PROTEINS: Structure, Function, and Genetics 43:103–112 (2001)

Clusters in α/β Barrel Proteins: Implications for Protein Structure, Function, and Folding: A Graph Theoretical Approach

N. Kannan,[1] S. Selvaraj,[2,3] M. Michael Gromiha,[2] and S. Vishveshwara[1]*
[1]*Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India*
[2]*RIKEN Tsukuba Institute, Institute of Physical and Chemical Research (RIKEN), Tsukuba, Ibaraki, Japan*
[3]*Department of Physics, Bharathidasan University, Tiruchirapalli, India*

ABSTRACT **The α/β barrel fold is adopted by most enzymes performing a variety of catalytic reactions, but with very low sequence similarity. In order to understand the stabilizing interactions important in maintaining the α/β barrel fold, we have identified residue clusters in a dataset of 36 α/β barrel proteins that have less than 10% sequence identity within themselves. A graph theoretical algorithm is used to identify backbone clusters. This approach uses the global information of the nonbonded interaction in the α/β barrel fold for the clustering procedure. The nonbonded interactions are represented mathematically in the form of an adjacency matrix. On diagonalizing the adjacency matrix, clusters and cluster centers are obtained from the highest eigenvalue and its corresponding vector components. Residue clusters are identified in the strand regions forming the β barrel and are topologically conserved in all 36 proteins studied. The residues forming the cluster in each of the α/β protein are also conserved among the sequences belonging to the same family. The cluster centers are found to occur in the middle of the strands or in the C-terminal of the strands. In most cases, the residues forming the clusters are part of the active site or are located close to the active site. The folding nucleus of the α/β fold is predicted based on hydrophobicity index evaluation of residues and identification of cluster centers. The predicted nucleation sites are found to occur mostly in the middle of the strands.**

Key words: clusters; graph theory; eigenvalues; eigenvectors; protein folding; nucleation sites; active site

INTRODUCTION

The α/β barrel fold was first discovered in the structure of triose phosphate isomerase (TIM); this fold is now found to be one of the largest occurring folds in the three superfamilies of eukaryotes, bacteria, and archaea.[1] Nearly 10% of the enzyme structures solved so far are known to possess this fold. Different enzymes with very low sequence homology that perform a variety of functions are known to adopt the TIM barrel fold, forming one of the largest structural superfamilies. The TIM barrel fold appears to be engineered in such a way that it is tolerant to very high sequence variation and on which different functionalities can be designed.[2]

A large number of investigations carried out for the past two decades, in an effort to understand the structural and functional principles underlying the TIM barrel fold, have recently been reviewed.[3] Although we have a good understanding of the functional aspects related to this fold, recently demonstrated by converting an enzyme phosphoribosylanthranilate with an α/β barrel fold to another α/β barrel protein indole-3-glycerol-phosphate synthase with a completely different function,[4] we still do not have a complete understanding of the stabilizing forces underlying the α/β barrel fold.

Previous studies addressing this question concluded that the packing of side-chains within the closed β barrel to be one of the most important factors in stabilizing the TIM barrel fold[5–7]; Lesk et al.[5] also described the specific arrangement of side-chains and their interactions in the β barrel region in three different enzymes. These investigators concluded that of the three enzymes studied, TIM and *Rhodospirillum rubrum* ribulose 1,5-bisphosphate carboxyoxygenase (RUBISCO) showed a different pattern of side-chain packing as compared with the enzyme glycolate oxidase (GOX). This difference in the side-chain packing was attributed to the evolution of these proteins. Clusters of amino acids in a protein structure are known to be important in stabilizing the protein fold.[8] Selvaraj and Gromiha[9] recently analyzed the amino acid clustering pattern in 36 α/β proteins and identified physicochemically similar clusters in pairs of α/β proteins.

The present study adopts a graph theoretical approach to identify residue clusters in a data set of 36 TIM barrel proteins. Techniques derived from graph theory have been used to compare protein structures and to identify specific side-chain patterns.[10,11] Graph theoretical concepts have also been useful in comparing secondary structures,[12] sheet topologies,[13] and protein structure modeling.[14] We recently showed the efficacy of a graph spectral method in

* Indian Institute of Science, Bangalore 560 012, India. E-mail: sv@mbu.iisc.ernet.in

identifying conserved residue clusters among structurally similar proteins with very low sequence homology[15] and in identifying topologically similar backbone clusters in a family of protein structures.[16] The advantage of a graph spectral method is that it uses the global information of nonbonded interactions in the protein molecule to identify clusters. The nonbonded interaction is represented mathematically in the form of a connected graph; clusters and cluster centers are obtained from the eigenvalues and eigenvectors of the constructed graph (see Materials and Methods).

This procedure has been used in the present study to identify backbone clusters in a dataset of 36 α/β barrel proteins. Clusters in the β barrel region were identified in all the proteins. The identified clusters are topologically conserved across protein families and sequentially conserved within their own family, emphasizing the fact that the identified clusters are crucial for stabilizing the TIM barrel fold. Residues that form the active site of the protein are also found to occur as part of clusters. Nucleation sites that are important from the folding perspective of the α/β barrel fold are predicted.

MATERIALS AND METHODS

Data Set

A representative data set of 36 $(\alpha/\beta)_8$ barrel proteins of resolution >2.5 Å and the average sequence identity less than 10% between any two proteins considered for the present study. The Protein Data Bank (PDB) codes of the 36 selected proteins are 1a0c, 1a3x, 1ad1, 1ado, 1ads, 1b4k, 1b54, 1b5t, 1bf6, 1bpl, 1bqc, 1btm, 1c0d, 1ceo, 1cnv, 1dos, 1edg, 1juk, 1luc, 1nar, 1pdy, 1pym, 1qap, 1qat, 1qtw, 1smd, 1tax, 1tml, 1ttp, 2ebn, 2myr, 2tmd, 2tps, 5ptd, 7a3h, and 8ruc and were obtained from the recent release of the PDB.[17]

Identification of Clusters by a Graph Theoretical Approach

Residue clusters in the protein structures adopting the α/β barrel fold are identified by a novel graph theoretical approach. The protein molecule is represented in the form of connected graph wherein the C_α atoms of the residues from the nodes of the graph and all the C_α atoms that occur within a spatial distance of 6.5 Å to a given $C_{\alpha i}$ are connected by an edge corresponding to an edge weight of 1. The two sequence neighbors on either side of the $C_{\alpha i}$ atom, i.e., ($C_{\alpha i-2}$, $C_{\alpha i-1}$, $C_{\alpha i+1}$, and $C_{\alpha i+2}$) are not considered in assigning the edge weight. This way of constructing the graph represents the nonlocal interactions of the molecule in the most appropriate way. The connected graph is shown for the protein molecule indole-3-glycerol phosphate synthase (1juk) in Figure 1.

