Biochemistry 324

Bioinformatics

Pairwise sequence alignment

How do we compare genes/proteins?

- When we have sequenced a genome, we try and identify the function of **“unknown” genes** by finding a **similar gene** of **known function**
- To do this we need to **find “similar” genes**
- The similarity of genes is defined by the **similarity of sequences**
- Sequence similarities are obtained by **aligning** sequences
- **Homologous** sequences share an **evolutionary history**
  - **Homology is qualitative**, i.e. sequences are either homologous or they are not, they are not 25% homologous, for instance
- Homologous sequences have **identities**, degree of **conservation** and **similarities**, which are quantitative
  - **Identities** describe the percentage of residues that are identical at corresponding positions after alignment of the sequences
  - **Conservation** described the percentage of residues at corresponding positions that have similar physicochemical properties (i.e., polar, acidic, etc.)
  - **Similarity** describes the **conserved + identical** residues at corresponding positions in the aligned sequences

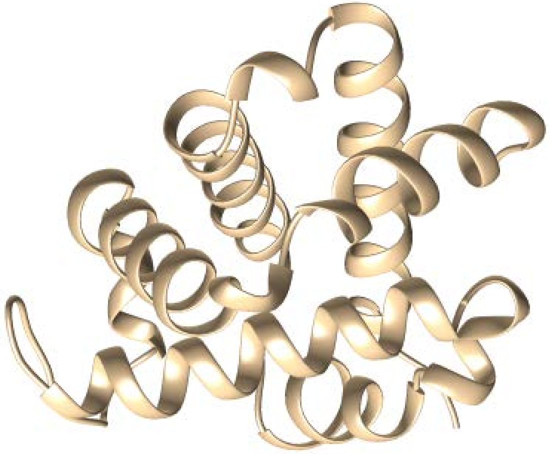
Homologies are often seen at the structural level

```
>AAR96398.1 hemoglobin beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRLFESFGDLFTPDAVMGNPKVKAHGKKVLG
AFSDGPAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH

>NP_976312.1 myoglobin [Homo sapiens]
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDLKKHGATVL
TALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFR
KDMASNYKELGFQG
```

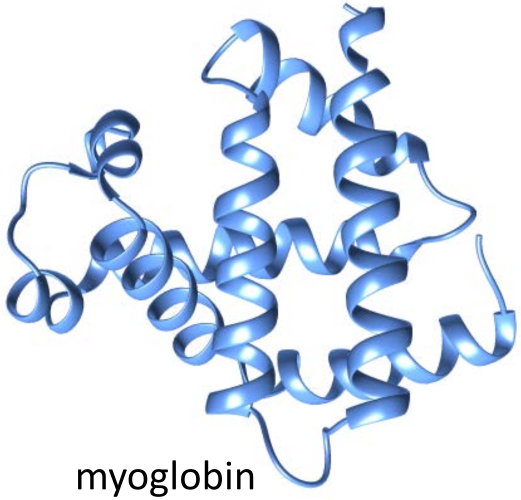
```
hemoglobin         1 MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRLFESFGD     48
                      :.|:..|...|..:||||..|  ..|.|.|.||...:|.|...|:.|..
myoglobin          1 -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKH     49

hemoglobin        49 LFTPDAVMGNPKVKAHGKKVLGAFSDGPAHLDNLKGTFATLSELHCDKLH     98
                     |.:.|.:..:..:|.||..||.|.........:.:.....|::.|..|..
myoglobin         50 LKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHK     99

hemoglobin        99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-    147
                     :..:....:...::.||......:|....|.|..|.:......:|..|.
myoglobin        100 IPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKE    149

hemoglobin       148 -----    147

myoglobin        150 LGFQG    154
```

```
# Identity:      36/155 (23.2%)
# Similarity:    57/155 (36.8%)
```

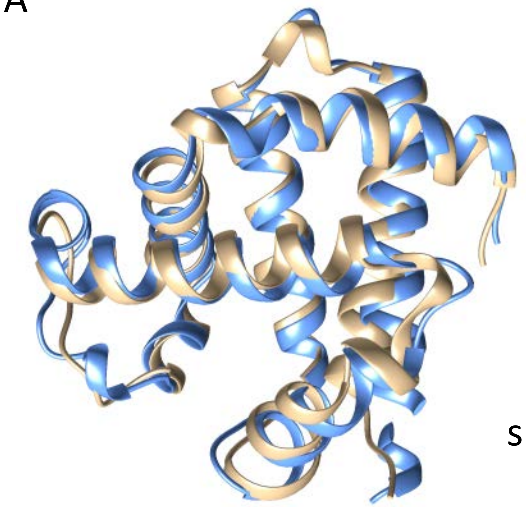
Protein structure is resistant to change at sequence level



β-hemoglobin chain A



myoglobin

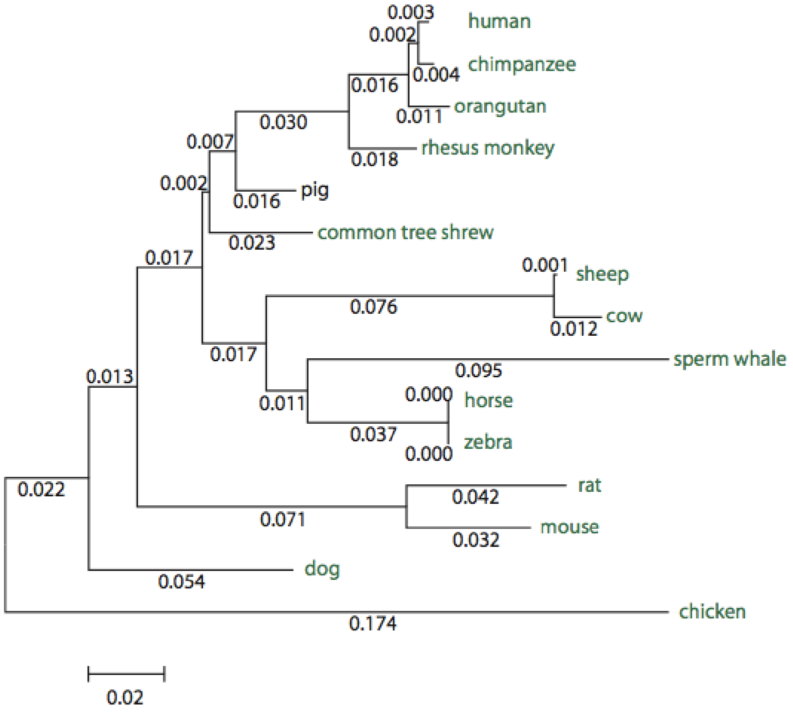


superimposed

Orthologous proteins

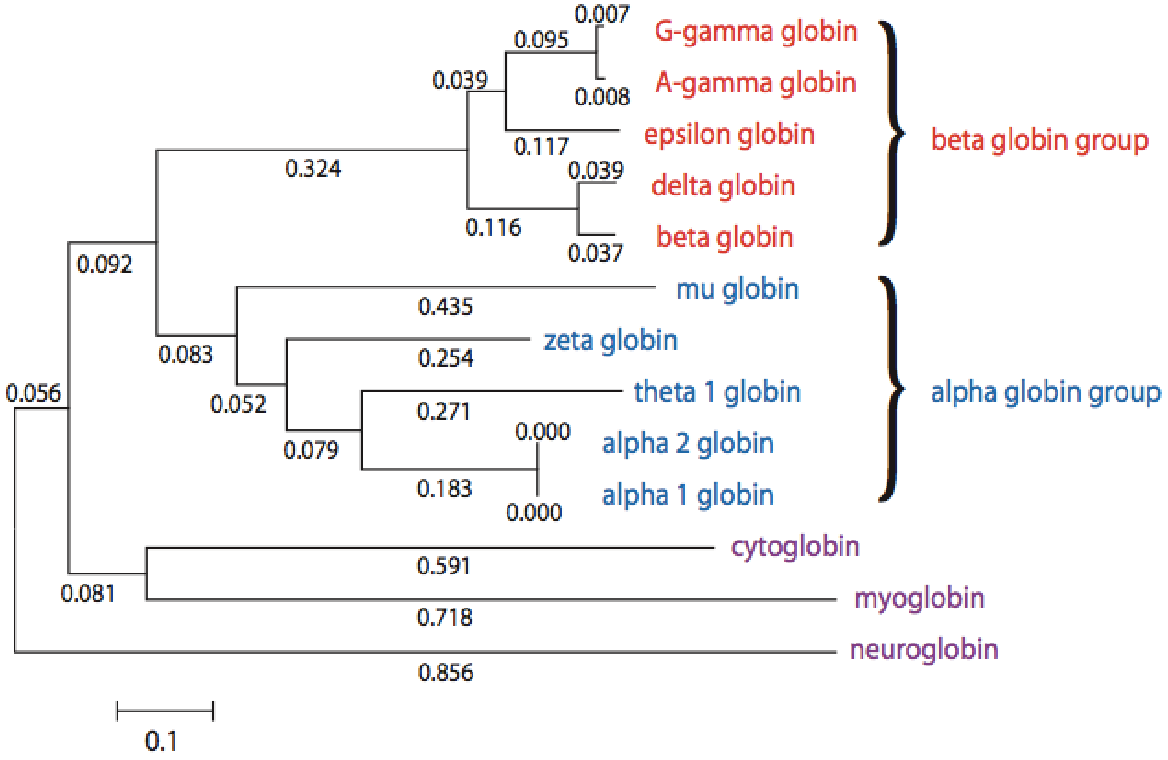
Orthologous proteins (or genes)
Homologous proteins that are found in **different species** that share a **common evolutionary ancestor**, and *may* have related functions

