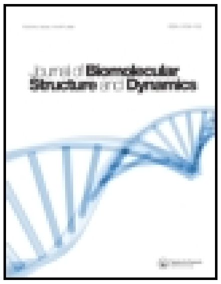

Journal of Biomolecular Structure and Dynamics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tbsd20

Understanding the role of domain-domain linkers in the spatial orientation of domains in multi-domain proteins

Ramachandra M. Bhaskara [a], Alexandre G. de Brevern [b c d e] & Narayanaswamy Srinivasan [a]

[a] Molecular Biophysics Unit, Indian Institute of Science, Bangalore, 560012, India

[b] DSIMB, INSERM UMR-S 665, Paris, F-75739, France

[c] Univ Paris Diderot, Sorbonne Paris Cité, UMR 665, F-75739, Paris, France

[d] INTS, Paris, F-75739, France

[e] Laboratoire d'Excellence GR-Ex, F75737, Paris, France

Version of record first published: 19 Dec 2012.

To cite this article: Ramachandra M. Bhaskara , Alexandre G. de Brevern & Narayanaswamy Srinivasan (2012): Understanding the role of domain-domain linkers in the spatial orientation of domains in multi-domain proteins, Journal of Biomolecular Structure and Dynamics, DOI:10.1080/07391102.2012.743438

To link to this article: http://dx.doi.org/10.1080/07391102.2012.743438

Understanding the role of domain–domain linkers in the spatial orientation of domains in multi-domain proteins

Ramachandra M. Bhaskara[a], Alexandre G. de Brevern[b,c,d,e] and Narayanaswamy Srinivasan[a]*

[a]*Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India;* [b]*DSIMB, INSERM UMR-S 665, Paris F-75739, France;* [c]*Univ Paris Diderot, Sorbonne Paris Cité, UMR 665, F-75739 Paris, France;* [d]*INTS, Paris F-75739, France;* [e]*Laboratoire d'Excellence GR-Ex, F75737 Paris, France*

Communicated by Ramaswamy H. Sarma

(*Received 25 August 2012; final version received 22 October 2012*)

Inter-domain linkers (IDLs)' bridge flanking domains and support inter-domain communication in multi-domain proteins. Their sequence and conformational preferences enable them to carry out varied functions. They also provide sufficient flexibility to facilitate domain motions and, in conjunction with the interacting interfaces, they also regulate the inter-domain geometry (IDG). In spite of the basic intuitive understanding of the inter-domain orientations with respect to linker conformations and interfaces, we still do not entirely understand the precise relationship among the three. We show that IDG is evolutionarily well conserved and is constrained by the domain–domain interface interactions. The IDLs modulate the interactions by varying their lengths, conformations and local structure, thereby affecting the overall IDG. Results of our analysis provide guidelines in modelling of multi-domain proteins from the tertiary structures of constituent domain components.

Keywords: multi-domain proteins; inter-domain linkers; inter-domain orientation; linker flexibility; interface constraints

Introduction

Two domains contiguous in multi-domain proteins are tethered by a stretch of polypeptide segment which is referred as the inter-domain linker (IDL). The IDLs vary in their size, composition and structure (George & Heringa, 2002). Analysis of IDLs of multi-domain proteins traditionally has been useful in expanding our understanding of nature of these fragments. Studying compositional and conformational properties of IDLs has provided us with the knowledge of designing flexible and strong tethers with unique desired properties for chimeric proteins (Arai, Ueda, Kitayama, Kamiya, & Nagamune, 2001; McClendon et al., 2008; Nomura et al., 2012; Wriggers, Chakravarty, & Jennings, 2005). This has also aided in the advancement and design of novel experimental methods in molecular biology (Tang, Jiang, Parakh, & Hilvert, 1996). These design principles involve generation of fusion proteins for antibody binding, adapter domains, fluorescent protein tagging and immunoassaying (Arai et al., 2001; Bird et al., 1988; Maeda et al., 1996).

Apart from serving as a covalent link between the domains of a multi-domain protein, linker regions could affect the function of the protein in various ways (Wriggers et al., 2005). The IDL regions equip proteins with unique ways of coupling biological functions of the tethered domains. The extent of functional coupling is determined by the structural communications between the domains (Bashton & Chothia, 2002). These features aid in modulating protein function precisely (Altschuh, Tessier, & Vernet, 1994). This is one of the predominant features of multi-domain proteins which have functional sites within the tethered domains (Wei, Ye, & Dunaway-Mariano, 2001). Most of the two-domain enzymatic proteins have a catalytic and a regulatory domain (Traut, 1988). The functional regulatory signals from the regulatory domain are often transmitted through the IDL segments (Abbott et al., 2000; Morra, Potestio, Micheletti, & Colombo, 2012). In experiments which introduce perturbations in these linkers the regulation and overall functional effects vary widely (Lu, Chai, & Fu, 2009; Strang, Wales, Brown, & Wild, 1993; Takizawa et al., 2011; Valentini et al., 2000). On the contrary, certain multi-domain proteins have functional sites which are dynamic. These functional sites are formed at the inter-domain interface

*Corresponding author. Email: ns@mbu.iisc.ernet.in

ISSN 0739-1102 print/ISSN 1538-0254 online

http://dx.doi.org/10.1080/07391102.2012.743438
http://www.tandfonline.com

and are dependent on the extent of interaction between the tethered domains. This serves as an elegant strategy to regulate protein function. Such linker regions often undergo major conformational transitions to juxtapose the domains in active conformation. These and other features of IDLs form a central part of allosteric regulation of various proteins (Altschuh et al., 1994; Valentini et al., 2000). These have been extensively studied by mutational analysis (Fiorani et al., 2003). Few proteins have active sites and catalytic residues within the IDL regions. These have been termed as soft linkers, which can undergo easy structural transitions to aid catalysis. In few cases structural changes are triggered by post-translational modifications of the linker segments (Bonet-Costa et al., 2012). Linkers can also act as structural scaffolds and aid in proteins looping around other macromolecules (Nomura et al., 2012; van Leeuwen, Strating, Rensen, de Laat, & van der Vliet, 1997). They have also been instrumental in molecular signalling events in prokaryotes by adopting specific helical conformation (Aravind & Ponting, 1999). Flexibility and hydrophilicity in the linkers are important in preventing disturbance of domain functions (Wriggers et al., 2005). More rigid linkers (helical and proline rich) (Adzhubei & Sternberg, 1994) also may keep domains apart and act as spacers and prevent unfavourable interaction during folding (Briggs & Smithgall, 1999; George & Heringa, 2002; Gokhale, Tsuji, Cane, & Khosla, 1999; Ikebe et al., 1998). Softness in linker regions control interdomain orientations. The flexible motion of domains with respect to one another is controlled by hinge regions localized in linkers (Ikebe et al., 1998). These bending and shearing motions vary from protein to protein and are involved in function (Bennett, Choe, & Eisenberg, 1994; Bennett & Eisenberg, 1994; Winkler, Schutt, Harrison & Bricogne, 1977).

All the above discussed functional diversity of multidomain proteins stems from the amino acid composition, the conformational preferences and the flexibility/rigidity of IDLs. The combination of sequence, structure and dynamic features enables selective advantage of linker regions to aid in the global functions of a protein favourably. Previous studies on IDLs show varying geometric properties and biophysical features which act as important signatures of the linkers (Wriggers et al., 2005). With the increasing sequence and structural databases and novel experimental methods to study linkers, the coherence among studied properties is diminishing and this has made generalizations unattainable.

The enormous diversity in linker properties has made prediction of the linker region from solely the sequence information a daunting task. Although there exist many machine learning (Adzhubei & Sternberg, 1994; Benros, de Brevern, & Hazout, 2009; Ebina, Toh, & Kuroda, 2009; Miyazaki, Kuroda, & Yokoyama, 2006) approaches to predict IDL regions, the problem is tightly coupled to the domain boundary (Chen et al., 2010; Eickholt, Deng, & Cheng, 2011; Ezkurdia, Grana, Izarzugaza, & Tress, 2009; Yoo et al., 2008; Yoo, Sikder, Zhou, & Zomaya, 2008) and domain prediction methods (Dong, Wang, Lin, & Xu, 2006; Dumontier, Yao, Feldman, & Hogue, 2005; Eickholt et al., 2011; Kong & Ranganathan, 2004; Zhang, Liu, Dong, & Jin, 2011). Most linker prediction methods perform poorly if there are errors associated with domain assignment itself (Ezkurdia et al., 2009).

In this current research, we attempt to understand the roles of IDLs in multi-domain proteins with respect to domain orientation in particular. We discuss how flexibility, sequence and conformational preferences of IDL regions affect the inter-domain interactions and interdomain orientations and hence affect the functional coupling of the domains. Inter-domain interactions are important in maintaining stability in multi-domain proteins (Bhaskara & Srinivasan, 2011). We use the concept of protein blocks (PB) (de Brevern, 2005; de Brevern, Etchebest, & Hazout, 2000; Fourrier, Benros, & de Brevern, 2004; Joseph et al., 2010) to understand the local conformational transitions to demarcate linkers form other loop regions. We also address the length dependence of linker conformations in relation to preservation of inter-domain geometry (IDG).

