PROTEIN SUBSTITUTION MODELS

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@uniroma1.it

CB_23_24_L9, Rome 12th oct 2023

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

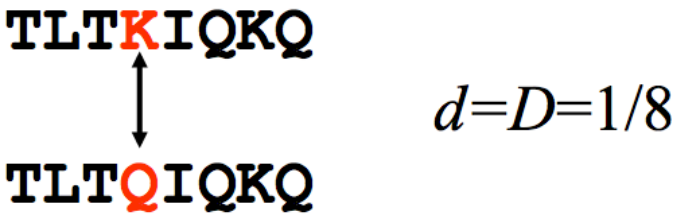
Protein substitution models

Ideally, we should have developed substitution models for protein sequence evolution as we did for DNA sequences. However, too many parameters in such models make them difficult to use and for theoretic analysis.

Therefore, traditionally, protein substitution models have been developed empirically based on statistical data from groups of closely related proteins (they are called **protein families**).

The first such empirical protein substitution model is the PAM (point accepted mutation) model, developed by Dayhoff, Schwarts and Orcu in 1978.

The basic assumption of the models is that if we only compare closely related sequences, then the observed different amino acids at a site car be thought being caused by a single substitution, i.e., $d=D$.



$d=D=1/8$

The PAM protein substitution models

To derive the statistics about amino acid substitutions, we need to use multiple closely related sequences, and make multiple alignments of them.

Because the sequences are closely related to one another, the multiple alignments can be reliably generated.

```
A:  T L K K V Q K T
B:  T L K K V Q K T
C:  T L K K I Q K Q
D:  I I T K L Q K Q
E:  T I T K L Q K Q
F:  T L T K I Q K Q
G:  T L T Q I Q K Q
```

However, to avoid over-counting of substitutions when comparing these multiple sequences, we have to assume that we know the phylogenetic relationships of the sequences, so that the minimal number of substitutions can be counted.

The method that creates a phylogenetic tree by minimizing the number of substitutions needed is called a **parsimonious method**, and the resulting tree is called a **parsimonious tree**.

