Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 250 (2008) 291–297
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

An amino acid map of inter-residue contact energies using metric multi-dimensional scaling

Sourav Rakshit, G.K. Ananthasuresh*

Mechanical Engineering, Indian Institute of Science, Bangalore 560012, India

Received 3 April 2007; received in revised form 7 August 2007; accepted 17 September 2007
Available online 26 September 2007

Abstract

We present an amino map based on their inter-residue contact energies using the Miyazawa–Jernigan matrix. This work is based on the method of metric multi-dimensional scaling (MMDS). The MMDS map shows, among other things, that the MJ contact energies imply the hydrophobic–hydrophilic nature of the amino acid residues. With the help of the map we are able to compare and draw inferences from uncorrelated data sets such as BLOSUM and PAM with MJ methods. We also use a hierarchical clustering method on our MMDS distance matrix to group the amino acids and arrive at an optimum number of groups for simplifying the amino acid set.

Keywords: MDS map; Reduced amino acid alphabet; Hierarchical clustering; MJ matrix; Hudrophobicity; BLOSUM; PAM

1. Introduction

In this work, we present a map based on the inter-residue contact energies given by the *Miyazawa–Jernigan* (MJ) matrix (Miyazawa and Jernigan, 1996). By presenting the data in a visual form, namely a map based on the metric multi-dimensional scaling (MMDS), we hope to reduce the complexity of finding out the inter-relations among the residues which might not be directly evident from the MJ matrix. Each amino acid is represented as a point on the MMDS map. The distance between two points on the map quantifies the dissimilarity in their contact energies. The larger the distance the larger the dissimilarity. This map brings out hidden relationships among the amino acids that are not easily discerned from the MJ matrix. The MMDS method is frequently used for a visual representation from a set of data representing the relation among a number of objects. Similar work was reported by French and Robson (1983) who had derived a map using MMDS for amino acids from Dayhoff's "relatedness odds matrix" (1972). Luthra et al. (2007), who presented a summary of different techniques to reduce the number of amino acid alphabet, note that French and Robson's (1983) work was one of the first attempts in reducing the amino acid alphabet. French and Robson (1983) were able to conclude from their map that hydrophobicity and molecular volume are two key properties that are conserved in the evolution of proteins. The MMDS map presented in this paper verifies that hydrophobicity is the key feature that characterizes the amino acid residues and that the inter-residue contact energies represent a rough hydrophobicity scale (Cornette et al., 1987; Chan, 1999; Venkatarajan and Braun, 2001). This map is based on the revised MJ matrix reported in 1996 and hence includes extensive structure and sequence information. Additionally, with the help of this map, we compare (the similarities/differences among amino acid residues as represented by) the MJ matrix with BLOSUM62 and PAM250 matrices. A novel feature of our map is that it can be used as a visual method of reducing the amino acid set. We support this by determining the groups using a hierarchical clustering method (Johnson and Wichern, 2006). By using the above method we are also able to arrive at an optimum number of groups for reducing the amino acid set. Recently, Agrafiotis et al. (2001) coupled MDS with nonlinear mapping (NLM) and neural networks and used it for mapping of large combinatorial

*Corresponding author. Tel.: +91 80 2293 2334.
E-mail addresses: srakshit@mecheng.iisc.ernet.in (S. Rakshit), suresh@mecheng.iisc.ernet.in (G.K. Ananthasuresh).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.09.032

libraries and ensemble of molecular conformations but not for the classification of amino acids. On the other hand, Venkatarajan and Braun (2001) used principal component analysis (Johnson and Wichern, 2006) for creating amino acid maps using large data sets. They used 237 physical–chemical properties of amino acids to form a vector in a 237-dimensional space for each amino acid and reduced the resulting matrix to a five-dimensional space. This was done by using the first five eigenvalues and eigenvectors. They showed that the principal components correspond to important properties such as hydrophobicity–hydrophilicity and molecular volume. As discussed in this paper, the same conclusion is drawn from the MMDS map presented in this paper in addition to some other results.

The rest of the paper is organized as follows. In the next section, we explain the method of multi-dimensional scaling that has been used in constructing the map. In the following section, we present the results and how the map can be used for finding out features that are not immediately apparent from the MJ matrix. The final section contains the concluding remarks.

