The construction of the Dayhoff matrix

A model of evolution:

In the absence of a valid model derived from first principles, an empirical approach seems more appropriate to score amino acid similarity. This approach is based on the assumption that once the evolutionary relationship of two sequences is established, the residues that did exchange are similar. This is the principle behind the mutation matrices compiled by Margaret O. Dayhoff and colleagues at the National Biomedical Research Foundation in the early 70s (Dayhoff, M.O. *et al.* (1978) Atlas of Protein Sequence and Structure. Vol. 5, Suppl. 3 National Biomedical Research Foundation, Washington D.C. U.S.A). They developed a precise and rigorous approach to implement a model of evolutionary change in their muatation data matrix.Their model allows to quantify the odds that a given alignment of two protein sequences would be observed by chance or would demonstrate an echo of a common ancestral sequence.

We will discuss this approach in some detail as a paradigm for the construction of any mutation data matrix. Note that any scoring matrix is a tool to measure the degree of conformity to the model underlying the construction of the matrix. Two amino acids will be termed similar, if they conform well to the model behind the matrix. An alignment is optimal, if it scores better in the application of the matrix than any other alignment. And two proteins will be significantly related, if the evolutionary process explains the relationship between their sequences better than chance. Since different mutation data matrices are appropriate for different alignment problems and the matrices sometimes even have to be edited by hand to solve difficult problems, the following section discusses some of the principles behind the construction of mutation data matrices and their use.

How is the Dayhoff mutation data matrix constructed ? Since we are looking for a matrix that will recognize significant evolutionary relationships, we must first define a model of evolution. The model used here states that **proteins evolve through a succesion of independent point mutations, that are accepted in a population and subsequently can be observed in the sequence pool.** Formally, the fact that the mutations are treated independent of their neigborhood and of their history makes this a Markov model. We can define an evolutionary distance between two sequences as the number of point mutations that was necessary to evolve one sequence into the other. (More specifically, the distance is the minimal number of mutations.) Two aspects of this process cause the evolutionary distance to be unequal in general to the number of observed differences between the sequences:

- First, there is a chance that a certain residue may have mutated, than reverted, hiding the effect of the mutation. This phenomenon is important in the evaluation of biological clocks and the question of how many mutational events may become fixed per time unit. In the context of the discussion of mutation matrices we do not need to consider this effect.
- Second, specific residues may have mutated more than once, thus the number of point mutations is likely to be larger than the number of differences between the two sequences. This has to be taken into account.

First step: Pair Exchange Frequencies

PAMs:

M.O. Dayhoff and colleagues introduced the term **"accepted point mutation"** for a mutation that is stably fixed in the gene pool in the course of evolution. Thus a measure of evolutionary distance between two sequences can be defined:

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

In order to identify accepted point mutations, a complete phylogenetic tree including all ancestral sequences has to be constructed. (There are standard procedures for this, which we will discuss at a later point.) To avoid a large degree of ambiguities in this step, Dayhoff and colleagues restricted their analysis to sequence families with more than 85% identity. Since the evolutionary distance between these highly homologous proteins is small, the construction of the phylogenetic tree can be achieved without to many complicating assumptions. For each of the observed and inferred sequences, the amino acid pair exchanges are tabulated into a 20x20 matrix. It is assumed, that the likelihood of an amino-acid X being replaced by an amino acid Y is the same as Y replacing X. Hence the matrix is constructed

symmetrically. (This assumption is probably largely true, if it were not, the amino acid composition of proteins would be in evolutionary disequilibrium.) Note that this process is different from comparing observed sequences directly with each other ! A_{ij} is the number of accepted mutations observed where amino acid ***i*** replaces amino acid ***j***.

Second step: Frequencies of Occurrence

If the properties of amino acids differ and if they occur with different frequencies, all statements we can make about the average properties of sequences will depend on the **frequencies of occurence** of the individual amino acids. These frequencies of occurence are approximated by the frequencies of observation. They are the number of occurences of a given amino acid divided by the number of amino-acids observed.

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

The sum of all f_i is one.

