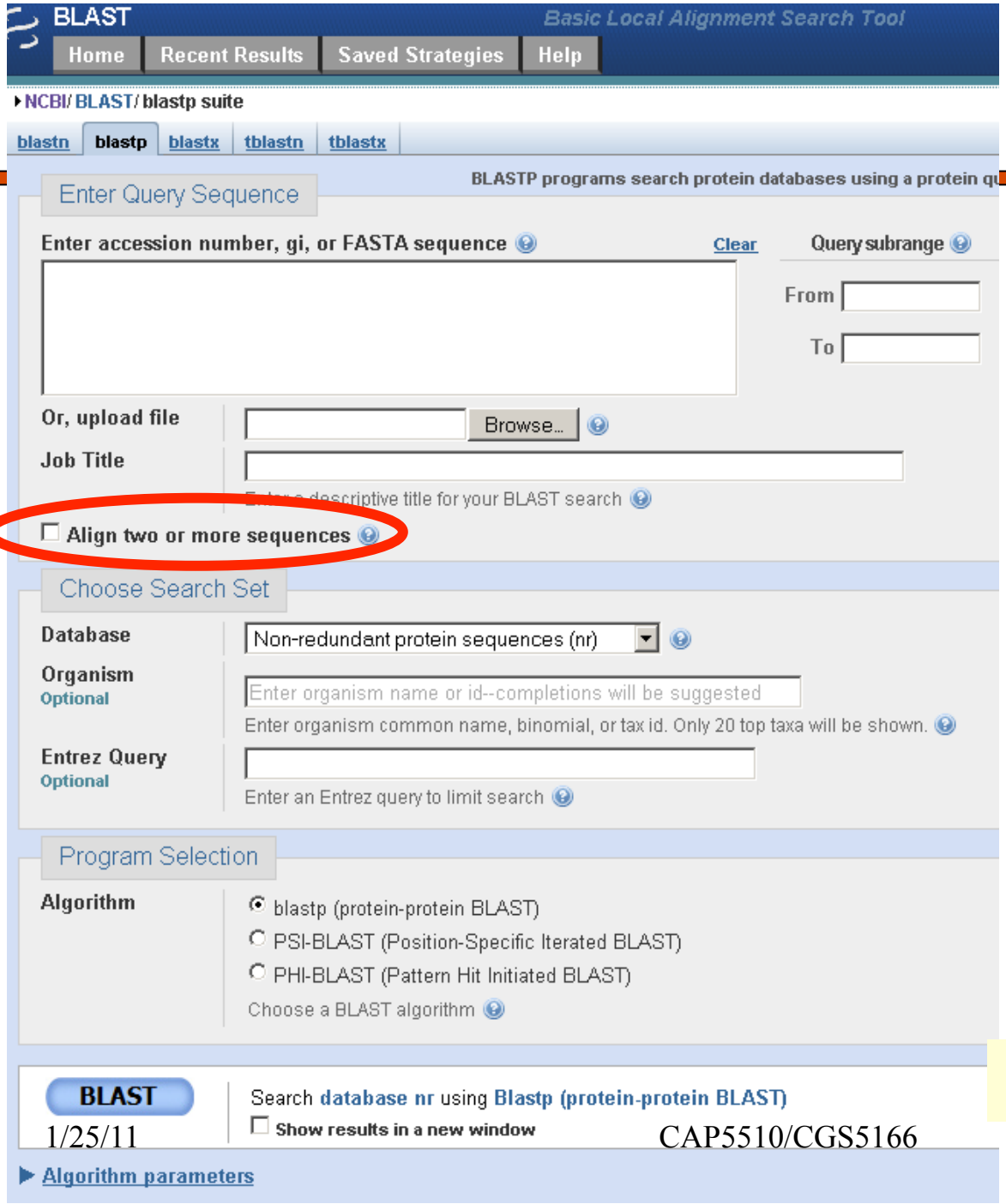


FIGURE 11.7 The initiation of a **BLAST** search. The search begins with query words of a given length (here, three amino acids) being compared against a scoring matrix to determine additional three-letter words "in the neighborhood" of the original query word. Any occurrences of these neighborhood words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

