ELSEVIER

Biophysical Chemistry 115 (2005) 177–180

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

Classification of amino acids induced by their associated matrices

J.G. Esteve, F. Falceto*

Department of Theoretical Physics, University of Zaragoza, 50009 Zaragoza, Spain

Received 25 June 2004; received in revised form 9 November 2004; accepted 10 December 2004
Available online 15 January 2005

Abstract

In this paper we carry out an analysis of different types of potential and substitution matrices for amino acids, oriented to give a classification of the latter. The cluster decomposition is obtained, in a fully unsupervised way, from the subdominant ultrametric associated to the distance between amino acids induced by the corresponding matrix. In the comparative study, by looking at the classifications obtained from diverse matrices, we can get information on how they account for the different chemical–physical properties of the amino acids.

Keywords: Amino acids; Classification; Potential and substitution matrices

1. Introduction

This work is an application of the methods developed in Ref. [1] to obtain an unsupervised classification of amino acids based on potential and substitution matrices. The classification will group together those amino acids that have a similar behavior (or that are "more or less equivalent") with respect to such matrices. Consequently, looking at the composition of the clusters, we can learn which properties of the amino acids are accounted by the matrices.

In general, a classification is an ordering of objects into groups on the basis of their relationships, similarities or dissimilarities. The general problem of classification starts with N objects each one characterized by D coordinates $X_k^{(i)}$(i=1, N; k=1, D), which measure the values of some properties for every object. Using this coordinates we define a distance between objects and the problem is to merge them into groups in a consistent way, i.e. we form clusters with those objects which are in some sense "neighbors".

A situation well suited for this program is when the distance between objects has the property of ultrametricity, since then there is a natural and easy way of doing the classification, the so-called hierarchical classification which has been widely used in biology. Conversely, a classification, or equivalently the cluster decomposition, is equivalent to defining an ultrametric (i.e. a metric in which the distance between any two elements is not greater than the maximum of the distances of any of them to a third one). One can see that given an ultrametric and a radius r the set of closed balls of radius r is a partition of the full space into disjoint sets, i.e. a complete classification. The cluster decomposition, of course, is finer and finer as the radius diminishes.

Given a potential or substitution matrix we consider the rows in the (symmetric) matrix as the coordinates for the corresponding amino acid in a 20-dimensional space. This provides a distance between amino acids and from this we produce its uniquely determined classification. The problem is that given the distance in the 20-dimensional space it is not in general an ultrametric, however there is a canonical way of producing such an ultrametric by simply considering the largest ultrametric in which the distance between any two elements is smaller than in the original metric. This is always defined and produces the desired classification.

All the ambiguity is then the choice of the initial distance in the 20-dimensional space. We consider the following family of distances

$$d_p(X^{(i)},X^{(j)}) = \left(\sum_{k=1}^{20} (X_k^{(i)} - X_k^{(j)})^p \right)^{1/p}, \quad p \geq 1$$

* Corresponding author. Tel.: +34 976 761273; fax: +34 976 761264.
E-mail addresses: esteve@unizar.es (J.G. Esteve), falceto@unizar.es (F. Falceto).

0301-4622/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2004.12.023

and we take the classification from the value of p that produces an ultrametric which is closer to the corresponding metric, where the distance between the metric and their associated ultrametric is defined as [1]

$$d_u = 1 - \delta/d$$

being d and δ the sum over all pair of amino acids of their distances and ultrametric distances, respectively, so by definition d_u varies between 0 (that corresponds to an original metric which is ultrametric), and (N-2)/(N+1) (6/7 in our case).

In the following we consider two families of matrices of very different origin. A first family is based on the so-called statistical potential that measures the probability of having a given pair of amino acids close to each other in the native state of a certain set of proteins. To this group corresponds the Miyazawa–Jernigan matrix and some variants of it. The second family is based on the probability of substitution (or better to the closeness of the common ancestor) for changes of amino acids of similar proteins from different species. We will show that the classification obtained from the two families has different peculiarities that can be explained reasonably well by taking into account the different origin of the two families of matrices.

In this paper we focus on the relation between the classification obtained from different matrices and the properties (genetic, chemical, ...) of the amino acids. Another possible application of the classifications obtained below is the construction of simplified alphabets (see Refs. [2–5] for related works).

