Markov Chain Properties of Amino Acid Replacement Matrices

Carolin Kosiol
School of Mathematics
Trinity College Dublin
kosiol@maths.tcd.ie

26th April 2002

A Markov Chain for Amino Acids

- Although the PAM (Point Accepted Mutation), the JTT (Jones, Taylor, Thorton) and the WAG (Whelan And Goldman) amino acid replacement matrices are based on different models of molecular evolution they all share an assumption, the Markov Chain property.

- 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. The probabilities of transition can be written as a matrix.

- Since there are 20 amino acids, the Markov matrix is a 20×20 matrix, which we denote by M. An element M_{ij} gives the probability that the amino acid j will be replaced by the amino acid i in an evolutionary interval.

The Point Accepted Mutation (PAM) Matrix

$$\frac{1}{10000}\left(\begin{array}{cccccccccccccccccccccc}
 & \mathbf{A} & \mathbf{R} & \mathbf{N} & \mathbf{D} & \mathbf{C} & \mathbf{Q} & \mathbf{E} & \mathbf{G} & \mathbf{H} & \mathbf{I} & \mathbf{L} & \mathbf{K} & \mathbf{M} & \mathbf{F} & \mathbf{P} & \mathbf{S} & \mathbf{T} & \mathbf{W} & \mathbf{Y} & \mathbf{V} & \\
\mathbf{A} & 9867 & 2 & 9 & 10 & 3 & 8 & 17 & 21 & 2 & 6 & 4 & 2 & 6 & 2 & 22 & 35 & 32 & 0 & 2 & 18 & \mathbf{Ala} \\
\mathbf{R} & 1 & 9914 & 1 & 0 & 1 & 10 & 0 & 0 & 10 & 3 & 1 & 19 & 4 & 1 & 4 & 6 & 1 & 8 & 0 & 1 & \mathbf{Arg} \\
\mathbf{N} & 4 & 1 & 9822 & 36 & 0 & 4 & 6 & 6 & 21 & 3 & 1 & 13 & 0 & 1 & 2 & 20 & 9 & 1 & 4 & 1 & \mathbf{Asn} \\
\mathbf{D} & 6 & 0 & 42 & 9859 & 0 & 6 & 53 & 6 & 4 & 1 & 0 & 3 & 0 & 0 & 1 & 5 & 3 & 0 & 0 & 1 & \mathbf{Asp} \\
\mathbf{C} & 1 & 1 & 0 & 0 & 9973 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 5 & 1 & 0 & 3 & 2 & \mathbf{Cys} \\
\mathbf{Q} & 3 & 9 & 4 & 5 & 0 & 9876 & 27 & 1 & 23 & 1 & 3 & 6 & 4 & 0 & 6 & 2 & 2 & 0 & 0 & 1 & \mathbf{Gln} \\
\mathbf{E} & 10 & 0 & 7 & 56 & 0 & 35 & 9865 & 4 & 2 & 3 & 1 & 4 & 1 & 0 & 3 & 4 & 2 & 0 & 1 & 2 & \mathbf{Glu} \\
\mathbf{G} & 21 & 1 & 12 & 11 & 1 & 3 & 7 & 9935 & 1 & 0 & 1 & 2 & 1 & 1 & 3 & 21 & 3 & 0 & 0 & 5 & \mathbf{Gly} \\
\mathbf{H} & 1 & 8 & 18 & 3 & 1 & 20 & 1 & 0 & 9913 & 0 & 1 & 1 & 0 & 2 & 3 & 1 & 1 & 1 & 4 & 1 & \mathbf{His} \\
\mathbf{I} & 2 & 2 & 3 & 1 & 2 & 1 & 2 & 0 & 0 & 9871 & 9 & 2 & 12 & 7 & 0 & 1 & 7 & 0 & 1 & 33 & \mathbf{Ile} \\
\mathbf{L} & 3 & 1 & 3 & 0 & 0 & 6 & 1 & 1 & 4 & 22 & 9947 & 2 & 45 & 13 & 3 & 1 & 3 & 4 & 2 & 15 & \mathbf{Leu} \\
\mathbf{K} & 2 & 37 & 25 & 6 & 0 & 12 & 7 & 2 & 2 & 4 & 1 & 9924 & 20 & 0 & 3 & 8 & 11 & 0 & 1 & 1 & \mathbf{Lys} \\
\mathbf{M} & 1 & 1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 5 & 8 & 4 & 9875 & 1 & 0 & 1 & 2 & 0 & 0 & 4 & \mathbf{Met} \\
\mathbf{F} & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 2 & 8 & 6 & 0 & 4 & 9944 & 0 & 2 & 1 & 3 & 28 & 0 & \mathbf{Phe} \\
\mathbf{P} & 13 & 5 & 2 & 1 & 1 & 8 & 3 & 2 & 5 & 1 & 2 & 2 & 1 & 1 & 9924 & 12 & 4 & 0 & 0 & 2 & \mathbf{Pro} \\
\mathbf{S} & 28 & 11 & 34 & 7 & 11 & 4 & 6 & 16 & 2 & 2 & 1 & 7 & 4 & 3 & 17 & 9840 & 38 & 5 & 2 & 2 & \mathbf{Ser} \\
\mathbf{T} & 22 & 2 & 13 & 4 & 1 & 3 & 2 & 2 & 1 & 11 & 2 & 8 & 6 & 1 & 5 & 32 & 9869 & 0 & 2 & 9 & \mathbf{Thr} \\
\mathbf{W} & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 9976 & 1 & 0 & \mathbf{Trp} \\
\mathbf{Y} & 1 & 0 & 3 & 0 & 3 & 0 & 1 & 0 & 4 & 1 & 1 & 0 & 0 & 21 & 0 & 1 & 1 & 2 & 9947 & 1 & \mathbf{Tyr} \\
\mathbf{V} & 13 & 2 & 1 & 1 & 3 & 2 & 2 & 3 & 3 & 57 & 11 & 1 & 17 & 1 & 3 & 2 & 10 & 0 & 2 & 9901 & \mathbf{Val} \\
 & \mathbf{Ala} & \mathbf{Arg} & \mathbf{Asn} & \mathbf{Asp} & \mathbf{Cys} & \mathbf{Gln} & \mathbf{Glu} & \mathbf{Gly} & \mathbf{His} & \mathbf{Ile} & \mathbf{Leu} & \mathbf{Lys} & \mathbf{Met} & \mathbf{Phe} & \mathbf{Pro} & \mathbf{Ser} & \mathbf{Thr} & \mathbf{Trp} & \mathbf{Tyr} & \mathbf{Val} &
\end{array}\right)$$

