CSE 549

Lecturer: Sael Lee

AMINO ACID SEQUENCE ALIGNMENT II

Slides provided by courtesy of Dr. D. Kihara @ Purdue

SCORING MATRICES

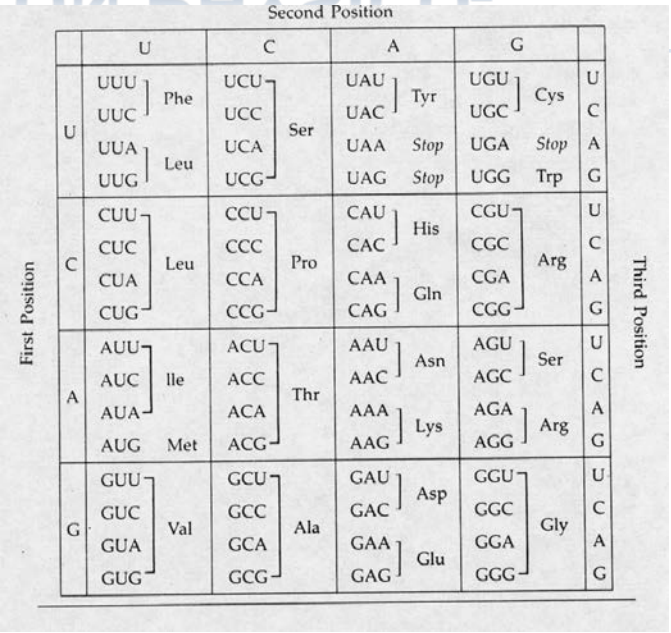
SCORING MATRICES FOR AA SEQUENCE ALIGNMENT

- Define scores for amino acid pairs in sequence alignments
- Reflect “similarity” of amino acid residues

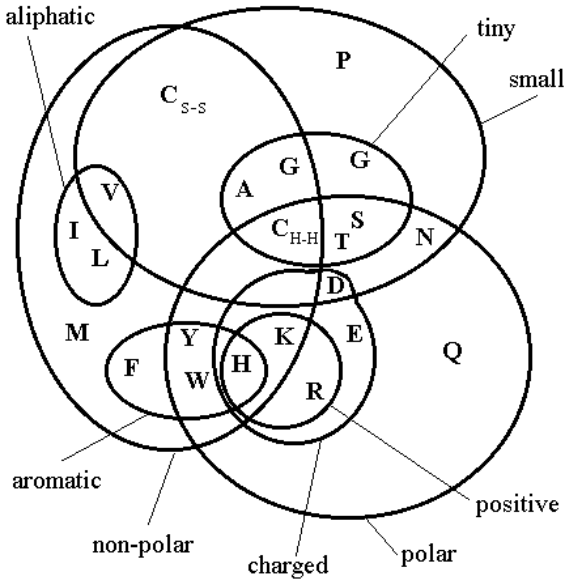
- *Amino acid scoring matrix/Amino acid similarity matrix* => symmetric
- *Amino acid substitution matrix* => not necessarily symmetric,
  - reflecting the difference of the mutation probability of A to B from B to A (A, B: two different amino acids)

SCORING MATRICES BASED ON PHYSICO-CHEMICAL PROPERTIES

- Identity Matrix
  - Same: 1, otherwise: 0
- Codon based
  - Similarity of tri-nucleotides coding each amino acid (next slide)
- Classification of amino acids



First Position	Second Position: U	C	A	G	Third Position
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	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



PAM MATRICES (DAYHOFF, 1978)

- PAM: A Point Accepted Mutations.
  - Models the replacement of a single AA in the primary structure of a protein with another single AA that is accepted by natural selection.
    - Does not include silent mutations , mutations which are lethal, or mutations which are rejected by natural selection in other ways.
- PAM matrix: 20x20 AA substitution matrix
  - Each entry indicates the likelihood of the AA of that row being replaced with the AA of that column through a series of one or more PAM during a specified evolutionary interval, compared to these two AA being aligned by chance.

PAM MATRIX CONT.