2. Method

Metric multi-dimensional scaling (Mead, 1992) is a multi-variate statistical analysis technique that is used for making a visual representation from a $n \times n$ matrix representing the interaction between a set of n objects that one is interested to study. The ijth entry in the matrix represents the interaction between ith and jth objects. If the ijth entry in the matrix represents dissimilarity between the ith and jth objects, then the matrix is called the *dissimilarity* or *distance* or *proximity matrix*. Here, as there can be no dissimilarity between an object and itself all the diagonal elements are zero. On the other hand, if the ijth entry into the matrix represents similarity between the ith and jth objects, then the matrix is called the *similarity matrix*. In this case the diagonal elements are non-zero. The results are represented as a plot of n points representing the n objects on a space of two or higher dimensions. This method was first suggested by Torgerson (1952) and then developed and used by Kruskal and Wish (1978) in representing as varied and qualitative things as cultural similarity among nations and dialects of Salish Indians. More recently, this map was used for classifying engineering materials based on ergonomic and aesthetic considerations (Ashby and Johnson, 2002). The key feature of this method is that it reveals the hidden structure among the objects that lies buried in the mass of data stored in a matrix form. Similar points are huddled together in the plot and the distances among the points give a measure of similarity among the objects. Furthermore, one can often identify variation of key parameters on which these objects depend along different directions in the map.

Mathematically, constructing an MMDS map can be shown to be a least-square minimization problem. Let n objects be represented by a set of n points on a plane. Let the distance between ith and jth points be d_{ij} and its corresponding entry in the proximity matrix is δ_{ij}. The MMDS technique attempts to minimize all such distances in the sense of least squares, i.e.,

$$\underset{\bar{\mathbf{x}},\bar{\mathbf{y}}}{\text{Minimize}} \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \left(d_{ij}^{(2)} - \delta_{ij}\right)^2, \tag{1}$$

where $\bar{\mathbf{x}} = \{x_1, x_2, \ldots, x_n\}$ and $\bar{\mathbf{y}} = \{y_1, y_2, \ldots, y_n\}$ are the x and y coordinates of the n points in the map, and

$$d_{ij}^{(2)} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \tag{2}$$

where the superscript in brackets indicates the dimensionality of the MMDS map (in this case it is two as we have chosen a planar representation). Therefore, we can write Eq. (1) as

$$\underset{\bar{\mathbf{x}},\bar{\mathbf{y}}}{\text{Minimize}} \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \left(\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} - \delta_{ij}\right)^2. \tag{3}$$

The solution of the minimization problem in Eq. (3) gives the coordinates of the points and helps create the MMDS map. It should be noted that the MMDS map is unaffected by the orientation of the chosen coordinate system, i.e., the final set of points may be oriented differently in different runs with different initial guesses but the relative positions of the points do not change. This happens because MMDS deals with only the distances between the points which are devoid of any directional information. We have used MATLAB's optimization toolbox program `fminunc` (unconstrained optimization which uses sequential quadratic programming combined with trust region method) to solve the above least-square minimization problem in constructing the map. However, the ijth entry of the MJ matrix cannot be directly used as δ_{ij} in Eq. (3). The treatment of the MJ matrix to get the δ_{ij}s is discussed next.

The ijth entry in the MJ matrix represents the contact energy between ith and jth amino acids. The diagonal entries represent contact energy between same amino acids. Therefore, the extent to which the ijth entry matches the corresponding diagonal entries (both ith and jth diagonal entries) represents the similarity between the ith and jth amino acids. Thus, the MJ matrix can be taken as a similarity matrix. To convert the MJ matrix to a proximity matrix, we do the following operation:

$$\delta_{ij} = \left|M_{ij} - (M_{ii} + M_{jj}/2)\right|. \tag{4}$$

Here we take the absolute value as δ_{ij} represents distance between two points and hence is always positive.

This symmetric transformation ensures that all the diagonal entries are zero. With this matrix now one can select multiple dimensions (one and above) for minimizing Eq. (3). For a given dimension it may not be possible to position all the points on the map such that the distances among them exactly match the corresponding distances given by the proximity matrix. The extent to which it

deviates from actual data is given by a measure called *stress* (Kruskal and Wish, 1978).