Amino acid frequencies:

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

The frequencies in the middle column are taken from Dayhoff (1978), the frequencies in the right column are taken from the 1991 recompilation of the mutation matrices by Jones *et al.* (Jones, D.T. Taylor, W.R. & Thornton, J.M. (1991) CABIOS 8:275-282) representing a database of observations that is approximately 40 times larger than that available to Dayhoff. Reassuringly, the changes are small. Note the higher abundance of hydrophobic residues, this may reflect the addition of many membrane proteins to the sequence databases in the time since 1978.

Third step: Relative Mutabilities

To obtain a complete picture of the mutational process, the amino-acids that do not mutate must be taken into account too. We need to know: what is the chance, on average, that a given amino acid will mutate at all. This is the relative mutability of the amino acid. It is obtained by multiplying the number of observed changes by the amino acids frequency of occurence.

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

Relative mutabilities of amino acids:

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

All values are taken relative to alanine, which is arbitrarily set at 100. Again, the 1991 data are from Jones *et al.* and the 1978 data from Dayhoff *et al.* The difference for some amino acids are quite significant, especially for those amino acids, for which hardly any exchanges have been observed in 1978 (C and W). Serine and threonine are the most mutable amino acids, cysteine and tryptophane are the most immutable.

Fourth step: Mutation Probability Matrix

With these data the probability that an amino acid in row ***i*** of the matrix will replace the amino acid in column ***j*** can be calculated: it is the mutability of

amino acid ***j***, multiplied by the relative pair exchange frequency (the pair exchange frequency for ***ij*** divided by the sum of all pair exchange frequencies for amino acid ***i***).

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

A_{ij} is a pair exchange frequency from the tabulated matrix of accepted point mutations, m_j is the mutability of amino acid ***j***

The diagonal elements represent the probability, that the amino acid will remain unchanged: they are (the number of ocurrences - the number of changes) divided by the number of occurences, or, simplified,

$$M_{ij} = 1 - m_j$$

Fifth step: The Evolutionary Distance Scale

Since the M_{ij} represent the probabilites for amino acids to remain conserved, if we scale all cells of our matrix by a constant factor λ we can scale the matrix to reflect a specific overall probability of change. We may chose λ so that the expected number of changes is 1 %, this gives the **matrix for the evolutionary distance of 1 PAM**. An average protein will have the composition:

$$n_i = f_i N$$

with N being the length of the protein and n_i the number of amino acids of a certain type. The number of amino acids of type ***i*** that will change in the evolutionary interval represented by this matrix is

$$n_i M_{ij}$$

and the number of total changes is the sum over all individual changes:

$$n_i M_{ij}$$

after introduction of the scaling factor λ, the mutation probability elements become

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

and the diagonal elements become

$$M_{ij} = 1 - \lambda m_j$$

It is apparent that the number of amino acids that are expected to change according to the matrix depends only on the factor λ. But this is true only for the evolutionary distances, from which the data were compiled. In the case of the Dayhoff matrices, where sequences with les than 15 % differences were used, scaling for 1 to 5 PAM should be permissible, from the original data. For higher evolutionary distances, extrapolation can not be simply done by adjusting λ, since this would neglect overlapping mutations.

In the framework of this model, **a mutation probability matrix for any distance can be obtained by multiplying the 1 PAM matrix with itself the required number of times**. Thus the 3 PAM matrix is simply the cube of the 1 PAM matrix. Obviously, the accuracy of this process of extrapolation will depend on the accuracy of the 1 PAM matrix - but up to that accuracy, it is rigorously correct.

