Supplementary Information for

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

Ramachandra M. Bhaskara¹, Alexandre G. de Brevern^{2,3,4,5} and Narayanaswamy Srinivasan¹

Affiliations:

¹*Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, INDIA*

²*INSERM UMR-S 665, DSIMB, F-75739 Paris, FRANCE*

³*Univ Paris Diderot, Sorbonne Paris Cité, UMR 665, F-75739 Paris, FRANCE*

⁴*INTS, F-75739 Paris, FRANCE*

⁵*Laboratoire d'Excellence GR-Ex, F75737 Paris, FRANCE*

Contact Information:

E: mail: N.S. ns@mbu.iisc.ernet.in

Telephone: +91-80-22932837

Fax numbers: +91-80-23600535

Molecular Biophysics Unit

Indian Institute of Science

Bangalore 560012

INDIA

URL: <http://nslab.mbu.iisc.ernet.in/>

This document contains the following

Figures S1-S9

Tables S1-S6

Supplemental Methods

Supplementary Results

Supplementary References

Supplementary Figure S1

Fig. S1. Amino acid propensities for IDL and ISS segments: Propensities values of all 20 amino acids to occur in the inter-domain linkers (IDLs) and inter-secondary structures (ISS). The propensity values are rescaled so that the preferred amino-acids have values above 0. Glycine and Proline show the highest preference to occur in the IDLs. In addition to them, polar amino-acids (Ser, Thr, Asp Asn, and His) also have high preference to be found in IDLs. These can serve as distinguishing features of IDLs as they are not correlated ($r=0.27$; $df=19$; $P=0.241$; Spearman rank correlation).

Supplementary Figure S2

Fig. S2. Hydrogen-bonding in IDLs: Proportion of Hydrogen bonds formed by Short (Red), Medium (Blue) and Long (Green) IDLs. H-bonds were assigned by using the HBOND program implemented in JOY software package (Mizuguchi, Deane, Blundell, Johnson & Overington 1998). H-bonds where the donor and acceptor are within the IDL segments are represented by LL, and those where either the donor atom or the acceptor atom is from the domain are marked by LD. The proportion of LL steadily increases; and the LD decreases with the increase in the length of the IDLs.

Supplementary Figure S3

Fig. S3. Propensities of Protein Blocks to occur in IDL and ISS segments: Propensities values of all 16 protein blocks to occur in the inter-domain linkers (IDLs) and inter-secondary structures (ISS). The propensity values are rescaled so that the preferred amino-acids have values above 0. PB *a*, *b*, *c*, *d*, *f*, *j* and *p* have the higher preference to occur in the IDLs.

Supplementary Figure S4

Fig. S4. diPB frequencies in IDLs: Frequency distributions of 256 di-PBs in (A) Short, (B) Medium and (C) Long IDLs. We can see that diPBs (cd , dd and de) corresponding to β -strands and termini of sheets and strands are present in all the three types of IDLs. Helices (ll , lm , mm and mn) are only observed in medium and long IDLs. Di-PBs corresponding to loop regions are seen in all the three Figures at lower overall frequencies.

Supplementary Figure S5

Fig. S5. Distribution of χ in IDLs: Circular histograms of the distributions of IDG (χ) in (A) Short, (B) Medium and (C) Long IDL containing proteins.

Supplementary Figure S6

Fig. S6. Interface contacts in homologous proteins: The number of aligned interface contacts computed by iAlign program (Gao & Skolnick 2010) between the homologous protein pairs. IDG-C set have 20 more interactions than the IDG-NC set indicating that interface is restraining the deviation of the IDG (χ).

Fig. S7. Distribution of average contacts per residue for linkers and non-linker regions:

The number of contacts per residue is computed at a threshold distance of $4.5\text{\AA}$ for each atom of a residue. $i+3$ and $i-3$ residues contacts are not included and are averaged for a given stretch of residues. The average number of contacts per residue for the inter domain linkers ($n = 290$) is significantly lower (Unpaired Students t -test, $t = -16.77$; $df = 289.593$; $P < 2.2\text{E-}16$) than the average contacts per residue from non-linker regions ($n = 115962$). Linker segments were marked using SCOP domain boundary definitions (see Methods). Non-linker segments are defined for each multi-domain proteins using randomly assigned domain boundary. This was iteratively done for 100 times. The lowest per residue contacts for multi-domain proteins in our dataset is 0.12 and the highest is 5.11. The mean value for the linker region is 3.31 ± 1.34 and for randomly selected segment (non-linker) is 4.56 ± 0.83 .

Fig. S8. Relationship between the IDL loop size and the end-to-end distance: The IDL loop size is correlated with the end-to-end distance (Pearson correlation coefficient; $r=0.76$; $P < 0.0001$). The end-to-end distance is measured as the physical distance between the C α atoms of the end residues of the IDL segment. This indicates that Short linkers form close contacts and closed loops more frequently in the IDLs. Note that there are few medium and long sized linkers whose end-to-end distances are below 10Å indicating that they too form closed loop structures.