This stress is given by

$$\text{Stress}\,(q) = \sqrt{\frac{\sum_i \sum_{j,i<j} \left(d_{ij}^{(q)} - \delta_{ij}\right)^2}{\sum_i \sum_{j,i<j} \delta_{ij}^2}}, \tag{5}$$

where q is the dimension of MMDS map. For example, when q is two, d_{ij} is given by Eq. (2). The values of stress calculated for one, two and three dimensions are shown in Fig. 1. We selected two as the dimension because stress is lowest there. We did not investigate stress in higher dimensions (four and above) because it is difficult to view cluster of points mapped in higher dimensions. Next, we show the scatter diagram for two dimensions in Fig. 2. The scatter diagram is a graphical representation of how well

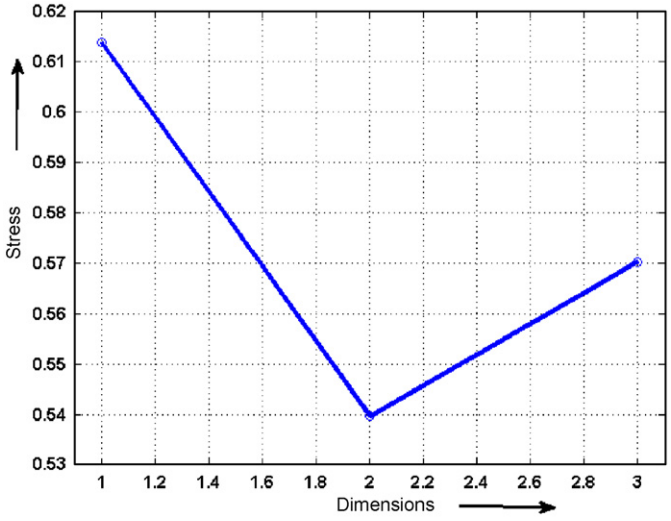

Fig. 1. Plot of stress against the number of dimensions.

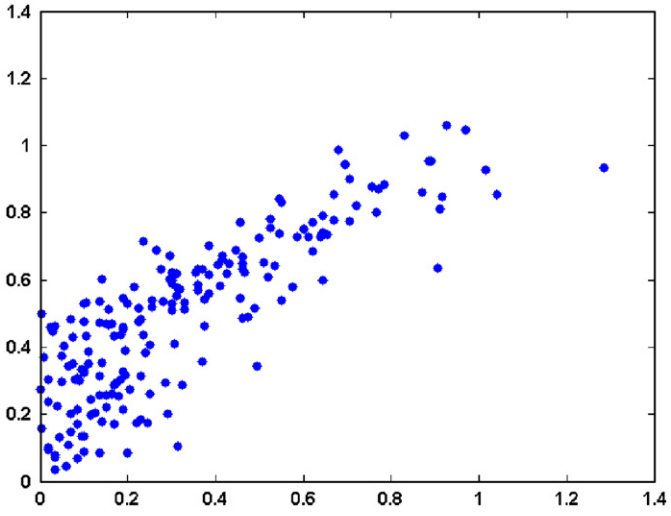

Fig. 2. Scatter diagram showing the discrepancies between entries in the distance matrix and corresponding distances calculated from MMDS map.

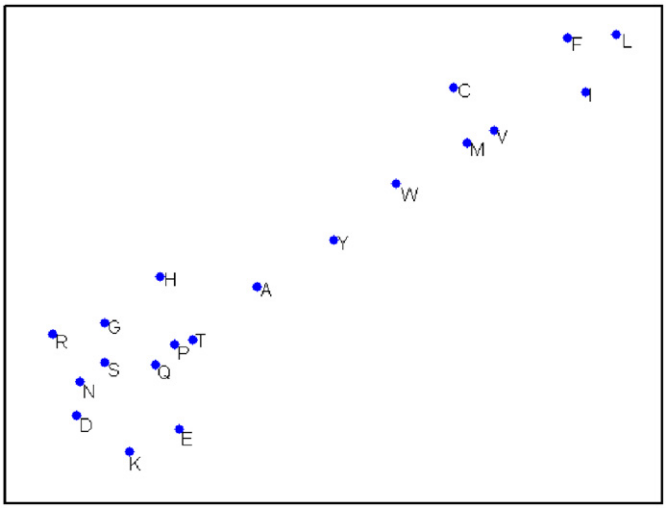

Fig. 3. MMDS amino acid map constructed using the matrix where we subtracted diagonal elements from the corresponding rows.