Find BLAST from the home page of NCBI and select protein BLAST…





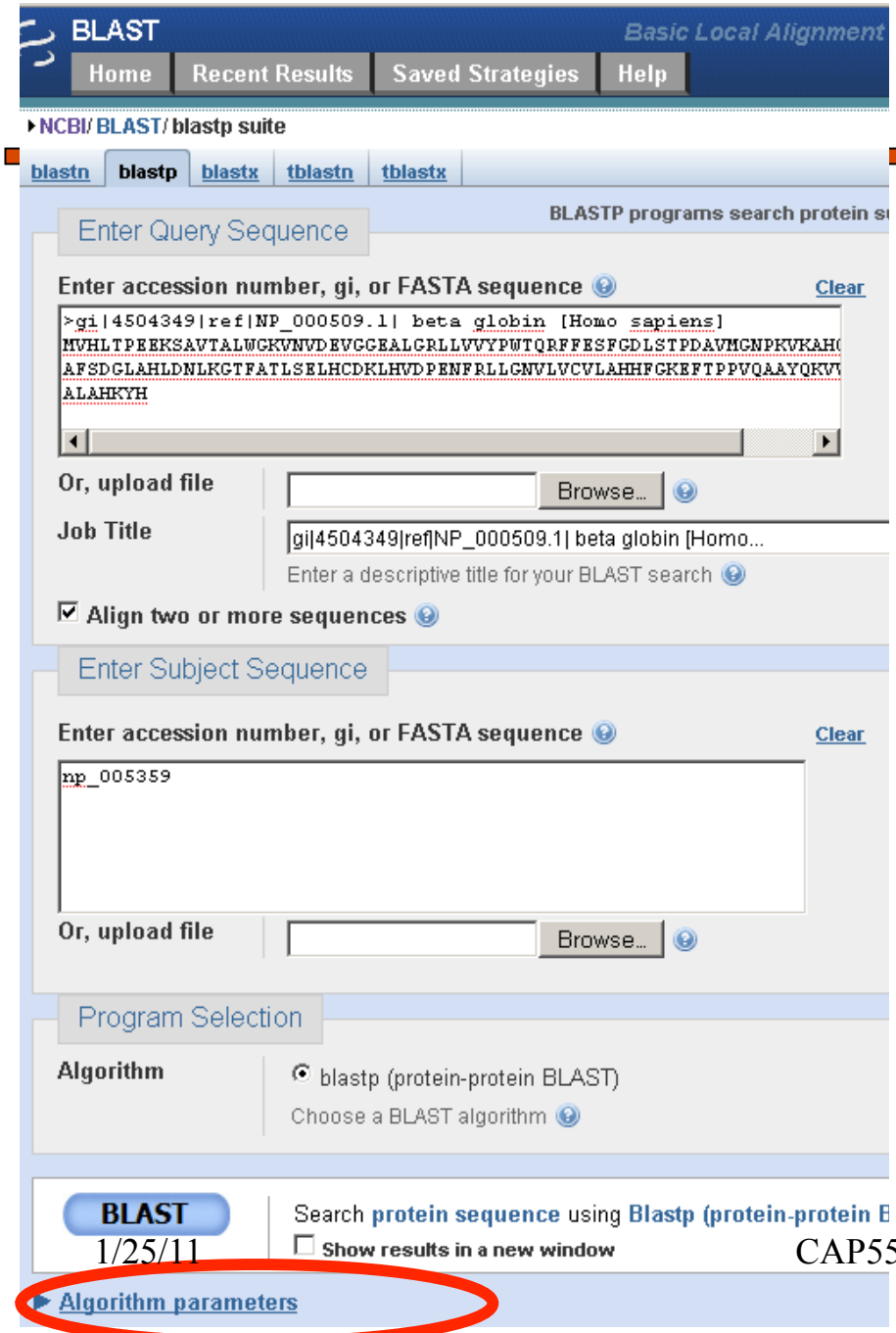
Choose align two or more sequences…

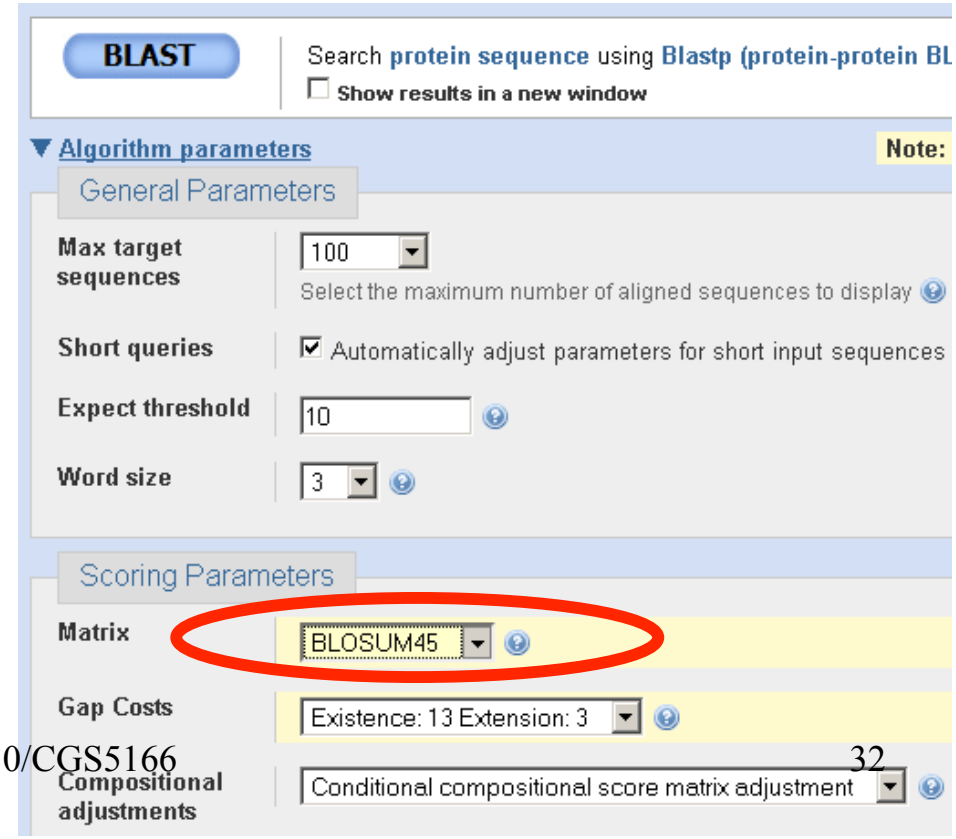
Slide: Courtesy J. Pevsner

Slide: Courtesy J. Pevsner

Enter the two sequences (as accession numbers or in the fasta format) and click BLAST.

Optionally select "Algorithm parameters" and note the matrix option.





Pairwise alignment result of human beta globin and myoglobin

Myoglobin RefSeq

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

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

 GENE ID: 4151 MB | myoglobin [Homo sapiens] (Over 10 PubMed links)