This connected nonbonded representation can be represented mathematically in the form of an adjacency matrix A_{ij}, where the matrix elements are given by

$A_{ij} = 1$ if $i \neq j$ (i and j are within a distance of 6.5 Å)

$A_{ij} = 0$ if $i = j$

Matrix A is of the order of $n \times n$ if there are n vertices in the graph or n residues in the protein structure.

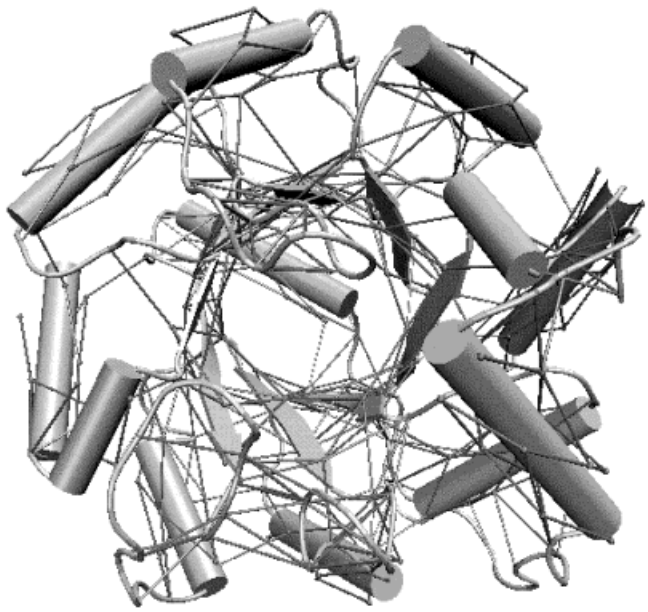

Fig. 1. Connected graph representation for the protein molecule 1juk.

Diagonalization of matrix and obtaining the eigenvalues and eigenvector components

The general eigenvalue equation for the adjacency matrix A is given by $(A - \lambda I)X = 0$, where I is the identity matrix, X is the eigenvector, and λs is the characteristic eigenvalue of the equation. On diagonalizing the matrix A, n eigenvalues and eigenvectors are obtained. The eigenvalues as well as their corresponding vector component magnitudes are sorted. The highest among the sorted eigenvalues and its corresponding vector component magnitude are considered. The residues with high vector component magnitude form a cluster,[16] and the residue with the largest vector component magnitude forms the cluster center.

We can also obtain the number of edges connected to a vertex (degree) by eigenvalue analysis. The highest eigenvalue has been used as a branching parameter[18] and has been used in deriving many physical and chemical properties of saturated hydrocarbons. The boundary limit of the highest eigenvalue of graph is $D_{\min} \leq \lambda_{\max} \leq D_{\max}$, where $D_{\max}$ and $D_{\min}$ are the maximum and minimum degrees of the graph, respectively. The present analysis of the graphs on protein structure is designed to elucidate the nature of nonbonded interactions, where the maximum branching at a vertex ranges from four to nine,[19] and the residues with the largest branching correspond approximately to the stabilization centers defined by Dosztanyi et al.[20] In the graph theoretical algorithm presented, vector components corresponding to the highest eigenvalue correspond to such stabilization centers.

On diagonalizing the adjacency matrix, the highest eigenvalue and the corresponding vector components are known to contain information regarding the clustering of nodes in the graph, and hence the clustering of residues in the protein structure.[16,21,22]

TABLE I. First Three Highest Eigenvalues and Their Corresponding Vector Components of the Protein Molecule Indole-3-Glycerol Phosphate Synthase[†]

Residues forming the vector component	5.323[a] Vector component magnitude	Residues forming the vector component	4.932[a] Vector component magnitude	Residues forming the vector component	4.239[a] Vector component magnitude
81(S,SM)(S2)	0.262	178(G,SM)(S6)	0.280	49(I,SM)(S1)	0.286
178(G,SM)(S6)	0.224	81(S,SM)(S2)	0.236	109(M,SC)(S3)	0.215
109(M,SC)(S3)	0.218	208(V,SM)(S7)	0.235	231(L,SM)(S8)	0.212
49(I,SM)(S1)	0.215	179(I,SC)(S6)	0.218	131(L,SM)(S4)	0.207
108(L,SM)(S3)	0.210	108(L,SM)(S3)	0.211	230(F,SM)(S8)	0.206
80(L,SM)(S2)	0.185	109(M,SC)(S3)	0.200	157(L,SM)(S5)	0.205
208(V,SM)(S7)	0.184	177(I,SM)(S6)	0.198	48(I,SM)(S1)	0.201
230(F,SM)(S8)	0.171	159(E,SC)(S5)	0.195	77(A,T)(S2)	0.190
50(A,SM)(S1)	0.169	207(K,SM)(S7)	0.190	50(A,SM)(S1)	0.188
110(K,SC)(S3)	0.167	160(I,SC)(S5)	0.167	158(I,SC)(S5)	0.180
157(L,SM)(S5)	0.164	82(I,SC)(S2)	0.167	130(V,SM)(S4)	0.176
82(I,SC)(S2)	0.162	210(E,SC)(S7)	0.166	110(K,SC)(S3)	0.163
179(I,SC)(S6)	0.158	209(A,SM)(S7)	0.163	80(L,SM)(S2)	0.161
177(I,SM)(S6)	0.157	80(L,SM)(S2)	0.161	78(V,T)(S2)	0.150
130(V,SM)(S4)	0.157	158(I,SC)(S5)	0.150	47(A,SN)(S1)	0.148
209(A,SC)(S7)	0.157	107(I,SM)(S3)	0.150	83(L,SC)(S2)	0.148
210(E,SC)(S7)	0.154	51(E,SC)(S1)	0.146	177(I,SM)(S6)	0.146
231(L,SM)(S8)	0.153	229(A,SN)(S8)	0.143	176(F,SN)(S6)	0.137
51(E,SC)(S1)	0.151	127(A,T)(T)	0.142	232(I,SC)(S8)	0.137
107(I,SM)(S3)	0.147	110(K,SC)(S3)	0.141	122(A,HM)(HM)	0.135
158(I,SC)(S5)	0.145	128(D,T)(T)	0.129	132(L,SM)(S4)	0.123
131(L,SM)(S4)	0.145	52(Y,SC)(S1)	0.123	79(G,SN)(S2)	0.122
207(K,SM)(S3)	0.141	230(F,SM)(S8)	0.122	106(P,SN)(S3)	0.119
159(E,SC)(S5)	0.139	176(F,SN)(S6)	0.117	156(P,SN)(S5)	0.114
127(A,T)(T)	0.131	83(L,SC)(S2)	0.116	209(A,SC)(S7)	0.112
229(A,SN)(S8)	0.128	157(L,SM)(S5)	0.115	210(E,SC)(S7)	0.109

[†]All entries in columns 1, 3, and 5 refer to residue number (residue name, location in the secondary structure (S, strand; H, helix; SM, middle of strand; SN, N-terminal of strand; SC, C-terminal of strand; T, turn region; strand number in the β barrel).
[a]Eigenvalue.

