Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Network approach for capturing ligand-induced subtle global changes in protein structures

Anshul Sukhwal,‡ Moitrayee Bhattacharyya and Saraswathi Vishveshwara*

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

‡ Current address: National Center for Biological Sciences, Bangalore 560 065, India.

Correspondence e-mail: sv@mbu.iisc.ernet.in

Received 6 December 2010
Accepted 24 February 2011

Ligand-induced conformational changes in proteins are of immense functional relevance. It is a major challenge to elucidate the network of amino acids that are responsible for the percolation of ligand-induced conformational changes to distal regions in the protein from a global perspective. Functionally important subtle conformational changes (at the level of side-chain noncovalent interactions) upon ligand binding or as a result of environmental variations are also elusive in conventional studies such as those using root-mean-square deviations (r.m.s.d.s). In this article, the network representation of protein structures and their analyses provides an efficient tool to capture these variations (both drastic and subtle) in atomistic detail in a global milieu. A generalized graph theoretical metric, using network parameters such as cliques and/or communities, is used to determine similarities or differences between structures in a rigorous manner. The ligand-induced global rewiring in the protein structures is also quantified in terms of network parameters. Thus, a judicious use of graph theory in the context of protein structures can provide meaningful insights into global structural reorganizations upon perturbation and can also be helpful for rigorous structural comparison. Data sets for the present study include high-resolution crystal structures of serine proteases from the S1A family and are probed to quantify the ligand-induced subtle structural variations.

1. Introduction

A comparison of protein structures conventionally involves the evaluation of root-mean-square deviations (r.m.s.d.s). R.m.s.d.s primarily capture changes at the backbone level and are mainly useful in identifying the commonality or differences at the fold or secondary-structure level. Small synchronized variations at the level of side-chain interactions as a result of ligand binding or environmental changes are of immense functional relevance (Bhattacharyya & Vishveshwara, 2010; Pargellis *et al.*, 2002) and such variations can also permeate to distal sites in the protein (Ghosh & Vishveshwara, 2008). However, side-chain reorientations are rarely addressed as a collective general feature owing to a lack of a simple understanding of the same in molecular perspective at a global level. Variations at the side-chain level are usually visited for a specific set of residues that are either near the active site or exhibit large backbone-level conformational variations (Shi *et al.*, 2006; Done *et al.*, 2008; Latz *et al.*, 2007).

Binding of ligands or changes in the environment are usually associated with conformational changes in proteins,

Table 1
Summary of data set I (72 structures from 22 different proteins).

The names of the proteins forming the reduced data set (data set II) are highlighted in bold.

Protein name (PDB code)	PDB codes of other chosen structures
Apolipoprotein a	1i71, 1kiv, 3kiv, 4kiv
Complement factor B (1rrk)	**1rtk**, **1rrk**
Complement C2	2i6q, 2odp, 2odq
Complement factor D (1dst)	**1bio**, **1dic**, 1hfd, 1dst
Kallikrein-6 (1gvl)	1lo6, 1l2e, 1gvl
Prostatin (3e1x)	3dfj, **3dfl**, 3e0n, **3e0p**, **3fvf**, **3e1x**, 3gyl
Azurocidin	1a7s, 1ae5, 1fy1, 1fy3
Ancrod	2aip, 2aiq
Plasma kallikrein	2anw, 2any
Chymase (1nn6)	**1klt**, **1t31**, **1pjp**, **1nn6**
Kallikrein-7	2qxh, 2qxj
Trypsin-1 (1utk)	**1hj8**, **1utj**, **1utl**, **1utm**, **1utk**
Urokinase-type plasminogen activator (2o8t)	**2o8u**, **2o8w**, **2r2w**, **1sqo**, **2o8t**
Trypsin (1os8)	**1oss**, 1sgt, 2fmj, **1os8**
Coagulation factor XI	1zhm, 1zsj, 2fda
Chymotrypsin-like elastase family member 1	1b0e, 1l1g
Anionic trypsin-2	1and, 1j14
Cationic trypsin	1auj, 1az8, 1bju, 1c1p, 1btw
Cathepsin G	1cgh, 1t32
Human leucocyte elastase	1hne, 1ppg
Granzyme M	2zgc, 2zgh, 2zgj
Protein elastase	1ela, 1qr3

both drastic and subtle (Campanacci *et al.*, 2003; Yang *et al.*, 2007). Often, the drastic changes at the backbone or secondary-structure level are more easily identified compared with the global subtle rewiring of the side chains. The dogma of local perturbations induced by ligand binding usually predominates in structural studies, and the percolation of these local changes to distal sites causing a global synchronization in response to ligand binding is largely unexplored. However, the global permeation of these subtle local changes is often of functional relevance, as seen in the case of allosteric proteins or where such changes are accompanied by ligand binding at secondary sites (Nettles *et al.*, 2004; Ghosh & Vishveshwara, 2007; Bhattacharyya *et al.*, 2009).

In this study, a novel graph theoretical metric is proposed based on protein side-chain interactions to capture the intricate rearrangements at the side-chain level which are largely elusive from backbone-level analyses. Here, protein-structure networks (PSNs; Kannan & Vishveshwara, 1999) and network parameters (such as cliques or communities; Palla *et al.*, 2005) are used to unravel the global reorientations in structures upon ligand binding. Furthermore, it is found that a comparative study of clique-forming residues for two structures is a better determinant of structural similarity (both at the backbone and side-chain level) in contrast to the conventional methods of analyses for structural deviations. The permeation of the effect of ligand binding to distal sites can also be efficiently monitored using the concept of cliques or communities for the PSN in a global manner instead of focusing only on the ligand-binding pocket. This global percolation of ligand-induced stress in the structure may or may not be of biological relevance depending on the system of interest. However, the metric used is general and can be applied to practically any well chosen data set of interest with the aim of studying ligand-induced subtle global conformational changes.

The S1A family of serine proteases was chosen for our analysis (Barrett & Rawlings, 1995). Serine proteases are digestive enzymes but they also exert a functional role in inflammation, blood clotting, the immune system and neural plasticity (Pham, 2006; Yoshida & Shiosaka, 1999; Walsh & Ahmad, 2002; Choo *et al.*, 2010). Most of the members of the chymotrypsin (S1) family are endopeptidases which differ widely in specificity. The linear order of catalytic triad residues in the polypeptide chains of the enzymatically active members of family S1 is His, Asp and Ser. Serine proteases are inhibited by a diverse group of compounds (serpins) including synthetic inhibitors and natural peptides (Rubin, 1996; Whisstock *et al.*, 2010). The diversity in function offered by this family of enzymes as well as the availability of a large number of high-resolution crystal structures (both native and ligand-bound structures) aptly suits the need of our present study. However, the concepts developed and the methods used in this article are highly generalized and can find potential use in structural studies as well as in studying structure–function relationships in any class of proteins of interest.

2. Methods

2.1. Creation of the data set

The data set is curated from the 'Serine Protease Home' at http://www.biochem.wustl.edu/~protease/. This database contains information from five clans, 30 families and around 700 serine protease sequences. The S1A family (Barrett & Rawlings, 1995) was selected for our investigations. The protein structures chosen for the present study are high-resolution (<2.5 Å) crystal structures with the chain length of the native protein being almost identical to that of the different ligand-bound forms (this condition is imposed for efficient comparison of network parameters). Using these criteria, 22 proteins from the S1A family of serine proteases (a total of 72 structures) are chosen for the study (data set I; Table 1 and Table S1[1]) of the relationship of r.m.s.d.-common clique residues. Furthermore, for the study of ligand-induced global conformational variations, an additional constraint is imposed where the native structure and at least one of the ligand-bound structures are available at high resolution. A reduced data set of eight native proteins (a total number of 27 structures are chosen including the native and the liganded forms; data set II) is obtained (highlighted in Table 1 and Table S1[1]). Additionally, a data set of 109 high-resolution (<2 Å) crystal structures of cationic trypsin (data set III; Table S2[1]) is curated to exhibit the long-range effect of ligand binding for a large number of different ligand-bound structures from the same protein.

