

---

# A method for computing the inter-residue interaction potentials for reduced amino acid alphabet

ABHINAV LUTHRA<sup>1</sup>, ANUPAM NATH JHA<sup>2</sup>, G K ANANTHASURESH<sup>3,\*</sup> and SARASWATHI VISHVESWARA<sup>2,\*</sup>

<sup>1</sup>Department of Biotechnology, Indian Institute of Technology-Guwahati, Guwahati 781 039, India

<sup>2</sup>Molecular Biophysics Unit, <sup>3</sup>Department of Mechanical Engineering, Indian Institute of Science, Bangalore 560 012, India

\*Corresponding authors (Email, suresh@mecheng.iisc.ernet.in; sv@mbu.iisc.ernet.in)

Inter-residue potentials are extensively used in the design and evaluation of protein structures. However, dealing with all (20×20) interactions becomes computationally difficult in extensive investigations. Hence, it is desirable to reduce the alphabet of 20 amino acids to a smaller number. Currently, several methods of reducing the residue types exist; however a critical assessment of these methods is not available. Towards this goal, here we review and evaluate different methods by comparing with the complete (20×20) matrix of Miyazawa-Jernigan potential, including a method of grouping adopted by us, based on multi dimensional scaling (MDS). The second goal of this paper is the computation of inter-residue interaction energies for the reduced amino acid alphabet, which has not been explicitly addressed in the literature until now. By using a least squares technique, we present a systematic method of obtaining the interaction energy values for any type of grouping scheme that reduces the amino acid alphabet. This can be valuable in designing the protein structures.

[Luthra A, Jha A N, Ananthasuresh G K and Vishveswara S 2007 A method for computing the inter-residue interaction potentials for reduced amino acid alphabet; *J. Biosci.* **32** 883–889]

## 1. Introduction

The objective of this paper is two-fold: first, we give an overview of the methods used to reduce the amino acid alphabet; second, we present a method for computing the inter-residue interaction energy values for reduced amino acid alphabet. The motivation for this work is based on the fact that the computational complexity of protein folding and design significantly reduces if fewer amino acid types than the naturally occurring 20 are identified. Therefore, such a reduced set of amino acid alphabet continues to be of interest in protein studies. Reduced number of amino acid types has at least two benefits. On the one hand, it has been experimentally observed that proteins containing fewer than 20 amino acids can fold to stable native conformations (e.g. Regan and DeGrado 1988; Riddle *et al* 1997). On the other hand, probable folding families can be predicted for a newly discovered sequence by interchanging

similar amino acids in known homology models (e.g. Murphy *et al* 2000).

Grouping 20 amino acids is a computationally daunting task: there are, for example, 524,287 and 749,206,090,500 possible groupings if all amino acids are to be reduced to two and five groups respectively (Cannata *et al* 2002). This is given by the *Stirling number* that identifies the number of ways in which  $n$  objects can be put into  $m$  groups wherein each group contains a variable number of objects. For  $m = 8$ , this number attains the maximum value of 15,170,932,662,679. Clearly, identifying the best grouping by exhaustively evaluating all possible groupings using a scoring scheme is computationally impractical. Therefore, several methods for doing this with reasonable computation are developed in the literature. Such methods are based on physical, chemical and biological properties and insights gained from protein studies. As described in this § 1.1, there is no exact agreement among the different groupings

**Keywords.** Inter-residue potential; multidimensional scaling; protein folding and design; reduced MJ matrix

Abbreviations used: DPA, direct pairwise alignment; NPS, normalized positional shift; PMA, multiple structural alignment; RMSD, root mean square deviations.

because each method uses a different scoring scheme and a different method for grouping based on that score. In § 2, we use an alternate method for amino acid grouping based on multi dimensional scaling (MDS) method (Kruskal and Wish 1978), which provides a visual representation that gives a clue for rational grouping that accounts for the overall similarity and disparity among the amino acids.