Methods

Data-set of 3D structures of multi-domain proteins

The data-set of multi-domain proteins was obtained by mining the PDB (http://www.pdb.org) (Berman, Kleywegt, Nakamura, & Markley, 2012) using the following criteria: the presence of a single polypeptide chain in the asymmetric unit and the biological unit; the crystallographic resolution ≤ 3.0 Å and the presence of only two continuous Structural Classification of Proteins (SCOP) (Murzin, Brenner, Hubbard, & Chothia, 1995) domains within each polypeptide chain. We did not consider protein structures without SCOP domain annotations in this data-set. The PDB accession codes are provided in Table S4. The table also gives the details and criteria used to curate the PDB to obtain ($n = 290$) the final data-set of two-domain proteins. This set was non-redundant at 30%.

Data-sets of homologous structures and sequences

Each of the 290 proteins sequences were queried against the entire PDB, using Position specific iterative blast (PSI-BLAST) (Altschul et al., 1997) at an *E*-value cut-off of 10^{-5} with low-complexity regions masked for five iterations. We ensured that the sets of homologous protein sequences picked by BLAST for every query was reliable by filtering the hits using the following criteria: sequence identity $\geq 30\%$ and query and hit coverage $\geq 80\%$. A total of 691 homologous protein sequences with known 3D structure for 255 sequences were picked. Thirty-five sequences were unique in the initial data-set and had no homologous proteins matching our criteria. This summed up to a total of 928 unique multi-domain protein–homologue pairs. This data-set was used to compare the features of multi-domain proteins with their homologous proteins. Domain definitions for the homologous proteins were taken from SCOP. In the absence of SCOP domain definitions, domain boundaries were marked from the alignments of these homologues with their corresponding SCOP annotated multi-domain protein. Apart from having homologues of known structure from the PDB, we also queried each of the 290 sequences to obtain homologues from UniProt database (Consortium, 2012) using PSI-BLAST (Altschul et al., 1997). We used the same criteria for selection and pruning of hits to obtain the final set of homologues sequences as described above. We were not able to obtain homologues for three sequences using the above-mentioned criteria.

Identification of IDLs in multi-domain proteins

The IDLs definition was guided by the SCOP domain definitions for the multi-domain protein data-set. The rationale is that IDLs have very little or no interactions with either of the domains which they tether. Linker fragments connecting the two domains for each of the multi-domain contain the *i*th (i.e. C-ter residue of the 1st SCOP domain) and $i+1$ residue (i.e. N-ter residue of the 2nd SCOP domain); they can have a maximum length of 40 residues. We scanned 20 residues towards the N- and C-terminus of the *i*th residue to generate all possible fragments. We then computed average number of heavy atom contacts for each residue within a sphere of 4.5 Å for all the fragments. The contacts within $i+3$ and $i-3$ residues, while computing averages for the *i*th residue, were neglected. The fragments generated using SCOP boundaries showed fewer contacts per residue than when a random boundary position was chosen. The fragment with the lowest average contacts was chosen as the IDL.

IDL and ISS sequence and structural properties

Amino acid distributions for the IDLs and inter-secondary structural linker (ISS) regions of multi-domain proteins were computed and compared with background distributions. The inter-secondary structure segments/loops were identified after secondary structure assignments by Stride (Frishman & Argos, 1995; Heinig & Frishman, 2004). We used protein structural alphabets to condense the 3D structural information into a 1D sequence. The structural alphabets are computed by breaking the structure into a series of five residue overlapping fragments called PB. The back bone features of the polypeptide fragments are then used to assign the structural alphabet to each block. A catalogue of this method has been published (Joseph et al., 2010) and used for a variety of studies (de Brevern et al., 2000; de Brevern, Valadie, Hazout, & Etchebest, 2002; Fourrier et al., 2004). We analysed the distributions of 16 PBs and their transitions (in terms of 256 consecutive di-PBs) in both the IDLs and the ISS regions.

Quantifying IDG

A dihedral angle (χ; -180 to +180) to quantify the inter-domain orientation/geometry (IDG) of multi-domain proteins was defined. We used the two centre of masses coordinates corresponding to interacting domains and the two C-α atom coordinates (residues i and $i+1$) of SCOP domain boundaries to define the dihedral angle. The distribution of IDGs was then analysed by fitting to von Mises distribution (Dowe et al., 1996). The variation in the IDG for a given multi-domain protein–homologue pair was computed as the absolute difference ($\Delta\chi$) of the smallest angle between the two dihedrals. We also analysed the frequency distributions of this difference in IDGs. Homologous protein pairs with $\Delta\chi \leq 30°$ were grouped in IDG-C set and those with $\Delta\chi > 30°$ in IDG-NC, respectively.

Global comparison of multi-domain protein structures

Dali structural alignment programme (Hasegawa & Holm, 2009; Holm & Sander, 1998) was used to compute the structure-based sequence alignment and the global RMSD (Å) for all the 928 pairs. Global distance test-total score (GDT_TS) (Zemla, 2003) computed at 1, 2, 4 and 8 Å was also used for a more robust comparison of homologous protein structures. The structure-based sequence alignments were done with Dali to map the PB sequences and derive a PB–PB alignment for each pair of proteins. The pair-wise structure-based amino acid sequence alignments and PB–PB alignments permit to compute the AA- and PB-scores. The AA-scores and PB-scores were done for both the interfaces and the IDLs separately.

Comparison of interfaces and IDLs

Structural comparisons of domain–domain interfaces among homologous protein pairs were performed using the iAlign algorithm (Gao & Skolnick, 2010). The *Z*-score cut-off of 6.0 was applied to filter the interface alignments. Interface alignments were analysed only if there were at least four residues in both the interfaces. We obtained IS-score, int-RMSD (Å) and No. of aligned contacts after the alignments. The IDL fragments of homologous pairs were compared by locally aligning them as a block to obtain an IDL alignment score (lin-Ali score). This alignment was then used to obtain the topological equivalences for superposition and subsequent lin-RMSD (Å) calculations. The fitting was performed using the McLachlan algorithm (1982) as implemented in the programme ProFit (Martin, A. C. R. and Porter, C. T., http://www.bioinf.org.uk/software/profit/).

Amino acid and PB substitution scores

We employed AA-score and PB-scores to quantify the amino acid and PB similarities, respectively, for interfaces and IDL segments. These scores quantify the extent of amino acid change and the PB change for a given sets of positions. AA-score depicts the conservation of amino acid nature for a set of positions and PB-score depicts the conservation of local conformation at a given set of positions. We used the structure-based amino acid sequence alignments and mapped PB–PB alignments of interface and IDL fragments to compute the substitution scores (Equations (1) and (2)), respectively.

$$\text{AAscore} = \frac{2 \times \sum_n M_{ij}}{\sum_n M_{ii} + \sum_n M_{jj}} \tag{1}$$

$$\text{PBscore} = \frac{2 \times \sum_n M_{pq}}{\sum_n M_{pp} + \sum_n M_{qq}} \tag{2}$$

where i and j represent aligned amino acids and p and q represent the aligned PB at a given alignment position n. M_{ij}, M_{ii} and M_{jj} are the BLOSUM62 substitution values and M_{pq}, M_{pp} and M_{qq} are the PB-substitution values. Gaps in both the amino acid and the PB alignments are penalized by using a value of −4.00. The PB substitution matrix used to score the conservation of local conformation is provided in supplementary Table S5.

Internal protein domain flexibility

An ensemble of 100 low-energy conformers for each multi-domain protein was obtained by using CONCOORD algorithm (de Groot et al., 1997) using a set of distance constraints. Yamber2 and Engh–Huber parameters were used for setting van der Waals and bonded constraints. Care was taken so that the conformers generated did not have any short contacts and did not violate the predefined bounds by more than 1 nm in total. The IDG (χ) was then computed for each of the conformers and a circular mean and circular variance was computed. The circular variance (Allen & Johnson, 1991) of the IDG (Equation (3)) was served as a measure of flexibility of inter-domain orientation.

$$Var(\chi) = 1 - \sqrt{(\sum_{i=1}^{n} \cos \chi_i)^2 + (\sum_{i=1}^{n} \sin \chi_i)^2 / n} \tag{3}$$

where χ_i is the IDG of the ith conformer and n is the total number of conformers. The ensemble of conformers was then converted into PB sequences and a multiple alignment of 100 PB sequences was generated for each multi-domain protein. The PB entropy (Equation (4)) (de Brevern et al., 2000) was then computed at each position along the sequence and averaged over the length of IDL segments.

$$S_{\text{PB}} = -\sum_{i=1}^{16} P_i \ln P_i \tag{4}$$

where P_i is the frequency of PB i in each of the aligned column. This value ranged from 0 to a maximum of 4.0. This measure provided with the local flexibility of the main-chain.

Free energy computations

The free energy contributions towards the folding for the individual domains and the full length multi-domain proteins were computed using the FoldX (Schymkowitz et al., 2005; Van Durme et al., 2011) empirical effective energy function. This interaction energy was applied as a measure of extent of interactions in addition to the number of contacts.

Statistics and significance tests

All variables were tested for normality using Kolmogorov–Smirnov tests (linear data) and Watson's test (angles). All normally distributed values were compared using Student's *t*-test and its circular variable equivalent with Watson's two-sample test for homogeneity (Mardia, 2000). All proportion and frequency data were arcsine transformed before checking for normality. All propensities were computed as a ratio of observed frequencies and background frequencies (see Supplementary Methods). The expected probabilities for amino acids, PBs and diPBs were computed from the observed frequencies in the whole data-set. Pearson's correlation coefficients were computed in establishing the relationships among various parameters. We used Fisher's *Z*-test to assess the significance in comparing the correlation coefficients. All statistics were performed using the statistics module from the R package (R-devel, 2011).