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: MV5039). Services for accessing this material are described at the back of the journal.

2.2. Construction of a protein structure network

A protein structure network or protein structure graph (PSN/PSG) efficiently portrays the noncovalent side-chain interactions from a global perspective. The details of the construction of such a graph at a particular interaction cutoff ($I_{\min}$) and the implications of such graphs have been discussed in detail previously (Kannan & Vishveshwara, 1999; Brinda & Vishveshwara, 2005). Protein structure networks are constructed by considering amino-acid residues as nodes, and edges are constructed between the nodes on the basis of noncovalent interactions between them (as evaluated from the normalized number of contacts between them) for each system. The noncovalent interaction between side-chain atoms of amino-acid residues are considered (with the exception of Gly where the C^{α} atom is considered), ignoring the interaction between sequence neighbours. The interaction between two residues i and j has been quantified previously in our laboratory as

$$I_{ij} = \frac{n_{ij}}{(N_i \times N_j)^{1/2}} \times 100, \qquad (1)$$

where n_{ij} is number of distinct atom pairs between the side chains of amino-acid residues i and j, which come within a distance of 4.5 Å, and N_i and N_j are the normalization factors for residues i and j. The pair of amino-acid residues having interaction strength (I_{ij}) greater than a user-defined cutoff ($I_{\min}$) are connected by edges to give a protein structure network (PSN) for a given interaction strength $I_{\min}$. Generally, $I_{\min}$ values in the PSNs vary from 1 to 15%. The lower the $I_{\min}$ value, the higher is the connectivity.

In order to choose an optimum $I_{\min}$ for analyses, the largest cluster profile is constructed for structures in data set II as described in detail in Brinda & Vishveshwara (2005) at different $I_{\min}$ values (1–8%). Clusters are obtained using the depth-first search algorithm. The optimal interaction strength in a protein structure is exhibited at the $I_{\min}$ at which the size of the largest noncovalently connected cluster (LClu) undergoes a transition.

2.3. Network parameters associated with high connections

In network theory, cliques or communities (Palla *et al.*, 2005, 2007; Adamcsek *et al.*, 2006; schematically represented in Figs. 1*a* and 1*b*) represent tightly connected regions of the network. This concept has been used to study various networks, both social (Palla *et al.*, 2007) and biological (Kose *et al.*, 2001; Alexander *et al.*, 2009). In the context of PSN, these parameters are used to identify the rigid regions in the protein structures and to recognize the ligand-induced conformational changes (Bhattacharyya *et al.*, 2009; Ghosh & Vishveshwara, 2008; Bhattacharyya & Vishveshwara, 2010). In this study, the PSNs are constructed as described above. These PSNs are analyzed in term of cliques or communities on the basis of the following definitions.

(i) ***k*-Clique**. A k-clique is defined as a set of k nodes (points represented by amino acids) in which each node is connected to all the other nodes. Fig. 1(*a*) schematically shows a $k = 3$ clique in which all the three nodes are connected to each other.

(ii) ***k*-Clique community**. A k-clique community is defined as a union of smaller k-cliques that share node(s). According to mathematical literature, a k-clique community has been defined as the assemblage of k-cliques that can be percolated through a series of adjacent k-cliques. In the present study, a k-clique community is one in which two k-cliques share

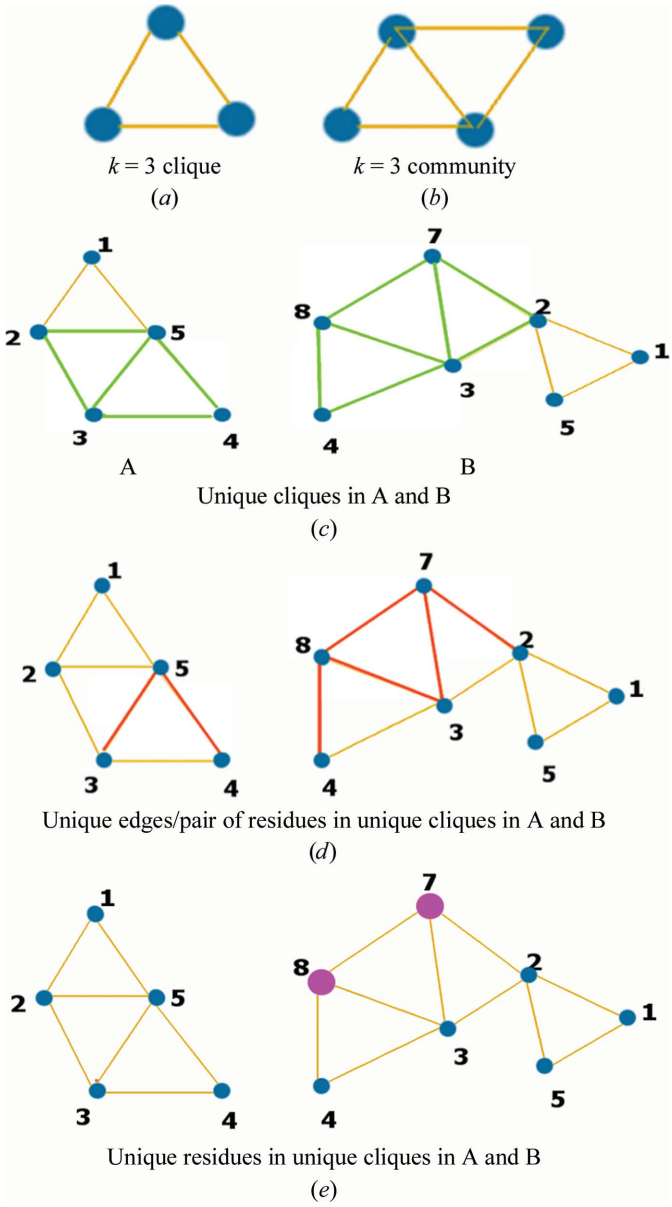

Figure 1
(*a*, *b*) A schematic description of the $k = 3$ clique and community formed by two $k = 3$ cliques sharing $k - 1$ edges, respectively. (*c*) Unique cliques in A and B with respect to each other. The $k = 3$ clique (nodes 1–2–5) are common to both A and B, whereas the other $k = 3$ cliques are exclusive to each of them. These unique cliques are highlighted as green triangles for A and B. (*d*) Depiction of the unique edge or pair of residues in the unique cliques of A and B [as shown in (*c*)]. The edges 3–4 and 4–5 are unique to A, whereas edges 3–8, 4–8, 2–7, 3–7 and 7–8 are unique to B (as highlighted with red lines). (*e*) Unique residues of unique cliques in A and B (node 7 and 8 for B, none for A) highlighted as purple spheres.

$k-1$ nodes. Fig. 1(*b*) schematically shows a $k = 3$ community where two $k = 3$ cliques share an edge (*i.e.* $3 - 1 = 2$ nodes).

(iii) ***k*-Clique or *k*-clique community finding algorithm**. The clique or community search is based on the algorithm proposed by Palla *et al.* (2005). *Cfinder* is used to obtain the cliques or communities from PSNs. In the majority of cases $k = 3$ cliques are obtained at the chosen $I_{min} = 3\%$. *k*-Clique communities with an overlap of $k - 1$ nodes are obtained using *Cfinder* (Adamcsek *et al.*, 2006).

The I_{min} values for analyses are also optimized based on the largest community profile for the structures in data set II as a function of different I_{min} values in the range 1–8%. The region of transition in the largest community size is considered to be of interest.

(iv) **Unique cliques (UC)**. Unique cliques are defined as those exclusively present in a particular structure with respect to another structure. In this study, the cliques present only in the ligand-bound states with respect to the corresponding native structures are termed unique. Such unique cliques represent tightly packed rigid regions in the liganded structures. Fig. 1(*c*) schematically highlights the unique cliques in models A and B with respect to each other. This is a comparison at the three-node (clique) level.

