Eur Biophys J (2007) 37:105–109
DOI 10.1007/s00249-007-0182-y

BIOPHYSICS LETTER

Amino acid partitioning using a Fiedler vector model

S. J. Shepherd · C. B. Beggs · S. Jones

Received: 21 December 2006 / Revised: 23 March 2007 / Accepted: 7 May 2007 / Published online: 4 July 2007

Abstract This paper presents a new Fiedler vector model for categorising amino acids, which is based on the Miyazawa-Jernigan matrix. The model splits the amino acid residues into two hydrophobic groups (LFI) and (MVWCY) and two polar groups (HATGP) and (RQSNEDK). In so doing, it independently confirms the findings of Wang and Wang and Cieplak et al. and demonstrates the validity of using eigenvectors to partition amino acid groups.

Keywords Amino acids · Residues · Partitioning · Eigenvector · Fiedler vector · Peptide

Introduction

The heterogeneity of the 20 types of amino acid residue introduces much complexity into the protein folding process, making simulation and prediction extremely difficult. However, despite this apparent complexity, there is growing evidence that folding behaviour is dominated by generic characteristics of the amino acids, such as hydrophobicity (Li and Liang 2007). This has led a number of researchers (Chan 1999; Li et al. 1997; Wang and Wang 1999) to identify simplified amino acid alphabets with a view to determining the minimal number of residue types required to form native proteins. This is of importance because it has been demonstrated that small proteins, such as the SH3 domain can be largely built using just five amino acid types (Riddle et al. 1997). Indeed, despite dramatic changes in sequence due to amino acid substitution, it has been shown that these reduced 'alphabet' proteins still folded in the same way as the naturally occurring 'wild-type' SH3 domain (Alm and Baker 1999; Riddle et al. 1997). Therefore, much effort has been expended developing 'minimalist' models, in which the 20 residue types are grouped into reduced sets according to similarities in their physical and chemical properties (Chan 1999; Cieplak et al. 2001; Esteve and Falceto 2005; Li et al. 1997; Wang and Wang 1999). Notable amongst these are Wang and Wang (1999), who used a clustering approach, which utilized a matrix mismatch minimisation method to divide the residues into several groups, each with different chemical and physical properties. In this paper we present an alternative method based on the Miyazawa-Jernigan (MJ) matrix (Miyazawa and Jernigan 1996), which produces results similar to those of Wang and Wang (1999) and Cieplak et al. (2001). In so doing, we independently confirm the findings of Wang and Wang and Cieplak et al. and demonstrate the validity of using eigenvectors to partition amino acid groups. The reduction algorithm described in this paper demonstrates that amino acids can be unambiguously divided into clearly defined groups, each possessing different characteristics.

Electronic supplementary material The online version of this article (doi:10.1007/s00249-007-0182-y) contains supplementary material, which is available to authorized users.

S. J. Shepherd · C. B. Beggs (✉)
Medical Biophysics Group, School of Engineering, Design and Technology, University of Bradford, BD7 1DP Bradford, UK
e-mail: c.b.beggs@bradford.ac.uk

S. J. Shepherd
e-mail: S.J.Shepherd@Bradford.ac.uk

S. Jones
Division of Biomedical Sciences, School of Life Sciences, University of Bradford, BD7 1DP Bradford, UK
e-mail: S.Jones8@Bradford.ac.uk

Miyazawa–Jernigan matrix

Proteins fold into specific three-dimensional structures which accord with the minimum free energy of their respective polypeptide sequence (Li et al. 1997). This native structure is dictated by the physical interactions between the various amino acids in the sequence. However, because proteins contain thousands of atoms, which interact with huge numbers of water molecules, it is extremely difficult to calculate the free energy function from first principles (Li et al. 1997). Therefore, if the three-dimensional structure of proteins is to be predicted, an alternative approach is required. One such approach, substitutes real interactions between atoms with general interaction potentials between amino acids based on the frequency of contact of the two amino acids in reference proteins (Esteve and Falceto 2005). A classic example of an approach using statistical potentials is the MJ matrix (Miyazawa and Jernigan 1996), which is a 20 x 20 matrix of inter-residue contact energies between different types of amino acids. This matrix tabulates the 'so-called' statistical potential—a measure of the probability of having a given pair of amino acids close to each other in the native state. It has been widely applied in protein design and folding applications, and produces results, which appear robust (Li et al. 1997; Li and Liang 2007).