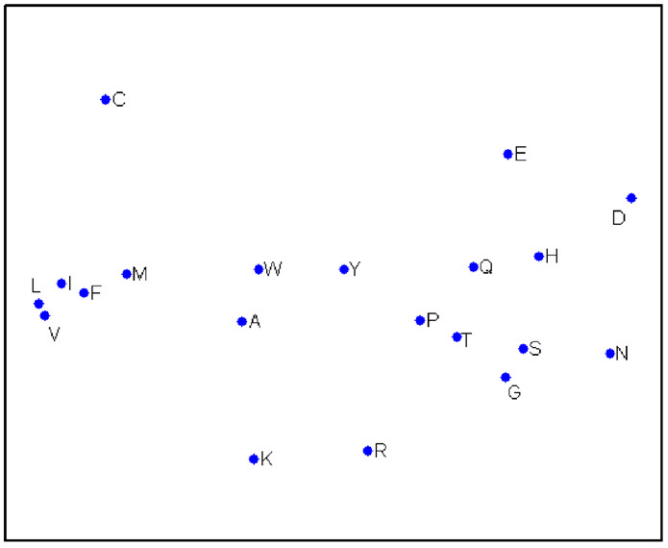

Fig. 4. Amino acid map constructed using the metric multi-dimensional scaling method and the modified Miyazawa–Jernigan matrix as the proximity matrix.

the distances given in the proximity matrix correlate with the distances calculated between corresponding points in the MMDS map. The correlation coefficient of the distances from MMDS map and proximity matrix is 0.828 and the RMS error is 0.224. We have formulated an alternative proximity matrix by subtracting the diagonal elements from the corresponding rows and performed MMDS on it. The resultant map in two dimensions is shown in Fig. 3. This map gives a better correlation coefficient (0.991) and RMS error (0.202). However, by treating the MJ matrix in this manner we lose the symmetry of the proximity matrix and the elements become dependant on the order of the amino acids in the diagonal. Hence, we forgo this method and stick to the conventional way of forming a proximity matrix which we have already described.

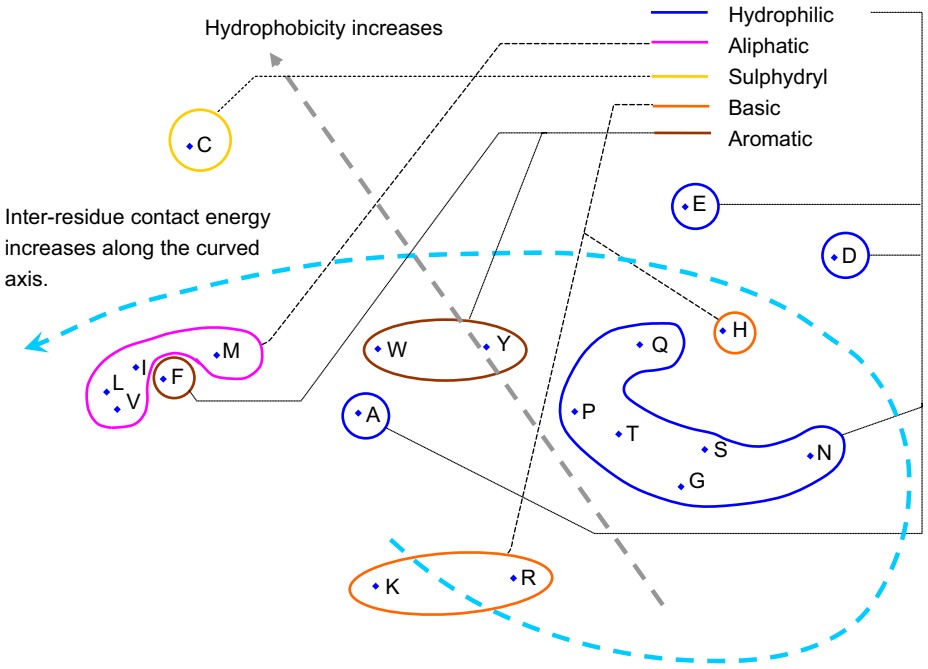

Fig. 5. The straight axis corresponds to an increase in hydrophobicity. The curved axis shows the direction along which inter-residue contact energies increase. Dayhoff's classification of amino acids in five groups based on chemical properties is shown with legends in top right corner.

