Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Theoretical Biology 228 (2004) 97–106

Journal of Theoretical Biology

www.elsevier.com/locate/jtbi

A new criterion and method for amino acid classification

Carolin Kosiol[a,*], Nick Goldman[b], Nigel H. Buttimore[a]

[a] *School of Mathematics, Trinity College, University of Dublin, Dublin 2, Ireland*
[b] *EMBL-EBI, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK*

Received 23 September 2003; received in revised form 4 December 2003; accepted 10 December 2003

Abstract

It is accepted that many evolutionary changes of amino acid sequence in proteins are conservative: the replacement of one amino acid by another residue has a far greater chance of being accepted if the two residues have similar properties. It is difficult, however, to identify relevant physicochemical properties that capture this similarity. In this paper we introduce a criterion that determines similarity from an evolutionary point of view. Our criterion is based on the description of protein evolution by a Markov process and the corresponding matrix of instantaneous replacement rates. It is inspired by the conductance, a quantity that reflects the strength of mixing in a Markov process. Furthermore we introduce a method to divide the 20 amino acid residues into subsets that achieve good scores with our criterion. The criterion has the time-invariance property that different time distances of the same amino acid replacement rate matrix lead to the same grouping; but different rate matrices lead to different groupings. Therefore it can be used as an automated method to compare matrices derived from consideration of different types of proteins, or from parts of proteins sharing different structural or functional features. We present the groupings resulting from two standard matrices used in sequence alignment and phylogenetic tree estimation.

Keywords: Protein evolution; Groupings of amino acids; Markov models; Conductance

1. Introduction

In this paper we develop a criterion and grouping method to classify amino acids, starting from amino acid replacement matrices. Our criterion and grouping method identify disjoint sets of amino acids with a high probability of change amongst the elements of each set but small probabilities of change between elements of different sets. These sets, or groupings, are easy to understand and may be readily compared for the numerous amino acid replacement matrices that have arisen in the last 30 years from studies of protein evolution.

Dayhoff et al. (1978) introduced a Markov model of protein evolution that resulted in the development of the widely used amino acid replacement matrices known as the PAM matrices. Jones et al. (1992) employed much the same methods, but based the estimation of the JTT matrix on a larger sequence database. In contrast, Whelan and Goldman (2001) used a maximum likelihood estimation method to generate the WAG matrix. The PAM, JTT and WAG matrices have some significantly different entries. However the differences between the models and matrices are not fully understood. Groupings of the 20 amino acids derived directly from the matrices are a suitable tool to analyse and compare these different models, and are much more comprehensible than graphical or tabular representations of a rate matrix.

The PAM, JTT and WAG models give increasingly good descriptions of the average patterns and processes of evolution of large collections of sequences (Whelan and Goldman, 2001), but such "average" models can fail to describe proteins with particular functions and structures (Goldman et al., 1998). For example, in functional domains where hydrophobicity is important, glycine (G) is rarely replaced by arginine (R). However, the replacement between the two occurs more frequently when hydrophobicity is not important. Likewise many seemingly different amino acids are good helix-formers

*Corresponding author. EMBL–EBI Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK. Tel.: +44-1223-494655; fax: +44-1223-494468.

E-mail address: kosiol@ebi.ac.uk (C. Kosiol).

0022-5193/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2003.12.010

(A, L) and many seemingly similar amino acids may differ much in their contribution to forming helices (L, M) (Goldman et al., 1998). Thus, when evolution "sees" a helix structure as important, rates of amino acid replacements will depend on the propensity for helix-formation of the amino acids. In such cases the amino acid replacement models have been improved by incorporating functional and structural properties of the proteins. For example, Overington et al. (1990) observed different numbers of amino acid replacements for different structural environments, Jones et al. (1994) demonstrated that transmembrane proteins have markedly different replacement dynamics and Goldman et al. (1998) considered different categories, e.g. α-helices, β-sheets, turns and loops, with each category further classified by whether it is exposed to solvent or buried in the protein core, and inferred an amino acid replacement matrix for each of the categories. Furthermore, Adachi and Hasegawa (1996) and Yang et al. (1998) have implemented models of amino acid replacement derived from mitochondrially encoded proteins, and Adachi et al. (2000) compiled a chloroplast-derived amino acid replacement matrix.

To date, little progress has been made in comparing amino acid replacement matrices such as those described above. A reliable and biologically meaningful method to summarize the information they contain, and that could lead to comparisons and contrasts of replacement patterns in different parts of proteins and in proteins of different types, would assist in a better understanding of protein sequence evolution. In addition to its uses in studying models of protein sequence evolution, other important applications of amino acid groupings have already been established; for example, by Wang and Wang (1999) in protein design and modelling and by Coghlan et al. (2001) to develop filtering algorithms for protein databases.

Several methods have been proposed to classify amino acids, although rarely using evolutionary information. Grantham (1974) introduced an amino acid distance formula that considers the chemical composition of the side chain, the polarity and the molecular volume. This approach was extended by Xia and Li (1998) in a study of the relevance of 10 amino acid properties to protein evolution. Grantham and Xia and Li presented their results in the form of distance matrices, whereas French and Robson (1983) arranged their results in two-dimensional diagrams using multidimensional scaling. Taylor (1986) also adopted a graphical approach and developed Venn diagrams of amino acids sets. The unions and intersections of a Venn diagram allow determination of (potentially hierarchical) sets of amino acids that might be conserved. The number of possible subsets is large, however, and includes many that have little physical meaning. The interpretation of these Venn diagrams requires detailed expert knowledge.

Accordingly, a recent approach from Cannata et al. (2002) is interesting since it automates the group-finding process. These authors propose a branch and bound analysis based on amino acid replacement probability matrices, but their method suffers from two problems. First, the classification criterion used has no clear evolutionary meaning. Second, the approach leads to different groupings for different time periods of the same matrix (e.g., PAM120 and PAM250), whereas we desire a criterion that is dependent on the replacement patterns of evolution but independent of the time scale on which evolution may be observed.

