Protein Engineering vol.15 no.4 pp.265–277, 2002

Analysis of homodimeric protein interfaces by graph-spectral methods

K.V.Brinda[1], N.Kannan[2] and S.Vishveshwara[1,3]

[1]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India and [2]Cold Spring Harbor Laboratory, 1 Bungtown Road, PO Box 100, NY 11724, USA

[3]To whom correspondence should be addressed.
E-mail: sv@mbu.iisc.ernet.in

The quaternary structures impart structural and functional credibility to proteins. In a multi-subunit protein, it is important to understand the factors that drive the association or dissociation of the subunits. It is a well known fact that both hydrophobic and charged interactions contribute to the stability of the protein interface. The interface residues are also known to be highly conserved. Though they are buried in the oligomer, these residues are either exposed or partially exposed in the monomer. It is felt that a systematic and objective method of identifying interface clusters and their analysis can significantly contribute to the identification of a residue or a collection of residues important for oligomerization. Recently, we have applied the techniques of graph-spectral methods to a variety of problems related to protein structure and folding. A major advantage of this methodology is that the problem is viewed from a global protein topology point of view rather than localized regions of the protein structure. In the present investigation, we have applied the methods of graph-spectral analysis to identify side chain clusters at the interface and the centers of these clusters in a set of homodimeric proteins. These clusters are analyzed in terms of properties such as amino acid composition, accessibility to solvent and conservation of residues. Interesting results such as participation of charged and aromatic residues like arginine, glutamic acid, histidine, phenylalanine and tyrosine, consistent with earlier investigations, have emerged from these analyses. Important additional information is that the residues involved are a part of a cluster(s) and that they are sequentially distant residues which have come closer to each other in the three-dimensional structure of the protein. These residues can easily be detected using our graph-spectral algorithm. This method has also been used to identify important residues ('hot spots') in dimerization and also to detect dimerization sites on the monomer. The residues predicted using the present algorithm have correlated well with the experiments indicating the efficacy of this method in predicting residues involved in dimer stability.

Keywords: dimerization sites/eigen vectors/
expanded clusters/interface clusters/interface hot spots

Introduction

Protein–protein interactions are extremely common in nature. Most proteins are functional only as dimers. Many others interact with other proteins to carry out their cellular functions. Classic examples include protein–receptor complexes, antigen–antibody complexes and innumerable other proteins involved in signal transduction. Hence, it is of utmost importance to understand the factors that affect the dimer interface stability. Different methods have been used to study protein interfaces. These include simple methods like detecting the change in the accessible surface area (ASA) when a monomer dimerizes (Chothia and Janin, 1975; Janin *et al.*, 1988) and conservation of amino acid residues at protein interfaces (Hu *et al.*, 2000; Valdar and Thornton, 2001). There are other methods which use geometric properties, surface complementarities between interacting monomers, change in conformational energies and other energy considerations, to predict interacting surfaces and to dock one monomer onto the other. The reviews by Sternberg *et al.* (Sternberg *et al.*, 1998) and Lengauer and Rarey (Lengauer and Rarey, 1996) discuss these methods in detail. Other aspects like amino acid and charge complementarities between interacting surfaces, electrostatic and hydrogen bonding abilities at interfaces and hydrophobic patches occurring at interfaces have also been looked at (Jones and Thornton, 1996, 1997; Xu *et al.*, 1997; Palma *et al.*, 2000). It is believed that correlated mutations contain information regarding interacting residues in proteins (Pazos *et al.*, 1997). The preferences of amino acid residues at the interface have also been analyzed (Jones and Thornton, 1996; Bogan and Thorn, 1998; Larsen *et al.*, 1998; Glaser *et al.*, 2001). Hydrophobic and charged interactions are known to play a major role in stabilizing the dimer (Larsen *et al.*, 1998). It has been proposed that tryptophan, arginine and tyrosine are the preferred amino acid residues at the interface (Bogan and Thorn, 1998). We also know that hydrophobic residues dominate large interfaces whereas charged residues dominate small interfaces (Glaser *et al.*, 2001). All these methods either look at one-to-one interactions between residues or surface geometries or change in conformational energy. However, our present analysis takes into account the overall topology of the protein and uses this input to detect side chain clusters in protein structures. Consequently, our method allows us to detect possible dimerization sites on the monomer as well as to recognize important residues involved in dimerization.

The present study has been directed towards analyzing protein interfaces in a set of 20 homodimers using a graph-spectral method (Kannan and Vishveshwara, 1999; Patra and Vishveshwara, 2000). Graph theory has been frequently used in the analysis of protein structures. For instance, graph-theoretic techniques have been used for the comparison of secondary structural motifs (Mitchell *et al.*, 1990), analysis of sheet topologies (Koch *et al.*, 1992) and identification of specific side chain patterns in three-dimensional structures of proteins (Artymiuk *et al.*, 1994). Thermal fluctuations in proteins have also been evaluated using Kirchoff's adjacency matrix based on proximity of residues in three-dimensional space (Bahar *et al.*, 1997). The present algorithm uses a graph-

theoretic method to determine side chain clusters in proteins. It is a well known fact that side chain clusters in proteins aid in protein folding and in stabilizing the three-dimensional structure of proteins (Heringa and Argos, 1991). Previous investigations related to protein structure and stability, that were carried out in our laboratory, showed that aromatic side chain clusters in thermophilic proteins were involved in imparting thermal stability to proteins (Kannan and Vishveshwara, 2000). Residue clusters identified in α–β barrel proteins were found to be topologically conserved and were essentially involved in imparting structural stability (Kannan *et al.*, 2001b). These studies provided us insights into the possible role of side chain clusters in maintaining protein structure and stability and hence motivated us to take a closer look at dimer interfaces in terms of side chain clusters.

Graph-spectral parameters like the eigen values and the eigen vector components provide us significant information about the side chain clusters. These eigen values and their vector components along with other properties such as the difference in the accessible surface area (δASA) upon dimerization (Chothia and Janin, 1975) and the conservation of residues in homologous proteins, have been used to predict a few residues at the interface of these proteins which may play a significant role in dimer interface stabilization. We have also analyzed the clusters in the 20 monomers to identify exposed and conserved clusters that could possibly be dimerization sites on the monomer.

Our studies confirm that the interface cluster residues comprise of both charged and hydrophobic residues which implies that both charged and hydrophobic interactions are required for stabilizing the dimer interface. Most charged residues in the interface clusters are neutralized by oppositely charged residues, which can either belong to the same chain or to the other chain. This leads us to believe that dimer interfaces are essentially neutral with charges nullified by complementary residues. We find that there is high correlation between the residues with a high eigen vector component, large δASA and high conservation in homologs. Considering these factors, we propose that the residues which satisfy all the three above-mentioned criteria, will probably play a very significant role in the stability of the dimer interface. We would like to emphasize that detection of interface residue clusters and determination of their eigen values and eigen vector components, gives insight to the structural characteristics of protein interfaces. Based on these observations we have attempted to predict mutations at the interface of these proteins which may possibly disrupt the dimer interface. Identification of dimerization sites on the monomer, based on the detection of exposed clusters that are conserved, has also yielded good results.

Materials and methods

Data set and detection of clusters

Construction of protein graphs. The crystallographic coordinates of the 20 homodimers (Table I), whose resolution is better than 2.5 Å, have been obtained from the RCSB protein data bank (Berman *et al.*, 2000). Side chain clusters were determined for all the 20 dimers and their corresponding monomers using a graph-theoretic algorithm. A brief description of the methodology is given here (Kannan and Vishveshwara, 1999). Each residue of the protein is represented in the form of a node in the graph. A protein with n residues is represented by a graph of n nodes. Non-glycine amino acids are represented by their Cβ atoms whereas glycines are represented by Cα atoms. The graph is constructed by connecting two non-adjacent nodes (residues $i - 2$ to $i + 2$ are excluded) by edges if the side chain interaction criterion (contact criterion) between them is satisfied. The contact criterion is a user-defined parameter, which essentially defines the condition that has to be satisfied by the residues so as to be connected to other residues. This criterion specifies the number of side chain atoms from a pair of sequentially non-adjacent residues that have to come within a distance of 4.5 Å so that they can be considered as interacting residues. If a contact criterion of 6% is specified by the user, then the output consists of clusters in which all the residues in a cluster will have at least 6% of overlap with one or more residues in the cluster.

Matrix construction. The connected protein graph can be represented in terms of an adjacency matrix, A, where:

$a_{ij} = 1/d_{ij}$, if i and j are connected and
$a_{ij} = 1/100$, otherwise.
d_{ij} = distance between i and j.

The degree matrix, D, is a diagonal matrix obtained by summing up the elements of each column. The Laplacian matrix, L is defined as $D - A$. This Laplacian matrix is of dimension $n \times n$, where n is the number of residues in the protein.