2. Potential matrices

A widely used simplification in the prediction of the three-dimensional structure of proteins, from its amino acid sequence, is based on the so-called statistical potentials. Here the real interactions between atoms are substituted by a general (distance independent) interaction potential between amino acids that depends on the frequency of contact of the two amino acids in a bunch of globular proteins that are chosen to form the data basis. In this way we obtain a 20 by 20 matrix that is used to predict the three dimensional structure of proteins. Examples of such matrices are those of Miyazawa and Jernigan (MJ99) [6,7] and that of Li, Tang and Wingreen (LTW) [8], from which we obtain the classifications that are depicted in Figs. 1 and 2.

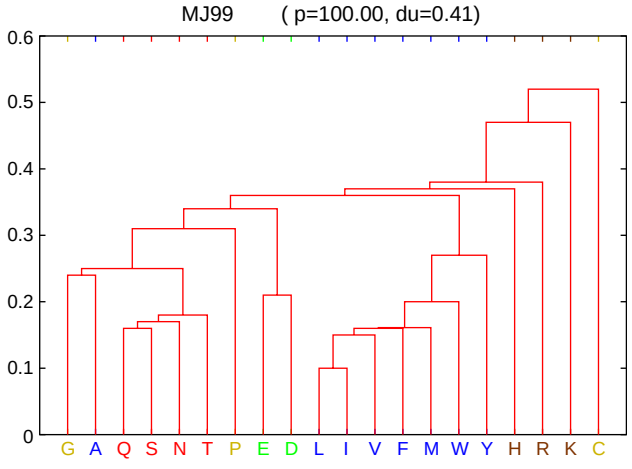

Fig. 1. Amino acid classification based on the Miyazawa–Jernigan potential. The value p=100 corresponds to that which minimizes d_u.

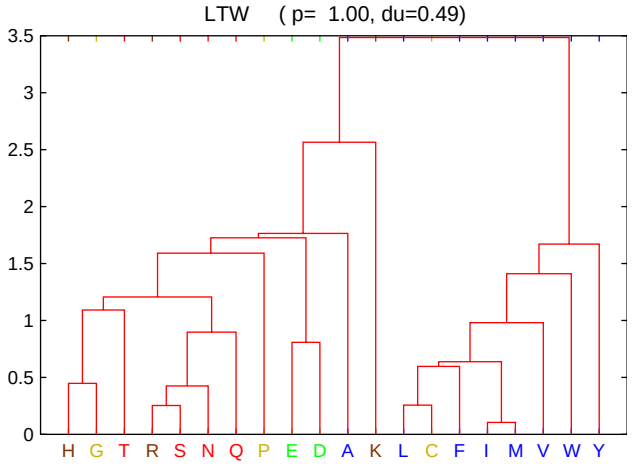

Fig. 2. Amino acid classification based on the Li–Tan–Wingreen matrix. The optimal parameters are p=1, which corresponds to a d_u=0.49.

In Fig. 1 we see that the classification of the amino acids with the subdominant ultrametric induced by the Miyazawa–Jernigan potential reproduces, with some peculiarities, the four group classification (hydrophobic, polar, negatively and positively charged). Note, however, that the positively charged amino acids do not form a unique group (perhaps due to their different pK_R); another peculiarity is that the Cysteine constitutes its own group, probably as a consequence of the fact that it is the only amino acid that form sulphur bridges. It is also worth pointing out that the Alanine seems to fit better in the group of polar amino acids than in the hydrophobic group and that the subgroup formed by $\{M,V,F,I,L\}$ is the most compact group inside the hydrophobic amino acids, as well as the subgroup $\{S,T,Q,N\}$ is the most compact group among the polar amino acids, finally we note that $\{I,L\}$ are the most similar amino acids.