The method we propose in this paper has its origin in the convergence diagnosis of Markov chain Monte Carlo (Behrends, 2000). In Section 2 we give a general introduction to the Markov models used to describe amino acid evolution. Section 3 introduces the conductance, a measure for the grouping of amino acids into non-empty sets, whose value will indicate the quality of the classification. Unfortunately, the measure itself does not suggest a method that would determine optimal groupings. In Sections 4 and 5 we explain the relationship between the eigenvalues and eigenvectors and the structure of the amino acid replacement matrix. Mathematically speaking, we are looking for a structure of the Markov matrix that is almost of block diagonal type. Markov matrices that show an almost block diagonal structure also show a low conductance value. The identification of the block structure leads to an algorithm that produces groupings for a given amino acid matrix. This algorithm is given in Section 6. We apply the conductance measure and the grouping algorithm to standard amino acid replacement matrices in Section 7 and finally discuss our results in Section 8.

2. Markov processes and amino acid evolution

Proteins are sequences of amino acids. The Markov model asserts that one protein sequence is derived from another protein sequence by a series of independent mutations, each changing one amino acid in the first sequence to another amino acid in the second during evolution. Thereby we assume independence of evolution at different sites.

The continuous-time Markov process is a stochastic model in which $P_{ij}(t)$ gives the probability that amino acid i will change to amino acid j at any single site after any time $t > 0$. Since there are 20 amino acids, i and j take the values $1, 2, \ldots, 20$ and we can write the probability $P_{ij}(t)$ as a 20×20 matrix that we denote $P(t)$. $P(t)$ is a Markov matrix, and so the rows sum to 1. In evolutionary studies it is necessary to be able to compute the probability matrix $P(t)$ for any real evolutionary time (distance) $t \geqslant 0$. This is achieved using an instantaneous rate matrix $Q = (Q_{ij})_{i,j=1,\ldots,20}$, which is related to $P(t)$ via

$$P(t) = e^{tQ} = I + tQ + \frac{(tQ)^2}{2!} + \frac{(tQ)^3}{3!} + \cdots$$

Q has its off-diagonal entries $Q_{ij,i \neq j}$ equal to the instantaneous rates of replacement of i by j. Its diagonal entries Q_{ii} are defined by the mathematical requirement that each row sum is zero. Typically in phylogenetic applications, Q is normalized so that the mean rate of replacement at equilibrium ($\sum_i \sum_{j \neq i} \pi_i Q_{ij}$, where π_i is the equilibrium frequency of amino acid i) is 1, meaning that times t are measured in units of expected numbers of changes per site. The probability matrix $P(t)$ for any time $t > 0$ is fully determined by the instantaneous rate matrix Q. Spectral decomposition shows that Q and $P(t)$ for any time $t > 0$ have the same eigenvectors (see Liò and Goldman, 1998).

Markov processes for amino acid sequence evolution can have two important properties: connectedness and reversibility. In a connected process there is a time $t > 0$ such that

$$P_{ij}(t) > 0 \quad \text{for all } i, j \in \{1, 2, \ldots, 20\}.$$

Connected Markov processes have a unique equilibrium distribution π such that $\pi^{\mathrm{T}} Q = 0$, or equivalently:

$$\pi^{\mathrm{T}} P(t) = \pi^{\mathrm{T}} \quad \text{for } t > 0,$$

where the superscript T denotes transpose. The vector π is also the limiting distribution of amino acids when time approaches infinity. Reversibility means that

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t) \quad \text{for all } i, j \in \{1, \ldots, 20\} \text{ and } t > 0.$$

A consequence of reversibility is that the process of protein sequence evolution is statistically indistinguishable from the same process observed in reverse.

3. A measure for amino acids sets

Our goal is to identify sets of amino acids with a high probability of change amongst the elements of each set but small probability of change between elements of different sets. Our starting point is to consider amino acid replacement matrices $P(t)$, for example the PAM series (Dayhoff et al., 1978). In order for groupings to be interpretable in terms of the processes of evolutionary replacement of amino acids, and not levels of divergence between protein sequences, we expect that groupings should perform equally under measures based on (e.g.) PAM120 or PAM250 (the PAM matrices $P(t)$ for t approximately equal to 1.20 and 2.50, respectively (Dayhoff et al., 1978)), and that optimal groupings derived from these matrices should be the same. The measure presented here has been inspired by the conductance, a measure of the strength of mixing of a Markov process that is used in the convergence diagnosis of Markov chain Monte Carlo methods (see Sinclair, 1992). Below, we redefine the conductance in terms of the instantaneous rate matrix Q instead of the probability matrix $P(t)$, to fulfill the requirement for independence of our measure and particular times t.

Let Q define a Markov process that is connected and reversible with equilibrium distribution π, and is normalized so that the mean rate of replacement at equilibrium is 1. (The mean rate of replacement is given by $\sum_i \sum_{j \neq i} \pi_i Q_{ij}$. Dividing Q by this mean rate of replacement provides a matrix with a mean rate of unity, so that evolutionary distances t are measured in units of expected numbers of changes per site (Liò and Goldman, 1998). Now consider an amino acid sequence of N sites. The expected number of changes of i to j per unit time is $N \pi_i Q_{ij}$, or $\pi_i Q_{ij}$ per site. Similar analysis can be carried out for sets of amino acids. Let $A_1, \ldots, A_K$ be K proper subsets of the set of amino acids $A = \{1, \ldots, 20\}$, where $A_k \cap A_l = \emptyset$ for $k, l = 1, \ldots, K, k \neq l$, and $\bigcup_k A_k = A$. If π_i is the i-th component of the equilibrium distribution π, we expect to observe

$$N \cdot \sum_{i \in A_k, j \in A_l} \pi_i Q_{ij}$$

changes per unit time from subset A_k to subset A_l, $k \neq l$, in the whole sequence, or

$$F_{kl} = \sum_{i \in A_k, j \in A_l} \pi_i Q_{ij}$$

changes per site. The quantity F_{kl} is called the *flow* from subset A_k to subset A_l.

When the Markov process is close to equilibrium, the frequencies of the amino acids remain more or less the same. The frequency of amino acids of subset A_k, called the *capacity* C_k of A_k, is then

$$C_k = \sum_{i \in A_k} \pi_i.$$

The ratio

$$\Phi_{kl} = \frac{F_{kl}}{C_k}, \quad k \neq l,$$

is called the *conductance* (Behrends, 2000). This is the expected number of changes from subset A_k to subset A_l per site per unit time when commencing at subset A_k.