Graph spectra. This Laplacian matrix, L, is diagonalized to yield the eigen values and the eigen vector components. The vector components corresponding to the second lowest eigen value gives the clustering information (Hall, 1970). The centers of these clusters can be identified from the eigen vector components of the top eigen values (Kannan and Vishveshwara, 1999; Patra and Vishveshwara, 2000). The cluster centers identified correspond to the nodes with the highest connectivity (degree) in the cluster, which, in most protein clusters that we have dealt with, also correspond to the geometric center of the cluster (unpublished results). Only clusters with three or more residues are considered in this analysis. Table II shows the complete set of clusters obtained in the monomers and dimer of yeast triose phosphate isomerase when a contact criterion of 12% is used. The residues with the same vector component in the second lowest eigen value form a cluster. The residue with the highest magnitude of a vector component in the corresponding top eigen value is the center of the cluster. Thus, we can see that monomer A has two clusters, monomer B has three clusters and the dimer has eight clusters for the chosen contact criterion of 12%. The residues forming the center of the clusters are marked in bold.

Identification of interface clusters

After determining side chain clusters in the dimer, those at the interface are identified and differentiated from the others based on the fact that the interface clusters would have contributions from both chains of the dimer. The contact criterion to select the interface clusters in a protein has been optimized as follows. Initially, we begin with a high contact criterion, for example 14%, to obtain clusters. We then gradually reduce the criterion by 1% until at least one or two clusters comprising of residues from both monomers are obtained. We have used contact criteria varying from 6 to 14% and have detected

Table I. Mutations predicted to influence dimerization

Protein	PDB code	Predicted hot spots	Cl. no.–Res no. (PVC)[a]	δASA (%)[b]	Conservation of residue	Experimentally observed[e] effect on dimer stability	Reference
Interleukin 8	1il8	E24	1–5(1)	43.1	T[c]	–	
		I28	2–5(1)	22.4	T	–	
		E29	3–10(1)	88.0	T	–	
		F65	4–10(1)	17.6	T	–	
		L25	4–10(2)	21.1	T	+ve	Horcher *et al.*, 1998
		V27	5–3(1)	56.3	T	+ve	Horcher *et al.*, 1998
Mannose binding protein	1msb	F111	1–3(1)	55.4	P[d]	–	
		N115	2–3(1)	59.3	P	–	
Phospholipase A2	1pp2	F5	1–9(1)	8.1	T	+ve	Liu *et al.*, 1995
		H34	2–8(1)	34.4	P	–	
		N67	3–12(1)	53.0	P	–	
		I9	1–9(2)	2.6	T	+ve	Liu *et al.*, 1995
Uteroglobin	1utg	M41	1–3(1)	43.1	T	–	
		T52	2–3(1)	44.1	T	–	
		I56	3–3(1)	37.9	T	–	
Triose phosphate isomerase	1ypi	F102	1–7(1)	48.6	P	–	Mainfroid *et al.*, 1996
		T75	2–3(1)	99.6	T	–	
		R98	1–7(2)	26.4	T	+ve	
Cytochrome C	2ccy	Q13	1–3(1)	22.4	P	–	
Citrate synthase	2cts	H246	1–6(1)	22.7	T	–	
		L250	1–6(2)	52.2	T	–	
		R421	2–3(1)	88.5	P	–	
		P422	3–3(1)	72.2	T	–	
		K423	2–3(2)	85.4	P	–	
Gene 5 DNA binding protein	2gn5	F68	1–6(1)	30.6	P	–	
		V4	1–6(2)	35.6	P	–	
Tyrosyl tRNA synthetase	2ts1	F164	1–4(1)	88.2	T	–	
		R137	2–5(2)	55.9	T	+ve	Bedouelle and Winter, 1986
		L88	3–5(1)	34.5	P	+ve	Bedouelle and Winter, 1986
Thymidylate synthase	2tsc	W133	1–3(1)	39.5	T	–	
		V135	2–5(1)	85.6	T	–	
		N134	2–5(2)	6.2	T	–	
Rubisco	2rus	N93	1–3(1)	27.4	P	–	
		E239	2–6(1)	49.0	T	–	
Tryptophan repressor	2wrp	W19	1–3(1)	60.6	T	–	
		V23	1–3(2)	44.8	T	–	
		Y30	2–4(1)	54.9	T	–	
		E47	3–5(1)	33.0	T	–	
		R54	4–5(1)	43.0	T	–	
Aspartate amino transferase	3aat	K68	1–3(1)	83.8	T	–	
		N297	2–7(1)	18.0	T	–	
Catabolic gene activator protein	3gap	F76	1–4(2)	41.3	T	–	
		S117	1–4(1)	53.0	T	–	
		L113	2–3(1)	27.4	T	–	
Glutathione reductase	3grs	F87	1–5(1)	60.8	T	–	
		R478	2–4(1)	41.6	P	–	
Isocitrate dehydrogenase	3icd	Y160	1–3(1)	26.1	P	+ve	Hurley *et al.*, 1996
		E164	2–3(1)	62.1	T	–	
Iron super oxide dismutase	3sdp	E159	1–4(1)	70.0	T	–	
		R167	1–4(2)	69.5	P	–	
		E21	2–5(2)	9.4	P	–	
Malate dehydrogenase	4mdh	M54	1–3(2)	50.8	P	–	
		D58	1–3(1)	88.2	T	–	
		R229	2–5(1)	39.0	T	–	
		K247	3–4(1)	55.1	P	–	
Spo0B	1ixm	F78	1–3(1)	37.5	P	–	
		R29	2–8(1)	50.1	T	–	
Cardiotoxin	1cdt	N45	1–4(1)	43.2	P	–	
		K50	1–4(2)	17.0	T	–	

[a]Cl. no.–Res. no. (PVC), cluster number of the cluster to which the residue belongs–number of residues in the cluster (position of the vector component of the residue in the highest eigen value vector of the cluster: 1, top; 2, second highest).
[b]δASA, change in the ASA of the residue on dimerization.
[c]T, totally conserved residues.
[d]P, partially conserved residues.
[e]A blank (–) indicates that no experimental information is available.

Table II. Cluster residues and their vector components of yeast triose phosphate isomerase (1ypi) at 12% contact criteria

Clusters in monomer			Clusters in dimer			Nature of cluster
Cluster no. and residues	Vector component		Residue	Vector component		
	HEV[a]	2nd LEV[b]		HEV	2nd LEV	
1. F108(A)	0.127	0.274	F108(A)[d]	0.055	0.146	Expanded interface cluster
N65(A)	0.237	0.274	Y67(B)[d]	0.069	0.146	
F102(A)	0.340	0.274	N65(A)	0.143	0.146	
E104(A)	0.371	0.274	E104(A)	0.196	0.146	
R98(A)[c]	**0.821**	0.274	E77(B)	0.400	0.146	
			R98(A)	0.598	0.146	
			F102(A)	**0.645**	0.146	
2. **R3(A)**	**0.810**	0.456	**R3(A)**	**0.753**	0.087	Non-interface cluster (first monomer)
R189(A)	0.495	0.456	R189(A)	0.649	0.087	
D227(A)	0.315	0.456	D227(A)	0.104	0.087	
3.			H95(A)	0.338	0.262	New interface cluster
			N10(A)	0.475	0.262	
			T75(B)	**0.813**	0.262	
4.			E97(B)	0.334	0.083	New interface cluster
			H95(B)	0.479	0.083	
			T75(A)	**0.812**	0.083	
5.			Y49(B)	0.214	0.007	New interface cluster
			D48(A)	0.576	0.007	
			K17(B)	**0.789**	0.007	
6. E37(B)	0.228	0.024	E37(B)	0.195	0.392	Non-interface cluster (second monomer)
R205(B)	0.565	0.024	R205(B)	0.589	0.392	
F6(B)	**0.793**	0.024	**F6(B)**	**0.784**	0.392	
7. H185(B)	0.069	0.306	H185(B)	0.213	0.001	Non-interface cluster (second monomer)
D227(B)	0.186	0.306	D227(B)	0.293	0.001	
I206(B)	0.327	0.306	I206(B)	0.349	0.001	
R189(B)	0.617	0.306	R189(B)	0.528	0.001	
R3(B)	**0.688**	0.306	**R3(B)**	**0.684**	0.001	
8. E104(B)	0.202	0.364	Y67(A)	0.062	0.152	Expanded interface cluster
F102(B)	0.324	0.364	E104(B)	0.125	0.152	
N65(B)	0.336	0.364	N65(B)	0.203	0.152	
R98(B)	**0.861**	0.364	E77(A)	0.390	0.152	
			F102(B)	**0.627**	0.152	
			R98(B)	**0.627**	0.152	

[a]HEV, highest eigen value for the corresponding cluster.
[b]2nd LEV, second lowest eigen value.
[c]Cluster centers and their respective vector components are represented in bold.
[d](A) and (B) are two different monomer chains.

clusters varying from 5 to 14 in number. The number of residues per cluster varies from 3 to 15 according to the contact criterion used and the size of the protein. The same contact criterion has been used for a chosen monomer–dimer pair so that the clusters obtained in the two can be compared. These side chain clusters were then visualized using the package VMD (Humphrey *et al.*, 1996). Having obtained a set of side chain clusters for the monomer as well as the dimer of the same protein, these clusters and their eigen values and eigen vector components were then critically analyzed to identify the cluster centers and also the changes in the side chain clusters that are expected to occur on dimerization.

