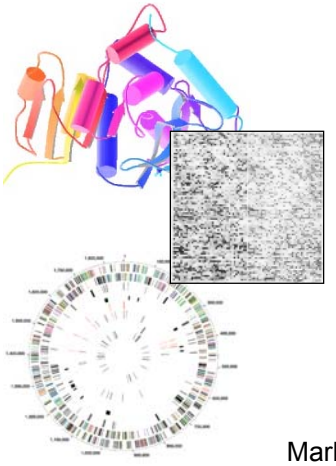
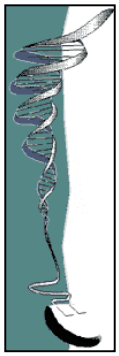
BIOINFORMATICS
Sequences





Mark Gerstein, Yale University
bioinfo.mbb.yale.edu/mbb452a
(last edit in fall 2002)

Sequence Topics (Contents)

- Basic Alignment via Dynamic Programming
- Suboptimal Alignment
- Gap Penalties
- Similarity (PAM) Matrices
- Multiple Alignment
- Profiles, Motifs, HMMs
- Local Alignment
- Probabilistic Scoring Schemes
- Rapid Similarity Search: Fasta
- Rapid Similarity Search: Blast
- Practical Suggestions on Sequence Searching
- Transmembrane helix predictions
- Secondary Structure Prediction: Basic GOR
- Secondary Structure Prediction: Other Methods
- Assessing Secondary Structure Prediction
- Features of Genomic DNA sequences

Molecular Biology Information: Protein Sequence

- 20 letter alphabet
  - ◊ ACDEFGHIKLMNPQRSTVWY but not BJOUXZ
- Strings of ~300 aa in an average protein (in bacteria), ~200 aa in a domain
- ~200 K known protein sequences

```
d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL--------NKPVIMGRHTWESI
d3dfr__ TAFLWAQDRDGLIGKDGHLPWH-LPDDLHYFRAQTV--------GKIMVVGRRTYESF

d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
d4dfra_ ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLD--------KPVIMGRHTWESI
d3dfr__ TAFLWAQDRNGLIGKDGHLPW-HLPDDLHYFRAQTVG--------KIMVVGRRTYESF

d1dhfa_ VPEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__ VPEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_ ---G-RPLPGRKNIILS-SQPGTDDRV-TWVKSVDEAIAACGDVP------EIMVIGGGRVYEQFLPKA
d3dfr__ ---PKRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLDQ----ELVIAGGAQIFTAFKDDV

d1dhfa_ -PEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
d8dfr__ -PEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
d4dfra_ -G---RPLPGRKNIILSSSQPGTDDRV-TWVKSVDEAIAACGDVPE----- IMVIGGGRVYEQFLPKA
d3dfr__ -P--KRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLD----QELVIAGGAQIFTAFKDDV
```


Aligning Text Strings

Core

Raw Data ???

```
T  C  A  T  G
   C  A  T  T  G
```

2 matches, 0 gaps

```
T  C  A  T  G
         |  |
C  A  T  T  G
```

3 matches (2 end gaps)

```
T  C  A  T  G  .
   |  |  |
.  C  A  T  T  G
```

4 matches, 1 insertion

```
T  C  A  -  T  G
   |  |     |  |
.  C  A  T  T  G
```

4 matches, 1 insertion

```
T  C  A  T  -  G
   |  |  |     |
.  C  A  T  T  G
```

Dynamic Programming

- What to do for Bigger String?

SSDSEREEHVKRFRQALDDTGMKVPMATTNLFTHPVFKDGGFTANDRDVRRYALRKTIRNIDLAVELGAETYVAWGGREGAESGGAKDVRDALDRMKEAFDLLGEYVTSQGYDIRFAIEP
KPNEPRGDILLPTVGHALAFIERLERPELYGVNPEVGHEQMAGLNFPHGIAQALWAGKLFHIDLNGQNGIKYDQDLRFGAGDLRAAFWLVDLLESAGYSGPRHFDFKPPRTEDFDGVWAS

- Needleman-Wunsch (1970) provided first automatic method
  - ◊ Dynamic Programming to Find Global Alignment
- Their Test Data (J->Y)
  - ◊ ABCNYRQCLCRPM
    AYCYNRCKCRBP

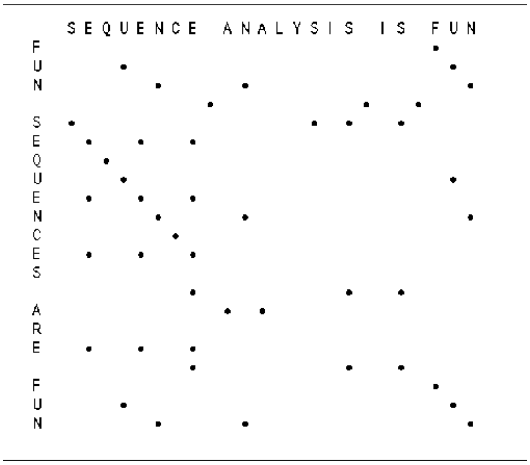
Step 1 -- Make a Dot Plot (Similarity Matrix)

Core

Put 1's where characters are identical.

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

A More Interesting Dot Matrix



(adapted from R Altman)

Step 2 -- Start Computing the Sum Matrix

Core

```
new_value_cell(R,C) <=
  cell(R,C)                        { Old value, either 1 or 0    }
  + Max[
        cell (R+1, C+1),           { Diagonally Down, no gaps    }
        cells(R+1, C+2 to C_max),{ Down a row, making col. gap }
        cells(R+2 to R_max, C+1) { Down a col., making row gap }
       ]
```

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Step 3 -- Keep Going

Core

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					*2*	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						*5*	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Step 4 -- Sum Matrix All Done

Core

Alignment Score is 8 matches.

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						*5*	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	*8*	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Step 5 -- Traceback

Core

Find Best Score (8) and Trace Back

A B C N Y - R Q C L C R - P M
A Y C - Y N R - C K C R B P

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	**8**	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	**7**	6	6	6	4	4	3	3	2	1	0	0
C	6	6	**7**	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	**6**	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	**5**	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	**4**	3	3	1	0	0
K	3	3	3	3	3	3	3	3	**3**	2	1	0	0
C	2	2	3	2	2	2	2	3	2	**3**	1	0	0
R	2	1	1	1	1	2	1	1	1	1	**2**	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	**1**	0

Step 5 -- Traceback

Core

A B C N Y - R Q C L C R - P M
A Y C - Y N R - C K C R B P

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	**8**	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	**7**	6	6	6	4	4	3	3	2	1	0	0
C	6	6	**7**	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	**6**	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	**5**	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	**4**	3	3	1	0	0
K	3	3	3	3	3	3	3	3	**3**	2	1	0	0
C	2	2	3	2	2	2	2	3	2	**3**	1	0	0
R	2	1	1	1	1	2	1	1	1	1	**2**	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	**1**	0

Step 6 -- Alternate Tracebacks

Core

A B C - N Y R Q C L C R - P M
A Y C Y N - R - C K C R B P

Also, Suboptimal Aligments

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
Y	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
Y	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

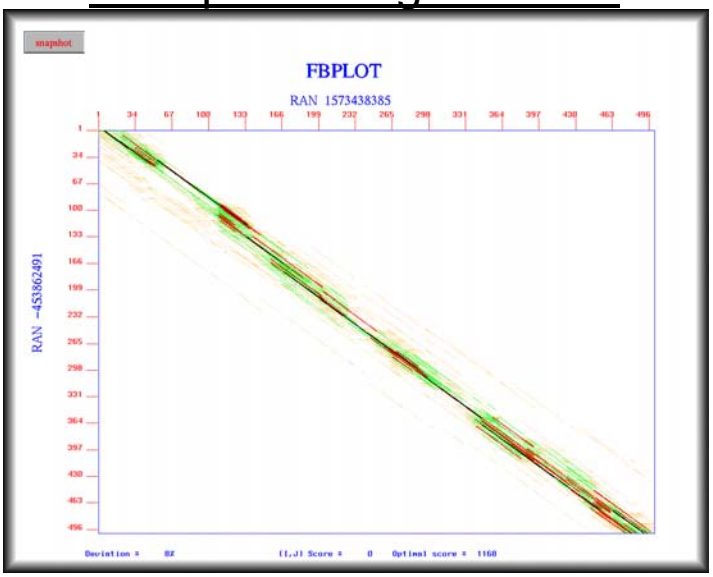
Suboptimal Alignments

```
;
; Random DNA sequence generated using the seed :   -453862491
;
;   500 nucleotides
;
; A:C:G:T =   1 :   1 :   1 :   1
;
RAN  -453862491
AAATGCCAAA TCATACGAAC AGCCGACGAC GGGAGCAACC CAAGTCGCAG TTCGCTTGAG CTAGCGCGCT
CCCACCGGGA TATACACTAA TCATTACAGC AGGTCTCCTG GGCGTACAGA CTAGCTGAAC GCGCTGCGCC
AATTCCAACT TCGGTATGAA GGATCGCCTG CGGTTATCGC TGACTTGAGT AACCAGATCG CTAAGGTTAC
GCTGGGGCAA TGATGGATGT TAACCCCTTA CAGTCTCGGG AGGGACCTTA AGTCGTAATA GATGGCAGCA
TTAATACCTT CGCCGTTAAT ATACCTTTAA TCCGTTCTTG TCAATGCCGT AGCTGCAGTG AGCCTTCTGT
CACGGGCATA CCGCGGGGTA GCTGCAGCAA CCGTAGGCTG AGCATCAAGA AGACAAACAC TCCTCGCCTA
CCCCGGACAT CATATGACCA GGCAGTCTAG GCGCCGTTAG AGTAAGGAGA CCGGGGGGCC GTGATGATAG
ATGGCGTGTT 1
;
; Random DNA sequence generated using the seed :   1573438385
;
;   500 nucleotides
;
; A:C:G:T =   1 :   1 :   1 :   1
;
RAN  1573438385
CCCTCCATCG CCAGTTCCTG AAGACATCTC CGTGACGTGA ACTCTCTCCA GGCATATTAA TCGAAGATCC
CCTGTCGTGA CGCGGATTAC GAGGGGATGG TGCTAATCAC ATTGCGAACA TGTTTCGGTC CAGACTCCAC
CTATGGCATC TTCCGCTATA GGGCACGTAA CTTTCTTCGT GTGGCGGCGC GGCAACTAAA GACGAAAGGA
CCACAACGTG AATAGCCCGT GTCGTGAGGT AAGGGTCCCG GTGCAAGAGT AGAGGAAGTA CGGGAGTACG
TACGGGGCAT GACGCGGGCT GGAATTTCAC ATCGCAGAAC TTATAGGCAG CCGTGTGCCT GAGGCCGCTA
GAACCTTCAA CGCTAACTAG TGATAACTAC CGTGTGAAAG ACCTGGCCCG TTTTGTCCCT GAGACTAATC
GCTAGTTAGG CCCCATTTGT AGCACTCTGG CGCAGACCTC GCAGAGGGAC CGGCCTGACT TTTTCCGGCT
TCCTCTGAGG 1

Parameters: match weight = 10, transition weight = 1, transversion weight = -3
Gap opening penalty = 50   Gap continuation penalty = 1
Run as a local alignment (Smith-Waterman)
```

(courtesy of Michael Zucker)

Suboptimal Alignments II



(courtesy of Michael Zucker)

Core

Gap Penalties

The score at a position can also factor in a penalty for introducing gaps (i. e., not going from i, j to i- 1, j- 1).

Gap penalties are often of linear form:

$$GAP = a + bN$$

GAP is the gap penalty

a = cost of opening a gap

b = cost of extending the gap by one (affine)

N = length of the gap

(Here assume b=0, a=1/2, so GAP = 1/2 regardless of length.)

Step 2 -- Computing the Sum Matrix **with Gaps**

Core

```
new_value_cell(R,C) <=
  cell(R,C)                                  { Old value, either 1 or 0    }
  + Max[
       cell (R+1, C+1),                      { Diagonally Down, no gaps    }
       cells(R+1, C+2 to C_max) - GAP ,{ Down a row, making col. gap }
       cells(R+2 to R_max, C+1) - GAP  { Down a col., making row gap }
      ]
```

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

	A	B	C	N	Y	R	Q	C	L	C	R	P	M
A	1												
Y					1								
C			1					1		1			
Y					1								
N				1									
R						1					1		
C			1					1		1			
K													
C			1					1		1			
R						1					*1.5*	0	0
B	1	*1.5*	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

GAP =1/2

All Steps in Aligning a 4-mer

C R B P

C R P M

- C R P M

C R - P M

	C	R	P	M
C	**1**			
R		**1**		
B				
P			**1**	

	C	R	P	M
C	1			
R		*2*	0	0
B	1	1	0	0
P	0	0	1	0

	C	R	P	M
C	**3**	1	0	0
R	1	**2**	0	0
B	1	1	0	0
P	0	0	**1**	0

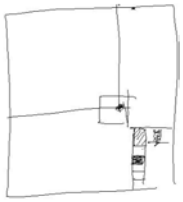
Bottom right hand corner of previous matrices

Key Idea in Dynamic Programming

◊ The best alignment that ends at a given pair of positions (i and j) in the 2 sequences is the score of the best alignment previous to this position PLUS the score for aligning those two positions.

◊ An Example Below

- Aligning R to K does not affect alignment of previous N-terminal residues. Once this is done it is **fixed**. Then go on to align D to E.
- How could this be violated?
  Aligning R to K changes best alignment in box.



```
ACSQRP--LRV-SH  RSENCV
A-SNKPQLVKLMTH  VKDFCV
```

```
ACSQRP--LRV-SH  -R  SENCV
A-SNKPQLVKLMTH  VK  DFCV
```


End of class 2002,10.16
(Bioinfo-2)
[started class in intro]

Similarity (Substitution) Matrix

Core

- Identity Matrix
  - ◊ Match L with L => 1
    Match L with D => 0
    Match L with V => 0??
- S(aa-1,aa-2)
  - ◊ Match L with L => 1
    Match L with D => 0
    Match L with V => .5
- Number of Common Ones
  - ◊ PAM
  - ◊ Blossum
  - ◊ Gonnet

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	8	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	7	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	6	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	10	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	6	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Where do matrices come from?

+ —> More likely than random
0 —> At random base rate
- —> Less likely than random

1 Manually align protein structures (or, more risky, sequences)

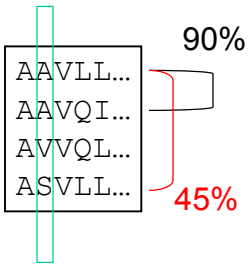
2 Look at frequency of a.a. substitutions at structurally constant sites. -- i.e. pair i-j exchanges

3 Compute log-odds

$S(aa\text{-}1,aa\text{-}2) = \log_2 (freq(O) / freq(E))$

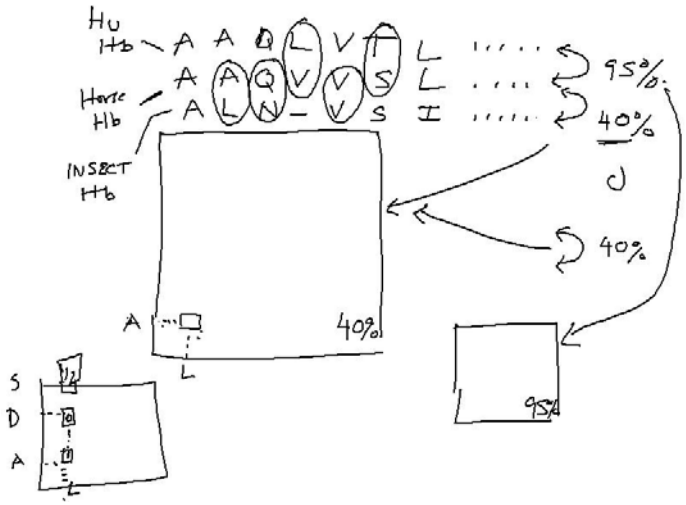
O = observed exchanges,
E = expected exchanges

- odds = freq(observed) / freq(expected)
- Sij = log odds
- freq(expected) = f(i)*f(j)
  = is the chance of getting amino acid i in a column and then having it change to j
- e.g. A-R pair observed only a tenth as often as expected



Core

Relationship of type of substitution to closeness in identity of the sequences in the training alignment



More on this….

To help us understand the knowledge incorporated in amino acid similarity scores we should briefly look at how they are calculated (4). First we compute an amino acid similarity ratio, Rij for every pair of amino acids i and j.

$$R_{ij} = q_{ij} / p_i p_j$$

Where qij is the relative frequency with which amino acids i and j are observed to replace each other in homologous proteins. pi and pj are the frequencies at which amino acids i and j occur in the set of proteins in which the

substitutions are observed. Their product, pipj, is the frequency at which they would be expected replace each other if the replacements were random. If the observed replacement rate is equal to the theoretical replacement

rate, then the ratio is one (Rij = qij / pipj = 1.0). If the replacements are favored during evolution (i.e. a conservative replacement) the ratio will be greater than one and if there is selection against the replacement the ratio

will be less than one.

The similarity reported in the evolutionary-based tables for any pair of amino acids i and j, Sij is the logarithm to the base 2 of this ratio, Rij, although it is often scaled by some constant factor.

$$S_{ij} = \log_2(R_{ij}) = \log_2(q_{ij} / p_i p_j)$$

Scores above zero (Sij > 0.0) indicate that two amino acids replace each other more often during evolution than we would expect if the replacements were random. Likewise, scores below zero indicate that amino acids

replace each other less often than we would expect if the replacements were random. Thus a positive alignment score means that the pattern of identities and substitutions described by an alignment are more likely to result

from previously observed evolutionary processes than to result from random replacements.

Amino Acid Frequencies of Occurrence

	1978	1991
L	0.085	0.091
A	0.087	0.077
G	0.089	0.074
S	0.070	0.069
V	0.065	0.066
E	0.050	0.062
T	0.058	0.059
K	0.081	0.059
I	0.037	0.053
D	0.047	0.052
R	0.041	0.051
P	0.051	0.051
N	0.040	0.043
Q	0.038	0.041
F	0.040	0.040
Y	0.030	0.032
M	0.015	0.024
H	0.034	0.023
C	0.033	0.020
W	0.010	0.014

Principles of Scoring Matrix Construction, in detail

The Dayhoff Matrix: Proteins evolve through a succesion of independent point mutations, that are accepted in a population and subsequently can be observed in the sequence pool. (Dayhoff, M.O. *et al.* (1978) Atlas of Protein Sequence and Structure. Vol. 5, Suppl. 3 National Biomedical Reserach Foundation, Washington D.C. U.S.A).

First step: Pair Exchange Frequencies

A **PAM** (**P**ercent **A**ccepted **M**utation) is one accepted point mutation on the path between two sequences, per 100 residues.

$$f_i = \frac{observations\ of\ i}{observations\ of\ any\ amino\ acid}$$

Principles of Scoring Matrix Construction, in detail #2

Second step: Frequencies of Occurrence

Amino acid frequencies:

	1978	1991
L	0.085	0.091
A	0.087	0.077
G	0.089	0.074
S	0.070	0.069
V	0.065	0.066
E	0.050	0.062
T	0.058	0.059
K	0.081	0.059
I	0.037	0.053
D	0.047	0.052
R	0.041	0.051
P	0.051	0.051
N	0.040	0.043
Q	0.038	0.041
F	0.040	0.040
Y	0.030	0.032
M	0.015	0.024
H	0.034	0.023
C	0.033	0.020
W	0.010	0.014

Third step: Relative Mutabilities

$$m_i = f_i \text{ (number of times } i \text{ is observed to change)}$$

Relative mutabilities of amino acids:

	1978	1991
A	100	100
C	20	44
D	106	86
E	102	77
F	41	51
G	49	50
H	66	91
I	96	103
K	56	72
L	40	54
M	94	93
N	134	104
P	56	58
Q	93	84
R	65	83
S	120	117
T	97	107
V	74	98
W	18	25
Y	41	50

All values are taken relative to alanine, which is arbitrarily set at 100.

Principles of Scoring Matrix Construction, in detail #3

Fourth step: Mutation Probability Matrix

$$M_{ij} = m_j \frac{A_{ij}}{\sum_{i=1}^{20} A_{ij}}$$

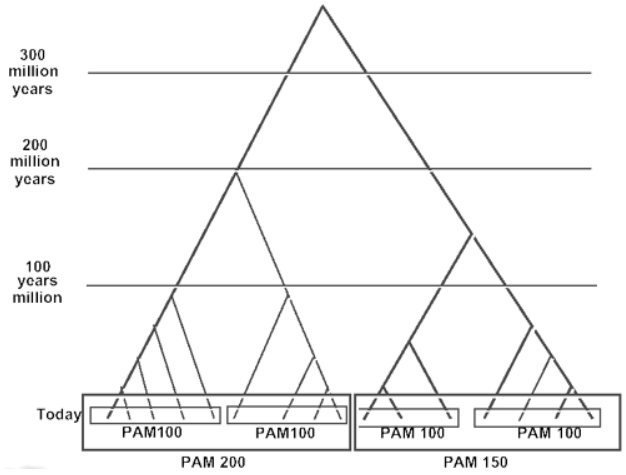
The probability that an amino acid in row ***i*** of the matrix will replace the amino acid in column ***j*** : the mutability of amino acid ***j***, multiplied by the pair exchange frequency for ***ij*** divided by the sum of all pair exchange frequencies for amino acid ***i:***

Last step: the log-odds matrix

log to base 10: a value of +1 would mean that the corresponding pair has been observed 10 times more frequently than expected by chance. The most commonly used matrix is the matrix from the 1978 edition of the Dayhoff atlas, at PAM 250: this is also frequently referred to as the **MDM78 PAM250** matrix.