Two facts should be pointed out:

The matrix at time PAM 0 is simply the unitary matrix...

$$M^0 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

The Matrix at infinite evolutionary disatance is:

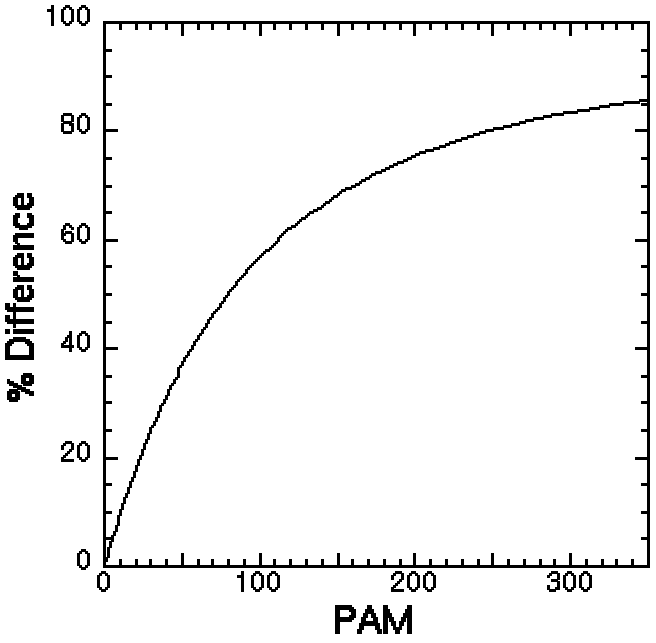
$$M^\infty = \begin{pmatrix} f_1 & f_1 & \cdots & f_1 \\ f_2 & f_2 & \cdots & f_2 \\ \vdots & \vdots & \ddots & \vdots \\ f_{20} & f_{20} & \cdots & f_{20} \end{pmatrix}$$

Thus at large distances, the amino acids are simply expected to be replaced according to the the average composition - the protein mutates beyond recognizabiltiy.

Estimation of evolutionary distance:

As noted above, the diagonal matrix elements give an indication of the evolutionary interval.It is clear, that the evolutionary interval is only equivalent to the % difference between two sequences at very low PAM distances. The correspondence is given below:

%Difference	PAM
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328



Note that the PAM250 matrix corresponds to approxiamtely 20% identities. the 50% identities level is approximately PAM100

The asymptote of this function is the % difference score of two randomly aligned sequences: since the probability for the chance occurence of an amino acid pair is the product of the probabilites for each amino acid, the probability for amino acid identities is the square of the probabilities for each amino acid. Thus the expected percent difference score is $100(1-\sum_{i=1}^{20} f_i^2)$. For the Jones and Taylor tabulation of frequencies this is 94.2%. This is close to the value of 95% that you would expect if all amino acids occured with equal frequencies. This means: even for randomly aligned sequences, you would expect to get around 5% identities.

Sixth step: Relatedness odds

The mutation probability matrix gives the probability M_{ij}, that an amio acid ***i*** will replace an amino acid of type ***j*** in a given evolutionary interval, in two related sequences, given the evolutionary model we have applied. By comparison, the probability that that same event is observed **by random chance** p_i^{ran} is simply given by the frequency of occurence of amino acid ***i***.

$$p_i^{ran} = f_i$$

Then the relative odds that a given event is due to evolution, rather than chance are

$$R_{ij} = \frac{M_{ij}}{p_i^{ran}}$$

Final step: the log-odds matrix

Finally we have a tool to quantify the probability that two sequences are homologous, i.e. related by evolution. The relatedness odds for one aligned pair are given by the corresponding matrix element. The relatedness odds for a second pair are multiplied with those of the first pair to give the joint probability. This is done for all aligned pairs of the whole sequence. Since multiplication is a computationally expensive process, it is preferrable to add the logarithms of the matrix elements. This matrix, the log odds matrix, is the foundation of quantitative sequence comparisons under an evolutionary model. Since the Dayhoff matrix was taken as the log to base 10, a value of +1 would mean that the corresponding pair has been observed 10 times more frequently than expected by chance. A value of -0.2 would mean that the observed pair was observed 1.6 times less frequently than chance would predict. The most commonly used matrix is the matrix from the 1978 edition of the Dayhoff atlas, at PAM 250: this is also frequently referred to as the **MDM78 PAM250** matrix.

The MDM78 PAM 250 matrix:

Alternate symbol comparison table for the comparison of peptide sequences.