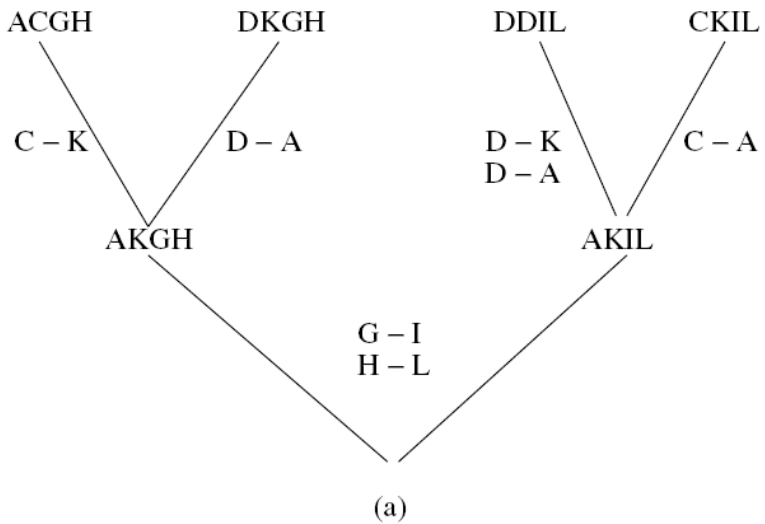
- Different PAM matrices correspond to different lengths of time in the evolution of the protein sequence.
  - EX> PAM1: one accepted mutation per 100 residues
  - (n in the PAM_n matrix represents the number of mutations per 100 amino acids,)
- Start from a set of well manually curated sequence alignments
  - >85% sequence identity
  - 71 groups of homologous sequences
- Construct phylogenetic trees and estimate the history of the mutation events in the family
  - 1572 observed mutations in the phylogenetic trees of 71 families of closely related proteins.

THE MODEL OF THE EVOLUTION

- The probability of a mutation in a position is independent on
  - Position and neighbour residues
  - Previous mutations in the position
- The biological (evolutionary) clock is assumed (meaning constant rate of mutations)
- This means that evolutionary time can be measured in number of mutations (here substitutions)

PAM: COLLECTION OF DATA FROM PHYLOGENETIC TREES

ACGH
DKGH
DDIL
CKIL



PAM SCORING MATRICES

	A	C	D	G	H	I	K	L
A		1	2					
C	1						1	
D	2						1	
G						1		
H								1
I				1				
K		1	1					
L					1			

(b)

Figure 5.4 (a) A small phylogenetic tree of four observed sequences, and two derived parent sequences. (b) The mutations are on the edges. The numbers of different mutations are shown in the table.

COMPUTING PROBABILITY OF *A* CHANGING TO *B* IN A CERTAIN TIME T

- Count for each branch in the phylogenetic trees, the number of mismatches recorded and compute fequencey
  - f_{ab} : frequency of mutation from *a* => *b or b* => *a* (assume symmetry i.e. $f_{ab} = f_{ba}$)
- Compute mutability of *a:* $f_a = \Sigma_{b \neq a} f_{ab}$
  - the total number of mutation involving *a*
- Compute $f = \Sigma_a f_a$:
  - <u>twice</u> the total number of mutations
- Compute p_a where $\Sigma_a p_a = 1$:
  - the frequency of amino acid *a,*
- Compute m_a : the relative mutability of *a*
  - the probability that *a* will mutate in the evolutionary time τ

CALCULATING M_A AND M_{AB} IN THE TIME T

- Consider the time τ = 1 PAM
  - the time while one mutation is accepted per 100 res.
- The probability that mutation is from *a* is:

  $$\tfrac{1}{2}\, f_a/(f/2) = f_a/f\,,$$

  (1/2 comes from $f_{ab} = f_{ba}$)
- Among 100 res., there are $100p_a$ occurrences of *a*
- The relative mutability of *a* is
  - $m_a = (1/\ 100p_a)\ f_a/f$
- The prob. that *a* will be mutated to *b* in the time τ
  - $M_{ab} = m_a\,(f_{ab}/f_a)$ for $a \neq b$; $M_{aa} = 1 - m_a$