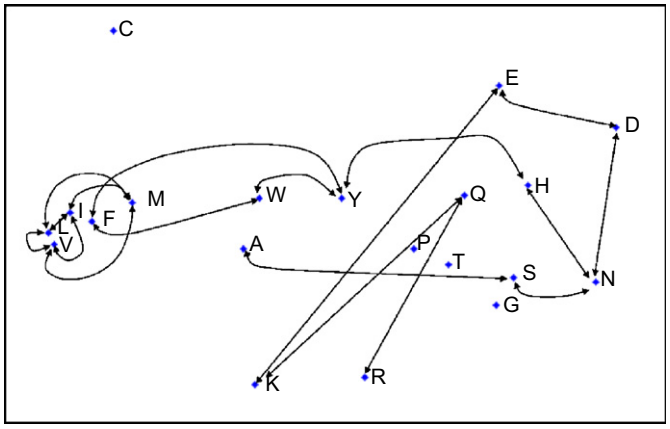

Fig. 6. The residues that have a positive log odd score in the BLOSUM62 matrix are connected by double-ended arrows.

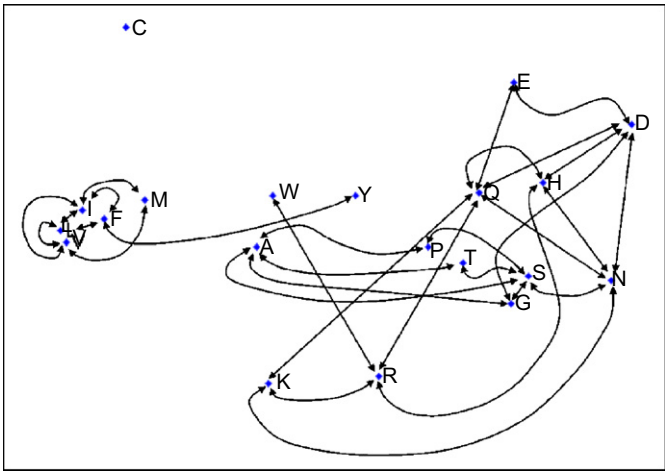

Fig. 7. The residues that have a positive log odd score in the PAM250 matrix are connected by double-ended arrows.

3. Results and discussion

Fig. 4 shows the map created using MMDS on MJ matrix. We note that the residues lie along an axis that corresponds to an approximate increase in hydrophobicity (Cornette et al., 1987). This axis is shown in Fig. 5. The curved axis in Fig. 5 shows the direction of increase in inter-residue contact energies. We also show the classification of amino acids according to their chemical properties as done by Dayhoff et al. (1972) in this figure.

In Fig. 6, we show the residues that favorably substitute one another in the BLOSUM database on the map. All the residues that have a positive log odd score in the BLOSUM62 matrix are connected by double-ended arrows in this figure. This figure shows that both substitutionally (BLOSUM62) and energetically (MJ matrix) Cysteine stands separate from other amino acid residues. The hydrophobic residues and the hydrophilic ones do not substitute one another favorably. This substitution is also unfavorable from contact energy viewpoint as shown by the map. According to the map, Proline, Threonine and Glutamic acid, being near to one another, should be favorable for substitution; this inference is not supported by BLOSUM. However, our conclusion can be supported from the viewpoint of conservation of molecular volume in evolutionary substitution (French and Robson, 1983). Proline, Threonine and Glutamic acid can be grouped together in one class characterized by their smallness of volume (Schulz and Schirmer, 1979).

Fig. 7 shows the residues that substitute one another favorably in the Percent Accepted Mutations (PAM) matrix. We connect the residues that have a positive log odd score in the PAM250 matrix by double-ended arrows. Here too, we see that Cysteine stands separate form all other amino acids in terms of evolutionary substitution (PAM250) and inter-residue contact energy (MJ). The hydrophobic and hydrophilic residues are likely to have different lineages in evolution as they do not substitute one another favorably (Dayhoff et al., 1972; Miyata et al., 1979). Here, we feel that it is worth mentioning that this map represents a unique and a novel way of drawing