Analysis of other properties

The ASAs of all the monomers as well as the dimers were determined using Connolly's ASA program with a probe of radius 1.4 Å (Connolly, 1993). The percentage difference in the ASA of residue *i* when the monomer dimerizes has been calculated as:

$$\%\delta\mathrm{ASA}(i) = \delta\mathrm{ASA}(i) \,/\, \text{total ASA}(i) \times 100$$

The total ASA for each residue type has been obtained from the literature (Miller *et al.*, 1987).

The conservation of the interface cluster residues in various species has been looked at using the ClustalW program (Thompson *et al.*, 1994). The sequences of these proteins from various species were obtained from the Swiss-Prot data bank (Bairoch and Apweiler, 2000).

Analysis of monomer clusters and identification of possible dimerization sites

It is believed that the three-dimensional structure of a monomer encodes the information required for dimerization. The features which have earlier been considered important for analyzing protein interfaces are (i) the nature and composition of amino acids (Jones and Thornton, 1996; Bogan and Thorn, 1998; Larsen *et al.*, 1998; Glaser *et al.*, 2001), (ii) solvent accessibility (exposed or buried) (Chothia and Janin, 1975) and (iii) conservation of interface residues (Hu *et al.*, 2000; Valdar and Thornton, 2001). In the present study, we have incorporated these features in the amino acid clusters detected by our method and have attempted to identify the clusters involved in dimerization. Our analysis has shown that a strong interface is formed only when there is a seeding cluster (of three or more residues) in at least one of the monomers, which gets strengthened on dimerization. These are considered as

'expanded' clusters. We also find that some 'new' clusters are formed during dimerization. Invariably, the size of such clusters is small and results in fewer interactions between the two monomers, when compared with expanded clusters. These are explained in a later section. Our procedure is able to identify the seeding clusters in the monomer which get expanded on dimerization. The details of the identification of such seeding clusters in the monomers are given below.

Our analysis has shown that the seeding clusters are formed even when a high contact criterion (10–12%) is used indicating that these seeding (expanded) clusters consist of very strongly interacting residues. Thus, the first step in the process is to identify side chain clusters in a given monomer using a high contact criterion. The contact criterion is gradually reduced starting from 12 down to 8% until the first surface clusters emerge in the monomer. This leads to the use of different contact criteria for different monomers. The clusters thus obtained are then characterized based on their location, the extent of conservation of the component residues and number of preferred residues present in them. The 'surface clusters' are identified by carrying out the ASA calculations using Connolly's algorithm (Connolly, 1993). Residues with $>20\%$ ASA are considered as exposed (E), those between 5 and 20% are considered partially exposed (P) and the rest are considered as buried (B). A cluster is considered as an exposed cluster if at least two of the residues are exposed or partially exposed. In the next step, we look for the presence of at least one of the preferred amino acids (arginine, histidine, phenylalanine, tyrosine and glutamic acid, which are the preferred amino acids in the seeding clusters based on our analysis) in these clusters. The clusters are then investigated for conserved residues. Residues could be classified based on their conservation not only as totally (T) and partially (P) conserved but also they could have undergone conserved mutations (M). A cluster with at least two of the residues conserved (T, P or M) is considered as a 'conserved cluster'. After considerable reduction of contact criterion, if no clusters satisfying the above criteria are obtained, one can look for such clusters in the other monomer, which could have clusters satisfying these criteria. (Although, in principle, the two monomers of a homodimer are identical, there could be differences in the coordinates of the two monomers if the crystallographic asymmetric unit is a dimer.) Thus, 'exposed, conserved clusters with the preferred residues' are identified as possible sites of dimerization. If there is more than one cluster identified as a dimerization site, then they can be ranked based on the extent of conservation and the number of preferred residues.

Results and discussion

Side chain cluster analysis

Identification of side chain clusters. Side chain clusters have been determined for all the 20 monomers and dimers using a graph-theoretic algorithm. The number and size of clusters obtained depend on the contact criterion used. If a low contact criterion is specified, then we could end up with the whole protein as a single cluster. Similarly, if a high contact criterion is used, one might lose essential information regarding side chain interactions because high cut-off will yield very few or no clusters at all. Hence, the contact criterion has been optimized for each protein so that we get clusters that are discriminated from the bulk of the protein and, therefore, different contact criteria have been used for different proteins. However, the same value is used for a chosen monomer–dimer pair. In this analysis, we have used contact criteria varying from 6 to 14% and we obtain side chain clusters varying from 5 to 15 in number per protein and the size of clusters varying from 3 to 15 residues per cluster.

Differences between clusters in monomers and dimers. After determining side chain clusters in the dimer, those at the interface are identified and differentiated from the others based on the fact that interface clusters would have component residues that are contributed from both chains of the dimer. The differences between the side chain clusters in the monomer and the dimer were then determined. The interface clusters in the dimer can be categorized (with reference to the clusters in the monomer) as: (i) a new cluster formed on dimerization; and (ii) an existing cluster in the monomer which expands or gets strengthened on dimerization, in which case the monomer consists of a set of seeding clusters to which more residues from the other monomer are added on dimerization.

These two cases can be understood from Tables II and III. Table II shows the clusters obtained in the monomer and dimer of triose phosphate isomerase using a contact criterion of 12%. Clusters 2, 6 and 7 are non-interface clusters whereas the others are interface clusters. Clusters 3, 4 and 5 are new clusters that emerged in the dimer whereas clusters 1 and 8 are clusters which already had a seeding in the monomer that got strengthened after dimerization. Table III elucidates the clusters in the monomer and dimer of malate dehydrogenase at 11% cut-off. Malate dehydrogenase has three interface clusters. Clusters 2 and 15 are new interface clusters whereas cluster 1 is an expanded interface cluster, which had a seeding in the monomer. All others are non-interface clusters, which are present in the monomers as well. Figures 1 and 2 show the clusters in the monomers and dimers of triose phosphate isomerase and malate dehydrogenase, respectively. The rectangular regions enclose the seeding clusters in the monomers and the interface clusters in the dimers.

Most of the proteins that have been analyzed have both these types of clusters at the interface. All 20 of them have some new clusters that are formed on dimerization. Greater than 70% of them have at least one additional cluster at the interface for which there was already a seeding cluster in the monomer, which got strengthened after dimerization. The expanded clusters impart more stability to the dimer interface than new clusters as they are involved in creating a bigger network of interactions between the two monomers involved in dimerization. These expanded (seeding) clusters have been further analyzed to predict possible dimerization sites on the monomer. The details of this are discussed in a different section.

Preference of amino acid residues in interface clusters. In both the above-mentioned cases, namely formation of new clusters and expansion of existing clusters, new charged and hydrophobic interactions, essential for dimer formation, are introduced. It has been observed that, if there is a charged (positive or negative) residue in the interface cluster, more often than not, there is an oppositely charged residue too in the same cluster, which neutralizes this charge and thus stabilizes the cluster. This oppositely charged residue can be from the same chain as the first charged residue or it can be from the other chain. Though dimer interfaces do have charged residues, they are essentially neutral because of the nullification

Table III. Clusters in malate dehydrogenase (4mdh) at 11% contact criterion

Cluster number	Clusters in monomer[a]	Clusters in dimer	Nature of cluster[b]
1	D251, R156, K247	Y355, K247, D251, R156	4
2		M388, R229, D392, R161, L157	3
3	K553, F641, K645	K553, F641, K645	2
4	K294, W257, F29	K294, W257, F29	1
5	E318, T189, N185	E318, T189, N185	1
6	W183, P288, I181	W183, P288, I181	1
7	F307, K311, Y191, I223	F307, K311, Y191, I223	1
8	W591, K628, F363	W591, K628, F363	2
9	F222, K169, W217	F222, K169, W217	1
10	F556, W551, K503	F556, W551, K503	2
11	D492, Q524, H520, Q561	D492, Q524, H520, Q561	2
12	Y543, L552, V546	Y543, L552, V546	2
13	S145, K121, D116	S145, K121, D116	1
14	R490, D585, K581	R490, D585, K581	2
15		M54, R563, D58	3
16	I515, P622, W517	I515, P622, W5172	

[a]The residues 1–333 belong to the first monomer whereas the residues 335–667 belong to the second monomer in malate dehydrogenase.
[b]Nature of cluster: 1, non-interface cluster in first monomer; 2, non-interface cluster in second monomer; 3, new interface cluster; 4, expanded interface cluster.