- Degree of similarity of myoblobin among different species



Paralogous proteins

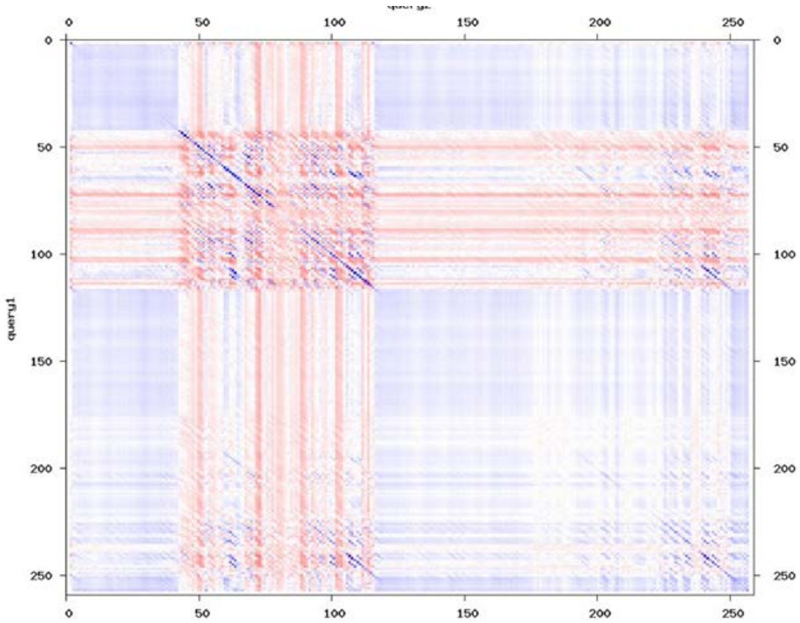
Paralogous proteins (or genes)
Homologous proteins that are coded by two genes in a **single genome** that **arose by gene duplication**, followed by gene drift



- The globin gene family in humans

Dotplots

- A **dotplot** is a quick way to **compare** two sequences
- **Residues or nucleotides** at the **intersect** of the vertical and horizontal sequences are indicated by **colours** to show **identity**, **conservation**, etc.
- **Diagonals** show **identity/conservation**
- The human brain is used to identify patterns in the dotplot that are interpreted as:
  - Repeats
  - Deletions
  - Inverted repeats



http://ffas.sanfordburnham.org

Global and local sequence alignments

- **Global alignment** is the optimal alignment of two or more sequences over the full length of all sequences, introducing gaps as needed to compensate for sequence length differences
- The **Needleman-Wunsch (“needle”) algorithm** performs global alignment
- **Local alignment** is the optimal alignment of short, local sequence lengths without any regard for the position of the aligned sequence within the large, full sequence
- The **Smith-Waterman (“water”) algorithm** performs local alignments
- Try the tools at http://www.ebi.ac.uk/Tools/emboss/

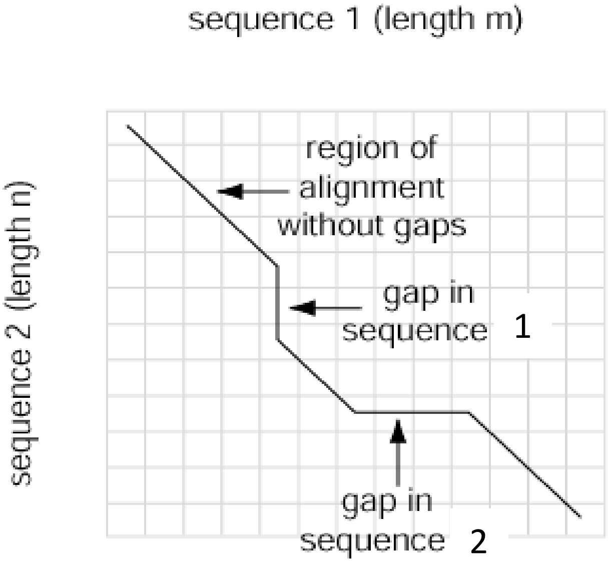
“Needle”

```
Beta-globin          1 MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFGD     48
                        :.|:..|...|..:||||..|  ..|.|.|.||...:|.|...|:.|..
Myoglobin            1 -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKH     49

Beta-globin         49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH     98
                       |.:.|.:..:..:|.||..||.|....|....:.:.....|::.|..|..
Myoglobin           50 LKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHK     99

Beta-globin         99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-    147
                       :..:....:...::.||......:|....|.|..|.:......:|..|.
Myoglobin          100 IPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKE    149

Beta-globin        148 -----    147

Myoglobin          150 LGFQG    154
```

“Water”

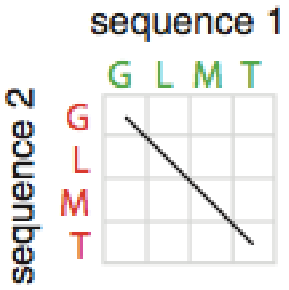
```
beta-globin          4 LTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFGDLST     51
                       |:..|...|..:||||..|  ..|.|.|.||...:|.|...|:.|..|.:
myoglobin            3 LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKS     52

beta-globin         52 PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDP    101
                       .|.:..:..:|.||..||.|....|....:.:.....|::.|..|..:..
myoglobin           53 EDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPV    102

beta-globin        102 ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKY    146
                       :....:...::.||......:|....|.|..|.:......:|..|
myoglobin          103 KYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNY    147
```

How do we generate a sequence alignment?

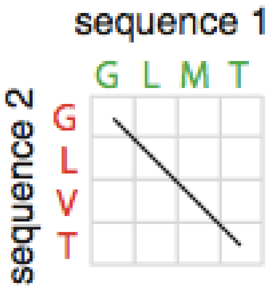


- Make a **matrix of size m+1 × n+1** for sequence 1 and 2 of **lengths m** and **n**
- When comparing 2 sequences, **trace a path through the matrix** with one sequence along the horizontal axis, and the other sequence along the vertical axis
- At every comparison, one of 4 results are possible:
  - Identical (stay on diagonal)
  - Mismatch (stay on diagonal)
  - Insert gap in sequence 1 (move along vertical)
  - Insert gap in sequence 2 (move along horizontal)

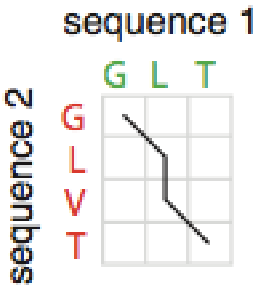
Four outcomes per comparison aligning 2 sequences



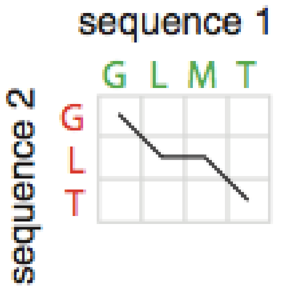
1 GLMT
2 GLMT



1 GLMT
2 GLVT



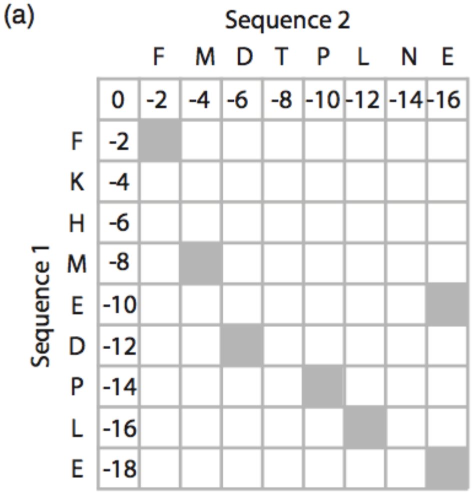
1 GL-T
2 GLVT



1 GLMT
2 GL-T

How do we find the alignment?