Different Matrices are Appropriate at Different Evolutionary Distances

Core



(Adapted from D Brutlag, Stanford)

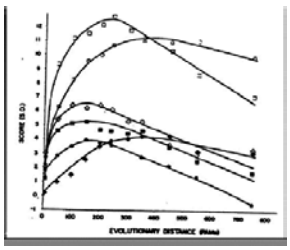
PAM-250 (distant)

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A	Ala	.18																			
R	Arg	-.15	.61																		
N	Asn	.02	0	.20																	
D	Asp	.03	-.13	.21	.39																
C	Cys	-.20	-.36	-.36	-.51	1.19															
Q	Gln	-.04	.13	.08	.16	-.54	.40														
E	Glu	.03	-.11	.14	.34	-.53	.25	.38													
G	Gly	.13	-.26	.03	.06	-.34	-.53	.25	.38												
H	His	-.14	.16	.16	.07	-.34	.29	.07	-.21	.65											
I	Ile	-.05	-.20	-.18	-.24	-.23	-.20	-.20	-.26	-.24	.45										
L	Leu	-.19	-.30	-.29	-.40	-.60	-.18	-.34	-.41	-.21	.24	.59									
K	Lys	-.12	.34	.10	.01	-.54	.07	-.01	-.17	0	-.19	-.29	.47								
M	Met	-.11	-.04	-.17	-.26	-.52	-.10	-.21	-.28	-.21	.22	.37	.04	.64							
F	Phe	-.35	-.45	-.35	-.56	-.43	-.47	-.54	-.48	-.18	.10	.18	-.53	.02	.91						
P	Pro	.11	-.02	-.05	-.10	-.28	.02	-.06	-.05	-.02	-.20	-.25	-.11	-.21	-.46	.59					
S	Ser	.11	-.03	.07	.03	0	-.05	0	.11	-.08	-.14	-.28	-.02	-.16	-.32	.09	.16				
T	Thr	.12	-.09	.04	-.01	-.22	-.08	-.04	0	-.13	.01	-.17	0	-.06	-.31	.03	.13	.26			
W	Trp	-.58	.22	-.42	-.68	-.78	-.48	-.70	-.70	-.28	-.51	-.18	-.35	-.42	.04	-.56	-.25	-.52	1.73		
Y	Tyr	-.35	-.42	-.21	-.43	.03	-.40	-.43	-.52	-.01	-.09	-.09	-.44	-.24	.70	-.49	-.28	-.27	-.02	1.01	
V	Val	.02	-.25	-.17	-.21	-.19	-.19	-.18	-.14	-.22	.37	.19	-.24	.18	-.12	-.12	-.10	.03	-.62	-.25	.43
		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

PAM-78

		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
Cys	C	12																			
Ser	S	0	2																		
Thr	T	-2	1	3																	
Pro	P	-3	1	0	6																
Ala	A	-2	1	1	1	2															
Gly	G	-3	1	0	-1	1	5														
Asn	N	-4	1	0	-1	0	0	2													
Asp	D	-5	0	0	-1	0	1	2	4												
Glu	E	-5	0	0	-1	0	0	1	3	4											
Gln	Q	-5	-1	-1	0	0	-1	1	2	2	4										
His	H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
Arg	R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
Lys	K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
Met	M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
Ile	I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
Leu	L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
Val	V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
Phe	F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Tyr	Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
Trp	W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

Change in Matrix with Ev. Dist.



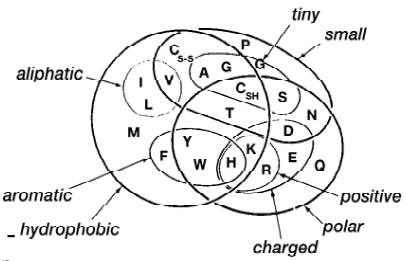
(Adapted from D Brutlag, Stanford)

Other Matrices: How to score the exchange of two amino acids in an alignment?

- Simplest way: the identity matrix

- A very crude model : to use the genetic code matrix,
the number of point mutations necessary to transform one
codon into the other.

Other similarity scoring matrices might be constructed from any property of amino acids that can be quantified -partition coefficients between hydrophobic and hydrophilic phases
-charge
-molecular volume, etc.
Unfortunately, all these biophysical quantities suffer from the fact that they provide only a partial view of the picture - there is no guarantee, that any particular property is a good predictor for conservation of amino acids between related proteins.



(graphic adapted from W Taylor)

The BLOSUM Matrices

Some concepts challenged: Are the evolutionary rates uniform over the whole of the protein sequence?
(No.)

The BLOSUM matrices: Henikoff & Henikoff (Henikoff, S. & Henikoff J.G. (1992) *PNAS* **89**:10915-10919) .

-Use blocks of sequence fragments from different protein families which can be aligned without the introduction of gaps.
Amino acid pair frequencies can be compiled from these blocks

Different evolutionary distances are incorporated into this scheme with a clustering procedure: two sequences that are identical to each other for more than a certain threshold of positions are clustered.

BLOSUM62 is the BLAST default

More sequences are added to the cluster if they are identical to any sequence already in the cluster at the same level.

All sequences within a cluster are then simply averaged.

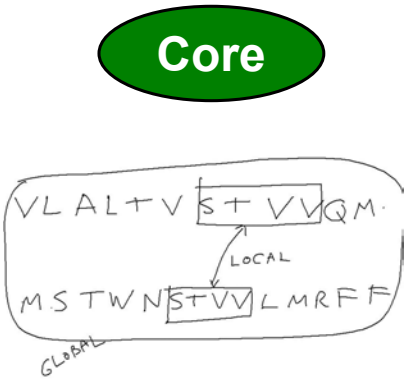
(A consequence of this clustering is that the contribution of closely related sequences to the frequency table is reduced, if the identity requirement is reduced.)

This leads to a series of matrices, analogous to the PAM series of matrices. BLOSUM80: derived at the 80% identity level.

Modifications for Local Alignment

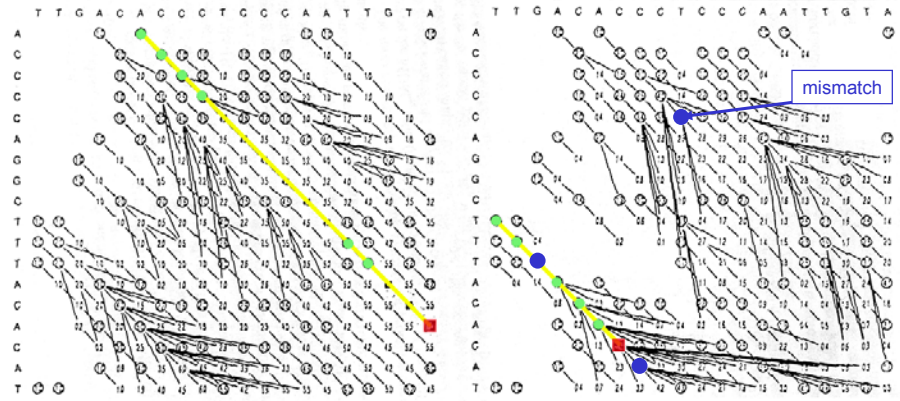
1 The scoring system uses negative scores for mismatches
2 The minimum score for at a matrix element is zero
3 Fine the best score anywhere in the matrix (not just last column or row)

- These three changes cause the algorithm to seek high scoring subsequences, which are not penalized for their global effects (mod. 1), which don't include areas of poor match (mod. 2), and which can occur anywhere (mod. 3)



(Adapted from R Altman)

Global (NW) vs Local (SW)Alignments




```
TTGACACCCTCCCAATTGTA...
||||      ||   |
.....ACCCCAGGCTTTACACAT
1234444444566 67
```

```
T T G A C A C C...
|  | - |  |  |  | -
T T T A C A C A...
1 2 1 2 3 4 5 4
0 0 4 4 4 4 4 8
```

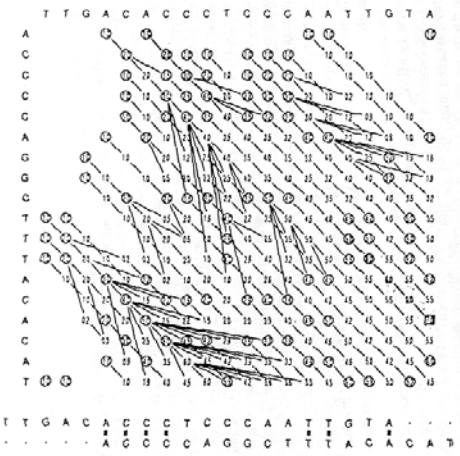
Match Score = +1
Gap-Opening=-1.2, Gap-Extension=-.03
for local alignment Mismatch = -0.6

Adapted from D J States & M S Boguski, "Similarity and Homology," Chapter 3 from Gribskov, M. and Devereux, J. (1992). Sequence Analysis Primer. New York, Oxford University Press. (Page 133)

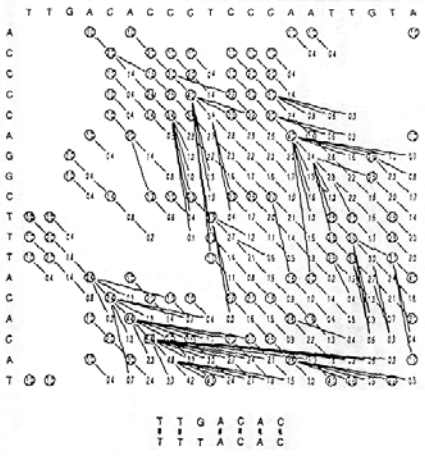
Shows Numbers

Match Score = 1, Gap-Opening=-1.2, Gap-Extension=-.03, for local alignment Mismatch = -0.6

Global



Local



Adapted from D J States & M S Boguski, "Similarity and Homology," Chapter 3 from Gribskov, M. and Devereux, J. (1992). Sequence Analysis Primer. New York, Oxford University Press. (Page 133)

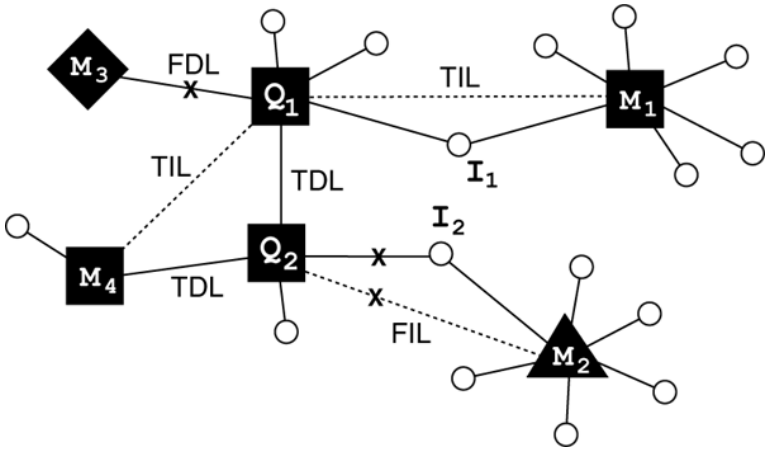
Local vs. Global Alignment

Core

- GLOBAL
  = best alignment of entirety of both sequences
  - ◊ For optimum global alignment, we want best score in the final row or final column
  - ◊ Are these sequences generally the same?
  - ◊ Needleman Wunsch
  - ◊ find alignment in which total score is highest, perhaps at expense of areas of great local similarity
- LOCAL
  = best alignment of segments, without regard to rest of sequence
  - ◊ For optimum local alignment, we want best score anywhere in matrix (will discuss)
  - ◊ Do these two sequences contain high scoring subsequences
  - ◊ Smith Waterman
  - ◊ find alignment in which the highest scoring subsequences are identified, at the expense of the overall score

(Adapted from R Altman)

Transitive Sequence Comparison



Multiple Sequence Alignments

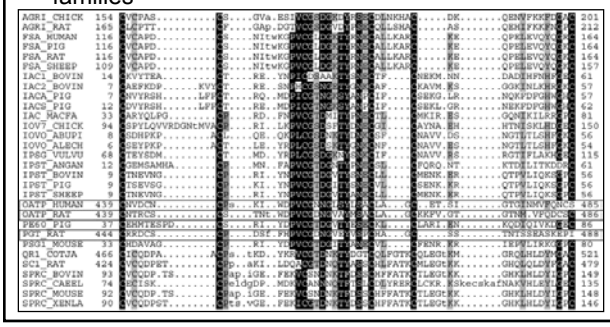
Core

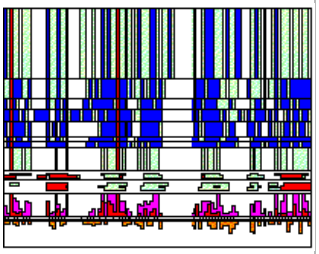
- One of the most essential tools in molecular biology

It is widely used in:

- Phylogenetic analysis
- Prediction of protein secondary/tertiary structure
- Finding diagnostic patterns to characterize protein families
- Detecting new homologies between new genes and established sequence families

- Practically useful methods only since 1987
- Before 1987 they were constructed by hand
- The basic problem: no dynamic programming approach can be used
- First useful approach by D. Sankoff (1987) based on phylogenetics





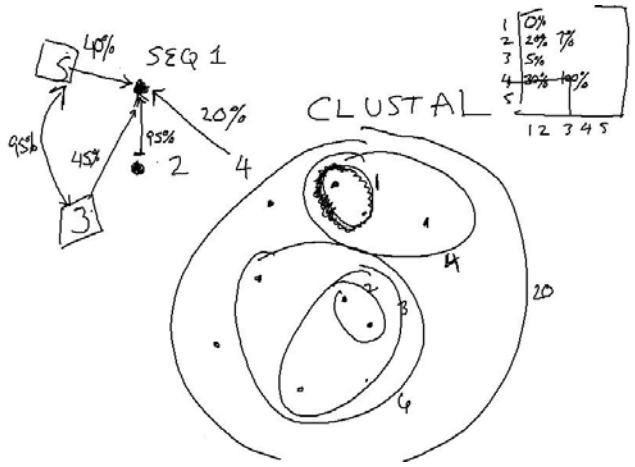
(LEFT, adapted from Sonhammer et al. (1997). "Pfam," Proteins 28:405-20. ABOVE, G Barton AMAS web page)

Progressive Multiple Alignments

```
AGRI_CHICK   154 CVCPAS.........CS....GVa.ESIVCGSDGKDYRSECDLNKHAC.....DK......
AGRI_RAT     165 CLCPTT.........CF....GAp.DGTVCGSDGVDYPSECQLLSHAC.....AS......
FSA_HUMAN    116 CVCAPD.........CS....NItwKGPVCGLDGKTYRNECALLKARC.....KE......
FSA_PIG      116 CVCAPD.........CS....NItwKGPVCGLDGKTYRNECALLKARC.....KE......
FSA_RAT      116 CVCAPD.........CS....NItwKGPVCGLDGKTYRNECALLKARC.....KE......
FSA_SHEEP    109 CVCAPD.........CS....NItwKGPVCGLDGKTYRNECALLKARC.....KE......
IAC1_BOVIN    14 CKVYTEA.........CT....RE..YNPICDSAAKTYSNECTF....CNEKM.NN......
IAC2_BOVIN     7 CAEFKDP......KVYCT....RE..SNPHCGSNGETYGNKCAF....CKAVM.KS......
IACA_PIG       7 CNVYRSH......LFFCT....RQ..MDPICGTNGKSYANPCIF....CSEKG.LR......
IACS_PIG      12 CDVYRSH......LFFCT....RE..MDPICGTNGKSYANPCIF....CSEKL.GR......
IAC_MACFA     33 CARYQLPG.........CP....RD..FNPVCGTDMITYPNECTL....CMKIR.ES......
IOV7_CHICK    94 CSPYLQVVRDGNtMVACP....RI..LKPVCGSDSFTYDNECGI....CAYNA.EH......
IOVO_ABUPI     8 CSDHPKP........ACL....QE..QKPLCGSDNKTYDNKCSF....CNAVV.DS......
IOVO_ALECH     6 CSEYPKP........ACT....LE..YRPLCGSDSKTYGNKCNF....CNAVV.ES......
IPSG_VULVU    68 CTEYSDM.........CT....MD..YRPLCGSDGKNYSNKCIF....CNAVV.RS......
IPST_ANGAN    12 CGEMSAMHA.......CP....MN..FAPVCGTDGNTYPNECSL....CFQRQ.NT......
IPST_BOVIN     9 CTNEVNG.........CP....RI..YNPVCGTDGVTYSNECLL....CMENK.ER......
IPST_PIG       9 CTSEVSG.........CP....KI..YNPVCGTDGITYSNECVL....CSENK.KR......
IPST_SHEEP     9 CTNEVNG.........CP....RI..YNPVCGTDGVTYANECLL....CMENK.ER......
OATP_HUMAN   439 CNVDCN..........CPs...KI..WDPVCGNNGLSYLSACLA...GC..ET.SI......
OATP_RAT     439 CNTRCS..........CS....TNt.WDPVCGNGVAYVSACLA...GCKKFV.GT......
PE60_PIG      37 CEHMTESPD.......CS....RI..YDPVCGTDGVTYPSECKL....CLARI.EN......
PGT_RAT      444 CRRDCS..........CP....DSf.FHPVCGNGVEYVSPCHA...GC.....SS......
PSG1_MOUSE    33 CHDAVAG.........CP....RI..YDPVCGTDGITYANECVL....CFENR.KR......
QR1_COTJA    466 CICQDPA........ACPs..tKD..YKRVCGTDNKTYDGTCQLFGTKCQLEGtKM......
SC1_RAT      424 CVCQDPET........CPp..aKI..LDQACGTDNQTYASSCHLFATKCMLEGtKK......
SPRC_BOVIN    93 CVCQDP.TS.......CPap.iGE..FEKVCSNDNKTFDSSCHFFATKCTLEGtKK......
SPRC_CAEEL    74 CECISK..........CPeldgDP..MDKVCANNNQTFTSLCDLYRERCLCKR.KSkecska
SPRC_MOUSE    92 CVCQDP.TS.......CPap.iGE..FEKVCSNDNKTFDSSCHFFATKCTLEGtKK......
SPR_HUMAN     92 CVCQDP.TS.......CPap.iGE..FEKICGTDNKTYDSSCHFFATKCTLEGtKK......
```

(adapted from Sonhammer et al. (1997) "Pfam," Proteins 28:405-20)

Clustering approaches for multiple sequence alignment



Problems with Progressive Alignments

- Local Minimum Problem
  - Parameter Choice Problem

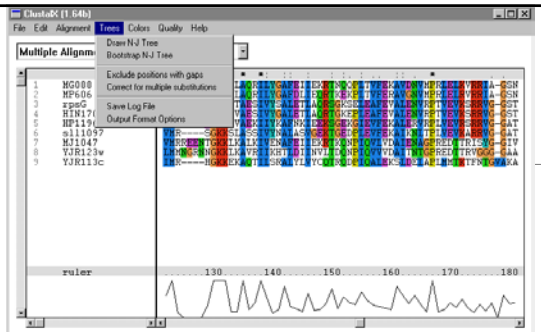
1. Local Minimum Problem

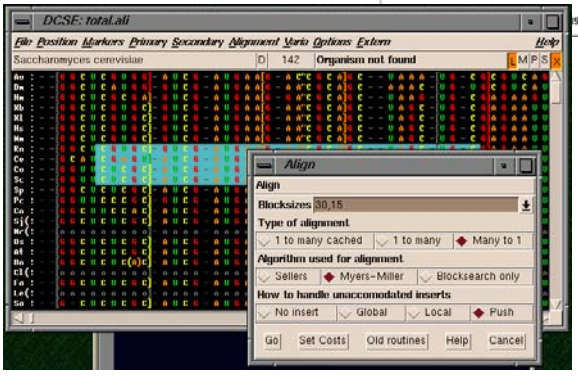
- It stems from greedy nature of alignment (mistakes made early in alignment cannot be corrected later)
- A better tree gives a better alignment (UPGMA neighbour-joining tree method)

2. Parameter Choice Problem

- It stems from using just one set of parameters (and hoping that they will do for all)

Popular Multiple Alignment Programs





Extra

End of class 2002,10.23
(Bioinfo-3)
[after midterm]

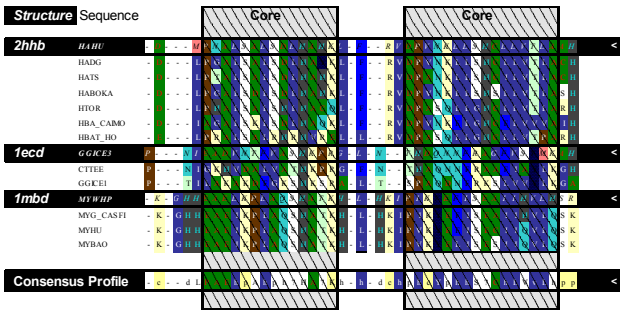
Fuse multiple alignment into:

- **Motif**: a short signature pattern identified in the conserved region of the multiple alignment
- **Profile**: frequency of each amino acid at each position is estimated
- **HMM**: Hidden Markov Model, a generalized profile in rigorous mathematical terms

Profiles
Motifs
HMMs

Core

Can get more sensitive searches with these multiple alignment representations (Run the profile against the DB.)