SUBSTITUTION MATRIX M^1

Table 5.1 Substitution (mutation probability) matrix for the evolutionary distance of 1 PAM. To simplify the appearance, the elements are shown multiplied by 10 000. The probabilities for not changing are replaced by *, the values vary between 9822 (N) and 9976 (W). An element of this matrix, M_{ab}, gives the probability that the amino acid in row a will be replaced by the amino acid in column b after a given evolutionary interval, in this case 1 accepted point mutation per 100 amino acids. Thus there is a 0.56% probability that D (Asp) will be replaced by E (Glu). The amino acids are alphabetically ordered on their names. Reproduced from Dayhoff (1978) with permission of the National Biomedical Research Foundation.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	*	1	4	6	1	3	10	21	1	2	3	2	1	1	13	28	22	0	1	13
R	2	*	1	0	1	9	0	1	8	2	1	37	1	1	5	11	2	2	0	2
N	9	1	*	42	0	4	7	12	18	3	3	25	0	1	2	34	13	0	3	1
D	10	0	36	*	0	5	56	11	3	1	0	6	0	0	1	7	4	0	0	1
C	3	1	0	0	*	0	0	1	1	2	0	0	0	0	1	11	1	0	3	3
Q	8	10	4	6	0	*	35	3	20	1	6	12	2	0	8	4	3	0	0	2
E	17	0	6	53	0	27	*	7	1	2	1	7	0	0	3	6	2	0	1	2
G	21	0	6	6	0	1	4	*	0	0	1	2	0	1	2	16	2	0	0	3
H	2	10	21	4	1	23	2	1	*	0	4	2	0	2	5	2	1	0	4	3
I	6	3	3	1	1	1	3	0	0	*	22	4	5	8	1	2	11	0	1	57
L	4	1	1	0	0	3	1	1	1	9	*	1	8	6	2	1	2	0	1	11
K	2	19	13	3	0	6	4	2	1	2	2	*	4	0	2	7	8	0	0	1
M	6	4	0	0	0	4	1	1	0	12	45	20	*	4	1	4	6	0	0	17
F	2	1	1	0	0	0	0	1	2	7	13	0	1	*	1	3	1	1	21	1
P	22	4	2	1	1	6	3	3	3	0	3	3	0	0	*	17	5	0	0	3
S	35	6	20	5	5	2	4	21	1	1	1	8	1	2	12	*	32	1	1	2
T	32	1	9	3	1	2	2	3	1	7	3	11	2	1	4	38	*	0	1	10
W	0	8	1	0	0	0	0	0	1	0	4	0	0	3	0	5	0	*	2	0
Y	2	0	4	0	3	0	1	0	4	1	2	1	0	28	0	2	2	1	*	2
V	18	1	1	1	2	1	2	5	1	33	15	1	4	0	2	2	9	0	1	*

CALCULATE M^Z BY MATRIX MULTIPLICATION

Example Z=2

- 2 mutations per 100 residues
- A residue *a* can be changed to residue *b* after 2 PAM of following reasons:
  1. *a* is mutated to *b* in first PAM, unchanged in the next, with probability $M_{ab}M_{bb}$
  2. *a* is unchanged in first PAM, changed in the next, probability $M_{aa}M_{ab}$
  3. *a* is mutated to an amino acid *x* in the first PAM, and then to *b* in the next, probability $M_{ax}M_{xb}$, *x* being any amino acid unequal (*a*,*b*)

These three cases are disjunctive, hence

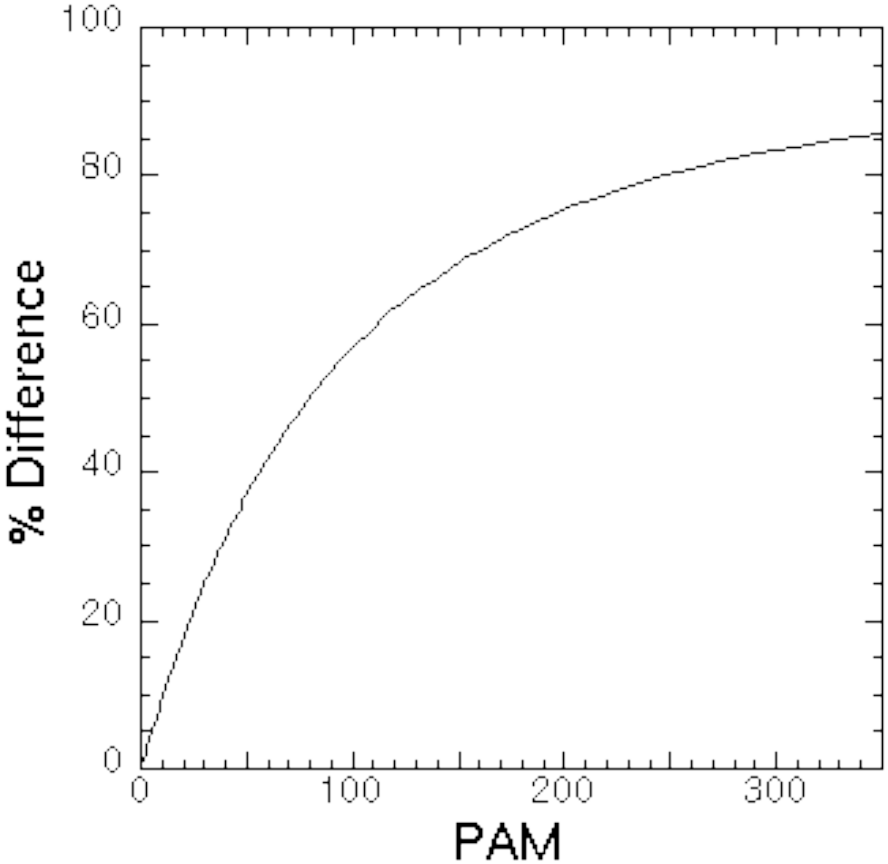
$$M_{ab}^2 = M_{ab}M_{bb} + M_{aa}M_{ab} + \sum_{x \notin \{a,b\}} M_{ax}M_{xb} = \sum_{x \in M} M_{ax}M_{xb}$$