After the amino acids are divided into a few groups, it is useful to characterize the inter-residue interaction energy values among the reduced amino acid alphabet, i.e. among the identified groups. Wang and Wang (1999) identify one representative amino acid for each group. This allows them to compute the inter-group energetic interactions. But most other works do not explicitly report how the inter-group energetic interactions are quantified. Therefore, a systematic approach to this problem is warranted so that a matrix similar to the Miyazawa-Jernigan (MJ) matrix (Miyazawa and Jernigan 1996) is obtained for any given reduced alphabet. Such a matrix is useful and almost essential for computational methods that design a sequence for a desired conformation. Thus, the problem addressed in this paper can be stated as follows: given a grouping that divides 20 amino acids into  $m$  groups, obtain a reduced inter-residue interaction energy matrix  $\mathbf{R}_m$  of size  $m \times n$ . The method we propose for this purpose is described in § 2. Section 3 contains results while concluding remarks are in the last section. Next, we present a consolidated review of the work done on grouping the amino acids in the last few decades so that different methods can be compared and contrasted with each other.

### 1.1 A review of grouping of amino acids

Table 1 summarizes some of the classifications of amino acids from the literature starting with the pioneering

work of Dayhoff during 1965-1978. As can be seen in the table, the criteria and methodologies for grouping are significantly different and are constantly evolving. The grouping criterion in the first column of table 1 refers to the underlying rationale for a certain classification. The second column refers to the computational (or otherwise) technique used to identify the best grouping using a grouping criterion. That is, the grouping criterion enables comparison of different groupings. The grouping criterion is often not easily discernible from the papers nor is it explicitly stated. Hence, the remarks in the second column in table 1 are our best interpretation of what is stated in the respective papers.

The list of groupings shown in Table 1 is not exhaustive. But it is adequate to notice that there is no universal rule for grouping because different concepts such as physical, chemical and biological or quantitative measures are used. Dayhoff *et al* (1978) used the likelihood of substituting one amino acid with another and thus developed a "similarity matrix". It is pertinent to note that the structural information about the proteins that was available at the time of this pioneering work was rather limited. Taylor (1986) noted this fact and proposed a Venn diagram representation of the information contained in Dayhoff's matrix and augmented it with local structural information and physico-chemical properties. This paved the way for quantitative comparisons using the formal logic of set theory. French and Robson (1983) used the similarity matrix of Dayhoff *et al.* (1978) and organized this data innovatively into a visual 2D map on the basis of MDS (Kruskal and Wish 1978).

Some newer measures for characterizing the similarity of amino acids have been proposed in the last decade. Wang and Wang (1999) proposed a metric that they called *mismatch* among blocks of the MJ matrix. Later they (Wang and Wang 2002) used this metric and an energy landscape technique to classify amino acids into eight to 10 groups.

**Table 1.** A few different methods for amino-acid grouping

Criterion for grouping	Grouping technique	Reference
Likelihood matrix of amino acid mutability	Not clear	Dayhoff <i>et al</i> (1978)
Tendencies towards different secondary structures and physico-chemical properties	Not clear	Robson and Suzuki (1976)
Similarity matrix of Dayhoff	Multi-dimensional scaling (MDS)	French and Robson (1983)
Physico-chemical, structural and mutation data	Venn diagram representation	Taylor (1986)
A special metric called "mismatch" between pairs	Appears to be exhaustive enumeration	Wang and Wang (1999)
BLOSUM50 similarity score	Correlation coefficient	Murphy <i>et al</i> (2000)
"Mismatch"	Energy landscape based	Wang and Wang (2001)
Miyazawa-Jernigan matrix	Eigenanalysis	Cieplak <i>et al</i> (2001)
Substitution matrices such as PAM and BLOSUM	Branch and bound algorithm	Cannata <i>et al</i> (2002)
BLOSUM62 similarity score	Enumeration and evaluation	Li <i>et al</i> (2003)
A special metric called "conductance"	Perturbation theory applied to Markov matrices	Koisol <i>et al</i> (2004)