Profiles

2hhb	**Human Alpha Hemoglobin**	R	V	D	C	V	A	Y	K	
	HAHU	R	V	D	C	V	A	Y	K	**100**
	HADG	R	V	D	C	V	A	Y	K	**89**
	HTOR	R	V	D	C	A	A	Y	Q	**76**
	HBA_CAIMO	R	V	D	P	V	A	Y	K	**73**
	HBAT_HORSE	R	V	D	P	A	A	Y	Q	**62**
1mbd	**Whale Myoglobin**	A	I	C	A	P	A	Y	E	
	MYWHP	A	I	C	A	P	A	Y	E	**100**
	MYG_CASFI	R	I	C	A	P	A	Y	E	**85**
	MYHU	R	I	C	V	C	A	Y	D	**75**
	MYBAO	R	I	C	V	C	A	Y	D	**71**

Eisenberg Profile Freq. A	1	0	0	2	2	9	0	0	↑ Identity
Eisenberg Profile Freq. C	0	0	4	3	2	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Eisenberg Profile Freq. V	0	5	0	2	3	0	0	0	
Eisenberg Profile Freq. Y	0	0	0	0	0	0	9	0	

Consensus = Most Typical A.A.	R	V	D	C	V	A	Y	E
Better Consensus = Freq. Pattern (PCA)	R	iv	cd	š	š	A	Y	µ

š = (A,2V,C,P); µ=(4K,2Q,3E,2D)

Entropy => Sequence Variability	**3**	**7**	**7**	**14**	**14**	**0**	**0**	**14**

Profile : a position-specific scoring matrix composed of 21 columns and N rows (N=length of sequences in multiple alignment)

Core

2hhb	**Human Alpha Hemoglobin**	R	V	D	C	V	A	Y	K	
	HAHU	R	V	D	C	V	A	Y	K	**100**
	HADG	R	V	D	C	V	A	Y	K	**89**
	HTOR	R	V	D	C	A	A	Y	Q	**76**
	HBA_CAIMO	R	V	D	P	V	A	Y	K	**73**
	HBAT_HORSE	R	V	D	P	A	A	Y	Q	**62**
1mbd	**Whale Myoglobin**	A	I	C	A	P	A	Y	E	
	MYWHP	A	I	C	A	P	A	Y	E	**100**
	MYG_CASFI	R	I	C	A	P	A	Y	E	**85**
	MYHU	R	I	C	V	C	A	Y	D	**75**
	MYBAO	R	I	C	V	C	A	Y	D	**71**

Eisenberg Profile Freq. A	1	0	0	2	2	9	0	0	↑ Identity
Eisenberg Profile Freq. C	0	0	4	3	2	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Eisenberg Profile Freq. V	0	5	0	2	3	0	0	0	
Eisenberg Profile Freq. Y	0	0	0	0	0	0	9	0	

Consensus = Most Typical A.A.	R	V	D	C	V	A	Y	E
Better Consensus = Freq. Pattern (PCA)	R	iv	cd	š	š	A	Y	µ

š = (A,2V,C,P); µ=(4K,2Q,3E,2D)

Entropy => Sequence Variability	**3**	**7**	**7**	**14**	**14**	**0**	**0**	**14**

Profiles formula for position M(p,a)

M(p,a) = chance of finding amino acid a at position p

$M_{simp}(p,a)$ = number of times a occurs at p divided by number of sequences

However, what if don't have many sequences in alignment? $M_{simp}(p,a)$ might be baised. Zeros for rare amino acids. Thus:

$M_{cplx}(p,a) = \Sigma_{b=1 \text{ to } 20} M_{simp}(p,b) \times Y(b,a)$

Y(b,a): Dayhoff matrix for *a* and *b* amino acids

$S(p,a) \sim \Sigma_{a=1 \text{ to } 20} M_{simp}(p,a) \ln M_{simp}(p,a)$

Core

2hhb	**Human Alpha Hemoglobin**	R	V	D	C	V	A	Y	K	
	HAHU	R	V	D	C	V	A	Y	K	100
	HADG	R	V	D	C	V	A	Y	K	89
	HTOR	R	V	D	C	A	A	Y	Q	76
	HBA_CAIMO	R	V	D	P	V	A	Y	K	73
	HBAT_HORSE	R	V	D	P	A	A	Y	Q	62
1mbd	**Whale Myoglobin**	A	I	C	A	P	A	Y	E	
	MYWHP	A	I	C	A	P	A	Y	E	100
	MYG_CASFI	R	I	C	A	P	A	Y	E	85
	MYHU	R	I	C	V	C	A	Y	D	75
	MYBAO	R	I	C	V	C	A	Y	D	71

Eisenberg Profile Freq. A	1	0	0	2	2	9	0	0	↑ Identity
Eisenberg Profile Freq. C	0	0	4	3	2	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Eisenberg Profile Freq. V	0	5	0	2	3	0	0	0	
Eisenberg Profile Freq. Y	0	0	0	0	0	0	9	0	

Consensus = Most Typical A.A.	R	V	D	C	V	A	Y	E
Better Consensus = Freq. Pattern (PCA)	R	iv	cd	š	š	A	Y	µ

š = (A,2V,C,P); µ=(4K,2Q,3E,2D)

Entropy => Sequence Variability	3	7	7	14	14	0	0	14

Profiles formula for entropy H(p,a)

$H(p,a) = - \Sigma_{a=1 \text{ to } 20} f(p,a) \log_2 f(p,a)$,
where f(p,a) = frequency of amino acid a occurs at position p ($M_{simp}(p,a)$)

Say column only has one aa (AAAAA):
$H(p,a) = 1 \log_2 1 + 0 \log_2 0 + 0 \log_2 0 + \ldots = 0 + 0 + 0 + \ldots = 0$

Say column is random with all aa equiprobable (ACD..ACD..ACD..):
$H_{rand}(p,a) = .05 \log_2 .05 + .05 \log_2 .05 + \ldots = -.22 + -.22 + \ldots = -4.3$

Say column is random with aa occurring according to probability found in the sequence databases (ACAAAADAADDDDAAAA….):
$H_{db}(a) = - \Sigma_{a=1 \text{ to } 20} F(a) \log_2 F(a)$,
where F(a) is freq. of occurrence of a in DB

Core

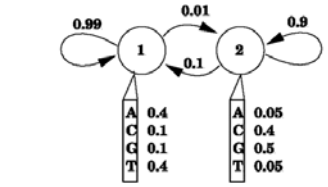
$H_{corrected}(p,a) = H(p,a) - H_{db}(a)$

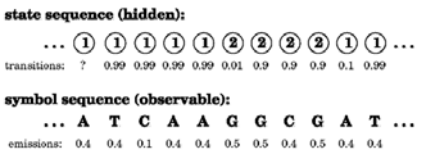
Hidden Markov Model:
- a composition of finite number of states,
- each corresponding to a column in a multiple alignment
- each state emits symbols, according to symbol-emission probabilities

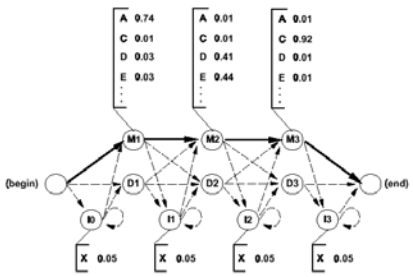
HMMs

Starting from an initial state, a sequence of symbols is generated by moving from state to state until an end state is reached.

Core

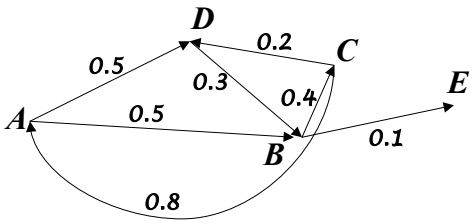






(Figures from Eddy, Curr. Opin. Struct. Biol.)

Markov Models



MM

Path: ***A*** ***D*** ***B*** ***C***

Probability = Init(A)*0.5*0.3*0.4

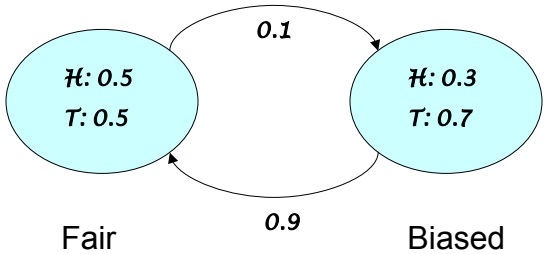
Extra

Hidden Markov models

The path is unknown (hidden): H H T H T T H T T T

Probability = ?

Two coin toss

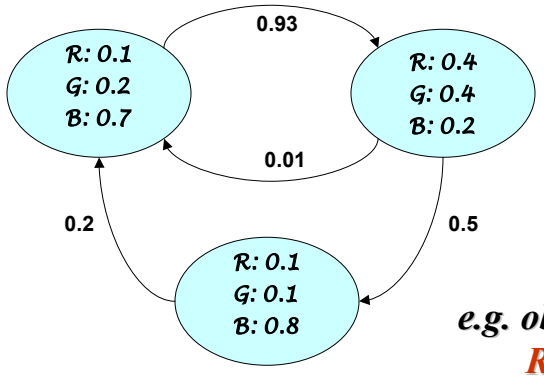


HMM

Extra

Extra

More HMMs



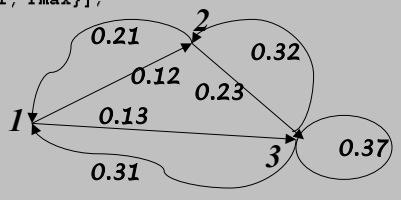
e.g. observed:
RRR

Probability of a given sequence =
Sum probability over ALL paths giving that sequence

Extra

Example:
simple fully interconnected model (N=3)

```
Nmax = 3;
Mmax = 3;
Tmax = 3;
a = Table[0.0, {i, 1, Nmax}, {j, 1, Nmax}];
b = Table[0.0, {i, 1, Nmax}, {j, 1, Mmax}];
init = Table[0.0, {i, 1, Nmax}];
mobs = Table[0.0, {i, 1, Tmax}];
a = {
  {0.75, 0.12, 0.13},
  {0.21, 0.56, 0.23},
  {0.31, 0.32, 0.37}
 };
b = {
  {0.10, 0.45, 0.45},
  {0.20, 0.40, 0.40},
  {0.30, 0.35, 0.35}
 };
init = {0.5, 0.2, 0.3};
mobs = {1, 1, 1};
Print["HMM and observation sequence have been initialized"];
```



Scoring by Brute Force method:

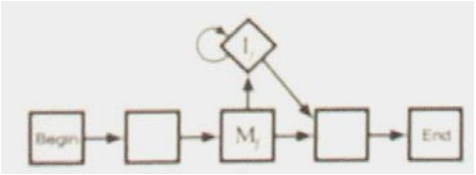
Extra

```
0.5 0.1 0.75 0.1 0.75 0.1      0.2 0.2 0.56 0.2 0.56 0.2
0.5 0.1 0.75 0.1 0.12 0.2      0.2 0.2 0.56 0.2 0.23 0.3
0.5 0.1 0.75 0.1 0.13 0.3      0.2 0.2 0.23 0.3 0.31 0.1
0.5 0.1 0.12 0.2 0.21 0.1      0.2 0.2 0.23 0.3 0.32 0.2
0.5 0.1 0.12 0.2 0.56 0.2      0.2 0.2 0.23 0.3 0.37 0.3
0.5 0.1 0.12 0.2 0.23 0.3      0.3 0.3 0.31 0.1 0.75 0.1
0.5 0.1 0.13 0.3 0.31 0.1      0.3 0.3 0.31 0.1 0.12 0.2
0.5 0.1 0.13 0.3 0.32 0.2      0.3 0.3 0.31 0.1 0.13 0.3
0.5 0.1 0.13 0.3 0.37 0.3      0.3 0.3 0.32 0.2 0.21 0.1
0.2 0.2 0.21 0.1 0.75 0.1      0.3 0.3 0.32 0.2 0.56 0.2
0.2 0.2 0.21 0.1 0.12 0.2      0.3 0.3 0.32 0.2 0.23 0.3
0.2 0.2 0.21 0.1 0.13 0.3      0.3 0.3 0.37 0.3 0.31 0.1
0.2 0.2 0.56 0.2 0.21 0.1      0.3 0.3 0.37 0.3 0.32 0.2
0.2 0.2 0.56 0.2 0.56 0.2      0.3 0.3 0.37 0.3 0.37 0.3
                               Brute Force Method  Score=0.00635752
```


Sequence profile elements

- Insertions:

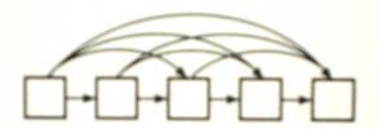
```
C  A  -  T  G
|  |     |  |
C  A  T  T  G
```

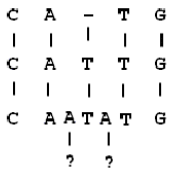


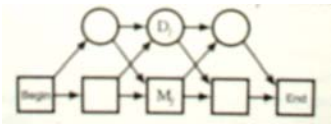
Extra

Sequence profile elements

- Deletions:

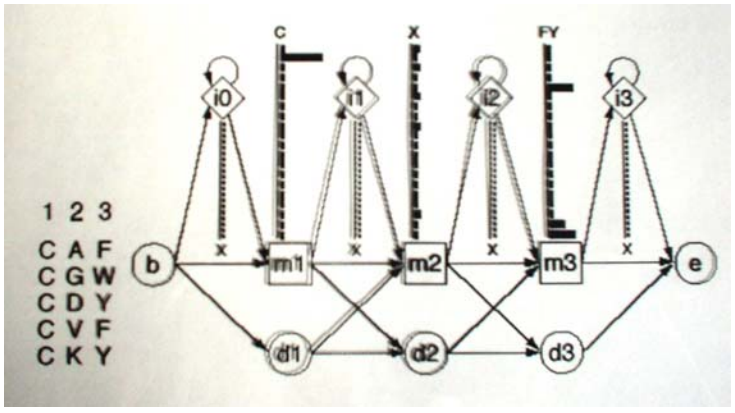






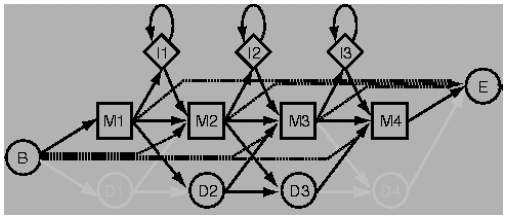
Extra

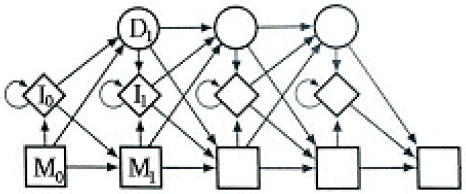
Result: HMM sequence profile



Extra

Different topologies:





Extra

Algorithms

$$P(O,\lambda) = \sum_{i=1}^{N} \alpha_t(i) \qquad \alpha_1(i) = \pi_i\, b_i(O_1)$$

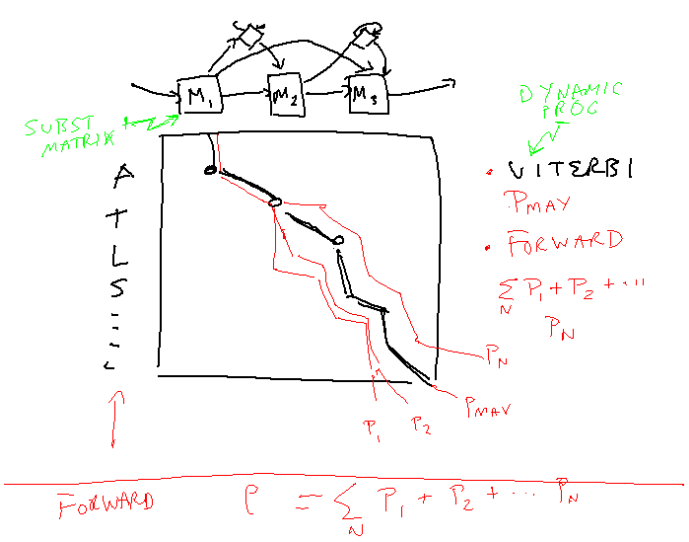
$$\alpha_{t+1}(j) = b_j(O_{t+1}) \left(\sum_{i=1}^{N} \alpha_t(j)\, a_{i,j} \right)$$

Extra

Forward Algorithm – finds probability P that a model λ emits a given sequence O by summing over all paths that emit the sequence the probability of that path

Viterbi Algorithm – finds the most probable path through the model for a given sequence
(both usually just boil down to simple applications of dynamic programming)

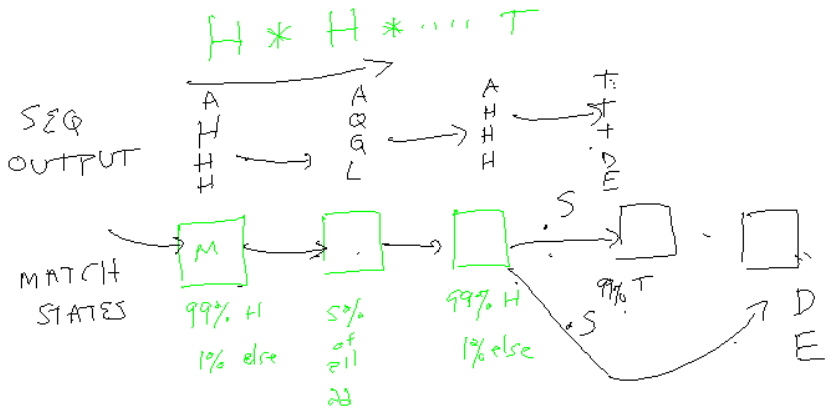
HMM algorithms similar to those in sequence alignment



EM ALG. SEQS → HMM

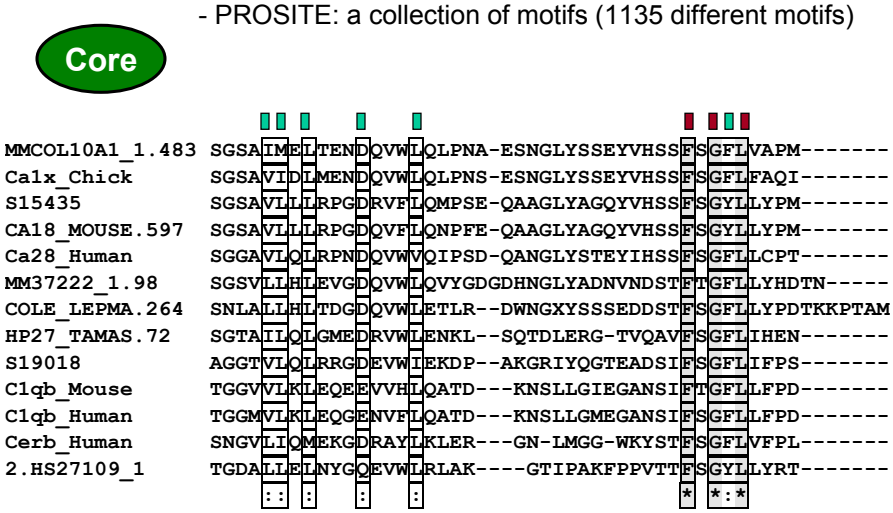
Core

The Hidden Part of HMMs



Motifs

- several proteins are grouped together by similarity searches
- they share a conserved motif
- motif is stringent enough to retrieve the family members from the complete protein database
- PROSITE: a collection of motifs (1135 different motifs)



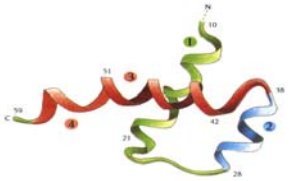
Motifs

· Each element in a pattern is separated from its neighbor by a "-".
· The symbol "x" is used for a position where any amino acid is accepted.
· Ambiguities are indicated by listing the acceptable amino acids for a given position, between brackets "[]".
· Ambiguities are also indicated by listing between a pair of braces "{}" the amino acids that are not accepted at a given position.
· Repetition of an element of the pattern is indicated by with a numerical value or a numerical range between parentheses following that element.

PKC_PHOSPHO_SITE	Protein kinase C phosphorylation site	[ST]-x-[RK]	Post-translational modifications
RGD	Cell attachment sequence	R-G-D	Domains
SOD_CU_ZN_1	Copper/Zinc superoxide dismutase	[GA]-[IMFAT]-H-[LIVF]-H-x(2)-[GP]-[SDG]-x-[STAGDE]	Enzymes_Oxidoreductases
THIOL_PROTEASE_ASN	Eukaryotic thiol (cysteine) proteases active site	[FYCH]-[WI]-[LIVT]-x-[KRQAG]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G-[LFYW]-[LIVMFYG]-x-[LIVMF]	Enzymes_Hydrolases
TNFR_NGFR_1	TNFR/CD27/30/40/95 cysteine-rich region	C-x(4,6)-[FYH]-x(5,10)-C-x(0,2)-C-x(2,3)-C-x(7,11)-C-x(4,6)-[DNEQSKP]-x(2)-C	Receptors

Modules

HMMs, Profiles, Motifs, and Multiple Alignments used to define modules



•Another example of the helix-loop-helix motif is seen within several DNA binding domains including the homeobox proteins which are the master regulators of development

(Figures from Branden & Tooze)

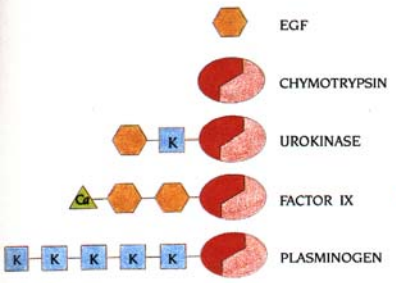


Figure 2.19 Organization of polypeptide chains into domains. Small protein molecules like the epidermal growth factor, EGF, comprise only one domain. Others like the serine proteinase chymotrypsin are arranged in two domains that are both required to form a functional unit (Chapter 15). Many of the proteins that are involved in blood coagulation and fibrinolysis, such as urokinase, factor IX, and plasminogen have long polypeptide chains that comprise different combinations of domains homologous to EGF and serine proteinases and, in addition, calcium-binding domains and Kringle domains.

Domains that are homologous to the epidermal growth factor, EGF, which is a small polypeptide chain of 53 amino acids;

Serine proteinase domains that are homologous to chymotrypsin, which has about 245 amino acids arranged in two domains;

K Kringle domains that have a characteristic pattern of three internal disulphide bridges within a region of about 85 amino acid residues;

Ca Calcium-binding domain (see Figure 2.13).

•Several motifs (β-sheet, beta-alpha-beta, helix-loop-helix) combine to form a compact globular structure termed a domain or tertiary structure

•A domain is defined as a polypeptide chain or part of a chain that can independently fold into a stable tertiary structure

•Domains are also units of function (DNA binding domain, antigen binding domain, ATPase domain, etc.)