With respect to the LTW matrix it should be noted that, although their matrix elements are very similar to those of the MJ99 matrix (a linear fit gives $M_{i,j}^{LTW}$=1.00003 $M_{i,j}^{MJ99}$–0.0017 and a correlation coefficient of 0.998), however, as it is apparent from Fig. 2, it accounts only for the hydrophobic-polar character of the amino acids and misses some other chemical properties as for example the special character of the Cysteine (note the $\{L,C\}$ group) or the peculiarities of the $\{H,R,K\}$ group and that now the $\{I,M\}$ are the most similar amino acids. We believe that

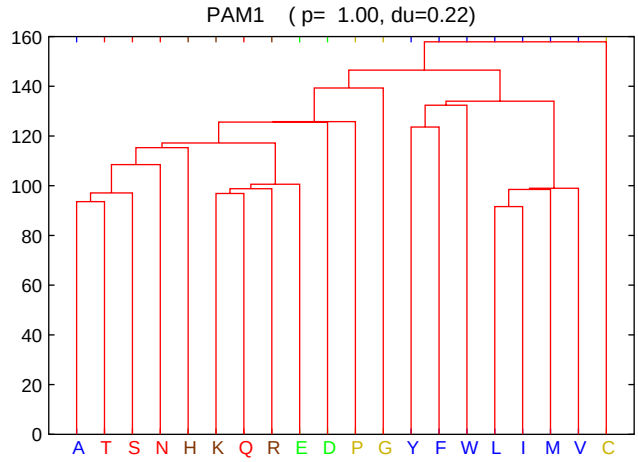

Fig. 3. Amino acid classification induced by the Dyhoff–PAM1 matrix.

this is a paradigmatic example of the usefulness of the classification method, as it can distinguish clearly between two nearly equal potential matrices and they tell us what the second matrix is missing.

3. Substitution matrices

If we assume a Markovian model of evolution for the genetic code i.e. the changes in every point of the amino acid sequence of a protein are random and independent of the history and/or the neighbors; it is a natural question which is the probability for such transitions between different amino acids. PAM mutation matrices give these probabilities and are obtained by comparing similar sequences in different species and examining the frequency of the individual changes of amino acids. If we consider two sequences similar when they differ only in a 1% of the amino acids we obtain the PAM1 mutation matrix. By letting the evolution clock run for longer times we obtain PAM matrices based on proteins that differ in a higher percentage of the genetic code. Within the Markovian model we get PAMk=(PAM1)^k. A widely used PAM matrix is PAM250 that gives the frequency of substitution between sequences that agree in approximately 17% of their genetic codes.

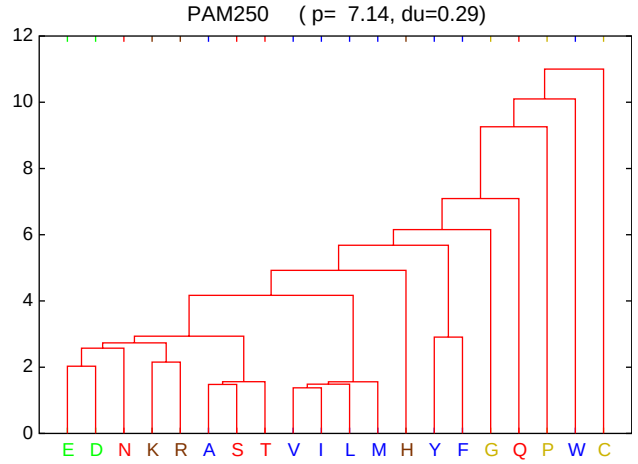

Fig. 4. Amino acid classification obtained from the PAM250 matrix.

Note, however, that PAM matrices are based in sequences that exist today in nature. From the evolutionary point of view it would be more natural to compare two such sequences with its common ancestor. From PAM matrices we can obtain the so-called Dyhoff–PAM matrices [9] (in honor of Margaret O. Dayhoff) that are used to find the alignment which maximizes the probability, for two given sequences, of having evolved from a common ancestor. This makes aligning sequences using Dayhoff–PAM matrices a soundly based algorithm.

In Figs. 3 and 4 we show the amino acid classification induced by the PAM1 and PAM250 matrices. The first difference we can see between the two dendograms above is that the one corresponding to PAM1 has all branches starting from the upper half of the tree while the one of PAM250 has the branches spread through all of it. This reflects the fact that in short times (PAM1) an amino acid has a large probability of being unchanged, which implies that the diagonal elements of the matrix are much larger than the rest and so the cluster structure is not very neat.