Using the above definition we can define a new matrix $\Phi = (\Phi_{kl})_{k,l=1,\ldots,K,k \neq l}$:

$$\Phi = \begin{pmatrix} \Phi_{11} & \Phi_{12} & \ldots & \Phi_{1K} \\ \Phi_{21} & \Phi_{22} & \ldots & \Phi_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{K1} & \Phi_{K2} & \ldots & \Phi_{KK} \end{pmatrix},$$

where the diagonal entries Φ_{kk} are given by the mathematical requirement that each row sums to zero. The matrix Φ is itself an instantaneous rate matrix. If we have "perfect" subsets, no changes between the subsets can be observed and $F_{kl} = 0$ for all $k, l, k \neq l$. Thus Φ

would be a null matrix. In the more general "imperfect" case, the expression

$$\varphi = \sum_{k} \sum_{l \neq k} \Phi_{kl}$$

measures the difference between Φ and the null matrix. We therefore use φ as our measure of the quality of the partition of the set A of 20 amino acids into K groups $A_1, \dots, A_K$.

Example 1. To set ideas, we consider a simple illustrative system of seven amino acids with rate matrix Q having the following block diagonal form:

$$\begin{pmatrix} -0.35 & 0.35 & 0 & 0 & 0 & 0 & 0 \\ 0.35 & -0.35 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2.1 & 2.1 & 0 & 0 & 0 \\ 0 & 0 & 2.1 & -2.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.7 & 0.35 & 0.35 \\ 0 & 0 & 0 & 0 & 0.35 & -0.7 & 0.35 \\ 0 & 0 & 0 & 0 & 0.35 & 0.35 & -0.7 \end{pmatrix}.$$

The block diagonal structure of the rate matrix suggests the partition into $A_1 = \{1,2\}$, $A_2 = \{3,4\}$ and $A_3 = \{5,6,7\}$. Since this Markov process is reversible, the flow from set A_k to A_l is same as the flow from set A_l to set A_k:

$$F_{A_1 \to A_2} = F_{12} = F_{21} = 0,$$

$$F_{A_1 \to A_3} = F_{13} = F_{31} = 0,$$

$$F_{A_2 \to A_3} = F_{23} = F_{32} = 0.$$

The equilibrium distribution in this example is not unique, since the corresponding Markov process is not connected. The rates Φ_{kl}, however, are independent of any choice of equilibrium distribution. Since the F_{kl} are all zero, we get

$$\begin{pmatrix} \Phi_{11} & \Phi_{12} & \Phi_{13} \\ \Phi_{21} & \Phi_{22} & \Phi_{23} \\ \Phi_{31} & \Phi_{32} & \Phi_{33} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

for any equilibrium distribution. Finally, the conductance measure is given by

$$\varphi = \sum_{k} \sum_{l \neq k} \Phi_{kl} = 0.$$

In the general case, however, choosing a partition into sets is often not obvious. One may wish to consider all possible partitions. The total number of partitions of a set of n elements into non-empty subsets is the n-th Bell number, B_n (Weisstein, 1999). The Bell numbers are given by the recurrence

$$B_{n+1} = \sum_{i=0}^{n} \binom{n}{i} B_i,$$

where B_0 is defined to equal 1.

Example 2. To determine the number of possible partitions of the set of four letters $\{ATGC\}$, the fourth Bell number is computed as follows:

$$B_1 = \binom{0}{0} B_0 = 1,$$

$$B_2 = \binom{1}{0} B_0 + \binom{1}{1} B_1 = 2,$$

$$B_3 = \binom{2}{0} B_0 + \binom{2}{1} B_1 + \binom{2}{2} B_2 \\ = 1 + 2 + 2 = 5,$$

$$B_4 = \binom{3}{0} B_0 + \binom{3}{1} B_1 + \binom{3}{2} B_2 + \binom{3}{3} B_3 \\ = 1 + 3 + 6 + 5 = 15.$$

The 15 possible partitions into non-empty subsets are

$$\begin{array}{lll} \{ATGC\}, & \{AT\}\{GC\}, & \{A\}\{TC\}\{G\}, \\ \{A\}\{TGC\}, & \{AC\}\{GT\}, & \{T\}\{AG\}\{C\}, \\ \{ATG\}\{C\}, & \{AG\}\{TC\}, & \{G\}\{AT\}\{C\}, \\ \{AGC\}\{T\}, & \{A\}\{GT\}\{C\}, & \{G\}\{AC\}\{T\}, \\ \{ATC\}\{G\}, & \{A\}\{GC\}\{T\}, & \{A\}\{G\}\{C\}\{T\}. \end{array}$$

Cannata et al. (2002) have pointed out that for 20 amino acids there exist 51,724,158,235,372 (roughly 51×10^{12}) possible partitions. Furthermore, they list how these partitions are distributed among the partitions into particular numbers ($K = 1, \dots, 20$) of sets. For example, under the restriction of partitioning only into exactly eight sets, as many as 15×10^{12} partitions still have to be considered. This means that exhaustive enumeration of the groupings and calculation of the conductance measure to find the optimal grouping of 20 amino acids is out of the question. In Sections 4 and 5 we describe a heuristic algorithm that seeks optimal or near-optimal groupings of amino acids. One advantage of our algorithm is that the computational cost of searching for a high-quality partition of the 20 amino acids into K subsets is independent of the value of K and the algorithm can easily be run for all non-trivial values of K ($2, \dots, 19$) given any matrix Q. Once partitions of amino acids have been determined algorithmically one may calculate the conductance measure φ in order to exhibit the quality of the groupings.

4. Block structure of matrices

Example 1 has indicated that blocks within matrices can act as "traps" for the flow between the sets and that choosing a partition accordingly results in a low conductance score φ. In this section we state results that link certain properties of the eigenvalues and eigenvectors of an amino acid replacement matrix to a block diagonal or perturbed block diagonal structure of the matrix. The main idea is to identify an almost block diagonal structure of the replacement matrix in order to find good candidates with low conductance score φ among all possible partitions. The eigenvectors are especially suitable to identify time-independent groupings, since the eigenvalues for different time distances t of the probability matrix $P(t) = \mathrm{e}^{tQ}$ (for example, PAM120 and PAM250) are different, but the eigenvectors remain the same.