M^{250}

Table 5.2 The mutation probability matrix for the evolutionary distance of 250 PAMs. To simplify the appearance, the elements are shown multiplied by 100. In comparing two sequences of average amino acid frequency at this evolutionary distance, there is a 13% probability that a position containing A (Ala) in the first sequence will contain A in the second. There is a 3% chance that it will contain R (Arg), and so forth. Reproduced from Dayhoff (1978) by permission of the National Biomedical Research Foundation.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	3	4	5	2	3	5	12	2	3	6	6	1	2	7	9	8	0	1	7
R	6	17	4	4	1	5	4	5	5	2	4	18	1	1	5	6	5	2	1	4
N	9	4	6	8	1	5	7	10	5	2	4	10	1	2	5	8	6	0	2	4
D	9	3	7	11	1	6	11	10	4	2	3	8	1	1	4	7	6	0	1	4
C	5	2	2	1	52	1	1	4	2	2	2	2	0	1	3	7	4	0	3	4
Q	8	5	5	7	1	10	9	7	7	2	6	10	1	1	5	6	5	0	1	4
E	9	3	6	10	1	7	12	9	4	2	4	8	1	1	4	7	5	0	1	4
G	12	2	4	5	2	3	5	27	2	2	3	5	1	1	5	9	6	0	1	5
H	6	6	6	6	2	7	6	5	15	2	5	8	1	3	5	6	4	1	3	4
I	8	3	3	3	2	2	3	5	2	10	15	5	2	5	3	5	6	0	2	15
L	6	2	2	2	1	3	2	4	2	6	34	4	3	6	3	4	4	1	2	10
K	7	9	5	5	1	5	5	6	3	2	4	24	2	1	4	7	6	0	1	4
M	7	4	3	3	1	3	3	5	2	6	20	9	6	4	3	5	5	0	2	10
F	4	1	2	1	1	1	1	3	2	5	13	2	2	32	2	3	3	1	15	5
P	11	4	4	4	2	4	4	8	3	2	5	6	1	1	20	9	6	0	1	5
S	11	4	5	5	3	3	5	11	3	3	4	8	1	2	6	10	8	1	2	5
T	11	3	4	5	2	3	5	9	2	4	6	8	1	2	5	9	11	0	2	7
W	2	7	2	1	1	1	1	2	2	1	6	4	1	4	1	4	2	55	3	2
Y	4	2	3	2	4	2	2	3	3	3	7	3	1	20	2	4	3	1	31	4
V	9	2	3	3	2	3	3	7	2	9	13	5	2	3	4	6	6	0	2	7

ESTIMATED SEQUENCE DIFFERENCE



- The number of differences in 100residues between two evolutionary related sequences over the time t can be estimated as

$$100(1-\sum_{a\in M} p_a M_{aa}^{\tau})$$

Amino acids

CONVERTING FROM A SUBSTITUTION MATRIX TO A SCORING MATRIX

- In a substitution matrix not symmetric in general,
  - $M_{ab} \neq M_{ba}$ (*a* in sequence q, *b* in sequence d)
- To remove the effect of the frequent occurrence of *b* in sequence d, the **odds scoring matrix** is
  - $O_{ab} = M_{ab}/p_b$
  - O_{ab} is symmetric ($O_{ab} = O_{ba}$, p. 110, middle)
- **Log-odds matrix R:**
  - $R_{ab} = \mathit{log}\ O_{ab}$

1PAM

A	7																			
R	-10	9																		
N	-7	-9	9																	
D	-6	-17	-1	8																
C	-10	-11	-17	-21	10															
Q	-7	-4	-7	-6	-20	9														
E	-5	-15	-5	0	-20	-1	8													
G	-4	-13	-6	-6	-13	-10	-7	7												
H	-11	-4	-2	-7	-10	-2	-9	-13	10											
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9										
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7									
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7								
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12							
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9						
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8					
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7				
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8			
W	-20	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-19	13		
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-8	10	
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

PAM-250 SCORING MATRIX

ble 5.3 Log-odds matrix for 250 PAMs. Elements are shown multiplied by 10. The neutral score is zero. A score of −10 means that the pair ›uld be expected to occur only one-tenth as frequently in related sequences as random chance would predict, and a score of +2 means that the pair ›uld be expected to occur 1.6 times as frequently. The order of the amino acids has been arranged to illustrate the patterns in the mutation data ·ouped according to the chemistries of the side groups). Reproduced from Dayhoff (1978) by permission of the National Biomedical Research undation.

