SEQUENCE COMPARISON

(2)

SEQUENCE COMPARISON

- BASED ON ALIGNMENT
  - IDENTITY
  - SIMILARITY
    - GAP
- BIOLOGICAL MEANING

IDENTITY, SIMILARITY, HOMOLOGY

- <u>Identity</u> is a measure of an alignment
  - sequence A can be "32 % identical to" sequence B
- <u>Similarity</u> is a measure of how close are two sequences
- <u>Homology is a qualitative trait: exists or does not exist</u>
  - evolutionary relation between sequences

ALIGNMENT SCORE

Quantitative evaluation of similarity

Must account for <u>the identity</u> of all the characters in both sequences <u>and the gap penalties</u>.

PERCENT SEQUENCE IDENTITY (PID)

THERE IS NO UNIVERSAL DEFINITION

- "PID1":100 * (identical positions) / (aligned positions + internal gap positions)
- "PID2":100 * (identical positions) / (aligned positions)
- "PID3":100 * (identical positions) / (length shorter sequence)

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGATAAC
           |   |        |      |
       CGAAATCGCATCAGCATACGATCGCATGC
```

PID= 4/29 *100

```
CGCTTCGGACGAAATCGCATCAGCATACGATCGCATGCCGGGCGGATAAC
         |||||||||||||||||||||||||||||
         CGAAATCGCATCAGCATACGATCGCATGC
```

PID= 29/29 *100

PERCENTAGE IDENTITY (PID).

PID is also strongly length dependent, so, the shorter a pair of sequences is, the higher the PID you might expect by chance.

GAP PENALTY

In the scoring of an alignment gaps are penalties

the gap opening penalty
>>
the gap extension penalty.

Similarity score

```
   IPLMTRWDQEQESDFGHKLP-IYTREWCTRG
   ||||||||| |||||||||| ||||||      Score = 25 - 2
CHKIPLMTRWDQ-QESDFGHKLPVIYTREW
```

gap creation penalty =-1