Suppose the eigenvalues λ_i of $P(t)$, where $1 \leqslant i \leqslant 20$, are ordered according to

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_{20}|.$$

By the Perron–Frobenius theorem (Behrends, 2000) all eigenvalues are real and are contained in $[-1, 1]$. Since $P(t)$ is reversible it is known that for every right eigenvector there is a corresponding left eigenvector that corresponds to the same eigenvalue. The greatest eigenvalue λ_1 is unity and is called the Perron root. The right eigenvector corresponding to λ_1 is $e = (1, \ldots, 1)^{\mathrm{T}}$. The corresponding left eigenvector $\pi = (\pi_1, \ldots, \pi_{20})^{\mathrm{T}}$ represents the equilibrium distribution under the assumption that it is normalized so that $\pi^{\mathrm{T}} e = 1$. In matrix notation we have

$$\pi^{\mathrm{T}} P(t) = \pi^{\mathrm{T}} \quad \text{and} \quad P(t)e = e \quad \text{for } t > 0.$$

The above results are true for a general Markov matrix. We will now focus on matrices where we can decompose the 20 amino acids into invariant subsets $A_1, \ldots, A_K$ of amino acids. This means that whenever the Markov process is in one of the invariant sets, e.g. A_1, it will remain in A_1 thereafter. If we use an appropriate ordering of the amino acid residues the amino acid replacement matrix $P(t)$ appears in block diagonal form

$$B = \begin{pmatrix} D_{11} & 0 & \cdots & 0 \\ 0 & D_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_{KK} \end{pmatrix},$$

where each block D_{kk} $(k = 1, \ldots, K)$ is a Markov matrix, reversible with respect to some corresponding equilibrium sub-distribution. Again, due to the Perron–Frobenius theorem, each *block* possesses a unique right eigenvector $e_k = (1, \ldots, 1)^{\mathrm{T}}$ of length $\dim(D_{kk})$ corresponding to its Perron root $\lambda_k = 1$.

In terms of the total amino acid replacement matrix $P(t)$, the eigenvalue $\lambda_1 = 1$ is K-fold and the K corresponding right eigenvectors can be written as linear combinations of the K vectors of the form

$$(0, \ldots, 0, e_k^{\mathrm{T}}, 0, \ldots, 0)^{\mathrm{T}}, \quad k = 1, \ldots, K.$$

As a consequence, right eigenvectors corresponding to $\lambda = 1$ are constant on each invariant set of states.

Example 3. To obtain a block diagonal probability matrix we calculate $B = P(t) = \mathrm{e}^{Qt}$, where Q is the block diagonal rate matrix of Example 1 and $t = 1$:

$$P(1) = \begin{pmatrix} 0.75 & 0.25 & 0 & 0 & 0 & 0 & 0 \\ 0.25 & 0.75 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.51 & 0.49 & 0 & 0 & 0 \\ 0 & 0 & 0.49 & 0.51 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.56 & 0.22 & 0.22 \\ 0 & 0 & 0 & 0 & 0.22 & 0.56 & 0.22 \\ 0 & 0 & 0 & 0 & 0.22 & 0.22 & 0.56 \end{pmatrix}.$$

The eigenvalues of $P(1)$ are

$$\lambda_1 = 1, \quad \lambda_2 = 1, \quad \lambda_3 = 1,$$

$$\lambda_4 = 0.5, \quad \lambda_5 = 0.34, \quad \lambda_6 = 0.34, \quad \lambda_7 = 0.02$$

and the right eigenvectors corresponding to $\lambda = 1$ are

$$x_1 = (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1),$$
$$x_2 = (0 \quad 0 \quad 1 \quad 1 \quad -1 \quad -1 \quad -1),$$
$$x_3 = (1 \quad 1 \quad -1 \quad -1 \quad -1 \quad -1 \quad -1).$$

Fig. 1 shows the eigenvectors x_1, x_2, and x_3 corresponding to $\lambda = 1$ as function of the seven states (corresponding to amino acids in our real application) s_i, $i \in \{1, \ldots, 7\}$. A constant level can be observed for each of the invariant sets $\{1, 2\}$, $\{3, 4\}$, and $\{5, 6, 7\}$. Moreover, the same pattern can be observed if we restrict our investigation to the sign structure σ_i, $i \in \{1, \ldots, 7\}$, of the states instead of the actual values. For example, the sign of state 1 is positive for eigenvectors x_1 and x_3 and is zero for eigenvector x_2. Thus, the sign structure σ_1 for state 1 can be written $(+, 0, +)$. Analogously, we determine the sign structure of all states:

$$\sigma_1 = (+, 0, +), \quad \sigma_2 = (+, 0, +),$$

$$\sigma_3 = (+, +, -), \quad \sigma_4 = (+, +, -),$$

$$\sigma_5 = (+, -, -), \quad \sigma_6 = (+, -, -),$$

$$\sigma_7 = (+, -, -).$$

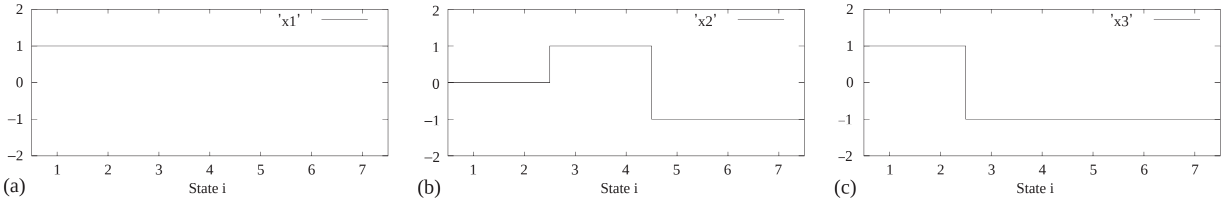

Fig. 1. The eigenvectors x_1, x_2, x_3 of Example 3, corresponding to $\lambda = 1$, as functions of the states.

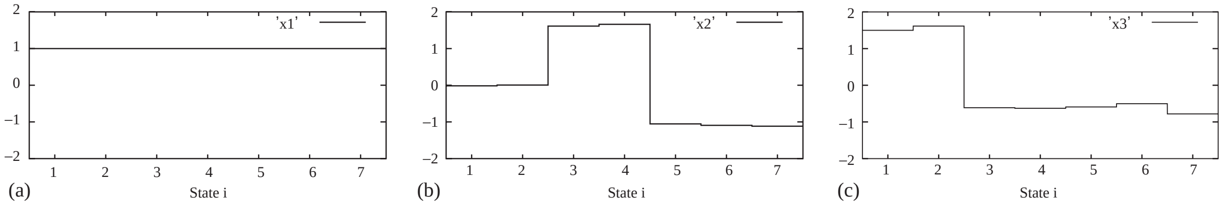

Fig. 2. The eigenvectors x_1, x_2, x_3 of Example 4, corresponding to $\lambda = 1, 0.85, 0.76$, as functions of the states.

