A New Criterion and Method for Amino Acid Classification

Carolin Kosiol[1,*], Nick Goldman[2] and Nigel H. Buttimore[1]

[1]School of Mathematics, University of Dublin, Trinity College, Dublin 2, Ireland

[2]EMBL - EBI, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK

April 30, 2003

ABSTRACT

Motivation: It is known that many evolutionary changes of amino acid sequence in proteins are conservative: a substitution of one amino acid by another residue has a far greater chance of being accepted if the two residues have similar properties. Yet it is difficult to identify the relevant physicochemical properties when classifying amino acids. In this paper we introduce a criterion and a method for finding groups of amino acids, which determine similarity from an evolutionary point of view.

Results: We present a criterion that assesses the quality of an amino acid grouping. 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 introduced to reflect 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. We show that 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. We present the groupings resulting from two standard matrices used in sequence alignment and phylogenetic tree estimation.

Availability: The C code for calculating the conductance measure and for finding amino acid groupings is available at http://www.ebi.ac.uk/goldman-srv/AIS.

Contact: kosiol@ebi.ac.uk

INTRODUCTION

Proteins are made up of different amino acids, which are denoted by 20 different letters of the alphabet. For particular tasks, however, it can be useful to simplify this alphabet and group one or more letters together. For instance we could split the 20 amino acids into two groups: a hydrophilic and a hydrophobic set.

Groupings have been applied to various fields. Wang and Wang [14] use sets of amino acids in protein design and modeling. Brazma *et al.* [2] have pointed out that the search for patterns occurring at an unexpected rate is an established strategy for the identification of functional constraints in protein sequences. Some amino acids can often be replaced without any significant functional alteration. A possible approach for seeking unusal degenerate patterns could be to rewrite the sequence using a simplified alphabet and search all the simplified sequences for unexpected patterns. Along those lines Coghlan *et al.* [4] use groupings of amino acids to develop filtering algorithms for protein databases.

*Now at EMBL – EBI, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK.

Several methods have been proposed to classify amino acids. Grantham [9] introduces an amino acid distance formula that considers the chemical composition of the side chain, the polarity and the molecular volume to help explain protein evolution. This approach has been extended by Xia and Li [17] in a study of the relevance of 10 amino acid properties affecting protein evolution. Grantham and Xia and Li present their results in the form of distance matrices, whereas French and Robson [7] arrange their results in two-dimensional diagrams using multidimensional scaling. Taylor [12] also adopts this graphical approach and develops Venn diagrams of amino acids sets. The unions and intersections of the Venn diagram allow determination of 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.* [3] 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 approach leads to different groupings for different time periods of the same matrix (e.g., PAM120 and PAM250). Second, the classification criterion used has no clear evolutionary meaning.

Our interest in groupings is motivated by the study of Markov models for protein sequence evolution. A grouping of 20 amino acids is much more comprehensible than the tabular representation of a rate matrix. We hope to use groupings in further studies as a tool to analyse and compare different models. It is therefore crucial that the classification criterion is biologically meaningful. In this paper we develop a criterion and grouping method to identify sets of amino acids with a high probability of change between elements of the set but small probabilities of change between different sets. The method we use has its origin in the convergence diagnosis of Markov chain Monte Carlo (MCMC) [1]. In section 1 we give a general introduction to the Markov models used to describe amino acid evolution. In section 2 we introduce the conductance, a measure for the grouping of amino acids into non-empty sets, whose value will indicate the quality of the division. Unfortunately the measure itself does not suggest a method that would determine optimal groupings. In section 3 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 4. We apply the conductance measure and the grouping algorithm to standard amino acid replacement matrices in section 5 and finally discuss our results in section 6.

SYSTEM AND METHODS

1 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 $1, 2, \ldots, 20$. We can write the $P_{ij}(t)$ as a 20×20 matrix, which we denote by $P(t)$. The matrix

$P(t)$ is a Markov matrix, and so the rows sum to 1. It can be represented as

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

where the matrix Q is known as the instantaneous rate matrix and has its off-diagonal entries Q_{ij} equal to the rates of replacement of i by j. The diagonal entries Q_{ij} are defined by the mathematical requirement that each row sums to zero (see Liò and Goldman [10]).

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

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

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

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

The vector π is also the limiting distribution when time approaches infinity. Reversibility means that

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

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

2 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 the 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 [5]. 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, 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 [11]). Below, we redefine the conductance in terms of the instantaneous rate matrix Q instead of the Markov 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 [10].)