Results

IDLs are distinct in sequence and structural properties from inter-secondary structures

We studied the properties of both IDL segments and ISSs. These comparisons would aid in the development of methods for identification and demarcation of IDLs from sequence databases. The IDLs in our data-set were significantly longer (12.53 ± 8.2) than ISS linkers (4.34 ± 1.9; Unpaired Student's *t*-test; $t = 41.74$; $df = 2981$; $p < .01$). The distribution of IDL lengths in our data-set

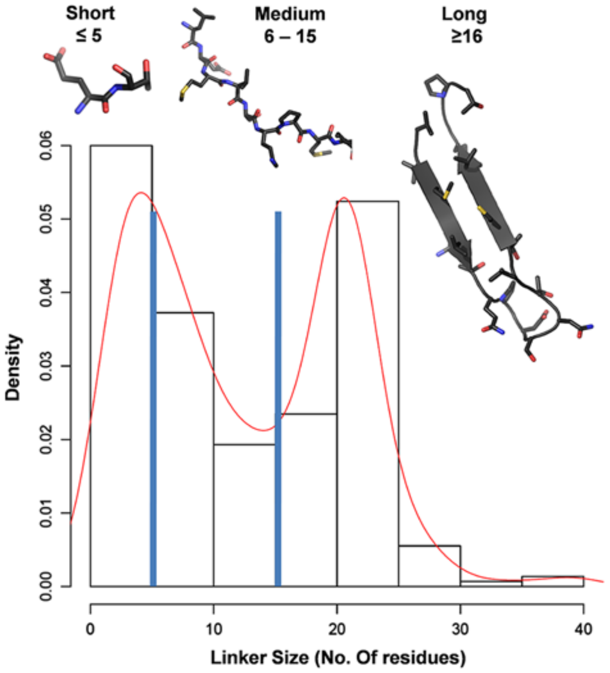

Figure 1. Length distribution of IDL in multi-domain proteins. Bimodal distribution of probability density of IDL lengths ($n = 290$). The blue lines (at 5 and 15) are the boundaries used to define short, medium and long size IDLs. Representative structures of these three classes of IDLs are also shown.

of globular two-domain proteins is bimodal with peaks at 15 and 22 residues, respectively (Figure 1). In order to understand the length dependence of IDL properties better, we classified them into three classes i.e. short linkers (2–5 residues), medium-sized linkers (6–15 residues) and long linkers (>16 residues) (Figure 1). We found almost equal proportion of short (30.0%, $n = 87$) and medium-sized IDLs (28.2%, $n = 82$) and slightly more proportion of longer IDLs (41.7%, $n = 141$) in the data-set.

Six amino acids (Glu, Gly, Ile, Pro, Lys and Trp) have high preferences in the IDLs (see Figure S1). Of these, only Trp and Lys residues have high propensities in comparison to ISSs (see Figure S1) and hence their presence can act as a demarcating feature when using composition information to identify IDLs (See Supplementary Results).

We investigated if IDLs comprised of standard secondary structures. Substantial number of linker residues are present in the standard secondary structures such as β-turns (23.1%; $n = 839$), β-sheets (28.5%; $n = 1035$) and α-helices (22.5%; $n = 817$). Residues with non-regular conformations, such as coils, were also present in same proportions (22.1%; $n = 803$). Although the occurrences of β-sheets, β-turns and coils were not very different from helices, they have high propensities (i.e. propensity > 1.00) to occur in the IDLs (Figure 2). The β-sheets/strands are found in IDLs with highest propensity (Figure 2). To determine if the IDLs formed stable structures within themselves (especially β-sheets and α-helices); we computed the proportion of hydrogen bonds (H-bonds) per residue within the IDLs. We also tested if there is length dependence for adopting these secondary structures. Both the intra IDL hydrogen bonds (LL) and the hydrogen bonds formed between IDLs and domain segments (LD) have been measured (See Supplementary Results). The medium and long linkers are able to interact within the linker segments favourably and form more compact and regular conformations. We investigated if the conformations adopted by residues in IDLs are distinct from the conformations of residues in the linker of contiguous regular secondary structural regions (ISSs). The conformational preferences of residues were quantified using 16 structural alphabets or PBs. The preference of the main-chain to adopt a particular conformation was quantified by measuring the propensities of structural alphabets.

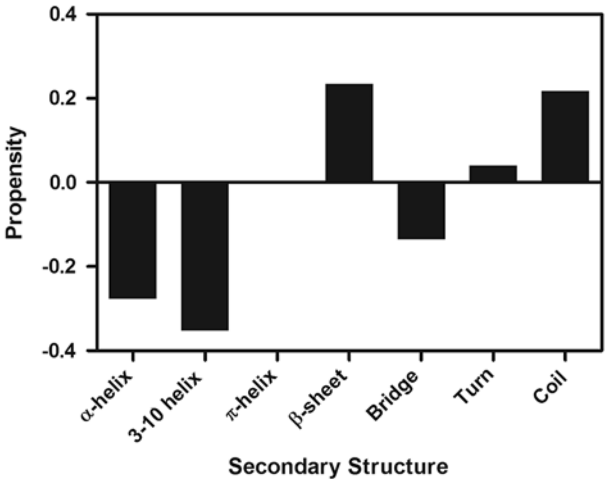

Figure 2. Propensities of gross secondary structures adopted by IDL segments. Propensities of IDL residues to adopt seven different secondary structures computed as the ratio of observed and background frequencies for each class. The propensities are rescaled to show preferred secondary structures above the value 0. β-sheets, turns and coil segments have high propensities (see also Figures S2 and S3).

Further, we quantified di-PB (consecutive structural alphabets) propensities in order to see how the conformation at a given position in the linker region propagates (Figure 3(A) and (B)). Specific structures, i.e. set of di-PBs, was distinct from the PB occurrences in the complete data-set of proteins (see Supplementary Methods). As most of these diPBs were also observed to have lower frequencies in the overall protein structures, we also quantified the most frequently observed di-PBs in the short, medium and long size IDLs. The frequency computations showed that short IDLs comprised mostly

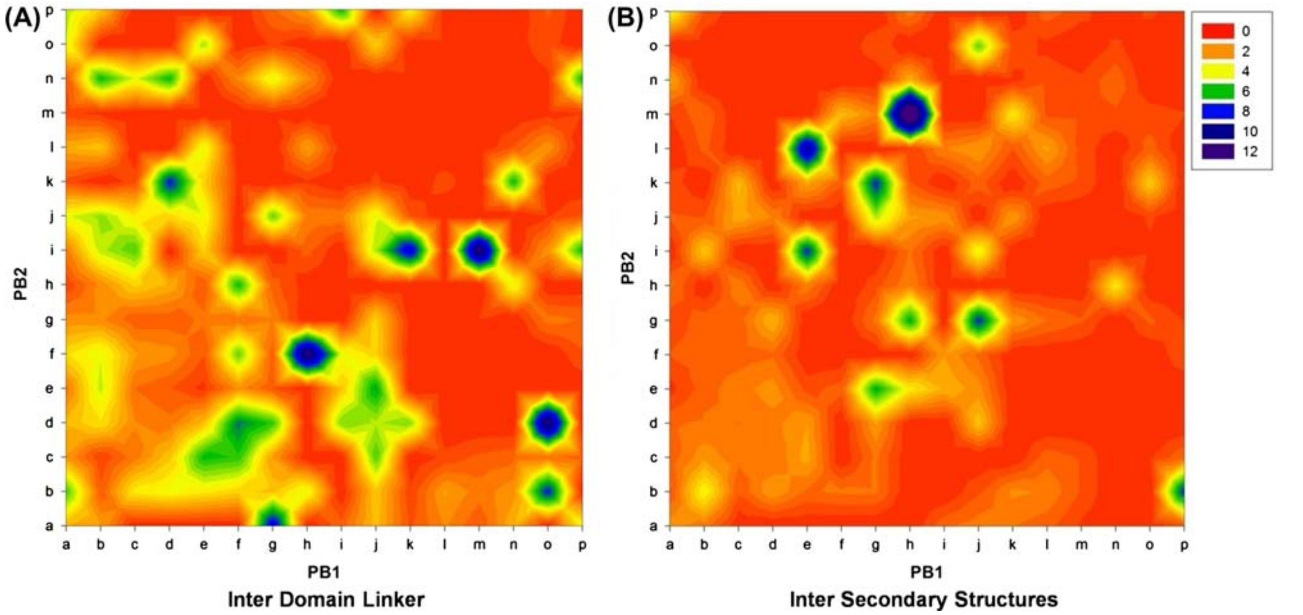

Figure 3. Conformational propagation and diPB propensity. The distribution of all 256 consecutive di-PB's propensities (1 being random) of (A) IDL and (B) ISS segments is shown. The most frequent and preferred conformations adopted by IDLs (green and blue coloured regions) are more regular (corresponding to β-sheets and α-helices) and distinct to the conformations of ISS loops (see also Figure S4).

of di-PBs corresponding to β-sheets and β-strands structures at the termini of β-strands (di-PBs: *cd, dd, de* and *df*). The next highest representation was observed in coil segments (di-PBs: *eh, fk, fb, gg, gh, hi, ia* and *kb*). The same pattern was also seen in medium and long size IDLs. In medium and long size IDLs, di-PBs corresponding to helical segments were also observed in relatively higher frequencies (di-PBs: *kl, lm, mm* and *mn*).