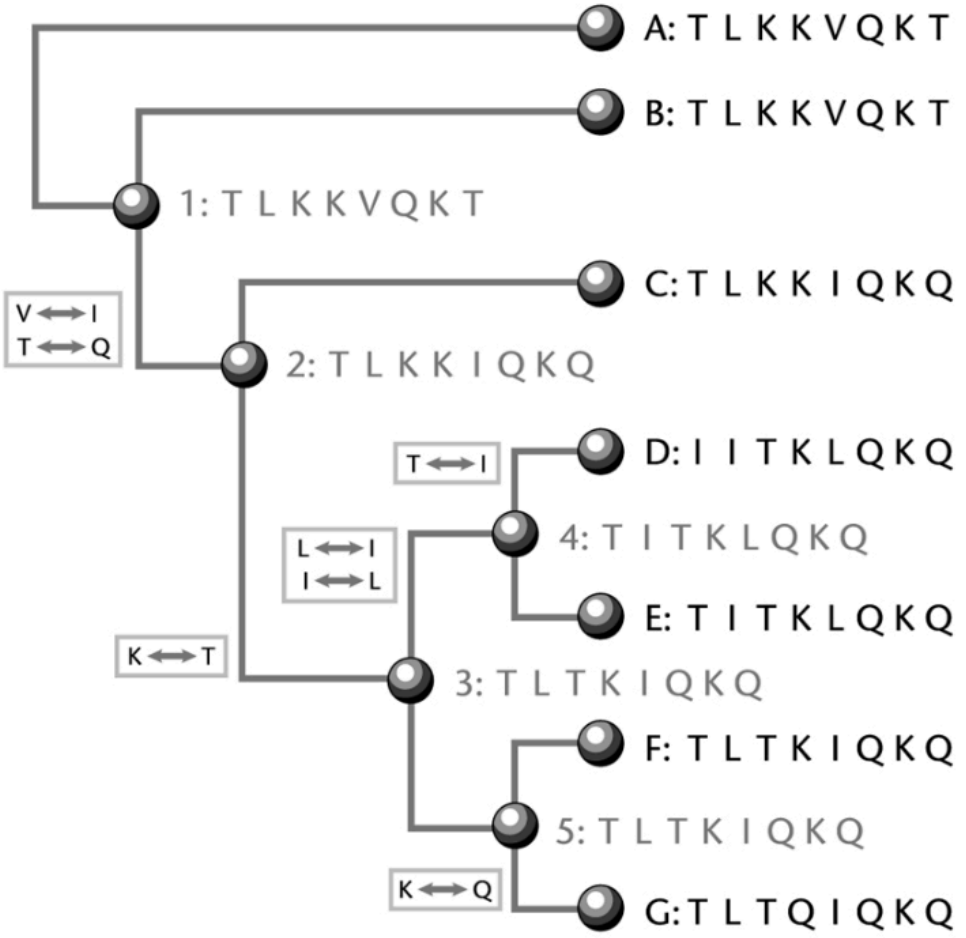
The PAM protein substitution models

On the other hand, a parsimonious tree can guarantee the minimal counts of substitutions among multiple sequences.

A:	T L K K V Q K T
B:	T L K K V Q K T
C:	T L K K I Q K Q
D:	I I T K L Q K Q
E:	T I T K L Q K Q
F:	T L T K I Q K Q
G:	T L T Q I Q K Q

Based on a parsimonious tree, we can generate an **observed substitution matrix** A, whose item A_{ij} is the number of times amino acid i has been substituted by j. The matrix is symmetric, i.e., $A_{ij} = A_{ji}$.

$$A = \begin{pmatrix} & \mathbf{I} & \mathbf{K} & \mathbf{L} & \mathbf{Q} & \mathbf{T} & \mathbf{V} \\ \mathbf{I} & - & - & 2 & - & 1 & 1 \\ \mathbf{K} & - & - & - & 1 & 1 & - \\ \mathbf{L} & 2 & - & - & - & - & - \\ \mathbf{Q} & - & 1 & - & - & 1 & - \\ \mathbf{T} & 1 & 1 & - & 1 & - & - \\ \mathbf{V} & 1 & - & - & - & - & - \end{pmatrix}$$



The PAM protein substitution models

The upper diagonal part of the following matrix is the original observed amino acid substitution matrix developed by Dayhoff and colleagues. Jones,Taylor; Thornton (1992)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	247	216	386	106	208	600	1183	46	173	257	200	100	51	901	2413	2440	11	41	1766
R	–1	5	116	48	125	750	119	614	446	76	205	2348	61	16	217	413	230	109	46	69
N	0	0	3	1433	32	159	180	291	466	130	63	758	39	15	31	1738	693	2	114	55
D	0	–1	2	5	13	130	2914	577	144	37	34	102	27	8	39	244	151	5	89	127
C	–1	–1	–1	–3	11	9	8	98	40	19	36	7	23	66	15	353	66	38	164	99
Q	–1	2	0	1	–3	5	1027	84	635	20	314	858	52	9	395	182	149	12	40	58
E	–1	0	1	4	–4	2	5	610	41	43	65	754	30	13	71	156	142	12	15	226
G	1	0	0	1	–1	–1	0	5	41	25	56	142	27	18	93	1131	164	69	15	276
H	–2	2	1	0	0	2	0	–2	6	26	134	85	21	50	157	138	76	5	514	22
I	0	–3	–2	–3	–2	–3	–3	–3	–3	4	1324	75	704	196	31	172	930	12	61	3938
L	–1	–3	–3	–4	–3	–2	–4	–4	–2	2	5	94	974	1093	578	436	172	82	84	1261
K	–1	4	1	0	–3	2	1	–1	1	–3	–3	5	103	7	77	228	398	9	20	58
M	–1	–2	–2	–3	–2	–2	–3	–3	–2	3	3	–2	6	49	23	54	343	8	17	559
F	–3	–4	–3	–5	0	–4	–5	–5	0	0	2	–5	0	8	36	309	39	37	850	189
P	1	–1	–1	–2	–2	0	–2	–1	0	–2	0	–2	–2	–3	6	1138	412	6	22	84
S	1	–1	1	0	1	–1	–1	1	–1	–1	–2	–1	–1	–2	1	2	2258	36	164	219
T	2	–1	1	–1	–1	–1	–1	–1	–1	1	–1	–1	0	–2	1	1	2	8	45	526
W	–4	0	–5	–5	1	–3	–5	–2	–3	–4	–2	–3	–3	–1	–4	–3	–4	15	41	27
Y	–3	–2	–1	–2	2	–2	–4	–4	4	–2	–1	–3	–2	5	–3	–1	–3	0	9	42
V	1	–3	–2	–2	–2	–3	–2	–2	–3	4	2	–3	2	0	–1	–1	0	–3	–3	4

The PAM protein substitution models

Based on this observed substitution matrix A_{ij}, we want to develop a protein substitution model M, whose element M_{ij} is the probability that amino acid i will be substituted by j in a small unit of time λ.