Dayhoff table (Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. [1979] in Atlas of Protein Sequence and Structure, Dayhoff, M. O. Ed, pp. 345-352 (Figure 84), National Biomedical Research Foundation, Washington D.C.) rescaled by dividing each value by 5.0

May 24, 1990 16:47

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	
0.4	0.0	-0.4	0.0	0.0	-0.8	0.2	-0.2	-0.2	-0.2	-0.4	-0.2	0.0	0.2	0.0	-0.4	0.2	0.2	0.0	-1.2	-0.6	0.0	A
	0.5	-0.9	0.6	0.4	-1.0	0.1	0.3	-0.4	0.1	-0.7	-0.5	0.4	-0.2	0.3	-0.1	0.1	0.0	-0.4	-1.1	-0.6	0.4	B
		2.4	-1.0	-1.0	-0.8	-0.6	-0.6	-0.4	-1.0	-1.2	-1.0	-0.8	-0.6	-1.0	-0.8	0.0	-0.4	-0.4	-1.6	0.0	-1.0	C
			0.8	0.6	-1.2	0.2	0.2	-0.4	0.0	-0.8	-0.6	0.4	-0.2	0.4	-0.2	0.0	0.0	-0.4	-1.4	-0.8	0.5	D
				0.8	-1.0	0.0	0.2	-0.4	0.0	-0.6	-0.4	0.2	-0.2	0.4	-0.2	0.0	0.0	-0.4	-1.4	-0.8	0.6	E
					1.8	-1.0	-0.4	0.2	-1.0	0.4	0.0	-0.8	-1.0	-1.0	-0.8	-0.6	-0.6	-0.2	0.0	1.4	-1.0	F
						1.0	-0.4	-0.6	-0.4	-0.8	-0.6	0.0	-0.2	-0.2	-0.6	0.2	0.0	-0.2	-1.4	-1.0	-0.1	G
							1.2	-0.4	0.0	-0.4	-0.4	0.4	0.0	0.6	0.4	-0.2	-0.2	-0.4	-0.6	0.0	-0.4	H
								1.0	-0.4	0.4	0.4	-0.4	-0.4	-0.4	-0.4	-0.2	0.0	0.8	-1.0	-0.2	-0.4	I
									1.0	-0.6	0.0	0.2	-0.2	0.2	0.6	0.0	0.0	-0.4	-0.6	-0.8	0.1	K
										1.2	0.8	-0.6	-0.6	-0.4	-0.6	-0.6	-0.4	0.4	-0.4	-0.2	-0.5	L
											1.2	-0.4	-0.4	-0.2	0.0	-0.4	-0.2	0.4	-0.8	-0.4	-0.3	M
												0.4	-0.2	0.2	0.0	0.2	0.0	-0.4	-0.8	-0.4	0.2	N
													1.2	0.0	0.0	0.2	0.0	-0.2	-1.2	-1.0	-0.1	P
														0.8	0.2	-0.2	-0.2	-0.4	-1.0	-0.8	0.6	Q
															1.2	0.0	-0.2	-0.4	0.4	-0.8	0.6	R
																0.4	0.2	-0.2	-0.4	-0.6	-0.1	S
																	0.6	0.0	-1.0	-0.6	-0.1	T
																		0.8	-1.2	-0.4	-0.4	V
																			3.4	0.0	-1.2	W
																				2.0	-0.8	Y
																					0.6	Z

The variant of the MDM78 PAM250 used in the GCG package

The matrix in most widespread use is entirely derived from the Dayhoff **MDM78** (Gribskov, M. & Burgess, R.R. (1986) *NAR* **14**:6745-6763). It has been renormalized to give a score of 1.5 for identical residues, and the scores for non-identical residues have been adjusted to give a mean of -0.17 and a standard deviation of 0.364. The negative expectation value is necessary for local alignment routines, since a positive expectation value would allow to increase the score

simply by extending random alignments. Note that these normalizations constitute a significant departure from the evolutionary model as discussed above. Specifically, the interpretation in terms of a specific evolutionary distance is now no longer possible.