Amino acid grouping

Previous researchers have grouped amino acid residues using a variety of methodologies. For example, Wang and Wang (1999) used the MJ matrix to minimize the intra-group interactions and maximize inter-group interactions. Using this method they divided the 20 amino acids into one hydrophobic group of residues (CMFILVWY) and four polar groups (AHT), (GP), (DE) and (NQRKS). By comparison Cieplak et al. (2001), used the Manhattan distance metric of the MJ matrix to divide the amino acid residues into two hydrophobic groups (LFI) and (MVWCY) (i.e. H_1 and H_2) and three polar groups (HA), (TGPRQSNED) and (K) (i.e. P_1, P_2, and P_3), as shown in Fig. 1.

In their model Cieplak et al. (2001) used reduction algorithms based on 'distance' measures, such as the Euclidean distance R_{ij} between amino acids i and j, and the Manhattan distance. The quantity R_{ij} is a measure of the fidelity of substitution of one amino acid by the other and is defined as $R_{ij}^2 = \Sigma_k \ (M_{ik} - M_{jk})^2$, whereas the Manhattan distance involves the sum of the absolute measures $(M_{ik} - M_{jk})$. In general, these algorithms select for amino acids that are separated by the shortest distance and combine them into one group. The effective coupling of new groups with other groups is obtained simply by taking an arithmetic average over the individual component amino acid interactions. While this methodology has advantages over that used by Wang and Wang (1999), it can be seen in Fig. 1 that the only clear demarcation is that between the H_2 and P_1 groups, with the other partitions being rather vague and subjective.

Fiedler vector partitioning model

In order to gain a clearer understand of the partitioning problem we developed a Fiedler vector model based on the MJ matrix, which grouped the amino acids into well-defined categories. We computed iteratively successive Fiedler vectors for the partitioned groups, starting with the complete un-partitioned amino acid residue set—hence we first split the set in two, then in four and finally in eight. In the model, the rank-1 approximation of the MJ matrix given by Cieplak et al. (2001) was used. This formulation is as accurate as the more complex expression presented in Eq. 2 of Li et al's paper (Li et al. 1997), but is easier to use computationally. Its use was justified because the dominant eigenvalue is much larger than the secondary and subsequent eigenvalues, suggesting that the contribution of other components is minimal.

In the eigensystem of the full MJ matrix, each eigenvalue represents an amount of energy λ_i, which is distributed among the different residues according to the corresponding eigenvector e_i. However, in our reduced model, there is, by construction, only one eigenvalue and one eigenvector. This approximation reduces a complex, multi-body interaction to a simplified 'charge-based' system describing the relative attraction between the various residues. Accordingly, the adjacency matrix A, from the MJ vector v is formed in the conventional manner (Cieplak et al. 2001). Having found A, it is then possible to compute the Laplacian L (Koren and Harel 2003). There is an intimate relationship between the combinatorial and topological characteristics of a graph and the algebraic properties of its Laplacian. The idea at the heart of spectral graph theory is that there is a direct connection between the *spectrum* of

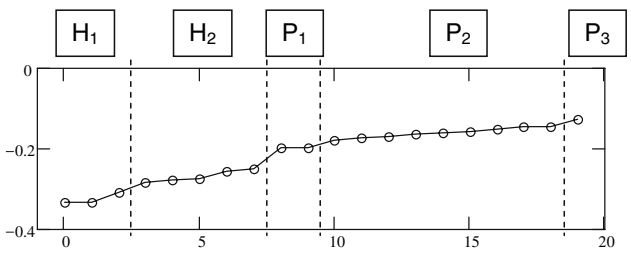

Fig. 1 The principal eigenvector of the Miyazawa–Jernigan matrix showing the partitions between the five amino acid groups identified by Cieplak et al. (2001)

the Laplacian and the *isoperimetric number* of the graph. Furthermore, a connected graph having a large subdominant eigenvalue (relative to the dominant eigenvalue) can be separated into two sets of vertices such that the two induced sub-graphs have a high degree of connectivity. In many cases, this is the maximum connectivity, that is, the graph has been 'cut' optimally via its 'weakest' links. Fiedler (1973) showed that this cut is given by the signum of the second smallest eigenvector of L, namely the *Fiedler vector*. That is, those nodes, which are members of one partition are denoted by the corresponding elements of the Fiedler vector whose signs are positive, and those nodes which are members of the other partition are denoted by the corresponding elements of the Fiedler vector whose signs are negative.