For a small period of time $t=\lambda n$, we can assume that the probability of substitution changes linearly with time, i.e., $p_{ij}(t)=\alpha t=\alpha\lambda n$.

Therefore, M_{ij} should be proportional to λ and A_{ii}, but inversely proportional to the total number of amino acid i in the sequences, N_i, i.e.,

$$M_{ij} = p_{ij}(\lambda) = \frac{\lambda A_{ij}}{N_i}, \quad (i \neq j).$$

Let π_i be the frequency of amino acid i in the sequences, then,

$$\pi_i = \frac{N_i}{N_{tot}}.$$

where N_{tot} is the total number of amino acids in the sequences.

The PAM1 protein substitution model

- If we define the time λ to be a PAM unit, which is the time required for an average of 1% of amino acids being substituted in the sequences, then we have,

$$\sum_i \pi_i \sum_j M_{ij} = \sum_i \sum_j \pi_i \frac{\lambda A_{ij}}{N_i} = \sum_i \sum_j \pi_i \frac{\lambda A_{ij}}{\pi_i N_{tot}} = \sum_i \sum_j \frac{\lambda A_{ij}}{N_{tot}}$$

$$= \frac{\lambda}{N_{tot}} \sum_i \sum_j A_{ij} = \frac{\lambda A_{tot}}{N_{tot}} = 0.01. \quad (i \neq j)$$

where A_{tot} is the total number of substitutions, *i.e.,* the sum of all elements in A. Therefore,

$$\lambda = \frac{0.01 N_{tot}}{A_{tot}}$$

- Using this one PAM unit value, we can compute all M_{ij} values using the formula,

$$M_{ij} = \frac{\lambda A_{ij}}{N_i} = \frac{0.01 N_{tot} A_{ij}}{A_{tot} N_i}, \ (i \neq j).$$

- For M_{ii}, the probability that amino acid i remains unchanged in the time of a PAM unit, we define,

$$M_{ii} = 1 - \sum M_{ij}.$$

The PAM1 protein substitution model

- This substitution matrix obtained at one PAM unit is called a **PAM1 matrix.** The PAM1 matrix obtained by Jones et al (1992) are shown below, where values have been multiplied by 100,000 for convenience.

$$M_{ij} = \frac{0.01 A_{ij} N_{tot}}{N_i A_{tot}}, \quad (i \neq j)$$

$$M_{ii} = 1 - \sum_{j \neq i} M_{ij}$$

$$\pi_i = \frac{N_i}{N_{tot}}$$

$$m_i = \frac{1 - M_{ii}}{0.01}$$

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	98759	27	24	42	12	23	66	129	5	19	28	22	11	6	99	**264**	**267**	1	4	193
R	41	98962	19	8	21	**125**	20	102	74	13	34	**390**	10	3	36	69	38	18	8	11
N	43	23	98707	**284**	6	31	36	58	92	26	12	150	8	3	6	**344**	137	0	23	11
D	63	8	**235**	98932	2	21	**478**	95	24	6	6	17	4	1	6	40	25	1	15	21
C	44	52	13	5	99450	4	3	41	17	8	15	3	10	28	6	**147**	28	16	**68**	41
Q	43	154	33	27	2	98955	**211**	17	130	4	64	**176**	11	2	81	37	31	2	8	12
E	82	16	25	**398**	1	**140**	99042	83	6	6	9	102	4	2	10	21	19	2	2	31
G	**135**	70	33	66	11	10	70	99369	5	3	6	16	3	2	11	**129**	19	8	2	32
H	17	164	171	53	15	**223**	15	15	98867	10	49	31	8	18	58	51	28	2	**189**	8
I	28	12	21	6	3	3	7	4	4	98722	**212**	12	113	31	5	28	149	2	10	**630**
L	24	19	6	3	3	29	6	5	12	**122**	99328	9	90	101	53	40	16	8	8	**117**
K	28	**334**	108	14	1	**122**	107	20	12	11	13	99101	15	1	11	32	57	1	3	8
M	36	22	14	10	8	19	11	10	8	**253**	**350**	37	98845	18	8	19	123	3	6	201
F	11	3	3	2	14	2	3	4	11	41	**230**	1	10	99357	8	65	8	8	**179**	40
P	**150**	36	5	7	3	66	12	16	26	5	97	13	4	6	99278	**190**	69	1	4	14
S	**297**	51	214	30	44	22	19	139	17	21	54	28	7	38	140	98548	**278**	4	20	27
T	**351**	33	100	22	9	21	20	24	11	134	25	57	49	6	59	**325**	98670	1	6	76
W	7	**65**	1	3	23	7	7	41	3	7	**49**	5	5	22	4	21	5	99684	24	16
Y	11	12	30	23	43	10	4	4	**134**	16	22	5	4	**222**	6	43	12	11	99377	11
V	**226**	9	7	16	13	7	29	35	3	**504**	161	7	71	24	11	28	67	3	5	98772