The sign structure is the same for states of the same invariant set $\{1,2\}$, $\{3,4\}$ or $\{5,6,7\}$.

Stated more formally, and reverting to consideration of 20-state (amino acid) matrices, if we associate with every state its particular sign structure

$$s_i \mapsto (\text{sign}(x_1)_i, \ldots, \text{sign}(x_K)_i) \quad i = 1, \ldots, 20,$$

then the following statements hold:

- invariant sets are collections of states with common sign structure,
- different invariant sets exhibit different sign structures.

A proof is given in Deuflhard et al., 2000. This indicates that the set of K right eigenvectors of the amino acid replacement matrix can be used to identify K invariant sets of amino acid residues via the sign structure.

5. Perturbation theory

The standard amino acid replacement matrices like PAM (Dayhoff et al., 1978). and WAG (Whelan and Goldman, 2001) do not exhibit block diagonal structure. As mentioned in the introduction most amino acid replacement matrices are connected. This means that for any time $t > 0$ all the entries of their probability matrices $P(t)$ are non-zero. Therefore, it is impossible to identify perfect invariant sets. However, it is still possible to identify *almost* invariant sets of amino acids via the sign structures σ_i as the following example illustrates:

Example 4. We add a perbutation matrix E to the block diagonal matrix B of Example 3:

$$P := 0.8B + 0.2E,$$

where the perturbation matrix E is given as follows:

$$\begin{pmatrix} 0.01 & 0.09 & 0.10 & 0.25 & 0.08 & 0.30 & 0.17 \\ 0.09 & 0.10 & 0.25 & 0.08 & 0.30 & 0.17 & 0.01 \\ 0.10 & 0.25 & 0.08 & 0.30 & 0.17 & 0.01 & 0.09 \\ 0.25 & 0.08 & 0.30 & 0.17 & 0.01 & 0.09 & 0.10 \\ 0.08 & 0.30 & 0.17 & 0.01 & 0.09 & 0.10 & 0.25 \\ 0.30 & 0.17 & 0.01 & 0.09 & 0.10 & 0.25 & 0.08 \\ 0.17 & 0.01 & 0.09 & 0.10 & 0.25 & 0.08 & 0.30 \end{pmatrix}.$$

The eigenvalues of P are now calculated as

$$\lambda_1 = 1, \quad \lambda_2 = 0.85, \quad \lambda_3 = 0.76,$$
$$\lambda_4 = 0.41, \quad \lambda_5 = 0.31, \quad \lambda_6 = 0.24, \quad \lambda_7 = -0.02.$$

The eigenvalue spectrum of the perturbed block diagonal amino acid replacement matrix can then be divided into three parts: the Perron root $\lambda_1 = 1$, a cluster of two eigenvalues $\lambda_2 = 0.85$, $\lambda_3 = 0.76$ close to one, and the remaining part of the spectrum, which is bounded away from 1. The right eigenvectors x_1, x_2, x_3 corresponding to $\lambda = 1, 0.85, 0.76$ are:

$$(1, \quad 1, \quad 1, \quad 1, \quad 1, \quad 1, \quad 1),$$

$$(-0.02, \quad 0.01, \quad 1.61, \quad 1.66, \quad -1.05, \quad -1.09, \quad -1.12),$$

$$(1.50, \quad 1.61, \quad -0.61, \quad -0.63, \quad -0.59, \quad -0.50, \quad -0.78).$$

Fig. 2 shows that for the perturbed block diagonal Markov matrix, nearly constant level patterns can be observed on the three almost invariant sets $\{1,2\}$, $\{3,4\}$, and $\{5,6,7\}$. In order to have an automated procedure for determining the sign structure, we need to define a threshold value θ that will separate components with clear sign information from those that might have been perturbed to such an extent that the sign information has been lost. Elements $x_k(s)$ of x_k satisfying $|x_k(s)| > \theta$ are taken to have clear sign information, $\sigma_s(k) = +$ or $-$, whereas $\sigma_s(k) = 0$ if $|x_k(s)| < \theta$. For example, by choosing $\theta = 0.25$ in the above example, we ensure that all states $\{1, \ldots, 7\}$ still have clear defined sign structure and that at least one of the eigenvectors, apart from x_1, has a sufficiently large component $|x_k(s)| > \theta$. In this example, the small components of the eigenvectors $x_2(1) = -0.02$ and $x_2(2) = 0.01$ are neglected and we obtain the following sign structure:

$$\sigma_1 = (+, 0, +), \quad \sigma_2 = (+, 0, +),$$

$$\sigma_3 = (+, +, -), \quad \sigma_4 = (+, +, -),$$

$$\sigma_5 = (+, -, -), \quad \sigma_6 = (+, -, -),$$

$$\sigma_7 = (+, -, -).$$

This sign structure is identical to the sign structure of the unperturbed Markov matrix, leading to the same grouping of the states $\{1, 2\}$, $\{3, 4\}$, and $\{5, 6, 7\}$. Example 4 indicates that the sign structure of eigenvectors corresponding to eigenvalues in the cluster around the Perron root λ_1 can be used to identify sets of amino acids that are almost invariant. An exact formulation and proof of the behaviour of the eigenvectors under the influence of perturbation is given by Deuflhard et al. (2000).

6. Algorithm

This section transforms the results of Sections 4 and 5 above to an algorithm that has three steps:

(1) Find states with stable sign structure.
(2) Define equivalence classes.
(3) Sort states to seek almost invariant sets.

Step 1: Find states with stable sign structure

We start from the heuristic that the sign of an eigenvector component is "more likely" to remain stable under perturbation, the "larger" this component is. In order to make the positive and negative parts of the eigenvectors comparable in size, we scale them as follows:

For $k = 1, \ldots, K$, we split $x_k = x_k^+ + x_k^-$ componentwise, where $x_k^+(s) = \max(0, x_k(s))$ and $x_k^-(s) = \min(0, x_k(s))$, and we set $\tilde{x}_k = x_k^+/\|x_k^+\|_\infty + x_k^-/\|x_k^-\|_\infty$ (where $\|v\|_\infty$ is the maximum norm of vector v, defined as $\max_i |v(i)|$).