IDG and domain orientations in multi-domain proteins

We measured IDG for all the multi-domain proteins in terms of the torsional angle (χ) which varied from −180° to +180° (see Methods, Figure 4(A)). This angle represents the relative orientation of the two domains with respect to each other. The farther the angles from 0°, the more non-collinear are the centre of masses of the interacting domains with respect to the IDLs. The inter-domain orientations in the two-domain proteins with respect to the IDL were small and close to zero ($< \pm 60°$) in most (41.3%; $n = 120$) of the cases. Most of the proteins, however, showed a slight negative gauche tendency (close to −60°) in their orientations (Figure 4). We found that the distribution of the inter-domain angle (χ) followed a von Mises distribution (Circular analogue of the normal distribution on a line; Watson's goodness of fit test; test-statistic = .7872; $p = .187$; $\mu = -49.97°$; $\rho_{(\text{circular})} = .226$; var($\chi$) = .773, Figure 4(B)). To explore if there is any preference of inter-domain orientation towards a certain angle in proteins within the three groups of IDLs, we computed the goodness of fit to von Mises distribution of angles for all of them. Although there is no statistically significant difference in the distribution of angles in short, medium and long size IDL containing proteins, as the length of the IDLs increases the mean inter-domain orientations are farther away from 0°. The inter-domain orientation is very well conserved in the homologous multi-domain proteins (Figure 4(C), see Methods). We assessed the similarities in the χ angles by measuring the difference of the angles for every multi-domain–homologue pair. This value, however, showed a strong peak close to 0°. The major half of the cases (61.09%; $n = 567$) showed a value less than or equal to 30° indicating that there were little or no variations in the inter-domain orientations of homologous protein structures (Figure 4(C)). Hence domain orientation is conserved in homologous multi-domain proteins. To understand if it is only a direct consequence of the conservation of sequence and structural divergence among the homologous proteins, correlations between difference in χ values and variations in sequence (sequence identity) and structural features (overall *C–α* RMSD and GDT_TS) for each multi-domain–homologue pair were measured (Figure 5). It suggests a slight negative correlation of sequence identity (Pearson $r = -.290$; $n = 928$) (Figure 5(A)) and structural variation (for GDT_TS: Pearson $r = -.294$; $n = 928$; for RMSD: Pearson $r = .245$; $n = 928$) with the variance in inter-domain orientation (Figure 5(B) and (C)). It highlights that

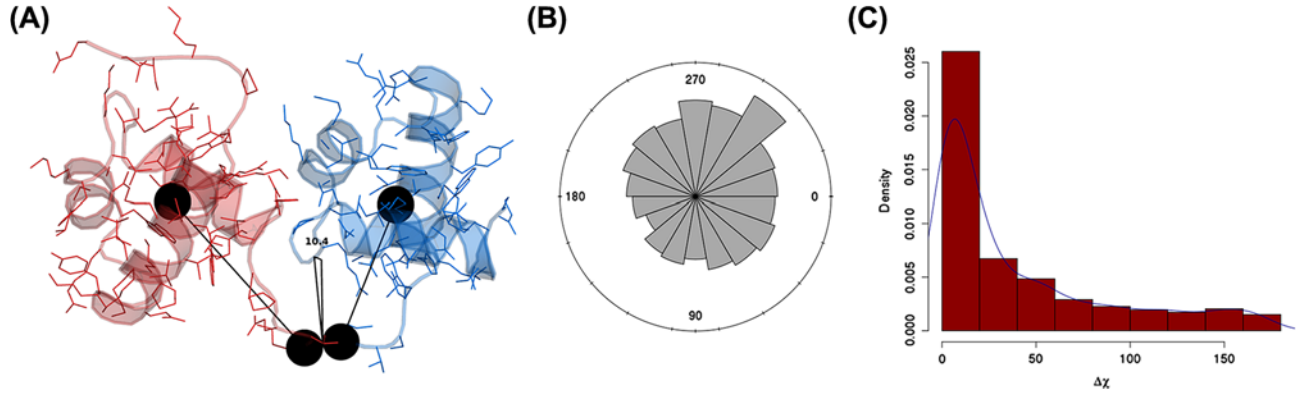

Figure 4. IDG distributions and evolutionary conservation. (A) The structure of representative multi-domain protein C-Myb (PDB code: 1GV2), DNA-binding protein showing two all alpha domains having a χ-angle of 10.4°. The centre of mass coordinates of the two domains (red and blue) and the two C-α atom coordinates of SCOP domain boundary residues are used to define the IDG torsion angle (χ). (B) Circular histogram showing the von Mises distribution of χ angles (Watson's goodness of fit test; test-statistic = .7872; $p = .187$; $\mu = -49.97°$; $\rho_{(\mathrm{circular})} = .226$; var($\chi$) = .773) in all the multi-domain proteins ($n = 290$). The area of each sector is proportional to the abundance of the IDG within that angular bin (see also Figure S5 and Table S3). (C) Probability density of variation of IDG in homologous multi-domain proteins. For a large fraction (61.09%; $n = 567$) of the homologous protein pairs IDG is conserved ($\Delta\chi \leq 30°$).

global changes in sequence and structure affect inter-domain orientations.

At a finer level, in order to establish quantitative relationships between inter-domain orientation changes and sequence and structural properties, we studied the changes in properties of IDLs and domain–domain interfaces in detail. We hypothesize that the extent to deviations observed for inter-domain orientations of the two interacting domains would depend on both the interaction patterns at the interface and the IDL flexibility/conformational freedom. The precise orientation adopted is the result of combination of IDL and interface properties dictated by the overall topology/structure. The relative contributions of these two features in preserving domain orientations would be reflected in the strong evolutionary selection pressure for conservation. In order to understand the above-mentioned relative contributions of interface and IDL properties, the multi-domain–homologue pairs were classified into IDG conserved (IDG-C; $\Delta\chi \leq 30°$) and non-conserved categories (IDG-NC; $\Delta\chi > 30°$).

Effect of IDLs and domain–domain interactions on evolutionary conservation of IDG

We quantified the effects of domain interactions and conformational preferences of IDLs on the inter-domain orientation. In understanding the extent of involvement of interface in the IDG, we computed interface similarities for pairs of homologous multi-domain proteins in terms of sequence (namely int-AA score, see Methods Section), local main-chain structure (int-PB score) and overall interactions patterns (IS-score and int-RMSD). For the IDLs, we determined the fragment sequences properties (lin-AA score, lin-Alignment score) and structural and conformational (lin-PB score, lin-RMSD) properties. We reasoned that features which affect inter-domain orientations would have differential distributions when we compare cases with conserved (IDG-C) and non-conserved inter-domain geometries (IDG-NC).

Among homologous multi-domain proteins showing conserved IDG-C, interfaces were well aligned. The IDG-C set had significantly higher IS-scores (.69 ± .19) than the IDG-NC set (.48 ± .21; Student's *t*-test: $t = 14.52$; $df = 878$; $p < .0001$; Figure 6(A)). This score reflects the underlying pattern in the structural (int-RMSD: Student's *t*-test: $t = 10.89$; $df = 878$; $p < .0001$; Figure 6(B)) and interaction similarities (No. of aligned contacts: Student's *t*-test: $t = 6.54$; $df = 878$; $p < .0001$; Figure S6) observed in the given homologous protein pairs in spite of similar sized interfaces (2812 ± 1697 Å^2 for IDG-C and 2562 ± 1646 Å^2 for IDG-NC; Unpaired Student's *t*-test: $t = -1.31$; $df = 259$; $p = .188$) of proteins in the two sets compared. In addition to these, we also found that the amino acid sequence and local structural conservation is higher in the IDG-C set.

A similar trend was observed for distributions of IDL properties for the two sets of homologous protein pairs. The IDL segments in both the groups showed similar length distributions (12.8 ± 8.3 for IDG-C; 12.2 ± 8.3 for IDG-NC; Unpaired Student's *t*-test: $t = .57$; $df = 254$; $p = .566$). We observed high sequence and structure conservation of the IDL segments in homologous protein pairs where IDG is conserved. The IDG-C set has higher IDL alignment score (2.93 ± 1.69) than the IDG-NC (2.06 ± 1.46) set (Unpaired Student's *t*-test: $t = 8.13$; $df = 926$; $p < .0001$; Figure 6(C)). The structural conservation