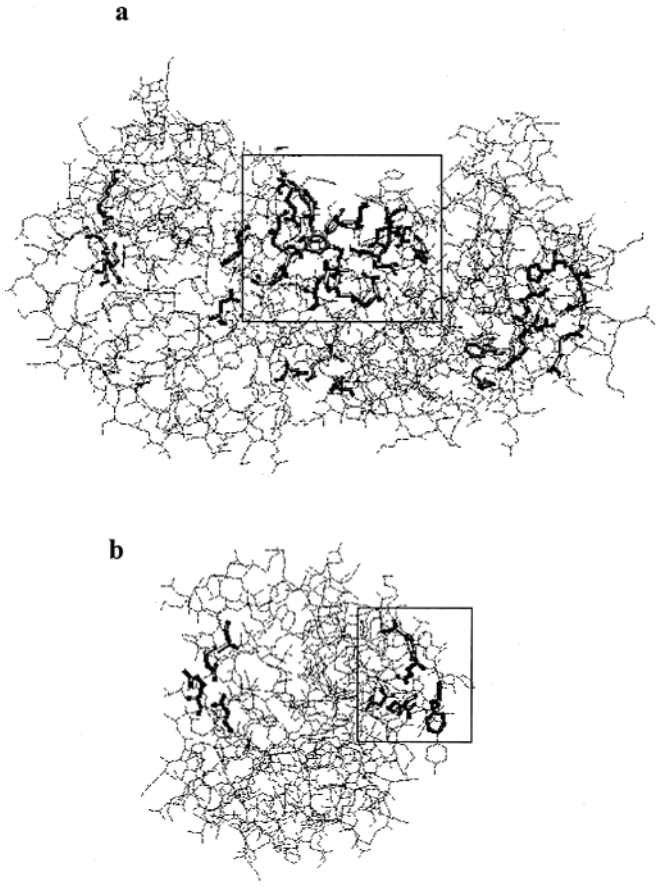

Fig. 1. Side chain clusters in triose phosphate isomerase (**a**) dimer and (**b**) monomer. The cluster residues are shown in bold. The rectangular box in (a) corresponds to the interface clusters and the one in (b) shows the seeding clusters that get expanded on dimerization.

of charges. Apart from these charged residues, most interface clusters also have hydrophobic residues. For example, in triose phosphate isomerase, when cluster 1 comprising of residues F102, E104, N65, R98, F108, gets strengthened to a cluster comprising of residues F108, E104, N65, R98, F102, E325, Y315 (Table II), new charged (E325) and hydrophobic (Y315) residues are added on to the cluster which induce new interactions in the dimer which were initially absent in the monomer. Similarly, in the case of malate dehydrogenase, clusters 2 and 15 with residues M388, R229, D392, R161, L157 and M54, R563, D58, respectively, are new clusters that are formed on dimerization (Table III). These clusters also induce some new charged and hydrophobic interactions in the dimer that were absent in the monomer. Hence, both charged and hydrophobic residues are equally involved in the formation of new interface clusters as well as strengthening of existing clusters, both of which occur when a protein dimerizes.

The composition of the amino acids at the interface of proteins considered in this data set is given in Table IV. The values that are obtained by considering all residues, which have lost even a small amount of ASA upon dimerization, are reported. Also reported are values obtained by considering the composition of interface clusters. The normalized values (percentage compositions) for both these cases are shown as histograms in Figures 3a and b, respectively. It is evident that the composition patterns are different in both. Figure 3b is more discriminatory than Figure 3a. This shows that by using a high contact criterion, we are able to identify residues that contribute significantly to the stability of the dimer, which has been the aim of the present analysis. In principle, we could pick up all those residues that have lost ASA on dimerization by our method using a low cut-off value, in which case Figure 3b would be similar to Figure 3a.

Arginine, histidine, phenylalanine, tyrosine and glutamic acid are found to be the most preferred residues in the interface clusters as shown in Figure 3b. There is a significant contribution from tryptophan as well as methionine in the interface clusters when compared with the other amino acids. This preference of amino acids in interface clusters is consistent with the preferences obtained by earlier studies which were carried out on the basis of residue-wise interactions (Jones and Thornton, 1996; Bogan and Thorn, 1998; Larsen *et al.*, 1998; Glaser *et al.*, 2001). Clearly, there seems to be a preference for charged and aromatic side chains in the interface clusters. Glycine, alanine, valine and cysteine are among those residues that are rarely found in the interface side chain clusters. A comparison of Figures 3a and b shows that although

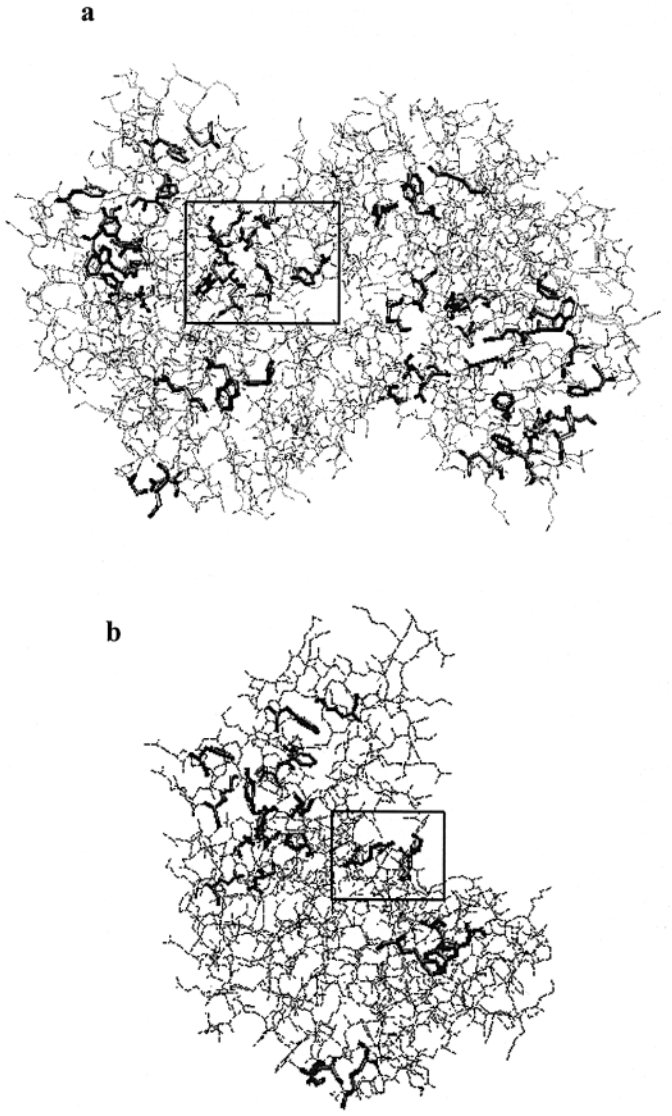

Fig. 2. Side chain clusters in malate dehydrogenase (**a**) dimer and (**b**) monomer. The cluster residues are shown in bold. The rectangular box corresponds to the interface clusters in the dimer (a) and the seeding cluster in the monomer (b).

these small side chain residues occur at the interface (Figure 3a), they do not contribute to the interface stability by participating in significant interactions across the interface (Figure 3b). Thus, the present analysis focuses on detecting strongly interacting residues at the interface.

Predicting residues involved in dimer stability

Previous investigators have used the loss of accessible area upon dimerization (Chothia and Janin, 1975) and other features such as conserved amino acid residues at the interface (Hu *et al.*, 2000; Valdar and Thornton, 2001) and correlated mutations that occur in protein sequences (Pazos *et al.*, 1997) for predicting the residues involved in stabilizing the dimers. These methods, no doubt, have aided in identifying important residues for oligomerization. Further, a rigorous method for analyzing protein interfaces using surface patch analysis has also been developed by Thornton's group (Jones and Thornton, 1997). Several other methods which use surface properties and

Table IV. Amino acid preferences at the interface

Amino acid	Total no. of residues in the data set	No. of residues that have lost ASA[a]	No. of residues which are part of interface clusters
Ala	369	68	0
Gly	366	66	0
Pro	186	36	7
Met	105	30	11
Cys	79	20	2
Val	278	65	5
Leu	397	92	19
Ile	247	49	14
Ser	231	59	5
Thr	234	49	9
Asp	237	41	18
Glu	286	54	35
Asn	162	46	17
Gln	145	34	12
Lys	291	56	18
Arg	180	52	37
His	107	26	19
Phe	155	44	25
Tyr	137	35	21
Trp	66	11	7

[a]The residues that have lost even a small amount of ASA upon dimerization have been considered.

geometric complementarity of monomers have also been used to identify protein interfaces. These methods have been discussed in detail in the reviews by Sternberg *et al.* (Sternberg *et al.*, 1998) and Lenguaer and Rarey (Lenguaer and Rarey, 1996). However, most of these techniques examine the interface interactions at a one-to-one residue level or at a surface geometry level. Our present method of identification of side chain clusters using graph theory considers the connectivity input in a global way and the identified clusters as a network of connections between the monomers. Thus, the interface clusters detected by the graph-spectral method give extended side-chain network information. Further, the technique also identifies the center of such a networked cluster, identifying probably the most important residue(s) for dimerization. In the present study, we have combined a graph-spectral algorithm with traditional methods of investigating the features such as loss in accessible area upon dimerization and conservation of interface residues in order to predict residues important in dimer stability. The details of the investigations as applied to the present data set are given below.