By means of a heuristic threshold value $0 < \delta < 1$, which is common for all eigenvectors, we then select those states that exhibit a "stable" sign structure according to

$$\mathcal{S} = \{s \in \{1, \ldots, N\}\colon \max_{k=1,\ldots,K} |\tilde{x}_k(s)| > \delta\}.$$

Only those states in $\mathcal{S}$ can be assigned to groups using the following procedure; states $s \notin \mathcal{S}$ are unclassifiable. Step 1 is a check that all of the states (i.e. amino acids) have at least one of the eigenvectors x_k, $k > 1$, with a "significantly" large component $x_k(s)$. We have chosen $\delta = 0.5$ for the amino acid replacement matrices we have investigated. In the case of the occurrence of unclassifiable states our algorithm aborts. However, one could then lower the value of δ at the expense of a higher risk of a false assignment of the states into subsets. This situation never arose in our examples of residue matrices.

Step 2: Define sign equivalence classes

Based on the sign structures of the states in $\mathcal{S}$, we proceed to define K equivalence classes with respect to sign structures. As already indicated the underlying idea is that only "significantly" large entries in the scaled vectors $\tilde{x}_k$ are permitted to contribute to a sign structure $\sigma_{(s,\theta)}$ for a state s with respect to some heuristic threshold value θ (with $0 < \theta < 1$) by virtue of

$$\sigma_{(s,\theta)} = (\sigma_1, \ldots, \sigma_K),$$

$$\text{with } \sigma(k) = \begin{cases} \operatorname{sign}(\tilde{x}_k(s)) & \text{if } |\tilde{x}_k(s)| > \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Two sign structures are defined to be equivalent if, and only if, their pointwise multiplication yields only non-negative entries. Sign structures of states that are not equivalent are said to be inequivalent.

Step 3: Sort states to seek almost invariant sets

In step 2 we have assigned a sign structure to all stable states. It is now necessary to sort the states with respect to their sign structure, compute the number of invariant sets and finally determine the invariant sets. Various methods can be applied to this challenge, and we have decided to transform the problem to a graph colouring problem. Therefore, we construct a graph where every stable state is represented by a vertex and in which inequivalent states are connected by edges. Colouring this graph determines K colour sets $\mathcal{S}_1, \ldots, \mathcal{S}_K$ and we assign each of the states in $\mathcal{S}$ to one sign structure class. An introduction to graph colouring and code that performs this task is available (Trick, 1994).

By combining the three steps above, we arrive at the following procedure to compute a partition into a particular number K of almost invariant sets:

```
Specify desired number of sets K
Read in the K eigenvectors with largest eigenvalues
Step 1: Find states with stable sign structure:
Set $\theta^- = 0$ and $\theta^+ = 1$
Step 2: Set $\tilde{\theta} = (\theta^- + \theta^+)/2$ (bisection search to find $\theta$
giving required number of subsets)
Determine the sign structures $\sigma_{(s,\theta)}$ with respect to $\tilde{\theta}$
Step 3: Calculate invariant sets and the number
of invariant sets, $\mathcal{K}(\tilde{\theta})$. Then:
  if ($\mathcal{K}(\tilde{\theta}) = K$) write out invariant sets
  else if ($\mathcal{K}(\tilde{\theta}) > K$) set $\theta^+ = \tilde{\theta}$ and goto Step2
  else set $\theta^- = \tilde{\theta}$ and goto Step2
```

7. Results

The above algorithm has been implemented in a program called Almost Invariant Sets (AIS). The C code for finding amino acid groupings and for calculating the conductance measure is available at

`http : //www.ebi.ac.uk/goldman/AIS`.

We now apply our code to standard amino acid replacement matrices as they are widely used in practice. We start with the PAM matrix (Dayhoff et al., 1978). The eigenvalues of the PAM1 matrix are given in Fig. 3. The spectrum of the PAM1 matrix (Fig. 3) does not exhibit a clearly identifiable cluster around the Perron root $\lambda_1 = 1$. Rather all 20 eigenvalues of the PAM1 matrix are close to 1. We decided firstly to calculate a grouping into five amino acid sets since we could compare this grouping to the one derived from the physicochemical properties of the amino acids by Taylor (1986) (see also French and Robson (1983), illustrated in Fig. 4 and summarized as follows:

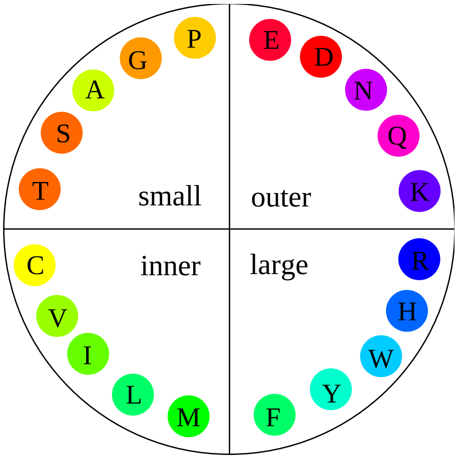

Fig. 4. Representation of the PAM matrix. This projection of the matrix by multidimensional scaling is an idealization adapted from French and Robson (1983) by Taylor (1986). The vertical axis of the circle corresponds to hydrophobicity, and consequently to whether the amino acid is mostly found in the inner or outer parts of proteins, and the horizontal axis corresponds to the molecular volume (small or large) of the amino acid. Amino acids that are close together exchange frequently. Colours used are those proposed by Taylor (1997).

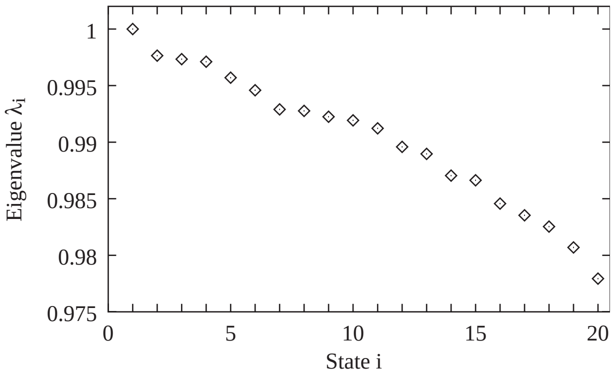

Fig. 3. Eigenvalues of the PAM1 matrix.