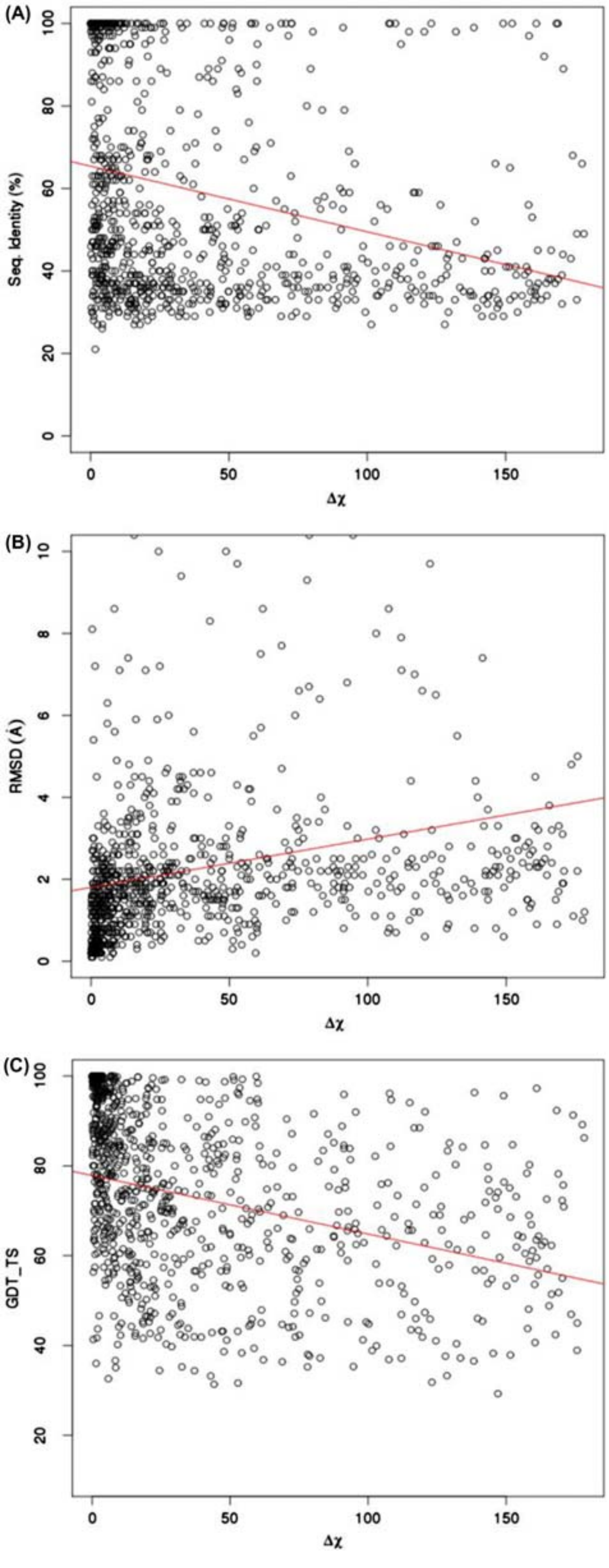

Figure 5. Relationship between the variation in global sequence and structural divergence and the variation in IDG. The variation in IDG as a function of (A) global sequence identity, (B) overall RMSD and (C) GDT_TS among homologous protein pairs ($n = 928$). A small correlation is observed for sequence identity and GDT_TS indicating that the relationship is not quantitative.

measured as local RMSD of the IDL segments also showed lower values for the 1DG-C set than the IDG-NC set (Unpaired Student's *t*-test: $t = 2.52$; $df = 926$; $p = .0118$; Figure 6(D)).

We quantified the sequence variation and structural variation for homologous interface and IDL pairs in AA- and PB-scores, respectively (See Methods). The int-AA score for IDG-C is significantly higher than that of IDG-NC (Student's *t*-test: $t = 8.42$; $df = 878$; $p < .0001$; Figure 6 (E)). The distributions of the int-PB scores are also high for interfaces pairs in homologous protein pairs (Student's *t*-test: $t = 7.51$; $df = 878$; $p < .001$) with domain orientation conserved (Figure 6(F)). In comparing the homologous IDL segments, the AA-score (Unpaired Student's *t*-test: $t = 10.41$; $df = 926$; $p < .0001$; Figure 6(E)) and the PB-scores (Unpaired Student's *t*-test: $t = 10.30$; $df = 926$; $p < .0001$; Figure 6(F)) showed significant differences in the distributions of IDG-C and IDG-NC sets. These comparisons of interface and IDL sequence and structure properties between IDG-C and IDG-NC sets showed that the variations in the interface and IDL affect IDG. In order to understand which properties contribute most to the variation of IDG among homologues, the correlation coefficients for the change in IDG ($\Delta\chi$) and each of the interface and IDL sequence and structural conservation parameters were computed (Table 1). Interface alignment, expressed as IS-score, is most well correlated (Pearson $r = -.46$) with the change in IDG. This value was significantly higher (Table 1) than that of IDL fragment alignment score (Pearson $r = -.26$) indicating that the interface variations determined the IDG variations better. This was also reflected in the local structural superposition of the interfaces as measured by local RMSD. Correlations coefficients for AA- and PB-scores are almost similar for interfaces and IDLs (Table 1). This conclusively indicated that the interface properties have more constraints in determining the IDG among homologous protein pairs.

Flexibility of IDLs and domain motions

The flexibility of IDLs induces domain motions and affects the IDG. To obtain a measure of extent of internal flexibility of IDG of multi-domain proteins, we generated and sampled an ensemble of 100 low-energy conformers using CONCOORD (see Methods). We measured 'circular variance' of the χ-angle for the conformers of each multi-domain protein. This value ranging from 0 to 1 depicts how tightly or loosely the inter-domain geometries cluster together across the entire ensemble of conformers. We then determined the relationship between circular variance of IDG-Var and the interaction energy between the two domains. We also investigated if there is a length dependence of this relationship for the three different IDLs in our data-set. In general, the spread of the IDG (circular variance of χ)

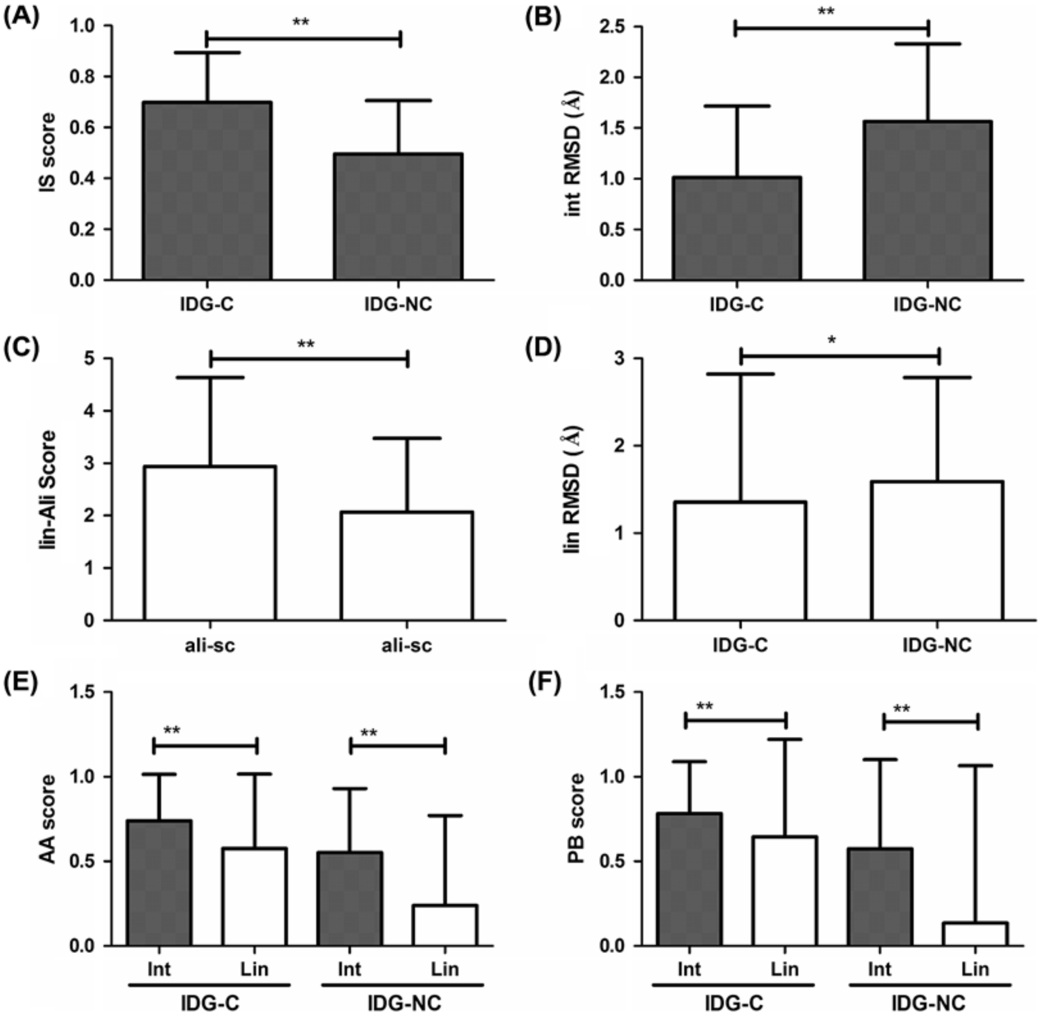

Figure 6. Relative contributions of interface and IDL properties in conservation and divergence of IDG. (A) IS-scores for IDG-C and IDG-NC sets showing that interfaces are conserved better in IDG-C set (see also Figure S6). (B) int-RMSD values are also lower in IDG-C than IDG-NC. Interface interaction patterns and geometries are more preserved and are indicative of conservation of IDG among homologous proteins. (C) Local IDL segment alignment scores are also high for IDG-C than the IDG-NC data-set. (D) The IDL local RMSD although significantly different in the two sets of homologous proteins have high variances and are less correlated to the overall change in IDG. (E) Amino acid and (F) PB substitution scores for interfaces and IDLs in the IDG-C and IDG-NC data-sets.