Advantage of a graph theoretical approach

Conservation of residues in protein sequences belonging to the same family of the α/β barrel fold is usually obtained by a sequence alignment procedure.[23] However, a simple sequence alignment procedure between protein analogues (sequences belonging to same fold but to a different family) will not work in identifying conserved features because of the possibility of multi-amino acid correlated mutations; therefore, sophisticated procedures such as a conservation of conservatism (COC) approach have been reported.[24] In the present graph theoretical approach, the conserved residues in the sequences emerge as a consequence of the analysis of the structures, rather than by sequence alignment procedures. Using a graph theoretical approach, the identified cluster residues happen to be conserved regions of protein sequences.

The major advantage of a graph theoretical approach over other methods of identifying clusters is that the global nonbonded interaction of the protein molecule is taken into account for the clustering procedure. The algorithm devised in the present study is computationally efficient, as a single numeric computation is sufficient to identify clusters. Moreover, the specific contribution of each residue (in terms of interaction of the residue with other residues in the cluster) is obtained from the vector component magnitudes corresponding to the highest eigenvalue. This aspect is demonstrated by an example. The top three eigenvalues and their corresponding vector components are shown for the protein molecule (1juk) in Table I. The highest eigenvalue of the graph is 5.323, and the corresponding vector component magnitude is given in the second column. Most of the nodes (residues) that have the largest vector component magnitudes are found to occur in the strand regions that form the β barrel (column 1). The largest vector component magnitude of 0.262 (column 2) is found for the serine 81 residue, which occurs in the middle of the strand 2, denoted 81 (S,SM) (S2) in the first column of Table I. Similarly, most of the other residues that form the top vector components of the largest eigenvalue are found to emanate from different strands S6, S3, S1, S7, S5, and S8, and the residues are found to cluster in barrel region in the native structure of the protein (Fig. 2). The vector components of the next two highest eigenvalues (4.932 and 4.239) are also found to occur in the β-strand regions. Because the highest eigenvalue of a connected graph is an indicator of the extent of branching in the graph,[21] in the present case we find that the residues corresponding to the vector components of the top three eigenvalues are found to occur mostly in the strand region forming the barrel.

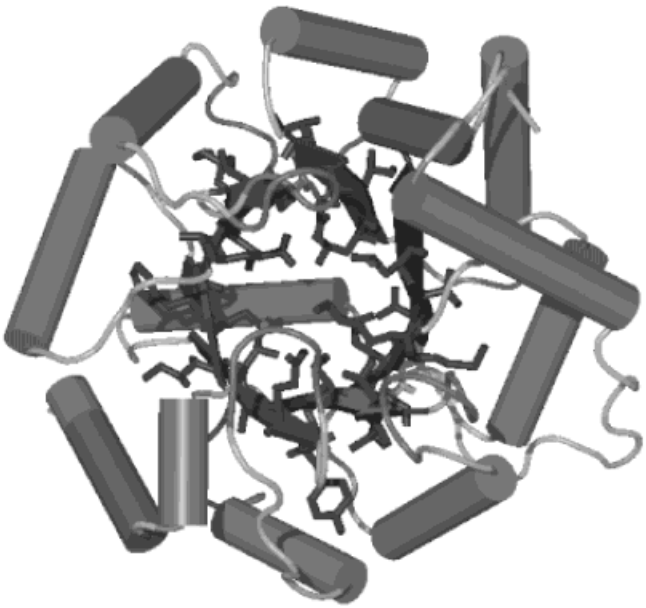

Fig. 2. Identified clusters occurring in the β barrel region. Cluster residues are shown in bond representation.[41]

Calculating Hydrophobicity Index and Nucleation Sites

The amino acid residues in a given structure are assigned hydrophobicity values based on the Nozaki–Tanford–Jones hydrophobicity index.[25,26] The surrounding hydrophobicity of a residue is defined by the sum of the hydrophobicity indices of various residues that appear within a distance of 8-Å radius from the given residue.[27] The residues having surrounding hydrophobicity value equal to or greater than twice the average value for all the residues in the protein are thought to form a domain, and the residue with the highest surrounding hydrophobicity within a domain is believed to be the nucleation site.[28] In the present study, we have also considered the residues close to the highest surrounding hydrophobicity as possible nucleation sites.

Assigning Residues to Different Secondary Structural Elements

The secondary structure location of the residues were identified using the DSSP program.[29] Further, the residues were classified to the C-terminus, N-terminus, and the middle M of the secondary structures as follows. The first two residues in the beginning of the strand was classified to strand N-terminal (SN), and the two residues at the end of the strand was assigned to strand C-terminal (SC). The remaining residues were assigned to strand middle (SM). The same procedure was used to classify residues to HN, HC, and the HM to different regions of the helix. For secondary structures that were less than five residues, only one residue in the N- and C-terminal was assigned to SN or SC (of strand) and HN or HC (of helix), respectively. The remaining residues were classified as SM for strands and HM for helix.

Identification of Conserved Residues Within a Family

The superfamily alignment for each protein has been carried out using the PIR multiple alignment tool, with diverse homology. The conservation of residues in each residue position within the multiple alignment was analyzed. Those positions having residues of the similar type (hydrophobic, charged, or polar) are identified as conserved residues.

RESULTS AND DISCUSSION

Topological Conservation of Clusters in α/β Barrel Proteins

The α/β barrel proteins are known to adopt the same fold in spite of very low sequence similarity. This could be possible only if the specific stabilizing interactions important in maintaining the fold are conserved in topologically equivalent positions. Such stabilization centers are usually identified by determining the extent of long-range contacts made by the residue in the native structure.[20] The constructed graph network in the present study depicts the nonbonded long-range interactions of the protein structure. The resulting highest eigenvalue indicates the topological nature of the protein (see Materials and Methods). The magnitude of the vector components of the highest eigenvalue correspond to the residues that make the largest number of nonbonded interactions in the native structure and can be associated with possible stabilization centers.