```
   IPLMTRWDQEQESDFGHKLP----IYTREWCTRG
   ||||||||| ||||||||||    ||||||
CHKIPLMTRWDQ-QESDFGHKLPVGSSIYTREW
```

gap extension penalty=-0.1

Score = 25 - 1-(1+(0.1*3)

SEQUENCE COMPARISON - SOME EMPIRICAL RULES

- Two protein sequences with more than 25 % identity are homologues
- Two DNA sequences with more than 70 % identity are homologues
- Homologous sequences have
  - A common ancestor (proteins and DNA)
  - A similar 3D structure (proteins)
  - Often a similar function (proteins)

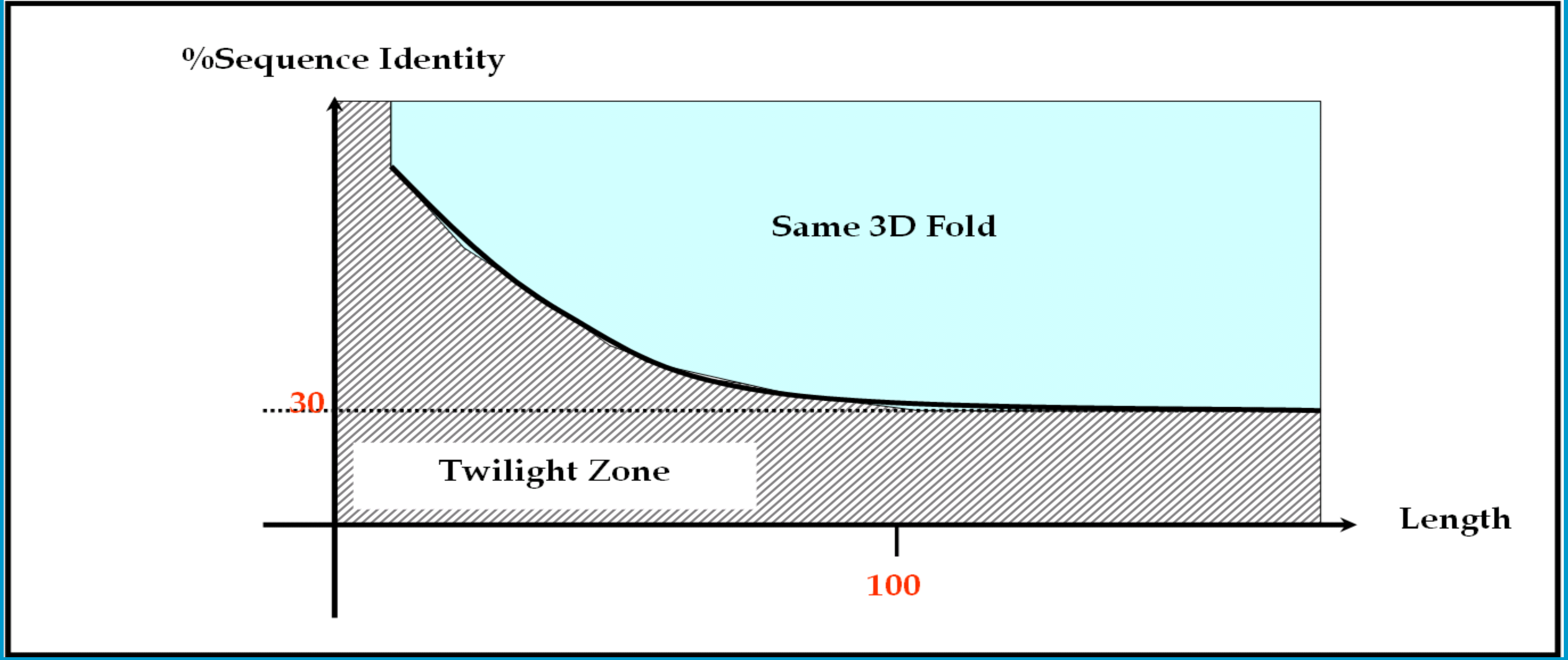
AT FIRST RELATION BETWEEN SEQUENCES SIMILARITY AND HOMOLOGY WAS EXPLORED IN PROTEINS

Homology

When two proteins have less than 25% identity

- They can be homologous or non-homologous
- Within this range of identity, it's impossible to say which is true

- This range of identity is called the "Twilight Zone"



However to establish homology between two sequences more precise tools are needed

- Dot plots for graphic analysis
- Quantitative measure of similarity
- Local or global alignment for residue/residue analysis

Dot plot

is the most basic method for comparing two sequences (pairwise comparison)

Is based on a visual approach and allows for

- general exploration
- discovering repeats
- finding long insertion or deletion
- comparison between any pair of sequences
  - DNA, Proteins, RNA

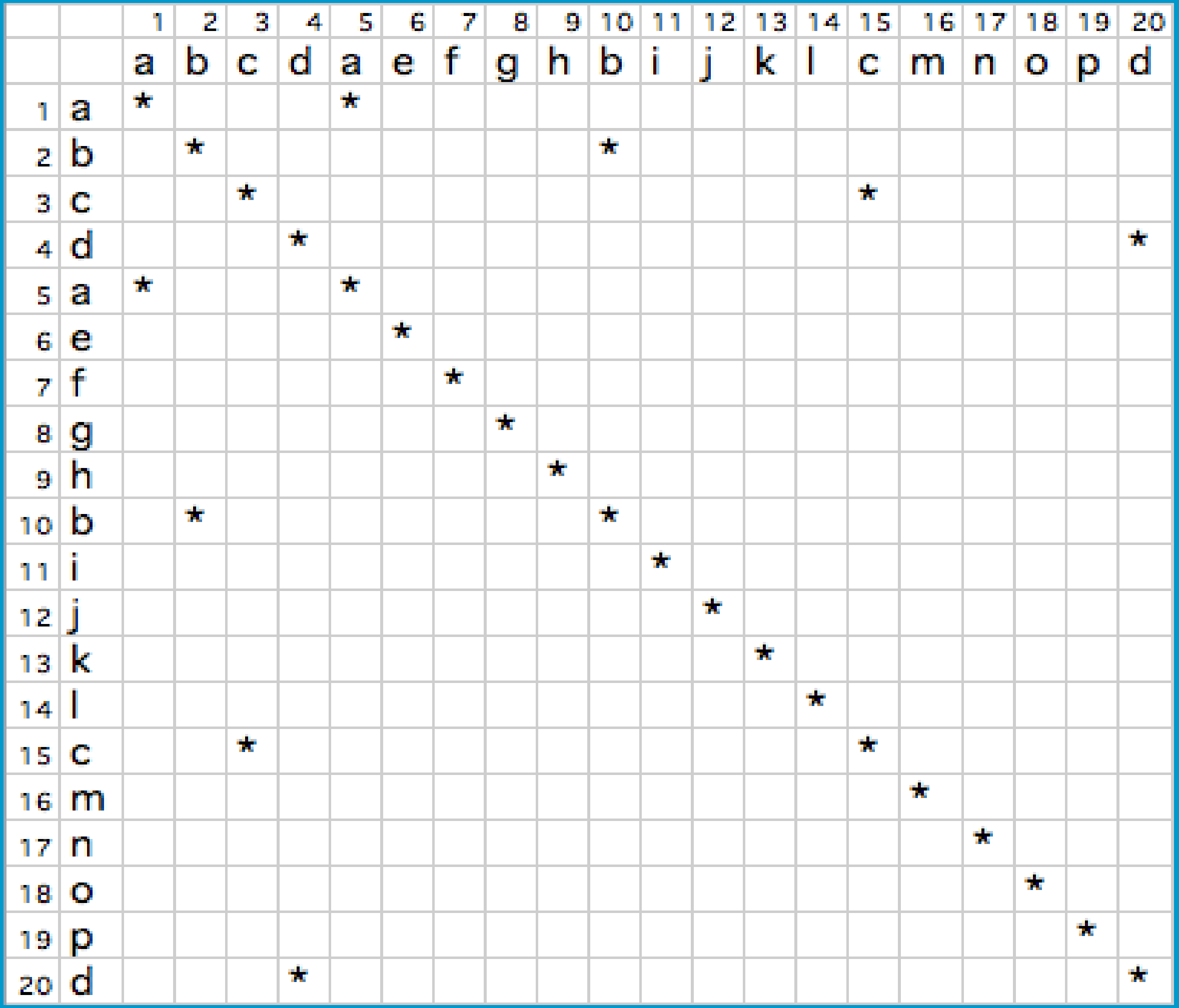
dot matrices