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	−2	1	3																	
P	−3	1	0	6																
A	−2	1	1	1	2															
G	−3	1	0	−1	1	5														
N	−4	1	0	−1	0	0	2													
D	−5	0	0	−1	0	1	2	4												
E	−5	0	0	−1	0	0	1	3	4											
Q	−5	−1	−1	0	0	−1	1	2	2	4										
H	−3	−1	−1	0	−1	−2	2	1	1	3	6									
R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6								
K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5							
M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6						
I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5					
L	−6	−3	−2	−3	−2	−4	−3	−4	−3	−2	−2	−3	−3	4	2	6				
V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4			
F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9		
Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10	
W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17

BLOSUM (HENIKOFF & HENIKOFF)

- BLOSUM (BLOcks SUbstitution Matrix) matrix is a substitution matrix used to score alignments between evolutionarily divergent protein sequences introduced by Henikoff and Henikoff in 1992

- Make multiple alignments consist of sequences sharing more than X% sequence identity
- Discover blocks not containing gaps (used over 2,000 blocks)

```
...KIFIMK.......GDEVK...
...NLFKTR       GDSKK...
   KIFKTK       GDPKA
   KLFESR       GDAER
   KIFKGR       GDAAK
```

- For each column in each block, counted the number of occurrences of each pair of AA
  - 210 different pairs (combination with repetition: (20+2-1)! /(2!(20-1)!))

BLOSUM CONT

- A block of length *w* from an alignment of *n* sequences has $T=w*n(n-1)/2$ possible occurrences of amino acid pairs
  - Let h_{ab} be the number of occurrences of the pair (*ab*) in all blocks ($h_{ab}=h_{ba}$)
  - T total number of pairs
  - $f_{ab}=h_{ab}/T$
- Constructing logodds matrix : $R_{ab}=log(f_{ab}/e_{ab})$
  - with background probabilities of finding the amino acids a and in any protein sequence as p_a
  - $e_{aa}=p_a p_a$
  - $e_{ab}=p_a p_b + p_b p_a = 2\, p_a p_b$ *for* $a \neq b$

COMPARING PAM AND BLOSUM

- PAM: based on an evolutionary model (tree)
- PAM1 is multiplied to obtain PAMx (the larger x, the more distant)

- BLOSUM: Based on common regions in protein families
- Simple to compute
- BLOSUMx (e.g. x=45, 62, 80, the larger more closer)

ANALYSIS OF SCORING MATRICES

- PAMx or BLOSUMy is designed for aligning sequences of that range
  - i.e. BLOSUM50 cannot align very distantly related sequences by definition
- Starts from a set of pairwise (multiple) alignments
  - alignments > scoring matrix > alignment
- Can develop a scoring matrix from any set of alignments following the BLOSUM's method
- There are many AAindex database

http://www.genome.ad.jp/dbget/aaindex.html

REFERENCES

- Protein Bioinformatics, Chapter 5

- Tomii K, & Kanehisa M. “Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins”. Protein Engineering, 9: 27-36, 1996.