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 amino acids $A = \{1, \ldots, 20\}$, where $A_k \cap A_l = \emptyset$ for $k, l = 1, \ldots, K$ and $\bigcup_k A_k = A$. If π_i is the ith 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 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 A_k to 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* of A_k, is then

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

The ratio

$$\Phi_{kl} = \frac{F_{kl}}{C_k}$$

is called the *conductance* [1]. This is the expected number of changes from A_k to A_l per site per unit time when commencing at subset A_k.

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

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

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, $k \neq l$. Subsequently Φ would be a null matrix. 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 $A_1, \ldots, A_K$.

Example 1

To set ideas, we consider a simple illustrative system of 7 amino acids with rate matrix Q having the following block diagonal form:

$$\begin{pmatrix} -.35 & .35 & 0 & 0 & 0 & 0 & 0 \\ .35 & -.35 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2.1 & 2.1 & 0 & 0 & 0 \\ 0 & 0 & 2.1 & -2.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -.7 & .35 & .35 \\ 0 & 0 & 0 & 0 & .35 & -.7 & .35 \\ 0 & 0 & 0 & 0 & .35 & .35 & -.7 \end{pmatrix} \quad .$$

The block diagonal structure of the rate matrix suggests the partition into $A_1 = \{1, 2\}$, $A_2 = \{3\}$ and $A_3 = \{4, 5, 6, 7\}$. 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:

$$\begin{aligned} F_{A_1 \to A_2} &= F_{12} = F_{21} = 0 \\ F_{A_1 \to A_3} &= F_{13} = F_{31} = 0 \\ F_{A_2 \to A_3} &= F_{23} = F_{23} = 0. \end{aligned}$$

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

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

As *Example 1* shows, 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 nth Bell number, B_n [15]. 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:

$$\begin{aligned} B_1 &= \binom{0}{0} B_0 = 1 \\ B_2 &= \binom{1}{0} B_0 + \binom{1}{1} B_1 = 2 \\ B_3 &= \binom{2}{0} B_0 + \binom{2}{1} B_1 + \binom{2}{2} B_2 \\ &= 1 + 2 + 2 = 5 \\ B_4 &= \binom{3}{0} B_0 + \binom{3}{1} B_1 + \binom{3}{2} B_2 + \binom{3}{3} B_3 \\ &= 1 + 3 + 6 + 5 = 15 \end{aligned}$$

The 15 possible partitions into non-empty subsets are

$\{ATGC\}$, $\{AT\}\{GC\}$, $\{A\}\{TC\}\{G\}$,

$\{A\}\{TGC\}$, $\{AC\}\{GT\}$, $\{T\}\{AG\}\{C\}$,

$\{ATG\}\{C\}$, $\{AG\}\{TC\}$, $\{G\}\{AT\}\{C\}$,

$\{AGC\}\{T\}$, $\{A\}\{GT\}\{C\}$, $\{G\}\{AC\}\{T\}$,

$\{ATC\}\{G\}$, $\{A\}\{GC\}\{T\}$, $\{A\}\{G\}\{C\}\{T\}$.

Cannata *et al.* [3] 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 ($K = 1, \ldots, 20$) of sets. For example, under the restriction only to admit partitions into exactly 8 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 the next two sections 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 $(2, \ldots, 19)$ 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.

3 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 will 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 $P(t) = e^{tQ}$ are different, but the eigenvectors remain the same.

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

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_{20}| \quad .$$

By the Frobenius-Perron theorem 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 $e = (1, \ldots, 1)^T$. The corresponding left eigenvector $\pi = (\pi_1, \ldots, \pi_{20})^T$ represents the equilibrium dis-

tribution under the assumption that it is normalized so that $\pi^T e = 1$. In matrix notation we have:

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

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

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

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

$$(0, \ldots, 0, e_k^T, 0, \ldots, 0)^T, \quad k = 1, \ldots, K,$$

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 $B = P(t) = e^{Qt}$, where Q is the block diagonal rate matrix of *Example 1* and $t = 1$:

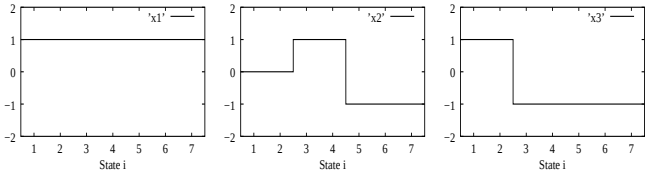
$$P(1) = \begin{pmatrix} .75 & .25 & 0 & 0 & 0 & 0 & 0 \\ .25 & .75 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .51 & .49 & 0 & 0 & 0 \\ 0 & 0 & .49 & .51 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .56 & .22 & .22 \\ 0 & 0 & 0 & 0 & .22 & .56 & .22 \\ 0 & 0 & 0 & 0 & .22 & .22 & .56 \end{pmatrix} .$$