Eigen vector component. The eigen vector component of an interface cluster residue can be used as an important criterion to determine residues, which may be involved in stabilizing the dimer interface. As mentioned earlier, the graph-theoretic algorithm used in this analysis gives the eigen values and the corresponding eigen vector components for each residue involved in cluster formation. Interface clusters and their centers have been identified as mentioned earlier. It has been observed earlier (Kannan and Vishveshwara, 1999; Patra and Vishveshwara, 2000) that the residues at the center of the cluster have a high vector component corresponding to the highest eigen value of that cluster. Similarly, the residues with a low vector component in this eigen value are away from the core of the cluster. The higher the magnitude of the vector component of a residue within a cluster, the more important is its role in the formation and stabilization of the cluster because it represents the core of the cluster. We would

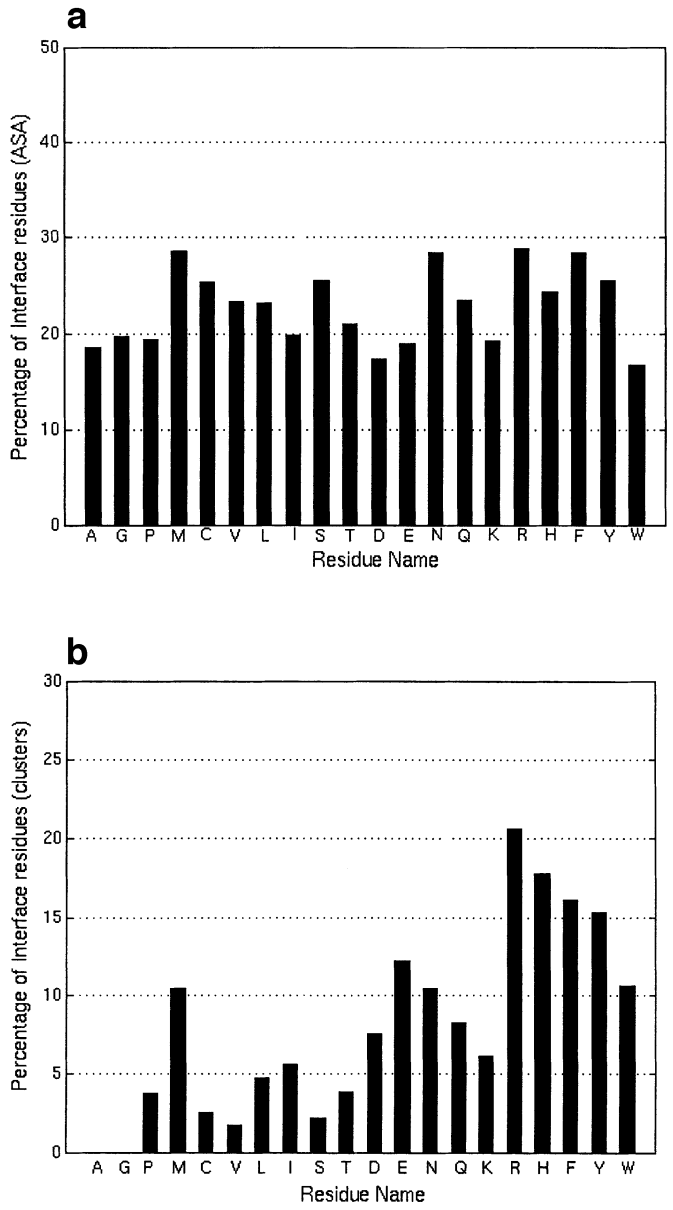

Fig. 3. Histogram representing the occurrence of the 20 amino acids at protein interfaces. Amino acids are represented using their single letter codes. (**a**) The percentage of amino acids occurring at the interface based on loss of accessible surface area. (**b**) The percentage of amino acids occurring in the interface clusters.

like to emphasize the fact that these high vector component residues make stronger and more number of contacts with their spatial neighbors than other residues in the cluster. The mutation of such residues could result in the loss of spatial contacts leading to loss of interactions across the dimer interface. Thus, interface clusters and their centers give us an insight to residues involved in dimer stability. Hence, this is one of the important factors that has been used to predict residues, which might stabilize the dimer interface.

Accessible surface area. The change in the accessible area of a residue when a monomer dimerizes can give us some information as to which are the interface residues. This property has been used as one of the criteria to predict dimer destabilizing mutants.

A measure of the loss in the ASA when two monomers associate to form a dimer gives an estimate of the hydrophobic free energy involved in dimer formation (Chothia and Janin, 1975). A good estimate of this can be made if the structures of both the monomer and the dimer are known. However, in the absence of the isolated monomer structure, it is reasonable to approximate the structure of the monomer to be very similar to that in the dimer. An empirical correlation between the extent of the ASA lost upon dimerization and the hydrophobic free energy contribution towards dimer formation has been given by Chothia and Janin (Chothia and Janin, 1975), according to which a loss of 1 $Å^2$ corresponds to a contribution of 0.025 kcal/mol of hydrophobic free energy. Hence, the more the loss in the ASA of a residue on dimerization, the more the residue contributes to the hydrophobic free energy of dimerization. This property has been used as a contributing factor to determine the residues, which might play a major role in stabilizing the dimer. In the present analysis, any residue, which is a part of the identified interface cluster and has lost even a small amount of the ASA upon dimerization, has been taken into consideration for identification of 'hot spots' at the dimer interface.

Conservation of interface cluster residues. It has been pointed out (Hu *et al.*, 2000; Valdar and Thornton, 2001) that the extent of conservation is high in the interface residues and so conserved residues are more likely to be important in protein dimerization. Hence, it is important to know whether the interface cluster residues are conserved or not. Therefore, this has been used as one of the factors to identify residues that might be important for dimer stability. The homologous set of sequences for each of the 20 homodimers were obtained from Swiss-Prot data bank and aligned using the ClustalW algorithm (Thompson *et al.*, 1994). The conserved residues were then identified from these aligned sequences. The sequence alignment of cardiotoxin (1cdt) is shown in Figure 4 as an example.

We find that most of the interface cluster residues are conserved indicating that they play an important role in stabilizing the dimer. It is important to note that a cluster, with most of its residues conserved in homologs, implies enormous significance because the cluster residues are sequentially distant residues and they make spatial contact in the three-dimensional structure of the protein. Such conserved clusters could be structurally or functionally important for the protein. The fact that most of the interface clusters are highly conserved indicates that these cluster residues, which are sequentially distant but close in the three-dimensional structure, are important from the structural perspective. Though most of the interface cluster residues are conserved, the ones that are cluster centers are essentially highly conserved. This factor precisely strengthens our argument that cluster centers of the interface clusters are strongly involved in dimer stabilization.

Those residues that satisfy all the three above-mentioned conditions (conserved, high vector component and high δASA) have been predicted as 'hot spots' on the dimer interface. The protein dimer will possibly lose its structural credibility upon mutating such residues. The hot spots identified in all the 20 proteins of the data set are listed in Table I. The position of the vector components, the percentage of δASA and the extent of conservation of these residues are also given in Table I. In most of the cases, there is high correlation between residues that have high vector components, high δASA and high conservation. There are a few cases where the δASA is

```
CX4_NAJMO       ----------------LKCN-KLIPIAYKTCPEGKNLCYKMMLA-SKKMVPVKRG
CX1_NAJAT  MKTLLLTLVVVTIVCLDLGYTLKCN-KLIPIASKTCPAGKNLCYKMFMM-SDLTIPVKRG
CX3_NAJAT  MKTLLLTLVVVTIVCLDLGYTLKCN-KLVPLFYKTCPAGKNLCYKMFMV-ATPKVPVKRG
CX1_NAJSP  MKTLLLTTVVVTIVCLDLEYTLKCN-KLVPLFYKTCPAGKNLCYKMFMV-ATPKVPVKRG
CX1_NAJMO       ----------------LKCN-QLIPPFWKTCPKGKNLCYKMTMR-AAPMVPVKRG
CX1_NAJPA       ----------------LKCN-QLIPPFWKTCPKGKNLCYKMTMR-AAPMVPVKRG
CX2_NAJMO       ----------------LKCN-QLIPPFWKTCPKGKNLCYKMTMR-GASKVPVKRG
CX6_NAJAT       ----------------LKCN-QLIPPFYKTCAAGKNLCYKMFMV-AAPKVPVKRG
CXN_NAJAT  MKTLLLTLVVATIVCLDLGYTLKCN-QLIPPFYKTCAAGKNLCYKMFMV-AAPKVPVKRG
CX8_NAJAT  MKTLLLTLVVVTIVCLDLGYTLKCN-QLIPPFYKACAAGKNLCYKMFMV-AAPKVPVKRG
CX5_NAJAT  MKTLLLTLVVVTIVCLDLGYTLKCN-KLIPIASKTCPAGKNLCYKMFMV-ATPKVPVKRG
CX2_NAJAT  MKTLLLTLVVVTIVCLDLGYTLKCN-KLVPLFYKTCPAGKNLCYKMFMV-SNLTVPVKRG
CX4_NAJAT  MKTLLLTLVVVTIVCLDLGYTRKCN-KLVPLFYKTCPAGKNLCYKMFMV-SNLTVPVKRG
CXH_NAJNA  MKTLLLTMVVVTIVCLDLGYTLKCHNTQLPFIYKTCPEGKNLCFKATLKKFPLKFPVKRG
CX7_NAJAT  MKTLLLTLVVVTIVCLDLGYTLKCHNTQLPFIYNTCPEGKNLCFKATLK-FPLKFPVKRG
                                 **:    *   ::*. *****:*  :      .*****

CX4_NAJMO       CINVCPKNSALVKYVCCSTDRCN
CX1_NAJAT       CIDVCPKNSLLVKYVCCNTDRCN
CX3_NAJAT       CIDVCPKSSLLVKYVCCNTDRCN
CX1_NAJSP       CIDVCPKSSLLVKYVCCNTDRCN
CX1_NAJMO       CIDVCPKSSLLIKYMCCNTNKCN
CX1_NAJPA       CIDVCPKSSLLIKYMCCNTDKCN
CX2_NAJMO       CIDVCPKSSLLIKYMCCNTDKCN
CX6_NAJAT       CIDVCPKSSLLVKYVCCNTDRCN
CXN_NAJAT       CIDVCPKSSLLVKYVCCNTDRCN
CX8_NAJAT       CIDVCPKSSLLVKYVCCNTDRCS
CX5_NAJAT       CIDVCPKNSLLVKYVCCNTDRCN
CX2_NAJAT       CIDVCPKNSALVKYVCCNTDRCN
CX4_NAJAT       CIDVCPKNSALVKYVCCNTDRCN
CXH_NAJNA       CADNCPKNSALLKYVCCSTDKCN
CX7_NAJAT       CAATCPRSSSLVKVVCCKTDKCN
                *   **:.* *:* :**.*: *.
```