We can consider a matrix where the first or last row is sequence 1 (written from left to right) and first column is sequence 2 (written from high to bottom)

We can put a dot in each cell where a row character matches a column character.
In case of perfect identity we obtain a continuous diagonal line.

abcdaefghbijklcmnopd

abcdaefghbijklcmnopd



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
		a	b	c	d	a	e	f	g	h	b	i	j	k	l	c	m	n	o	p	d
1	a	*				*															
2	b		*								*										
3	c			*												*					
4	d				*																*
5	a	*				*															
6	e						*														
7	f							*													
8	g								*												
9	h									*											
10	b		*								*										
11	i											*									
12	j												*								
13	k													*							
14	l														*						
15	c			*												*					
16	m																*				
17	n																	*			
18	o																		*		
19	p																			*	
20	d				*																*

Dotplot showing identities between the strings

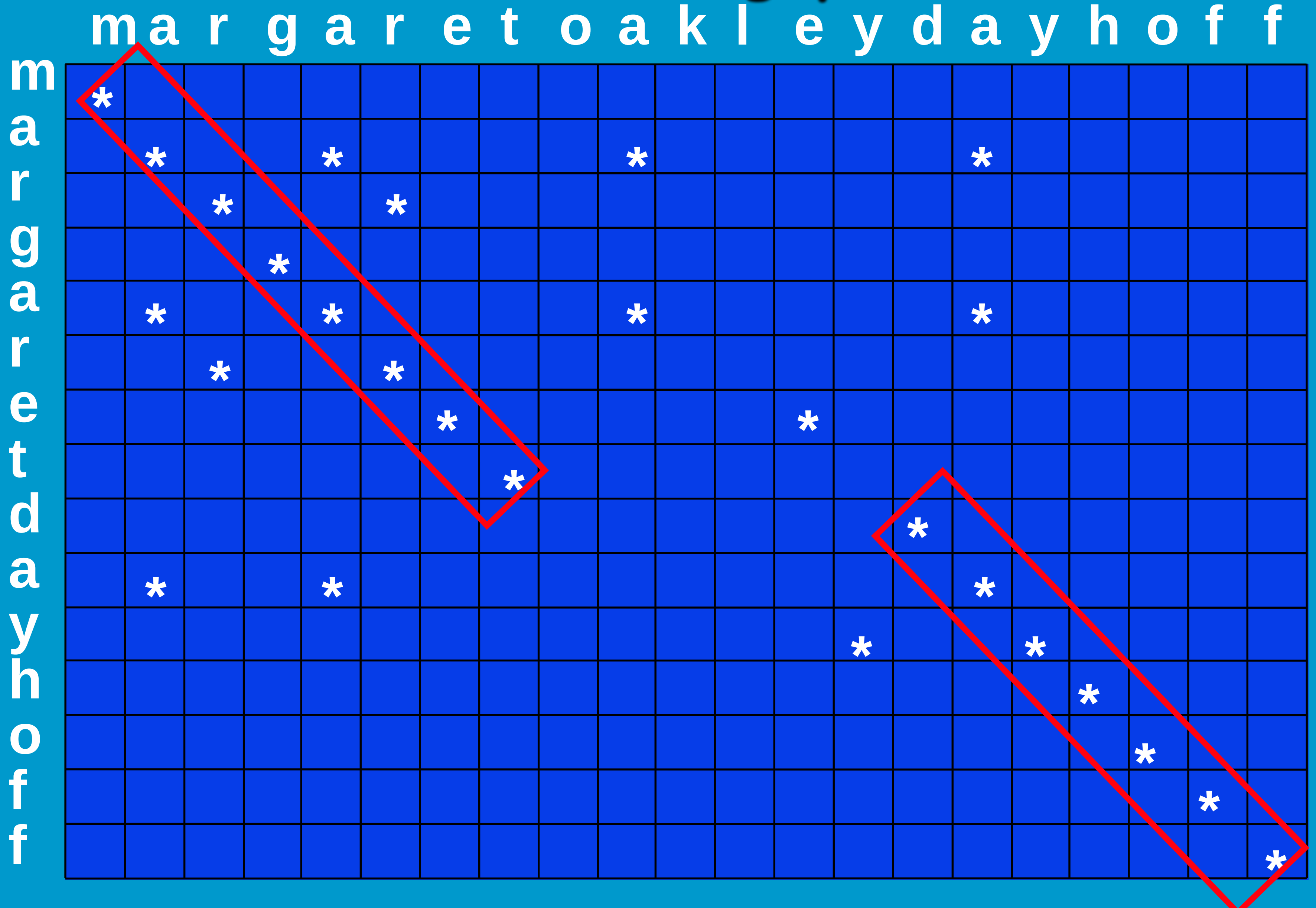
MARGARETOAKLEYDAYHOFF

MARGARETDAYHOFF

MARGARETOAKLEYDAYHOFF

MARGARET------DAYHOFF

Sequence similarity can be evidenced even with gaps.



	m	a	r	g	a	r	e	t	o	a	k	l	e	y	d	a	y	h	o	f	f
m	*																				
a		*			*					*						*					
r			*			*															
g				*																	
a		*			*					*						*					
r			*			*															