- Step 1: make a matrix where identical residues are indicated



(a)

Sequence 2

Sequence 1

		F	M	D	T	P	L	N	E
	0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2	▨							
K	-4								
H	-6								
M	-8		▨						
E	-10								▨
D	-12			▨					
P	-14					▨			
L	-16						▨		
E	-18								▨

Step 2: Define a scoring scheme:

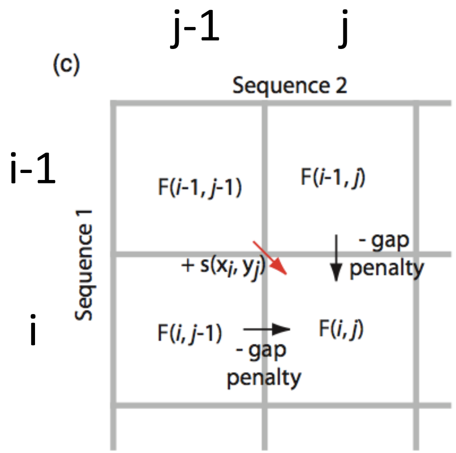
Match = +1

Mismatch = -2

Gap (horizontal or vertical) = -2

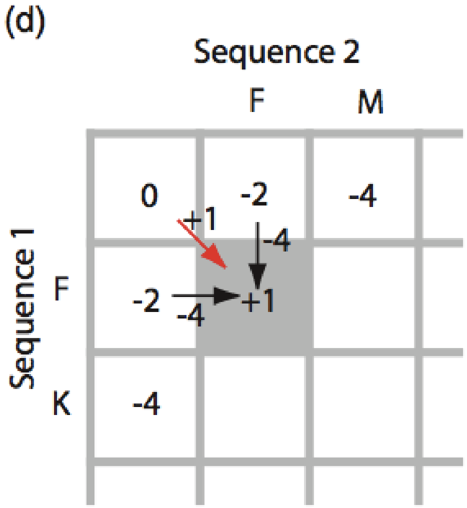
- Scores are **cumulative**
- Since we can **start with a gap** in either sequence 1 or 2, we indicate that with to top row and first column

Calculating the score as we complete the matrix

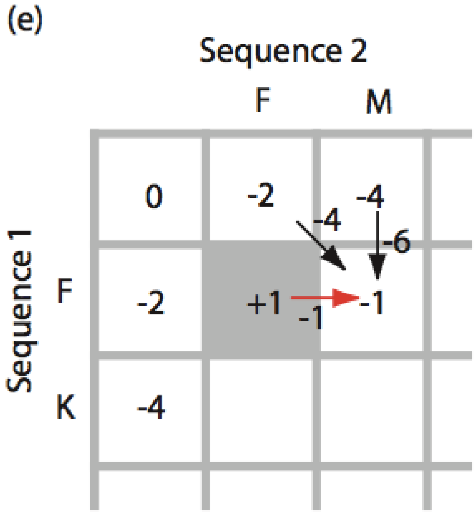


- When comparing two residues (or nucleotides) at position i,j they can be:
- Identical (score = +1) or a mismatch or gap (score = -2)
- We can arrive at position i,j from
  - (i-1,j-1), a previous match/mismatch
  - A gap in sequence 1, (i-1,j), or
  - A gap in sequence 2, (i,j-1)
- We now write at position (i,j) the **maximum** of
  - (i-1,j)+gap penalty (-2)
  - (i,j-1)+gap penalty (-2)
  - (i-1,j-1)+score (-2 for mismatch or +1 for match)

Choose the maximum score for a position, and move right

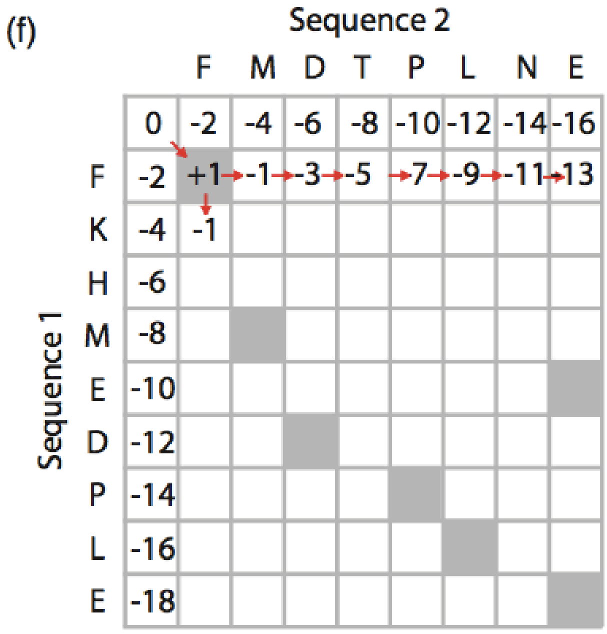


- When comparing the first residues F,F, we have a match, i.e., value = +1
- We can arrive at a score
  - Diagonally 0+1=+1
  - Vertically -2-2=-4
  - Horizontally -2-2=-4
- The maximum score is +1
- Choose +1



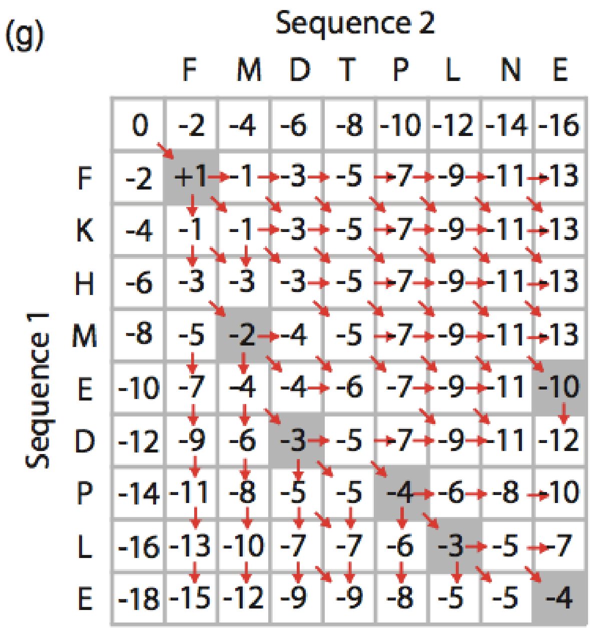
- Now, proceed along the row
- You can arrive at the next block F,M:
  - From a match, introducing a gap (+1-2=-1)
  - From a gap, introducing a a mismatch (-2-2=-4)
  - From a gap, introducing a gap (-4-2=-6)
- Maximum score = -1

Fill in each row in turn



(f)

		F	M	D	T	P	L	N	E
	0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2	+1	-1	-3	-5	-7	-9	-11	-13
K	-4	-1							
H	-6								
M	-8								
E	-10								
D	-12								
P	-14								
L	-16								
E	-18								

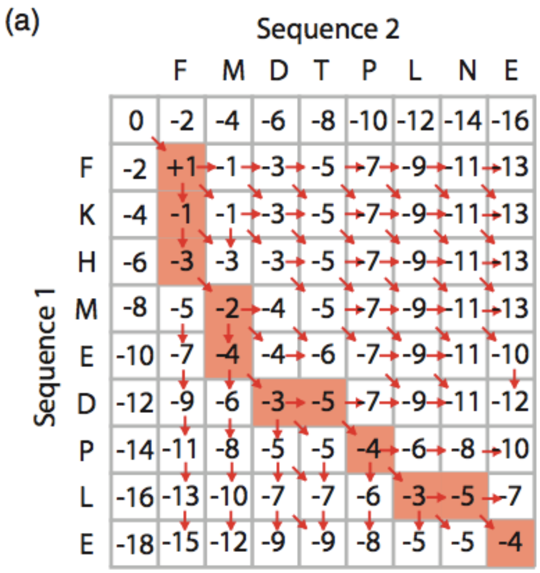


(g)

		F	M	D	T	P	L	N	E
	0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2	+1	-1	-3	-5	-7	-9	-11	-13
K	-4	-1	-1	-3	-5	-7	-9	-11	-13
H	-6	-3	-3	-3	-5	-7	-9	-11	-13
M	-8	-5	-2	-4	-5	-7	-9	-11	-13
E	-10	-7	-4	-4	-6	-7	-9	-11	-10
D	-12	-9	-6	-3	-5	-7	-9	-11	-12
P	-14	-11	-8	-5	-5	-4	-6	-8	-10
L	-16	-13	-10	-7	-7	-6	-3	-5	-7
E	-18	-15	-12	-9	-9	-8	-5	-5	-4

- The red arrows indicate the **cell from which we came**, that gave the best score

What is the optimal alignment?



		F	M	D	T	P	L	N	E
	0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2	+1	-1	-3	-5	-7	-9	-11	-13
K	-4	-1	-1	-3	-5	-7	-9	-11	-13
H	-6	-3	-3	-3	-5	-7	-9	-11	-13
M	-8	-5	-2	-4	-5	-7	-9	-11	-13
E	-10	-7	-4	-4	-6	-7	-9	-11	-10
D	-12	-9	-6	-3	-5	-7	-9	-11	-12
P	-14	-11	-8	-5	-5	-4	-6	-8	-10
L	-16	-13	-10	-7	-7	-6	-3	-5	-7
E	-18	-15	-12	-9	-9	-8	-5	-5	-4

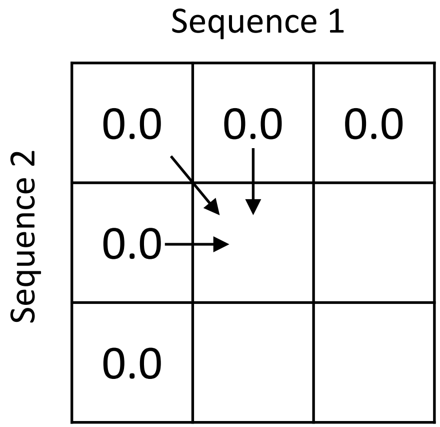
- Starting from the bottom right, find the continuous path tracing the red arrows
- This path is the optimal alignment for the two sequences

```
Sequence 1 FKHMED-PL-E
Sequence 2 F--M-DTPLNE
```