(v) **Unique edges of unique cliques (UEUC)**. A comparison of the edges constituting the unique cliques in the liganded structures with respect to the ones in the native state gives the unique edges of unique cliques (UEUC) for the liganded structures. Such UEUCs reflect increased pairwise connections (in liganded structures with respect to the native) between the residues in the liganded states. Fig. 1(*d*) schematically highlights the UEUC in A and B with respect to each other. This is a more rigorous comparison at the two-node (edge) level.

(vi) **Unique residues of unique cliques (URUC)**. A comparison of the residues constituting the unique cliques in the liganded structures with respect to those in the native state results in the unique residues of unique cliques (URUC). URUC represents the participation of new residues in UC for the liganded state with respect to the cliques in the native structure. Fig. 1(*e*) schematically highlights the URUC in B with respect to A (there are no URUC in A with respect to B). This is the most rigorous structural comparison at the single-node (residue) level. The rewiring of the network subsequent to ligand binding is responsible for such rearrangements in connectivity at the clique/edge/residue level.

2.4. Dividing the protein structure into three 'tiers'

The average distance of the boundaries of protein structures from the active-site triad (Ser, His, Asp) are calculated using *VMD* (Humphrey *et al.*, 1996). For all the structures in data set II this average distance (d) is found to be approximately 24 Å. The protein structures in data set II are theoretically divided into three tiers, near ($<d/3$), mid ($d/3$–$2d/3$) and far ($>2d/3$), based on their proximity from the active-site triad. The native and ligand-bound structures are compared through cliques. The uniqueness is identified through (i) an entire clique motif (UC), (ii) unique edges of unique cliques (UEUC) and (iii) unique residues of unique cliques (URUC) (a detailed explanation of these parameters is given in the preceding section). The differences in the number of these three parameters between the native and the ligand-bound structures are identified in each tier in order to capture the long-range effects of ligand binding. A normalization of these numbers based on the volume of each tier (Table S4) is performed in order to eliminate the bias arising from small differences in volume, especially for the 'far' tier.

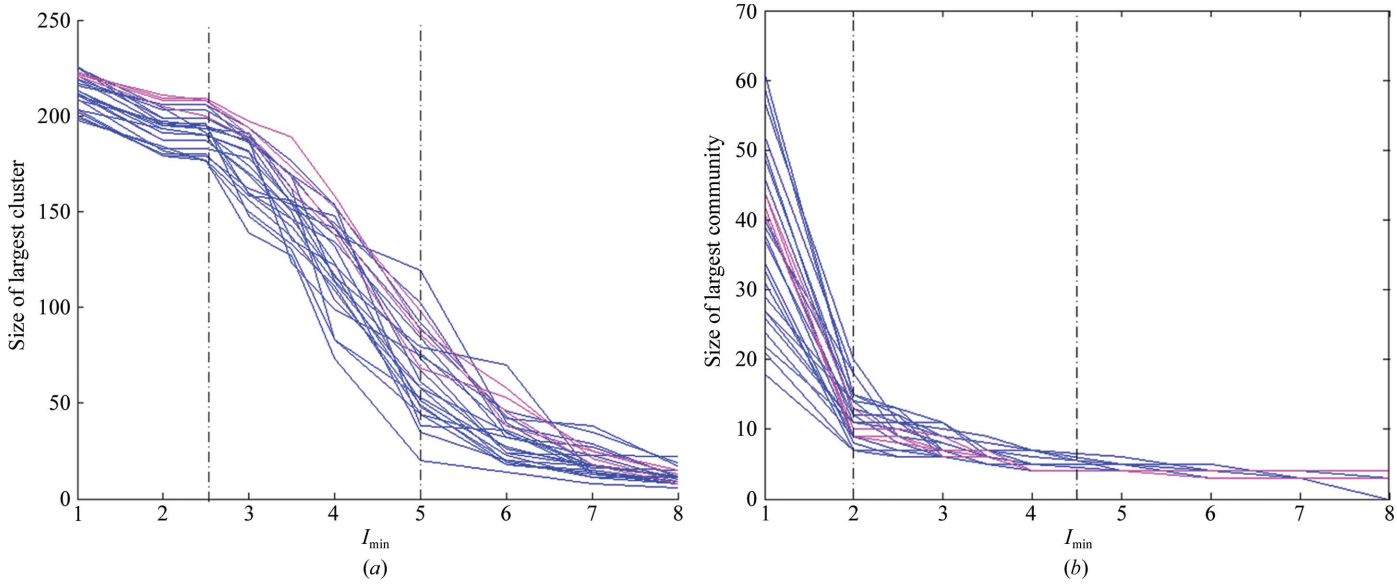

Figure 2
(*a*) Plot of size of largest cluster *versus* I_{min} (1–8%) showing a transition between 2.5 and 5%. (*b*) Plot of size of largest community *versus* I_{min} (1–8%) showing a transition between 2 and 4%.

2.5. Comparison of URUC with results from standard network parameters

The standard network parameters used for comparison are residue-wise connectivity (*i.e.* the number of connections made by a residue) and the residue-wise clustering coefficient. The residue-wise changes in the number of connections (Δconn) and the clustering coefficient (Δccfs) in going from native structures to the liganded states are evaluated for all the structures in data sets II, data set III and the elastase data set using

$$\Delta\mathrm{conn}^{i} = \mathrm{conn}^{i}_{\mathrm{lig}} - \mathrm{conn}^{i}_{\mathrm{nat}} \quad (2a)$$

$$\Delta\mathrm{ccfs}^{i} = \mathrm{ccfs}^{i}_{\mathrm{lig}} - \mathrm{ccfs}^{i}_{\mathrm{nat}}, \quad (2b)$$

where $\mathrm{conn}^{i}_{\mathrm{lig}}$ and $\mathrm{conn}^{i}_{\mathrm{nat}}$ are the number of connections and $\mathrm{ccfs}^{i}_{\mathrm{lig}}$ and $\mathrm{ccfs}^{i}_{\mathrm{nat}}$ the clustering coefficient of residue i in the liganded and the corresponding native state, respectively ($i = 1, \ldots, N$, where N is the total number of residues).

Again, the clustering coefficient of a vertex (or node) is a measure of cliquishness (Watts & Strogatz, 1998; *i.e.* it quantifies the tendency of its neighbours to be a clique) of that node and is defined as $\mathrm{ccfs}^{\mathrm{vertex}}$ = no. of edges between the vertex's neighbours/total possible no. of edges between the vertex's neighbours (Watts & Strogatz, 1998).

3. Results and discussion

Functionally diverse proteins are included (Table S1) from the S1A family in our data set with widely different sequence lengths for each protein class. The PSNs are constructed for each structure and used for further analysis as detailed in the subsequent paragraphs.

3.1. General network properties and the choice of I_{min}

I_{min} is a measure of the extent of connectivity in the PSNs. A lower I_{min} is associated with higher connectivity and *vice versa*. Previous studies from our group have shown that the optimal interaction strength in a protein structure is exhibited at an I_{min} at which the size of the largest noncovalently connected cluster (LClu) undergoes a transition (Brinda & Vishveshwara, 2005). Here, the largest cluster profile is obtained for all the structures in data set II as a function of I_{min} (Fig. 2*a*) and a profile similar to that of the proteins from earlier studies is obtained with a transition in the size of the LClu between I_{min} of about 2.5 and 5%. Additionally, a largest community (LComm) profile (see §2) for structures in data set II is obtained at different I_{min} values (Fig. 2*b*). The size of the LComm also shows a transition in the I_{min} range of about 2–4%. At lower I_{min} values (pretransition region) the network is densely connected, whereas at higher I_{min} values (post-transition region) the network connectivity is very sparse. These values of I_{min} are the extremes of the range. The transition regions in LClu and LComm profiles reflect the most meaningful connections and thus all our further investigations are mainly focused on the I_{min} range 2–5%, emphasizing the results at I_{min} = 3%. However, a case study on elastase shows that consistent results are obtained for different I_{min} values corresponding to the transition region.