C1Q - Example

Extra

```
Ca28_Human
ELSAHATPAFTAVLTSPLPASGMPVKFDRTLYNGHSGYNPATGIFTCPVGGVYYFAYHVH
VKGTNVWVALYKNNVPATYTYDEYKKGYLDQASGGAVLQLRPNDQVWVQIPSDQANGLYS
TEYIHSSFSGFLLCPT
C1qb_Human
DYKATQKIAFSATRTINVPLRRDQTIRFDHVITNMNNNYEPRSGKFTCKVPGLYYFTYHA
SSRGNLCVNLMRGRERAQKVVTFCDYAYNTFQVTTGGMVLKLEQGENVFLQATDKNSLLG
MEGANSIFSGFLLFPD
Cerb_Human
VRSGSAKVAFSAIRSTNHEPSEMSNRTMIIYFDQVLVNIGNNFDSERSTFIAPRKGIYSF
NFHVVKVYNRQTIQVSLMLNGWPVISAFAGDQDVTREAASNGVLIQMEKGDRAYLKLERG
NLMGGWKYSTFSGFLVFPL
COLE_LEPMA.264
RGPKGPPGESVEQIRSAFSVGLFPSRSFPPPSLPVKFDKVFYNGEGHWDPTLNKFNVTYP
GVYLFSYHITVRNRPVRAALVVNGVRKLRTRDSLYGQDIDQASNLALLHLTDGDQVWLET
LRDWNGKYSSSEDDSTFSGFLLYPDTKKPTAM
HP27_TAMAS.72
GPPGPPGMTVNCHSKGTSAFAVKANELPPAPSQPVIFKEALHDAQGHFDLATGVFTCPVP
GLYQFGFHIEAVQRAVKVSLMRNGTQVMEREAEAQDGYEHISGTAILQLGMEDRVWLENK
LSQTDLERGTVQAVFSGFLIHEN
HSUPST2_1.95
GIQGRKGEPGEGAYVYRSAFSVGLETYVTIPNMPIRFTKIFYNQQNHYDGSTGKFHCNIP
GLYYFAYHITVYMKDVKVSLFKKDKAMLFTYDQYQENNVDQASGSVLLHLEVGDQVWLQV
YGEGERNGLYADNDNDSTFTGFLLYHDTN
2.HS27109_1
ENALAPDFSKGSYRYAPMVAFFASHTYGMTIPGPILFNNLDVNYGASYTPRTGKFRIPYL
GVYVFKYTIESFSAHISGFLVVDGIDKLAFESENINSEIHCDRVLTGDALLELNYGQEVW
LRLAKGTIPAKFPPVTTFSGYLLYRT
4.YQCC_BACSU
VVHGWTPWQKISGFAHANIGTTGVQYLKKIDHTKIAFNRVIKDSHNAFDTKNNRFIAPND
GMYLIGASIYTLNYTSYINFHLKVYLNGKAYKTLHHVRGDFQEKDNGMNLGLNGNATVPM
NKGDYVEIWCYCNYGGDETLKRAVDDKNGVFNFFD
5.BSPBSXSE_25
ADSGWTAWQKISGFAHANIGTTGRQALIKGENNKIKYNRIIKDSHKLFDTKNNRFVASHA
GMHLVSASLYIENTERYSNFELYVYVNGTKYKLMNQFRMPTPSNNSDNEFNATVTGSVTV
PLDAGDYVEIYVYVGYSGDVTRYVTDSNGALNYFD
```

```
MMCOL10A1_1.483       SGMPLVSANHGVTG-------MPVSAFTVILS--KAYPA---VGCPHPIYEILYNRQQHY
Ca1x_Chick            ----------ALTG-------MPVSAFTVILS--KAYPG---ATVPIKFDKILYNRQQHY
S15435                ----------GGPA-------YEMPAFTAELT--APFPP---VGGPVKFNKLLYNGRQNY
CA18_MOUSE.597        HAYAGKKGKHGGPA-------YEMPAFTAELT--VPFPP---VGAPVKFDKLLYNGRQNY
Ca28_Human            ----------ELSA-------HATPAFTAVLT--SPLPA---SGMPVKFDRTLYNGHSGY
MM37222_1.98          ----GTPGRKGEPGE---AAYMYRSAFSVGLETRVTVP-----NVPIRFTKIFYNQQNHY
COLE_LEPMA.264        ------RGPKGPPGE---SVEQIRSAFSVGLFPSRSFPP---PSLPVKFDKVFYNGEGHW
HP27_TAMAS.72         -------GPPGPPGMTVNCHSKGTSAFAVKAN--ELPPA---PSQPVIFKEALHDAQGHF
S19018                ----------NIRD-------QPRPAFSAIRQ---NPMT---LGNVVIFDKVLTNQESPY
C1qb_Mouse            --------------D---YRATQKVAFSALRTINSPLR----PNQVIRFEKVITNANENY
C1qb_Human            --------------D---YKATQKIAFSATRTINVPLR----RDQTIRFDHVITNMNNNY
Cerb_Human            --------------V---RSGSAKVAFSAIRSTNHEPSEMSNRTMIIYFDQVLVNIGNNF
2.HS27109_1           ---ENALAPDFSKGS---YRYAPMVAFFASHTYGMTIP------GPILFNNLDVNYGASY
                                                .* .                 :     :

MMCOL10A1_1.483       DPRSGIFTCKIPGIYYFSYHVHVKGT--HVWVGLYKNGTP-TMYTY---DEYSKGYLDTA
Ca1x_Chick            DPRTGIFTCRIPGLYYFSYHVHAKGT--NVWVALYKNGSP-VMYTY---DEYQKGYLDQA
S15435                NPQTGIFTCEVPGVYYFAYHVHCKGG--NVWVALFKNNEP-VMYTY---DEYKKGFLDQA
CA18_MOUSE.597        NPQTGIFTCEVPGVYYFAYHVHCKGG--NVWVALFKNNEP-MMYTY---DEYKKGFLDQA
Ca28_Human            NPATGIFTCPVGGVYYFAYHVHVKGT--NVWVALYKNNVP-ATYTY---DEYKKGYLDQA
MM37222_1.98          DGSTGKFYCNIPGLYYFSYHITVYMK--DVKVSLFKKDKA-VLFTY---DQYQEKNVDQA
COLE_LEPMA.264        DPTLNKFNVTYPGVYLFSYHITVRNR--PVRAALVVNGVR-KLRTR---DSLYGQDIDQA
HP27_TAMAS.72         DLATGVFTCPVPGLYQFGFHIEAVQR--AVKVSLMRNGTQ-VMERE---AEAQDG-YEHI
S19018                QNHTGRFICAVPGFYYFNFQVISKWD--LCLFIKSSSGGQ-PRDSLSFSNTNNKGLFQVL
C1qb_Mouse            EPRNGKFTCKVPGLYYFTYHASSRGN---LCVNLVRGRDRDSMQKVVTFCDYAQNTFQVT
C1qb_Human            EPRSGKFTCKVPGLYYFTYHASSRGN---LCVNLMRGRER--AQKVVTFCDYAYNTFQVT
Cerb_Human            DSERSTFIAPRKGIYSFNFHVVKVYNRQTIQVSLMLNGWP----VISAFAGDQDVTREAA
2.HS27109_1           TPRTGKFRIPYLGVYVFKYTIESFSA--HISGFLVVDGIDKLAFESEN-INSEIHCDRVL
                          . *     * * * :

MMCOL10A1_1.483       SGSAIMELTENDQVWLQLPNA-ESNGLYSSEYVHSSFSGFLVAPM-------
Ca1x_Chick            SGSAVIDLMENDQVWLQLPNS-ESNGLYSSEYVHSSFSGFLFAQI-------
S15435                SGSAVLLLRPGDRVFLQMPSE-QAAGLYAGQYVHSSFSGYLLYPM-------
CA18_MOUSE.597        SGSAVLLLRPGDQVFLQNPFE-QAAGLYAGQYVHSSFSGYLLYPM-------
Ca28_Human            SGGAVLQLRPNDQVWVQIPSD-QANGLYSTEYIHSSFSGFLLCPT-------
MM37222_1.98          SGSVLLHLEVGDQVWLQVYGDGDHNGLYADNVNDSTFTGFLLYHDTN-----
COLE_LEPMA.264        SNLALLHLTDGDQVWLETLR--DWNGXYSSSEDDSTFSGFLLYPDTKKPTAM
HP27_TAMAS.72         SGTAILQLGMEDRVWLENKL--SQTDLERG-TVQAVFSGFLIHEN-------
S19018                AGGTVLQLRRGDEVWIEKDP--AKGRIYQGTEADSIFSGFLIFPS-------
C1qb_Mouse            TGGVVLKLEQEEVVHLQATD---KNSLLGIEGANSIFTGFLLFPD-------
C1qb_Human            TGGMVLKLEQGENVFLQATD---KNSLLGMEGANSIFSGFLLFPD-------
Cerb_Human            SNGVLIQMEKGDRAYLKLER---GN-LMGG-WKYSTFSGFLVFPL-------
2.HS27109_1           TGDALLELNYGQEVWLRLAK----GTIPAKFPPVTTFSGYLLYRT-------
                       .  :: :   : . :                    * *:*.
```

Extra

Clustal Alignment

EGF Profile Generated from SEARCHWISE

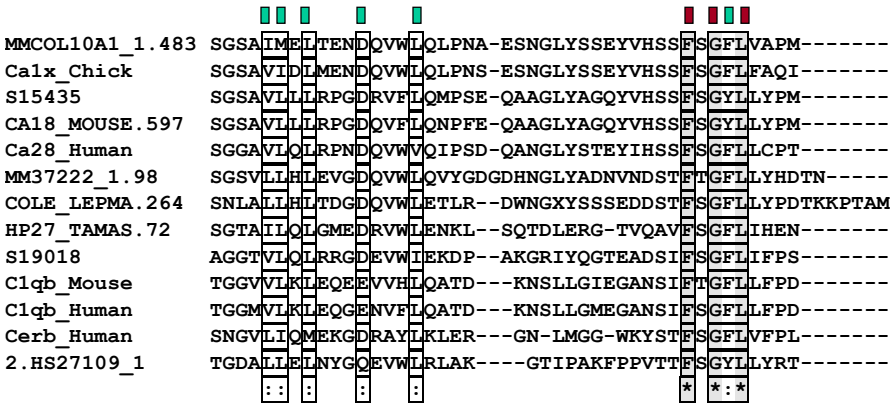
Extra

Cons. Cys

Cons	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Gap
V	-1	-2	-9	-5	-13	-18	-2	-5	-2	-7	-4	-3	-5	-1	-3	0	0	-1	-24	-10	100
D	0	-14	-1	-1	-16	-10	0	-12	0	-13	-8	1	-3	0	-2	0	0	-8	-26	-9	100
V	0	-13	-9	-7	-15	-10	-6	-5	-5	-7	-5	-6	-4	-4	-6	-1	0	-1	-27	-14	100
D	0	-20	18	11	-34	0	4	-26	7	-27	-20	15	0	7	4	6	2	-19	-38	-21	100
P	3	-18	1	3	-26	-9	-5	-14	-1	-14	-12	-1	12	1	-4	2	0	-9	-37	-22	100
C	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	100
A	2	-7	-2	-2	-21	-5	-4	-12	-2	-13	-9	0	-1	0	-3	2	1	-7	-30	-17	100
s	2	-12	3	2	-25	0	0	-18	0	-18	-13	4	3	1	-1	7	4	-12	-30	-16	25
n	-1	-15	4	4	-19	-7	3	-16	2	-16	-10	7	-6	3	0	2	0	-11	-23	-10	25
p	0	-18	-7	-6	-17	-11	0	-17	-5	-15	-14	-5	28	-2	-5	0	-1	-13	-26	-9	25
c	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	25
L	-5	-14	-17	-9	0	-25	-5	4	-5	8	8	-12	-14	-1	-5	-7	-5	2	-15	-5	100
N	-4	-16	12	5	-20	0	24	-24	5	-25	-18	25	-10	6	2	4	1	-19	-26	-2	100
g	1	-16	7	1	-35	29	0	-31	-1	-31	-23	12	-10	0	-1	4	-3	-23	-32	-23	50
G	6	-17	0	-7	-49	59	-13	-41	-10	-41	-32	3	-14	-9	-9	5	-9	-29	-39	-38	100
T	3	-10	0	2	-21	-12	-3	-5	1	-11	-5	1	-4	1	-1	6	11	0	-33	-18	100
C	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	100
I	-6	-13	-19	-11	0	-28	-5	8	-4	6	8	-12	-17	-4	-5	-9	-4	6	-12	-1	100
d	-4	-19	8	6	-15	-13	5	-17	0	-16	-12	5	-9	2	-2	-1	-1	-13	-24	-5	31
i	0	-6	-8	-6	-4	-11	-5	3	-5	1	2	-5	-8	-4	-6	-2	0	4	-14	-6	31
g	1	-13	0	0	-20	-3	-3	-12	-3	-13	-8	0	-7	0	-5	2	0	-7	-29	-16	31
L	-5	-11	-20	-14	0	-23	-9	9	-11	8	7	-14	-17	-9	-14	-8	-4	7	-17	-5	100
E	0	-20	14	10	-33	5	0	-25	2	-26	-19	11	-9	4	0	3	0	-19	-34	-22	100
S	3	-13	4	3	-28	3	0	-18	2	-20	-13	6	-6	3	1	6	3	-12	-32	-20	100
Y	-14	-9	-25	-22	31	-34	10	-5	-17	0	-1	-14	-13	-13	-15	-14	-13	-7	17	44	100
T	0	-10	-6	-1	-11	-16	-2	-7	-1	-9	-5	-3	-9	0	-1	1	3	-4	-16	-8	100
C	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	100
R	0	-13	0	2	-19	-11	1	-12	4	-13	-8	3	-8	4	5	1	1	-8	-23	-13	100
C	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	100
P	0	-14	-8	-4	-15	-17	0	-7	-1	-7	-5	-4	6	0	-2	0	1	-3	-26	-10	100
P	1	-18	-3	0	-24	-13	-3	-12	1	-13	-10	-2	15	2	0	2	1	-8	-33	-19	100
G	4	-19	3	-4	-48	53	-11	-40	-7	-40	-31	5	-13	-7	-7	4	-7	-29	-39	-36	100
y	-22	-6	-35	-31	55	-43	11	-1	-25	6	4	-21	-34	-20	-21	-22	-20	-7	43	63	50
S	1	-9	-3	-1	-14	-7	0	-10	-2	-12	-7	0	-7	0	-4	4	4	-5	-24	-9	100
G	5	-20	1	-8	-52	66	-14	-45	-11	-44	-35	4	-16	-10	-10	4	-11	-33	-40	-40	100
E	2	-20	10	12	-31	-7	0	-19	6	-20	-15	5	4	7	2	4	2	-13	-38	-22	100
R	-5	-17	0	1	-16	-13	8	-16	9	-16	-11	5	-11	7	15	-1	-1	-13	-18	-6	100
C	5	115	-32	-30	-8	-20	-13	-11	-28	-15	-9	-18	-31	-24	-22	1	-5	0	-10	-5	100
E	0	-26	20	25	-34	-5	6	-25	10	-25	-17	9	-4	16	5	3	0	-18	-38	-23	100
T	-4	-11	-13	-8	-1	-21	2	0	-4	-1	0	-6	-14	-3	-5	-4	0	0	-15	0	100
D	0	-18	5	4	-24	-11	-1	-11	2	-14	-9	1	-6	2	0	0	0	-6	-34	-18	100
I	0	-10	-2	-1	-17	-14	-3	-4	-1	-9	-4	0	-11	0	-4	0	2	-1	-29	-14	100
D	-4	-15	-1	-2	-13	-16	-3	-8	-5	-6	-4	-1	-7	-2	-7	-3	-2	-6	-27	-12	100

Motifs

- several proteins are grouped together by similarity searches
- they share a conserved motif
- motif is stringent enough to retrieve the family members from the complete protein database
- PROSITE: a collection of motifs (1135 different motifs)

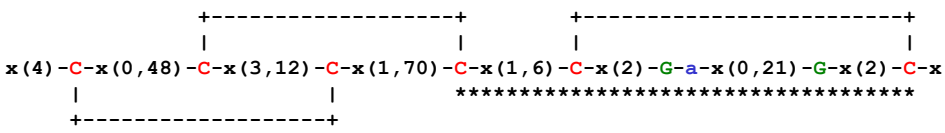


Prosite Pattern -- EGF like pattern

A sequence of about thirty to forty amino-acid residues long found in the sequence of epidermal growth factor (EGF) has been shown [1 to 6] to be present, in a more or less conserved form, in a large number of other, mostly animal proteins. The proteins currently known to contain one or more copies of an EGF-like pattern are listed below.

- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation.
- Caenorhabditis elegans developmental proteins lin-12 (13 copies) and glp-1 (10 copies).
- Calcium-dependent serine proteinase (CASP) which degrades the extracellular matrix proteins type ….
- Cell surface antigen 114/A10 (3 copies).
- Cell surface glycoprotein complex transmembrane subunit .
- Coagulation associated proteins C, Z (2 copies) and S (4 copies).
- Coagulation factors VII, IX, X and XII (2 copies).
- Complement C1r/C1s components (1 copy).
- Complement-activating component of Ra-reactive factor (RARF) (1 copy).
- Complement components C6, C7, C8 alpha and beta chains, and C9 (1 copy).
- Epidermal growth factor precursor (7-9 copies).

Extra



'C': conserved cysteine involved in a disulfide bond.
'G': often conserved glycine
'a': often conserved aromatic amino acid
'*': position of both patterns.
'x': any residue
-Consensus pattern: C-x-C-x(5)-G-x(2)-C
[The 3 C's are involved in disulfide bonds]

http://www.expasy.ch/sprot/prosite.html

The Score

$$S = \sum_{i,j} S(i,j) - nG$$

Core

S = Total Score

S(i,j) = similarity matrix score for aligning i and j

Sum is carried out over all aligned i and j

n = number of gaps (assuming no gap ext. penalty)

G = gap penalty

Simplest score (for identity matrix) is S = # matches

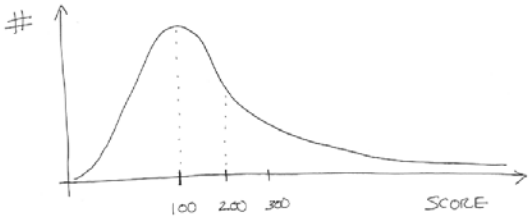
What does a Score of 10 mean? What is the Right Cutoff?

Score in Context of Other Scores

- How does Score Rank Relative to all the Other Possible Scores

  Core

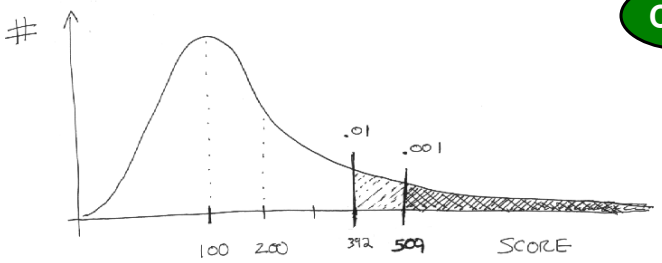
  - ◊ P-value
  - ◊ Percentile Test Score Rank
- All-vs-All comparison of the Database (100K x 100K)
  - ◊ Graph Distribution of Scores
  - ◊ ~10^{10} scores much smaller number of true positives
  - ◊ N dependence



P-value in Sequence Matching

- P(s > S) = .01
  - ◊ P-value of .01 occurs at score threshold S (392 below) where score s from random comparison is greater than this threshold 1% of the time
- Likewise for P=.001 and so on.

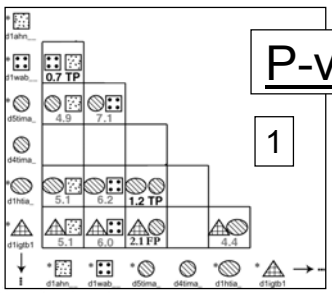
Core



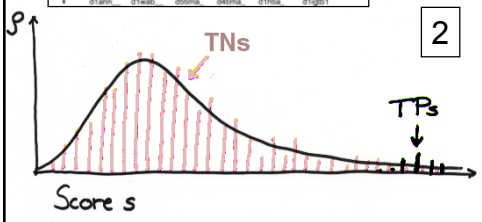
P-values

1

Core

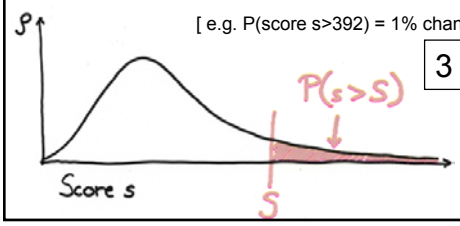


2



[e.g. P(score s>392) = 1% chance]

3

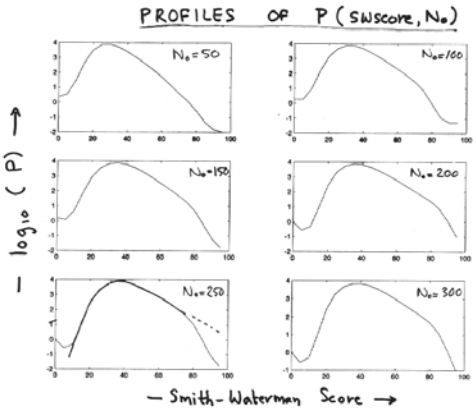


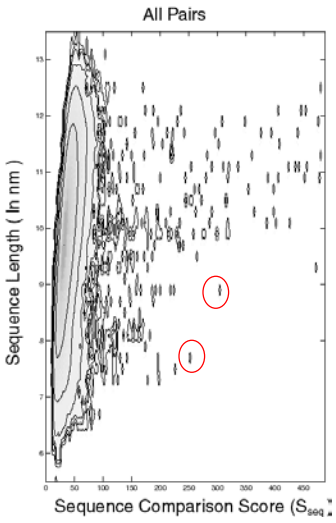
- Significance Statistics
  - ◊ For sequences, originally used in Blast (Karlin-Altschul). Then in FASTA, &c.
  - ◊ Extrapolated Percentile Rank: How does a Score Rank Relative to all Other Scores?
- Our Strategy: Fit to Observed Distribution
  1) All-vs-All comparison
  2) Graph Distribution of Scores in 2D (N dependence); 1K x 1K families -> ~1M scores; ~2K included TPs
  3) Fit a function ρ(S) to TN distribution (TNs from scop); Integrating ρ gives P(s>S), the CDF, chance of getting a score better than threshold S randomly
  4) Use same formalism for sequence & structure

What Distribution Really Looks Like

- N Dependence
- True Positives

Extra





End of class 2002,10.28
(Bioinfo-5)
[..]