Fig. 4. Multiple sequence alignment of cardiotoxin using ClustalW. The first sequence shown in this alignment is the sequence of 1cdt, the protein in our data set. *, completely conserved residues; :, conserved mutations; ., partially conserved residues. The residues K50 and N45 (shown in bold), which are part of a new interface cluster, are completely conserved and partially conserved, respectively.

less but the vector component and the extent of conservation are very high and are also experimentally found to have an effect on dimerization (discussed in the next section). Hence, in the present study, residues which are conserved and have high vector components have been given higher weightage.

Experimental evidence for the predicted mutations. Single and multiple mutations carried out on these proteins were analyzed to correlate our results with the experimentally verified mutations. Although some of these mutations were designed for other purposes, we have reported only those which have relevant information regarding dimer interface stability. An extensive literature survey shows that only in five of the 20 proteins considered in our data set, experimental data is available on mutations that affect the dimer formation and stabilization (Table I). In these cases, the predicted mutations have correlated well with the experiments. In the case of phospholipase A2, we have predicted two mutations (Phe5 and Ile9) which have been carried out experimentally and were found to disrupt the dimer (Liu *et al.*, 1995). In triose phosphate isomerase, Arg98, which is one of the predicted residues, has been mutated and found to destabilize the interface (Mainfroid *et al.*, 1996). Similarly, in the case of interleukin8, tyrosyl tRNA synthetase and isocitrate dehydrogenase, some predicted mutations have already been experimentally verified and found to disrupt dimer formation. In all the other proteins present in the data set, no relevant experimental information is available because most of the other predicted mutations in all these proteins are yet to be carried out. Our prediction of the Glu104 mutation in triose phosphate isomerase, as having no effect on dimerization, was a false negative as this mutation was known to effect dimerization (Daar *et al.*, 1986). We had failed to predict this residue, as it did not form the cluster center even though it was part of the interface cluster with a reasonably high vector component value (Table II). This residue could have been a part of the set of predicted residues if we had used a less stringent criterion regarding the magnitude of the vector component. This could be done when the cluster size is big. If the cluster size was small, considering the first few vector components would be sufficient for predicting the crucial residues important for dimer stability. All our predicted mutations are listed in Table I. In the terminology of the Laplacian matrix, these predicted 'hot spots' have very high magnitude in the corresponding diagonal element of the Laplacian matrix. Apart from these, other experimentally tested mutations, which have not been predicted by us, are listed in Table V. Interestingly, most of these mutations do not have any effect on dimerization. These negative results can be rationalized from our present investigations. A careful analysis shows that these residues have not been predicted for the following reasons. They are either not a part of any interface cluster, in which case they may or may not have a high magnitude in the Laplacian matrix, or do not have a high vector component, due to very low magnitude in the corresponding diagonal element of the Laplacian matrix. Hence, looking at interface clusters in terms of their vector components can give us valuable information about residues involved in dimerization. Out of 55 predicted mutations, eight have been experimentally carried out and all of them have given positive results (Table I). Out of 16 non-predicted mutations that have

Table V. Experimentally tested mutations which have not been predicted by the present method

Protein	Experimentally tested mutation	Reason for not predicting by the present method[a]	Effect of mutation on dimerization	Reference
Cardiotoxin	K2	1	–ve	Lo *et al.*, 1998
	D57	1	–ve	
Interleukin 8	F21, L49, S14, Y13	1	–ve	Hammond *et al.*, 1996
Spo0B	H30	1	–ve	Tzeng *et al.*, 1998
Mannose binding protein	H189, I207	1	–ve	Iobst *et al.*, 1994
Phospholipase A2	Y52, Y73	2	–ve	Maliwal *et al.*, 1994; Dupreur *et al.*, 1992a,b
	Y22, F106	1	–ve	
	L19, L20	1	–ve	Lee *et al.*, 1996
	Q4, E6	2	–ve	Liu *et al.*, 1995
	L2, V3	1	–ve	Liu *et al.*, 1995
Uteroglobin	D46	2	–ve	Barnes *et al.*, 1996
	Y21, T60	2	–ve	Dunkel *et al.*, 1995
Triose phosphate isomerase	H95	2	–ve	Borchert *et al.*, 1995
	K13	1	–ve	Borchert *et al.*, 1995
	N65	2	–ve	Williams *et al.*, 1999
	E104	2	+ve	Daar *et al.*, 1986

[a]Reason for not predicting by the present method: 1, residue is not a part of interface cluster; 2, residue is a part of an interface cluster, however is not the first or second highest vector component residue.

been carried out only one has yielded a positive result (Table V). Thus, the algorithm has yielded correct results in >90% of the cases. These experiments do validate our argument that identifying interface clusters could essentially give us information on residues involved in dimer formation and stabilization, and that we have a rational method to identify residues that stabilize the dimer interface. Therefore, we can infer that these interface cluster residues and especially the cluster centers, are structurally and functionally important for the protein. The monomers fail to associate upon mutation of the residues, thereby hampering the formation of the functionally significant dimer.

Another convincing support for our prediction method is its experimental verification on the dimerization of the α-subunit of RNA polymerase (Kannan *et al.*, 2001a). A few residues were predicted that could destabilize the dimer, based on a similar analysis. The experiments were then carried out and it was found that mutation of the predicted residues indeed disrupts the dimer interface. All previous mutation experiments on the proteins in the data set have been carried out either randomly or by just considering the change in the accessible area of the residue when the monomer dimerizes and other residue-wise interactions. We now have provided a rational method, which actually takes into consideration the spatial interactions amongst residues and the clustering of such interactions at interfaces. This has proved to be more effective than predictions based on other traditional methods. Also, the residues, which satisfy the criteria of conservation and high δASA, can be scored on the basis of vector components to get a rank-ordered list of 'hot spots'.

Analysis of monomer clusters that get expanded on dimerization

Apart from predicting the hot spots from the interface clusters, we have also analyzed the clusters in the monomers that get expanded on dimerization. Most often, we have a crystal structure of the monomer and from the biochemistry of the monomer we know that it is a functional dimer or it interacts with other monomers for its activity. In such cases, it is relevant to predict the possible dimerization sites on the monomer. We have analyzed the side chain clusters in the monomers to get some insight to the possible dimerization sites.

Identification of exposed and conserved clusters in the monomer. As mentioned earlier, we do see that some clusters present in the monomer get expanded or strengthened after dimerization. This suggests that there is some kind of seeding that is present in the monomer, which gets strengthened upon dimerization. We have analyzed these strengthened clusters in all the 20 homodimers to look for a specific clustering pattern in these expanded clusters in terms of size and nature of amino acids. We find that most of the strengthened clusters are those which are formed when the contact criterion used is as high as 10–12%. The size of these clusters varies from three to six residues per cluster. Also, the preferred amino acid residues in such seeding clusters are arginine, phenylalanine, histidine, tyrosine and glutamic acid.