3.2. Low r.m.s.d.: do not ignore!

A global view of the synchronization at the level of side-chain interactions for different protein structures can be achieved through the network approach, which has highlighted certain common features that are general to all proteins. One such common feature investigated by us earlier is the existence of densely connected clusters (cliques or communities) of interacting side chains, which percolate through a large part of the protein structure network (Deb *et al.*, 2009). This analysis has opened up the possibility of a rigorous comparison of protein structures which shows subtle deviations through side-chain reorientations.

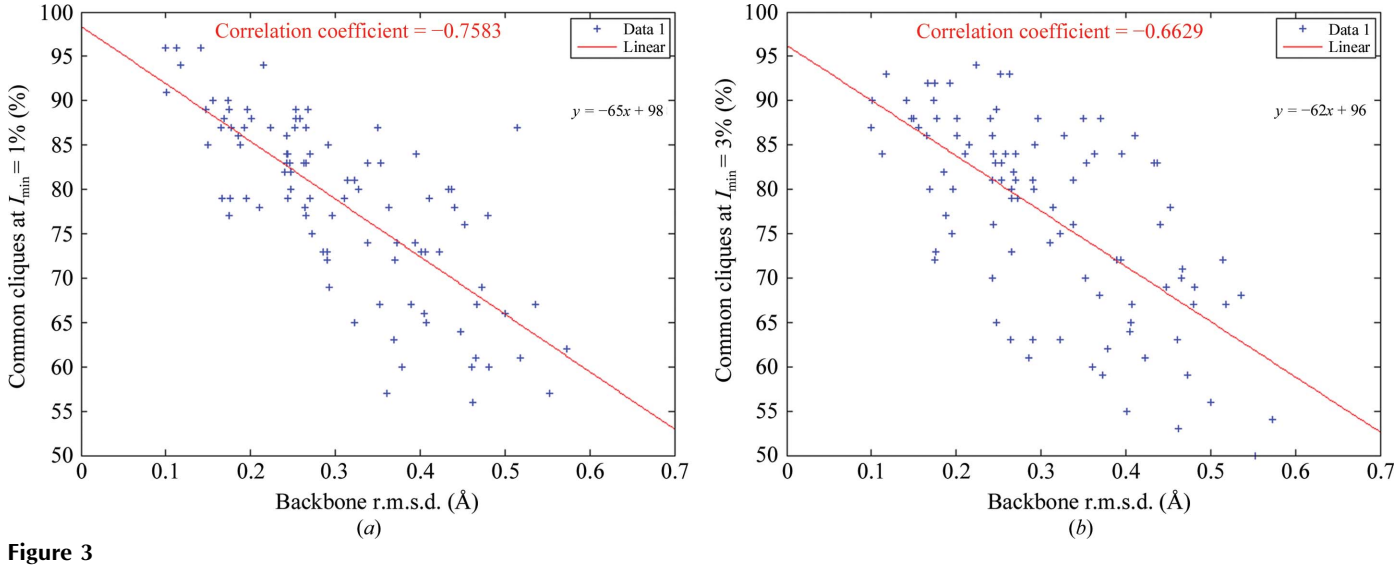

Figure 3
Percentage of common cliques *versus* backbone r.m.s.d. between all the structures of the selected proteins in data set I (22 proteins and 72 high-resolution crystal structures) at (*a*) I_{min} = 1% and (*b*) I_{min} = 3%.

Table 2
Comparison of URUC (in the liganded structures of data set II with respect to their corresponding native structures) with results from Δccfs and Δconn.

All the URUCs either overlap with the residues showing an increase in clustering coefficient or connectivity (in italic type) or both (in bold type). The % overlap with Δccfs and Δconn is also reported.

Protein	Ligand-bound structure(s)	Unique resolution of unique cliques (URUC) with respect to native†	Identity of URUC with respect to Δccfs (%)	Identity of URUC with respect to Δconn (%)
Complement factor D native (1dst)	1bio	**2L 14Y 29V 35W 39A 74R** *92L 94Q* **108L** *117V* **121T 135G 137R** *148V* **154C 160H 164I 170C** *186P* **188V** *191G* **193L 199S 202R 212I 216V 227L**	93	85
	1dic	**8E** *10H* ***12R*** **29V** *35W* **39A 58H** *64E* **74R** ***94Q*** **120G 122L** *143H* **147P 154C 160H 164I** *170C* **199S 202R 212I**	86	81
Trypsin native (1os8)	1oss	*33L 41G* **43G 69K** *70V 87K* **117A 122G 157A 169Q** *176M* **190I** *206Y* **223L**	86	71
Trypsin-1 native (1utk)	1hj8	**1I 4G 7C 12Q 19N 40H** *47E* **49R 73H 85I 114T 116C 118V** *135Q* **136C 137L 141I 154M** *211F*	84	100
	1utj	**14H 19N 31N** *34W* **40H 42Y** *47E* **49R 52E 62E 64F 70V 73H 85I 86M 88I 103V 154M 181V 184G 186L 188G 208V** *211F 219M*	88	96
	1utl	**14H 19N 39A** *42H 47E 49R* **52E 62E 64F 70V 72R 73H 75N** *83N 85I 86M 103V* **154M 181V** *184G 211F*	90	67
	1utm	**73H** *85I 211 F*	100	33
Prostatin native (3e1x)	3dfl	*18I 28G 43F* **52Y 86G 154R 158N** *161Y 171H 172F 174Q* **177M** *214W 215G*	86	50
	3e0p	**1I** *4G 38G 43F* **56L 57G** *60Q* **65S** *68A* **69K 100I 106I 118F 126V 138L** *141P* **143P** *145Q 147L* **202V** *207Y 213S* **228T** *233Y*	83	71
	3fvf	**1I** *4G 60Q* **93L 118F 126V 138L** *141P* **143P** *145Q 147L 154R* **158N** *172F* **202V** *207Y* **237I**	88	65
Complement factor B native (1rrk)	1rtk	*3N 33V* **38V 41R** *57V* **88T** *109H* **114M** *119H* **123G** *125P* **136L** *156G 157V 161V* **181K** *182V 196I* **220Q 223Q** *230R* **254T** *257D* **289K 292A** *294I 357F 358V 359S* 361E *368K* **384D 387Y** *399E* **418N** 420C 423D *427P* **436F** *438Q* **447V 449V** *452R 456V* **457P 459 H 460 A**	74	64
Chymase native (1nn6)	1KLT	**1I 2I** 3G *4G* **6E** *14Y* **15M** *17Y* **29F** *32G* **46C 54T 58H 59N** *60I 61T 88H* **108L 111P 121R** *123C 126V 128W 130R* **132G** *139D* **140T 141L** *142Q* 144V **148L 149M 164L** *169G 171P 175K 181D* **183G** *185P 187L* **188C** *191V* **193Q 200R** *202D* **204K** *215Y*	81	64
	1pjp	**1I 2I** 3G *4G* **6E** *14Y* **15M 29F 47A 54T 58H 59N** *61T* **69K 71E 73I 77R** *96K* **108L 113Q 117V 120G 122M** *123C 128W 130R* **132G** *139D* **140T 141L 142Q** 144V **147R** *165Q 175K 181D* **186L 187L 188C** *193Q 194G 198Y* **200R** *202D* **204K** *208V 209F* **213S**	85	71
	1t31	**1I 2I** 3G *4G* **6E** *14Y* **15M** *20I 28K* **29F** *30C 32G 33F* **35I 54T 58H 59N** *61T* **69K** *85T* **87H** *88H 90I* **108L 110F 121R 123C** *128W 130R* **132G** *138S 139D* **140T 141L 142Q** 144V **149M 159D** *161D* **164L** *169G 171P 175K 181D* **185P 186L 187L 188C** *195I* **200R** *202D 203A* **204K** *215Y* **216R 220N**	84	66
Kallikrein-6 native (1gvl)	1l2e	**2V 35W 37L 52G 54H 63Q 70R 88L 90R 124K 125T 130F 131P 164D 167Y 170D 171S 172C 173Q** *190V* **198G 199S** *206Y 216I*	100	88
Urokinase-type plasminogen activator native (2o8t)	2o8u	*14W* 18I *58R* **64N** *68E 94L* **113I** *136E* **144P** *148K 160C 165Y 170V 176C 182W 186S 196V* **219D** *224Y*	58	58
	2o8w	*15F 18I 40W* **45T 49I 57Y** *62G 63R* **69N 70T 73E** *87Y 98I* **100L** *101L* **120I 128D** *129P* **133T 138T** *143E* **151P** *155K* **156M** *161L 167C 172Y 173Y 176E* **177V** *181M* **182L 183C** *203V 218W* **226D** *230V* **232T** *233R 237F* **245T**	73	73
	2r2w	**7T 45T 49I** *62G* **70T** *73E 87Y 128D* **156M** *172Y 189W* **226D 233R** *235S*	79	71
	1sqo	**45T 49I** *63R* **69N** *73E 87Y 128D* **141G** *143E* **151P** *155K* **156M** *193S 196G* **226D** *233R 235S*	82	59

† The residue numbers are based on cleaned PDB files with the residues being renumbered from 1.