The highest eigenvalue of a graph can be considered a topological index,[22] as it is related to the extent of branching of nodes in the graph. For example, the highest eigenvalue obtained for the proteins belonging to the RNase A family and globin family was close to 4.8[16] when the graph was constructed based on the nonbonded connectivity of residues if C_α atoms occur within a spatial distance of 6.5 Å. In the present study, the highest eigenvalue for all the structures adopting the α/β fold is within the range of 5.3–5.7 (using the same distance criteria of 6.5 Å), which is characteristic of this particular fold. The vector components corresponding to the highest eigenvalues are from the β barrel region of the protein, indicating a high backbone packing density of the β barrel region.

The high vector components magnitude of the residues corresponding to the highest eigenvalue are found to cluster around the β barrel region and are found to be topologically conserved in all 36 proteins studied. The vector components and their secondary structural location are shown for 10 proteins belonging to different protein families (Table II). In all cases, the residues that form the top vector components are found to be located in the strand region, which forms the barrel and also occasionally from the loops connecting the strands to the helices. The number of residues forming the cluster is usually within the range of 20–35 residues and the nature of amino acid is mostly hydrophobic in the middle of the strands and charged toward the C-terminal. In most cases, the residue

TABLE II. Vector Components Corresponding to the Highest Eigenvalue for Proteins Belonging to Different Superfamily

Ketolisomerase 1a0c(5.347)[a]	Transferase 1a3x(5.548)	Seed protein 1cnv(5.499)	Oxidoreductase 1ads(5.143)	Phosphotriesterase 1bf6(5.386)	Isomerase 1btm(5.388)	β-amylase 1tml(5.279)	Thiamin biosynthesis 2tps(5.685)	Glycosidase 2myr(5.515)	Glycosyltransferase 1qap(5.638)
336(N,SM)[b]	240(K,SC)	82(A,SC)	159(S,SC)	124(A,SM)	208(Q,SC)	154(Y,SC)	184(V,SC)	455(L,SM)	256(E,SC)
46(A,SC)	261(M,SM)	32(F,SN)	182(N,SN)	80(C,SC)	164(A,SM)	113(I,SC)	185(G,SC)	34(A,SC)	277(I,SM)
337(F,SM)	238(I,SN)	80(F,SM)	108(L,SM)	81(T,SC)	163(I,SN)	185(A,SM)	206(S,SM)	33(V,SM)	276(F,SN)
44(S,SM)	25(G,SM)	127(H,SN)	183(Q,SN)	123(I,SM)	207(I,SM)	156(D,SC)	205(V,SM)	32(G,SM)	257(V,SC)
98(C,SM)	260(V,SM)	9(V,SC)	158(I,SM)	156(S,SM)	123(I,SM)	184(I,SM)	149(G,SM)	94(Y,SN)	255(L,SN)
45(I,SM)	211(F,SM)	225(E,SC)	208(A,SC)	157(T,SC)	209(Y,SM)	155(F,SC)	186(I,SC)	407(V,SM)	232(I,SM)
97(F,SM)	239(V,SC)	10(Y,SC)	157(G,SM)	54(M,T)	124(I,SC)	114(V,SC)	204(G,SN)	454(Y,SM)	233(M,SM)
96(Y,SN)	295(I,SC)	33(I,SM)	207(T,SC)	155(I,SN)	231(L,SN)	71(V,SC)	26(L,T)	456(A,SC)	148(L,SN)
335(L,SN)	294(V,SN)	126(I,SN)	109(I,SC)	82(G,T)	165(Y,SC)	112(I,SC)	29(Y,SN)	31(F,SN)	275(D,T)
295(D,SC)	26(T,SC)	162(A,SC)	206(V,SN)	183(T,SM)	125(C,SC)	70(L,SN)	150(L,SC)	326(G,SN)	278(S,SM)
296(A,SC)	24(I,SN)	31(V,SN)	184(I,SC)	108(E,HC)	227(I,T)	215(A,SN)	203(D,T)	408(T,SC)	234(L,SC)
43(F,SN)	296(C,SC)	185(F,SM)	259(V,SM)	105(M,HM)	229(G,SN)	153(I,SN)	28(V,SN)	327(L,SN)	149(L,SC)
137(L,SM)	335(T,HM)	263(A,SM)	258(VSM)	120(A,SN)	162(V,SN)	183(G,SN)	202(A,T)	459(L,T)	214(E,SC)
338(D,SM)	259(G,SN)	160(S,SM)	181(V,SN)	184(V,SM)	228(D,T)	216(V,SC)	148(V,SM)	406(Y,SM)	147(Q,SN)
139(G,SM)	212(A,SC)	186(V,SM)	77(K,SC)	56(N,SC)	5(I,SM)	39(G,SN)	53(A,T)	328(N,SM)	212(E,SN)
294(I,SM)	49(R,SC)	128(F,SC)	107(Y,SM)	122(I,SM)	210(G,SC)	72(V,SC)	30(F,SN)	95(R,SN)	146(T,SN)
297(N,SC)	48(V,SM)	187(R,SC)	110(H,T)	125(E,SM)	91(I,SM)	186(T,SC)	183(I,SN)	93(G,SN)	254(R,SN)
136(V,SN)	262(V,SM)	129(D,SC)	257(L,SN)	186(H,T)	230(P,T)	182(H,T)	27(S,T)	458(A,SC)	231(I,SN)
181(G,T)	331(L,SC)	116(A,SC)	156(I,SM)	53(E,SC)	90(V,SM)	69(I,SN)	221(A,HM)	409(E,SC)	259(G,T)
99(F,SM)	241(I,SC)	264(L,SC)	209(Y,SC)	207(Q,SM)	122(P,SN)	111(Y,SM)	198(I,HM)	92(T,T)	215(V,SC)
184(Y,SN)	210(V,SM)	223(F,SN)	76(S,SM)	12(H,SC)	92(L,SC)	157(A,T)	130(S,SC)	457(W,SC)	274(V,T)
138(W,SM)	330(M,SM)	8(A,SM)	78(L,SC)	182(V,SN)	7(G,SC)	217(I,SC)	48(A,HM)	255(T,SM)	171(G,T)
140(T,SC)	82(A,SM)	184(I,SM)	75(V,SM)	185(G,SC)	126(G,SC)	218(D,SC)	55(L,SN)	133(I,SN)	258(S,SC)
267(E,SC)	23(I,SN)	81(L,SM)	180(A,T)	109(I,HC)	186(C,HM)	255(F,SM)	129(V,SC)	453(G,SM)	150(D,SC)
	215(I,T)	224(L,SN)	205(V,SN)	206(V,SM)	232(V,SC)	257(W,T)	131(A,SC)	30(I,SN)	213(V,SC)
	329(V,SM)	262(I,SM)	186(C,SC)	205(Y,SM)	166(E,SC)	181(A,T)	L151(G,T)	475(L,T)	
	47(I,SN)	83(L,SC)	160(N,T)	158(H,SC)	93(G,T)	152(R,SN)	54(T,T)	183(L,SM)	
	83(L,SM)	226(L,SC)		241(S,SC)	6(A,SC)	41(W,SM)		96(F,SM)	
	237(I,SN)	11(W,SC)		52(I,SC)	222(L,HM)			325(L,SN)	
	332(S,T)	265(W,SC)		209(D,T)	38(V,SM)			137(V,SM)	
	84(D,SC)	267(R,HN)		126(I,SC)	40(C,SC)			256(M,SM)	
	50(M,SC)	34(S,SC)		159(T,SC)					
	326(A,T)			13(E,SC)					
	45(L,SN)								