e							*						*								
t								*													
d															*						
a		*			*											*					
y														*			*				
h																		*			
o																			*		
f																				*	
f																					*

Two identical sequences are characterised by a single unbroken diagonal line accross the plot.

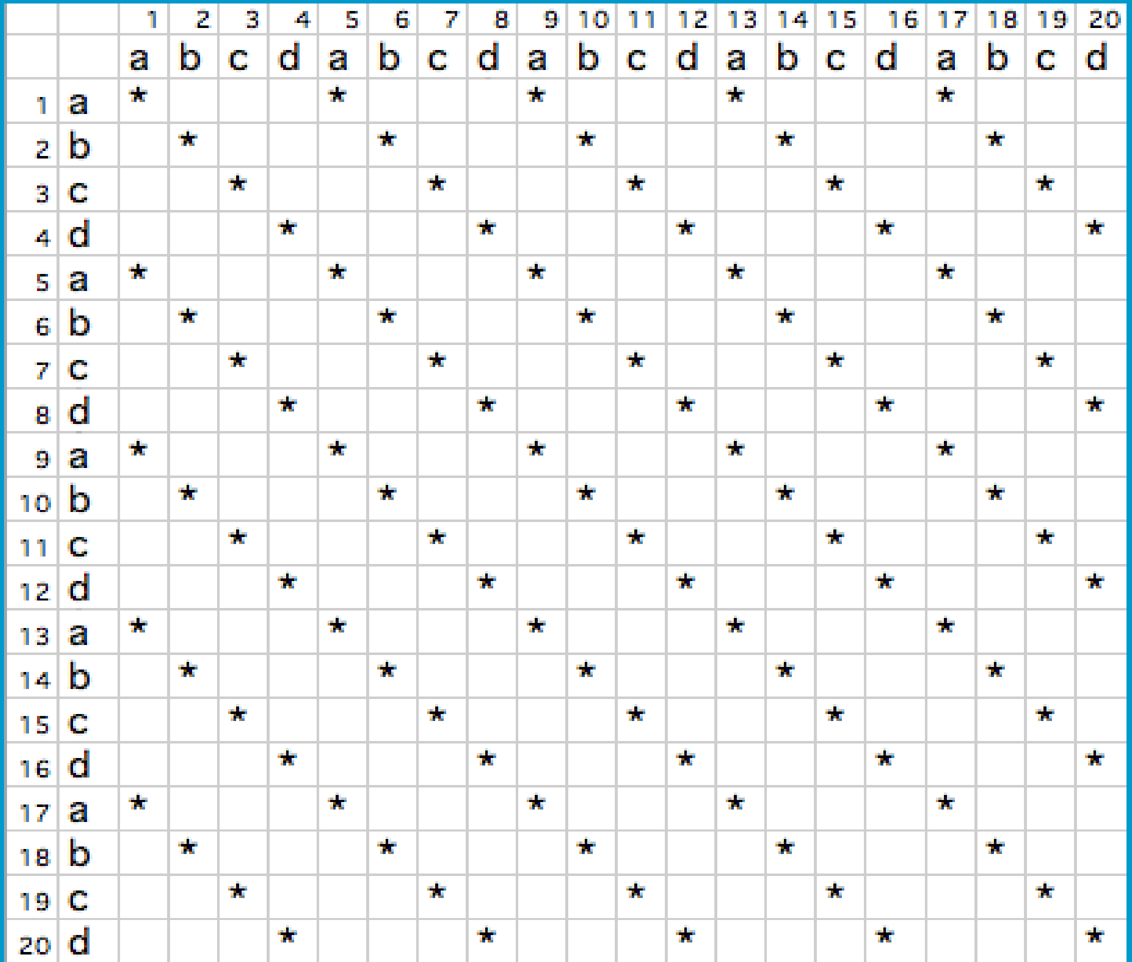
Two similar sequences will be characterised by a broken diagonal.

The interrupted regions indicate the locations of sequence mismatches

REPEATED SEQUENCES

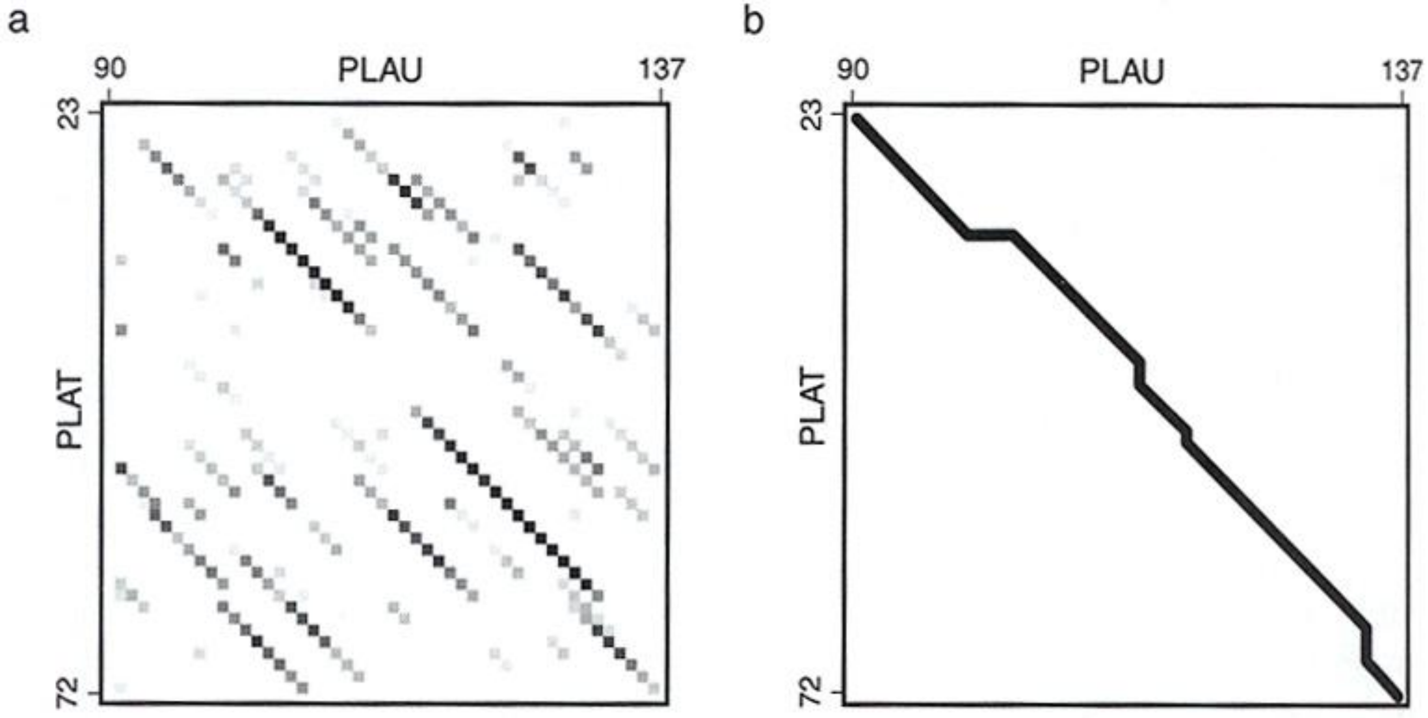
abcdabcdabcdabcdabcd

abcdabcdabcdabcdabcd



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
		a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d	a	b	c	d
1	a	*				*				*				*				*			
2	b		*				*				*				*				*		
3	c			*				*				*				*				*	
4	d				*				*				*				*				*
5	a	*				*				*				*				*			
6	b		*				*				*				*				*		
7	c			*				*				*				*				*	
8	d				*				*				*				*				*
9	a	*				*				*				*				*			
10	b		*				*				*				*				*		
11	c			*				*				*				*				*	
12	d				*				*				*				*				*
13	a	*				*				*				*				*			
14	b		*				*				*				*				*		
15	c			*				*				*				*				*	
16	d				*				*				*				*				*
17	a	*				*				*				*				*			
18	b		*				*				*				*				*		
19	c			*				*				*				*				*	
20	d				*				*				*				*				*

Some Typical Dot-plot Comparisons



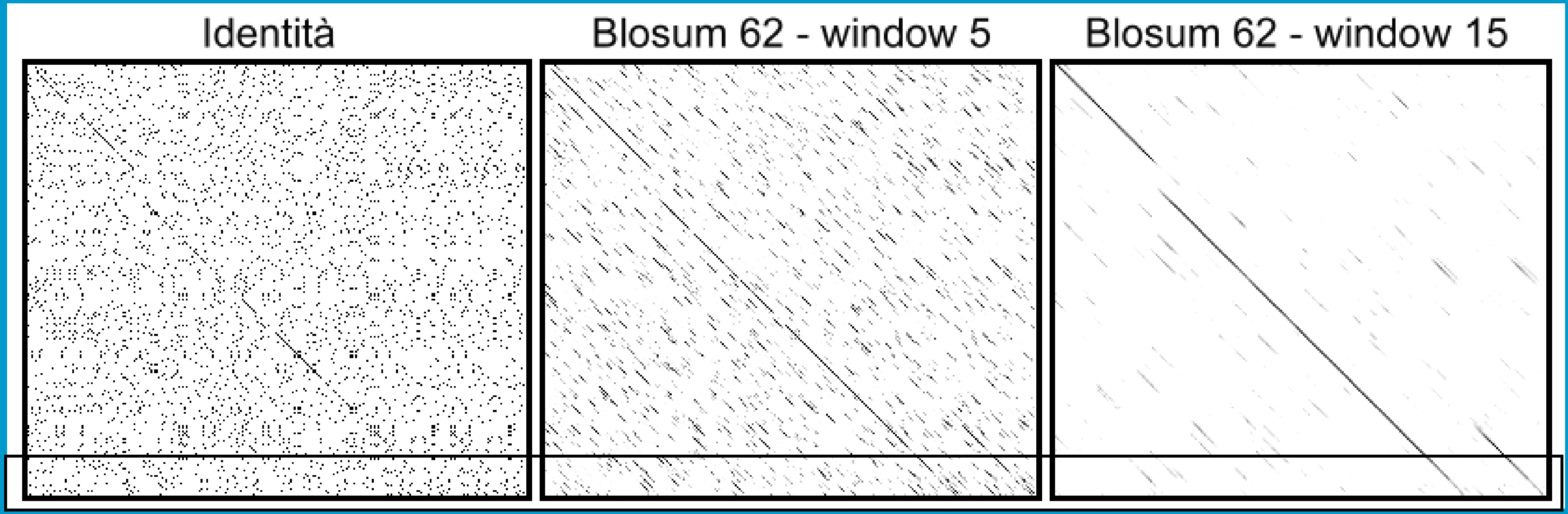
c