The eigenvalues of $P(1)$ are

$$\lambda_1 = 1 \quad \lambda_2 = 1 \quad \lambda_3 = 1$$
$$\lambda_4 = 0.5 \quad \lambda_5 = 0.34 \quad \lambda_6 = 0.34 \quad \lambda_7 = 0.02 .$$

The right eigenvectors corresponding to $\lambda = 1$ are

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



(a) (b) (c)

Figure 1: The eigenvectors x_1, x_2, x_3 of *Example 3*, corresponding to $\lambda = 1$.

Figure 1 shows the eigenvectors x_1, x_2 and x_3 corresponding to $\lambda = 1$ as function of the seven states 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:

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

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

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

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

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.* [6]. 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.

4 Perturbation Theory

The standard amino acid replacement matrices like PAM [5] and WAG [16] 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 the probability matrix 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 below:

$$\begin{pmatrix} .01 & .09 & .10 & .25 & .08 & .30 & .17 \\ .09 & .10 & .25 & .08 & .30 & .17 & .01 \\ .10 & .25 & .08 & .30 & .17 & .01 & .09 \\ .25 & .08 & .30 & .17 & .01 & .09 & .10 \\ .08 & .30 & .17 & .01 & .09 & .10 & .25 \\ .30 & .17 & .01 & .09 & .10 & .25 & .08 \\ .17 & .01 & .09 & .10 & .25 & .08 & .30 \end{pmatrix}.$$

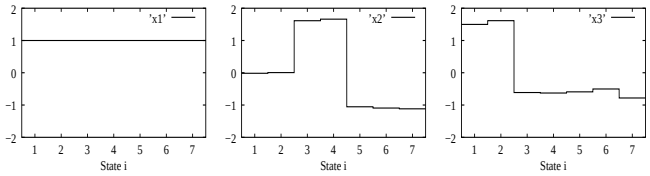
The eigenvalues of P are now calculated as

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

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:

$$\begin{array}{rrrrrrr} (1, & 1, & 1, & 1, & 1, & 1, & 1) \\ (-0.02, & 0.01, & 1.61, & 1.66, & -1.05, & -1.09, & -1.12) \\ (1.50, & 1.61, & -0.61, & -0.63, & -0.59, & -0.50, & -0.78) \end{array} .$$

Figure 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



(a) (b) (c)

Figure 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.

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 $\{1, 2\}$, $\{3, 4\}$ or $\{5, 6, 7\}$. *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 in Deuflhard *et al.* [6].

ALGORITHM

This section transforms the results of sections 3 and 4 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

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

Only those states in $\mathcal{S}$ can be assigned to groups using the following procedure; states $s \notin \mathcal{S}$ are unclassifable. *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 $x_k(s)$. For the amino acid replacement matrices studied we have chosen $\delta = 0.5$. 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 case never arose in our examples.

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

$$\begin{aligned} \sigma_{(s,\theta)} &= (\sigma_1, \ldots, \sigma_K) \\ \text{with } \sigma(k) &= \begin{cases} \text{sign}(\tilde{x}_k(s)) & \text{if } |\tilde{x}_k(s)| > \theta, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

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. The colouring of graphs is a standard problem. An introduction to graph coloring and code that performs this task can be found at Michael Trick's webpage [13].

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 $\tilde{\theta} = \frac{\theta^- + \theta^+}{2}$ (bisection search to find θ 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 $(\mathcal{K}(\tilde{\theta}) > K)$ $\theta^+ = \tilde{\theta}$ and **goto Step2**

else $\theta^- = \tilde{\theta}$ and **goto Step2**

IMPLEMENTATION

The above algorithm has been implemented in a C-program called Almost Invariant Sets (AIS). We now apply our code to standard amino acid replacement matrices as they are widely used in practice. We start with the PAM1 matrix [5]. The eigenvalues of the PAM1 matrix are given in the Figure 3.