Li *et al* (1997) used eigenvalue decomposition of the MJ matrix and developed a *q-scale* for amino acids. They showed that this scale correlates with the structure information. They also used this scale to separate 20 amino acids into two broad groups. Cieplak *et al* (2001) extended this eigenanalysis technique and classified amino acids into five groups. Cannata *et al* (2002) used amino acid substitutions matrices such as PAM (Schwartz and Dayhoff 1978) and BLOSUM (Henikoff and Henikoff 1992) and exhaustively evaluated all possible simplified alphabets using a branch and bound algorithm. They used an average of similarity scores of sequence alignment and account for different combinations of substitutions for the amino acids in each reduced group. Li *et al* (2003) also used substitution matrices and sequence-alignment scores but used a Monte Carlo method to obtain the best grouping. They also developed a measure that they call *normalized coverage* of structural information and conclude that 10 groups are the best grouping strategy. Recently, Koisol *et al* (2004) proposed a new measure called *conductance* that quantifies mixing in a Markov process that they used to effect mutative substitutions using PAM, JTT (Jones *et al* 1992) and WAG (Whelan and Goldman 2001) matrices. There are some similarities and considerable variations in the grouping schemes reported in the literature. While reviewing a few different methods of groupings, Chan (1999) noted that experimental verification of predicted folding of reduced-alphabet proteins is the ultimate validation of different methods. There is also no universal agreement on the minimum number of groups. Using energy landscape arguments, Wolynes (1997) suggested that at least five types are necessary for protein-like folding. Fan and Wang (2003) thought that the minimum number of groups is probably ten. Murphy *et al* (2000) also thought that at least 10 groups are necessary and that further reduction results in rapid degradation of information required for identification of structural homologues.

## 2. Methods

### 2.1 A method for creating MDS

The MDS is a widely used method in social sciences and psychometry (Kruskal and Wish 1978). It is particularly suitable to identify similar groups in a given collection of varied objects when the data about their similarities to each other is imprecise. Thus, it is well suited for amino acids where there is much debate about which amino acids are more similar to some as opposed to others. The highlight of MDS is that the criteria used to quantify similarity do not matter in the sense that all criteria can be taken together at once in specifying inter-pair similarities in a group. It also does not matter if objective opinions and qualitative considerations are quantified approximately. The method

is said to be robust against uncertainties. In 1983, French and Robson considered the similarity matrix of Dayhoff *et al* (1978) for this purpose. However, as stated earlier, the number of proteins with known sequences and structures was also small at that time. Thus, here, we revisit the application of MDS taking  $\mathbf{D}$ , which is the similarity matrix, to be the MJ matrix, which is rich in knowledge about the protein structure and biological and physico-chemical properties.

For the purpose of plotting the 2D map of MDS, we choose an origin and a coordinate system, and then determine the coordinates for the 20 points corresponding to the amino acids. For this purpose, we use the inter-residue energy values given by the MJ matrix. All the entries are normalized by taking their ratios with the largest magnitude so that they lie between 0 and -1. The diagonal entries are taken as the distances of the amino acid points from the origin and the off-diagonal entries indicate the pair-wise distances between the amino acid points. That is, if the entry  $D_{ij}$  is close to 0, it would mean that *i*th and *j*th amino acids are similar. Likewise, if it is close to 1, they are dissimilar. A value in between the two extremes (i.e. 0 and 1) indicates the extent of similarity or disparity.

In the 2D map, we need the (*x*, *y*) coordinates of each amino acid. Thus, there are 40 unknown variables. By using these unknowns as optimization variables the following problem is solved to arrive at pair-wise "distances" between points are as close to those in  $\mathbf{D}$  as possible.

$$\text{Minimize}_{x_i=1,20, y_i=1,20} F = \sum_{i=1}^{20} \sum_{j=1}^{20} (d_{ij} - D_{ij})^2, \quad (1)$$

where  $d_{ij}$  is the computed distance between two points corresponding to *i*th and *j*th amino acids. It is given by

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

Clearly, the minimization problem in eq. (1) will achieve the best value (i.e. 0) of the objective function *F* for arbitrary distances given in  $\mathbf{D}$  only when  $n = 3$ . Beyond that unless there is a consistent 2D structure in  $\mathbf{D}$ , the obtained solution can only be approximate. The results obtained with this method are presented in § 3.1.