```
PLAU   90   EPKKVKDHCSKHSPCQKGGTCVNMP--SGPH-CLCPQHLTGNHCQKEK---CFE   137
PLAT   23   ELHQVPSNCD----CLNGGTCVSNKYFSNIHWCNCPKKFGGQHCEIDKSKTCYE   72
```

Figure 8.5. Dot-matrix, path graph, and alignment. All three views represent the alignment of the EGF similarity domains in the human urokinase plasminogen activator (PLAU; SWISS-PROT P00749) and tissue plasminogen activator (PLAT; SWISS-PROT P00750) proteins.

DOT PLOT MATRICES

More sophisticated dotplots exploit advanced scoring schemes which filter out noise through implementation of a sliding window to improve the signal/noise ratio



EVOLUTIONARY MEANING OF SIMILARITY AND DISTANCE

DISTANCE BETWEEN SEQUENCES

Corresponds to the sum of individual differences between a pair of sequences.

By an evolutionary point of view, distance gives a measure of the amount of evolutionary change between sequences, since divergence from a common ancestor.

SIMILARITY AND DISTANCE

- similarity scores are measures of similarity between two sequences:
  - Similar sequences → high scores

- distances are measures of dissimilarity between two sequences:
  - Similar sequences → small distances

Distance between sequences

Hamming distance between two strings of equal length is the number of positions with mismatching characters

Levenshtein or “edit distance” between two strings of not necessarily equal length is the minimal number of “edit operations” (insertion, deletion, alteration of a character) required to change one string into the other.

QUANTITATIVE MEASURES OF SIMILARITY

TO OBTAIN QUANTITATIVE MEASURES OF SIMILARITY

We need

- Algorithm searching for the highest alignment score
- Gap (indel) penalties
- Substitution scores for each pair of elements

SUBSTITUTION SCORE

quantifies the distance between sequences.

Assumptions:

- the sequences have an evolutionary ancestral sequence.
- The best guess is the path that requires the fewest evolutionary events.
- All substitutions are not equally likely and should be weighted to account for this.
- Insertions and deletions are less likely than substitutions and should be weighted to account for this.

THE DISTANCE

BETWEEN DNA SEQUENCES

Distance between DNA sequences

Simplest rules

All base changes (mismatches) are considered equally.

Each base substitution has the same weight. 12 possible substitution – each base can be replaced by one of the three other bases.

Gap penalty

Insertions/deletions (indels) are generally given a larger weight than replacements.

Indels of multiple bases at one position are given less weight than multiple independent indels.

In calculating the score for an alignment we evaluate the identities between characters and sum matches, mismatches, gap penalties.

When we assign value “1” to each match and “0” to mismatch we equally penalize all substitutions.

The substitution matrix is a matrix defining the substitution rules for a string of characters

The identity matrix is the simplest substitution matrix

Identity matrix

The simplest substitution matrix is one in which each character is considered similar to itself, but not able to transform into any other character.

This matrix is the identity matrix :

$$\begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 1 & 0 & 0 & 0 & 0 \\ B & 0 & 1 & 0 & 0 & 0 \\ C & 0 & 0 & 1 & 0 & 0 \\ D & 0 & 0 & 0 & 1 & 0 \\ E & 0 & 0 & 0 & 0 & 1 \end{array}$$

matches weigth 1 and mismatches weigth 0.

In the biological string the elements have a probability to be substituted by another element

In DNA sequences the probability of nucleotide substitution varies between transitions and transversions

transition mutations

purine (A) ↔ purine (G) and

pyrimidine (T) ↔ pyrimidine (C)

are more common than transversion

purine (A, G) ↔ pyrimidine (T, C)

A substitution matrix could reflect this