Results of the partitioning study

The results obtained using our amino acid partitioning model, are presented in Figs. 2, 3 and 4. Figure 2 shows that the first spectral partition splits the amino acids into a hydrophobic group (LFIMVWCY) and a polar group (HATGPRQSNEDK). The second spectral partition of each subgroup yields the split (LFI) and (MVWCY) for the hydrophobic group and the split (HATGP) and (RQSNEDK) for the polar group (see Figs. 3, 4, respectively). It should be noted that the splitting of each group can be continued in this manner until the Fiedler vectors of the new sub-groups become essentially flat, indicating that the sub-groups genuinely represent residues of different character and that partitioning should stop at that point.

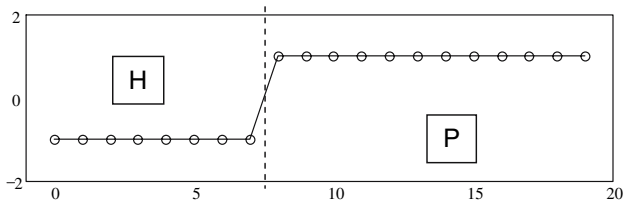

Fig. 2 The signum of the Fiedler vector of the Laplacian of the Miyazawa–Jernigan matrix showing the unambiguous partitioning of the hydrophobic (L F I M V W C Y) and the polar (H A T G P R Q S N E D K) amino acid groups

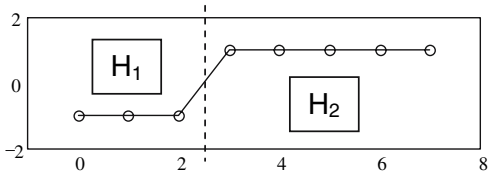

Fig. 3 The signum of the Fiedler vector of the Laplacian of the hydrophobic subset matrix showing the unambiguous partitioning of the (L F I) (i.e. H_1) and (M V W C Y) (i.e. H_2) groups

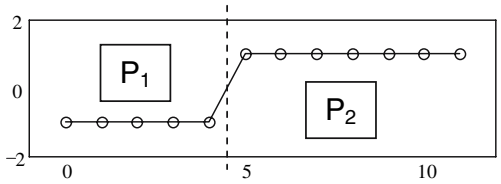

Fig. 4 The Fiedler vector of the Laplacian of the polar subset matrix showing the unambiguous partitioning of the (H A T G P) (i.e. P_1) and (R Q S N E D K) (i.e. P_2) groups

When the results in Figs. 2, 3 and 4 are compared with those of Cieplak et al (2001) presented in Fig. 1, it can be seen that our model produces demarcations between the various groups, which are clear and unambiguous, whereas only the demarcation between the H_2 and P_1 is clearly defined in Fig. 1. If the Fiedler vector model is applied to groups H_1, H_2, P_1 and P_2 in Figs. 3 and 4, then the residues can be further split into eight groups (LF), (I), (MVW), (CY), (HAT), (GP), (RQS) and (NEDK).

Discussion

Construction of reduced alphabets by grouping together amino acids on the basis of the values of their physical and chemical properties have proven to be useful in the study of proteins in a number of different contexts. For instance, reduced amino acid alphabets have been shown to be useful in studies of protein folding and protein recognition (Chan 1999; Murphy et al. 2000) and in protein design (Riddle et al. 1997). In particular, protein design experiments have shown that the use of specific subsets of amino acids can produce foldable proteins. For example, Schafmeister et al. (1997) demonstrated that a 108 amino acid, 4 helix bundle protein could be synthesized using a reduced alphabet of just 7 amino acids. This finding reinforces those of Riddle et al. (1997) who showed that a functional β-sheet protein (SH3 domain) could be largely encoded by a reduced set of five amino acids. In addition, the use of a reduced amino acid alphabet is advantageous, in so far that it facilitates better computational analysis of proteins because it dramatically reduces the computing power required.

Given that foldable proteins have been produced from reduced amino acid sets, it prompts the question of whether there is a minimal amino acid alphabet, which could be used to fold all proteins. From the results of simplified lattice model simulations, it has been postulated that a minimum of three different amino acid types is required for protein folding (Murphy et al. 2000). However, Riddle et al. (1997) found that three amino acids was too few a number to promote folding and that a minimum of five different residues were required. Likewise, Wang and Wang (1999) suggested that a minimum of five amino

acids, one from each of the groups identified in Table 1, is required for successful folding to occur. However, the true minimal alphabet may require additional complexity in order to create the large number of protein fold types actually observed in nature. Murphy et al. (2000) therefore estimate that foldable sequences for most proteins can be represented using a reduced alphabet containing 10–12 amino acids. This conclusion is based on the observation that there is little loss of the information necessary to pick out structural homologs in a clustered protein sequence database when the amino acid alphabet is reduced from 20 to 10 letters, whereas the information is rapidly degraded when further reductions in the alphabet are made. Likewise, Romero et al. (1999) using information-theoretic arguments, show that the minimal alphabet size necessary for specifying globular proteins that occur in nature is 10.