For our complete data set (data set I) of 72 high-resolution serine protease structures from the S1A family, it is seen that all-atom, backbone and C^{α} r.m.s.d. values between all the structures of a protein (the range of r.m.s. deviation is between 0.1 and 0.6 Å, with the r.m.s.d. values at the all-atom, backbone and C^{α} levels having a high correlation value of 0.98–0.99 amongst themselves) could indeed be correlated with differences in the side-chain network connections (at $I_{\min} = 1\%$ and $I_{\min} = 3\%$) in terms of common cliques to various extents (Fig. 3, Table S3). It is evident from Fig. 3 that backbone r.m.s.d.s between two structures are reflected in the percentage of common cliques between these structures in an inverse

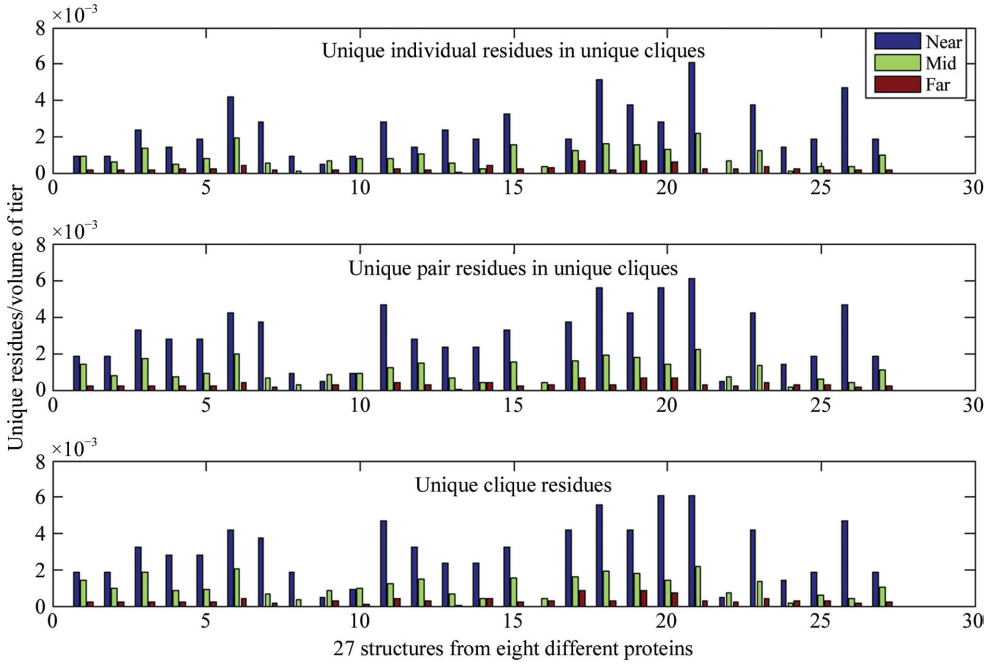

Figure 4
Long-range effect of ligand-induced conformational changes in 27 structures from eight different proteins of the S1A family of serine proteases. The effect is captured in terms of UC, UEUC and URUC in the three tiers around the ligand-binding site: 'near', 'mid' and 'far' with respect to a native structure.

manner for our complete data set I with a high anticorrelation value of ∼0.75 (at I_{min} = 1%) and ∼0.66 (at I_{min} = 3%) (similar values of anticorrelation are seen for the r.m.s.d.s at the C^{α} and all-atom levels; results not shown). A general trend that lower r.m.s.d. difference is accompanied by a higher percentage of common cliques is prominent from the data, implying that even a small r.m.s.d. value can give us a clue to the rewiring of side-chain interactions in a protein as a function of environment. This general methodology can be used in tracking the ligand-induced subtle conformational variations in proteins at the level of side-chain interactions in molecular detail from a global perspective. Such an analysis of comparing changes through r.m.s.d. and combining it with the network approach (*i.e.* the percentage of common cliques) has a potential application as a powerful tool in identifying subtle structural differences with high relevance to function in homologous proteins.

3.3. Ligand/environment-induced conformational variations at proximal/distal sites

In this article, the global effects of ligand binding towards side-chain interactions is efficiently captured using the concepts of graph theory, thereby throwing light on the readjustments in the protein structure network both proximal to and distal from the site of ligand binding. One such interesting graph parameter is cliques, which portray the transmission of perturbation to distal sites in an efficient manner. In order to quantify the changes at different distances from the site of ligand binding, the structure of every protein in our reduced data set II is divided into three regions (near, within *d*/3; mid, between *d*/3 and 2*d*/3; far: beyond 2*d*/3; where *d* is the average distance of the catalytic triad from the farthest points in any direction of all the proteins in the data set, approximately equal to 24 Å; described in detail in §2). The number of residues participating in unique cliques (UC), unique edges of unique cliques (UEUC) and unique residues of unique cliques (URUC) (detailed in Figs. 1*c*–1*e* and §2) in each of these three regions for all the structure in data set II are computed and the results are summarized in Fig. 4 and Tables S5–S7 (at I_{min} = 3%). It is evident from Fig. 4 that the conformational variations (as determined by the above-mentioned parameters) are at a maximum in the 'near' region, proximal to the site of ligand binding, as expected. However, the 'mid' and 'far' regions also bear evidence of perturbation (in terms of these three parameters), validating the argument that ligand-induced conformational changes are not only restricted to the site of binding but can also permeate to longer distances by the rewiring of the structure network at the level of side-chain interactions.

To further substantiate our observations with eight different proteins (*i.e.* 27 structures in data set II) as described above, similar calculations are performed on another data set (data set III) comprising 109 cationic trypsin structures [108 liganded structures and one native structure (1s0q)]. Similar results are obtained with the three parameters being distributed in the 'mid' and 'far' tiers, in addition to the 'near' tier (Fig. S1). Thus, it is evident that a larger data set containing structures from a single protein (cationic trypsin) in different ligand-bound states also yields similar outcomes in terms of global permeation of local perturbation upon ligand binding. The computation of the matrices and the cliques can now be evaluated through our webserver *GraProStr* available at http://vishgraph.mbu.iisc.ernet.in/GraProStr/index.html (Vijayabaskar *et al.*, 2011).

3.4. Comparison of results with standard network parameters

In order to link the results based on cliques in the preceding section with standard network parameters, the overlap between the unique residues of unique cliques (URUC) and the residues showing an increase in connectivity (Δconn > 0) or clustering coefficient (Δccfs > 0) (see §2) in a ligand-bound structure (with respect to the native structure) is computed. Interestingly, the URUC obtained for the liganded structures in data sets II and III also show an increase in connectivity or

clustering coefficient or both with respect to the corresponding native structures (Table 2, Fig. 5). The overlap is comparatively better with the residues showing Δccfs > 0. All the URUC for most of the structures in data sets II and III are a subset of a combined list of residues exhibiting either Δconn > 0 or Δccfs > 0 with respect to the corresponding native structures [Table 2, Fig. S2; the number of residues in the combined list is almost double the number of URUCs (details for elastase summarized in Table S8)]. The increase in both connectivity and clustering coefficient indicate that the residue has acquired more connections. However, a clique provides higher order connectivity information (Deb *et al.*, 2009) capturing the connectivity as a global property elucidated at a detailed node level and further screening the residues identified to be important from the parameters such as connectivity and clustering coefficient. These data clearly portray the validity and robustness of the results derived from comparison of cliques, demonstrating that cliques are a sensitive parameter for structure comparison, in contrast to standard network parameters such as connectivity or clustering coefficient alone.