Once the clusters in the monomer are determined using high contact criteria, these clusters are characterized further based on the ASA and extent of conservation of component residues as well as the number of preferred residues present in the cluster. The 'surface clusters' are identified by looking at the ASA of the cluster residues. After identifying the surface clusters, we examine these surface clusters thoroughly for the presence of frequently occurring residues at the interface, which include arginine, histidine, phenylalanine, tyrosine and glutamic acid. In most cases, the number of exposed clusters when a high contact criterion is used is as small as two to five clusters. The number of exposed clusters with the preferred amino acids would be still smaller. The 'conserved clusters' amongst these clusters are the likely sites of dimerization on the monomer. Even if there is more than one cluster which is exposed, conserved and also has the preferred amino acid residues, we could still rank these clusters according to their tendency to dimerize, based on the extent of conservation of the residues forming these clusters and the number of preferred residues present. Hence, we can localize the dimerization sites

Table VI. Identification of dimerization sites on the monomers

Protein	Contact Criterion (%)	Clusters in Monomer	Accessible Surface Area (ASA)[a]	Conservation of interface clusters[b]
Phospholipase A2	12	**D99, F5, Y52, H48** [c]	B,E,E,P	P,T,T,T
Triose Phosphate Isomerase	12	R3, R189, D227; **N65, R98, F102, E104, F108**;	E,E,B ; P,E,E,P,E ;	P,T,M; T,T,P,T,M;
Cytochrome C	11	R12, W58, F82, F125; **Q22, D39, 43R**;	B,P,B,B ; E,E,E ;	T,M,T,M; P,P,P;
Rubisco	14#	W184,I190,K226; **K31,E119,Y120**; E432,Q412,H430; K82,E73,Y72; R6,Y7,D68;	B,B,B ; E,E,E ; E,P,E ; P,E,P ; E,B,B ;	P,N,T; P,P,P; P,N,P; P,P,P; P,T,P;
Thymidylate Synthase	12	R234,R243,D246;R222,R225,H255; V130,F150,Y181,H147,Y94; **E137,W101,N134,Q97**;	E,E,B ; E,E,E ; B,B,B,B,B ; E,E,E,B ;	N,N,P; M,N,P; M,N,T,T,T; M,M,T,T;
Tryptophan Repressor	8 *	**M66, E70, R63, E59, E60 ; 43L,44T,47E**	E,E,E,E,E ; E,E,E;	P,P,P,P,M ; P,P,T;
Glutathione Reductase	11	D308,H312,P293;P160,T176,D178; D227,H434,K420;Y407,R413,D81; **H351,R347,R38**; D461,Y399,K457;	E,E,B ; E,E,E ; B,B,P ; E,E,B ; E,E,E ; E,E,P ;	N,N,P; P,P,T; T,P,T ; N,N,N M,P,P ; M,M,P ;
Iron Superoxide Dismutase	12	H27, H31, Y9 ; **Y34, H74, Q70** ;	E,P,E ; E,P,B ;	P,T,M ; T,T,P ;
Malate Dehydrogenase	11	K169,F222,W217;F29,W257,K294; **D251,R156,K247**;S145,K121,D116; E318,T189,N185;P288,W183,I181; K311,Y191,F307,I223;	E,B,P ; B,E,E ; E,E,E ; E,E,E ; E,E,B ; B,B,B ; E,B,E,B ;	M,P,M ; M,M,P T,M,P ; P,P,P M,P,P ; T,P,N N,P,PM ;
Citrate Synthase	10	Y318,P372,Y330,H377;S152,D257, L260,N149;W109,E113,R117;Q223, H340,F336,L227;Q106,W94,L6; **L250,H246,E420**;W306,L309,Q364; L58,R329,V374,H238;Y385,F224, Y219,M387,Y194,Y190,E389,Y392; W114,W209,P183,F213,I179;M176, Y172,K7; F397,L269,H235,Y231; K290,Q287,Y304,E291;	P,B,B,B ; B,B, E,E ; P,P,E ; E, E,B,B ; E,B,B ; E,E,E ; E,P,E; P,B,B,B ; E,B, P,B,E,B,E,B ; B,B,B,B,B ; B, B,E ; B,B,B,B ; E,E,P,E;	M,P,P,M ; P,P M,M ; N,N,P; N, P,M,P ; P,P,T P,M,T ; P,M,N ; M,P,P,P ; P,N, P,N,M,P,P,N ; N,P,P,N,M ; N M,M ; P,M,T,T ; N,N,P,N ;
Spo0B	10	352F,321F,362F; 377R,388E,349Y ; **318F,292L,283F ;**	B,B,E, ; P,E,E ; P,P,P ;	T,T,P ; N,P,N ; N,T,T;
Tyrosyl tRNA Synthetase	10	208K,207R,204E;279E,275K,9W;273L,277E, 274S;240W,245K,242D;263V,295Q,292R;**148 M,170M,167F**;255W,307H,251F;177F,34Y, 189Q,199I;123N,176D,126W;266Y,223V,231 F;10R,6E,57R,272F,56R;45H,48H,44L;	E,E,P ; E,E,E ; P,E, E ; P,E,P ; B,B,E ; P,B,P ; B,P,B ; B,B, B,B ; B,B,P ; B,P,P ; B,E,B,E,E ; E,P,B ;	N,N,N ; N,N,N ; N,P, N ; N,N,N ; P,N,N ; P,N,P ; P,P,N ; N.P, P,P ; P,T,M ; N,P,P ; P,N,N,N,N ; T,N,P ;
Aspartate amino Transferase	10	189H,193H,160Y;284F,101R,281D;319W, 260F,256Y;228F,40Y,326M;**263Y,258K, 225Y;**24F,27D,380Y;134W,184D,127R,156R, 179E;396N,32K,400L;329R,43E,332R;195P, 386F,362F;	P,B,P ; B,E,E ; B, B,P ; B,P,B ; E,E, P ; E,E,B ; B,B,P,E, E ; P,P,B ; P,E,E ; B, P,P ;	P,P,T ; N,N,N ; P, N,P ; P,P,P ; P,T, M ; N,P,M; T,N,N,N, N; N,P,N; P,P,N; T, N,P ;
Isocitrate Dehydrogenase	11	274E,265K,258I;369W,57Y,61R; **292R, 296Y,288Q**;144P,133Y,139S;404K,407E, 349D;327N,119R,155N;241F,238E,302M; 391Y,350K,345Y;**229H,242K,279D**;	E,E,E; B,P,E; E, P,P; P,P,E; E,E, P; B,P,B; B,B,B; P,E,E; B,E,P;	N,N,N; N,P,N; N. P,P; N,N,N; N,N. N; N,P,P; M,M,N; N,P,P; P,N,P;

#Since the protein is large, there were exposed clusters even at 14% cut-off.
*Since the protein is small, no clusters were obtained at contact criterion higher than 8%. The first clusters were seen at this contact criterion only.
[a]ASA: E, exposed (>20% surface area exposed); P, partially exposed (5–20% surface area exposed); B, buried (<5% surface area exposed).
[b]Conservation of residues: T, totally conserved; M, conserved mutations; P, partially conserved; N, not conserved.
[c]The exposed, conserved clusters with the preferred amino acids are underlined. The expanded interface clusters are shown in bold.

on the monomer to these clusters. If we do not find any cluster satisfying these criteria, then we could reduce the contact criterion down to 8% to get a new set of clusters which can again be subjected to the same set of rules. Table VI shows the monomer clusters in 14 proteins where at least one of the clusters has got expanded on dimerization. We have identified the possible clusters of dimerization (underlined) based on the above mentioned criteria. The selection criteria have done extremely well in identifying all the interface clusters (bold, underlined). A few clusters which have been identified as possible dimerization sites but are actually not involved in dimerization (underlined, not bold) indeed do not rank top if the conservation criterion is made more stringent.

Out of the 20 proteins used in the data set, 14 of them show such seeding clusters which are identified using the above mentioned criteria (Table VI). Four other proteins (1cdt, 1il8,

1utg and 2gn5) are very small proteins with approximately 60–80 residues per monomer. These proteins do not show seeding clusters, until the criterion is reduced to 4%. This could be because these monomers do not have a well defined buried core as in the case of bigger proteins. So the ratio of buried to exposed residues in such proteins is much smaller leading to difficulty in selectively identifying exposed residues and clusters, which are of importance for dimerization. Two other proteins, 1msb and 3gap, of sizes 115 and 209 residues, respectively, do show exposed, conserved clusters at a high contact criterion, but these clusters do not participate in the homodimer interface even when the contact criterion is reduced down to 8%. Possibly, these clusters are involved in interaction with other proteins. There is experimental evidence to show that one of these proteins (1msb) interacts with various other proteins like CD14, serine proteases, etc. (Wallis and Dodd, 2000; Chiba *et al.*, 2001).

Thus, the present analysis of looking for exposed and conserved clusters in the monomer with the preferred amino acids, using high contact criteria, could help us to identify dimerization sites on the monomer. This method actually narrows down the search space for the identification of dimerization sites on the monomer. Instead of analyzing all exposed surfaces or residues in the monomer, we can restrict our analysis to such exposed and conserved clusters. This method has the potential to evolve as an effective one for predicting possible dimerization sites on the monomer. The limitation of this method is that it does not perform well on smaller proteins with less than approximately 100 amino acid residues.