Needleman-Wunsch is **guaranteed** to find the optimal alignment
Example of **dynamic programming** – take a complex problem, break it down into smaller problems, and solve each only once, storing the result

Smith-Waterman Algorithm (Local Alignment)

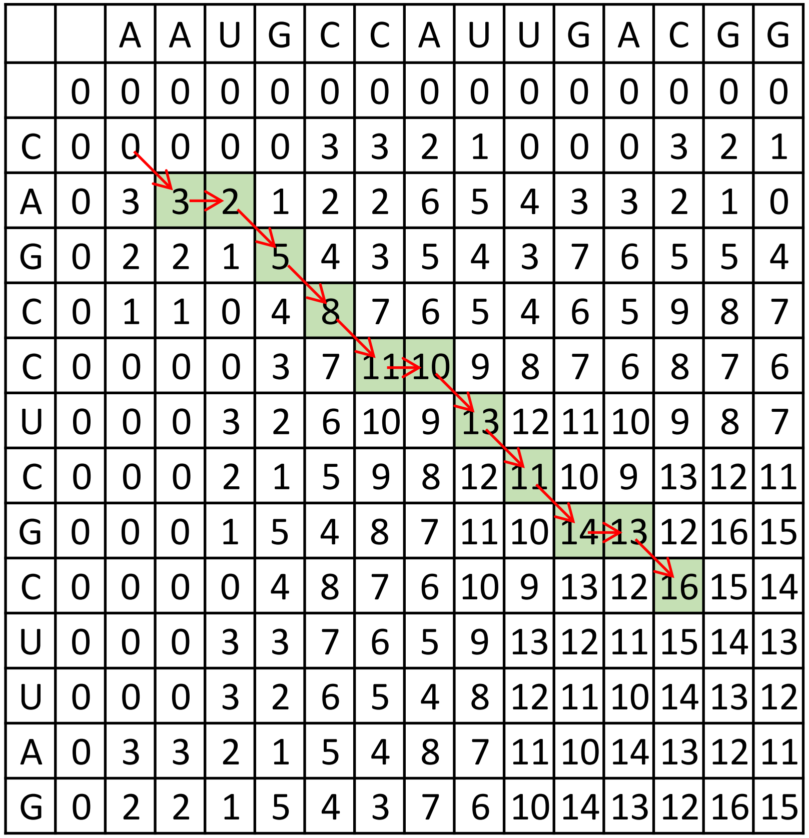
- Use RNA sequence as example here
- Construct matrix m+1,n+1 for m,n sequences
- Method is also with dynamic programming, like “needle”, but in “water” the **scoring scheme is different**:
- The maximum of:
  - Diagonal movement: score of (i-1,j-1) + value of match/mismatch
  - Horizontal movement: score of (i-1,j) + gap penalty
  - Vertical movement: score of (i,j-1) + gap penalty
  - If **all the above < 0**, then **insert the score 0**



0.0	0.0	0.0
0.0		
0.0		

Match = +3
Mismatch = -2
Gap = -1

Smith-Waterman Algorithm (Local Alignment)



		A	A	U	G	C	C	A	U	U	G	A	C	G	G
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	3	3	2	1	0	0	0	3	2	1
A	0	3	3	2	1	2	2	6	5	4	3	3	2	1	0
G	0	2	2	1	5	4	3	5	4	3	7	6	5	5	4
C	0	1	1	0	4	8	7	6	5	4	6	5	9	8	7
C	0	0	0	0	3	7	11	10	9	8	7	6	8	7	6
U	0	0	0	3	2	6	10	9	13	12	11	10	9	8	7
C	0	0	0	2	1	5	9	8	12	11	10	9	13	12	11
G	0	0	0	1	5	4	8	7	11	10	14	13	12	16	15
C	0	0	0	0	4	8	7	6	10	9	13	12	16	15	14
U	0	0	0	3	3	7	6	5	9	13	12	11	15	14	13
U	0	0	0	3	2	6	5	4	8	12	11	10	14	13	12
A	0	3	3	2	1	5	4	8	7	11	10	14	13	12	11
G	0	2	2	1	5	4	3	7	6	10	14	13	12	16	15

- Calculate the maximum score for each cell, keeping track of the path
- **Find the maximum score in the matrix**
- **Trace the path back** until you hit **0**

AAUGCCAUUGACGG
CA-GC-CUCG-CUUAG

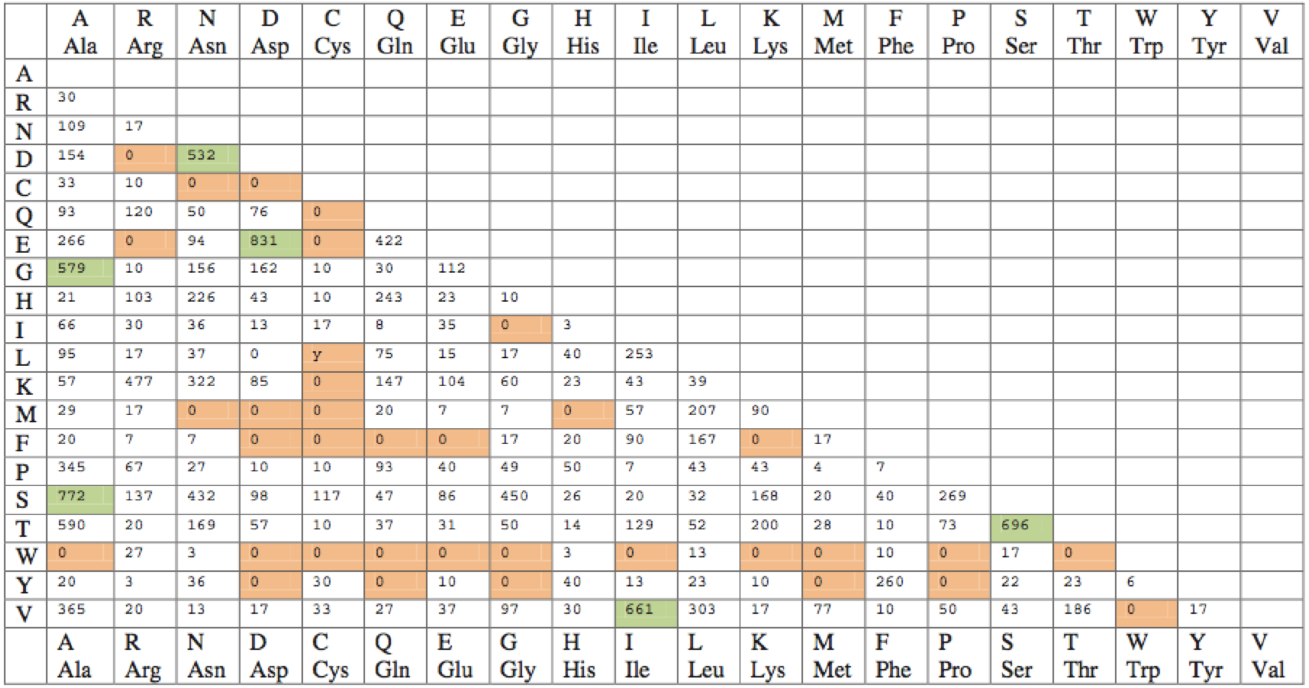
- Generate your own “water” matrices with your own scores: http://fridolin-linder.com/2016/03/30/local-alignment.html

Scoring matrices

- Once we have an alignment (global or local), **how do we calculate similarity**?
- Margaret Dayhoff developed a scheme to score alignments in proteins based on the **frequency of substitutions** observed in aligned, homologous proteins
- Mutations accepted by natural selection were referred to as point accepted mutations (**PAM**)
- Dayhoff looked at 1572 mutations in 71 groups of closely related proteins

Original amino acid

Substitutions



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	y	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