3.5. Conformational reorientation upon analogous ligand binding: a case study with elastase

The ligand-induced subtle global conformational variations are considered to be of functional relevance and such changes are often elusive from conventional structural studies. It has been shown previously that a judicious use of graph theoretical parameters can be of immense use in unravelling such subtle changes (Bhattacharyya & Vishveshwara, 2009). Here this point is elaborated for different ligand-bound structures of serine proteases in general, with a case study on elastase. High-resolution crystal structures (1.7–2.1 Å) of elastase bound to three analogous inhibitors with different binding modes are available (Mattos *et al.*, 1994). These three structures of elastase (1ela, 1elb and 1elc) exhibit different preferences of ligand-binding subsites (Mattos *et al.*, 1995). On superposing these three crystal structures on the corresponding high-resolution native structure of elastase (1esa, 1.6 Å), significant backbone conformational changes are not observed (backbone and all-atom r.m.s.d. less than 0.22 Å). However, a detailed comparison at the level of side-chain interactions yields significant changes between the four different systems under study. The superposition of the residues comprising the catalytic triad and those in the substrate-binding pockets clearly show variations at the level of side-chain organization upon non-analogous binding of analogous ligands (Fig. 6) and such 'plasticity' has been referred to as the 'subtle induced fit of the active site as a result of ligand binding' (Mattos *et al.*, 2006). This local perturbation is percolated to distal sites mediated through the connectivity network at the side-chain level, as captured by the URUC. Elastase and its inhibitor-bound forms have been extensively investigated. Multiple-solvent crystal structures of elastase used to probe the various ligand-binding sites on elastase have mainly identified the active site as the potential pocket (Mattos *et al.*, 2006; Mattos & Ringe, 1996). Although a large number of biochemical and structural studies have been performed focusing on the active sites of elastase (Mattos *et al.*, 1994, 1995, 2006), few studies have probed the effect of distal residues on the catalysis and specificity of the enzyme (Hung & Hedstrom, 1998). It has been proposed by Hedstrom that the catalysis and specificity in serine proteases in general are not controlled by a small set of residues, but are determined by the properties of the 'entire protein network' (Hedstrom, 2002). In this study, such a global perspective is provided by the protein structure network and the percolation

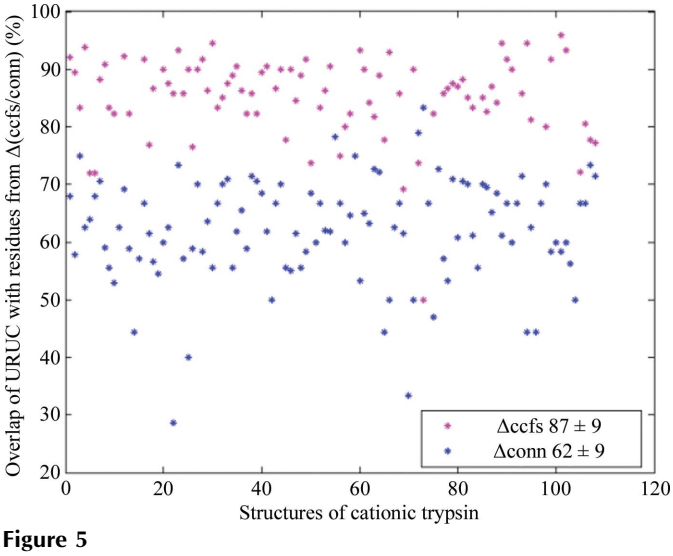

Figure 5
Plot of percentage overlap of URUC with residues showing increase in connectivity (conn) or clustering coefficient (ccfs) (with respect to the native structure 1s0q) for all the 108 liganded structures of cationic trypsin (data set III). The overlap is comparatively better with the residues showing Δccfs > 0.

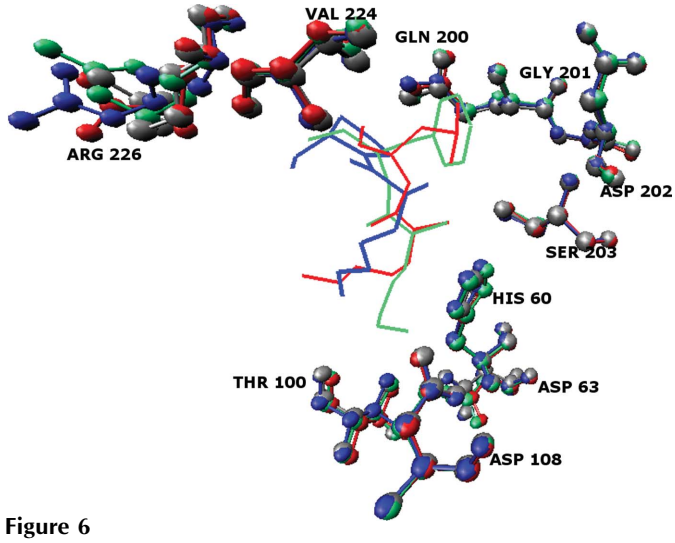

Figure 6
Pictorial depiction of the superposition of the active site (60H, 108D, 203S) and ligand-binding residues (63D, 100T, 200Q, 201G, 202D, 224V and 226R) for 1ela, 1elb and 1elc. Subtle variations are clearly exhibited at the level of side-chain orientations. The ligands for 1ela, 1elb and 1elc are depicted as blue, red and green lines and the corresponding residues are represented by CPK using the same colour code.

Table 3
Comparison of URUC [in three liganded structures of elastase with respect to 1esa (native)] with results from Δccfs and Δconn.

All the URUC either overlap with the residues showing an increase in clustering coefficient or connectivity (in italic type) or both (in bold type). The percentage overlap with Δccfs and Δconn is also reported.

Protein	Ligand-bound structure(s)	Unique residues of unique cliques (URUC) with respect to native	Identity of unique residues of unique cliques with respect to Δccfs (%)	Identity of unique residues of unique cliques with respect to Δconn (%)
Elastase native (1esa)	1ela	*34Q* **35Y 41W** *54W* **58A 59A 60H** *63D* **64R** *66L 69R 72V 85Q* **89V** *91K 92I* **94V 100T 112L** *113R* **124V** *136L 140S* **167T 225S** *228G* **231V** *234K* **255N**	90	62
	1elb	*27W 35Y* **41W** *48T 50I 56M* **59A 60H 64R** *66L* **89V** *92I* **94V 112L 124V** *143Y* **162Q** *198G 208H 215Y 224V* **229C**	77	68
	1elc	*21E 27W 34Q* **41W 58A 60H** *69R 72V 75H 85Q* **89V** *92I 93V 112L 124V 143Y 159T* **162Q** *251V* **255N**	80	45

of any local stress to distant sites is captured by comparison of the network parameters (cliques in this case).

Pretransition I_{min} values (*e.g.* 1%) again produce dense networks connecting most of the nodes in both the native and the ligand-bound structures. This leads to the identification of very few URUCs. However, a very stringent connectivity criterion ($I_{min} > 5\%$) gives rise to a sparse network in both the native and the ligand-bound structures, thus leading to small numbers of URUC. Meaningful comparison studies can be performed at the transition region between the densely and sparsely connected networks (as explained above). A profile of the number of unique cliques for 1ela, 1elb and 1elc with respect to 1esa (native) as a function of I_{min} clearly exhibits a peak at ∼2–3% (Figure S3a). Additionally, a comparison of URUC between each of the three ligand-bound systems of elastase with respect to 1esa (native) at different I_{min} values reveals a significant overlap between such residues in the I_{min} range of ∼2–4% (Figs. S3b and S4). Therefore, an I_{min} of 3% was chosen for all further analyses on elastase.