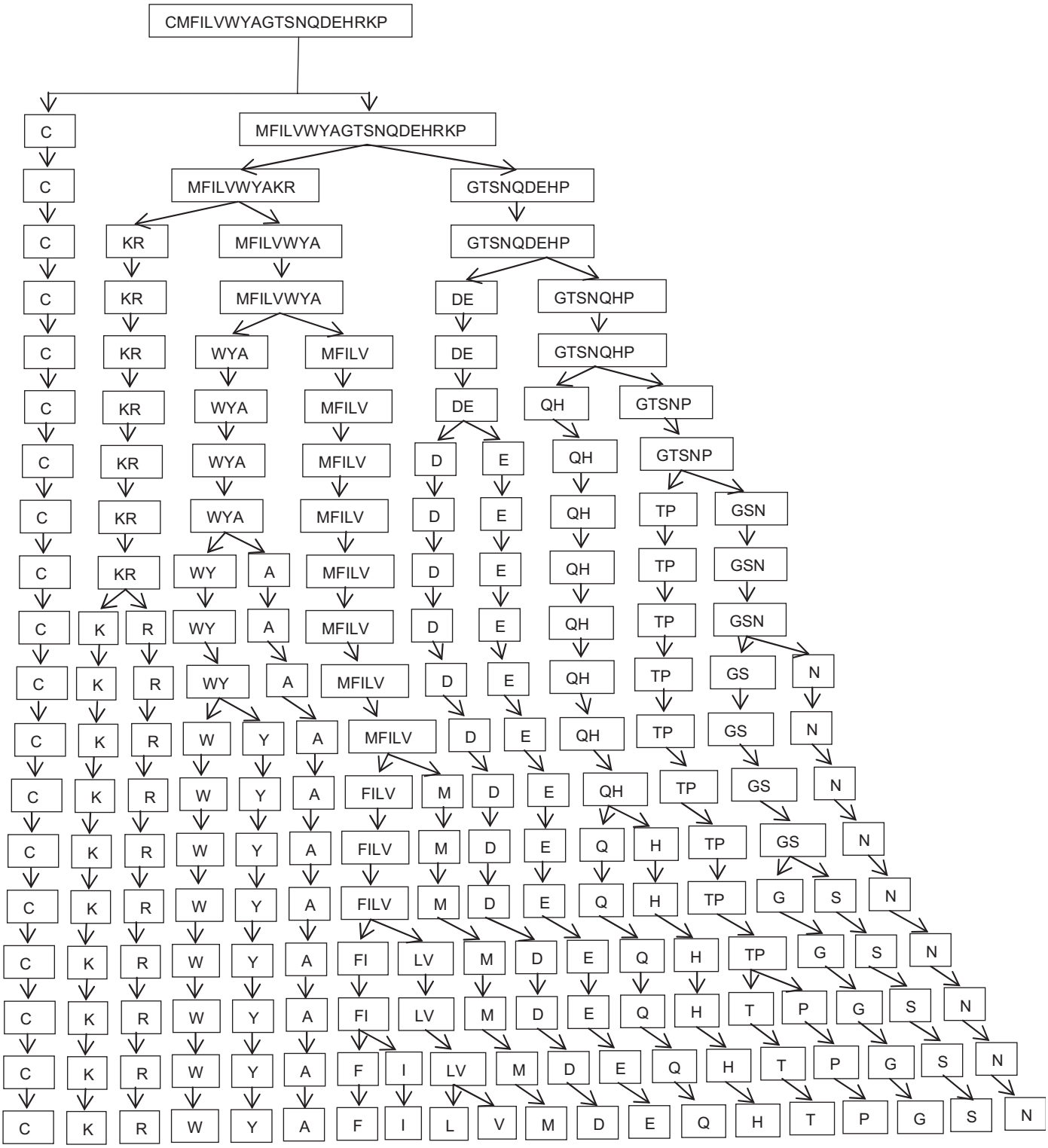

Fig. 8. Dendrogram showing hierarchical grouping of amino acids based on our distance matrix.

conclusions from two different data sets related to the amino acid residues, namely the block substitution (BLOSUM) or evolutionary (PAM250) data and inter-residues contact energy data derived from experimental data (MJ).

Reducing the amino acid residues to a small set is a topic of active interest among protein researchers. A few works based on the MJ matrix exist in the current literature (Wang and Wang, 2002, 1999; Cieplak et al., 2001; Li et al., 1997). All these works employ different methods for reducing the amino acid set. The MMDS method provides an easy visual method of grouping (see Fig. 4). To reinforce this, we use a hierarchical clustering method based on average distance of clusters (Johnson and Wichern, 2006) on our distance matrix to simplify the amino acid set. In this method, we find the minimum distance in the distance matrix and group the corresponding amino acids. Next, we find the distance between this group and all other amino acids by calculating the mean distance from this group to all other amino acids or groups.