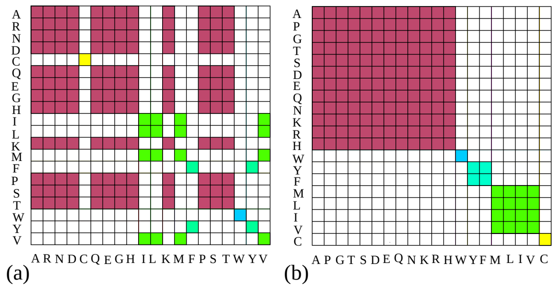

Fig. 5. Application of the AIS algorithm to PAM. Amino acid colours are combined for each group according to the scheme of Taylor (1997). (a) Hidden block Structure of the PAM matrix. (b) Sorted PAM matrix.

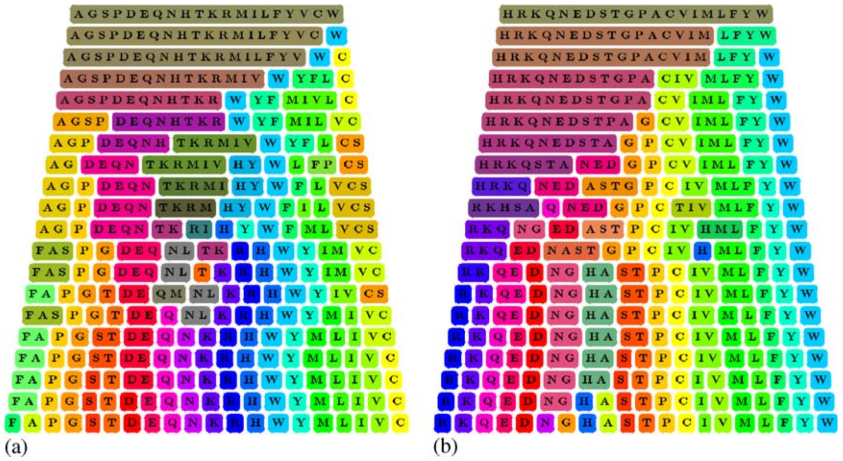
AGSPDEQNHTKRMILFYVCW
AGSPDEQNHTKRMILFYVC W
AGSPDEQNHTKRMILFYV W C
AGSPDEQNHTKRMIV W YFL C
AGSPDEQNHTKR W YF MIVL C
AGSP DEQNHTKR W YF MIL VC
AGP DEQNH TKRMIV W YF L CS
AG DEQN TKRMIV HY W L FP CS
AG P DEQN TKRMI HY W F L VCS
AG P DEQN TKRM HY W F I L VCS
AG P DEQN TK RI H Y W F ML VCS
FAS P G DEQ NL TK R H W Y IM VC
FAS P G DEQ NL T K R H W Y IM VC
FA P G T DE QM NL K R H W Y IV CS
FAS P G T DE Q NL K R H W Y M I VC
FA P G ST DE Q N K R H W Y M L I VC
FA P G ST DE Q N K R H W Y M L I V C
FA P G S T DE Q N K R H W Y M L I V C
FA P G S T D E Q N K R H W Y M L I V C
F A P G S T D E Q N K R H W Y M L I V C
(a)
HRKQNEDSTGPACVIMLFYW
HRKQNEDSTGPACVIM LFYW
HRKQNEDSTGPACVIM LFY W
HRKQNEDSTGPA CIV MLFY W
HRKQNEDSTGPA CV IML FY W
HRKQNEDSTPA G CV IML FY W
HRKQNEDSTA G P CV IML FY W
HRKQSTA NED G P CV IML FY W
HRKQ NED ASTG P C IV MLF Y W
RKHSA Q NED G P C TIV MLF Y W
RKQ NG ED AST P C IV HML F Y W
RKQ ED NAST G P C IV H ML F Y W
RK QE D NG HA ST P C IV ML F Y W
R K QE D NG HA ST P C IV ML F Y W
R K QE D NG HA ST P C IV M L F Y W
R K Q E D NG HA ST P C IV M L F Y W
R K Q E D NG HA S T P C IV M L F Y W
R K Q E D NG HA S T P C I V M L F Y W
R K Q E D NG H A S T P C I V M L F Y W
R K Q E D N G H A S T P C I V M L F Y W
(b)

Fig. 6. Complete results of the AIS algorithm for the PAM and WAG matrices. (a) The 20 best groupings according to the PAM matrix. (b) The 20 best groupings according to the WAG matrix.

Hydrophilic:	A P G T S D E Q N,
Basic:	K R H,
Aromatic:	W Y F,
Aliphatic:	M L I V,
Sulphydryl:	C.

The colours used in Figs. 4–6 follow the scheme proposed by Taylor (1997). The use of this scheme may itself lead to new insights into protein sequence evolution (Taylor, 1997), but this approach has not been further pursued here. The algorithm identified the five blocks for the PAM matrix shown in Fig. 5a. After reordering of the amino acids the inferred almost block diagonal structure of the PAM matrix is clearly visible. We read out the grouping from the ordered PAM matrix (Fig. 5b) as follows:

{A P G T S D E Q N K R H}{W}{Y F}{M L I V}{C}.

The groupings derived from the physicochemical properties and using our algorithm show similarities: only direct neighbours in the circle of amino acids of Fig. 4 constitute groups. The sets {M L I V} and {C} are identical. The hydrophilic group {A P G T S D E Q N} and the basic group {K R H} are merged by our algorithm into one set. Phenylalanine (F) and tyrosine (Y) remain members of the same set, but according to our algorithm tryptophan (W) is not a member of this aromatic group and forms its own group {W}. Tryptophan is known to show unique behaviour. To compare these groupings quantitatively we calculate the conductance measure for both:

$$\varphi_{AIS\ \ algorithm} = 1.306 < \varphi_{physicochem} = 1.738$$

and thus the grouping that was found by the algorithm outperforms the grouping suggested by physical and chemical properties of the amino acids.