### 2.2 Computation of the energies of the reduced amino acid alphabet

Accurate computation of the total free energy of a protein by *ab-initio* methods is extremely demanding. One of the methods generally used is based on molecular dynamics simulations (e.g. Wang *et al* 2001; Hansson *et al* 2002; Karplus and McCammon 2002), which is prohibitively expensive. Hence, knowledge based (from protein structures) statistical potential functions is the method of choice for coarse-grained studies. Several potentials have

$$\begin{array}{c}
 \begin{array}{ccccc}
 & 1..7 & 8..12 & 13..16 & 17..19 & 20 \\
 1..7 & \left[ \begin{array}{c|c|c|c|c} M_{1..7,1..7} & M_{1..7,8..12} & M_{1..7,13..16} & M_{1..7,17..19} & M_{1..7,20} \\ \hline 8..12 & M_{8..12,1..7} & M_{8..12,8..12} & M_{8..12,13..16} & M_{8..12,17..19} & M_{8..12,20} \\ \hline 13..16 & M_{13..16,1..7} & M_{13..16,8..12} & M_{13..16,13..16} & M_{13..16,17..19} & M_{13..16,20} \\ \hline 17..19 & M_{17..19,1..7} & M_{17..19,8..12} & M_{17..19,13..16} & M_{17..19,17..19} & M_{17..19,20} \\ \hline 20 & M_{20,1..7} & M_{20,8..12} & M_{20,13..16} & M_{20,17..19} & M_{20,20} \end{array} \right] \\
 \end{array} \\
 \rightarrow \mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{21} & r_{22} & r_{23} & r_{24} & r_{25} \\ r_{31} & r_{32} & r_{33} & r_{34} & r_{35} \\ r_{41} & r_{42} & r_{43} & r_{44} & r_{45} \\ r_{51} & r_{52} & r_{53} & r_{54} & r_{55} \end{bmatrix}
 \end{array}$$

**Figure 1.** Dividing the MJ matrix  $\mathbf{M}$  into a reduced matrix  $\mathbf{R}$ . The values in  $\mathbf{R}$ , i.e.,  $r_{ij}$ s, are to be computed to best represent the members of the group.

been developed (Miyazawa and Jernigan 1996; Mairov and Crippen 1992; Sippl 1995) to represent the residue-residue interactions in proteins. The one developed by Miyazawa and Jernigan (1996) on the basis of the frequency of inter-residue interactions observed in protein structures (MJ), is extensively used in the evaluation of the total energy of a protein sequence in a given structure. Hence, we have used the MJ matrix as basis for the matrix of reduced alphabet.

Let  $\mathbf{M}$  denote the MJ matrix of size  $20 \times 20$ . When a reduced alphabet of size  $m$  is identified, we can divide  $\mathbf{M}$  into  $m^2$  blocks as shown in figure 1 for the case of  $m = 5$ .

Figure 1 shows the MJ matrix separated into five groups with amino acids distributed as  $[\{C, M, F, I, L, V, W\}; \{Y, A, G, T, S\}; \{N, Q, D, E\}; \{H, R, K\}; \{P\}]$ . In this case we get 25 block matrices of different sizes. Since we want to reduce this to  $\mathbf{R}$  of size  $5 \times 5$ , each block matrix should have a single representative value for use in computations involving the reduced alphabet. Such a value is denoted as  $r_{ij}$  in figure 1. This is done using the least squares technique by solving the following minimization problem.

$$\text{Minimize}_{r_{ij}} f_{ij} = \sum_{i=a}^b \sum_{j=c}^d (M_{ij} - r_{ij})^2, \quad (3)$$

where the summation range-limits  $a, b, c$  and  $d$  are taken as per the indices  $i$  and  $j$  of  $r_{ij}$ . The value of  $r_{ij}$  that minimizes  $f_{ij}$  is the required representative value as it best represents the block matrix  $M_{a..b,c..d}$  in the sense of least squares; in fact, it is simply the average value the entries in  $M_{a..b,c..d}$ . The results of this method are presented in § 3.2.