This Matrix is the default matrix used by most routines of the **UWGCG program package** and it is stored in the runtime data directory (logical name GENRUNDATA:) under the following names:

```
COMPARPEP.CMP
NWSGAPPEP.CMP
PEPDISTANCES.CMP
PILEUPPEP.CMP
PLOTSIMPEP.CMP
PRETTYPEP.CMP
PROFILEPEP.CMP
REPEATPEP.CMP
SEGGAPPEP.CMP
SEGPEP.CMP
SWGAPPEP.CMP
```

The GCG matrix:

```
Default scoring matrix used by COMPARE for the comparison of
protein sequences.

Dayhoff table (Schwartz, R. M. and Dayhoff, M. O. [1979] in Atlas of
Protein Sequence and Structure, Dayhoff, M. O. Ed, pp. 353-358, National
Biomedical Research Foundation, Washington D.C.) rescaled by dividing
each value by the sum of its row and column, and normalizing to a mean
of 0 and standard deviation of 1.0.  The value for FY (Phe-Tyr) = RW =
1.425.  Perfect matches are set to 1.5 and no matches on any row are
better than perfect matches.

Table used by Gribskov and Burgess NAR 14(16) 6745-6763
```

December 29, 1986 12:46

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	..
1.5	0.2	0.3	0.3	0.3	-0.5	0.7	-0.1	0.0	0.0	-0.1	0.0	0.2	0.5	0.2	-0.3	0.4	0.4	0.2	-0.8	-0.3	0.2	A
	1.1	-0.4	1.1	0.7	-0.7	0.6	0.4	-0.2	0.4	-0.5	-0.3	1.1	0.1	0.5	0.1	0.3	0.2	-0.2	-0.7	-0.3	0.6	B
		1.5	-0.5	-0.6	-0.1	0.2	-0.1	0.2	-0.6	-0.8	-0.6	-0.3	0.1	-0.6	-0.3	0.7	0.2	0.2	-1.2	1.0	-0.6	C
			1.5	1.0	-1.0	0.7	0.4	-0.2	0.3	-0.5	-0.4	0.7	0.1	0.7	0.0	0.2	0.2	-0.2	-1.1	-0.5	0.9	D
				1.5	-0.7	0.5	0.4	-0.2	0.3	-0.3	-0.2	0.5	0.1	0.7	0.0	0.2	0.2	-0.2	-1.1	-0.5	1.1	E
					1.5	-0.6	-0.1	0.7	-0.7	1.2	0.5	-0.5	-0.7	-0.8	-0.5	-0.3	-0.3	0.2	1.3	1.4	-0.7	F
						1.5	-0.2	-0.3	-0.1	-0.5	-0.3	0.4	0.3	0.2	-0.3	0.6	0.4	0.2	-1.0	-0.7	0.3	G
							1.5	-0.3	0.1	-0.2	-0.3	0.5	0.2	0.7	0.5	-0.2	-0.1	-0.3	-0.1	0.3	0.5	H
								1.5	-0.2	0.8	0.6	-0.3	-0.2	-0.3	-0.3	-0.1	0.2	1.1	-0.5	0.1	-0.2	I
									1.5	-0.3	0.2	0.4	0.1	0.4	0.8	0.2	0.2	-0.2	0.1	-0.6	0.4	K
										1.5	1.3	-0.4	-0.3	-0.1	-0.4	-0.4	-0.1	0.8	0.5	0.3	-0.2	L
											1.5	-0.3	-0.2	0.0	0.2	-0.3	0.0	0.6	-0.3	-0.1	-0.1	M
												1.5	0.0	0.4	0.1	0.3	0.2	-0.3	-0.3	-0.1	0.4	N
													1.5	0.3	0.3	0.4	0.3	0.1	-0.8	-0.8	0.2	P
														1.5	0.4	-0.1	-0.1	-0.2	-0.5	-0.6	1.1	Q
															1.5	0.1	-0.1	-0.3	1.4	-0.6	0.2	R
																1.5	0.3	-0.1	0.3	-0.4	0.0	S
																	1.5	0.2	-0.6	-0.3	0.1	T
																		1.5	-0.8	-0.1	-0.2	V
																			1.5	1.1	-0.8	W
																				1.5	-0.6	Y
																					1.1	Z

formatado de: http://www.lmb.uni-muenchen.de/Groups/Bioinformatics/04/ch_04_3.html