Thus, if L (Leucine) and I (Isoleucine) are clubbed together to form a group {L,I}, then the distance $d_{\{L,I\}G}$ of this group from G (Glycine) is given by $(d_{LG}+d_{IG})/2$. We continue this procedure starting from 20 amino acids and go on grouping until we arrive at a single group. The resulting dendrogram is shown in Fig. 8.

In Fig. 9, we plot the minimum distance between the groups as we go on decreasing the number of groups. We see that the highest ratio of increase in the minimum distance to the current minimum distance occurs when the number of clusters changes from 19 to 18 and from five to four or four to three. A sudden increase in minimum distance indicates that the groups are losing their compact size as they are merged together. Since 18 is a large number for reducing the amino acid set, we conclude that five or four is the best number for simplifying the amino acid set. Our conclusion is further supported by the fact that grouping amino acid residues into five sets is most common in the literature (Dayhoff et al., 1972; Li et al., 1997, 2003; Wolynes, 1997; Murphy et al., 2000; Wang and Wang, 2002, 1999; Cieplak et al., 2001; Cannata et al., 2002; Koisol et al., 2004). In Fig. 10, we show these five groupings on our MMDS map. Although our grouping is based on hierarchical clustering and our database is contact energies from a statistical database we find distinctive chemical properties within each group. D and E are acidic whereas K and R have basic properties. Q, H, P, T, S, G, N all have small molecular volume and are hydrophilic in nature. On the other hand, L, V, I, M, F, W, Y (all except A in that group) are characterized by their largeness in size and hydrophobic nature. Lastly, C stands alone because of its unique ability to form disulphide bonds.

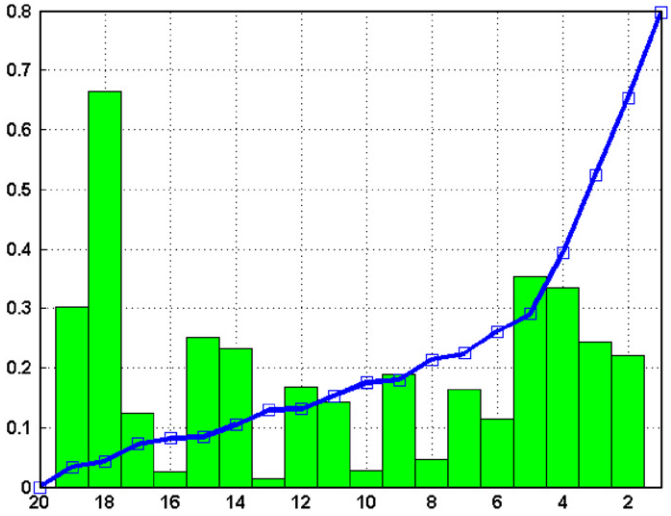

Fig. 9. Minimum distance between groups as a function of the number of groups. The ratio of increase in minimum distance (between groups) as we reduce the number of groups to the minimum distance (between groups) in current number of groups is highest for 18 and five groups.

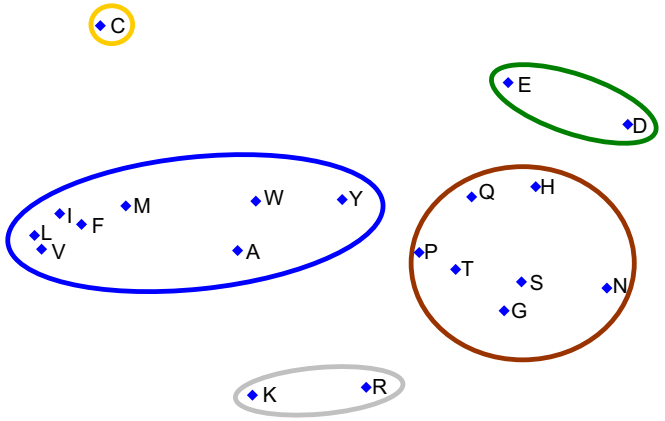

Fig. 10. Grouping of amino acids into five groups based on hierarchical clustering method. This grouping coincides if one goes for clustering amino acids into five groups on MDS map based on visual inspection alone. The hierarchical clustering in Fig. 9 can also be done from our MDS map by mere visual inspection.