Mutability of amino acids

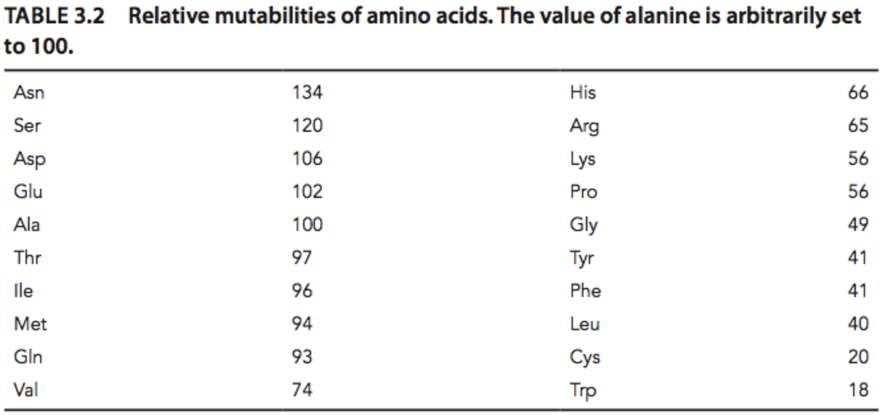
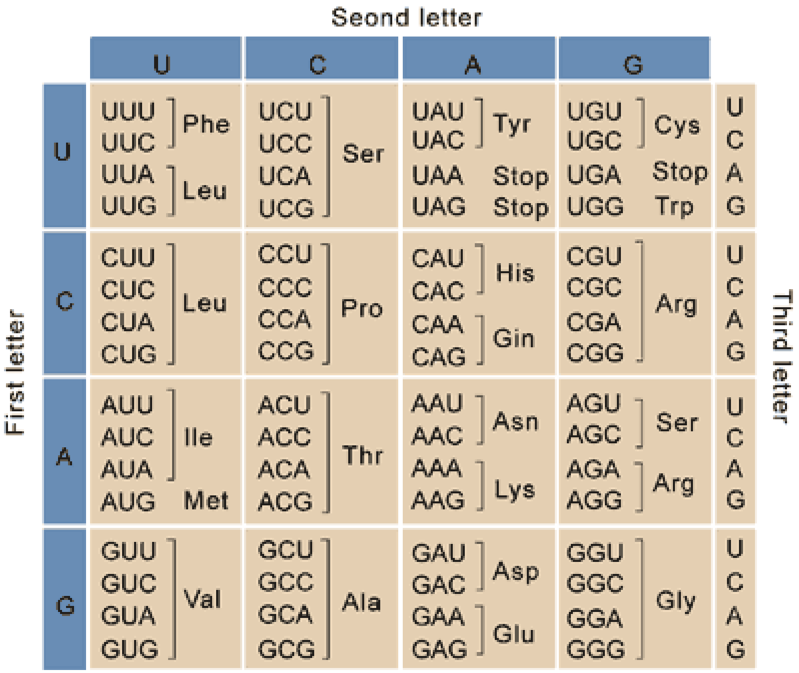


TABLE 3.2 Relative mutabilities of amino acids. The value of alanine is arbitrarily set to 100.

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18



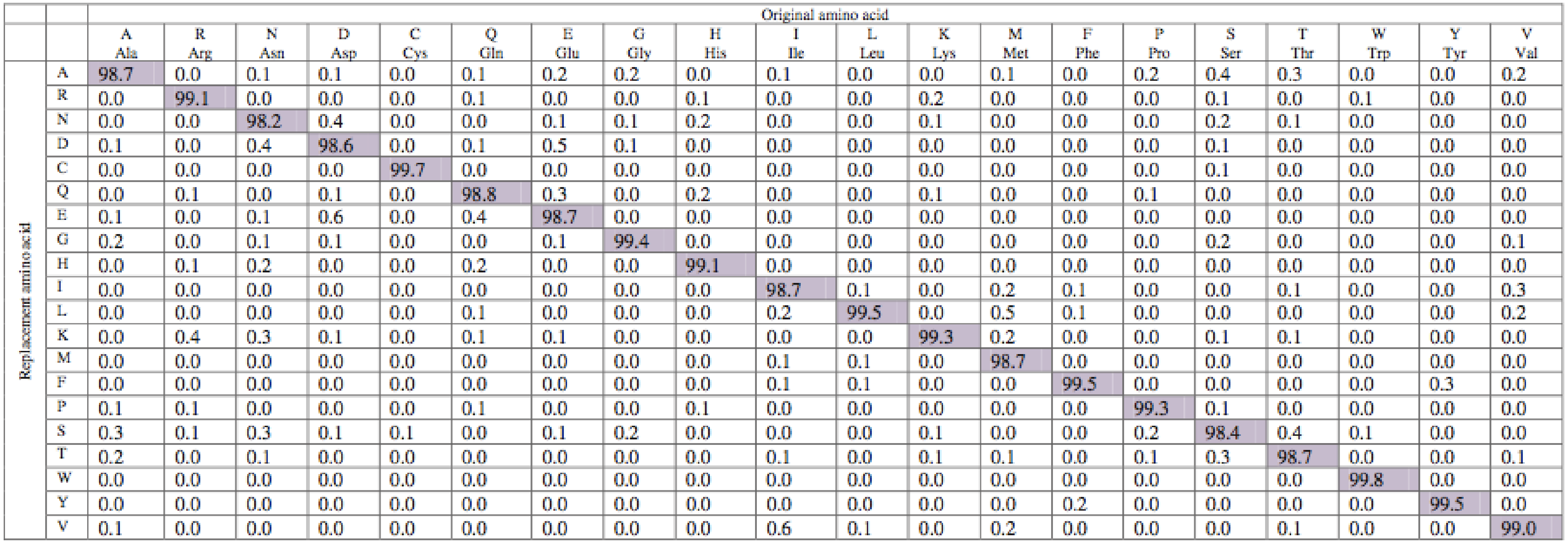
Seond letter

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

- Common amino acid substitutions require a single nucleotide change;
- Eg. GAC → GAA (D → E)
- The least mutable amino acids are often coded by only 1 or 2 codons
- (W, Y, C, F)
- A change of the last nucleotide of W codon changes the amino acid
- The low mutability of this amino acid means that mutations are not readily selected for

PAM1 matrix

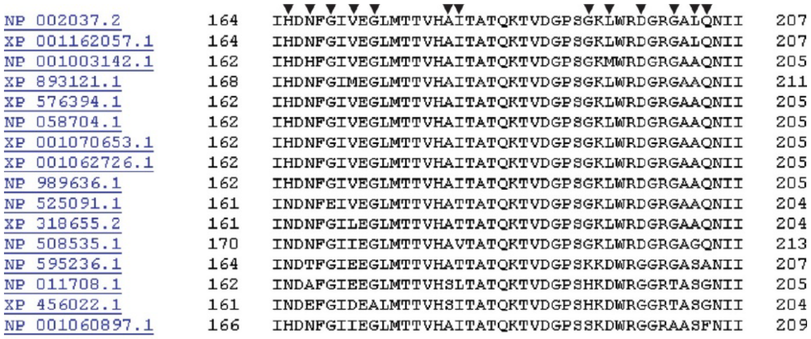
- Using data from accepted mutations and frequency of each amino acid in dataset, Dayhoff calculated substitution fraction (percentage change) when 1% (i.e., 1 in 100) amino acids are mutated
- **1%** is indication of **degree of change**, **not evolutionary distance** (proteins evolve at different rates)
- Each column adds to 100%



		Original amino acid																			
		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
Replacement amino acid	A	98.7	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.4	0.3	0.0	0.0	0.2
	R	0.0	99.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
	N	0.0	0.0	98.2	0.4	0.0	0.0	0.1	0.1	0.2	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0
	D	0.1	0.0	0.4	98.6	0.0	0.1	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	C	0.0	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	Q	0.0	0.1	0.0	0.1	0.0	98.8	0.3	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
	E	0.1	0.0	0.1	0.6	0.0	0.4	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	G	0.2	0.0	0.1	0.1	0.0	0.0	0.1	99.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.1
	H	0.0	0.1	0.2	0.0	0.0	0.2	0.0	0.0	99.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	98.7	0.1	0.0	0.2	0.1	0.0	0.0	0.1	0.0	0.0	0.3
	L	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.2	99.5	0.0	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.2
	K	0.0	0.4	0.3	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	99.3	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0
	M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	F	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	99.5	0.0	0.0	0.0	0.0	0.3	0.0
	P	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	99.3	0.1	0.0	0.0	0.0	0.0
	S	0.3	0.1	0.3	0.1	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.2	98.4	0.4	0.1	0.0	0.0
	T	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.3	98.7	0.0	0.0	0.1
	W	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.8	0.0	0.0
	Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	99.5	0.0
	V	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.1	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.0	99.0

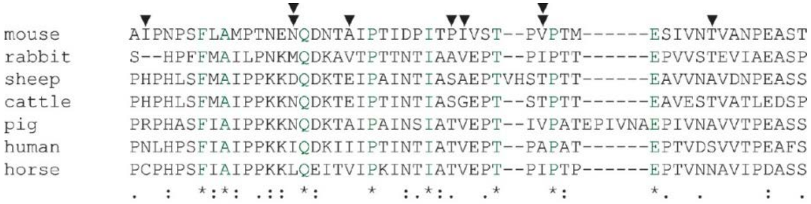
- Note that N changes more frequently than W

Different families of proteins evolve at different rates



```
                       ▼ ▼ ▼ ▼ ▼      ▼▼           ▼ ▼  ▼  ▼ ▼▼
NP_002037.2     164   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNII   207
XP_001162057.1  164   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNII   207
NP_001003142.1  162   IHDHFGIVEGLMTTVHAITATQKTVDGPSGKMWRDGRGAAQNII   205
XP_893121.1     168   IHDNFGIMEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   211
XP_576394.1     162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
NP_058704.1     162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
XP_001070653.1  162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
XP_001062726.1  162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
NP_989636.1     162   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII   205
NP_525091.1     161   INDNFEIVEGLMTTVHATTATQKTVDGPSGKLWRDGRGAAQNII   204
XP_318655.2     161   INDNFGILEGLMTTVHATTATQKTVDGPSGKLWRDGRGAAQNII   204
NP_508535.1     170   INDNFGIIEGLMTTVHAVTATQKTVDGPSGKLWRDGRGAGQNII   213
NP_595236.1     164   INDTFGIEEGLMTTVHATTATQKTVDGPSKKDWRGGRGASANII   207
NP_011708.1     162   INDAFGIEEGLMTTVHSLTATQKTVDGPSHKDWRGGRTASGNII   205
XP_456022.1     161   INDEFGIDEALMTTVHSITATQKTVDGPSHKDWRGGRTASGNII   204
NP_001060897.1  166   IHDNFGIIEGLMTTVHAITATQKTVDGPSSKDWRGGRAASFNII   209
```