Capacity and Ergodic Flow

Let $M = (M_{ij})_{i,j=1,\ldots,N}$ be an aperiodic and irreducible Markov matrix with equilibrium distribution $(p_1, \ldots, p_N)'$.

Consider proper subsets $T_1, \ldots, T_K$ of the state space $S = \{1, \ldots, N\}$, where $T_k \cap T_l = \emptyset$ for $k, l = 1, \ldots, K$ and $\bigcup_k T_k = S$.

We define the **capacity of** T_l

$$C_l := \sum_{j \in T_l} p_j$$

We also define the **ergodic flow from subset** T_l **to** T_k

$$F_{kl} = \sum_{i \in T_k,\, j \in T_l} M_{ij} p_j \quad .$$

The **conductance** is given by the ratio

$$\Phi_{kl} := \frac{F_{kl}}{C_l} \quad .$$

The conductance is a conditional probability related to the following: how likely is it that the Markov chain will leave the subset T_l to go to T_k in the next step under the condition that the chain is in fact in subset T_l now?

The Conductance Measure

We can define a new matrix $\Phi = (\Phi_{kl})_{k,\,l=1,\dots,K}$

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

The matrix Φ is itself a Markov matrix. To calculate the difference between Φ and the identity matrix we introduce the measure

$$\varphi = \frac{1}{N^2} \sum_k \sum_l (\Phi_{kl} - \delta_{kl})^2 \quad .$$

This measure can be used to control the quality of the decomposition of the state space into subsets.

But how can we find a good decomposition?

The Grouping Algorithm

- Given the system as a whole, we seek to identify the almost invariant subsets together with the (small) probabilities of transitions between them.

- Mathematically speaking, we are looking for a structure of the Markov matrix which is almost of block diagonal type.

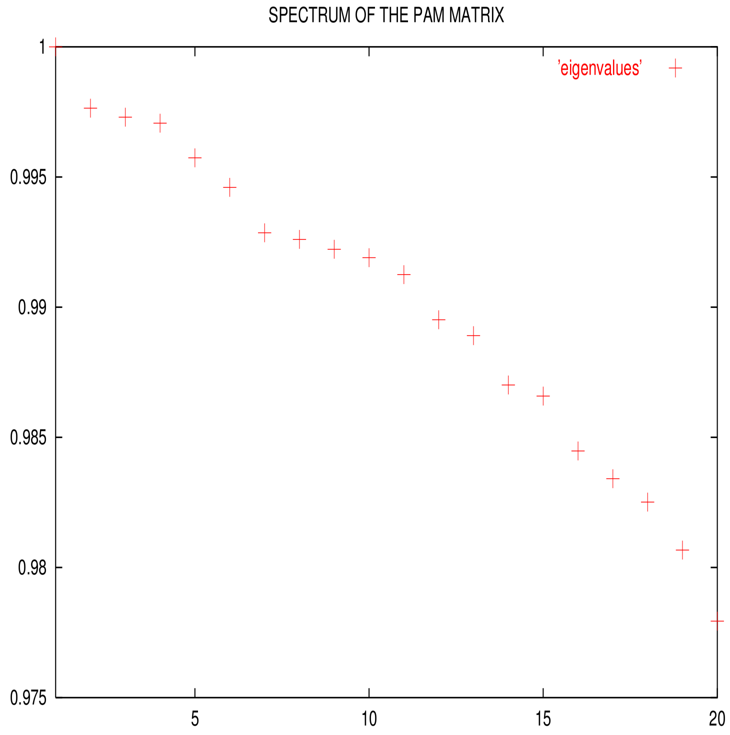
- The block structure depends on the spectrum (i.e. the eigenvalues and eigenvectors) of the Markov matrix. We use an algorithm developed by Deuflhardt et al. [3] to find the blocks:

  - Determine the number of blocks k.

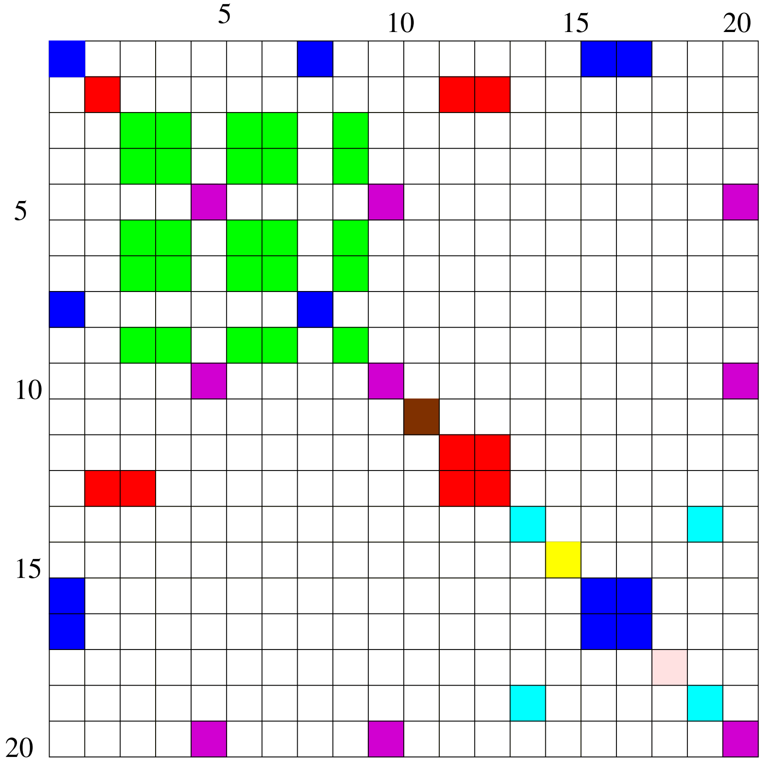
  - Derive a discriminating sign structure for k of the eigenvectors.

  - Transform the problem to a graph colouring algorithm.

Identification of 8 Almost Invariant Subsets of the PAM Matrix



(a) All eigenvalues are positive and close to 1.



(b) Hidden block structure of the PAM matrix.

8 Sets of Amino Acids for the PAM Matrix

According to the grouping algorithm:

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

According to chemical/physical properties [1]:

Small, Hydrophilic				Large, Hydrophobic			
Small, Aliphatic		Amide, Acidic		Reactive, Branched		Generally large	
Small	Smallest	Carbonyl	Hydroxyl	Sulfhyd	Aliphatic	Aromatic	Basic
P	A G	N D E Q	ST	C	I L M V	H F Y W	R K

Performance under the Conductance measure: $\varphi_{algo} = 4.7 \cdot 10^{-6} < \varphi_{chem} = 7 \cdot 10^{-6}$

A Blind Test: 10 anonymous matrices provided by the Biological Sequence Analysis Group, Department of Zoology, Cambridge.

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

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

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

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

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

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

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

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

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

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

Summary and Future Work

- The Conductance measure and the grouping algorithm are useful to find sets of amino acids for filter algorithms.
- Mutation matrices other than the PAM matrix, particularly WAG matrices, are now under investigation.
- All Mutation matrices are based on experimental data and contain measurement errors. This leads to the more general question: How does the grouping of the amino acids change if the mutation matrix changes?
- Finally we hope to improve the mutation and scoring matrices of alignment programs on the basis of our analysis.

References

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

[2] M. O. Dayhoff, ed. *Atlas of Protein Sequence and Structure,* Vol. 5, supp. 3, National Biomedical Research Foundation, Washington, D.C., 1978.

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