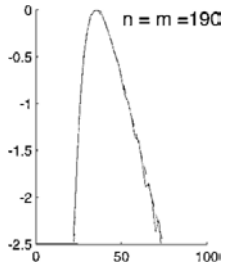
EVD Fits

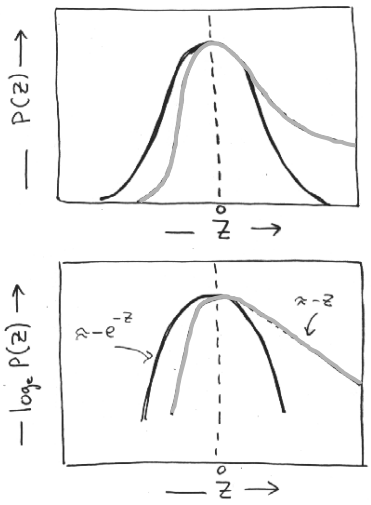
$$\rho(z) = \exp\left(-z - e^{-z}\right)$$

$$\left(\ln \rho(z) = -z - e^{-z}\right)$$

- Reasonable as Dyn. Prog. maximizes over pseudo-random variables
- EVD is **Max**(indep. random variables);
- Normal is **Sum**(indep. random variables)

Observed





$\rho(z) = \exp(-z^2)$

$\ln \rho(z) = -z^2$

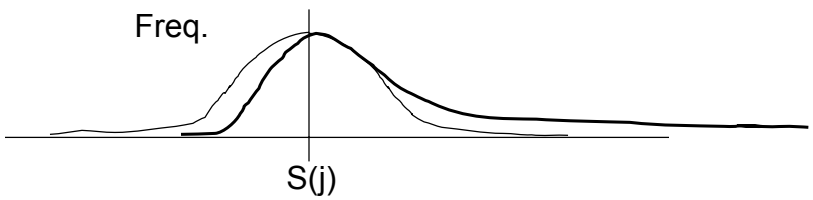
Extreme Value Distribution (EVD, long-tailed) fits the observed distributions best. The corresponding formula for the P-value:

$$P(z > Z) = \int \rho(z)dz = 1 - \exp(-e^{-Z})$$

Core

Extreme Value vs. Gaussian

- X = set of random numbers
  Each set indexed by j
  - ◊ j=1: 1,4,9,1,3,1
  - ◊ j=2: 2,7,3,11,22,1,22
- Gaussian S(j) = $\Sigma_j X_i$ [central limit]
- EVD S(j) = max(X_i)



EVD #2

3 Free Parm. fit to EVD involving: ***a, b,σ***.
These are the only difference betw. sequence and structure.

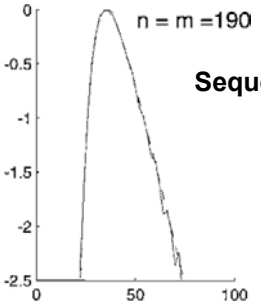
$$Z = \frac{S - (a \ln N + b)}{\sigma}$$

$$S = \sum_{i,j} M(i,j) - G$$

$$\rho(z) = \exp\left(-z - e^{-z}\right)$$

N, G, M also defined differently for sequence and structure.
N = number of residues matched.
G = total gap penalty.
M(i,j) = similarity matrix
(Blossum for seq. or $M_{str}(i,j)$ struc.)

Extra

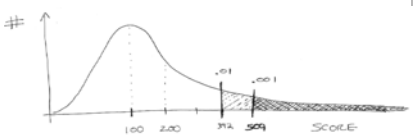


Explicit Form of the P-value in terms of Extreme Value Distribution

F(s) = E.V.D of scores
F(s) = exp(-Z(s) - exp(-Z(s)))

Z(s) = s/A + ln(NM) + B
= (s' - L)/W
s = Score from random S-W Alignment
L = most common one (mode)
W = width parameter (like SD)
N & M are lengths of 2 seq.
A & B are fit parameters

P(s>S) = CDF = integral[F(s)]
P(s>S) = 1 - exp(-exp(-Z(s)))
Given Score Threshold S (1%),
P (s > S) is the chance that a given random score **s** is greater than the threshold

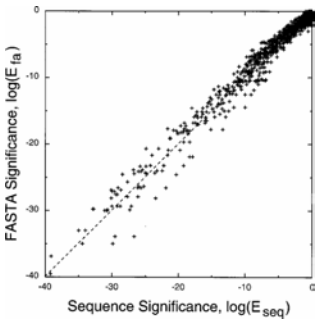


Extra

Use Sequence Scores to Validate

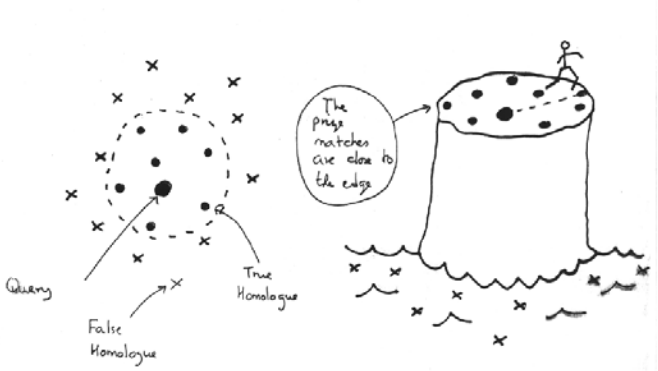
- Sequence P-value perfectly tracks FASTA e-value
  - ◊ Validates approach
  - ◊ Added Benefit: allows computation of an e-value without doing a db run
- Significance computation can be applied to **any** exisiting sequence or structure alignment

Extra



Objective is to Find Distant Homologues

- Score (Significance) Threshold
- Maximize Coverage with an Acceptable Error Rate



(graphic adapted from M Levitt)

Coverage v Error Rate (ROC Graph)

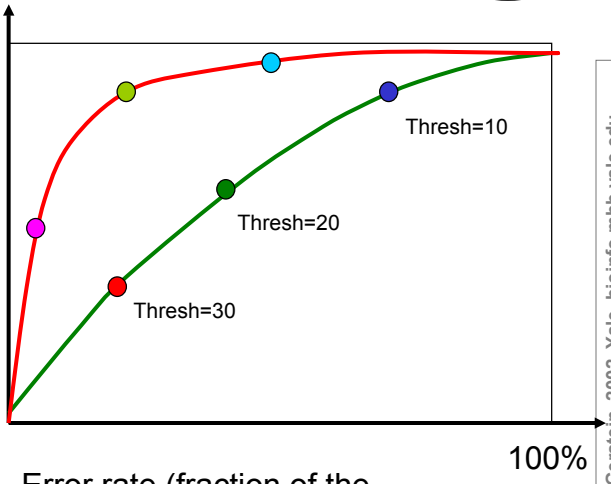
Core

Coverage (roughly, fraction of sequences that one confidently "says something" about)

[sensitivity=tp/p=tp/(tp+fn)]

Different score thresholds

Two "methods" (red is more effective)

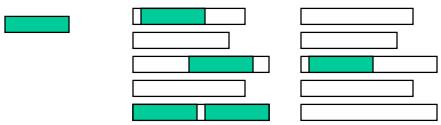


Error rate (fraction of the "statements" that are false positives)

[Specificity = tn/n =tn/(tn+fp)]
error rate = 1-specificity = fp/n

Significance Depends on Database Size

- The Significance of Similarity Scores Decreases with Database Growth
  - ◊ The score between any pair of sequence pair is constant
  - ◊ The number of database entries grows exponentially
  - ◊ The number of nonhomologous entries >> homologous entries
  - ◊ Greater sensitivity is required to detect homologies
    Greater s
- Score of 100 might rank as best in database of 1000 but only in top-100 of database of 1000000



Low-Complexity Regions

Core

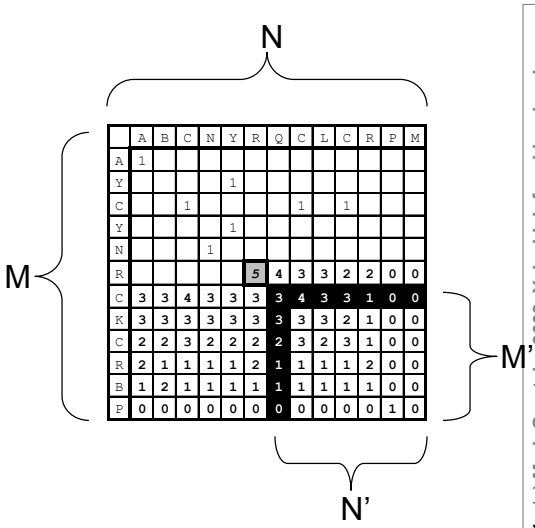
- Low Complexity Regions
  - ◊ Different Statistics for matching
    AAATTTAAATTTAAATTTAAATTTAAATTT
    than
    ACSQRPLRVSHRSENCVASNKPQLVKLMTHVKDFCV
  - ◊ Automatic Programs Screen These Out (SEG)
  - ◊ Identify through computation of sequence entropy in a window of a given size

    $H = \Sigma f(a) \log_2 f(a)$
- Also, Compositional Bias
  - ◊ Matching A-rich query to A-rich DB vs. A-poor DB

Computational Complexity

Core

- Basic NW Algorithm is $O(n^2)$ (in speed)
  - ◊ M x N squares to fill
  - ◊ At each square need to look back (M'+N') "black" squares to find max in block
  - ◊ M x N x (M'+N') -> $O(n^3)$
  - ◊ However, max values in block can be **cached**, so algorithm is really only $O(n^2)$
- $O(n^2)$ in memory too!
- Improvements can (effectively) reduce sequence comparison to O(n) in both

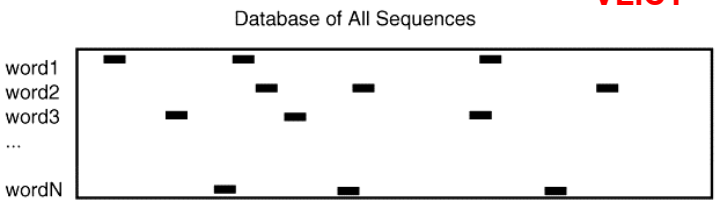


FASTA

Core

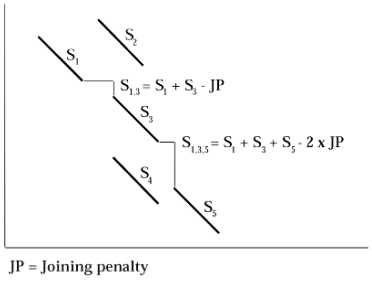
- Hash table of short words in the query sequence
- Go through DB and look for matches in the query hash (linear in size of DB)
- perl: $where{“ACT”} = 1,45,67,23....
- K-tuple determines word size (k-tup 1 is single aa)
- by Bill Pearson

VLICT =

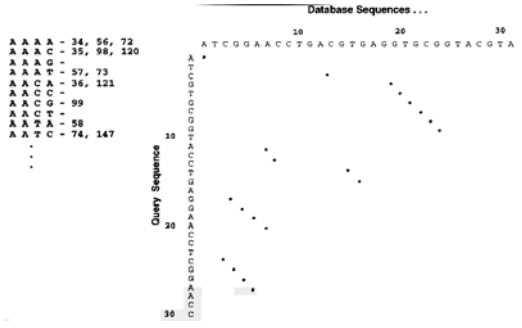


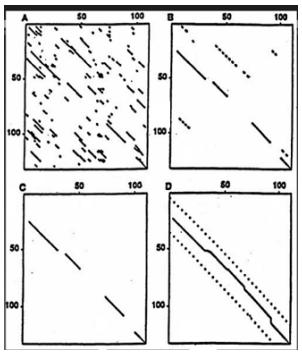
VLICTAVLM**VLICT**AAA**VLICT**MSDFFD

Join together query lookups into diagonals and then a full alignment



JP = Joining penalty





(Adapted from D Brutlag)

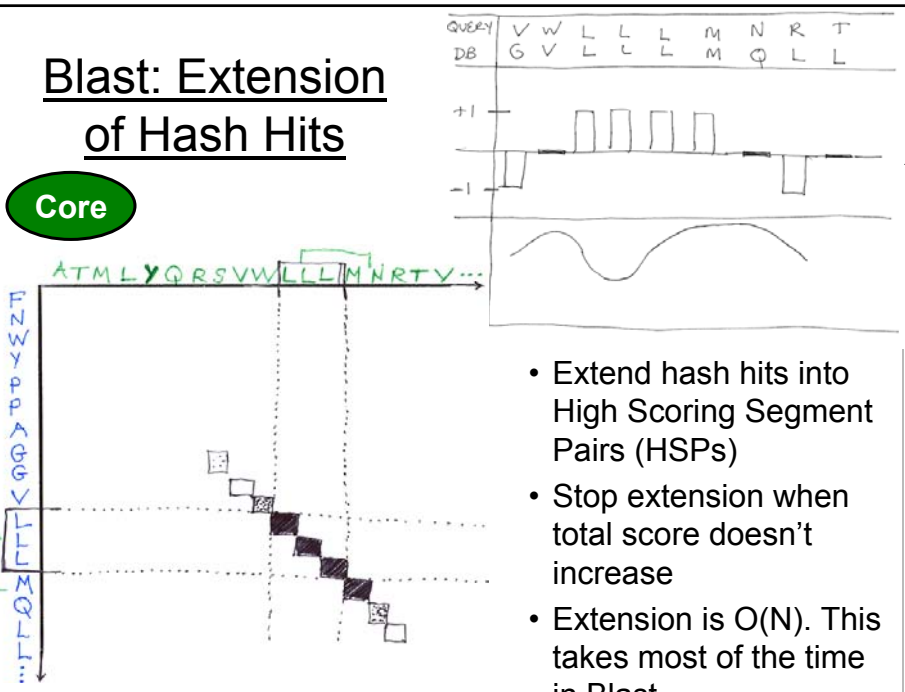
Basic Blast

Core

- Altschul, S., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410
- Indexes query (also tried indexing DB)
- Starts with all overlapping words from query
- Calculates "neighborhood" of each word using PAM matrix and probability threshold matrix and probability threshold
- Looks up all words and neighbors from query in database index
- Extends High Scoring Pairs (HSPs) left and right to maximal length
- Finds Maximal Segment Pairs (MSPs) between query and database
- Blast 1 does not permit gaps in alignments

Blast: Extension of Hash Hits

Core

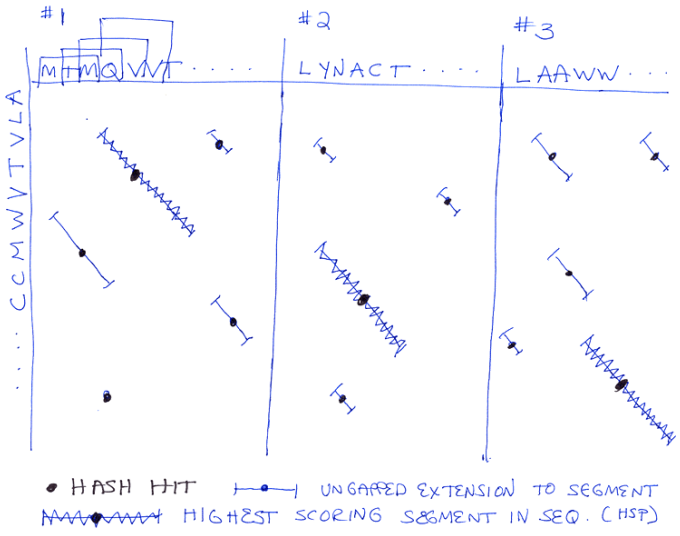


- Extend hash hits into High Scoring Segment Pairs (HSPs)
- Stop extension when total score doesn't increase
- Extension is O(N). This takes most of the time in Blast

- In simplest Blast algorithm, find best scoring segment in each DB sequence
- Statistics of these scores determine significance

Blasting against the DB

Number of hash hits is proportional to O(N*M*D), where N is the query size, M is the average DB seq. size, and D is the size of the DB



End of class 2002,10.30

(Bioinfo-6)

["broken projector"]

Analytic Score Formalism for Blast

Karlin-Altschul statistics for occurrence of high-scoring segments (HSPs) in random sequences

Extra

$$\Pr ob(S > X) \approx 1 - \exp\left\{-Ke^{-\lambda X}\right\}$$

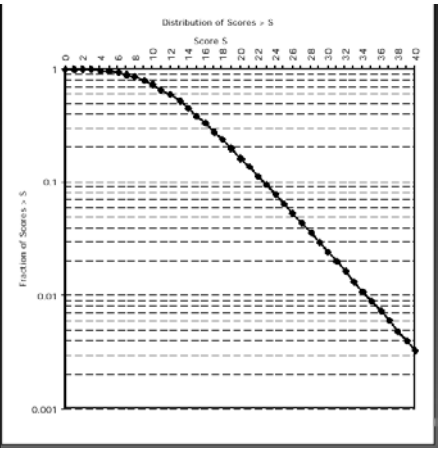
where λ is the root of the equation:

$$\sum_{i=1}^{r}\sum_{j=1}^{r} p_i p_j \exp\{\lambda s_{ij}\} = 1$$

p_i and p_j are the probabilities of each residue in each sequence,

s_{ij} are the similarity scores of two residues.

If the expected value of the scores for random sequences is < 0, i. e. $\left(\sum_{i=1}^{r}\sum_{j=1}^{r} p_i p_j s_{ij} < 0\right)$ then there are two solutions for λ, zero and one other positive root.



Blast2: Gapped Blast

Core

Nucleic Acids Research, 1997, Vol. 25, No. 17 3389–3402

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs

Stephen F. Altschul*, Thomas L. Madden, Alejandro A. Schäffer[1], Jinghui Zhang, Zheng Zhang[2], Webb Miller[2] and David J. Lipman

National Cent
Bethesda, MD
Institute, Natio
Engineering, F

Received June 2

ABSTRACT

The BLAST
searching pr
similarities.
definitional,

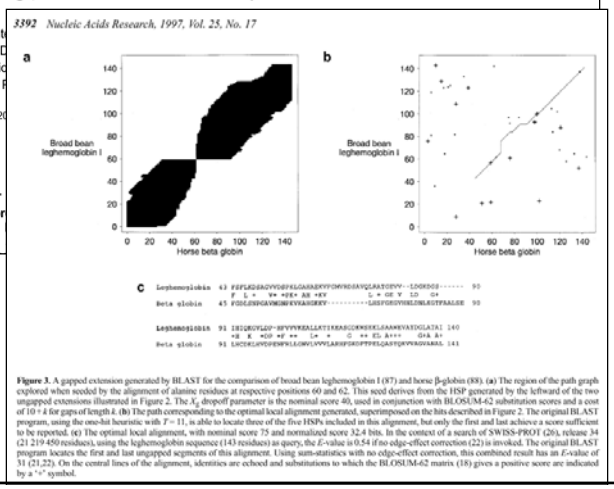
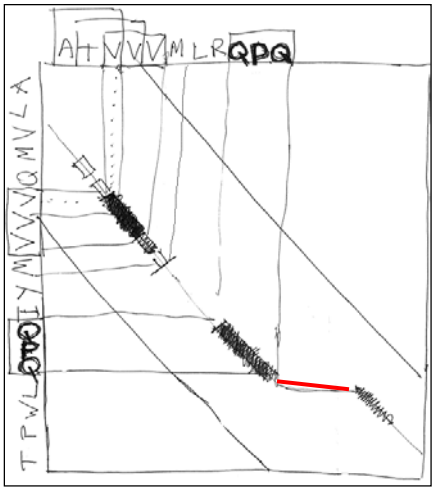


Figure 3. A gapped extension generated by BLAST for the comparison of broad bean leghemoglobin I (87) and horse β-globin (88). (**a**) The region of the path graph explored when seeded by the alignment of alanine residues at respective positions 60 and 62. This seed derives from the HSP generated by the leftward of the two ungapped extensions illustrated in Figure 2. The X_g dropoff parameter is the nominal score 40, used in conjunction with BLOSUM-62 substitution scores and a cost of $10 + k$ for gaps of length k. (**b**) The path corresponding to the optimal local alignment generated, superimposed on the hits described in Figure 2. The original BLAST program, using the one-hit heuristic with $T = 11$, is able to locate three of the five HSPs included in this alignment, but only the first and last achieve a score sufficient to be reported. (**c**) The optimal local alignment, with nominal score 75 and normalized score 32.4 bits. In the context of a search of SWISS-PROT (26), release 34 (21 219 450 residues), using the leghemoglobin sequence (143 residues) as query, the E-value is 0.54 if no edge-effect correction (22) is invoked. The original BLAST program locates the first and last ungapped segments of this alignment. Using sum-statistics with no edge-effect correction, this combined result has an E-value of 31 (21,22). On the central lines of the alignment, identities are echoed and substitutions to which the BLOSUM-62 matrix (18) gives a positive score are indicated by a '+' symbol.