Table 1. Relative contribution of interface and IDL properties in determining IDG: Pearson correlation coefficients representing the relationship between IDG (χ) and various interface and IDL parameters. The first two interface properties are significantly more correlated with IDG than the corresponding IDL properties indicating that interface variations and conservation better dictates IDG variations during evolutionary processes. Comparing the correlation coefficients using Fishers *Z*-test established statistical significance. The diff-*Z* denotes the difference in *Z*-scores of interface and IDL properties after Fishers transformation. The corresponding one-tailed *p*-value is also provided.

Parameter	Interface	IDL	Diff *Z*	*p*-value
Alignment score	−.46	−.26	−4.86	**8.68E-07**
local RMSD	.33	.07	5.75	**6.53E-09**
AA-score	−.27	−.30	.69	4.87E-01
PB-score	−.26	−.29	.59	4.90E-01

$p < .001$ is shown in bold.

decreases as the interaction free energy between the domains increases favourably. This was more stark only for short and long size IDLs (Figure 7(A)) as they are more positively correlated with the interaction free energy (Pearson *r*: short = .34; medium = .09; and long = .25) (Figure 7(A)). This might also be reflective of the observed bias in the lengths of IDL in multi-domain proteins (Figure 1). The apparent lack of correlation between the flexibility of IDG and the interaction energy in medium-sized linkers can make them less effective and costly for incorporation in multi-domain proteins.

These results showed mechanistically that the interface has more constraints on the IDG and the interacting domains making more contacts across the interface can stabilize the domain orientations better. The internal flexibility (IDG-Var) stems from overall effect of various dihedral angle shifts across the main-chain of the protein. We captured this shift in terms of PB-entropy for each site over the entire ensemble (See Methods). A slight positive correlation (Pearson *r* = .25) was detected between average IDL PB entropy. This showed that high PB entropy concentration in IDL segments can manifest as internal flexibility and aid in domain motions.

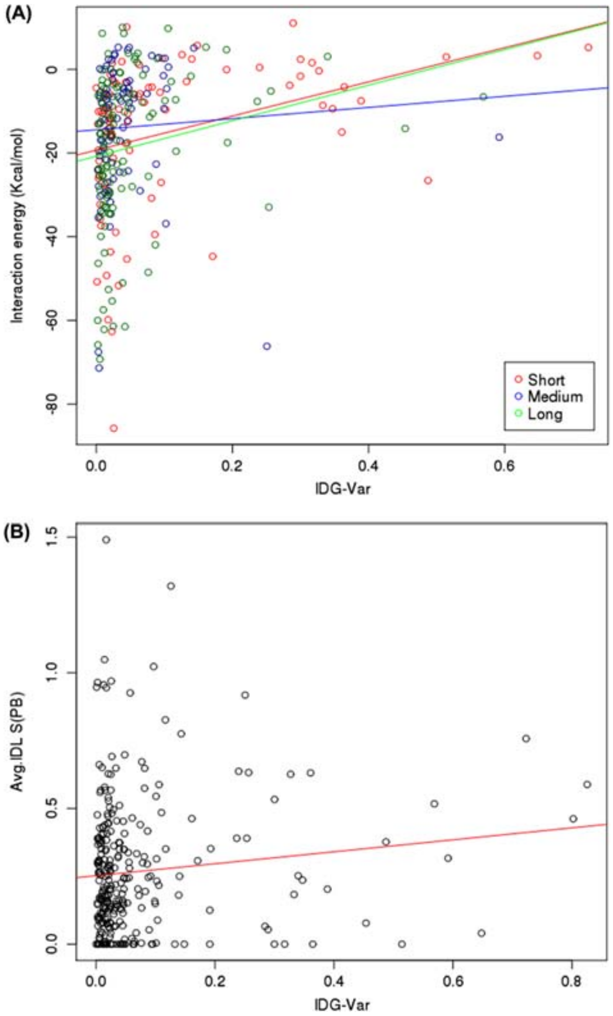

Figure 7. Effect of domain–domain interactions and PB transitions of IDLs on the internal flexibility IDG. (A) Circular variance of χ denotes the internal flexibility of IDG and is dependent on the extent of inter-domain interactions. It shows a small positive correlation with the interaction free energy (kcal/mol) in proteins with short and long size IDLs. (B) The internal flexibility is also affected by the local main-chain conformational changes measured as average PB entropy for the IDL segments. We observe a slight correlation here also.

Discussion

IDLs bridge the domains of the multi-domain proteins. Various roles have been attributed to IDLs from functional coupling of domains, catalytic (Altschuh et al., 1994) act as spacers (Adzhubei & Sternberg, 1994; George & Heringa, 2002) and interaction stabilizers. Previous attempts to understand IDL properties have aided in setting up some rules for the design of inactive and inert IDLs used frequently for the generation of fused and tagged proteins (Chung, Parker, Bianchet, Amzel, & Stivers, 2009).

The IDLs and ISS segments showed demarcating features in amino acid compositions from the rest of the proteins, but did not display features specific to IDLs alone. The IDLs are mainly preferred to be in β-strands/sheets, turns and coiled conformations. Converting protein 3D structures into one-dimensional PB has aided in understanding complex information such as similarities and differences in conformational properties in terms of strings and also opened up various string comparisons and matching tools (de Brevern et al., 2000; de Brevern, 2005; Gelly, Joseph, Srinivasan, & de Brevern, 2011). Di-PBs in IDLs and ISS regions showed distinct distributions indicating that IDLs can be identified if PB can be predicted. Further, these results also substantiated the observed secondary structural and hydrogen-bonding patterns observed for the three IDL groups. Analysis of the sequence and conformational properties of IDLs illustrated the diversity in their interaction patterns and length dependent features clearly. Short linkers (<5 residues) are mostly part of loop segments or at the termini of β-sheets/strands. Medium-sized (6–15 residues) linkers showed some short helices, but were able to form individual β-strands. Long (>16 residues) IDLs showed large loops, fully H-bonded anti-parallel β-sheets and formed more stable structures. Although there is a length dependence of the end-to-end distance with the linker size, few of the medium-sized linkers and, to some extent, the long linkers are able to form closed loop structures. These structures are conformationally diverse and form a basic element of protein structure (Berezovsky, Grosberg, & Trifonov, 2000; Berezovsky, Kirzhner, Kirzhner, Rosenfeld, & Trifonov, 2002). Apart from forming closed loop structures, these structures are stabilized by hydrogen-bonding and van der Waals forces. The van der Waals contacts between the linker residues make them compact and also aid in shielding the non-polar side chains from water, providing extra stabilization due to solvation. These structures are similar to the van der Waals locks (Berezovsky, 2003; Berezovsky & Trifonov, 2001). Particularly, these structures are essential in proteins with small IDLs, where they act as cementing structures across the interface (Berezovsky, 2003; Berezovsky & Trifonov, 2001).

Central to the conformational properties of IDLs is their roles in dictating the inter-domain interactions and orientations. Covalent linking of domains in multi-domain proteins brings various emergent geometric properties associated with protein structure. Crucial to the functioning of proteins among them is IDG. IDG shapes the extent of functional coupling between protein domains (Bashton & Chothia, 2002). Previous work on IDG showed the importance of interface interactions and their variations (Han, Kerrison, Chothia, & Teichmann, 2006). We underlined detailed biophysical explanations to the observed interplay of domain orientations, interactions and linker conformations to these claims. Most multi-domain proteins displayed relatively small χ angle and followed a von Mises distribution; a circular analogue of the normal distribution. For a large proportion of multi-domain proteins, the inter-domain torsional angle is evolutionarily conserved. This conservation among homologous proteins was only weakly correlated with global sequence and structural variations. Hence, the IDG seems affected to a large extent by changes in interacting surfaces and/or IDLs. In the analysis in context of variation of inter-domain orientation, we highlighted that IS-score (Gao & Skolnick, 2010) shows the highest correlation with variation in IDG. Although the variations in IDL alignments and structures are also correlated with IDG, they are significantly lower, indicating that the relative contributions of the interface are greater. The variations across evolutionarily related proteins are primarily due to variations in sequence. These sequence variations manifest as variations in conformational preferences and interactions in the 3D structure. In spite of variances in sequence (AA-scores) and conformational (PB-scores) properties of both interfaces and IDLs being correlated with variation in IDG, there appears to be no significant differences amongst them. Only interface structural variations show the highest correlation with changes in IDG. This accentuates the constraints in preserving the IDG caused by interacting surfaces and interactions. Interaction preservation constrains the interaction geometry. Although the variations in local sequence and structural properties of both interfaces and IDLs vary over evolutionary time, the manifestation of these changes as divergence in IDG is primarily governed by the interaction preservation/variance across the interfaces.