FIGURE 3.10 Multiple sequence alignment of a portion of the glyceraldehyde 3-phosphate dehydrogenase (GAPDH) protein from 13 organisms: *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Canis lupus* (dog), *Mus musculus* (mouse), *Rattus norvegicus* (rat; three variants), *Gallus gallus* (chicken), *Drosophila melanogaster* (fruit fly), *Anopheles gambiae* (mosquito), *Caenorhabditis elegans* (worm), *Schizosaccharomyces pombe* (fission yeast), *Saccharomyces cerevisiae* (baker's yeast), *Kluyveromyces lactis* (a fungus), and *Oryza sativa* (rice). Columns in the alignment having even a single amino acid change are indicated with arrowheads. The accession numbers are given in the figure. The alignment was created by searching HomoloGene at NCBI with the term gapdh.

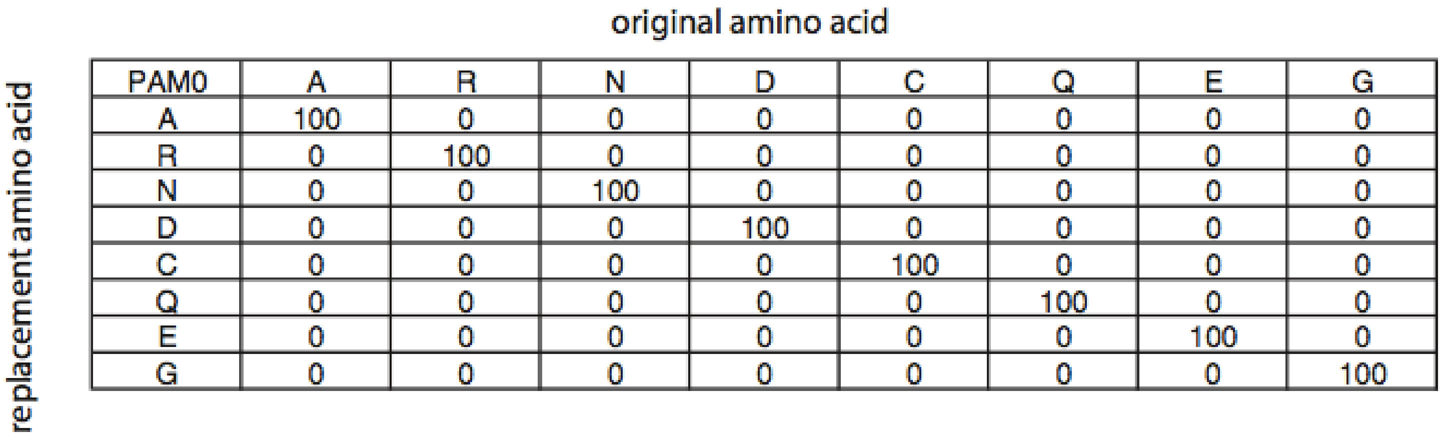


```
                        ▼                     ▼
           ▼            ▼    ▼        ▼▼      ▼              ▼
mouse     AIPNPSFLAMPTNENQDNTAIPTIDPITPIVST--PVPTM------ESIVNTVANPEAST
rabbit    S--HPFFMAILPNKMQDKAVTPTTNTIAAVEPT--PIPTT------EPVVSTEVIAEASP
sheep     PHPHLSFMAIPPKKDQDKTEIPAINTIASAEPTVHSTPTT------EAVVNAVDNPEASS
cattle    PHPHLSFMAIPPKKNQDKTEIPTINTIASGEPT--STPTT------EAVESTVATLEDSP
pig       PRPHASFIAIPPKKNQDKTAIPAINSIATVEPT--IVPATEPIVNAEPIVNAVVTPEASS
human     PNLHPSFIAIPPKKIQDKIIIPTINTIATVEPT--PAPAT------EPTVDSVVTPEAFS
horse     PCPHPSFIAIPPKKLQEITVIPKINTIATVEPT--PIPTP------EPTVNNAVIPDASS
          .  :  *:*: .:: *:    *  :.*:.  .*    *:       *.  .     :  .
```

FIGURE 3.11 Multiple sequence alignment of seven kappa caseins, representing a protein family that is relatively poorly conserved. Only a portion of the entire alignment is shown. Note that just eight columns of residues are perfectly conserved (indicated with asterisks), and gaps of varying length form part of the alignment. In several columns, there are four different aligned amino acids (arrowheads); in two instances there are five different residues (double arrowheads). The sequences were aligned with MUSCLE 3.6 (see Chapter 6) and were human (NP_005203), equine (*Equus caballus*; NP_001075353), pig (*Sus scrofa* NP_001004026), ovine (*Ovis aries* NP_001009378), rabbit (*Oryctolagus cuniculus* P33618), bovine (*Bos taurus* NP_776719) and mouse (*Mus musculus* NP_031812).

- Change in κ-caseins is more than 1 in every 100 amino acids
- Thus, using the PAM1 matrix will not give substitution scores that match the dataset, and we may miss some related proteins because the calculated similarity is incorrect
- The PAM250 matrix represents a dataset where **250 changes** have occurred over a **100 amino acid region**
- The **PAM250 matrix** is derived by successive **matrix multiplication** of the PAM1 matrix with itself, **250 times**

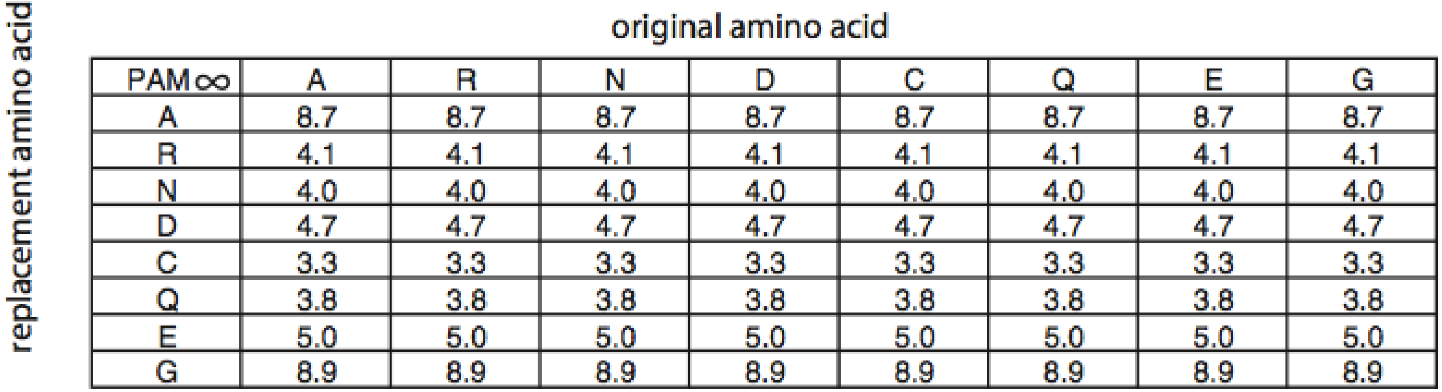
PAM matrices at the extremes



original amino acid

replacement amino acid

PAM0	A	R	N	D	C	Q	E	G
A	100	0	0	0	0	0	0	0
R	0	100	0	0	0	0	0	0
N	0	0	100	0	0	0	0	0
D	0	0	0	100	0	0	0	0
C	0	0	0	0	100	0	0	0
Q	0	0	0	0	0	100	0	0
E	0	0	0	0	0	0	100	0
G	0	0	0	0	0	0	0	100