Substitution matrix for DNA sequences

Several models exist based on substitution matrices where differences are hypothetized

- In sustitution rate between nucletides
- In nucletide frequency

JC69 model (Jukes-Cantor, ‘69)

The Jukes-Cantor model assumes equal base frequencies and equal mutation rates, therefore it does not have any free parameter

$$Q = m_r \times \left(\begin{array}{ccccc} & A & C & G & T \\ A & * & 0.25 & 0.25 & 0.25 \\ C & 0.25 & * & 0.25 & 0.25 \\ G & 0.25 & 0.25 & * & 0.25 \\ T & 0.25 & 0.25 & 0.25 & * \end{array}\right)$$

K80 model (Kimura, ‘80)

The Kimura model assumes equal base frequencies and accounts for the difference between transitions and transversions with one parameter.

α transition = 1

α transversion = α_1

$$Q = m_r \times \left(\begin{array}{ccccc} & A & C & G & T \\ A & * & 0.25\alpha_1 & 0.25 & 0.25\alpha_1 \\ C & 0.25\alpha_1 & * & 0.25\alpha_1 & 0.25 \\ G & 0.25 & 0.25\alpha_1 & * & 0.25\alpha_1 \\ T & 0.25\alpha_1 & 0.25 & 0.25\alpha_1 & * \end{array}\right)$$

THE DISTANCE BETWEEN AMINO ACID SEQUENCES ACCOUNTS FOR VARIETY OF SCORING SCHEMES

Identity matrix

The simplest possible substitution matrix would be one in which each amino acid is considered similar to itself, but not able to transform into any other amino acid.

This matrix is an identity matrix:

$$\begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 1 & 0 & 0 & 0 & 0 \\ B & 0 & 1 & 0 & 0 & 0 \\ C & 0 & 0 & 1 & 0 & 0 \\ D & 0 & 0 & 0 & 1 & 0 \\ E & 0 & 0 & 0 & 0 & 1 \end{array}$$

Substitution matrix

The identity matrix will succeed in the alignment of very similar amino acid sequences but will be miserable at aligning two distantly related sequences.

We need to figure out all the substitution probabilities in a more rigorous fashion.

It turns out that an empirical examination of previously aligned sequences works best.

Amino acid substitution

- From the standpoint of the genetic code some amino acid changes are due to the replacement of a single DNA base, while others require two or even three changes in the DNA sequence.

Amino acid substitution

- From a functional standpoint, some amino acids can replace one another with relatively little effect on the structure and function of the final protein, while other replacements can be devastating.

Substitution score

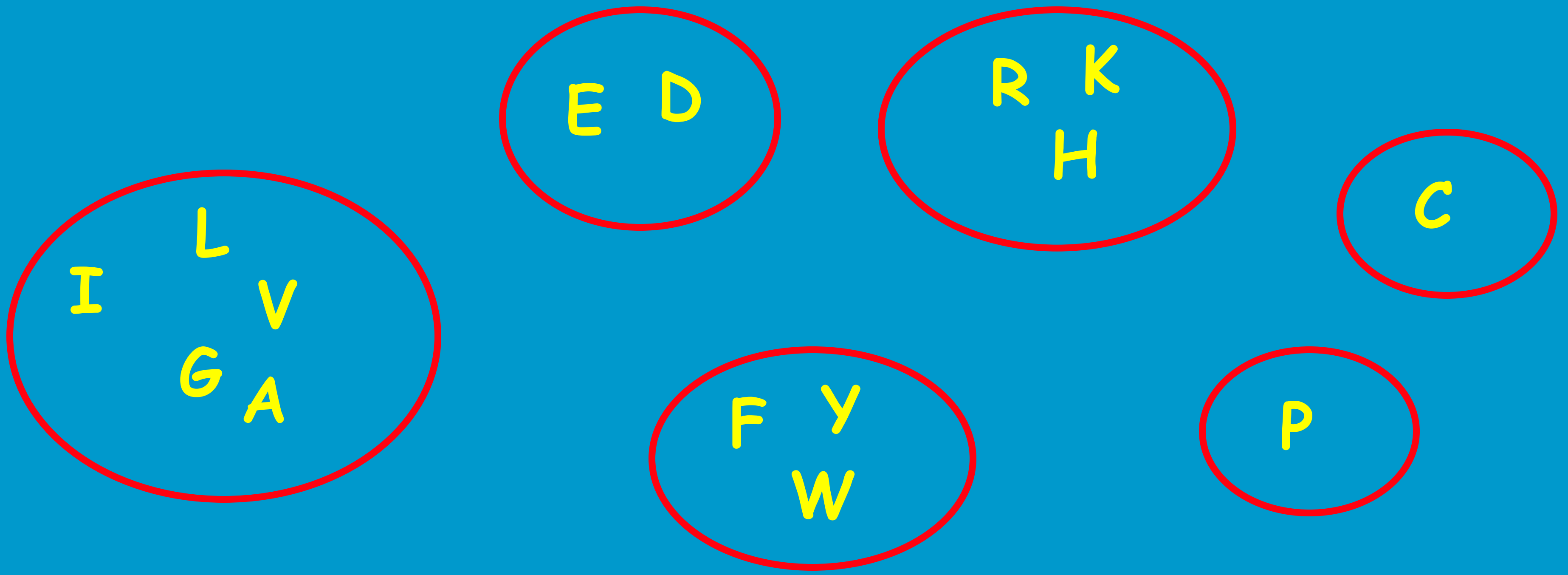
A substitution is more likely to occur between amino acids with similar biochemical properties.

For example:

substitution between the hydrophobic amino acids Isoleucine (I) and valine (V) occurs more frequently than

substitution between the hydrophobic amino acid Isoleucine with the hydrophilic amino acid Cystine (C).

Amino acid residue can be clustered according to similarity in chemico-physical characteristics.



Amino acid substitutions give different scores

There have been extensive studies looking at the frequencies in which amino acids substituted for each other during evolution. The studies involved carefully aligning all of the proteins in several families of proteins.

This can then be used to produce tables (scoring substitution matrices) of the relative frequencies with which amino acids replace each other over a short evolutionary period.

Distance between amino acid sequences

Changes in protein sequences during evolution tend to involve substitutions between aminoacids with similar properties which maintain the structural stability of the protein.

In scoring amino acid substitution we must take into account similarity between different even not identical residue (for example leucine and isoleucine).

We can take into account this similarity concept for each pair of the 20 naturally occurring amino acids and give a sustitution score to each pair of amino acids.

↓