Blast2: Gapped Blast

- Gapped Extension on Diagonals with two Hash Hits
- Statistics of Gapped Alignments follows EVD empirically

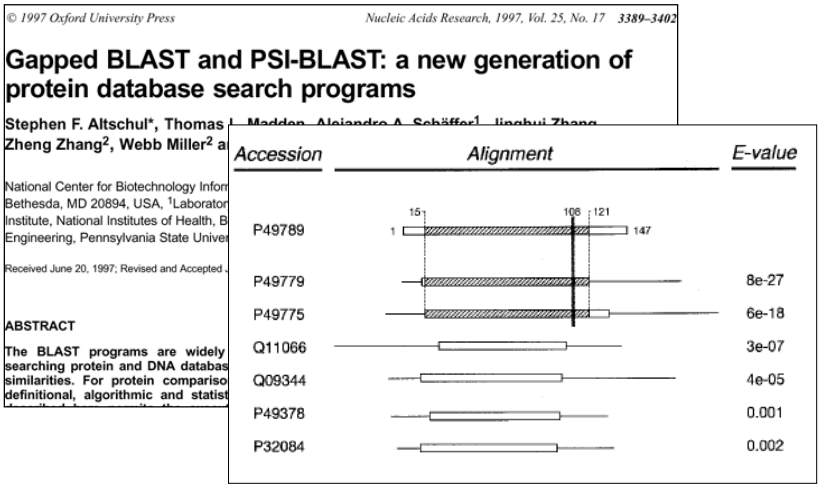


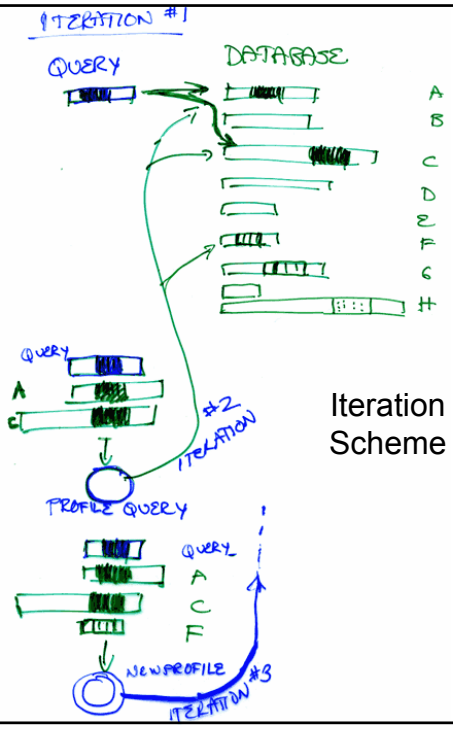
Core

Ψ-Blast

Parameters: overall threshold, inclusion threshold, interations

- Automatically builds profile and then searches with this
- Also PHI-blast





PSI-Blast

Core

Iteration Scheme

Sensitivity

Speed

Blast
FASTA
Smith-Waterman
PSI-Blast
Profiles
HMMs

Sequence Searching History

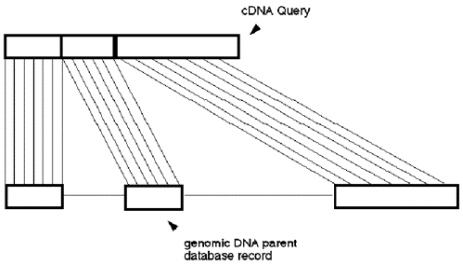
1970	NW		
1980	SW		
~1985	FASTA	PEARSON & LIPMAN	ENCBI
1989	BLAST	ALTSCHUL LIPMAN (KARLIN)	
1997	BLAST2 & γBLAST	"	

Extra

Practical Issues on DNA Searching

- Examine results with exp. between 0.05 and 10
- Reevaluate results of borderline significance using limited query
- Beware of hits on long sequences
- Limit query length to 1,000 bases
- Segment query if more than 1,000 bases

(graphic and some text adapted from D Brutlag)



- Search both strands
- Protein search is more sensitive, Translate ORFs
- BLAST for infinite gap penalty
- Smith-Waterman for cDNA/genome comparisons
- cDNA =>Zero gap-Transition matrices Consider transition matrices
- Ensure that expected value of score is negative

General Protein Search Principles

- Choose between local or global search algorithms
- Use most sensitive search algorithm available
- Original BLAST for no gaps
- Smith-Waterman for most sensitivity
- FASTA with k-tuple 1 is a good compromise
- Gapped BLAST for well delimited regions
- PSI-BLAST for families
- Initially BLOSUM62 and default gap penalties

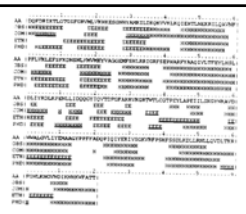
(some text adapted from D Brutlag)

- If no significant results, use BLOSUM30 and lower gap penalties
- FASTA cutoff of .01
- Blast cutoff of .0001
- Examine results between exp. 0.05 and 10 for biological significance
- Ensure expected score is negative
- Beware of hits on long sequences or hits with unusual aa composition
- Reevaluate results of borderline significance using limited query region
- Segment long queries ≥ 300 amino acids
- Segment around known motifs

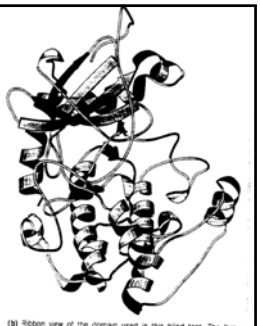
Secondary Structure Prediction Overview

- Why interesting?
  - ◊ Not tremendous success, but many methods brought to bear.
  - ◊ What does difficulty tell about protein structure?
- Start with TM Prediction (Simpler)
- Basic GOR Sec. Struc. Prediction
- Better GOR
  - ◊ GOR III, IV, semi-parametric improvements, DSC
- Other Methods
  - ◊ NN, nearest nbr.

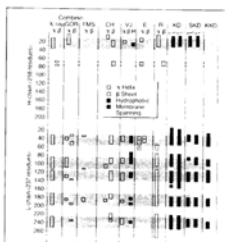
What secondary structure prediction tries to accomplish?



(a) Residue-by-residue comparison of experimentally observed (OBS) and predicted [COM[42], ETH[38], PHD (Ref. 35 and B. Rost and C. Sander, submitted)] structures of the catalytic subunit of the cAMP-dependent protein kinase (1cdk). AA is the amino acid sequence taken from Protein Data Bank entry 1cpk (residues 27-287). Secondary structure: H = α-helix, E = β-sheet (extended), Blank = loop. Predicted α-helices and β-strands that have insufficient overlap with an observed segment of the same type are underlined. Note the relatively good prediction of the location of segments for the ETH and PHD methods and overprediction of α-helices for the COM method.



(b) Ribbon view of the domain used in this blind test. The X-ray structure of catalytic subunit of the cAMP-dependent protein kinase. Drawn using Molscript[43].



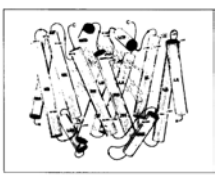


Figure 1
Column model for the core of the reaction center from *Rsp. viridis*. Reproduced, with permission, from Ref. 18.

Credits: Rost et al. 1993; Fasman & Gilbert, 1990

- Not Same as Tertiary Structure Prediction -- no coordinates
- Need torsion angles of terms + slight diff. in torsions of sec. str.

```
Sequence   RPDFCLEPPYTGPCKARIIRYFYNAKAGLVQTFVYGGCRAKRNNFKSAEDAMRTCGGA
Structure  CCGGGGCCCCCCCCCCCCEEEEEEETTTTEEEEEEECCCCCTTTTBTTHHHHHHHHHCC
```

Some TM scales:

GES

F	-3.7
M	-3.4
I	-3.1
L	-2.8
V	-2.6
C	-2.0
W	-1.9
A	-1.6
T	-1.2
G	-1.0
S	-0.6
P	+0.2
Y	+0.7
H	+3.0
Q	+4.1
N	+4.8
E	+8.2
K	+8.8
D	+9.2
R	+12.3

Goldman, Engleman, Steitz

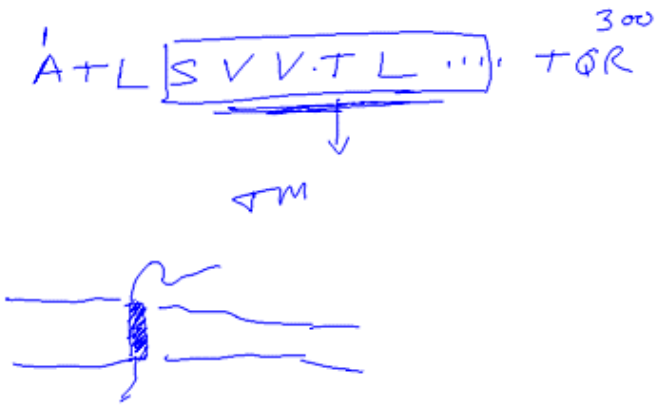
KD

KD – Kyte Dolittle

I	4.5
V	4.2
L	3.8
F	2.8
C	2.5
M	1.9
A	1.8
G	-0.4
T	-0.7
W	-0.9
S	-0.8
Y	-1.3
P	-1.6
H	-3.2
E	-3.5
Q	-3.5
D	-3.5
N	-3.5
K	-3.9
R	-4.5

For instance, ΔG from transfer of a Phe amino acid from water to hexane

TM Helix Identification: the Problem



How to use GES to predict proteins

- Transmembrane segments can be identified by using the GES hydrophobicity scale (Engelman et al., 1986). The values from the scale for amino acids in a window of size 20 (the typical size of a transmembrane helix) were averaged and then compared against a cutoff of -1 kcal/mole. A value under this cutoff was taken to indicate the existence of a transmembrane helix.
- H-19(i) = [H(i-9)+H(i-8)+...+H(i) + H(i+1) + H(i+2) + . . . + H(i+9)] / 19

Core

Graph showing Peaks in scales

Illustrations Adapted From: von Heijne, 1992; Smith notes, 1997

Core

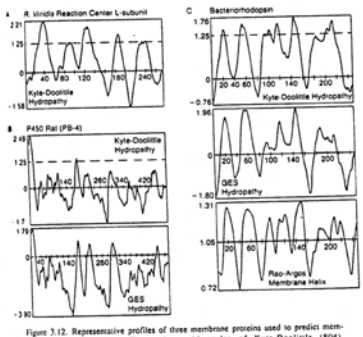
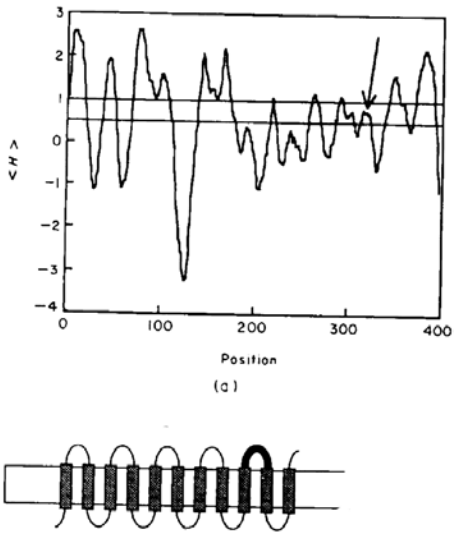


Figure 3.12. Representative profiles of three membrane proteins used to predict membrane-spanning helices. The amino acid scales of Kyte-Doolittle (804), Goldman–Engelman–Steitz (GES) (389), and Rao–Argos (1194) were used. A computer software package (SEQANAL) provided by Dr. A. Crofts (Univ. of Illinois) was used to generate these profiles. For comparative purposes, the Kyte–Doolittle and GES plots were obtained using a window of 19 residues and then smoothed using a second pass with a window of 7. The average value at each residue position is plotted as a function of residue number starting with the amino terminus on the left in each case. The values plotted for the Kyte–Doolittle and GES scales represent average hydropathy and transfer free energy per residue (kcal/mol). The Rao–Argos plot used a span of 7 residues and was smoothed over two additional passes with the same span of 7, as recommended by the authors. The scale values reflect the relative preference for being in a membrane-spanning helix. Note that the version of the GES algorithm which was used does not take into account possible ion pair formation. See text for details.



Removing Signal sequences

- Initial hydrophobic stretches corresponding to signal sequences for membrane insertion were excluded. (These have the pattern of a charged residue within the first 7, followed by a stretch of 14 with an average hydrophobicity under the cutoff).

Extra

Ex. P(i,α) probability that residue i has secondary structure α

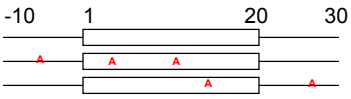
- Problem of DB Bias
- f(A) = frequency of residue A to have a TM-helical conf. in db
- f(A,i) = f(A) at position i in a particular sequence
- E(α)=statistical energy of helix over a window
- p(i, α) = probability that residue i is in a TM-helix

$$E_\alpha = \sum_i^N \ln f_\alpha^i$$

$$p_\alpha^i = \frac{e^{-Ea/RT}}{\sum_j^N e^{-E_j/RT}}$$

Core

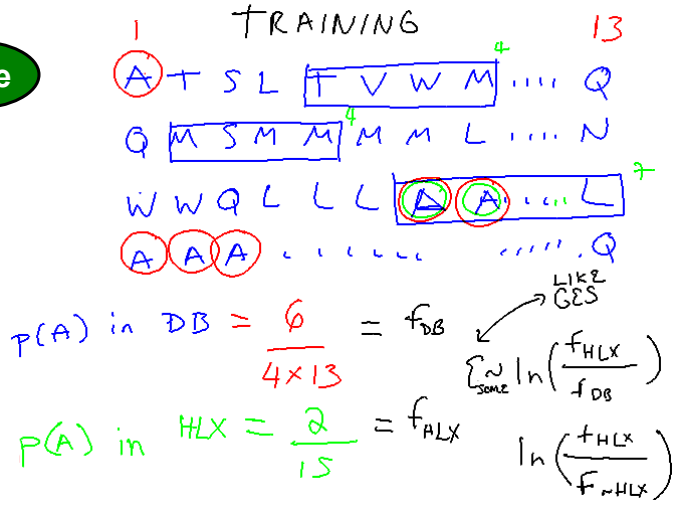
$F_{in\text{-}DB}(A) = 5/120$



$F_{in\text{-}TM}(A) = 3/60$

Example of Deriving a Scale from Frequencies

Core



Training, Testing, Running

TRAINING (SMALL SET) — DETERMINING THE RMS

TESTING — SEEING HOW WELL IT DOES

RUNNING — (LARGE SET) APPLYING THE PROC.

$$\sum_{\text{WINDOW}} \varepsilon_{\text{SCALE}}(i) \qquad \prod_i P_i(A)$$

Statistics Based Methods: Persson & Argos

- Propensity P(A) for amino acid A to be in the middle of a TM helix or near the edge of a TM helix

Core

$$P(A) = \frac{\frac{n(A,\mathrm{TM})}{\sum_A n(A,\mathrm{TM})}}{\frac{n(A,\mathrm{everywhere})}{\sum_A n(A,\mathrm{everywhere})}}$$

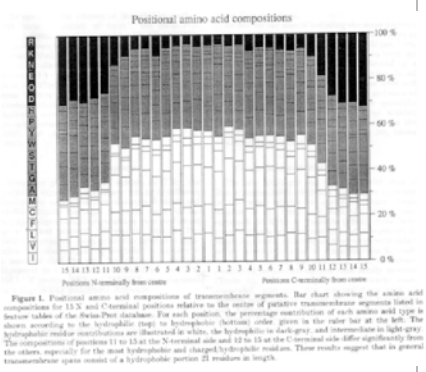
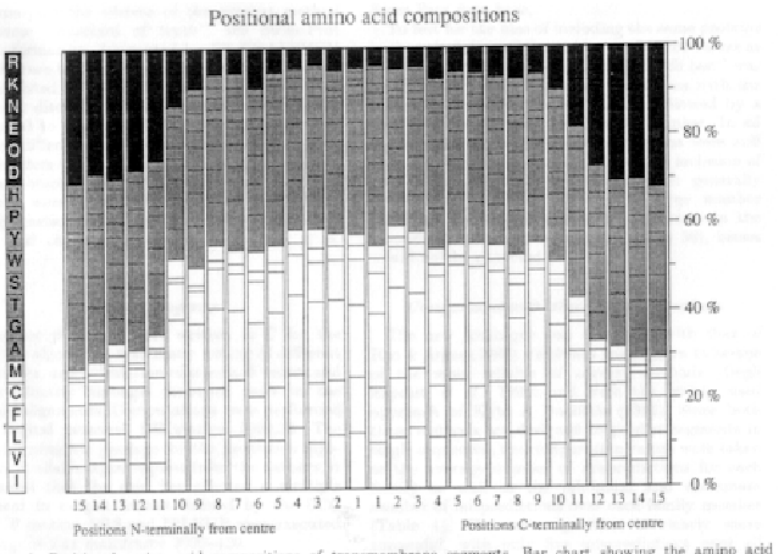


Illustration Credits: Persson & Argos, 1994

$P(A) = f_{TM}(A)/f_{SwissProt}(A)$



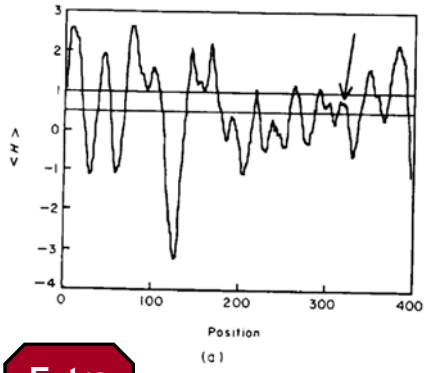
Extra

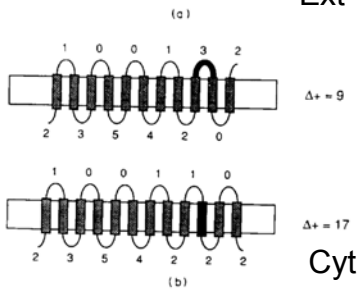
Scale Detail

Refinements: Charge on the Outside, Positive Inside Rule

- for marginal helices, decide on basis of R+K inside (cytoplasmic)

Ext





Cyt

Figure 4. (a) Hydrophobicity plot for the SecY protein. The upper and lower cutoffs are marked. A tentative transmembrane segment with a mean hydrophobicity falling between the 2 cutoffs is marked by an arrow. (b) Two possible topologies for the SecY protein based on the hydrophobicity plot. The putative transmembrane segment is shown in black. The number of Arg + Lys residues is shown next to each polar segment. Note that the correct alternative (bottom, including the putative transmembrane segment) has a much higher charge-bias than the incorrect one.

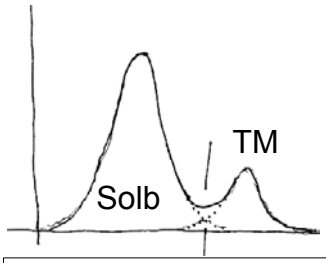
Extra

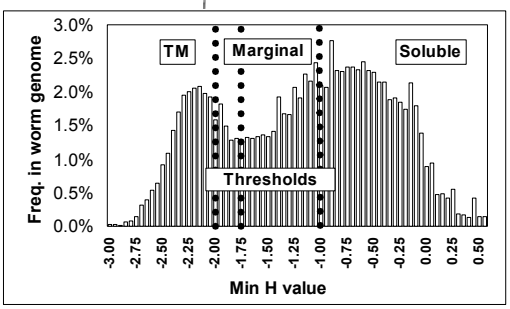
Credits: von Heijne, 1992

Refinements: MaxH

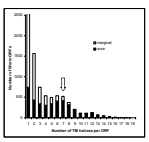
Extra

- How to train to find right threshold? Not that many TM helices
- Marginal TM helices are not that hydrophobic but 1/3 of TM's are very hydrophobic, so focus on these.





- Sosui, Klein & Delisi, Boyd
- Discriminant analysis: set threshold to be best partition of dataset



End of class 2002,11.04
(Bioinfo-7)
[quiz 1 up to here]

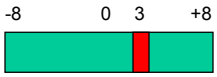
GOR: Simplifications

Core

- For independent events just add up the information
- $I(S_j ; R_1, R_2, R_3,...R_{last})$ = Information that first through last residue of protein has on the conformation of residue j (S_j)
  - ◊ Could get this just from sequence sim. or if same struc. in DB (homology best way to predict sec. struc.!)
- Simplify using a 17 residue window: $I(S_j=H ; R[j-8], R[j-7], ..., R[j], ... R[j+8])$
- Difference of information for residue to be in helix relative to not: I(dSj;y) = I(Sj=H;y)-I(Sj=~H;y)
  - ◊ odds ratio: I(dSj;y)= ln P(Sj;y)/P(~Sj;y)
  - ◊ I determined by observing counts in the DB, essentially a lod value

Basic GOR

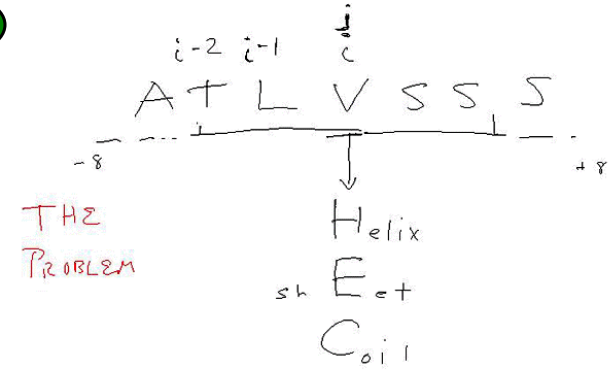
- Pain & Robson, 1971; Garnier, Osguthorpe, Robson, 1978
- I ~ sum of I(Sj,R[j+m]) over 17 residue window centered on j and indexed by m
  - ◊ I(Sj,R[j+m]) = information that residue at position m in window has about conformation of protein at position j
  - ◊ 1020 bins=17*20*3
- In Words
  - ◊ Secondary structure prediction can be done using the GOR program (Garnier et al., 1996; Garnier et al., 1978; Gibrat et al., 1987). This is a well-established and commonly used method. It is statistically based so that the prediction for a particular residue (say Ala) to be in a given state (i.e. helix) is directly based on the frequency that this residue (and taking into account neighbors at +/- 1, +/- 2, and so forth) occurs in this state in a database of solved structures. Specifically, for version II of the GOR program (Garnier et al., 1978), the prediction for residue i is based on a window from i-8 to i+8 around i, and within this window, the 17 individual residue frequencies (singlets).