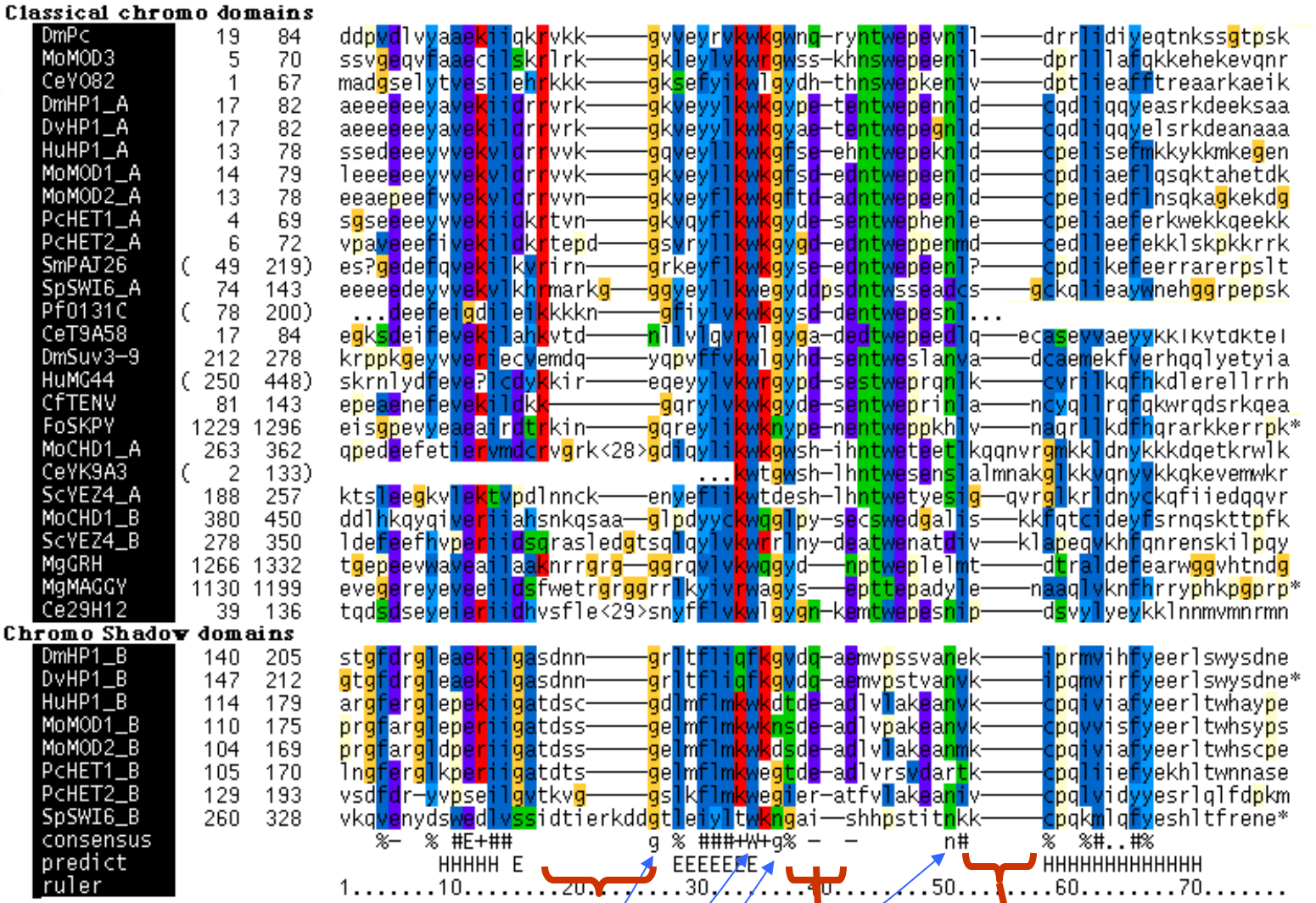
MULTIPLE ALIGNMENT

USE OF ALIGNMENTS

- High sequence similarity usually means significant structural and/or functional similarity.
- Homolog proteins (common ancestor) can vary significantly in large parts of the sequences, but still retain common 2D-patterns, 3D-patterns or common active site or binding site.
- Comparison of several sequences in a family can reveal what is common for the family. Conserved regions can be significant when regarding all of the sequences, but need not if regarding only two.

- Multiple alignment can be used to derive evolutionary history.
- Conserved positions : structurally/functionally important