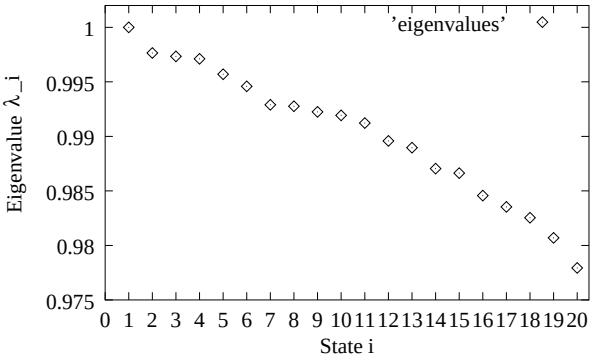


Figure 3: Eigenvalues of the PAM1 matrix

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 four amino acid sets since we could compare this grouping to a grouping according to physicochemical properties of the amino acids [7], [12], illustrated in Figure 4.

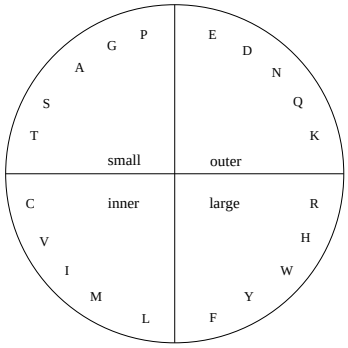
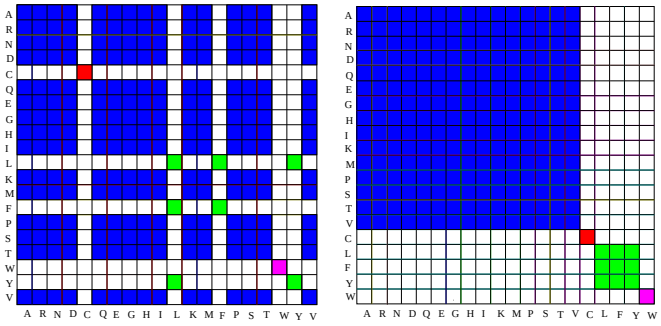


Figure 4: Representation of the PAM matrix. This projection of the matrix by multidimensional scaling is an idealization adapted from Robson and French [7] by Taylor [12]. Amino acids that are close together exchange frequently. The diagram divides the 20 amino acids into four sets of equal size.

The algorithm identified four blocks for the PAM matrix, as shown in Figure 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:

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

Our algorithm divides the residues into four sets of unequal size. There is little overlap of the grouping according to the physicochemical properties and the grouping according to our algorithm. However leucine, phenylalanine and tyrosine { L F Y } are direct neighbours in Figure 4 and cysteine { C } and Tryptophan { W } are known to show unique behaviour. To compare these groupings quantitatively we calculate the conductance measure for



(a) Hidden Block Structure of the PAM matrix.

(b) Sorted PAM matrix.

Figure 5: Application of the AIS algorithm to PAM.

both:

$$\varphi_{\text{AIS algorithm}} = 0.937 < \varphi_{\text{physicochem}} = 1.814$$

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 four subsets, the best partitions between 1 and 20 subsets have been calculated and are given in Figure 6.

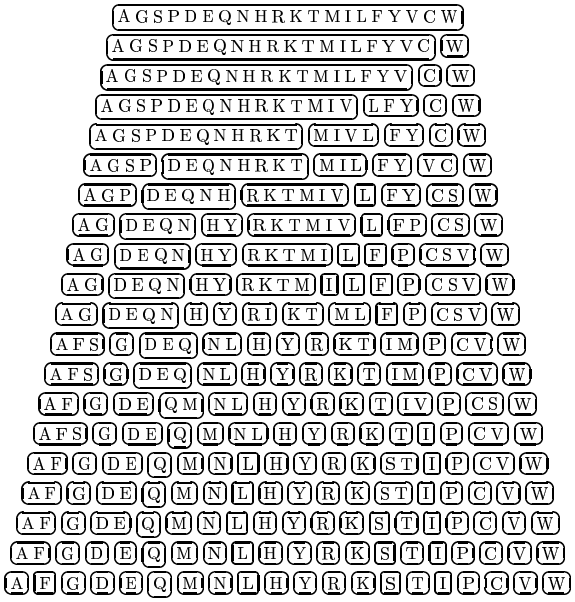


Figure 6: The 20 best groupings according to the PAM matrix.

Figure 7 shows how the conductance measure φ increases with the number of sets. The conductance mea-

sure 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.