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
π_i	0.077	0.051	0.043	0.052	0.020	0.041	0.062	0.074	0.023	0.053	0.091	0.059	0.024	0.040	0.051	0.069	0.059	0.014	0.032	0.066
m_i	1.241	1.038	1.293	1.068	0.550	1.045	0.958	0.631	1.133	1.273	0.672	0.899	1.155	0.643	0.722	1.452	1.330	0.316	0.623	1.228

The PAM1 protein substitution model

- Note that the mean substitution rate per λ time unit is 0.01.
- The sum of the each row except for the element in the diagonal is the total probability for that amino acid to be substituted by another one, which is equal to 1-M_{ii},
- Dividing this number by the mean substitution rate 0.01, gives the **relative mutability** of the amino acid i, denoted by m_i, i.e.,

$$m_i = \frac{1-M_{ii}}{0.01}$$

- A value of m_i grater than 1, means the amino acid i is more likely to be substituted by another than an average one, and vice versa.
- The time reversibility also holds for the PAM1 substitution matrix, as for any amino acids i, and j, we have,

$$\pi_i M_{ij} = \pi_i \frac{\lambda A_{ij}}{N_i} = \frac{N_i}{N_{tot}} \frac{\lambda A_{ij}}{N_i} = \frac{\lambda A_{ij}}{N_{tot}} = \frac{\lambda A_{ji}}{N_{tot}} = \frac{N_j}{N_{tot}} \frac{\lambda A_{ji}}{N_j} = \pi_j M_{ji}.$$

The PAM*n* protein substitution models

- Using the PAM1 substitution matrix, we can compute substitution probability that amino acid i will be substituted by j at a time interval of 2, 3, 4, …, and n PAM units. Let's denote this probability as $p(n\lambda)$.
- Let's first consider $p_{AV}(2\lambda)$, the probability that alanine (A) will be substituted by valine (V) after 2λ time.
- After one PAM unit time, A can remain unchanged, or be substituted by any of the other 19 amino acids, so we have to consider the following 20 possible scenarios,

$t=0$	A		A		A	…..		A
	↓		↓		↓			↓
$t=1$	A	M_{AA}	R	M_{AR}	N	M_{AN}	…..	V M_{AV}
	↓		↓		↓			↓
$t=2$	V	M_{AV}	V	M_{RV}	V	M_{NV}	…..	V M_{VV}

Therefore, $p_{AV}(2\lambda) = \sum_k M_{Ak} M_{kV} = M_{A\bullet} M_{\bullet V} = M^{(2)}_{AV}$.

In general, we have, $p_{ij}(2\lambda) = \sum_k M_{ik} M_{kj} = M^{(2)}_{ij}$.

The PAM*n* protein substitution models

- For $p_{AV}(n\lambda)$, we have to consider the following 20 scenarios,

$t = 0$	A		A		A		.....	A	
	↓		↓		↓			↓	
$t = n-1$	A	$M^{(n-1)}_{AA}$	R	$M^{(n-1)}_{AR}$	N	$M^{(n-1)}_{AN}$	.....	V	$M^{(n-1)}_{AV}$
	↓		↓		↓			↓	
$t = n$	V	M_{AV}	V	M_{RV}	V	M_{NV}	.....	V	M_{VV}

Therefore, $p_{AV}(n\lambda) = \sum_k M^{(n-1)}_{Ak} M_{kV} = M^{(n)}_{AV}$.

In general, we have, $p_{ij}(n\lambda) = \sum_k M^{(n-1)}_{ik} M_{kj} = M^{(n)}_{ij}$.