Substitution Score Matrix

describes the likelihood that two residue types would mutate to each other in evolutionary time.

Why use a substitution matrix?

- Determine likelihood of homology between two sequences.

- Substitutions that are more likely should get a higher score,

- Substitutions that are less likely should get a lower score.

Scoring Matrices

- Log-odds matrix where each cell gives the probability of aligning two residues
- Score for each residue given by:

$$s(a \Leftrightarrow b) = \frac{1}{\lambda}\log(\frac{f_{a\Leftrightarrow b}}{f_a f_b})$$

- Score of alignment = Sum of log-odds scores of residues

Substitution score

A substitution is more likely to occur between amino acids with similar biochemical properties.

For example the hydrophobic amino acids Isoleucine(I) and valine(V) get a positive score on matrices adding weight to the likeliness that one will substitute for another.

While the hydrophobic amino acid Isoleucine has a negative score with the hydrophilic amino acid Cystine (C) as the likeliness of this substitution occurring in the protein is far less.

Thus matrices are used to estimate how well two residues of given types would match if they were aligned in a sequence alignment.

Substitution matrices

the most famous of these is the PAM 250 matrix, created by Margaret Dayhoff in 1978.

PAM stands for "percent accepted mutations"

it based on the assumption that as mutations accumulate the sequences diverge.

Collecting statistics from pairs of closely related sequences and correcting for different amino acid abundances produces 1PAM substitution matrix.

1 PAM = 1 Percent Accepted Mutation
99% identical residues

To produce a matrix appropriate for more widely divergent sequences we can take powers of this matrix

PAM (Percent Accepted Mutation) quantifies the amount of evolutionary change in a protein sequence.

- 1.0 PAM unit, is the amount of evolution which will change, on average, 1% of amino acids in a protein sequence.
- A PAM(x) substitution matrix is a table in which scores for each amino acid substitution have been calculated based on the frequency of that substitution in closely related proteins that have experienced a certain amount (x) of evolutionary divergence.

Amino acid scoring matrices are traditionally PAM matrices which refer to various degrees of sensitivity depending on the evolutionary distance between sequence pairs.

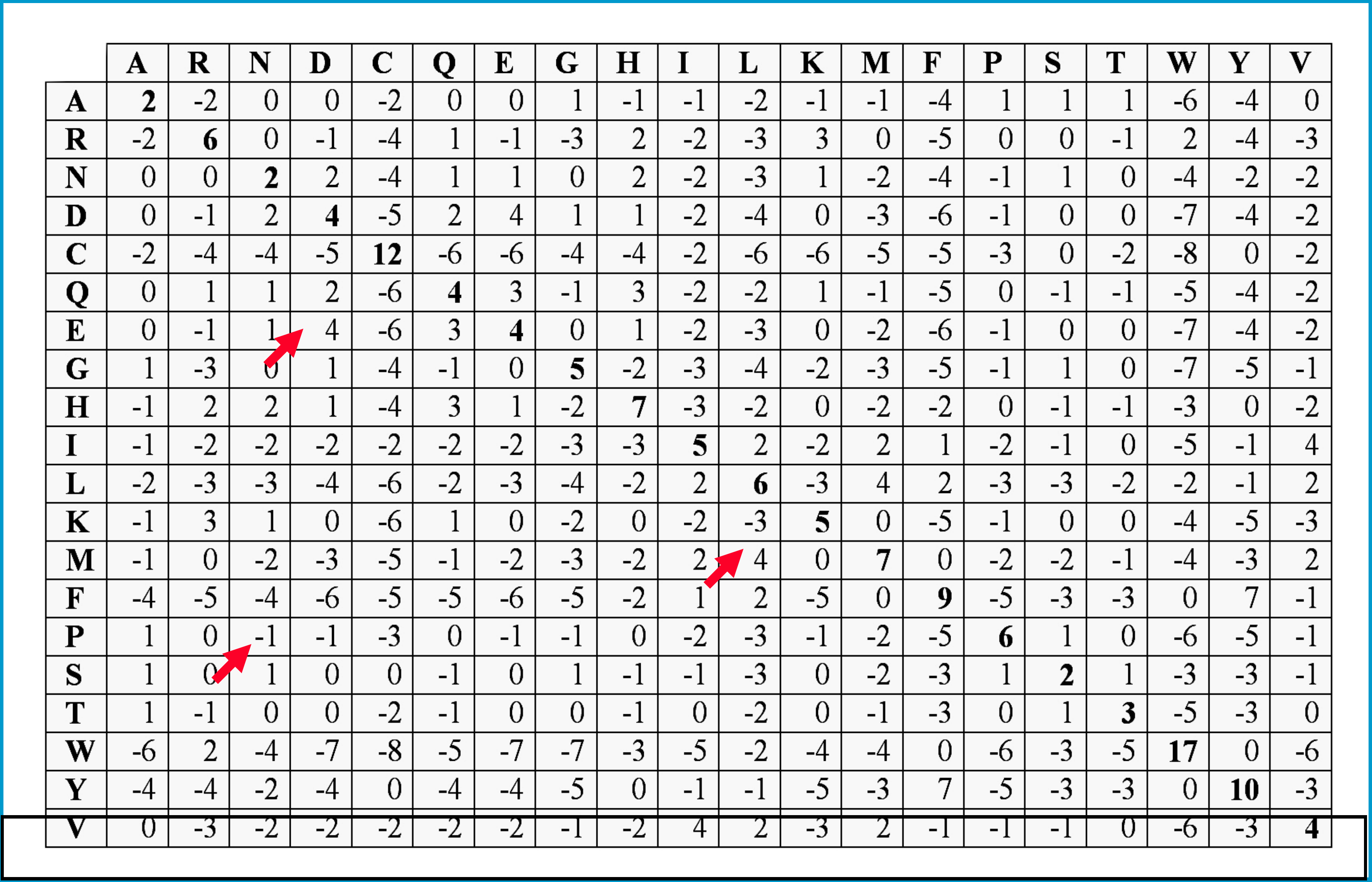
PAM	0	30	80	110	200	250
% IDENTITY	100	75	50	60	25	20

PAM (Percent Accepted Mutation) matrices

- The PAM matrices are based on mutations observed throughout a global alignment, this includes both highly conserved and highly mutable regions.
- PAM250 is for more distantly related sequences and is considered a good general matrix for protein database searching.

- For nucleotide sequence searching a simpler approach is used which either convert a PAM40 matrix into match/mismatch values which takes into consideration that a purine may be replaced by a purine and a pyrimidine by a pyrimidine.

substitution matrix PAM 250



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

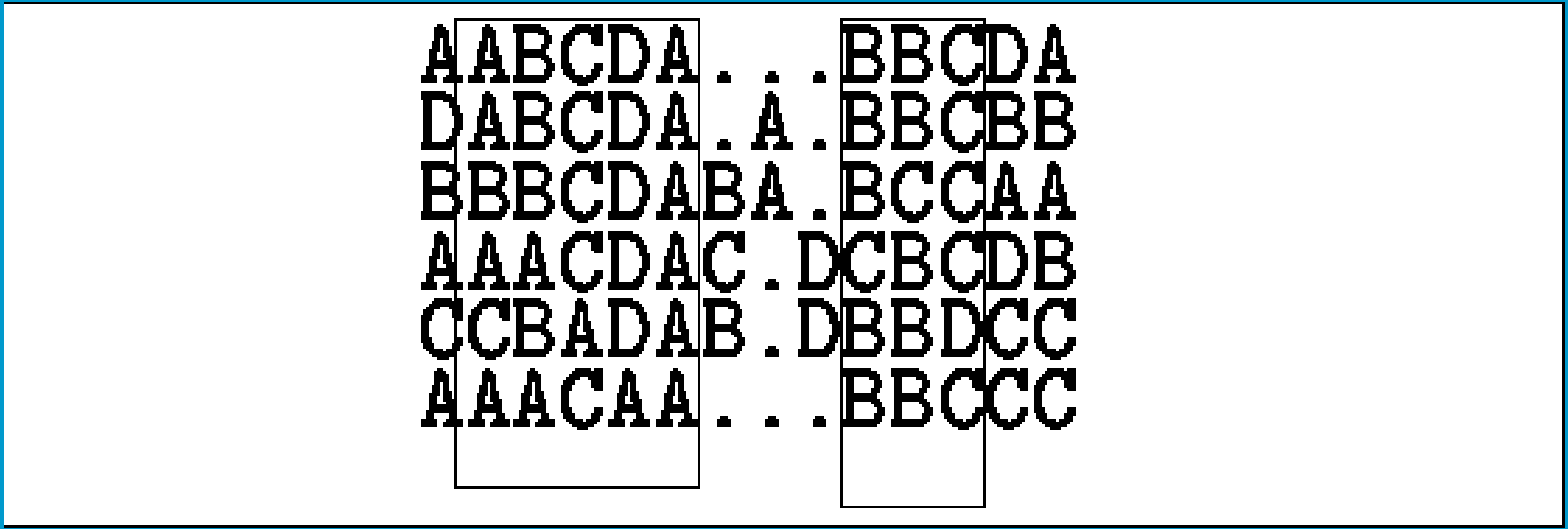
BLOSUM *(BLOck SUbstitution Matrix)*

- PAM methodology of comparing closely related species does not work very well for aligning evolutionarily divergent sequences.

- The BLOSUM series of matrices rectifies this problem. Henikoff and Henikoff (1992) constructed these matrices using multiple alignments of evolutionarily divergent proteins.

BLOSUM (BLOck SUbstitution Matrix)

The probabilities used in the matrix calculation are computed by looking at "blocks" of conserved sequences found in multiple protein alignments. These conserved sequences are assumed to be of functional importance within related proteins