### 3. Results

#### 3.1 Interpretation of the MDS map

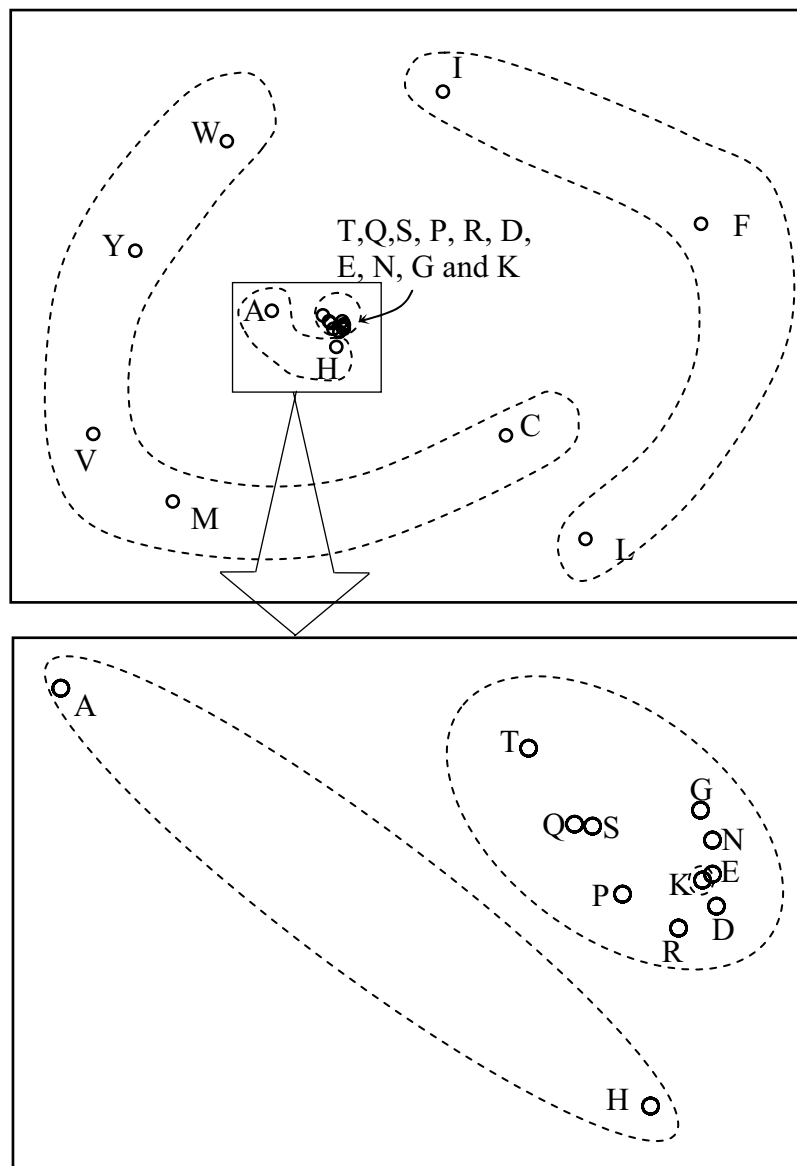
The solution of the problem in eq. (1) will give the best possible coordinates for all the 20 points corresponding to the 20 amino acids. This is shown in figure 2 where single-letter codes are indicated next to the points. As per the nature of this MDS map, similar amino acids cluster closer together than the dissimilar ones. The map shown in

figure 2 is not unique because the minimization problem in eq. (1) may have local minima. So, depending on the initial guess, slightly different maps will emerge. However, the differences among them are not significant. For example, the clustering seen in the centre of the plot is always seen while the ones around the boundary exchange places to some extent.

The visual presentation in figure 2 is immediately suggestive of how to group the amino acids both in terms of the number of groups and the numbers and identities of members in each group. In the map of figure 2, the central core can easily be mapped into two groups as shown. These two groups exactly match two groups of Cieplak *et al* (2001). The exception is *K*, which Cieplak *et al* (2001) put in a separate group alone. The remaining ones can be grouped in many ways. This is where further judgment and interpretation is needed. If we indeed follow the grouping of Cieplak *et al* (2001), it will be as shown in figure 2. Alternatively, one can use only the outer amino acids and create a separate 2D map only for those to see which among them are similar to each other. In any case, this map seems to agree with the conclusion of Wolynes (1997) who noted that three residue grouping is not adequate.