Moving on from the division into five subsets, the best partitions between 1 and 20 subsets have been calculated and are given in Fig. 6a. Fig. 7 shows how the conductance measure φ increases with the number of sets. The conductance measure grows moderately for a grouping into $n = 1$–4 sets. The growth then changes to a rapid rise for divisions into $n = 5$–15 groups, slows down for $n = 16$–17 groups and finally grows rapidly again for $n = 18$–20. Overall the conductance measure increases strictly monotonically and no local extrema or plateaus can be observed.

We have also applied the AIS algorithm to the WAG matrix of Whelan and Goldman (2001). In Fig. 6b we present the partitions of between 1 and 20 subsets found by the AIS algorithm. The 20 best groupings of the PAM and the WAG matrix are clearly distinguishable. For example, the most conserved group of WAG is {L M F Y} (along with its subgroups {L M F} and {L M}). In contrast, the set {C S V} (and {C V}) is the most stable among the groupings of the PAM matrix.

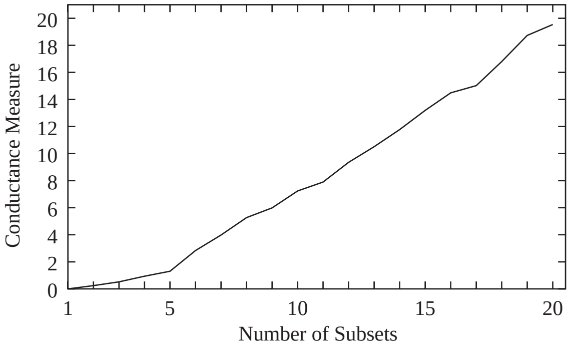

Fig. 7. The conductance measure for 20 best groupings according to the PAM matrix.

Generally in the case of the PAM matrix new sets evolve by splitting up the previous sets. Among the groupings according to the WAG matrix swaps between sets can frequently be observed in addition to simple splits.

8. Discussion and conclusion

The conductance measure and the grouping algorithm have been proven useful in finding disjoint sets of amino acids within which residues are more closely related. However, the criterion and the method only enable us to find the best grouping for a particular given number of subsets. No decision on the best number of subsets can be made, since neither the clustering of the eigenvalues around the Perron root $\lambda_1 = 1$ nor the graph of conductance measure as function of the number of subsets allow us to choose in a sensible way. To make progress, here it might be necessary to modify the definition of the conductance measure.

The groupings found for a particular number of subsets are also reasonable from a biochemical point of view as the comparison with the grouping of Taylor (1986) into five subsets shows. The advantage of our approach is that the algorithm automates the process of finding groupings and that the conductance allows a quantitative assessment of the partition in a biologically meaningful way. The grouping algorithm identifies sets of amino acids with a high probability of mutation between amino acids of the same set but small probabilities of change between different sets. The conductance measure quantifies the evolutionary changes between subsets that are of most interest. Furthermore, if the analysis is based on the normalized rate matrix of a Markov model, it is possible to directly compare the results of different models.

The analysis of the WAG matrix and the PAM matrix indicates that different amino acid replacement matrices lead like fingerprints to different groupings. In the future we will therefore use the groupings and their score according to the conductance measure as a tool to analyse and compare various Markov models of protein sequence evolution. We will also apply our method to larger 61 × 61 rate matrices of codon models (see, e.g. Goldman and Yang, 1994; Yang et al., 2000).

Acknowledgements

C.K. and N.B. were partially supported by PRTL Ireland, IITAC - Bioinformatics. C.K. is a member of Wolfson College, Cambridge University. N.G. is supported by a Wellcome Trust Senior Fellowship. It is a pleasure to thank J.C. Sexton for the helpful discussions on stochastic optimization methods. We are grateful to S. Whelan for providing us with amino acid replacement matrices.

References

Adachi, J., Hasegawa, M., 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. 42, 459–468.

Adachi, J., Waddell, W., Martin, M., Hasegawa, M., 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. 50, 348–358.

Behrends, E., 2000. Introduction to Markov Chains with Special Emphasis on Rapid Mixing. Vieweg, Wiesbaden.

Cannata, N., Toppo, S., Romaldi, C., Valle, G., 2002. Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices. Bioinformatics 18, 1102–1108.

Coghlan, A., Mac Dónaill, D.A., Buttimore, N.H., 2001. Representation of amino acids as five-bit or three-bit patterns for filtering protein databases. Bioinformatics 17, 676–685.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. A model of evolutionary change in proteins. In: Dayhoff, M.O. (Eds.), Atlas of Protein Sequence and Structure, Vol. 5, suppl. 3. National Biomedical Research Foundation, Washington, DC, pp. 345–352.

Deuflhard, P., Huisinga, W., Fischer, A., Schütte, C., 2000. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. Linear Algebra Appl. 315, 39–59 Doi:10.1016/S0024-3795(00)00095-1.

French, S., Robson, B., 1983. What is a conservative substitution? J. Mol. Evol. 19, 171–175.

Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11, 725–736.

Goldman, N., Thorne, J.L., Jones, D.T., 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149, 445–458.

Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. Science 185, 862–864.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Comp. Appl. Biosci. 8, 275–282.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1994. A mutation data matrix for transmembrane proteins. FEBS Lett. 339, 269–275 Doi:10.1016/0014-5793(94)80429-X.

Liò, P., Goldman, N., 1998. Models of molecular evolution and phylogeny. Genome Res. 8, 1233–1244.

Overington, J., Johnson, M.S., Šali, A., Blundell, T.L., 1990. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc. R. Soc. London B 241, 132–145.

Sinclair, A., 1992. Improved bounds for mixing rates of Markov chains and multicommodity flow. Combin. Probab. Comp. 1, 351–370.

Taylor, W.R., 1986. The classification of amino acid conservation. J. Theor. Biol. 119, 205–218.

Taylor, W.R., 1997. Residual colours: a proposal for aminochromography. Prot. Eng. 10, 743–746.

Trick, M.,1994. http://mat.gsia.cmu.edu/COLOR/color.html.

Wang, J., Wang, W., 1999. A computational approach to simplifying the protein folding alphabet. Nature Struct. Biol. 6, 1033–1038 Doi:10.1038/14918.

Weisstein, E.W., 1999. http://mathworld.wolfram.com/BellNumber.html.

Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18, 691–699.

Xia, X., Li, W., 1998. What amino acid properties affect protein evolution? J. Mol. Evol. 47, 557–564.

Yang, Z., Nielsen, R., Hasegawa, M., 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15, 1600–1611.

Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.-M.K., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.