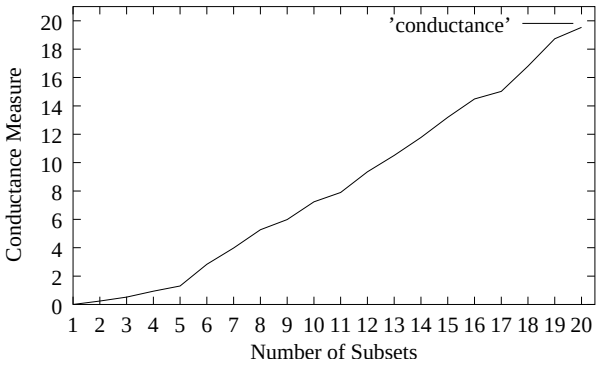


Figure 7: The Conductance measure for 20 best groupings according to the PAM matrix.

We have also applied the AIS algorithm to the WAG matrix of Whelan and Goldman [16]. In Figure 8 we present the partitions of between 1–20 subsets found by the AIS algorithm.

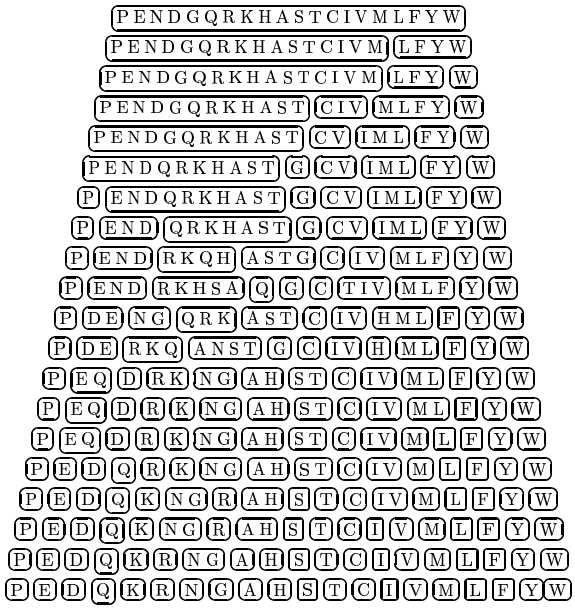


Figure 8: The 20 best groupings according to the WAG matrix.

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. 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.

DISCUSSION

The conductance measure and the grouping algorithm have been proven useful in finding sets of amino acids. 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 [12] into four 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 ma-

trix indicates that different amino acid replacement matrices lead like fingerprints to different groupings. In future studies we will therefore use the groupings and their score according 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., [8], [18]).

Acknowledgements

C. K. was partially supported by PRTL Ireland, IITAC - Bioinformatics. It is a pleasure to thank J.C. Sexton for helpful discussions on stochastic optimization methods. We are grateful to S. Whelan for providing us with amino replacement matrices.

References

[1] Behrends,E. (2000) *Introduction to Markov Chains with Special Emphasis on Rapid Mixing.* Vieweg, Wiesbaden.

[2] Brazma,A., Jonassen,I., Eidhammer,I. and Gilbert,D. (1998) Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol.*, **5**, 279–305.

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

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

[5] Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure, Vol. 5, supp. 3.* National Biomedical Research Foundation, Washington, D.C.

[6] Deuflhard,P., Huisinga,W., Fischer,A. and Schütte,C. (2000) Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Appl.*, **315**, 39–59.

[7] French,S. and Robson,B. (1983) What is a conservative substitution? *J. Mol. Evol.*, **19**, 171–175.

[8] Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, **11**, 725–736.

[9] Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.

[10] Liò,P. and Goldman,N. (1998) Models of molecular evolution and phylogeny. *Genome Res.*, **8**, 1233–1244.

[11] Sinclair,A. (1992) Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combinatorics, Probab., Comp.*, **1**, 351–370.

[12] Taylor,W.R. (1986) The classification of amino acid conservation. *J. Theoret. Biol.*, **119**, 205–218.

[13] Trick,M. (1994) `http://mat.gsia.cmu.edu/COLOR/color.html`

[14] Wang,J. and Wang,W. (1999) A computational approach to simplifying the protein folding alphabet. *Nat. Struct. Biol.*, **6**, 1033–1038.

[15] Weisstein,E.W. (1999) `http://mathworld.wolfram.com/BellNumber.html`

[16] Whelan,S. and 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.

[17] Xia,X. and Li,W. (1998) What amino acid properties affect protein evolution? *J. Mol. Evol.*, **47**, 557–564.

[18] Yang,Z., Nielsen,R., Goldman,N. and Pedersen,A.-M.K. (2000) Codon-substitution models for heterogeneous selectionpressure at amino acid sites. *Genetics*, **155**, 431–449.