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

Query  4    LTPEEKSAVTALWGKVNVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  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

Middle row displays identities; + sign for similar matches

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 4 6 6 5 4 5 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

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

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

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.

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

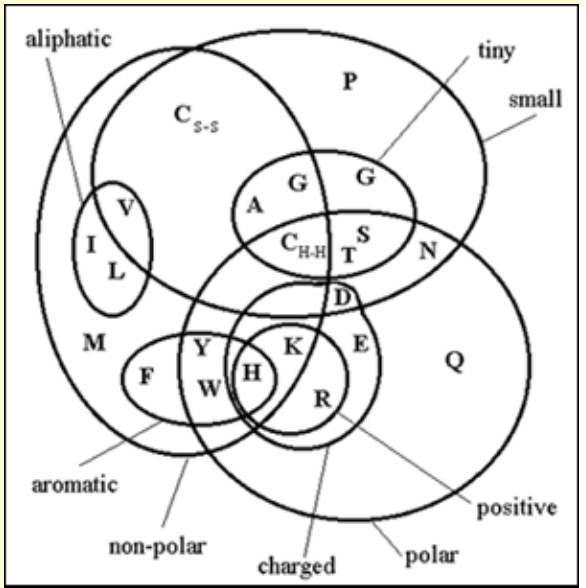
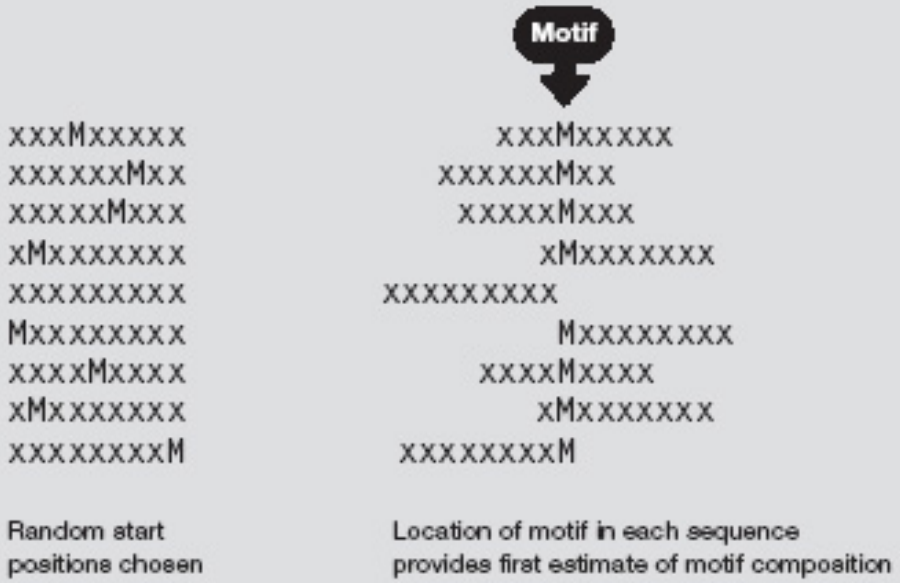