[a]The Protein Data Bank (PDB) code and the highest eigenvalue within parentheses.
[b]Residue number (residue name, location in the secondary structure).

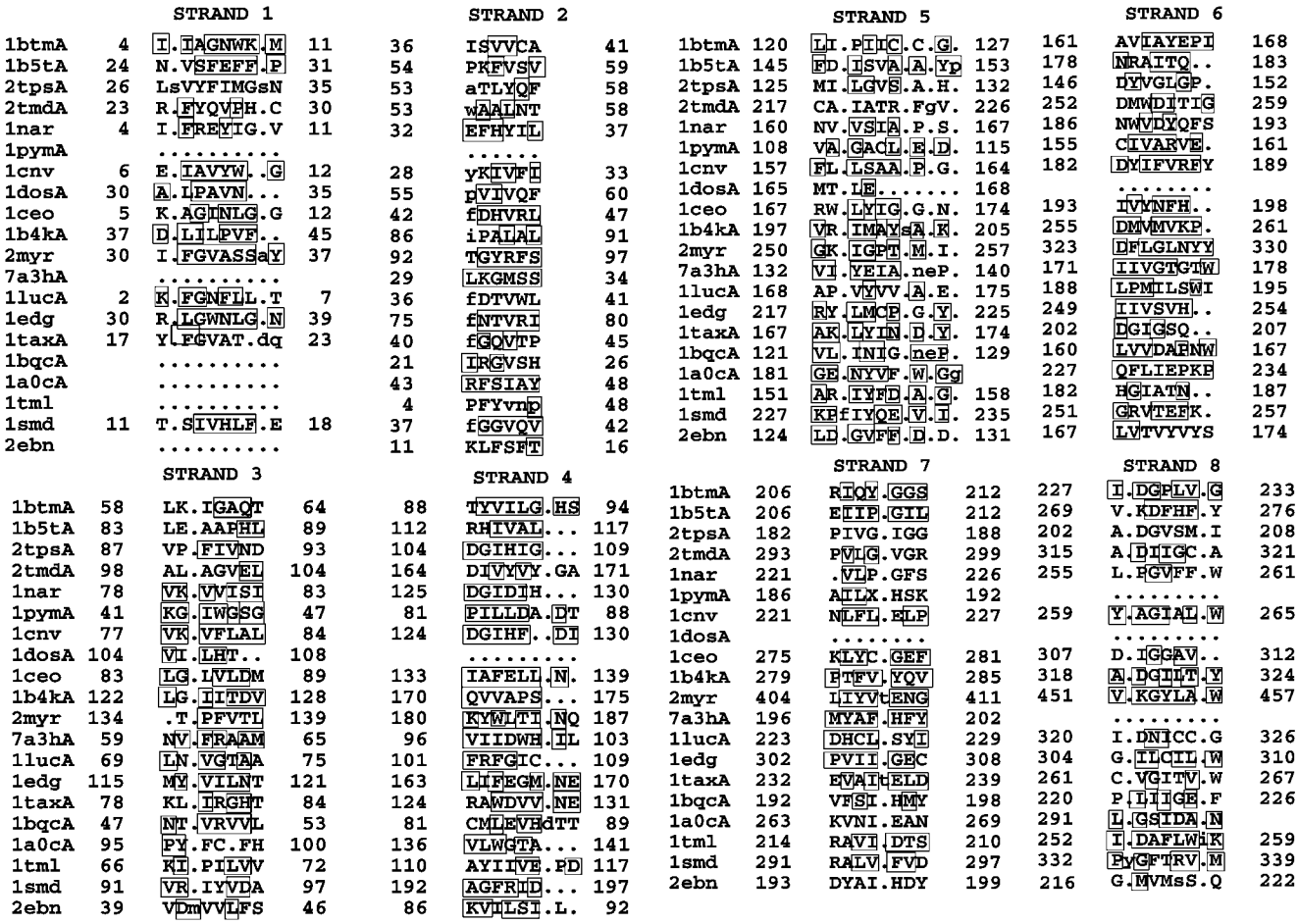

Fig. 3. Structural alignment of strand regions forming the barrel. Residues on the either side of N-terminal and C-terminal region of the strands are also shown in the alignment. Residues conserved within the protein family are denoted by a box symbol.

in the middle of the strand or in the C-terminal of the strand is found to have the largest or the second largest vector component magnitude (cluster center); e.g., 46A (1a0c), 240K (1a3x), 82A (1cnv), 159S (1ads), 80C (1bf6), 154Y (1tml), 184V (2tps), 34A (2myr), and 256E (1qap) occur as one among the top two vector component magnitudes.

It can also be noted that the top 5–6 residues of the highest eigen value are generally found to be located in the middle of the strands or in the C-terminal region of the strands. This indicates a dense backbone connectivity in these two regions of the strands as compared with the N-terminal region.

Sequential Conservation of Cluster Residues

It is known that α/β topology is adopted by sequences with low sequence homology. Hence it is interesting to analyze the nature of conservation among these sequences, expecially among those residues that occur in topologically similar locations. Since most of the cluster residues are found to be located in the β barrel region or in the loops connecting the strands and the helices, we investigated the structural conservation of these residues from a structure based sequence alignment. Of the 36 proteins, 20 were optimally superimposed using FSSP.[30] Interestingly, hydrophobic residues in the strand regions are considerably conserved among the aligned strands across families. Furthermore, within a family of homologous proteins amino acid residues are conserved to a large extent, as compared with its conservation across families. In this study we investigated the conservation of residues among each of the protein family and then marked its position in the structural alignment of 20 proteins (Fig. 3). The structural alignment of the strand regions and a few residues on either side of the strand are also shown in the alignment in Figure 3. The conserved residues in each of the families are shown in the structural alignment by a box symbol.