4. Conclusions

The metric multi-dimensional map (MMDS) of the amino acid residues gives an informative and revealing representation of the MJ matrix on a two-dimensional plane. It shows which amino acid residues are similar to one another in terms of statistics-based contact energy. We infer from the map that the inter-residue contact energies in the MJ matrix underscore the hydrophobic–hydrophilic character of the amino acid residues. We are able to represent on the same platform, i.e., the MMDS map, different data sets (BLOSUM and MJ or PAM and MJ) related to the amino acid residues. This map can serve as a simple way of grouping the amino acids into reduced number of sets. We support our claim by using a hierarchical clustering method which gives the same groups as would be determined by visual inspection from our MMDS map. Finally, by using this method we are able

to arrive at an optimum number for reducing the amino acid set.

Acknowledgments

The authors thank Prof. Saraswathi Vishveswara (Molecular Biophysics Unit, Indian Institute of Science) for helpful technical discussions. The authors are also grateful to an anonymous reviewer for his/her insightful suggestions, which helped improve this paper. This work was supported in part by the Swarnajayanthi fellowship of the Department of Science and Technology, Government of India, to the second author. This support is gratefully acknowledged.

References

Agrafiotis, D.K., Rassokhin, D.N., Lobanov, V.S., 2001. Multidimensional scaling and visualization of large molecular similarity tables. J. Comp. Chem. 22 (5), 488–500.

Ashby, M.F., Johnson, K., 2002. Materials and Design: The Art and Science of Material Selection in Product Design. Butterworth-Heinemann.

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

Chan, H.S., 1999. Folding alphabets. Nat. Struct. Biol. 6 (11), 994–996.

Cieplak, M., Holter, S.N., Maritan, A., Banavar, R.J., 2001. Amino acid classes and protein folding problem. J. Chem. Phys. 114, 1420–1423.

Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A., DeLisi, C., 1987. Hydrophobicity scales and computational techniques for detecting amphipatic structures in proteins. J. Mol. Biol. 195, 659–685.

Dayhoff, M.O., Eck, R.V., Park, C.M., 1972. A model of evolutionary change in proteins. In: Dayhoff, M.O. (Ed.), Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Maryland, pp. 89–100.

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

Johnson, R.A., Wichern, D.W., 2006. Applied Multivariate Statistical Analysis. Pearson Education Inc.

Koisol, C., Goldman, N., Buttimore, H.N., 2004. A new criteria and method for amino acid classification. J. Theor. Biol. 228, 97–106.

Kruskal, J.B., Wish, M., 1978. Multidimensional Scaling. Sage Publications.

Li, H., Tang, C., Wingreen, N.S., 1997. Nature of driving force for protein folding: a result from analyzing statistical potential. Phys. Rev. Lett. 79 (4), 765–768.

Li, T., Fan, Ke., Wang, J., Wang, W., 2003. Reduction of protein complexity by residue grouping. Protein Eng. 16, 323–330.

Luthra, A., Jha, A.N., Ananthasuresh, G.K., Vishveswara, S., 2007. J. Biosci. 32 (5).

Mead, A., 1992. Review of the development of Multidimensional Scaling methods. The Statistician 41 (1), 27–39.

Miyata, T., Miyazawa, S., Yasunaga, T., 1979. Two types of amino acid substitutions in protein evolution. J. Mol. Evol. 12, 219–236.

Miyazawa, S., Jernigan, J.L., 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J. Mol. Biol., 256–523.

Murphy, R.L., Wallqvist, A., Levy, M.R., 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng. 13, 149–152.

Schulz, G.E., Schirmer, R.H., 1979. Principles of protein structure. In: Cantor, C.R. (Ed.), Springer Advanced Texts in Chemistry. Springer, pp. 10–16.

Torgerson, W.S., 1952. Multidimensional scaling: I. Theory and method. Psychometrica 17 (4), 401–419.

Venkatarajan, M.S., Braun, W., 2001. New quantitative descriptors of amino-acids based on multidimensional scaling of a large number of physical–chemical properties. J. Mol. Model. 7, 445–453.

Wang, J., Wang, W., 1999. A computational approach to simplifying the protein folding alphabet. Nat. Struct. Biol. 6, 1033–1038.

Wang, J., Wang, W., 2002. Grouping of residue based on their interactions. Phys. Rev. E 65, 041911–041915.

Wolynes, P.G., 1997. As simple as can be. Nat. Struct. Biol. 4, 871–874.