#### 3.2 Energies for the reduced amino acid alphabet

This method described for computing the reduced inter-residue energies was applied to various amino acid groupings with a group size of five. The energy values in the reduced MJ matrices for these cases are shown in table 2. Table 3 shows the total root mean square error ( $e$ ) and the combined correlation coefficient ( $\gamma$ ). It can be observed in table 3 that the correlation coefficients for all others except that based on chemical properties are quite good. The large values of total RMS for PAM and WAG based groupings of Koisol *et al* (2004) are due to isolation of one or two amino acids rather than grouping them with others in a large group. The total RMS value is the least for the grouping of Cieplak *et al* (2001) with a correlation coefficient that is



**Figure 2.** A 2D multidimensional scaling (MDS) map of amino acids based on the MJ matrix. The grouping shown with dashed envelopes is as per Cieplak *et al.* (2001). They separate K as the fifth group but this map shows it contained within a large group. Except this, the grouping suggested by the visual map agrees well with Cieplak *et al.*'s grouping

closest to unity. Referring to the MDS map of figure 2, this grouping emerges as the most natural. It can be explained as follows.

The MJ matrix, which is derived from observed propensities of interactions, incorporates various properties of amino acids. Hence, the reduced matrices based on a single property such as chemical or evolutionary features show high total RMS error with respect to the MJ matrix. The grouping of Cieplak *et al.* (2001) and MDS are based on the reduction of the MJ matrix by mathematical methods such as eigenvalue analysis and multidimensional scaling.

Thus, the correlation is the highest and the error is the least for Cieplak *et al.*'s (2001) grouping in table 3.

The energy values for the reduced matrix of two groups [H (hydrophobic) and P (polar)] are  $HH = -2.3998$ ,  $HP = -1.0$ , and  $PP = 0.0$  (RT units). [This had a total RMS error equal to 0.0685, which is lower than those reported for 5 types of grouping (table 3)]. The energy value for HH is often taken as -2.3 in the literature [e.g. Li *et al.* (1997)]. Thus, the value obtained here, i.e., -2.3998, validates the method presented here for computing the reduced interaction energy values of the reduced amino acid alphabet.