Table 1 summarises the residue groupings determined by Wang and Wang (1999), Cieplak et al. (2001) and ourselves. The data in Table 1 reveals that our results are similar to those of both Wang and Wang and Cieplak et al, with all three sets of results agreeing that a major division exists between the hydrophobic (LFIMVWCY) and polar (HATGPRQSNEDK) groups of residues. However, with regard to the hydrophobic residues our results concur with those of Cieplak et al. rather than Wang and Wang, and suggest that two main hydrophobic groups exist, (LFI) and (MVWCY), which can be further divided into four sub-groups. By contrast, with regard to the polar residues, our results are closer to those of Wang and Wang than those of Cieplak et al. We divided the polar residues into two main groups, (HATGP) and (RQSNEDK), which were further divided into four sub-groups (HAT) (GP) (RQS) and (NEDK); divisions which although not identical to Wang and Wang's are nevertheless similar.

Interestingly, the results of Wang and Wang, Cieplak et al. and ourselves all agree that glycine (G), alanine (A) and proline (P) (all generally considered to be aliphatic hydrophobic residues) should be classified with the polar residues. In particular, all three-research teams grouped G and P together and grouped A with positively charged histidine (H). Similarly, all three-research teams collated the negatively charged residues aspartate (D) and glutamate (E) in the same group.

Collectively, the results presented in this paper show that Fielder vector model can be used to efficiently divide the amino acid set into unambiguous groups with similar characteristics. Although our results are not identical to those of Wang and Wang (1999) and Cieplak et al. (2001) they are similar, and reinforce their findings.

Table 1 Summary of amino acid partitioning results

Model	Hydrophobic residues	Polar residues
Wang and Wang	(FLIMVWCY)	(HAT) (GP) (DE) (NQRKS)
Cieplak et al.	(LFI) (MVWCY)	(HA) (TGPRQSNED) (K)
Fiedler vector (4 groups)	(LFI) (MVWCY)	(HATGP) (RQSNEDK)
Fiedler vector (8 groups)	(LF) (I) (MVW) (CY)	(HAT) (GP) (RQS) (NEDK)

Post-script

As a post-script we draw the readers attention to the online amino acid website *AAindex* (http://www.genome.ad.jp/dbget/aaindex.html) (Kawashima et al. 1999), which at this time of writing contains a highly comprehensive listing of 516 different amino acid interaction matrices that have been devised for various different purposes. Some like the MJ matrix are quite general in their scope, while others offer higher accuracy but only for specific fold conformations. The reader should be aware that the technique presented in this paper is generic in nature and can be applied equally effectively to any of these 516 matrices.

References

Alm E, Baker D (1999) Matching theory and experiment in protein folding. Curr Opin Struct Biol 9:189–196
Chan HS (1999) Folding alphabets. Nat Struct Biol 6:994–996
Cieplak M, Holter NS, Maritan A, Banavar JR (2001) Amino acid classes and the protein folding problem. J Chem Phys 114:1420–1423
Esteve JG, Falceto F (2005) Classification of amino acids induced by their associated matrices. Biophys Chem 115:177–180
Fiedler M (1973) Algebraic connectivity of graphs. Czeck Math J 23:298–305
Kawashima S, Ogata H, Kanehisa M (1999) AAindex: amino acid index database. Nucleic Acids Res 27:368–369
Koren Y, Harel D (2003) A two-way visualization method for clustered data proceedings of the 9th ACM international conference on knowledge discovery and data mining (KDD'03). ACM press, New York, pp 589–594
Li X, Liang J (2007) Knowledge-based energy functions for computational studies of proteins. In: Xu Y, Xu D, Liang J (eds) Computational methods for protein structure prediction and modelling, vol 1, basic characterization. Springer, Heidelberg
Li H, Tang C, Wingreen NS (1997) Nature of driving force for protein folding—a result from analysing the statistical potential Phys. Rev Lett 79:765–768
Miyazawa S, Jernigan RL (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 LR, Wallqvist A, Levy RM (2000) Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 13:149–152

Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D (1997) Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 4:805–809

Romero P, Obradovic Z, Dunker AK (1999) Folding minimal sequences: the lower bound for sequence complexity of globular proteins. FEBS Lett 462:363–367

Schafmeister CE, LaPorte SL, Miercke LJ, Stroud RM (1997) A designed four helix bundle protein with native-like structure. Nat Struct Biol 4:1039–1046

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