This approach is more or less similar to the conservatism of conservatism approach described by Mirny and Shakhnovich[24] to analyze the evolutionary signals specific to a given fold. Interestingly, more than 70% of the residues conserved in their respective families are topologically conserved in and around the barrel region (Fig. 3). All the conserved residues are also part of the cluster identified in the β barrel region; therefore, these resi-

TABLE III. Number of Conserved Residues in the Alignment of Proteins Belonging to the Same Family

PDB code	Tot no. of cons res in the family	Tot cons res in the strands	Tot cons res in the N-terminal region of the strands	Tot cons res in the C-terminal of the strands
1btm	50	12	3	26
1b5t	43	13	5	18
1nar	38	9	3	7
1cnv	85	14	15	33
1dos	12	6	1	2
1ceo	41	10	5	15
1b4k	48	9	7	17
1luc	37	5	11	10
1edg	71	13	9	26
1bqc	31	4	2	13
1a0c	53	11	4	28
2tps	19	9	1	8
2tmd	94	6	9	4
1pym	35	9	2	7
2myr	151	25	13	23
7a3h	72	22	11	21
1tax	55	19	11	16
1tml	48	13	3	20
1smd	126	24	16	22
2ebn	15	7	3	4

res, residues; cons, conserved; tot, total; PDB, Protein Data Bank.

dues could possibly be important in stabilizing the α/β barrel fold.

Sequence alignment of proteins belonging to the same family showed that nearly 75% of the conserved amino acids occur in the strand regions or in the regions close to the strand regions (+/− 8 residues on the N-terminal or C-terminal side of the β-strands) (Table III). This conservation is probably because the β barrel region forms the core of the protein and so this conservation of the core residues is important in maintaining the fold. It is interesting to note that the topologically conserved residues in the structural alignment of the 20 α/β barrel proteins also show the highest vector component magnitude (highly branched centers) for each of the strand regions. For example, in the case of isomerase (1btm in Table II), 164A and 209Y are among the residues that have the top eigenvector component and are found to emanate from the middle of strands 6 and 7, respectively. They are among the conserved residues in the sequences belonging to the isomerase family (indicated by box in Fig. 3) and is also topologically conserved (bold letters in the alignment (Fig. 3). Hence, we believe that the conserved residues that show high vector component magnitude could be possible stabilization centers of the α/β barrel protein. The recent study conducted by Mirny and Shakhnovich,[24] using the COC approach, also showed that most of the conserved residues occur in the β barrel region and are located in the loops connecting the C-terminal of the strands to the N-terminal of α-helices.

The composition of the amino acids conserved in the strands and the regions close to the strands (8 residues on the N-terminus of the strand and 8 residues on the C-terminus of the strand) are given in Table IV. A striking difference in the preference of conserved residues is observed in the strands forming the β barrel and in the regions on either side of the strands. Hydrophobic residues V, L, I, and F are more conserved in the strands. Eight residues on the C-terminal region of the strands are mostly in the loops connecting the C-terminal of the strands to the N-terminal of the α-helices. The active site is usually located in this region, so a large number of charged residues like H, D, and E, and which are important for the function are conserved along the C-terminal region. Eight residues on the N-terminal region of the protein constitute the loop connecting the C-terminal region of the α-helices and the N-terminal region of the strands and also a few residues in the C-terminal region of the α-helices as these loops are shorter.[31] The conserved residues in this region are mostly glycine and hydrophobic residues and very little of charged residues. This is in striking contrast to the conservation of residues on the C-terminal region (+8 residues from the C-terminal region) of the strands in which charged residues such as K, H, D, and E are conserved. This result can be useful in predicting sequence motifs capable of detecting proteins that adopt the α/β barrel topology. Attempts were made earlier to predict sequence motifs from the protein sequence that are likely to take the α/β barrel topology based on the multiple alignment of the strands,[32] but the success rate was very low, as the motifs were generated only on the basis of the conservation of residues in the strands. The present study shows that generating sequence motifs based on the conserved residues in the strand regions, as well on the either side of the strand region, can improve the predictive power of assigning the sequences to the α/β barrel fold.

TABLE IV. Composition of Conserved Residues in Strands and Regions Close to Strands

Residue	Tot comp of conserved residues	Tot cons in the strands	Tot cons along C-terminal	Tot cons along N-terminal
G	113	31	45	37
A	61	14	18	29
V	46	33	4	9
L	42	19	8	15
I	46	30	6	10
F	47	20	18	12
Y	48	13	20	15
W	30	8	20	2
C	7	1	3	3
M	12	3	6	3
K	22	2	13	7
H	34	9	23	2
D	59	13	34	12
E	42	11	26	5
N	35	11	21	3
Q	17	3	10	4
S	30	9	17	4
T	18	8	8	2
P	46	10	27	9
R	32	8	13	11

Tot, total; cons, conserved.

Possible Nucleation Sites

Recent experimental and theoretical studies to probe the transition state of protein folding have shown that a specific nucleus of residues cluster during the transition state of folding[33,34] and the subsequent formation of the structure is very fast and downhill in the free energy profile. Further, it was shown that the location of the specific nucleus is highly dependent on the fold, and the residues are conserved in all the sequences that take this fold.[35,36] Recently, Poupon and Mornon[37] showed that the folding nucleus of a protein fold can be predicted by identifying the conserved hydrophobic positions in the sequences that adopt the fold and it has also been shown by earlier studies that residues that form the folding nucleus have an usually high hydrophobic environment.[15,28,38]

Because not many experimental data are available on the folding of TIM barrel proteins, we have attempted to predict the folding nucleus of the proteins by identifying residues that satisfy the following four conditions. Residues were identified to form the folding nucleus if the residue was hydrophobic in nature, were conserved among the sequences within the protein family, exhibited a high hydrophobicity index (see Materials and Methods) and were also part of the detected clusters. Interestingly, in 17 α/β barrel proteins we identified residues that satisfy all the above four criteria (Table V). The nucleation sites were found to occur mostly in the middle of the strands (Fig. 4). Since these residues satisfy all four of the above imposed conditions, we propose that these residues can be the possible nucleation sites in the corresponding α/β protein.