- Using this formula, we can compute any n PAM units time substitution matrix, and each is called a PAMn matrix, e.g, PAM250 matrix; and n is called the **PAM distance**.

PAM distances

The evolutionary distance between two sequences that have ***n* PAM unit distance** is: d=0.01n substitutions / site

For two very long protein sequences, we assume that their amino acid frequency π_i is the same as the sequence set used to generate PAM1 substitution matrix. Therefore the expected difference between the two sequences at PAM distance n is,

$$D = \sum_i \pi_i (1 - M_{ii}^{(n)}).$$

This is simply the probability that a site in one sequence is amino acid i, and same site in the other sequence is not i.

Theoretically, we can use this relationship to calculate the PAM distance d=0.01n, by solving for n, but the calculation is not straight forward, because we cannot easily solve for n in this equation.

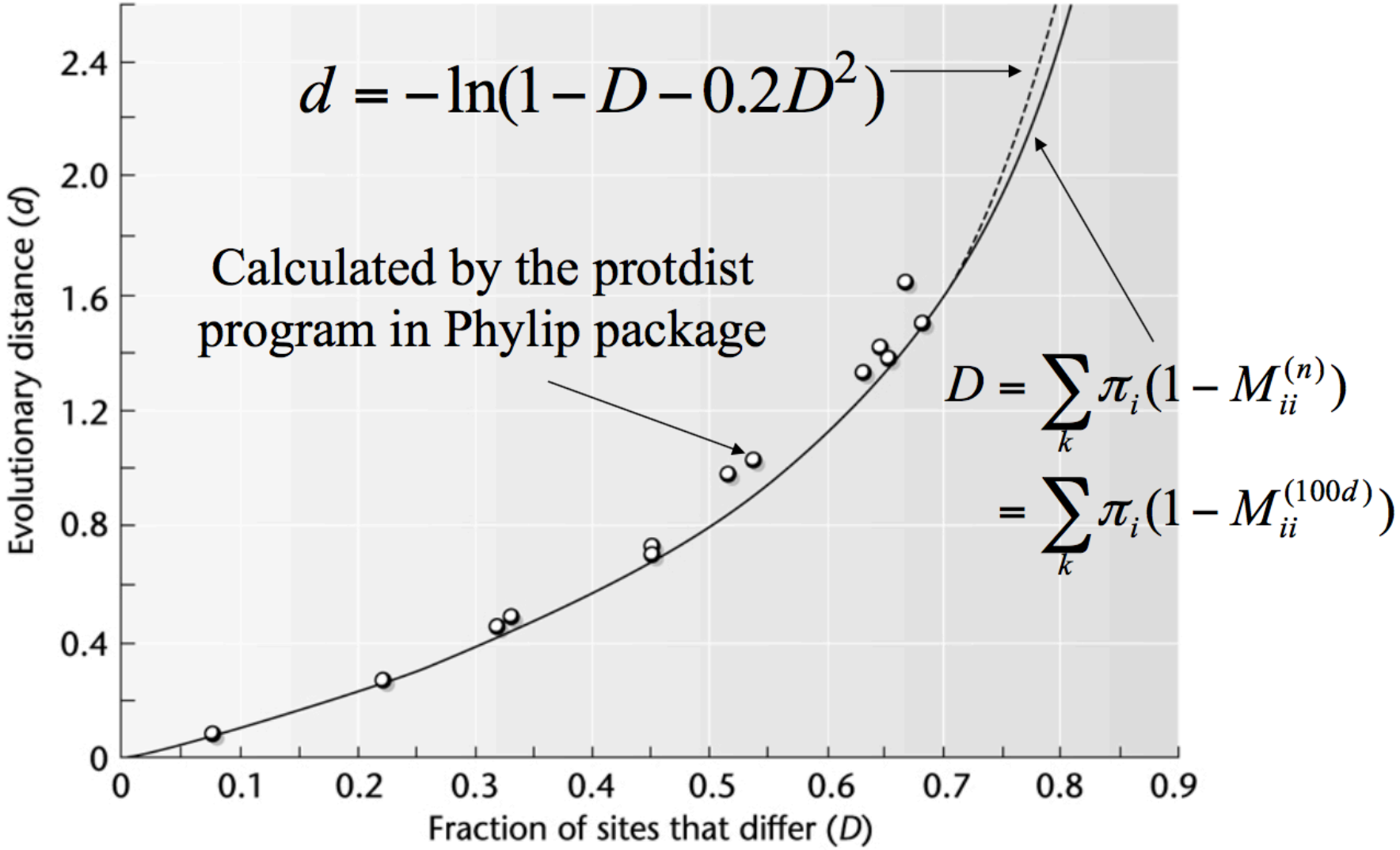
A pre-computed table or graph can be used for calibration purpose.