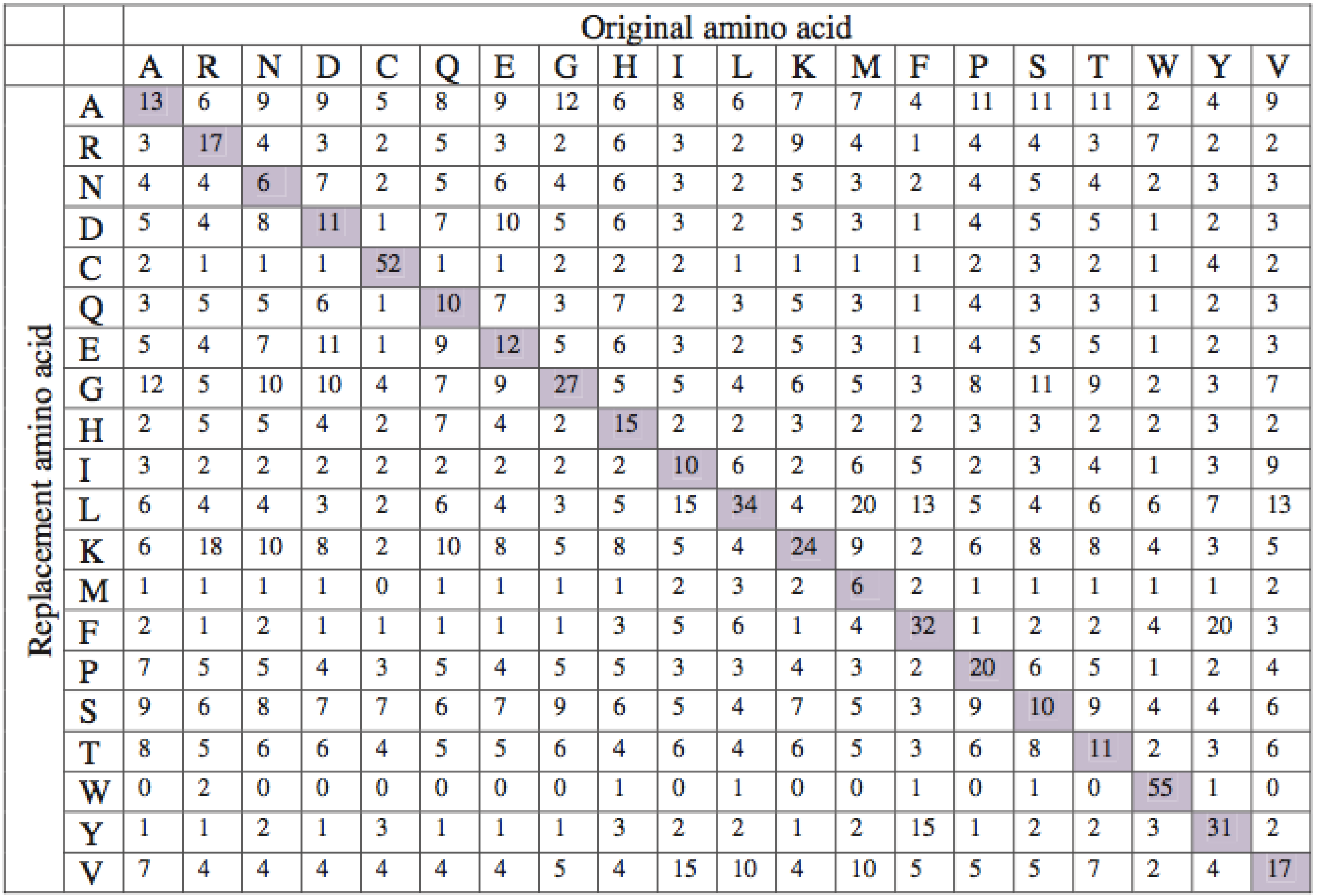
original amino acid

replacement amino acid

PAM∞	A	R	N	D	C	Q	E	G
A	8.7	8.7	8.7	8.7	8.7	8.7	8.7	8.7
R	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1
N	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
D	4.7	4.7	4.7	4.7	4.7	4.7	4.7	4.7
C	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
Q	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8
E	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
G	8.9	8.9	8.9	8.9	8.9	8.9	8.9	8.9

- When calculating a PAM0 matrix where there are 0 mutations per 100 amino acids, a diagonal of 100% is obtained
- The **PAM∞** matrix converges where the percentage change for every amino acids is its **relative abundance**

The PAM250 matrix



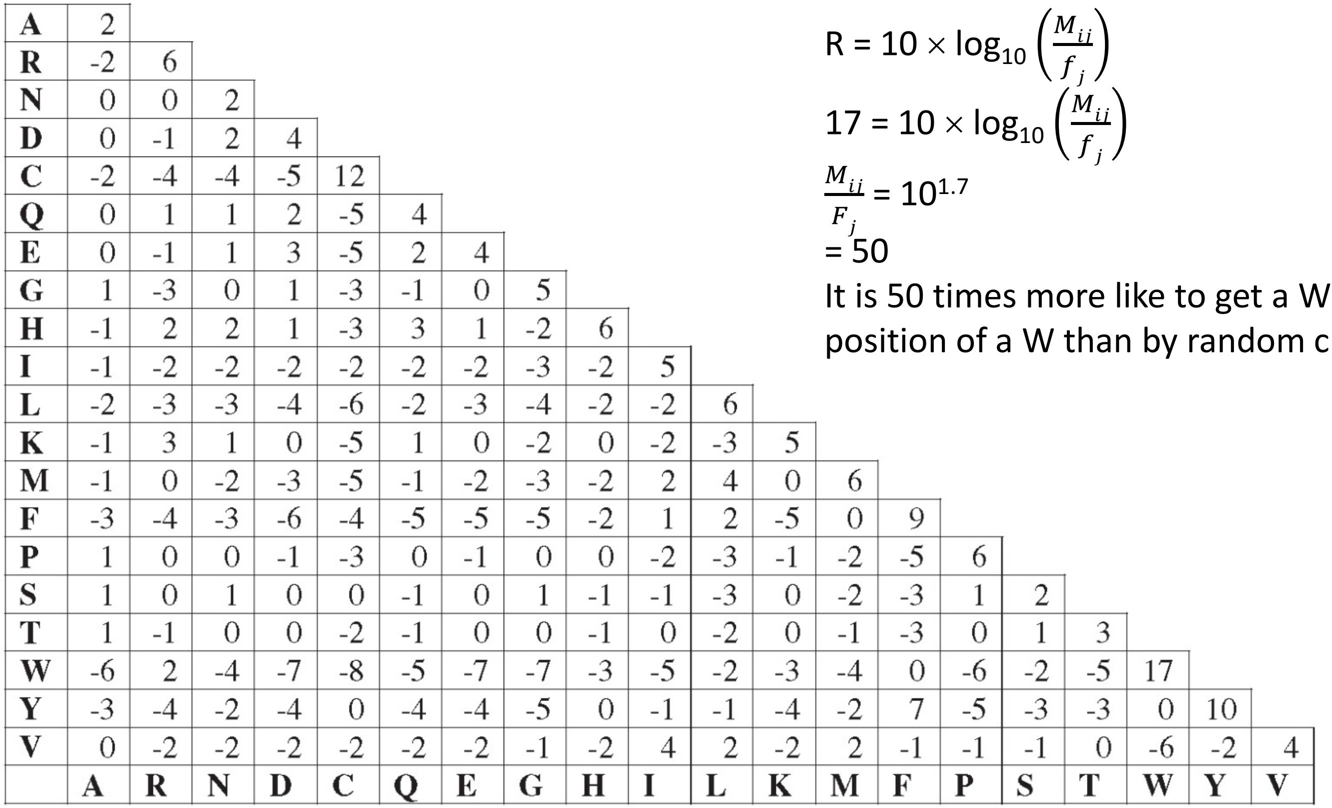
		Original amino acid																			
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Replacement amino acid	A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
	R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
	N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
	D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
	C	2	1	1	1	52	1	1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

- The **PAM250** matrix is used where proteins **share an identity of ~20%**
- Although one can get information on the chance of change for each different amino acid in the alignment, it is **difficult to interpret the score**
- Use a **relatedness odds ratio** to get a **more interpretable value**

The relatedness odds matrix

- The relatedness odds ratio is the ratio of the chance of having a mutation i→j at a position (M_{ij}), divided by the chance that residue j appears in the second sequence by chance (f_j)
- R = $\frac{M_{ij}}{f_j}$
- For instance, PAM250 shows that a C → L substitution has a probability of 0.02
- The frequency of occurrence of L is 0.085
- R = $\frac{0.02}{0.085}$ = 0.24, a chance **less than observed by random chance**
- One can also calculate the **log-odds ratio**:
- R =$10 \times \log_{10}\left(\frac{M_{ij}}{f_j}\right)$
- Thus, for the C → L substitution, the **log-odds ratio** is **-6.3**
- Where the log-odds ratio **> 0,** the **occurrence is more often** than by random chance
- Where the log-odds ratio **< 0,** the **occurrence is less** often than by random chance
- When calculating similarities between aligned sequences, the log-odds ratio of each position can be **added** (computationally less demanding)

The PAM250 log-odds ratio matrix



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

FIGURE 3.14 Log-odds matrix for PAM250. High PAM values (e.g., PAM250) are useful for aligning very divergent sequences. A variety of algorithms for pairwise alignment, multiple sequence alignment, and database searching (e.g., BLAST) allow you to select an assortment of PAM matrices such as PAM250, PAM70, and PAM30. Adapted from NCBI, ftp://ftp.ncbi.nlm.nih.gov/blast/matrices/.

W = +17

What does this mean?