**Table 2.** Reduced MJ matrices of interaction energies (RT units) for the amino acid alphabet\*

(1)	(2)	(3)
$\begin{bmatrix} 1.72 & 3.27 & 3.51 & 3.50 & 2.83 \\ & 5.06 & 5.41 & 5.66 & 4.95 \\ & & 5.68 & 5.92 & 4.98 \\ & Sym. & & 6.22 & 5.32 \\ & & & & 5.55 \end{bmatrix}$	$\begin{bmatrix} 1.72 & 2.97 & 3.62 & 3.51 & 3.27 \\ & 5.22 & 5.69 & 5.21 & 5.06 \\ & & 6.47 & 6.08 & 5.82 \\ & Sym & & 5.68 & 5.41 \\ & & & & 5.06 \end{bmatrix}$	$\begin{bmatrix} 5.73 & 3.92 & 3.58 & 3.17 & 2.94 \\ & 2.46 & 2.12 & 1.83 & 1.87 \\ & & 1.93 & 1.55 & 1.35 \\ & Sym & & 1.38 & 1.72 \\ & & & & 1.04 \end{bmatrix}$
(4)	(5)	(6)
$\begin{bmatrix} 6.00 & 5.72 & 3.73 & 3.12 & 3.44 \\ & 5.47 & 3.55 & 3.06 & 3.43 \\ & & 2.10 & 1.66 & 1.77 \\ & Sym & & 1.43 & 1.77 \\ & & & & 1.53 \end{bmatrix}$	$\begin{bmatrix} 5.29 & 5.70 & 3.43 & 3.50 & 3.13 \\ & 6.22 & 3.67 & 3.81 & 3.29 \\ & & 2.24 & 2.02 & 1.60 \\ & Sym & & 1.99 & 1.68 \\ & & & & 1.65 \end{bmatrix}$	$\begin{bmatrix} 6.50 & 5.44 & 3.18 & 4.30 & 4.09 \\ & 4.81 & 2.85 & 3.59 & 3.55 \\ & & 1.44 & 1.68 & 1.75 \\ & Sym & & 2.72 & 2.45 \\ & & & & 2.41 \end{bmatrix}$
(7)	(1)→ [ {AGTSNQDEHRKP}, {W}, {YF}, {MIVL}, {C} ] (2)→ [ {AGTSNQDEHRKP}, {CV}, {IML}, {FY}, {W} ] (3)→ [ {CMIFLYWV}, {AHT}, {GP}, {QNRSK}, {DE} ] (4)→ [ {CMFI}, {LVWY}, {AGTS}, {NQDE}, {HPRK} ] (5)→ [ {CFYW}, {MLIV}, {G}, {PATS}, {NHQEDRK} ] (6)→ [ {IVL}, {FYWH}, {KRDE}, {GACS}, {TMNQP} ] (7)→ [ {LFI}, {MVWCY}, {HA}, {TGPRQSNED}, {K} ]	
$\begin{bmatrix} 7.04 & 5.97 & 4.62 & 3.81 & 3.24 \\ & 4.97 & 3.73 & 3.01 & 2.44 \\ & & 2.64 & 2.05 & 1.33 \\ & Sym & & 1.70 & 1.22 \\ & & & & 0.12 \end{bmatrix}$		

\* Seven schemes of classifying of amino acids into five groups (the grouping details are given in the bottom right box) have been chosen, corresponding references for all seven groups are given in table 3.

**Table 3.** The total RMS values and combined correlation coefficients of computed reduced matrices (shown in table 2)

S. No.	Groupings	Total RMS error, $e$	Correlation coefficient, $\gamma$
1	PAM matrix and Koisol <i>et al</i> (2004)	0.7067	0.92865
2	WAG matrix and Koisol <i>et al</i> (2004)	0.6637	0.93225
3	Wang and Wang (1999)	0.2226	0.94081
4	Wang and Wang (2002)	0.3411	0.92741
5	Li <i>et al</i> (2003)	0.3408	0.92924
6	Chemical properties	0.5339	0.75584
7	Cieplak <i>et al</i> (2001)	0.2072	0.97863

#### 4. Summary

Designing sequences for a chosen structure as well as predicting the structures for newly sequenced proteins have been challenging tasks and there has been continued interest in classifying amino acids and identifying the reduced alphabet to achieve these goals. Because there are numerous attempts made by several researchers who followed varied criteria, the information about the grouping of amino acids is rather scattered in the literature. In this paper, we have summarized various attempts so that they can be compared and contrasted. Here we noted that there is no agreement in the literature on the minimum number of groups (i.e. types) of amino acids to achieve protein-like folding behaviour. So, in this paper, we re-visited the application of MDS to create a 2D map of amino acids. The MDS map is known to be

effective in identifying similar entities in a large collection when similarity data among them is imprecise or if there are too many criteria to be considered for comparison. The grouping suggested by the 2D MDS map created using the MJ matrix agrees closely with the grouping suggested by Cieplak *et al* (2001). There are opportunities for further investigation along the lines of MDS to gain more insight into the nature of grouping of amino acids. In addition, we also note that none of the papers explicitly report the inter-group interaction energy values although this data is essential in computational techniques that work with reduced amino acid alphabet. We have proposed a method to compute such normalized interaction energies for any given reduced alphabet by solving a least-squares optimization problem. The total RMS error and the correlation coefficient of this calculation indicate how effective a particular grouping

is. We have used the MJ matrix as the starting point to compute the reduced interaction energy matrix. Such matrices are computed for several groupings reported in the literature. Thus, the interaction energy values reported in this study can be effectively used in sequence design or structure prediction using reduced alphabet of amino acids.

### Acknowledgments

The help of Mr Sourav Rakshit in writing the computer program that was used to create the MDS map is gratefully acknowledged.