Kimura gave an empirical approximation of d using D,

$$d = 0.01n = -\ln(1 - D - 0.2D^2).$$

PAM distances

The relationship between the evolutionary distance *d* and the difference D between two amino acid sequences.



Log-odds scoring matrices for amino acids

We now develop the PAM scoring matrix for protein sequence alignment.

Let *a* and *b* be two very long sequences with *n* PAM distance, then the probability that a site in *a* has an amino acid *i* and the same site in sequence *b* has an amino acid *j* is, $\pi_i M_{ij}^{(n)}$.

The probability for this alignment to happen in two random sequences is $\pi_i \pi_j$.

We define the **odds ratio** of these two events as, $R_{ij} = \frac{\pi_i M_{ij}^{(n)}}{\pi_i \pi_j} = \frac{M_{ij}^{(n)}}{\pi_j}$

If $R_{ij}>1$, then amino acid *i* and *j* are more likely to be aligned with each other according to the PAM model than they would be by chance. The opposite conclusion holds if $R_{ij}<1$.

Because $R_{ji} = \frac{\pi_j M_{ji}^{(n)}}{\pi_i \pi_j} = \frac{\pi_i M_{ij}^{(n)}}{\pi_i \pi_j} = R_{ij}$,

```
a: TLTKIQKQ ...
b: TLTQIQKQ ...
```

the **odds ratio matrix R** is symmetric.

Log-odds scoring matrices for amino acids

- Let a_k and b_k be amino acids at the k-th site of the alignment between sequences a and b, then the relative likelihood that this alignment can be made according to the PAM model relative to the likelihood that it can be made by chance is,

$$L(a,b) = \prod_{k=1}^{l} R_{a_k b_k}.$$

```
a: TLTKIQKQ
b: TLTQIQKQ
```

- For the convenience of calculation, we take logarithm on $L(a,b)$,

$$\log L(a,b) = \log \prod_{k=1}^{l} R_{a_k b_k} = \sum_{k=1}^{l} \log R_{a_k b_k}.$$

- We define the score to align amino acid i and j as $S(a,b) = c \log R_{ij}$, where c is a scaling factor.
- S is called the **scoring matrix**, clearly, it is symmetric, ie, $S(i,j)=S(j,i)$.
- Then we define the score of an alignment between sequences a and b as,

$$S_{alignemnt}(a,b) = c \log L(a,b) = \sum_{k=1}^{l} S(a_k, b_k).$$

- The goal of a pair-wise alignment method is to find the alignment that maximizes this scoring function.

The relationship between alignment score and the physico-chemical prosperities of amino acids

The lower diagonal part of the following matrix is the log-odds scoring matrix corresponding to the PAM250 matrix, and $S(i,j)=10\log_{10}R_{ij}$

Two amino acids that have similar physico-chemical property tend to have a large positive score, and vice versa.

Shaded gray: positive score

Shaded black: switched by single mutations:

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	247	216	386	106	208	600	1183	46	173	257	200	100	51	901	2413	2440	11	41	1766
R	–1	5	116	48	125	750	119	614	446	76	205	2348	61	16	217	413	230	109	46	69
N	0	0	3	1433	32	159	180	291	466	130	63	758	39	15	31	1738	693	2	114	55
D	0	–1	2	5	13	130	2914	577	144	37	34	102	27	8	39	244	151	5	89	127
C	–1	–1	–1	–3	11	9	8	98	40	19	36	7	23	66	15	353	66	38	164	99
Q	–1	2	0	1	–3	5	1027	84	635	20	314	858	52	9	395	182	149	12	40	58
E	–1	0	1	4	–4	2	5	610	41	43	65	754	30	13	71	156	142	12	15	226
G	1	0	0	1	–1	–1	0	5	41	25	56	142	27	18	93	1131	164	69	15	276
H	–2	2	1	0	0	2	0	–2	6	26	134	85	21	50	157	138	76	5	514	22
I	0	–3	–2	–3	–2	–3	–3	–3	–3	4	1324	75	704	196	31	172	930	12	61	3938
L	–1	–3	–3	–4	–3	–2	–4	–4	–2	2	5	94	974	1093	578	436	172	82	84	1261
K	–1	4	1	0	–3	2	1	–1	1	–3	–3	5	103	7	77	228	398	9	20	58
M	–1	–2	–2	–3	–2	–2	–3	–3	–2	3	3	–2	6	49	23	54	343	8	17	559
F	–3	–4	–3	–5	0	–4	–5	–5	0	0	2	–5	0	8	36	309	39	37	850	189
P	1	–1	–1	–2	–2	0	–2	–1	0	–2	0	–2	–2	–3	6	1138	412	6	22	84
S	1	–1	1	0	1	–1	–1	1	–1	–1	–2	–1	–1	–2	1	2	2258	36	164	219
T	2	–1	1	–1	–1	–1	–1	–1	–1	1	–1	–1	0	–2	1	1	2	8	45	526
W	–4	0	–5	–5	1	–3	–5	–2	–3	–4	–2	–3	–3	–1	–4	–3	–4	15	41	27
Y	–3	–2	–1	–2	2	–2	–4	–4	4	–2	–1	–3	–2	5	–3	–1	–3	0	9	42
V	1	–3	–2	–2	–2	–3	–2	–2	–3	4	2	–3	2	0	–1	–1	0	–3	–3	4

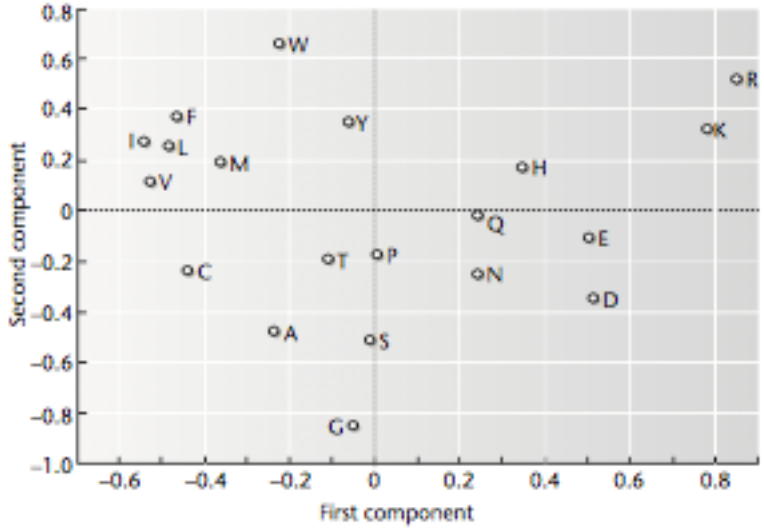
Table 2.2 Physico-chemical properties of the amino acids.

			Vol.	Bulk.	Polarity	pI	Hyd.1	Hyd.2	Surface area	Fract. area
Alanine	Ala	A	67	11.50	0.00	6.00	1.8	1.6	113	0.74
Arginine	Arg	R	148	14.28	52.00	10.76	−4.5	−12.3	241	0.64
Asparagine	Asn	N	96	12.28	3.38	5.41	−3.5	−4.8	158	0.63
Aspartic acid	Asp	D	91	11.68	49.70	2.77	−3.5	−9.2	151	0.62
Cysteine	Cys	C	86	13.46	1.48	5.05	2.5	2.0	140	0.91
Glutamine	Gln	Q	114	14.45	3.53	5.65	−3.5	−4.1	189	0.62
Glutamic acid	Glu	E	109	13.57	49.90	3.22	−3.5	−8.2	183	0.62
Glycine	Gly	G	48	3.40	0.00	5.97	−0.4	1.0	85	0.72
Histidine	His	H	118	13.69	51.60	7.59	−3.2	−3.0	194	0.78
Isoleucine	Ile	I	124	21.40	0.13	6.02	4.5	3.1	182	0.88
Leucine	Leu	L	124	21.40	0.13	5.98	3.8	2.8	180	0.85
Lysine	Lys	K	135	15.71	49.50	9.74	−3.9	−8.8	211	0.52
Methionine	Met	M	124	16.25	1.43	5.74	1.9	3.4	204	0.85
Phenylalanine	Phe	F	135	19.80	0.35	5.48	2.8	3.7	218	0.88
Proline	Pro	P	90	17.43	1.58	6.30	−1.6	−0.2	143	0.64
Serine	Ser	S	73	9.47	1.67	5.68	−0.8	0.6	122	0.66
Threonine	Thr	T	93	15.77	1.66	5.66	−0.7	1.2	146	0.70
Tryptophan	Trp	W	163	21.67	2.10	5.89	−0.9	1.9	259	0.85
Tyrosine	Tyr	Y	141	18.03	1.61	5.66	−1.3	−0.7	229	0.76
Valine	Val	V	105	21.57	0.13	5.96	4.2	2.6	160	0.86
Mean			109	15.35	13.59	6.03	−0.5	−1.4	175	0.74
Std. dev.			28	4.53	21.36	1.72	2.9	4.8	44	0.11