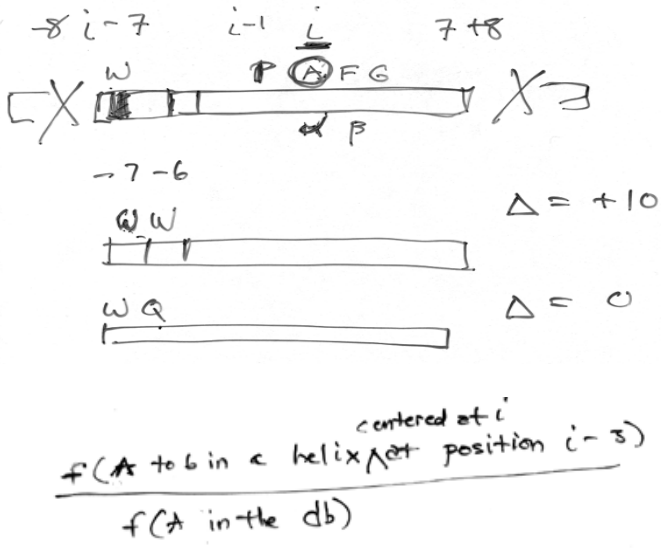
f(H,+3)/f(~H,+3)

The Secondary Structure Prediction Problem

Core



More GOR



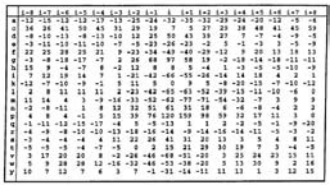
Directional Information

LOD= ln OBS / EXP

helix

strand

coil



Credits: King & Sternberg, 1996

Core

Table 3. *Directional informational parameters: h'Sj = x.x': Rj + mj for residue position versus residue type for α-helices**

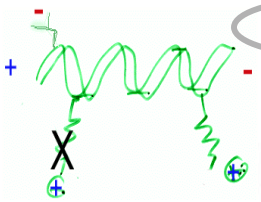
	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	19	21	22	24	34	36	44	47	60	60	53	50	44	40	31	23	24
c	-47	-45	-44	-47	-44	-36	-44	-55	-56	-58	-54	-55	-58	-58	-59	-53	-66
d	14	15	14	15	17	21	15	17	-7	-11	-31	-42	-28	-12	-8	1	-5
e	14	16	15	20	26	27	34	52	62	57	32	15	19	12	6	7	9
f	-19	-14	-10	-4	-2	-1	6	-1	10	10	12	12	-4	-5	2	0	2
g	5	2	1	-5	-22	-30	-50	-70	-92	-52	-28	-21	-13	-17	-8	-6	-6
h	-22	-20	-9	-10	-19	-10	-14	-7	-11	-4	0	-3	-2	2	6	11	12
i	7	7	0	0	1	1	2	-5	1	2	1	7	-6	-3	10	8	6
k	-2	-1	-1	-1	-6	-9	-6	5	17	17	21	27	35	33	21	22	23
l	0	-1	0	6	9	16	30	33	45	47	51	53	37	32	30	25	18
m	4	3	15	23	30	30	39	36	45	54	57	53	44	29	30	14	1
n	2	3	2	-5	-9	-10	-16	-17	-31	-16	-17	-16	-9	-8	-9	-10	-5
p	-12	-15	-14	-19	-23	-25	-30	-48	-82	-195	-145	-104	-67	-49	-43	-33	-17
q	-4	3	7	4	13	8	10	24	35	32	31	21	18	18	9	8	6
r	5	3	6	13	7	13	19	27	34	32	36	41	33	29	23	21	18
s	-10	-7	-10	-10	-16	-17	-25	-21	-39	-35	-39	-41	-32	-35	-34	-35	-33
t	1	-1	-6	-8	-6	-11	-16	-25	-48	-47	-48	-46	-34	-31	-34	-26	-24
v	-5	-12	-13	-14	-13	-19	-17	-20	-15	-22	-22	-20	-26	-19	-15	-10	-5
w	0	-4	-12	-19	-7	14	16	12	18	17	12	8	1	-6	1	3	-13
y	-22	-19	-17	-20	-16	-21	-30	-32	-8	-10	-4	-12	-17	-9	-10	-14	-15

*Note that the convention used is the reverse of that adopted by (Garnier et al., 1978), for example the first entry for alanine at position j-8 is the amount of information that an alanine residue eight positions toward the N terminus has for predicting an α-helix

Table 4. *Directional informational parameters for residue position versus residue type for β-strands*

	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	-8	-7	-13	-17	-23	-33	-26	-32	-43	-37	-30	-30	-26	-27	-26	-25	-25
c	3	13	-9	-20	-15	-3	9	33	47	51	21	19	9	-5	7	-5	-14
d	-7	-5	0	-9	-4	-14	-42	-73	-83	-59	-21	10	22	24	16	11	13
e	-14	-5	-5	-11	-21	-27	-45	-44	-57	-54	-46	-29	-25	-12	-12	-2	0
f	-9	-20	-32	-34	-30	-12	24	44	49	39	24	2	-9	-23	-24	-29	-23
g	-3	9	24	29	34	30	18	-23	-48	-27	6	27	39	38	33	23	23
h	6	11	17	22	12	16	0	-2	3	-2	5	3	8	4	-1	1	-3
i	-21	-30	-31	-21	-12	-3	26	58	76	64	33	11	-14	-24	-20	-14	-11
k	20	12	15	14	8	4	-8	-14	-25	-40	-39	-27	-20	-24	-20	-15	-15
l	-2	-10	-18	-27	-30	-27	-6	15	27	21	2	-19	-31	-29	-28	-26	-25
m	-22	-26	-29	-40	-31	-17	-7	23	24	28	17	2	-15	-31	-53	-36	-16
n	1	8	14	5	0	-6	-30	-65	-62	-28	-6	11	18	21	16	10	3
p	9	7	12	24	20	8	-22	-65	-108	-64	-8	17	25	30	32	31	21
q	6	12	8	16	8	-5	-22	-27	-30	-52	-49	-34	-22	-17	-9	2	20
r	0	8	3	-3	5	2	1	-14	-26	-32	-30	-35	-27	-26	-25	-25	-21
s	16	14	17	19	14	5	-3	-13	-15	-4	15	27	32	32	31	28	21
t	6	8	14	15	16	21	19	25	31	22	13	9	12	25	34	34	34
v	1	-11	-15	-11	4	25	51	75	91	81	49	19	-6	-12	-16	-11	-11
w	-8	-8	-28	-19	-9	5	23	44	45	30	13	-18	-22	-40	-15	-7	-9
y	13	13	4	14	12	20	24	37	48	31	20	-1	2	11	7	0	-4

Types of Residues



Credits: King & Sternberg, 1996

Table 3. *Directional informational parameters: I(Sj = x:x': Rj + m) for residue position versus residue type for α-helices*[a]

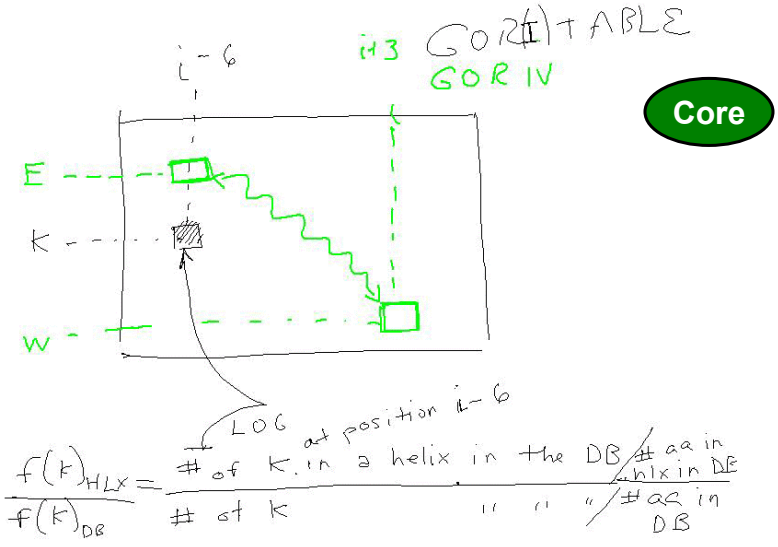
	i-8	i-7	i-6	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+6	i+7	i+8
a	19	21	22	24	34	36	44	4	60	60	53	50	44	40	31	23	24
c	-47	-45	-44	-47	-44	-36	-44	-5	-56	-58	-54	-55	-58	-58	-59	-53	-66
d	14	15	14	15	17	21	15	1	-7	-11	-31	-42	-28	-12	-8	1	-5
e	14	16	15	20	26	27	34	52	62	57	32	15	19	12	6	7	9
f	-19	-14	-10	-4	-2	-1	6	-1	10	10	12	12	-4	-5	2	0	2
g	5	2	1	-5	-22	-30	-50	-70	-92	-52	-28	-21	-13	-17	-8	-6	-6
h	-22	-20	-9	-10	-19	-10	-14	-7	-11	-4	0	-3	-2	2	6	11	12
i	7	7	9	0	1	1	2	-5	1	2	1	7	-6	-3	10	6	6
k	-2	-1	-1	-1	-6	-9	-6	5	17	17	21	27	35	33	21	22	23
l	0	-1	-2	6	9	16	30	33	45	47	51	53	37	32	30	25	18
m	4	3	15	23	30	30	39	36	45	54	57	53	44	29	30	14	1
n	2	3	2	-5	-9	-10	-16	-17	-31	-16	-17	-16	-9	-8	-9	-10	-5
p	-12	-15	-14	-19	-23	-25	-31	-48	-82	-195	-145	-104	-67	-49	-43	-33	-17
q	-4	3	7	4	13	8	10	24	35	32	31	21	18	18	9	8	6
r	5	3	6	13	7	13	19	27	34	32	36	41	33	29	23	21	18
s	-10	-7	-10	-10	-16	-17	-25	-21	-39	-35	-39	-41	-32	-35	-34	-35	-33
t	1	-1	-6	-8	-6	-11	-16	-25	-48	-47	-48	-46	-34	-31	-34	-26	-24
v	-5	-12	-13	-14	-13	-19	-17	-20	-15	-22	-22	-20	-26	-19	-15	-10	-5
w	0	-4	-12	-19	-7	14	16	12	18	17	12	8	1	-6	1	3	-13
y	-22	-19	-17	-20	-16	-21	-30	-32	-8	-10	-4	-12	-17	-9	-10	-14	-15

[a]Note that the convention used is the reverse of that adopted by (Garnier et al., 1978), for example the first entry for alanine at position j-8 is the amount of information that an alanine residue eight positions toward the N terminus has for predicting an α-helix at position j.

- Group I favorable residues and Group II unfavorable one:
- A, E, L -> H; V, I, Y, W, C -> E; G, N, D, S -> C
- P complex; largest effect on proceeding residue
- Some residues favorable at only one terminus (K)

Core

How to calculate an entry in the GOR I tables and a comparison to GOR IV



Core

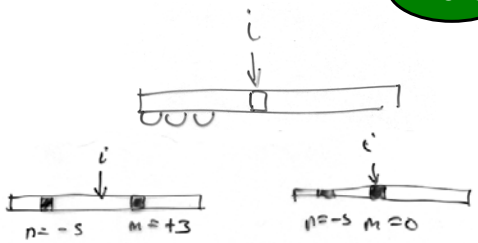
GOR IV

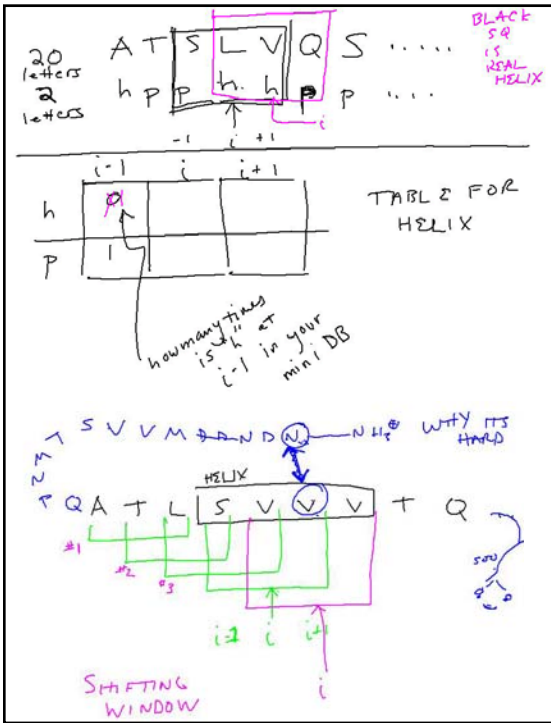
- I(Sj; R[j+m], R[j+n]) = the frequencies of all 136 (=16*17/2) possible di-residue pairs (doublets) in the window.
  - ◊ 20*20*3*16*17/2=163200 pairs
- Parameter Explosion Problem: 1000 dom. struc. * 100 res./dom. = 100k counts, over how many bins
- Dummy counts for low values (Bayes)

Core

All Singletons in 17 residue window

All Pairs





An example of mini-GOR

Also, why secondary structure prediction is so hard

Core

Assessment

- Q3 + other assess, 3x3
- Q3 = total number of residues predicted correctly over total number of residues
- GOR gets 65%
  - ◊ sum of diagonal over total number of residue -- (14K+5K+21K)/ 64K
- Under predict strands & to a lesser degree, helices: 5.9 v 4.1, 10.9 v 10.6

THE GOR METHOD

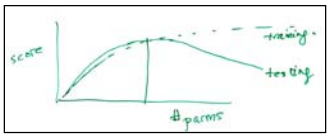
TABLE II
GLOBAL RESULTS FOR DATABASE PREDICTION

	Observed			
	H	E	C	Total
Predicted				
H	14,460	3094	4790	22,344
E	1124	4965	2089	8178
C	6002	5546	21,496	33,044
Total	21,586	13,605	28,375	63,566
Q_{pred} [a]	64.7	60.7	65.1	
Q_{obs} [b]	67.0	36.5	75.8	

Q_3 [c] = 64.4%

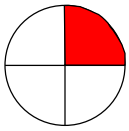
[a] Number of correctly predicted residues/number of predicted residues.
[b] Number of correctly predicted residues/number of observed residues
[c] Total number of correctly predicted residues/total number of residues.

Credits: Garnier et al., 1996



Training and Testing Set

- Cross Validation: Leave one out, seven-fold



4-fold

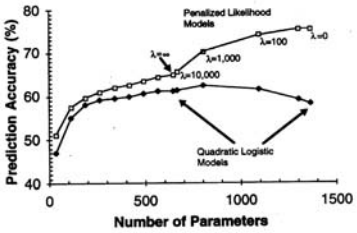


Figure 2. Comparison of prediction accuracy (correctly predicted residues as a proportion of total residues) versus effective number of parameters for linear-logistic models (number of parameters ≤640) and penalized likelihood models for crossvalidated (◆) and uncrossvalidated (❑) results. The values of the penalty parameter λ are shown.

Credits: Munson, 1995;
Garnier et al., 1996

TABLE I
DATABASE PROTEINS*

1aaj.x	1aak.x	1aap.a	1aba.x	1abk.x	1abm.a	1add.x
1ads.x	1alk.a	1aor.a	1apa.x	1apm.e	1arb.x	1atr.x
1avb.a	1ayh.x	1bab.a	1bbh.a	1bbp.a	1bet.x	1bge.a
1bll.e	1bmd.a	1bov.a	1bpb.x	1brx.d	1btc.x	1c2r.a
1caj.x	1cau.a	1cau.b	1ede.x	1edt.a	1cew.i	1cgt.x
1chm.a	1cmb.a	1cob.a	1col.a	1epc.a	1cpc.b	1cpt.x
1crl.x	1cse.i	1ctf.x	1ctm.x	1cus.x	1ddt.x	1dhr.x
1dog.x	1dsb.a	1eaf.x	1eco.x	1ede.x	1end.x	1epa.a
1fba.a	1fdd.x	1fha.x	1fia.a	1fkb.x	1fna.x	1fnr.x
1fxi.a	1gal.x	1gd1.o	1gdh.a	1gky.x	1glt.x	1gmf.a
1gof.x	1gox.x	1gp1.a	1gpb.x	1gsr.a	1gsr.a	1hbq.x
1hdx.a	1hiv.a	1hlb.x	1hfe.a	1hmy.x	1hoe.x	1hpl.a
1hrh.a	1hsl.a	1huw.x	1ifc.x	1ipd.x	1isu.a	1ith.a
1l29.x	1le4.x	1len.a	1lga.a	1lis.x	1lla.x	1lmb.3
1lts.a	1lts.d	1mdc.x	1mgn.x	1min.a	1min.b	1mjc.x
1mpp.x	1mup.x	1nar.x	1nba.a	1ndk.x	1noa.x	1nsb.a
1nxb.x	1ofv.x	1olb.a	1omf.x	1omp.x	1onc.x	1osa.x
1pda.x	1pfk.a	1pgb.x	1pgd.x	1phh.x	1php.x	1pii.x
1plf.a	1poc.x	1poh.x	1pox.a	1ppa.x	1ppf.e	1ppf.i
1ppn.x	1prc.c	1prc.h	1pre.l	1pre.m	1pts.x	1pya.a
1pya.b	1ppd.a	1reb.x	1rec.x	1rib.a	1rod.x	1rop.a
1rve.a	1s01.x	1sac.a	1sbp.x	1scs.a	1sgt.x	1sha.a
1shf.a	1sim.x	1slt.b	1snc.x	1spa.x	1stf.i	1tbe.a
1tca.x	1tie.x	1tnd.x	1tnd.a	1tpl.a	1trb.x	1trk.a
1tro.a	1ttb.a	1utg.x	1vaa.a	1vaa.b	1vmo.a	1wht.a
1wht.b	1wsy.a	1wsy.b	1ybb.x	1zaa.c	256b.a	2aai.b
2aza.a	2bop.a	2ccy.x	2cdv.x	2chs.a	2cmd.x	2cp4.x
2cpl.x	2cro.x	2ctc.x	2ets.x	2cyp.x	2dnj.a	2er7.e
2hbg.x	2hhm.a	2hip.a	2hpd.a	2ihl.x	2ilh2.x	2liv.x
2mhr.x	2mnr.x	2msb.a	2mta.c	2mta.h	2mta.l	2pf1.x
2pia.x	2pol.a	2por.x	2reb.x	2rn2.x	2rsl.a	2sar.a
2sas.x	2scp.a	2sga.x	2sn3.x	2spc.a	2tgi.x	2tmd.a
2tpr.a	2tsc.a	3aah.a	3aah.b	3adk.x	3b5c.x	3cd4.x
3chy.x	3cla.x	3cox.x	3dfr.x	3eca.a	3gap.a	3gbp.x
3ink.c	3rub.l	3rub.s	3sdh.a	3tgl.x	451c.x	4blm.a
4enl.x	4fgf.x	4gcr.x	4ts1.a	4xis.x	5fbp.a	5p21.x
5tim.a	6fab.h	6fab.l	6taa.x	8abp.x	8acn.x	8atc.a
8atc.b	8cat.a	81b.x	8rxn.a	8tln.e	9ldt.a	9rnt.x
9wga.a						

* The database was prepared by J. M. Levin and checked for homologous sequences with the help of V. Di Francesco. This database has been modified to restore the total length (the DSSP program omits residues whose coordinates are missing in the PDB file, and thus if this occurs in the middle of the polypeptide chain it is split into two or more chains). Residues having no coordinates were assigned the conformation X and were not taken into account for the prediction accuracy although the prediction was done with the whole sequence length. The PDB code is followed by the chain name a, b, c, d, h (heavy), l (light), x (one chain only), e (enzyme), or i (inhibitor).

End of class 2002,11.06
(Bioinfo-8)
[]

Is 100% Accuracy Possible?

Extra

Quoted from Barton (1995):

One problem that has arisen is how to evaluate secondary structure predictions. For prediction of a single protein sequence one might expect the best residue by residue accuracy to be 100%. It is not possible to define the secondary structure of a protein exactly, however. There is always room for alternative interpretations of where a helix or strand begins or ends so failure of a prediction to match exactly the secondary structure definition is not a disaster [24]. The problem of evaluation is more complicated for prediction from multiple sequences, as the prediction is a consensus for the family and so is not expected to be 100% in agreement with any single family member. The expected range in accuracy for a perfect consensus prediction is a function of the number, diversity and length of the sequences. Russell and I have calculated estimates of this range [11].

Simple residue by residue percentage accuracy has long been the standard method of assessment of secondary structure predictions. Although a useful guide, high percentage accuracies can be obtained for predictions of structures that are unlike proteins. For example, predicting myoglobin to be entirely helical (no strand or coil) will give over 80% accuracy but the prediction is of little practical use. Rost *et al.* [25] and Wang [26] explore these problems and suggest some alternative measures of predictive success based on secondary structure segment overlap. Although such measures help in an objective assessment of the prediction, there is no complete substitute for visual inspection. By eye, serious errors stand out and predictions of structures that are unlike proteins are usually recognizable. By eye, it is also straightforward to weight the importance of individual secondary structures. For example, prediction of what is in fact a core strand to be a helix would seriously hamper attempts to generate the correct tertiary structure of the protein from the predicted secondary structure, whereas prediction of a non-core helix as coil may have little impact on the integrity of the tertiary structure.

Types of Secondary Structure Prediction Methods

- Parametric Statistical
  - ◊ struc. = explicit numerical func. of the data (GOR)
- Non-parametric
  - ◊ struc. = NON- explicit numerical func. of the data
  - ◊ generalize Neural Net, seq patterns, nearest nbr, &c.
- Semi-parametric: combine both
- single sequence
- multi sequence
  - ◊ with or without multiple-alignment

Core

GOR Semi-parametric Improvements

- Filtering GOR to regularize

$[\neg a, \neg a, c, \mathbf{b}, *, \neg b] \rightarrow c$

$[\neg a, *, *, \mathbf{a}, b] \rightarrow b$

$[\neg a, *, *, \mathbf{a}, c] \rightarrow c$

$[a, *, *, \mathbf{a}, c, *, \neg c] \rightarrow c$

$[\neg a, \neg a, \mathbf{a}, a, c, \neg a] \rightarrow c$

$[\neg a, c, \neg c, \mathbf{a}, a, c, \neg a] \rightarrow c$

$[\neg a, c, c, a, a, \neg b, \neg a] \rightarrow c$

$[a, c, *, \mathbf{a}, a, a, \neg a] \rightarrow c$

$[*, c, *, \mathbf{a}, a, b, \neg a] \rightarrow c$

$[c, b, b \quad \mathbf{a}, a, *, a] \rightarrow b$

$[c, * \quad \mathbf{a}, a, \neg a, a] \rightarrow c$

a = α-helix, b = β-strand, c = coil, * = wildcard (α-helix or β-strand or coil) ¬ = not.

If the pattern on the left is met in a prediction, then the secondary structure in bold on the left is rewritten as the secondary structure on the right of the rule. For example:

$[b, b, b, \mathbf{a}, c] \rightarrow [b, b, b, \mathbf{c}, c]$

$[b, b, c, \mathbf{a}, c] \rightarrow [b, b, c, \mathbf{c}, c]$

$[b, b, b, a, b, b, b] \rightarrow [b, b, b, \mathbf{b}, b, b, b].$

Illustration Credits: King & Sternberg, 1996

Core

Multiple Sequence Methods

- Average GOR over multiple seq. Alignment
- The GOR method only uses single sequence information and because of this achieves lower accuracy (65 versus >71 %) than the current "state-of-the-art" methods that incorporate multiple sequence information (e.g. King & Sternberg, 1996; Rost, 1996; Rost & Sander, 1993).

Illustration Credits: Livingston & Barton, 1996

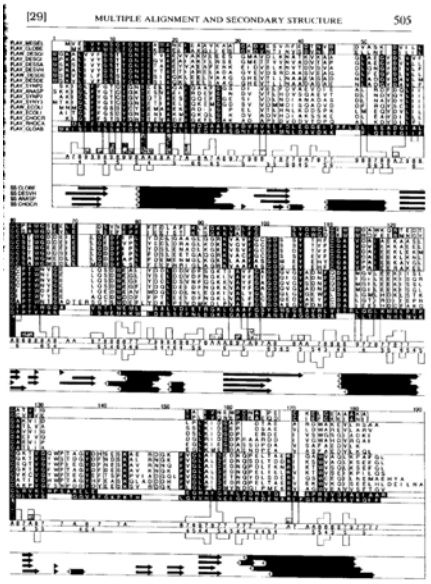


FIG. 5. Conservation analysis of the 17 flavodoxin sequences clustered in Fig. 3. The Taylor Venn diagram was used (Fig. 1) with a threshold of $T = 7$. See text for details.