### References

- Cannata N, Toppo S, Romualdi C and Valle G 2002 Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices; *Bioinformatics* **18** 1102–1108
- Chan S H 1999 Folding alphabets; *Nat. Struct. Biol.* **6** 994–996
- Cieplak M, Holter S N, Maritan A and Banavar R J 2001 Amino acid classes and the protein folding problem; *J. Chem. Phys.* **114** 1420–1423
- Dayhoff M O, Schwartz R M and Orcutt B C 1978 A model of evolutionary change in proteins; *Atlas Protein Sequence Struct.* **5** 345–352
- Fan K and Wang W 2003 What is the minimum number of letters required to fold a protein; *J. Mol. Biol.* **328** 921–926
- French S and Robson B 1983 What is Conservative Substitution?; *J. Mol. Evol.* **19** 171–175
- Hansson T, Oostenbrink C and van Gunsteren W 2002 Molecular dynamics simulations; *Curr. Opin. Struct. Biol.* **12** 190–196
- Henikoff S and Henikoff J G 1992 Amino acid substitution matrices from protein blocks; *Proc. Natl. Acad. Sci. USA* **89** 10915–10919
- Jones D T, Taylor W R and Thornton J M 1992 The rapid generation of mutation data matrices from protein sequences; *Comput. Appl. Biosci.* **8** 275–282
- Karplus M and McCammon J A 2002 Molecular dynamics simulations of biomolecules; *Nat. Struct. Biol.* **9** 646–652
- Kosiol C, Goldman N and Buttimore H N 2004 A new criterion and method for amino acid classification; *J. Theor. Biol.* **228** 97–106
- Kruskal J B and Wish M 1978 *Multidimensional scaling* (Sage Publications)
- Levitt M 1976 A Simplified representation of protein conformations for rapid simulation of protein folding; *J. Mol. Biol.* **104** 59–107
- Li H, Tang C and Wingreen N S 1997 Nature of driving force for protein folding: A result from analyzing the statistical potential; *Phys. Rev. Lett.* **79** 765–768
- Li T, Fan K, Wang J and Wang W 2003 Reduction of protein sequence complexity by residue grouping; *Protein Eng.* **16** 323–330
- Mairov V N and Crippen G M 1992 Contact potential that recognizes the correct folding of globular proteins; *J. Mol. Biol.* **227** 876–888
- Miyazawa S and Jernigan R 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** 623–644
- Murphy R L, Wallqvist A and Levy M R 2000 Simplified amino acid alphabets for protein fold recognition and implications for folding; *Protein Eng.* **13** 149–152
- Regan L and DeGrado W F 1988 Characterization of a helical protein designed from first principles; *Science* **241** 976–978
- Riddle S D, Santiago J V, Bray-Hall S T, Doshi N, Grantcharova V P, Yi Q and Baker D 1997 Functional rapidly folding proteins from simplified amino acid sequences; *Nat. Struct. Biol.* **4** 805–809
- Schwartz R M and Dayhoff M O 1978 Matrices for detecting distant relationships; *Atlas Protein Struct.* **5** 353–358
- Sippl M J 1995 Knowledge-based potentials for proteins; *Curr. Opin. Struct. Biol.* **5** 229–235
- Taylor R W 1986 The classification of amino acid conservation; *J. Theor. Biol.* **119** 205–218
- Wang W, Donini O, Reyes C M and Kollman P A 2001 **BIOMOLECULAR SIMULATIONS**: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions; *Annu. Rev. Biophys. Biomol. Struct.* **20** 211–243
- Wang J and Wang W 1999 A computational approach to simplifying the protein folding alphabet; *Nat. Struct. Biol.* **6** 1033–1038
- Wang J and Wang W 2002 Grouping of residue based on their contact interactions; *Phys. Rev. E* **65** 041911–5
- Whelan S and Goldman N 2001 A General empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach; *J. Mol. Evol.* **18** 691–699
- Wolynes G P 1997 As simple as can be?; *Nat. Struct. Biol.* **4** 871–874

ePublication: 16 June 2007