Vol., volume calculated from van der Waals radii (Creighton 1993); Bulk., bulkiness index (Zimmerman, Eliezer, and Simha 1968); Polarity, polarity index (Zimmerman, Eliezer, and Simha 1968); pI, pH of the isoelectric point (Zimmerman, Eliezer, and Simha 1968); Hyd.1, hydrophobicity scale (Kyte and Doolittle 1982); Hyd.2, hydrophobicity scale (Engelman, Steitz, and Goldman 1986); Surface area, surface area accessible to water in unfolded peptide (Miller *et al.* 1987); Fract. area, fraction of accessible area lost when a protein folds (Rose *et al.* 1985).

From Higgs and Attwood, bioinformatics and molecular Evolution chap2

Results from a PCA, principal component analysis)



Neutral, nonpolar	W, F, G, A, V, I, L, M, P
Neutral, polar	Y, S, T, N, Q, C
Acidic	D, E
Basic	K, R, H

	Vol	**Bulk.**	**Pol.**	**pI**	**Hyd.1**	**Hyd.2**	**S.A.**	**Fr.A.**
Comp. 1	(0.06,	–0.22,	0.44,	0.19,	–0.49,	–0.51,	0.10,	–0.45)
Comp. 2	(0.58,	0.48,	0.10,	0.25,	0.03,	–0.03,	0.56,	0.17)

BLOSUM scoring matrices for amino acids

BLUSOM scoring matrices are another popular models for amino acid evolution.

The models are derived from multiple sequence alignments without resorting to an evolutionary tree and substitution model.

Given a multiple sequence alignment, we first count the number of times amino acid i is aligned with j, denoted A_{ij}.

```
A:  T L K K V Q K T
B:  T L K K V Q K T
C:  T L K K I Q K Q
D:  I I T K L Q K Q
E:  T I T K L Q K Q
F:  T L T K I Q K Q
G:  T L T Q I Q K Q
```

$$A=\begin{pmatrix} & \mathbf{I} & \mathbf{K} & \mathbf{L} & \mathbf{Q} & \mathbf{T} & \mathbf{V} \\ \mathbf{I} & 8 & - & 16 & - & 6 & 6 \\ \mathbf{K} & - & 78 & - & 6 & 12 & - \\ \mathbf{L} & 16 & - & 22 & - & - & 4 \\ \mathbf{Q} & - & 6 & - & 62 & 10 & - \\ \mathbf{T} & 6 & 12 & - & 10 & 44 & - \\ \mathbf{V} & 6 & - & 4 & - & - & 2 \end{pmatrix}$$

If amino acids i and j occurs n and m times in a column, respectively, the total number of alignments between a and b in this column is mn.

If amino acid i occurs n times in a column, the total number of alignments among these n amino acids in the column is $n(n\text{-}1)$.

BLOSUM scoring matrices for amino acids

The frequency of aligning i with j in the multiple alignments is,

$$q_{ij} = \frac{A_{ij}}{A_{tot}}.$$

where A_{tot} is the sum of the items in the matrix A, *i.e.*, the total number of pair-wise alignments among amino acids in the multiple alignments.

The relative frequency of A_{ij} compared to two randomly selected sequence is,

$$R_{ij} = \frac{q_{ij}}{\pi_i \pi_j}.$$

Similarly, we can define the score to align amino acid i and j as,

$$S(i,j) = c \log R_{ij} = c \log \frac{q_{ij}}{\pi_i \pi_j}.$$

Based on the similarity levels of the sequences in the multiple alignments, we can define different scoring matrices at different evolutionary distances, such as,

BLOSUM62: sequence identity < 62%, and
BLOSUM80: sequence identity < 80%.

Make practice with multiple alignment software
(see ya on Monday!)

Clustal Omega
COBALT
EMBOSS
MUSCLE
3D COFFEE
…