TABLE V. Nucleation Sites

PDB code	Residue in the protein	Secondary structure location
1ads	156I	SM
1ado	145A	SN
1b4k	259V	SM
1btm	229G, 164A, 123I	SN, SM, SM
1dos	31L	SN
1pdy	144L, 392G	T, T
1pym	158A, 213V	SC, T
1qtw	5G, 214G	SM, SM
2tps	48A	HM
1a0c	97F	SM
1a3x	328C	T
1cnv	263A	SM
1edg	338I, 304I	SM, SM
1nar	37L	SM
1tml	114V	SC
5ptd	30G	SM
8ruc	40F	SM

SN, N-terminal strand; SC, C-terminal strand; SM, middle of strand; T, turn/loop; HM, middle of strand.

Fig. 4. Possible nucleation site in the α/β barrel fold. Residue 123I located in the middle of the strand of the protein triose phosphate isomerase (TIM).

Cluster Residues Close to the Active Site

In all the α/β barrel enzyme structures known so far the active site of the enzyme is found close to the C-terminal region of the strands or in the loops connecting the C-terminal of the strands and the N-terminal of the helices.[2] In the present study, a large number of residues that form part of the cluster emanate from the C-terminal region of the strands and the loops connecting the C-terminal of the strands to the N-terminal of the helices. In our investigation of the location of these residues within the context of the active site of the proteins, we found that

TABLE VI. Cluster Residues Close to the Active Site[†]

PDB code	Residues close to the active site
1a0c	295D, 46A
1a3x	**240K**
1ado	300S
1ads	**159S**
1b54	**181M**
1b5t	**88H**
1bf6	157T
1btm	**208Q**
1dos	**59Q**
1fcb	**252Q**
1juk	**110K**
1pdy	**294E**
1pym	**158A, 159R**
1qap	**256E**
1qat	**439G**
1qtw	**259I**
1smd	**196I**
1tml	156D
2myr	34A
2tmd	103E, **104L**
2tps	**184V, 186I**
5ptd	**31T, 81H**

[†]Boldface indicates residues that have the largest vector component value as compared with the C-terminal region of other strands.

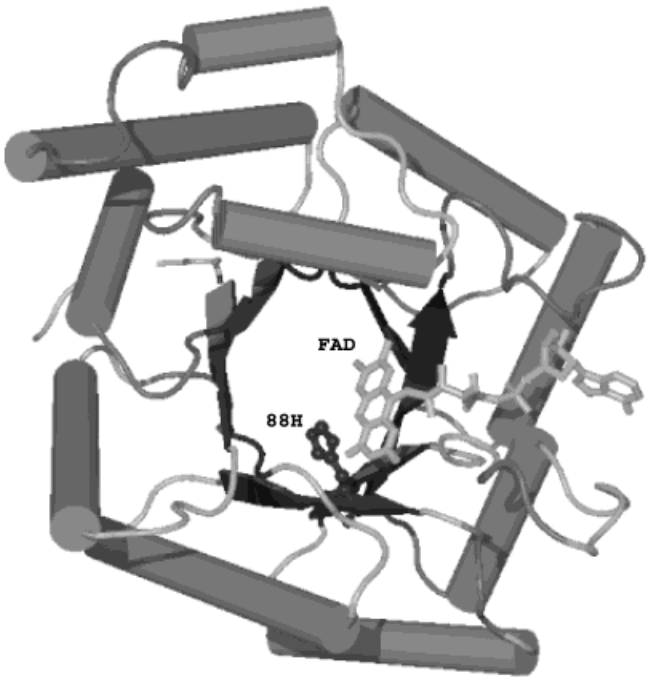

Fig. 5. Cluster residues close to the active site. Residue 88H shown close to the substrate in protein methylene tetrahydrofolate reductase.

out of 26 proteins for which active site information was known, in 23 cases at least one or two cluster residues were part of the active site or within a distance of 6.5 Å from any of the ligand atoms. The list of proteins and the residues of the cluster that form the active site or close to the active/binding are given in Table VI. Further, the residues that occur in the C-terminal region of the strands close to the active site had the largest vector component magnitude as compared with the residues that occur in the C-terminal region of the other strands not close to the active site. This points to the fact that the C-terminal region of the strands forming the active site have a high backbone packing density as compared with the C-terminal regions of the other strands. Also, more of polar and charged residues dominate the list (Table VI). A histidine residue occurring in the C-terminal region of the strands and forming the active site of the protein is shown interacting with the ligand (Fig. 5).

Russell et al.[39] had identified supersites that correspond to the functionally important sites in the protein structure and had identified the supersites to occur in the loops on top of the β barrel structure. In the present investigation, we find these functionally important residues to be part of the cluster, which are also important in stabilizing the native fold. Nature has designed the TIM barrel fold by keeping the functionally important residue close to the stabilizing units (clusters) and hence preserving the optimal orientation of the active site residues which are important for its function. Moreover, because the observed clusters are usually located close to the active/binding site of the protein, identification of such clusters can provide new insights into protein function, especially when the location of the active site is unknown.

Origin of α/β Barrel Fold

There is considerable debate in the literature regarding the evolution of α/β barrel proteins as convergent or divergent.[3] This debate is alive, as there are gray areas that can lead to argument in either directions. However, one factor seems to be apparent, the high designability of the β barrel structure. According to Li et al.,[40] highly designable structures are those that are taken up by a large number of sequences. The fact that nearly 10% of the enzyme structures known so far constitute the α/β barrel clearly indicates that it is a highly designable structure. Thus, it is more of the property of the structure than that of the sequences that has given rise to the abundance of the α/β barrel fold. The present analysis shows that the cluster residues identified in the β barrel region are to a large extent sequentially conserved across protein families, hence they could probably be the most critical residues in stabilizing the native fold. Moreover, the highest eigenvalue observed for all the α/β barrel structures fall within a range of 5.3–5.7, which is unique for the α/β fold. Hence the protein topology is captured here by the highest eigenvalue. The use of highest eigenvalue parameter to classifying protein folds and its ability to identify the native fold of sequence from other non-native folds is in progress.

CONCLUSIONS

A graph theoretical method has been used to identify backbone clusters in 36 α/β barrel proteins. Clusters are found to be located consistently in the barrel region in all proteins studied. The identified cluster residues are conserved within its own family members. Nearly 70% of the

conserved residues in each of the protein families are found to occur either in the beta strand region or in regions close to the β-strands. The cluster centers are mostly found in the middle of the strands and are also predicted as possible folding nucleus. The residues forming the clusters are also part of the active site or found close to the active site. The C-terminal region of the strands close to the active site have a high backbone packing density as compared with the C-terminal regions of the other strands. The present study identifies the conserved topological clusters and cluster centers as possible stabilization units in the β barrel fold. The results provided in this study can be useful in engineering such folds by retaining the identified stabilizing interactions.