If we analyze the clusters in the Dyhoff PAM250 dendogram we see some clearly separated groups, namely $\{V,I,L,M\}$ and $\{A,S,T\}$ whose members have codons that

Table 1
The genetic code

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

The first, second and third codons are depicted in the first column, first row and last column, respectively.

differ only in the first base (see Table 1); $\{E,D\}$ differ in the last base; $\{K,R\}$ and $\{Y,F\}$ differ in the middle base. Note also that these groups are homogeneous in the physical chemical properties of their members (the only exception again is, significantly, the Alanine). Finally we might consider the big group $\{E,D,N,K,R\}$ whose members, except for the Arginine, lay in the third column of the table below. In conclusion, the Dyhoff–PAM250 matrix classification favors the clustering of the amino acids that differ in only one codon (in the same position) in their genetic code. With respect to the classification induced by the Dyhoff–PAM1 matrix, it seems to account well for the hydrophobic-polar character of the amino acids (and also for the special case of the Cysteine), so the mutation probability for "small" time periods seems to be essentially related with the chemical character of the amino acids involved in the mutation.

4. Conclusions

Different relations between amino acids are, usually represented in terms of matrices. In such a case, the analysis of the classification induced by those matrices can be useful in order to know which characteristics are taken (or not) into account. In this paper we used a small number of representative matrices to check our methods. We are planning to perform a more comprehensive study including the comparison of different substitution matrices like BLOSUM [10], potential matrices like Risler [11] or Birkbeck65 [12] for instance, and those that take into account the secondary structure Zhang-Kim [13], Zhu-Braun [14].

Acknowledgments

We are grateful to B.E. Villarroya, M.A. Ciriano and C. Tejel for comments. This work has been supported the CICYT grants BFM2000-1057 and FPA2000-1252.

References

[1] J.G. Esteve, F. Falceto, A general clustering approach with application to the Miyazawa–Jernigan potentials for amino acids, Proteins: Struct. Funct. Bioinform. 55 (2004) 999–1004.
[2] Lynne Reed Murphy, Anders Wallqvist, Ronald M. Levy, Simplified amino acid alphabets for protein fold recognition and implications for folding, Protein Eng. 13 (2000) 149–152.
[3] Sergey I. Rogov, Alexei N. Nekrasov, A numerical measure of amino acid residues similarity based on the analysis of their surroundings in natural protein sequences, Protein Eng. 14 (2001) 459–463.
[4] Xin Liu, Di Liu, Ji Qi, Wei-Mou Zheng, Simplified amino acid alphabets based on deviation of conditional probability from random background, Phys. Rev., E 66 (2002) 021906.
[5] Tanping Li, Ke Fan, Jun Wang, Wei Wang, Reduction of protein sequence complexity by residue grouping, Protein Eng. 16 (2003) 323–330.
[6] S. Miyazawa, R.L. Jernigan, Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading, J. Mol. Biol. 256 (1996) 623–644.
[7] S. Miyazawa, R.L. Jernigan, Self-consistent estimation of inter-residue contact energies based on a residue mixture approximation for proteins, Proteins: Struct. Funct. Genet. 34 (1999) 49–68.
[8] Chao Tang, Hao Li, Ned S. Wingreen, Nature of driving force for protein folding: a result from analyzing the statistical potential, Phys. Rev. Lett. 79 (1997) 765–768.
[9] Gaston Gonnet, Scientific Computation, http://linneus20.ethz.ch:8080/.
[10] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 10915–10919.
[11] J.L. Risler, M.O. Delorme, H. Delacroix, A. Henaut, Amino acid substitutions in structurally related proteins: a pattern recognition approach, J. Mol. Biol. 204 (1988) 1019–1029.
[12] M.S. Johnson, J.P. Overington, A structural basis for sequence comparisons: an evaluation of scoring methodologies, J. Mol. Biol. 233 (1993) 716–738.
[13] Chao Zhang, Sung-Hou Kim, Environment-dependent residue contact energies for proteins, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 2550–2555.
[14] H. Zhu, W. Braun, Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of beta-sheet formation in proteins, Protein Sci. 8 (1999) 326–342.