Conclusions

The graph-theoretic algorithm, which considers the global topology of protein structures, has been successfully implemented to obtain side chain clusters at dimer interfaces of proteins. Analyses of these side chain clusters indicate that both charged and hydrophobic residues are involved in stabilizing the dimer interface. However, the interface is neutral because of the presence of oppositely charged residues. Arginine, histidine, phenylalanine, tyrosine and glutamic acid seem to be the most preferred residues at the dimer interface. Residues important for dimer formation and stabilization have been predicted using the present graph-theoretic algorithm and the predicted mutations have correlated well with experimental results. Hence, we have a robust method for predicting residues that play a significant role in stabilizing dimers. We have also ventured into predicting dimerization sites on the monomer, which has been extremely successful in this limited data set and hence, the algorithm could very well evolve as a good method for prediction of dimerization sites on the monomer. We would like to emphasize the fact that the major advantage of this method is that we are analyzing interfaces on the basis of the side chain clusters detected using a graph-theoretic algorithm, where the clustering residues are sequentially non-adjacent but spatially close to each other. Analysis of the clusters of such spatially connected residues yields better results than just analyzing pair-wise residue interactions.

Acknowledgements

The authors would like to thank the Super Computer Education and Research Centre and the Distributed Informatics Centre of the Indian Institute of Science, Bangalore, India, for the computational facilities provided by them. One of the authors (K.V.B.) would like to thank the Centre for Scientific and Industrial Research, India, for the fellowship offered.

References

Artymiuk,P.J., Poirrette,A.R., Grindley,H.M., Rice,D.W. and Willet,P., (1994) *J. Mol. Biol.*, **243**, 327–344.
Bahar,I., Atilgan,A.R. and Erman,B. (1997) *Fold Des.*, **2**, 173–181.
Bairoch,A. and Apweiler,R. (2000) *Nucleic Acids Res.*, **28**, 45–48.
Barnes,H.J., Nordlund-Moller,L., Nord,M., Gustafsson,J., Lund,J. and Gillner M. (1996) *J. Mol. Biol.*, **256**, 392–404.
Bedoulle,H. and Winter,G. (1986) *Nature*, **320**, 371–373.
Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) *Nucleic Acids Res.*, **28**, 235–242.
Bogan,A.A. and Thorn,K. (1998) *J. Mol. Biol.*, **280**, 1–9.
Borchert,T.V., Kishan,K.V., Zeelen,J.P., Schliebs,W., Thanki,N., Abagyan,R., Jaenicke,R. and Wierenga,R.K. (1995) *Structure*, **3**, 669–679.
Chiba,H., Sano,H., Iwaki,D., Murakami,S., Mitsuzawa,H., Takahashi,T., Konishi,M., Takahashi,H. and Kuroki,Y. (2001) *Infect. Immun.*, **69**, 1587–1592.
Chothia,C. and Janin,J. (1975) *Nature*, **256**, 705–708.
Connolly,M.L. (1993) *J. Mol. Graph.*, **11**, 139–141.
Daar,I.O., Artymiuk,P.J., Phillips,D.C. and Maquat,L.E. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 7903–7907.
Dunkel,R., Vriend,G., Beato,M. and Suske,G. (1995) *Protein Eng.*, **8**, 71–79.
Dupureur,C.M., Yu,B.Z., Mamone,J.A., Jain,M.K. and Tsai,M.D. (1992a) *Biochemistry*, **31**, 10576–10583.
Dupureur,C.M., Yu,B.Z., Jain,M.K., Noel,J.P., Deng,T., Li,Y., Byeon,I.J. and Tsai,M.D. (1992b) *Biochemistry*, **31**, 6402–6413.
Glaser,F., Steinberg,D.M., Vakser,I.A. and Ben-Tal,N. (2001) *Proteins Struct. Funct. Genet.*, **43**, 89–102.
Hall,K.M. (1970) *Manage. Sci.*, **17**, 219–229.
Hammond,M.E., Shyamala,V., Siani,M.A., Gallegos,C.A., Feucht,P.H., Abbott,J., Lapointe,G.R., Moghadam,M., Khoja,H., Zakel,J. and Tekamp-Olson,P. (1996) *J. Biol. Chem.*, **271**, 8228–8235.
Heringa,J. and Argos,P. (1991) *J. Mol. Biol.*, **220**, 151–171.
Horcher,M., Rot,A., Aschauer,H. and Besemer,J. (1998) *Cytokine*, **10**, 1–12.
Hu,Z., Ma,B., Wolfson,H. and Nussinov,R. (2000) *Proteins Struct. Funct. Genet.*, **39**, 331–342.
Humphrey,W., Dalke,A. and Schulten,K. (1996) *J. Mol. Graph.*, **141**, 33–38.
Hurley,J.H., Chen,R. and Dean,A.M. (1996) *Biochemistry*, **35**, 5670–5678.
Iobst,S.T., Wormald,M.R., Weis,W.I., Dwek,R.A., Drickamer,K. (1994) *J. Biol. Chem.*, **269**, 15505–15511.
Janin,J., Miller,S. and Chothia,C., (1988) *J. Mol. Biol.*, **204**, 155–164.
Jones,S. and Thornton,J.M. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
Jones,S. and Thornton,J.M. (1997) *J. Mol. Biol.*, **272**, 121–132.
Kannan,N. and Vishveshwara,S. (1999) *J. Mol. Biol.*, **292**, 441–464.
Kannan,N. and Vishveshwara,S. (2000) *Protein Eng.*, **13**, 753–761.
Kannan,N., Preethi,C., Pallavi,G., Vishveshwara,S. and Dipankar,C. (2001a) *Protein Sci.*, **10**, 46–54.
Kannan,N., Selvaraj,S., Gromiha,M.M. and Vishveshwara,S. (2001b) *Proteins Struct. Funct. Genet.*, **43**, 103–112.
Koch,I., Kaden,F. and Selbig J. (1992) *Proteins*, **12**, 314–323.
Larsen,T.A., Olson,A.J. and Goodsell,D.S. (1998) *Structure*, **6**, 421–427.
Lee,B.I., Yoon,E.T. and Cho,W. (1996) *Biochemistry*, **35**, 4231–4240.
Lengauer,T. and Rarey,M. (1996) *Curr. Opin. Struct. Biol.*, **5**, 402–406.
Liu,X., Zhu,H., Huang,B., Rogers,J., Yu,B.Z., Kumar,A., Jain,M.K., Sundaralingam,M. and Tsai,M.D. (1995) *Biochemistry*, **34**, 7322–7334.
Lo,C.C., Hsu,J.H., Sheu,Y.C., Chiang,C.M., Wu,W., Fann,W. and Tsao,P.H. (1998) *Biophys. J.*, **75**, 2382–2388.
Mainfroid,V., Mande,S.C., Hol,W.G., Martial,J.A. and Goraj,K. (1996) *Biochemistry*, **35**, 4110–4117.
Maliwal,B.P., Yu,B.Z., Szmacinski,H., Squier,T., Binsbergen,J., Slotboom,A.J. and Jain,M.K. (1994) *Biochemistry*, **33**, 4509–4516.
Miller,S., Janin,J., Lesk,A.M. and Chothia,C. (1987) *Nature*, **328**, 834–836.
Mitchell,E.M., Artymiuk,P.J., Rice,D.W. and Willet,P. (1990) *J. Mol. Biol.*, **212**, 151–166.
Palma,P.N., Krippahl,L., Wampler,J.E. and Moura J.J. (2000) *Proteins Struct. Funct. Genet.*, **39**, 372–384.
Patra,S.M. and Vishveshwara,S. (2000) *Biophys. Chem.*, **84**, 13–25.
Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia,A. (1997) *J. Mol. Biol.*, **271**, 511–523.
Sternberg,M.J., Gabb,H.A. and Jackson,R.M. (1998) *Curr. Opin. Struct. Biol.*, **8**, 250–256.

Thompson,J.D., Higgins,D.J. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
Tzeng,Y.L., Zhou,X.Z. and Hoch,J.A. (1998) *J. Biol. Chem.*, **273**, 23849–23855.
Valdar,W.S. and Thornton,J.M. (2001) *Proteins Struct. Funct. Genet.*, **42**, 108–124.
Wallis,R. and Dodd,R.B. (2000) *J. Biol. Chem.*, **275**, 30962–30969.
Williams,J.C., Zeelen,J.P., Neubauer,G., Vriend,G., Backmann,J., Michels,P.A., Lambeir,A.M. and Wierenga,R.K. (1999) *Protein Eng.*, **12**, 243–250.
Xu,D., Tsai,C.J. and Nussinov,R. (1997) *Protein Eng.*, **10**, 999–1012.

Received July 27, 2001; revised December 21, 2001; accepted January 28, 2002