Different aspects related to protein structure, function, and folding are addressed in this article within the context of α/β barrel fold. The developed graph theoretical procedure can be applied to any other protein family of interest. Further, a graph theoretical approach can provide insight into such problems as protein–protein interaction by identifying conserved clusters on the surface and can also be extended to identify domains in protein structures.

ACKNOWLEDGMENTS

The Supercomputer Education and Research Center and the Bioinformatics Center at the Indian Institute of Science Bangalore are acknowledged. The authors thank Professor Pujadas for providing the list of nonhomologous TIM barrel proteins.

REFERENCES

1. Wolf YI, Brenner SE, Bash PA, Koonin EV. Distribution of protein folds in the three superkingdoms of life. Genome Res 1999;9:17–26.
2. Farber GK, Petsko GA. The evolution of alpha/beta barrel enzymes. Trends Biochem Sci 1990;15:228–234.
3. Pujadas G, Palau J. TIM barrel fold: structural, functional and evolutionary characteristics in natural and designed molecules. Biol Bratislava 1999;54:231–254.
4. Altamirano MM, Blackburn JM, Aguayo C, Fersht AR. Directed evolution of new catalytic activity using the alpha/beta-barrel scaffold. Nature 2000;403:617–622.
5. Lesk AM, Branden C-I, Chothia C. Structural principles of alpha/beta barrel proteins: the packing of the interior of the sheet. Proteins 1989;5:139–148.
6. Wodak SJ, Lasters I, Pio F, Claessens M. Basic design features of the parrallel alpha beta barrel, a ubiquitous protein-folding motif. Biochem Soc Symp 1990;57:99–121.
7. Murzin AG, Lesk AM, Chothia C. Principles determining the structure of beta-sheet barrels in proteins. J Mol Biol 1994;236: 1382–1400.
8. Laurence TC, Evans PA. Conserved structural features on protein surfaces; small exterior hydrophobic clusters. J Mol Biol 1995;249: 251–258.
9. Selvaraj S, Gromiha MM. An analysis of the amino acid clustering pattern in the alpha/beta barrel proteins. J Protein Chem 1998;17: 407–415.
10. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willet P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains on protein structure. J Mol Biol 1994;243:327–344.
11. Artymiuk PJ, Rice DW, Mitchell EM, Willett P. Structural resemblance between the families of bacterial signal transduction proteins and of G proteins revealed by graph theoretical techniques. Protein Eng 1990;4:39–43.
12. Mitchell EM, Artymiuk PJ, Rice DW, Willett P. Use of techniques from graph theory to compare secondary structural motifs in proteins. J Mol Biol 1990;212:151–166.
13. Koch I, Kaden F, Selbig J. Analysis of sheet topologies by graph theory methods. Proteins 1992;12:314–323.
14. Samudrala R, Moult J. A graph-theoretic algorithm for comparative modeling of protein structures. J Mol Biol 1998;279:287–302.
15. Kannan N, Vishveshwara S. Identification of side-chain clusters in protein structures by graph spectral method. J Mol Biol 1999;292:441–464.
16. Patra SM, Vishveshwara S. Backbone cluster identification in proteins by a graph theoretical method. Biophys Chem 2000;84:13–25.
17. Berman HM, Westbrook J, Feng Z, Gilliand G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
18. Cvetkovic DM, Gutman I. Note on branching. Croat Chem Acta 1977;49:105–121.
19. Panjikar SK, Biswas M, Vishveshwara S. Determinants of backbone packing in globular proteins: an analysis of spatial neighbours. Acta Crystallogr [D] 1997;53:627–637.
20. Dosztanyi Z, Fiser A, Simon I. Stabilization centers in proteins: identification, characterization and predictions. J Mol Biol 1997; 272:597–612.
21. Patra SM, Vishveshwara S. Classification of polymer structures by graph theory. Int J Quant Chem 1998;71:349–356.
22. Randic M, Krilov G. On the characterization of the folding of proteins. Int J Quant Chem 1999;75:1017–1026.
23. Bork P, Gellerich J, Groth H, Hooft R, Martin F. Divergent evolution of a beta/alpha-barrel subclass: detection of numerous phosphate-binding sites by motif search. Protein Sci 1995;4:268–274.
24. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds. Reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999;291:177–196.
25. Nozaki Y, Tanford D. The solubility of amino acids and two glycine peptides in the aqueous ethanol and dioxane solutions. Establishment of hydrophobic scale. J Biol Chem 1971;246:2211–2217.
26. Jones DD. Amino acid properties and side-chain orientation in proteins: a cross correlation approach. J Theor Biol 1975;50:167–183.
27. Manavalan P, Ponnuswamy PK. Hydrophobic character of amino acids residues in globular proteins. Nature 1978;275:673–674.
28. Ponnuswamy PK, Prabhakaran M. Properties of nucleation sites in globular proteins. Biochem Biophys Res Commun 1980;97:1582–1590.
29. Kabsch W, Sander C. Dictionary of protein secondary structure—pattern recognition of hydrogen bonded and geometrical features. Biopolymers 1983;22:2577–2637.
30. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res 1994;22:3600–3609.
31. Scheerlinck J-PY, Lasters I, Claessens M, Maeyer MD, Pio F, Delhaise P, Wodak SJ. Recurrent alpha-beta loop structures in TIM barrel motifs show a distinct pattern of conserved structural features. Proteins 1992;12:299–313.
32. Pickett SD, Saqi MAS, Sternberg MJE. Evaluation of the sequence template method for protein structure prediction. J Mol Biol 1992;228:170–187.
33. Fersht AR. Nucleation mechanisms in protein folding. Curr Opin Struct Biol 1997;7:3–9.
34. Abkevich VI, Gutin AM, Shakhnovich EI. Specific nucleus as the transition state for protein folding: evidence from the lattice model. Biochemistry 1994;33:10026–10036.
35. Martinez J, Pissabarro T, Serrano L. Obligatory steps in protein folding and the conformational diversity of the transition state. Nature Struct Biol 1998;5:721–729.
36. Shakhnovich E, Abkevich V, Ptitsyn O. Conserved residues and the mechanism of folding. Nature 1996;379:96–98.
37. Poupon A, Mornon J-P. Predicting the protein folding nucleus from the sequence. FEBS Lett 1999;452:283–289.
38. Zehfus MH. Automatic recognition of hydrophobic clusters and their correlation with protein folding units. Protein Sci 1995;4: 1188–1202.
39. Russell R, Sasieni P, Stenberg M. Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol 1998;282:903–918.
40. Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. Science 1996;273: 666–669.
41. Humphrey W, Dalke A, Schulten K. VMD—visual molecular dynamics. J Mol Graph 1996;14:33–38.