Alignment of chromo domains



Aasland and Stewart, (1995) Nucl. Acids. Res. 23:3168-3173

Conserved positions

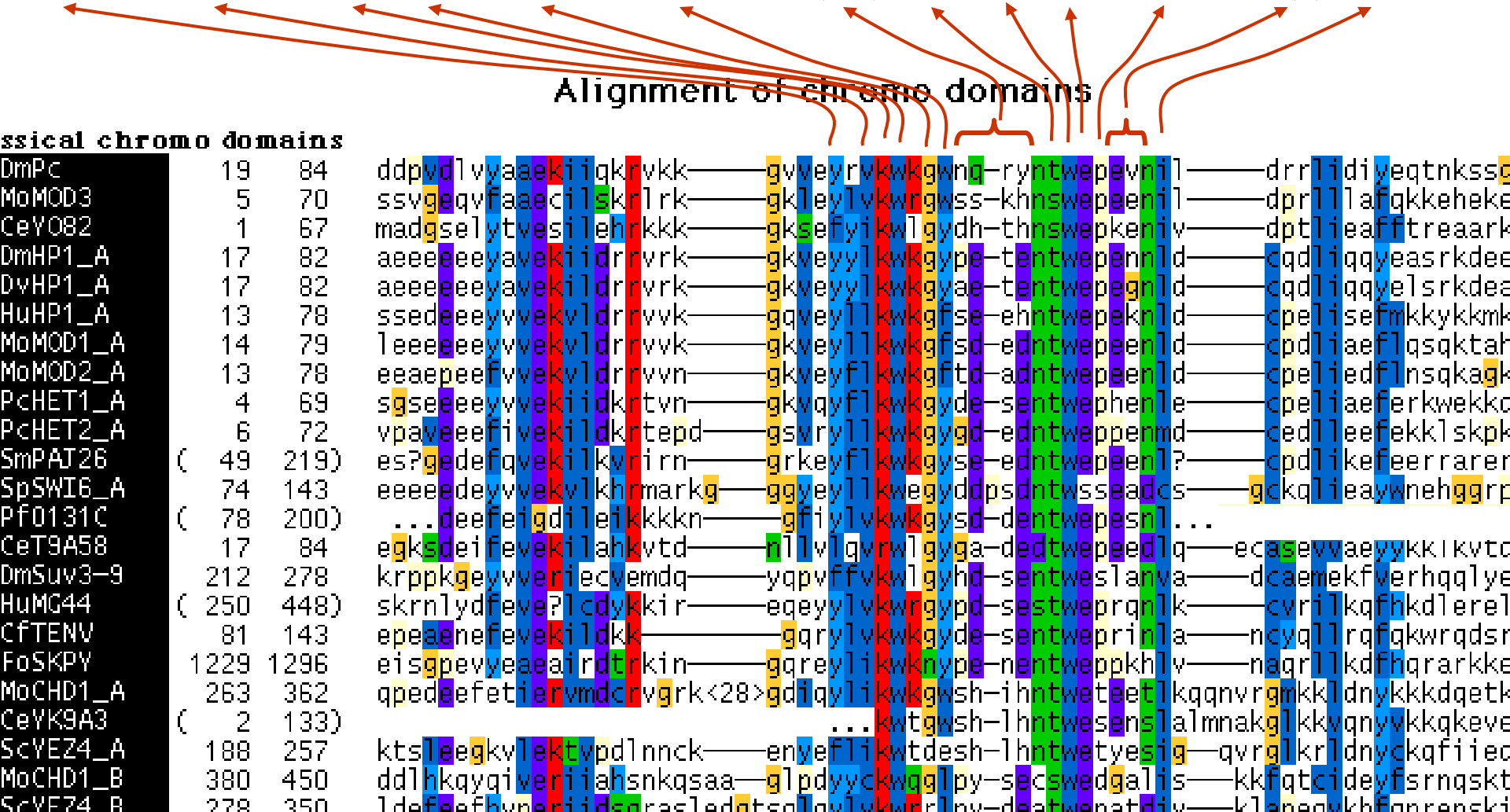
Loop? ***Loop?*** ***Loop?***

USE OF ALIGNMENTS - MAKE PATTERNS/PROFILES

- Can make a *profile* or a *pattern* that can be used to m atch against a sequence database and identify *new fa mily members*
- Profiles/patterns can be used to predict family memb ership of *new* sequences
- *Databases* of profiles/patterns
  - PROSITE
  - PFAM
  - PRINTS
  - …

PATTERN FROM ALIGNMENT

[FYL]-x-[LIVMC]-[KR]-W-x-[GDNR]-[FYWLE]-x(5,6)-[ST]-W-[ES]-[PSTDN]-x(3)-[LIVMC]



ALIGN BY USE OF DYNAMIC PROGRAMMING

- Dynamic programming finds *best* alignment of *k* sequences with given scoring scheme

- For two sequences there are three different column types

- For three sequences there are seven different column types

  x means an amino acid, - a blank

Sequence1	x	-	x	x	-	-	x
Sequence2	x	x	-	x	-	x	-
Sequence3	x	x	x	-	x	-	x

- Time complexity of $O(n^k)$ (sequence lengths = n)

SCORING MULTIPLE SEQUENCE ALIGNMENTS

Alignment

```
AR-L
ARSL
AWTL
AWT-
```

- Sum of the pairwise sequence score

$$S(MSA) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} S(s_i, s_j)$$

m: the number of sequences
s_i, s_j: sequence i, j
$S(s_i,s_j)$ = score of s_i,s_j