```
AABCDA...BBCDA
DABCDA.A.BBCBB
BBBCDABA.BCCAA
AAACDAC.DCBCDB
CCBADAB.DBBDCC
AAACAA...BBCCC
```

BLOSUM MATRICES

- The Blosum matrices are based on the replacement found in more highly conserved regions of the sequences forbidden to contain gaps.
- These more highly conserved regions are those discovered in database searches and they serve as anchor points in alignments involving complete sequences.
- It is expected that the replacements that occur in more highly conserved regions will be more restricted than those that occur in highly variable regions of the sequence.

BLOSUM (BLOck SUbstitution Matrix)

- For the BLOSUM62 matrix, the identity threshold was set at 62%. One would use a higher numbered BLOSUM matrix for aligning two closely related sequences and a lower number for more divergent sequences.
- It turns out that the BLOSUM62 matrix does an excellent job detecting similarities in distant sequences, and this is the matrix used by default in most recent alignment applications.

BLOSUM62

A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	5															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8												
[illegible]	[illegible]	[illegible]	[illegible]	-3	-1	-3	-3	-4	-3	4											
[illegible]	[illegible]	[illegible]	[illegible]	-4	-1	-2	-3	-4	-3	2	4										
[illegible]	[illegible]	[illegible]	[illegible]	-1	-3	1	1	-2	-1	-3	-2	5									
[illegible]	[illegible]	[illegible]	[illegible]	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5				
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X

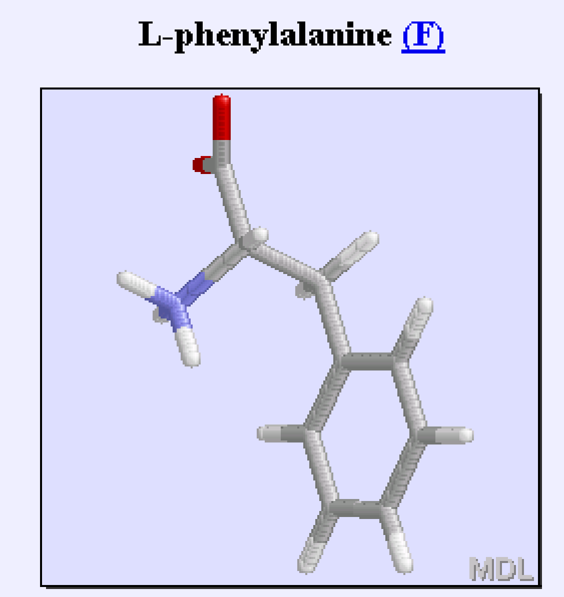
Positive for more likely substitutions

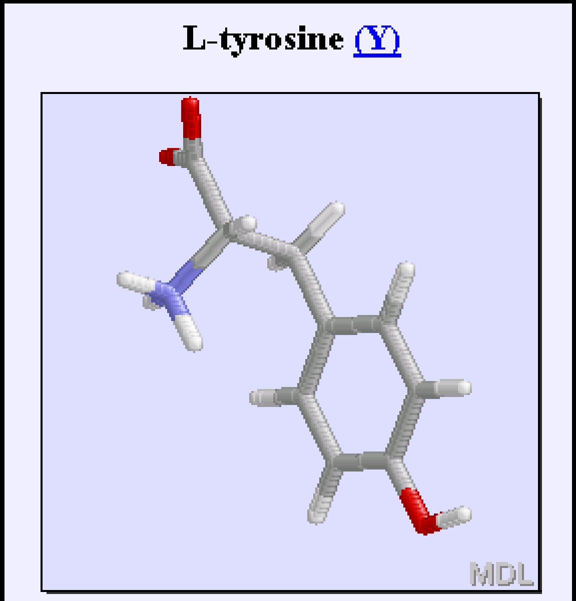
Common amino acids have low weights

Rare amino acids have high weights

BLOSUM62

A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	5															
E	-1																				
G	0																				
H	-2								8												
I	-1								-3	4											
L	-1								-3	2	4										
K	-1								-1	-3	-2	5									
M	-1								-2	1	2	-1									
F	-2								-1	0	0	-3									
P	-1								-2	-3	-3	-1									
S	1								-1	-2	-2	0									
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1									
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1		1	-1	-2	-2	0	-3	-1	4	
X	0	-1	-1	-											-2	0	0	-2	-1	-1	-1
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X


L-phenylalanine (F)

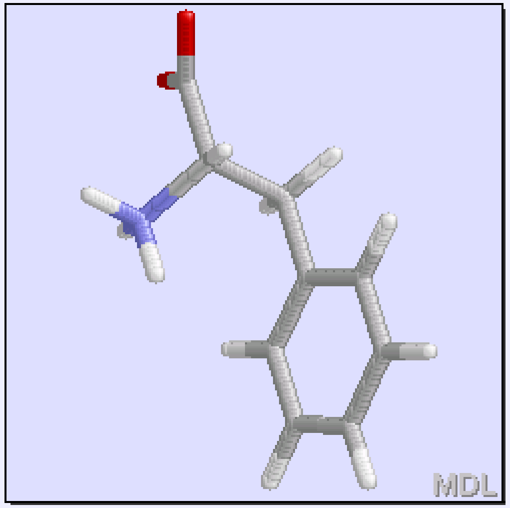

L-tyrosine (Y)

Positive for more likely substitutions

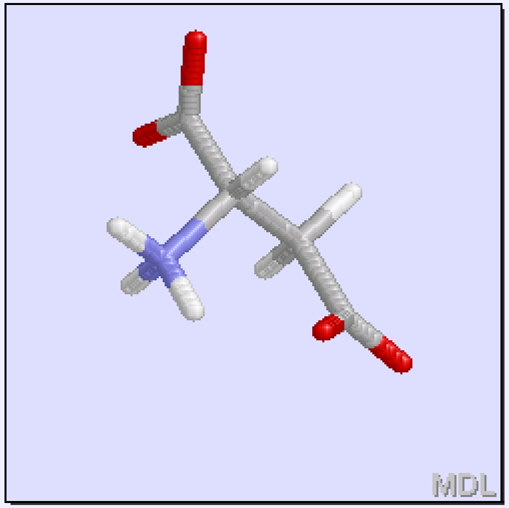
BLOSUM62

A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3																	
Q	-1	1	0	0																	
E	-1	0	0	2																	
G	0	-2	0	-1																	
H	-2	0	1	-1																	
I	-1	-3	-3	-3																	
L	-1	-2	-3	-4							4										
K	-1	2	0	-1							-2										
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1			0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T										-1	-1	-1	-1	-2	-1	1	5				
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X

L-phenylalanine (F)



L-aspartic acid (D)



Negative for less likely substitutions

- All substitution matrices derive from alignment procedure without gaps.
- Furthermore gap penalties for gap existence and extension must be considered.
  - Gap penalty bases on biological assumption and empirical alignments.
  - A reliable alignment can be tested by using different gap penalties.