$R = 10 \times \log_{10}\left(\frac{M_{ij}}{f_j}\right)$

$17 = 10 \times \log_{10}\left(\frac{M_{ij}}{f_j}\right)$

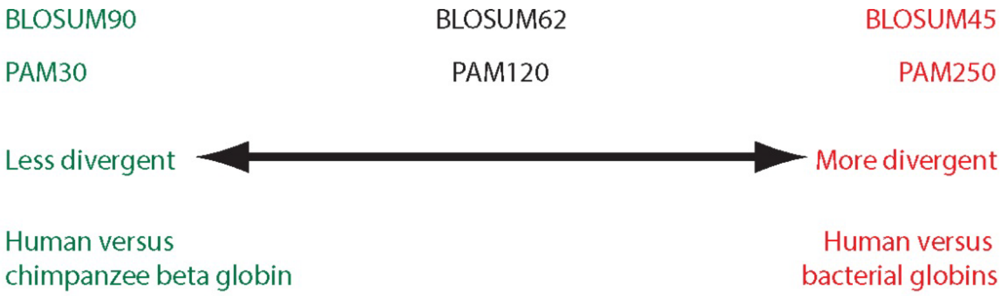
$\frac{M_{ij}}{F_j} = 10^{1.7}$

= 50

It is 50 times more like to get a W at the position of a W than by random chance

The BLOSUM matrices

- Henikoff and Henikoff used the BLOCK database of conserved regions of proteins that are distantly related
- The BLOSUM matrices use a **log$_2$ scoring scheme**
- BLOSUM62 used alignments of proteins that had at least 62% sequence identity
- There are also other BLOSUM matrices, eg. BLOSUM 50, BLOSUM70, BLOSUM90, based on 50%, 70% and 90% sequence identity
- The **BLOSUM** matrices are more successful at identifying more **distantly related proteins**
- The scores in the BLOSUM matrices are calculated from **empirical, aligned protein sequences**
- The scores in the PAM matrices are **derived** from the PAM1 matrix, with the assumption that substitution probabilities can be extrapolated



BLOSUM90 BLOSUM62 BLOSUM45

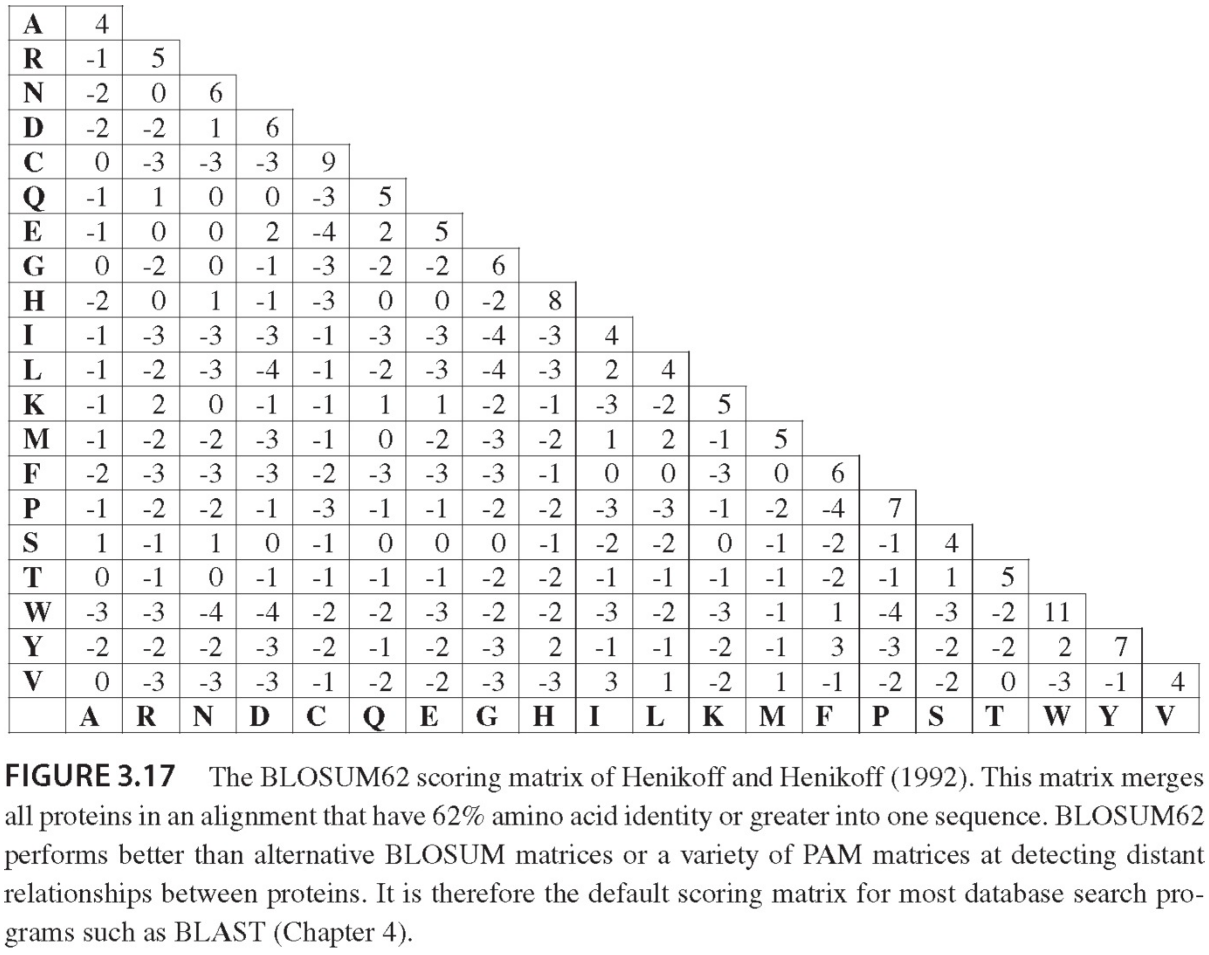
PAM30 PAM120 PAM250

Less divergent ⟷ More divergent

Human versus chimpanzee beta globin

Human versus bacterial globins

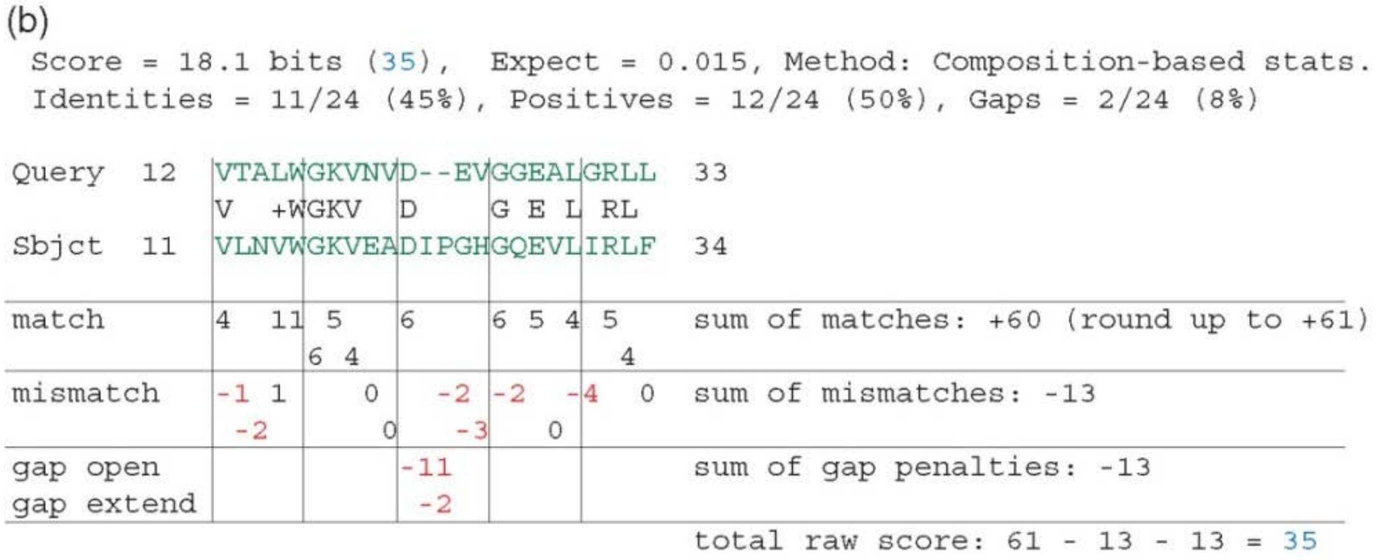
BLOSUM62 matrix



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

FIGURE 3.17 The BLOSUM62 scoring matrix of Henikoff and Henikoff (1992). This matrix merges all proteins in an alignment that have 62% amino acid identity or greater into one sequence. BLOSUM62 performs better than alternative BLOSUM matrices or a variety of PAM matrices at detecting distant relationships between proteins. It is therefore the default scoring matrix for most database search programs such as BLAST (Chapter 4).

Application of the BOSUM62 matrix



(b)

```
 Score = 18.1 bits (35),  Expect = 0.015, Method: Composition-based stats.
 Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query  12  VTALWGKVNVD--EVGGEALGRLL  33
           V  +WGKV  D    G E L RL
Sbjct  11  VLNVWGKVEADIPGHGQEVLIRLF  34

match      4  11 5   6    6 5 4 5    sum of matches: +60 (round up to +61)
                6 4              4
mismatch   -1 1     0  -2 -2  -4  0  sum of mismatches: -13
            -2       0  -3   0
gap open             -11          sum of gap penalties: -13
gap extend            -2
                                     total raw score: 61 - 13 - 13 = 35
```