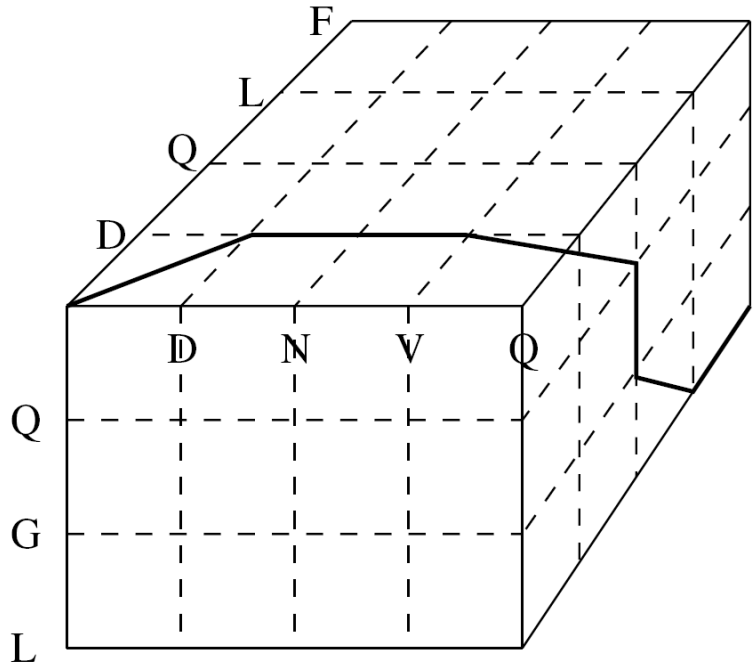
- Sum of scores for each row

$$S(MSA) = \sum_{k=1}^{r} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} R_{S_k^i S_k^j}$$

r: number of columns

USE OF K-DIMENTIONAL DYNAMIC PROGRAMMING

- Dynamic programming finds *best* alignment of *k* sequences given a scoring scheme

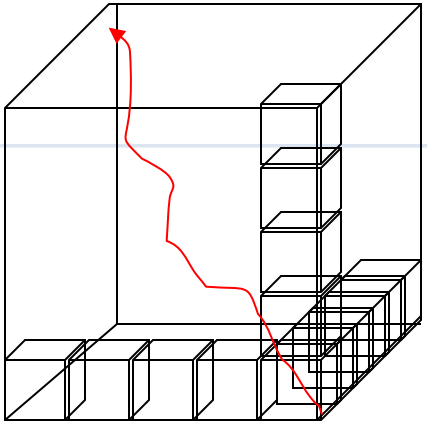


(a)

```
D--Q-LF
DNVQ---
---QGL-
```

(b)

MULTI-DIMENSIONAL DP



- 3 sequences:
  - Linear gap cost: $\gamma(d) = -gd$
  - Score of the whole MSA: $S(m) = \sum_i S(m_i)$

$$F(i,j,k) = \max \begin{cases} F(i-1,j-1,k-1) + S(x_i, y_j, z_k) \\ F(i,j-1,k-1) + S(-, y_j, z_k) \\ F(i-1,j,k-1) + S(x_i, -, z_k) \\ F(i-1,j-1,k) + S(x_i, y_j, -) \\ F(i-1,j,k) + S(x_i, -, -) \\ F(i,j-1,k) + S(-, y_j, -) \\ F(i,j,k-1) + S(-, -, z_k) \end{cases}$$

-d-d+s(y_j,z_k)
or
-d+s(y_j,z_k) etc.

-d-d+"s(-,-)"
or
-d + 0 etc.

MULTI-DIMENSIONAL DP: K SEQUENCES

$$F(i_1, i_2,,,, i_k) =$$

$$\max \begin{cases} F(i_1 - 1, i_2 - 1,,,, i_k - 1) + S(x_{i1}^1, x_{i2}^2,,, x_{ik}^k) \\ F(i_1, i_2 - 1,,,, i_k - 1) + S(-, x_{i2}^2,,, x_{ik}^k) \\ F(i_1 - 1, i_2,,,, i_k - 1) + S(x_{i1}^1, -,,, x_{ik}^k) \\ .... \\ F(i_1, i_2,,,, i_k - 1) + S(-, -, x_{i3}^3,,, x_{ik}^k) \\ .... \end{cases}$$

Complexity: O (n^k); if N=3, O(n^3)

REDUCING THE COMPUTATIONAL TIME BY A PRUNING ALGORITHM

- In order to obtain the optimal alignment, it is not necessary to calculating cells which certainly cannot lie on the best alignment path in the DP matrix.
- *dynamic pruning* – cells to avoid are found during the run
  - *forward recursion*
  - (*backward recursion* : conventional DP)