Domain motions are a result of changes in the dihedral angles (φ and ψ) along the main-chain of the proteins. These changes when localized in the IDLs can result in large-scale domain shifts. Internal flexibility of inter-domain orientation is dependent on the energetic costs and the topology of the protein. Proteins which have high circular variance ($\sim$1.00) of IDG are very flexible and undergo large-scale motions easily. We earlier showed that interaction patterns provide the biggest constraints to evolutionary conservation of IDG. An interesting hypothesis is that the interaction energy between the domains affects the flexibility (IDG-Var) of IDG. We found that as the interaction free energy across the domains increases the flexibility decreases. We found limited correlation for this inverse relationship. We also found length dependence for this relationship. This inverse correlation was only observed for short and long size IDL containing proteins. Medium-sized IDL containing proteins did not show any correlation of interaction energy and IDG-Var. This provided us with a possible mechanistic answer to the observed bias in the abundance of short and long size IDLs in multi-domain proteins. Short linkers are constrained and have closely interacting domains, which makes IDG more coupled to the interaction patterns and interaction energy at the interfaces. In the case of long size IDLs, most of them loop back and form stable interactions like anti-parallel β-sheets across the interface resembling closed loops (Berezovsky et al., 2000), thereby making IDG again more coupled to the interface. In the case of medium-sized linkers, the length is not small enough to constrain the interactions and not long enough to loop back to form stable interactions. Here, the IDG seems to be decoupled to the contacts at the interfaces. We previously showed that the inter-domain interactions are more instrumental in stabilizing multi-domain proteins (Bhaskara & Srinivasan, 2011). This stability is also seen in terms of retention of IDG (small IDG-var). Torsional angular shifts across the length of the IDLs were computed by measuring the average PB entropy over the ensemble. This again showed small correlation with the IDG-var, indicating that main-chain conformational switching can aid in the flexibility of IDG. These methods to quantify the main-chain flexibility can be employed to decipher the mechanism of domain motions.

George and Heringa studied IDL sequence and conformational properties and identified two main classes of IDLs, i.e. helical and non-helical (George & Heringa, 2002). The current analysis on IDL revisits the sequence and structural analysis in greater detail using PBs. Comparison of IDL sequence and conformational properties with ISS segments provides us with guidelines to develop methods to identify IDLs in newly sequenced genomes. We show this with proteins having continuous domains where there is ambiguity on domain boundaries. In case of more complex and discontinuous domains, defining the domain boundaries and, hence, the IDLs can be a difficult problem. We believe that our analysis in conjunction with hierarchical domain decomposition methods (Berezovsky, Esipova, & Tumanyan, 2000; Berezovsky, Namiot, Tumanyan, & Esipova, 1999; Koczyk & Berezovsky, 2008) would be useful in understanding the relative importance of multiple IDLs connecting complex domains.

Our results show that the interface interaction preservation across homologous proteins best preserves the IDG. This is instrumental in modelling of multi-domain proteins from the 3D structures of individual domains. To model the full-length proteins from individual domains, either the domain–domain interfaces need to be accurately identified or the accurate loop modelling of the IDLs needs to be performed. We show that the interface identification is quite essential in fixing the IDG, and therefore more instrumental in building accurate full-length protein models.

Conclusions

The IDLs play a central role in the interplay of domains of multi-domain proteins. We showed here, using a PB approach, that IDLs are distinct in structure from inter secondary structures (ISSs). In order to understand the effects of IDLs on IDG, we quantified inter domain geometry (IDG) using a simple geometric descriptor. The distribution of IDG is an evolutionarily conserved phenomenon among homologous multi-domain proteins. In understanding the mechanistic basis of this conserved pattern, we probed at global sequence and structural variations among homologues. To clearly decipher the relationship between protein divergence and conservation/ preservation of IDG, local properties of interfaces and IDL segments were quantified. We highlight that both the IDLs and the interfaces show effects in dictating the conservation of IDG. By measuring their extent of correlations with IDG, we quantified the relative contributions of interface and IDL properties. This illustrated that the interface has a more pronounced effect in dictating IDG in comparison to IDL, although variations in IDLs are high among homologous proteins. Moreover, a length dependence of correlation between the flexibility in IDG and the interaction energy across the interface exists.

Author contributions

RMB, AdB and NS designed the study and wrote the manuscripts text. RMB prepared the Figures 1–7; Table 1; Figures S1–S8; and Tables S1–S6. All authors reviewed the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.10.1080/07391102.2012.743438.

Acknowledgements

We thank Dr Agnel Praveen Joseph, Ms. Garima Agarwal and Ms. Smita Mohanty for discussions and suggestions. R.M.B. is supported by a fellowship from Council of Scientific and Industrial Research, New Delhi. This research is supported by Indo-French CEFIPRA project 3903-1. These works were supported by grants from the Ministry of Research (France), University of Paris Diderot, Sorbonne Paris Cité (France), National Institute of Blood Transfusion (INTS, France), National Institute of Health and Medical Research (INSERM, France) to AdB and Department of Biotechnology (DBT, India) to NS.

References

Abbott, M. B., Gaponenko, V., Abusamhadneh, E., Finley, N., Li, G., Dvoretsky, A., … Rosevear, P. R. (2000). Regulatory domain conformational exchange and linker region flexibility in cardiac troponin C bound to cardiac troponin I. *Journal of Biological Chemistry, 275*, 20610–20617.

Adzhubei, A. A., & Sternberg, M. J. (1994). Conservation of polyproline II helices in homologous proteins: Implications for structure prediction by model building. *Protein Science, 3*, 2395–2410.

Allen, F. H., & Johnson, O. (1991). Automated conformational analysis from crystallographic data. 4. Statistical descriptors for torsion angles. *Acta Crystallographica. Section B, Structural Science, 47*, 62–67.

Altschuh, D., Tessier, D. C., & Vernet, T. (1994). Modulation of the enzymatic activity of papain by interdomain residues remote from the active site. *Protein Engineering, 7*, 769–775.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, 25*, 3389–3402.

Arai, R., Ueda, H., Kitayama, A., Kamiya, N., & Nagamune, T. (2001). Design of the linkers which effectively separate domains of a bifunctional fusion protein. *Protein Engineering, 14*, 529–532.

Aravind, L., & Ponting, C. P. (1999). The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins. *FEMS Microbiology Letters, 176*, 111–116.

Bashton, M., & Chothia, C. (2002). The geometry of domain combination in proteins. *Journal of Molecular Biology, 315*, 927–939.

Bennett, M. J., Choe, S., & Eisenberg, D. (1994). Domain swapping: Entangling alliances between proteins. *Proceedings of the National Academy of Sciences USA, 91*, 3127–3131.

Bennett, M. J., & Eisenberg, D. (1994). Refined structure of monomeric diphtheria toxin at 2.3 A resolution. *Protein Science, 3*, 1464–1475.

Benros, C., de Brevern, A. G., & Hazout, S. (2009). Analyzing the sequence-structure relationship of a library of local structural prototypes. *Journal of Theoretical Biology, 256*, 215–226.

Berezovsky, I. N. (2003). Discrete structure of van der Waals domains in globular proteins. *Protein Engineering, 16*, 161–167.

Berezovsky, I. N., Esipova, N. G., & Tumanyan, V. G. (2000). Hierarchy of regions of amino acid sequence with respect to their role in the protein spatial structure. *Journal of Computational Biology, 7*, 183–192.

Berezovsky, I. N., Grosberg, A. Y., & Trifonov, E. N. (2000). Closed loops of nearly standard size: Common basic element of protein structure. *FEBS Letters, 466*, 283–286.

Berezovsky, I. N., Kirzhner, V. M., Kirzhner, A., Rosenfeld, V. R., & Trifonov, E. N. (2002). Closed loops: Persistence of the protein chain returns. *Protein Engineering, 15*, 955–957.

Berezovsky, I. N., Namiot, V. A., Tumanyan, V. G., & Esipova, N. G. (1999). Hierarchy of the interaction energy distribution in the spatial structure of globular proteins and the problem of domain definition. *Journal of Biomolecular Structure & Dynamics, 17*, 133–155.

Berezovsky, I. N., & Trifonov, E. N. (2001). Van der Waals locks: Loop-n-lock structure of globular proteins. *Journal of Molecular Biology, 307*, 1419–1426.

Berman, H. M., Kleywegt, G. J., Nakamura, H., & Markley, J. L. (2012). The protein data bank at 40: Reflecting on the past to prepare for the future. *Structure, 20*, 391–396.

Bhaskara, R. M., & Srinivasan, N. (2011). Stability of domain structures in multi-domain proteins. *Scientific Reports, 1*, 40.

Bird, R. E., Hardman, K. D., Jacobson, J. W., Johnson, S., Kaufman, B. M., Lee, S. M., Lee, T., Pope, S. H., Riordan, G. S., & Whitlow, M. (1988). Single-chain antigen-binding proteins. *Science, 242*, 423–426.

Bonet-Costa, C., Vilaseca, M., Diema, C., Vujatovic, O., Vaquero, A., Omenaca, N., … Azorin, F. (2012). Combined bottom-up and top-down mass spectrometry analyses of the pattern of post-translational modifications of *Drosophila melanogaster* linker histone H1. *Journal of Proteomics, 75*, 4124–4138.

Briggs, S. D., & Smithgall, T. E. (1999). SH2-kinase linker mutations release Hck tyrosine kinase and transforming activities in Rat-2 fibroblasts. *Journal of Biological Chemistry, 274*, 26579–26583.

Chen, P., Liu, C., Burge, L., Li, J., Mohammad, M., Southerland, W., … Wang, B. (2010). DomSVR: Domain boundary prediction with support vector regression from sequence information alone. *Amino Acids, 39*, 713–726.

Chung, S., Parker, J. B., Bianchet, M., Amzel, L. M., & Stivers, J. T. (2009). Impact of linker strain and flexibility in the design of a fragment-based inhibitor. *Nature Chemical Biology, 5*, 407–413.

Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research, 40*, D71–D75.

de Brevern, A. G. (2005). New assessment of a structural alphabet. *In Silico Biology, 5*, 283–289.

de Brevern, A. G., Etchebest, C., & Hazout, S. (2000). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins, 41*, 271–287.

de Brevern, A. G., Valadie, H., Hazout, S., & Etchebest, C. (2002). Extension of a local backbone description using a structural alphabet: A new approach to the sequence-structure relationship. *Protein Science, 11*, 2871–2886.

de Groot, B. L., van Aalten, D. M., Scheek, R. M., Amadei, A., Vriend, G., & Berendsen, H. J. (1997). Prediction of protein conformational freedom from distance constraints. *Proteins, 29*, 240–251.

Dong, Q., Wang, X., Lin, L., & Xu, Z. (2006). Domain boundary prediction based on profile domain linker propensity index. *Computational Biology and Chemistry, 30*, 127–133.

Dowe, D. L., Allison, L., Dix, T. I., Hunter, L., Wallace, C. S. & Edgoose, T. (1996). Circular clustering of protein dihedral angles by minimum message length. *Pacific Symposium on Biocomputing*, 242–255.

Dumontier, M., Yao, R., Feldman, H. J., & Hogue, C. W. (2005). Armadillo: Domain boundary prediction by amino acid composition. *Journal of Molecular Biology, 350*, 1061–1073.

Ebina, T., Toh, H., & Kuroda, Y. (2009). Loop-length-dependent SVM prediction of domain linkers for high-throughput structural proteomics. *Biopolymers, 92*, 1–8.

Eickholt, J., Deng, X., & Cheng, J. (2011). DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning. *BMC Bioinformatics, 12*, 43.

Ezkurdia, I., Grana, O., Izarzugaza, J. M., & Tress, M. L. (2009). Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins, 77*(Suppl. 9), 196–209.

Fiorani, P., Bruselles, A., Falconi, M., Chillemi, G., Desideri, A., & Benedetti, P. (2003). Single mutation in the linker domain confers protein flexibility and camptothecin resistance to human topoisomerase I. *Journal of Biological Chemistry, 278*, 43268–43275.

Fourrier, L., Benros, C., & de Brevern, A. G. (2004). Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC Bioinformatics, 5*, 58.

Frishman, D., & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins, 23*, 566–579.

Gao, M., & Skolnick, J. (2010). IAlign: A method for the structural comparison of protein-protein interfaces. *Bioinformatics, 26*, 2259–2265.

Gelly, J. C., Joseph, A. P., Srinivasan, N., & de Brevern, A. G. (2011). IPBA: A tool for protein structure comparison using sequence alignment strategies. *Nucleic Acids Research, 39*, W18–W23.

George, R. A., & Heringa, J. (2002). An analysis of protein domain linkers: Their classification and role in protein folding. *Protein Engineering, 15*, 871–879.

Gokhale, R. S., Tsuji, S. Y., Cane, D. E., & Khosla, C. (1999). Dissecting and exploiting intermodular communication in polyketide synthases. *Science, 284*, 482–485.

Han, J. H., Kerrison, N., Chothia, C., & Teichmann, S. A. (2006). Divergence of interdomain geometry in two-domain proteins. *Structure, 14*, 935–945.

Hasegawa, H., & Holm, L. (2009). Advances and pitfalls of protein structural alignment. *Current Opinion in Structural Biology, 19*, 341–348.

Heinig, M., & Frishman, D. (2004). STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research, 32*, W500–W502.

Holm, L., & Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucleic Acids Research, 26*, 316–319.

Ikebe, M., Kambara, T., Stafford, W. F., Sata, M., Katayama, E., & Ikebe, R. (1998). A hinge at the central helix of the regulatory light chain of myosin is critical for phosphorylation-dependent regulation of smooth muscle myosin motor activity. *Journal of Biological Chemistry, 273*, 17702–17707.

Joseph, A. P., Agarwal, G., Mahajan, S., Gelly, J. C., Swapna, L. S., Offmann, B., … De Brevern, A. G. (2010). A short survey on protein blocks. *Biophysical Reviews, 2*, 137–147.

Koczyk, G., & Berezovsky, I. N. (2008). Domain hierarchy and closed loops (DHcL): A server for exploring hierarchy of protein domain structure. *Nucleic Acids Research, 36*, W239–W245.

Kong, L., & Ranganathan, S. (2004). Delineation of modular proteins: Domain boundary prediction from sequence information. *Brief Bioinformatics, 5*, 179–192.

Lu, M., Chai, J., & Fu, D. (2009). Structural basis for autoregulation of the zinc transporter YiiP. *Nature Structural & Molecular Biology, 16*, 1063–1067.

Maeda, Y., Ueda, H., Hara, T., Kazami, J., Kawano, G., Suzuki, E., & Nagamune, T. (1996). Expression of a bifunctional chimeric protein A-Vargula hilgendorfii luciferase in mammalian cells. *BioTechniques, 20*, 116–121.

Mardia K. V., J. P., (Eds.). (2000). *Directional statistics. Probabillity and statistics*. New York, NY: Wiley.

McClendon, A. K., Gentry, A. C., Dickey, J. S., Brinch, M., Bendsen, S., Andersen, A. H., & Osheroff, N. (2008). Bimodal recognition of DNA geometry by human topoisomerase II alpha: Preferential relaxation of positively supercoiled DNA requires elements in the C-terminal domain. *Biochemistry, 47*, 13169–13178.

McLaclan, A. D. (1982). Rapid comparison of protein structures. *Acta Crystallographica, 38*, 871–873.

Miyazaki, S., Kuroda, Y., & Yokoyama, S. (2006). Identification of putative domain linkers by a neural network application to a large sequence database. *BMC Bioinformatics, 7*, 323.

Morra, G., Potestio, R., Micheletti, C., & Colombo, G. (2012). Corresponding functional dynamics across the Hsp90 Chaperone family: Insights from a multiscale analysis of MD simulations. *PLoS Computational Biology, 8*, e1002433.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology, 247*, 536–540.

Nomura, W., Masuda, A., Ohba, K., Urabe, A., Ito, N., Ryo, A., … Tamamura, H. (2012). Effects of DNA binding of the zinc finger and linkers for domain fusion on the catalytic activity of sequence-specific chimeric recombinases determined by a facile fluorescent system. *Biochemistry, 51*, 1510–1517.

R-devel, C. (2011). *R: A language and environment for statistical computing*. Vienna: R foundation for Statistical Computing.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: An online force field. *Nucleic Acids Research, 33*, W382–W388.

Strang, C. J., Wales, M. E., Brown, D. M., & Wild, J. R. (1993). Site-directed alterations to the geometry of the aspartate transcarbamoylase zinc domain: Selective alteration to regulation by heterotropic ligands, isoelectric point, and stability in urea. *Biochemistry, 32*, 4156–4167.

Takizawa, M., Miyauchi, K., Urano, E., Kusagawa, S., Kitamura, K., Naganawa, S., … Komano, J. (2011). Regulation of the susceptibility of HIV-1 to a neutralizing antibody KD-247 by nonepitope mutations distant from its epitope. *AIDS, 25*, 2209–2216.

Tang, Y., Jiang, N., Parakh, C., & Hilvert, D. (1996). Selection of linkers for a catalytic single-chain antibody using phage display technology. *Journal of Biological Chemistry, 271*, 15682–15686.

Traut, T. W. (1988). Enzymes of nucleotide metabolism: The significance of subunit size and polymer size for biological function and regulatory properties. *CRC Critical Reviews in Biochemistry, 23*, 121–169.

Valentini, G., Chiarelli, L., Fortin, R., Speranza, M. L., Galizzi, A., & Mattevi, A. (2000). The allosteric regulation of pyruvate kinase. *Journal of Biological Chemistry, 275*, 18145–18152.

Van Durme, J., Delgado, J., Stricher, F., Serrano, L., Schymkowitz, J., & Rousseau, F. (2011). A graphical interface for the FoldX forcefield. *Bioinformatics, 27*, 1711–1712.

van Leeuwen, H. C., Strating, M. J., Rensen, M., de Laat, W., & van der Vliet, P. C. (1997). Linker length and composition influence the flexibility of Oct-1 DNA binding. *EMBO Journal, 16*, 2043–2053.

Wei, M., Ye, D., & Dunaway-Mariano, D. (2001). Investigation of the role of the domain linkers in separate site catalysis by Clostridium symbiosum pyruvate phosphate dikinase. *Biochemistry, 40*, 13466–13473.

Winkler, F. K., Schutt, C. E., Harrison, S. C., & Bricogne, G. (1977). Tomato bushy stunt virus at 5.5-A resolution. *Nature, 265*, 509–513.

Wriggers, W., Chakravarty, S., & Jennings, P. A. (2005). Control of protein functional dynamics by peptide linkers. *Biopolymers, 80*, 736–746.

Yoo, P. D., Sikder, A. R., Taheri, J., Zhou, B. B., & Zomaya, A. Y. (2008). DomNet: Protein domain boundary prediction using enhanced general regression network and new profiles. *IEEE Transactions on Nanobioscience, 7*, 172–181.

Yoo, P. D., Sikder, A. R., Zhou, B. B., & Zomaya, A. Y. (2008). Improved general regression network for protein domain boundary prediction. *BMC Bioinformatics, 9*(Suppl. 1), S12.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research, 31*, 3370–3374.

Zhang, Y., Liu, B., Dong, Q., & Jin, V. X. (2011). An improved profile-level domain linker propensity index for protein domain boundary prediction. *Protein and Peptide Letters, 18*, 7–16.