The URUCs in the inhibitor-bound systems (with respect to the native elastase) clearly reveal that the effect of differential binding of the three analogous inhibitors at the active site is permeated in a magnified manner throughout the network (Fig. 7). An extensive comparison among the URUCs for 1ela, 1elb and 1elc clearly portrays the differences in the global structural organizations for the three systems (Table S9). The reason for such differences in reorientations at the level of side-chain interactions is the differential perturbation at the ligand-binding sites by these three inhibitors. Interestingly, many such unique residues are identified at regions not immediately proximal to the ligand-binding pocket (*i.e.* in the 'mid' and 'far' tiers) (Table S10). The other proteins and their different ligand-bound forms considered in data set II also show a similar trend (Fig. S5). Such an insight could not be obtained by mere inspection of the crystal structures. The URUCs obtained for the three ligand-bound structures of elastase are compared with standard network parameters such as Δconn or Δccfs (see §2). All URUCs exhibit an increase in connectivity or clustering coefficient or both for the three liganded structures of elastase (Table 3). Also, an extensive SCA analysis of the S1A family of serine proteases yielded groups of statistically coupled residues which are proposed to have a possible role in specificity and catalysis (Suel *et al.*, 2003). Strikingly, a majority of the residues in these groups coincide with the URUCs identified for the three systems 1ela, 1elb and 1elc (Table S11).

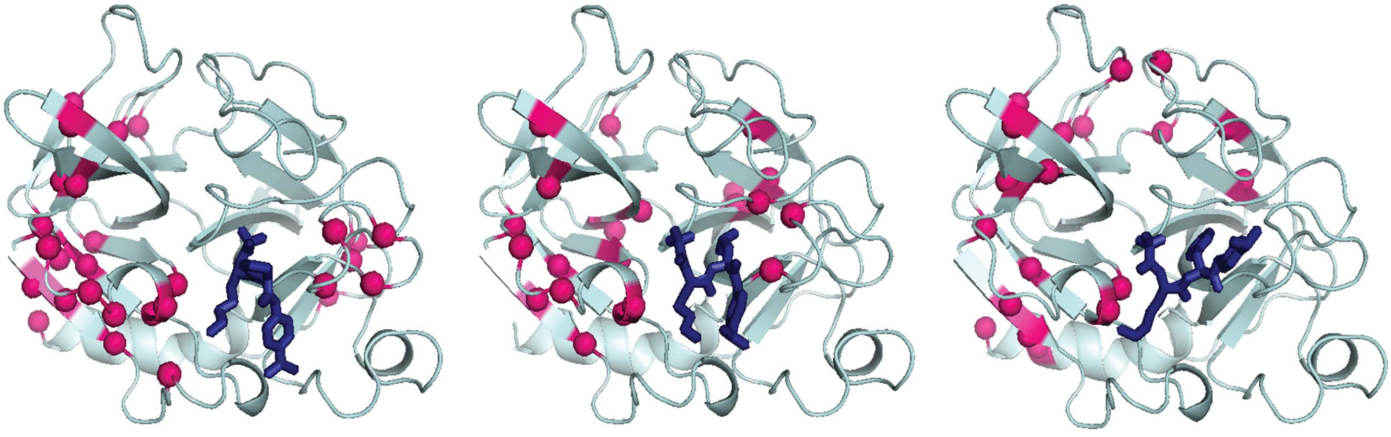

Figure 7
Effect of differential binding of three analogous inhibitors on global side-chain rewiring in elastase. The URUC with respect to the native elastase structure (1esa) for the three systems are topologically different. A slight change in the binding of the ligand shows distinct variations in conformational reorientation in the three systems at the side-chain level (Table S9). The protein backbone is represented as a light grey cartoon and the unique clique residues are depicted as van der Waals spheres. The ligands are represented as deep blue sticks.

4. Conclusions

Protein structure comparisons are generally performed at the level of backbone topology. Identification of a consolidated global noncovalent connectivity of the side chains has been a challenge to structural biologists. In this paper, a robust method is provided to identify subtle conformational changes owing to perturbations such as ligand binding or point mutations based on the concepts of graph theory.

Data sets of high-resolution crystal structures belonging to the serine protease S1A family are considered. It is demonstrated that small r.m.s.d. values between all the structures from the same protein show a correlation with differences in network parameters such as cliques (highly connected sets of residues). Structural differences between liganded and the corresponding native structures based on cliques are also obtained and our results show significant agreement with standard network parameters such as clustering coefficients and the number of connections evaluated at the residue level from protein structure networks.

This study has provided a reliable method for comparison of structural features at the detailed side-chain level, with very similar backbone topology. The outlined method may prove to be a powerful tool in investigating subtle changes in biologically important phenomena such as allostery, where ligand binding has a long-range effect that is sensed at a distal region. Furthermore, such an approach can also be employed in structural comparison in a functionally relevant fluctuating environment obtained from MD simulations or NMR studies to track the subtle global changes in a statistically relevant manner to provide structural biological insights.

5. Related literature

The following articles are also related to this work: Caughey (1989); Akaaboune *et al.* (1994); Debela *et al.* (1997); Demers *et al.* (1991); Feric *et al.* (2008); Gaboriaud *et al.* (1996); Gomis-Rüth *et al.* (2002); Huber & Bode (1978); Jin *et al.* (2005); Kershaw & Flier (2004); Lauritzen *et al.* (2005); Lu *et al.* (2006); Martini *et al.* (2010); Meyer-Hoffert *et al.* (2004); Moon *et al.* (2010); Seriramalu *et al.* (2010); Spraggon *et al.* (2009); Szmola & Sahin-Toth (2007); Talas *et al.* (2000); Tang *et al.* (2005); Torreira *et al.* (2009); Yamazaki & Aoki (1997); Zhao *et al.* (2007).

Support from the Department of Science and Technology for Mathematical Biology (DSTO773), Government of India, is acknowledged. MB thanks the Council for Scientific and Industrial Research (CSIR) for a fellowship.