DSC -- an improvement on GOR

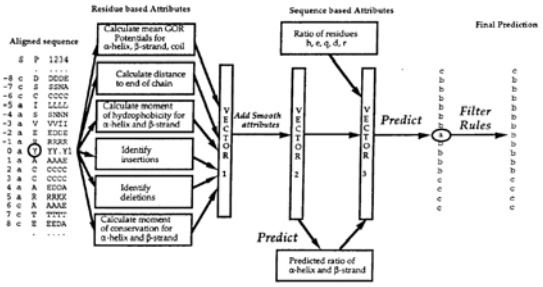


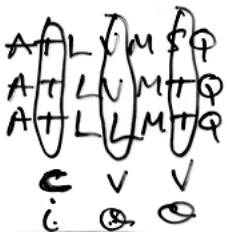
Fig. 1. DSC prediction method. For the aligned sequence: S is the observed secondary structure of the primary sequence, P. The residue at position 0 is predicted (circled).

Illustration Credits: King & Sternberg, 1996

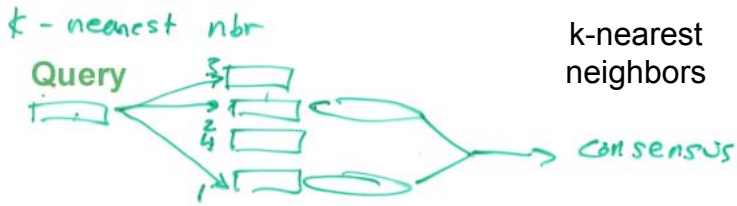
- GOR parms
- + simple linear discriminant analysis on:
  - ◊ dist from C-term, N-term
  - ◊ insertions/deletes
  - ◊ overall composition
  - ◊ hydrophobic moments
  - ◊ autocorrelate: helices
  - ◊ conservation moment

Conservation, k-nn

Extra



Patterns of Conservation



k-nearest neighbors

Neural Networks

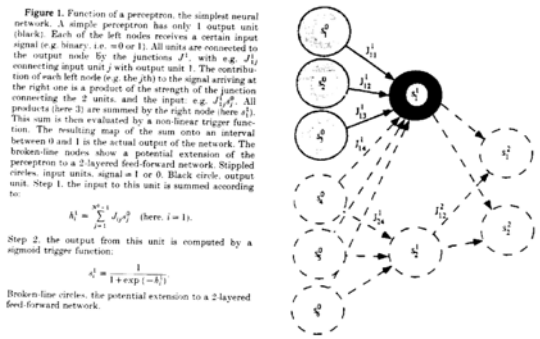


Figure 1. Function of a perceptron, the simplest neural network. A simple perceptron has only 1 output unit (black). Each of the left nodes receives a certain input signal (e.g. binary, i.e. =0 or 1). All units are connected to the output node by the junctions J^1, with e.g. J^1_{ij} connecting input unit j with output unit 1. The contribution of each left node (e.g. the jth) to the signal arriving at the right one is a product of the strength of the junction connecting the 2 units, and the input, e.g. $J^1_{ij}s^0_j$. All products (here 3) are summed by the right node (here s^1_i). This sum is then evaluated by a non-linear trigger function. The resulting map of the sum onto an interval between 0 and 1 is the actual output of the network. The broken-line nodes show a potential extension of the perceptron to a 2-layered feed-forward network. Stippled circles, input units, signal = 1 or 0. Black circle, output unit. Step 1, the input to this unit is summed according to:

$$h^1_i = \sum_{j=1}^{N^0+1} J_{ij}s^0_j \quad \text{(here, } i=1\text{)}.$$

Step 2, the output from this unit is computed by a sigmoid trigger function:

$$s^1_i = \frac{1}{1+\exp(-h^1_i)}$$

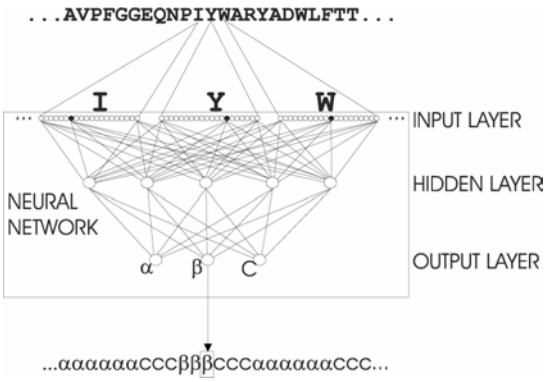
Broken-line circles, the potential extension to a 2-layered feed-forward network.

- Somehow generalize and learn patterns
- Black Box
- Rost, Kneller, Qian….
- Perceptron (above) is Simplest network
  - ◊ Multiply junction * input, sum, and threshold

Extra

Illustration Credits: Rost & Sander, 1993

More NN



- Hidden Layer
- Learning
  - ◊ Steepest descent to minimize an error function
- Jury Decision
  - ◊ Combine methods
  - ◊ Escape initial conditions

Extra

Illustration Credits: D Frishman handout

Yet more methods….

- struc class predict
  - ◊ Vect dist. between composition vectors
- threading via pair pot
- seq comparison
- ab initio from md
- ab initio from pair pot.

Extra

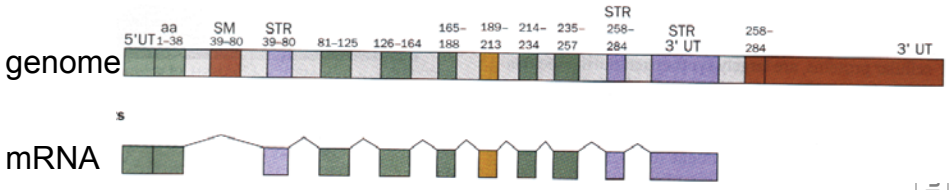
Mail Servers and Web Forms

Extra

Method	URL	Institution	Source code Availability
ANTHE-PROT	http://www.ibcp.fr/antheprot.html (currently unreachable)	Institute of Biology and Chemistry of Proteins (Lion)	YES
PSSP	http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html	Baylor College of Medicine (Houston)	NO
DSC	http://bonsai.lif.icnet.uk/bmm/dsc/dsc_form_align.html	Imperial Cancer Research Center (London)	YES
GOR	http://molbiol.soton.ac.uk/compute/GOR.html	University of Southampton	NO
nnPredict	http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html	University of California (San Francisco)	NO
Predict-Protein	http://www.embl-heidelberg.de/predictprotein/predictprotein.html	EMBL (Heidelberg)	NO
PRED-ATOR	http://www.embl-heidelberg.de/argos/predator/predator_form.html	EMBL (Heidelberg)	YES
PSA	http://bmerc-www.bu.edu/psa/	BioMolecular Engineering Research Center, Boston	NO
SSPRED	http://www.embl-heidelberg.de/sspred/sspred_info.html	EMBL (Heidelberg)	NO
GOR and DSC	http://genome.imb-jena.de/cgi-bin/GDEWWW/menu.cgi	IMB (Jena)	NO
GOR	http://absalpha.dcrt.nih.gov:8008/gor.html	DCRT/NIH (Washington)	NO
GOR	ftp://ftp.virginia.edu/pub/fasta	University of Virginia	YES
Mult-Predict	http://kestrel.ludwig.ucl.ac.uk/zpred.html	Ludwig Institute for Cancer Research (London)	NO

Illustration Credits: D Frishman handout

Additional Features of DNA sequences in Genomes



Gene finding

- composition of codons, nts
- Splice site finding

TABLE 30-2. THE "STANDARD" GENETIC CODE[a]

First position (5' end)	Second position U	Second position C	Second position A	Second position G	Third position (3' end)
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met[b]	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

CAI **Extra**

Genetic Code

- The genetic code is highly degenerate (64 codons to encode 20 amino acids)
- Three amino acids (Arg, Leu, Ser) are each specified by six codons, and many of the other amino acids are specified by two or four codons
- The arrangement of the codons within the genetic code is not random
- In most cases mutation of the third nucleotide in the codon would either cause no change in the amino acid (Arg, Val or Leu for example) or would create a fairly conservative change (Phe to Leu or Asp to Glu)
- Codons with second position pyrimidines encode mostly hydrophobic amino acids (tan), while those with second position purines encode mostly polar amino acids (blue, red, and purple)
- The genetic code is nonambiguous. Each codon encodes a single amino acid. The only exception is GUG which in some mRNAs is used as a start codon to encode Met
- The genetic code includes three stop codons, UAG, UAA, and UGA which are termed amber, ochre, and opal codons
- The genetic code is nearly but not absolutely universal. The genetic code in mitochondria and some ciliates use a slightly modified version of the code

(Page adapted from S Strobel, Biochemistry Lecture Notes)

Splicing

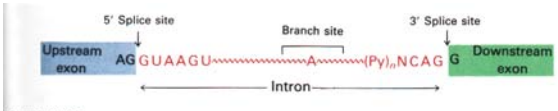


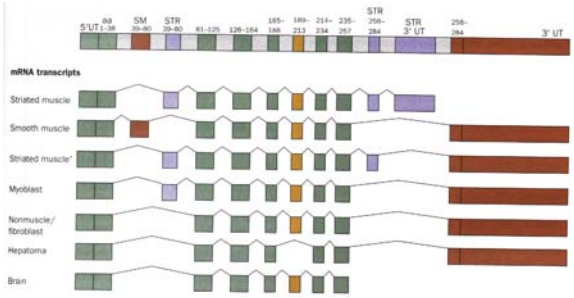
Figure 33-34
Splicing signals. Consensus sequences for the 5′ splice site and the 3′ splice site are shown.

Extra

- Splicing must be done accurately. Missplicing by even one nucleotide would result in a frameshift mutation throughout the remainder of the message
- The splice sites are defined largely by sequences within the intron
- The intron begins with the sequence GU and ends with AG and is part of a larger consensus sequence at both the5' and 3' splice sites (see figure)
- 30-50 nucleotides upstream of the 3' splice site is the branch site which includes an A that serves as the nucleophile in the reaction

(Page adapted from S Strobel, Biochemistry Lecture Notes)

Alternative Splicing: Multiple Proteins from One Gene



- A single transcript can be processed to include or not include specific exons within the gene. This is termed alternative splicing
- This makes it possible to generate multiple proteins from a single gene
- For example a single rate gene encodes seven tissue-specific variants of the muscle protein a-tropomyosin through the process of alternative splicing
- Sex determination in Drosophila is largely controlled by a series of alternative splicing events

(Page adapted from S Strobel, Biochemistry Lecture Notes)

Promotors

•The RNA polymerase recognizes a promoter sequence within the DNA

•The consensus promoter includes two six base pair regions upstream of the transcription start site (defined as nucleotide +1)

•The Pribnow box (consensus sequence of TATAAT) is 10 nt upstream

•There is second element 35 nt upstream (consensus sequence TTGACA)

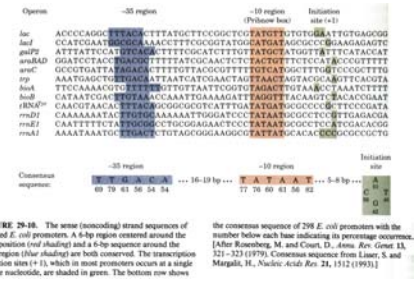


FIGURE 29-10. The sense (noncoding) strand sequences of selected *E. coli* promoters. A 6-bp region centered around the −10 position (*red shading*) and a 6-bp sequence around the −35 region (*blue shading*) are both conserved. The transcription initiation sites (+1), which in most promoters occurs at a single purine nucleotide, are shaded in green. The bottom row shows the consensus sequence of 298 *E. coli* promoters with the number below each base indicating its percentage occurrence. [After Rosenberg, M. and Court, D., Annu. *Rev. Genet.* **13**, 321–323 (1979). Consensus sequence from Lisser, S. and Margalit, H., *Nucleic Acids Res.* **21**, 1512 (1993).]

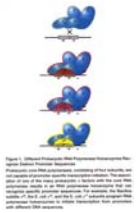
•The rates at which genes are transcribed vary directly with the rate that their promoters from stable initiation complexes with the holoenzyme

•The -10 and -35 regions of the promoter sequence are recognized by the sigma subunit of the RNA polymerase holoenzyme (which also includes two α and two β subunits

•Without the sigma subunit the RNA polymerase has no affinity for the DNA

•After entering the elongation phase of transcription, the sigma factor is removed from the polymerase complex

•Expression of different sigma factors makes it possible for a bacteria to efficiently respond to external stimuli (turn on sporulation genes, heat shock genes, etc.)



Extra

(Page adapted from S Strobel, Biochemistry Lecture Notes)

References

Argos P. (1976) Prediction of the secondary structure of mouse nerve growth factor and its comparison with insulin. *Biochemical and Biophysical Research Communications* 3:805-811.

Bairoch A and Apweiler R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res* 24:21-25.

Barton GJ. (1995) Protein secondary structure prediction. *Curr Opinion Struct Biol* 5:372-376.

Benner SA, Gerloff DL, and Jenny TF. (1994) Predicting protein crystal structures. *Science* 265:1642-1644.

Benner SA. (1995) Predicting the conformation of proteins from sequences. Progress and future progress. *J Mol Recogn* 8:9-28.

Boyd, D., Schierle, C. & Beckwith, J. (1998). How many membrane proteins are there? *Prot. Sci.* **7**, 201-205.

Crawford IP, Niermann T, and Kirschner K. (1987) Prediction of secondary structure by evolutionary comparison: application to the alpha subunit of thryptophan synthase. *Proteins: Struct Func Genet* 2:118-129.

Deleage G and Roux B. (1987) An algorithm for protein secondary structure prediction based on class prediction. *Protein Engineering* 4:289-294.

Eigenbrot C, Randal M, and Kossiakoff AA. (1992) Structural Effects Induced by Mutagenesis Affected by Crystal Packing Factors: the Structure of a 30-51 Disulfide Mutant of Basic Pancreatic Trypsin Inhibitor. *Proteins* 14:75.

Fasman, G. D. & Gilbert, W. A. (1990). The prediction of transmembrane protein sequences and their conformation: an evaluation. *Trends Biochem Sci* **15**, 89-92.

Frishman D, and Argos P. (1995) Knowledge-Base Protein Secondary Structure Assignment. *Proteins: Structure, Function, and Genetics* 23:566-79.

Frishman D, and Argos P. (1996) Incorporation of Non-Local Interactions in Protein Secondary Structure Prediction From the Amino Acid Sequence. *Protein Engineering* 2:in the press.

Frishman D, and Argos P. (1997) The Future of Protein Secondary Structure Prediction Accuracy. *Folding & Design* 2:159-62.

Frishman, D, and P Argos. (1996) 75% Accuracy in Protein Secondary Structure Prediction. Proteins 1997 Mar;27(3):329-335.

Garnier J and Levin JM. (1991) The protein structure code: what is its present status. *Comput Appl Biosci* 7:133-142.

Garnier, J. (1990). Protein structure prediction. *Biochimie* **72**, 513-24.

Garnier, J., Gibrat, J. F. & Robson, B. (1996a). GOR method for predicting protein secondary structure from amino acid sequence. *Meth. Enz.* **266**, 540-553.

References

Garnier, J., Gibrat, J. F. & Robson, B. (1996b). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* **266**, 540-53.

Garnier, J., Osguthorpe, D. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97-120.

Geourjon C, and Deléage G. (1995) SOPMA: Significant Improvements in Protein Secondary Structure Prediction by Consensus Prediction From Multiple Sequences. *Comput Appl Biosci* 11:681-84.

Gibrat, J., Garnier, J. & Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.* **198**, 425-443.

Gilbert RJ. (1992) Protein structure prediction from predicted residue prperties utilizing a digital encoding algorithm. *J Mol Graph* 10:112-119.

Holley LH, and Karplus M. (1989) Protein Secondary Structure Prediction With a Neural Network. *Proc Natl Acad Sci USA* 86:152-56.

Hunt NG, Gregoret LM, and Cohen FE. (1994) The origins of protein secondary structure. Effects of packing density and hydrogen bonding studied by a fast conformational search. *J Mol Biol* 241:214-225.

Kabsch W and Sander C. (1984) On the use of sequence homologies to predict protein structure. Identical pentapeptides can have completely different conformation. *Proc Natl Acad Sci USA* 81:1075-1078.

Kabsch W, and Sander C. (1983) Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 22:2577-637.

Kendrew JC, Klyne W, Lifson S, Miyazawa T, Nemethy G, Phillips DC, Ramachandran GN, and Sheraga HA. (1970). *Biochemistry* 9:3471-79.

King, R. D. & Sternberg, M. J. E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Prot. Sci.* **5**, 2298-2310.

King, R. D., Saqi, M., Sayle, R. & Sternberg, M. J. (1997). DSC: public domain protein secondary structure predication. *Comput Appl Biosci* **13**, 473-4.

Levin J, Pascarella S, Argos P, and Garnier J. (1993) Quantification of secondary structure prediction improvement using multiple alignments. *Protein Engineering* 6:849-854.

Levin JM, Robson B, and Garnier J. (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett* 205:303-308.

Levitt M, and Greer J. (1977) Automatic Identification of Secondary Structure in Globular Proteins. *J Mol Biol* 114:181-239.

Lim VI. (1974) Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J Mol Biol* 88:873-894.

Livingstone CD, Barton GJ (1996). Identification of functional residues and secondary structure from protein multiple sequence alignment. Methods Enzymol 266:497-512

Lupas A, Koster AJ, Walz J, and Baumeister W. (1994) Predicted secondary structure of the 20 S proteasome and model structure of the putative peptide channel. *FEBS Lett* 354:45-49.

Matthews B. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochem Biophys Acta* 405:442-451.

Mehta PK, Heringa J, and Argos P. (1995) A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Science* 4:2517-2525.

References

Muggleton S, King RD, and Sternberg MJE. (1992) Protein Secondary Structure Prediction Using Logic-Based Machine Learning. *Protein Engineering* 5:647-57.

Nishikawa K, and Ooi T. (1986) Amino Acid Sequence Homology Applied to the Prediction of Protein Secondary Structures, and Joint Prediction With Existing Methods. *Biochimica Et Biophysica Acta* 871:45-54.

Pauling L, Corey RB, and Branson HR. (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA* 37:205-211.

Persson, B. & Argos, P. (1997). Prediction of membrane protein topology utilizing multiple sequence alignments. *J Protein Chem* **16**, 453-7.

Presnell SR, Cohen BI, and Cohen FE. (1992) A segment-based approach to protein secondary structure prediction. *Biochemistry* 31:983-993.

Ptitsyn OB and Finkelstein AV. (1983) Theory of protein secondary structure and algorithm of its prediction. *Biopolymers* 22:15-25.

Qian N, and Sejnowski TJ. (1988) Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J Mol Biol* 202:865-84.

Rackovsky S. (1993) On the nature of the protein folding code. *Proc Natl Acad Sci U S A* 90:644-648.

Ramakrishnan C, and Soman KV. (1982) Identification of Secondary Structures in Globular Proteins - a New Algorithm. *Int J Pept Protein Res* 20:218-37.

Rao S, Zhu Q-L, Vaida S, and Smith T. (1993) The local information content of the protein structural database. *FEBS Lett* 2:143-146.

Rice CM, Fuchs R, Higgins DG, Stoehr PJ, and Cameron G N. (1993) The EMBL data library. *Nucleic Acids Res* 21:2967-2971.

Richards FM, and Kundrot CE. (1988) Identification of Structural Motifs From Protein Coordinate Data: Secondary Structure and First-Level Supersecondary Structure. *Proteins: Struct Func Genet* 3:71-84.

Robson B, and Garnier J. (1993) Protein Structure Prediction. *Nature* 361:506.

Rost B, and Sander C. (1993) Prediction of Protein Secondary Structure at Better Than 70% Accuracy. *J Mol Biol* 232:584-99.

Rost B, Sander C, and Schneider R. (1994) Redefining the goals of protein secondary structure prediction. *J Mol Biol* 235:13-26.

Rost B, Sander C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA* 90:7558-7562

Rost, B., Schneider, R. & Sander, C. (1993). Progress in protein structure prediction? *Trends Biochem Sci* **18**, 120-3.

Rumelhart DE, Hinton GE, and Williams R. (1986) Learning representations by back-propagating errors. *Nature* 323:533-536.

Salamov AA, and Solovyev VV. (1995) Prediction of Protein Secondary Structure by Combining Nearest-Meighbour Akgorithms Amd Multiple Sequence Alignments. *J Mol Biol* 247:11-15.

Salamov AA, and Solovyev VV. (1997) Protein Secondary Structure Prediction Using Local Alignments. *Journal of Molecular Biology* 268:31-36.

Sayle RA, and Milner-White EJ. (1995) RASMOL: Biomolecular Graphics for All. *Trends in Biochemical Sciences* 20:374-76.

References

Sayle RA, and Milner-White EJ. (1995) RASMOL: Biomolecular Graphics for All. *Trends in Biochemical Sciences* 20:374-76.

Sklenar H, Etchebest C, and Lavery R. (1989) Describing Protein Structure: a General Algorithm Yielding Complete Helicoidal Parameters and a Unique Overall Axis. *Proteins: Struct Func Genet* 6:46-60.

Solovyev VV and Salamov AA. (1994) Predicting alpha-helix and beta-strand segments of globular proteins. *Comput Appl Biosci* 10:661-669.

Stolorz P, Lapedes A, and Xia Y. (1992) Predicting Protein Secondary Structure Using Neural Net and Statistical Methods. *J Mol Biol* 225:363-77.

Sumpter BG, Getino C, and Noid DW. (1994) Theory and applications of neural computing in chemical science. *Ann Rev phys Chem* 45:439-481.

Sumpter BG, Getino C, and Noid DW. (19949 Theory and applications of neural computing in chemical science. *Ann Rev phys Chem* 45:439-481.

Taylor WR and Thornton JM. (1984) Recognition of super-secondary structure in proteins. *J Mol Biol* 173:487-5141984.

Thompson JD, Higgins DG, and Gibson TJ. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673-4680.

Thornton JM, Flores TP, Jones DT, and Swindells MB. (1991) Prediction of progress at last. *Nature* 354:105-106.

von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* **225**, 487-94.

Wasserman PD. (1989) *Neural Computing. Theory and Practice*. New York.

Zhang X, Mesirov JP, and Waltz DL. (1992) Hybrid System for Protein Secondary Structure Prediction. *J Mol Biol* 225:1049-63.

Zvelebil MJ, Barton GJ, Taylor WR, and Sternberg MJ. (1987) Prediction of Protein Secondary Structure and Active Sites Using the Alignment of Homologous Sequences. *J Mol Biol* 195:957-61.

End of Class 4 with 15' left