Figure 1. A Venn diagram showing the relationship of the 20 naturally occurring amino acids to a selection of physio-chemical properties thought to be important in the determination of protein structure.

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)

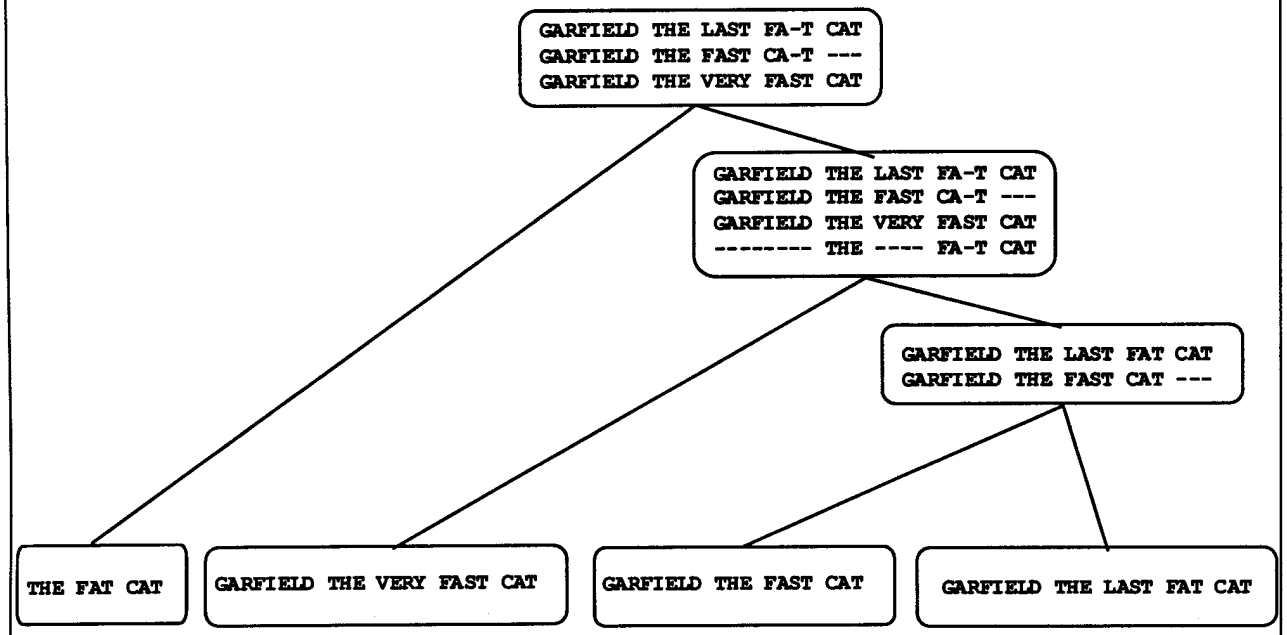
- Gibbs Sampling Method
  - Lawrence, Altschul, Boguski, Liu, Neuwald, Winton, *Science*, 1993
- Hidden Markov Model
  - Krogh, Brown, Mian, Sjolander, Haussler, *JMB*, 1994

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