References

Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I. & Vicsek, T. (2006). *Bioinformatics*, **22**, 1021–1023.
Akaaboune, M., Villanova, M., Festoff, B. W., Verdière-Sahuqué, M. & Hantaï, D. (1994). *FEBS Lett.* **351**, 246–248.
Alexander, R. P., Kim, P. M., Emonet, T. & Gerstein, M. B. (2009). *Sci. Signal.* **2**, e44.
Barrett, A. J. & Rawlings, N. D. (1995). *Arch. Biochem. Biophys.* **318**, 247–250.
Bhattacharyya, M., Ghosh, A., Hansia, P. & Vishveshwara, S. (2009). *Proteins*, **78**, 506–517.
Bhattacharyya, M. & Vishveshwara, S. (2009). *BMC Struct. Biol.* **9**, 8.
Bhattacharyya, M. & Vishveshwara, S. (2010). *BMC Struct. Biol.* **10**, 27.
Brinda, K. V. & Vishveshwara, S. (2005). *Biophys. J.* **89**, 4159–4170.
Campanacci, V., Lartigue, A., Hällberg, B. M., Jones, T. A., Giudici-Orticoni, M., Tegoni, M. & Cambillau, C. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 5069–5074.
Caughey, G. H. (1989). *Am. J. Physiol. Lung Cell. Mol. Physiol.* **257**, L39–L46.
Choo, Y. M., Lee, K. S., Yoon, H. J., Kim, B. Y., Sohn, M. R., Roh, J. Y., Je, Y. H., Kim, N. J., Kim, I., Woo, S. D., Sohn, H. D. & Jin, B. R. (2010). *PLoS One*, **5**, e10393.
Deb, D., Vishveshwara, S. & Vishveshwara, S. (2009). *Biophys. J.* **97**, 1787–1794.
Debela, M., Hess, P., Magdolen, V., Schechter, N. M., Steiner, T., Huber, R., Bode, W. & Goettig, P. (2007). *Proc. Natl Acad. Sci. USA*, **104**, 16086–16091.
Demers, C., Ginsberg, J. S., Brill-Edwards, P., Panju, A., Warkentin, T. E., Anderson, D. R., Turner, C. & Kelton, J. G. (1991). *Blood*, **78**, 2194–2197.
Done, S. H., Brannigan, J. A., Moody, P. C. & Hubbard, R. E. (1998). *J. Mol. Biol.* **284**, 463–475.
Feric, N. T., Boffa, M. B., Johnston, S. M. & Koschinsky, M. L. (2008). *J. Thromb. Haemost.* **6**, 2113–2120.
Gaboriaud, C., Serre, L., Guy-Crotte, O., Forest, E. & Fontecilla-Camps, J. C. (1996). *J. Mol. Biol.* **259**, 995–1010.
Ghosh, A. & Vishveshwara, S. (2007). *Proc. Natl Acad. Sci. USA*, **104**, 15711–15716.
Ghosh, A. & Vishveshwara, S. (2008). *Biochemistry*, **47**, 11398–11407.
Gomis-Rüth, F. X., Bayés, A., Sotiropoulou, G., Pampalakis, G., Tsetsenis, T., Villegas, V., Avilés, F. X. & Collect, M. (2002). *J. Biol. Chem.* **277**, 27273–27281.
Hedstrom, L. (2002). *Chem. Rev.* **102**, 4501–4524.
Huber, R. & Bode, W. (1978). *Acc. Chem. Res.* **11**, 114–122.
Humphrey, W., Dalke, A. & Schulten, K. (1996). *J. Mol. Graph.* **14**, 33–38.
Hung, S.-H. & Hedstrom, L. (1998). *Protein Eng.* **11**, 669–673.
Jin, L., Pandey, P., Babine, R. E., Gorga, J. C., Seidl, K. J., Gelfand, E., Weaver, D. T., Abdel-Meguid, S. S. & Strickler, J. E. (2005). *J. Biol. Chem.* **280**, 4704–4712.
Kannan, N. & Vishveshwara, S. (1999). *J. Mol. Biol.* **292**, 441–464.
Kershaw, E. E. & Flier, J. S. (2004). *J. Clin. Endocrinol. Metab.* **89**, 2548–2556.
Kose, F., Weckwerth, W., Linke, T. & Fiehn, O. (2001). *Bioinformatics*, **17**, 1198–1208.
Latz, E., Verma, A., Visintin, A., Gong, M., Sirois, C. M., Klein, D. C., Monks, B. G., McKnight, C. J., Lamphier, M. S., Duprex, W. P., Espevik, T. & Golenbock, D. T. (2007). *Nature Immunol.* **8**, 772–779.
Lauritzen, B., Lykkesfeldt, J., Djurup, R., Flodgaard, H. & Svendsen, O. (2005). *Pharmacol. Res.* **51**, 509–514.
Lu, H., Hou, Q., Zhao, T., Zhang, H., Zhang, Q., Wu, L. & Fan, Z. (2006). *J. Immunol.* **177**, 1171–1178.
Martini, P. G. *et al.* (2010). *BMC Immunol.* **11**, 43.
Mattos, C., Bellamacina, C. R., Peisach, E., Pereira, A., Vitkup, D., Petsko, G. A. & Ringe, D. (2006). *J. Mol. Biol.* **357**, 1471–1482.
Mattos, C., Giammona, D. A., Petsko, G. A. & Ringe, D. (1995). *Biochemistry*, **34**, 3193–3203.
Mattos, C., Rasmussen, B., Ding, X., Petsko, G. A. & Ringe, D. (1994). *Nature Struct. Mol. Biol.* **1**, 55–58.
Mattos, C. & Ringe, D. (1996). *Nature Biotechnol.* **14**, 595–599.
Meyer-Hoffert, U., Wingertszahn, J. & Wiedow, O. (2004). *J. Invest. Dermatol.* **123**, 338–345.

Moon, J.-Y., Yim, E.-Y., Song, G., Lee, N. H. & Hyun, C.-G. (2010). *Eurasia. J. Biosci.* **4**, 41–53.
Nettles, K. W., Sun, J., Radek, J. T., Sheng, S., Rodriguez, A. L., Katzenellenbogen, J. A., Katzenellenbogen, B. S. & Greene, G. L. (2004). *Mol. Cell*, **13**, 317–327.
Palla, G., Barabási, A. L. & Vicsek, T. (2007). *Nature (London)*, **446**, 664–667.
Palla, G., Derényi, I., Farkas, I. & Vicsek, T. (2005). *Nature (London)*, **435**, 814–818.
Pargellis, C., Tong, L., Churchill, L., Cirillo, P. F., Gilmore, T., Graham, A. G., Grob, P. M., Hickey, E. R., Moss, N., Pav, S. & Regan, J. (2002). *Nature Struct. Mol. Biol.* **9**, 268–272.
Pham, C. T. (2006). *Nature Rev. Immunol.* **6**, 541–550.
Rubin, H. (1996). *Nature Med.* **2**, 632–633.
Seriramalu, R., Pang, W. W., Jayapalan, J. J., Mohamed, E., Abdul-Rahman, P. S., Bustam, A. Z., Khoo, A. S. & Hashim, O. H. (2010). *Electrophoresis*, **31**, 2388–2395.
Shi, J., Koeppe, J. R., Komives, E. A. & Taylor, P. (2006). *J. Biol. Chem.* **281**, 12170–12177.
Spraggon, G., Hornsby, M., Shipway, A., Tully, D. C., Bursulaya, B., Danahay, H., Harris, J. L. & Lesley, S. A. (2009). *Protein Sci.* **18**, 1081–1094.
Suel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. (2003). *Nature Struct. Mol. Biol.* **10**, 59–69.
Szmola, R. & Sahin-Toth, M. (2007). *Proc. Natl Acad. Sci. USA*, **104**, 11227–11232.
Talas, U., Dunlop, J., Khalaf, S., Leigh, I. M. & Kelsell, D. P. (2000). *J. Invest. Dermatol.* **114**, 165–170.
Tang, J., Yu, C. L., Williams, S. R., Springman, E., Jeffery, D., Sprengeler, P. A., Estevez, A., Sampang, J., Shrader, W., Spencer, J., Young, W., McGrath, M. & Katz, B. A. (2005). *J. Biol. Chem.* **280**, 41077–41089.
Torreira, E., Tortajada, A., Montes, T., Rodríguez de Córdoba, S. & Llorca, O. (2009). *J. Immunol.* **183**, 7347–7351.
Vijayabaskar, M. S., Niranjan, V. & Vishveshwara, S. (2011). *Open Bioinformatics J.* **5**, 53–58.
Walsh, P. N. & Ahmad, S. S. (2002). *Essays Biochem.* **38**, 95–111.
Watts, D. J. & Strogatz, S. H. (1998). *Nature (London)*, **393**, 440–442.
Whisstock, J. C., Silverman, G. A., Bird, P. I., Bottomley, S. P., Kaiserman, D., Luke, C. J., Pak, S. C., Reichhart, J. M. & Huntington, J. A. (2010). *J. Biol. Chem.* **285**, 24307–24312.
Yamazaki, T. & Aoki, Y. (1997). *J. Leukoc. Biol.* **61**, 73–79.
Yang, C.-S., Xin, H.-W., Kelley, J. B., Spencer, A., Brautigan, D. L. & Paschal, B. M. (2007). *Mol. Cell. Biol.* **27**, 3390–3404.
Yoshida, S. & Shiosaka, S. (1999). *Int. J. Mol. Med.* **3**, 405–409.
Zhao, G., Yuan, C., Wind, T., Huang, Z., Andreasen, P. A. & Huang, M. (2007). *J. Struct. Biol.* **160**, 1–10.