Fig. S9. Stabilizing interactions within the IDLs: The IDL vary in their interactions to form appropriate stable structures. This has a length dependency. The major contributions to the stabilization of the IDL conformations are derived from the Van der Waals interactions, non-polar residue burial and the Backbone H-bonding energies. The different energy terms were computed using FoldX energy function (Schymkowitz et al. 2005). Only the linker structure was used in these computations to measure the stabilizing force contributions. Although the side-chains H-bonding contribution towards overall stability is less, they are shielded from water due to efficient packing, thereby contributing towards the solvation energy. These structures might resemble the van der Waals locks described by Berezhovsky and Trifonov (Berezhovsky & Trifonov 2001).

Supplementary Table. S1

Amino acid	Propensities			Binomial Probabilities			Remarks
	Short	Medium	Long	Short	Medium	Long	
ALA	0.82	0.67	0.93	6.77E-02	1.32E-03	1.94E-02	- - -
CYS	1.64	1.36	0.74	8.57E-02	5.83E-02	2.33E-02	-
ASP	1.16	0.84	0.96	7.36E-02	3.30E-02	2.98E-02	- -
GLU	1.17	1.27	0.97	6.62E-02	7.67E-03	2.90E-02	++/-
PHE	0.62	1.29	0.85	5.77E-02	1.72E-02	1.21E-02	+/-
GLY	0.91	1.11	1.01	8.64E-02	3.60E-02	2.97E-02	++
HIS	0.94	0.78	1.05	1.60E-01	6.46E-02	4.74E-02	+
ILE	1.14	1.10	1.04	8.26E-02	4.63E-02	2.97E-02	++
LYS	0.82	1.35	1.06	7.91E-02	3.01E-03	2.35E-02	+++
LEU	0.84	0.81	0.97	6.15E-02	1.18E-02	2.42E-02	- -
MET	0.40	1.06	0.82	8.24E-02	1.01E-01	2.98E-02	-
ASN	1.27	0.82	1.03	6.58E-02	3.85E-02	3.60E-02	+/-
PRO	1.68	1.08	1.32	6.25E-03	5.67E-02	9.60E-05	+++
GLN	0.76	0.88	0.98	1.01E-01	6.26E-02	4.10E-02	-
ARG	0.91	1.17	0.95	1.06E-01	3.32E-02	2.99E-02	+/-
SER	0.91	0.90	0.97	9.72E-02	4.90E-02	3.16E-02	- -
THR	0.94	0.88	1.06	1.05E-01	4.82E-02	2.56E-02	+/-
VAL	1.14	1.05	0.96	7.27E-02	5.12E-02	2.76E-02	-
TRP	0.42	1.27	1.17	9.71E-02	6.11E-02	3.09E-02	+
TYR	1.34	0.76	1.07	5.76E-02	2.93E-02	3.10E-02	+/-

Table. S1: Amino acid distribution in short medium and long IDLs:

Propensity values of all 20 amino acids to occur in the Short, Medium and Long IDLs. The preferred amino acids (Propensity value > 1.0) are marked in bold. The next three columns show the binomial probabilities of finding the amino acids in the Short, Medium and Long linkers respectively. The expected probabilities are computed from background amino acid distribution in the entire dataset. The bold values represent significance at $P < 0.01$. The last column represents the overall observed pattern for each amino-acid in all the three groups. The '+' and the '-' symbols represent significant overrepresentation and under representation at a $P < 0.05$.

Supplementary Table. S2

PBs	Propensities			Binomial Probabilities			Gross Str.	Remarks
	Short	Medium	Long	Short	Medium	Long		
<i>a</i>	0.97	0.94	0.85	1.21E-01	7.01E-02	1.14E-02	N-cap β	-
<i>b</i>	1.52	1.18	1.28	1.86E-02	3.31E-02	2.85E-04	N-cap β	++
<i>c</i>	1.45	1.27	1.16	6.06E-03	3.14E-03	1.19E-03	N-cap β	+++
<i>d</i>	1.60	1.47	1.32	1.38E-06	2.18E-10	2.63E-15	β	+++
<i>e</i>	0.89	1.73	1.48	1.55E-01	9.46E-04	7.15E-05	C-cap β	++
<i>f</i>	1.58	1.03	1.01	4.10E-03	5.46E-02	3.07E-02	C-cap β	+
<i>g</i>	2.72	0.77	0.92	2.97E-03	1.05E-01	6.36E-02	mainly coil	+
<i>h</i>	1.66	1.55	1.23	3.57E-02	6.26E-03	1.04E-02	mainly coil	+/-
<i>i</i>	1.50	1.37	1.09	9.05E-02	4.21E-02	4.89E-02	mainly coil	+
<i>j</i>	1.98	1.69	1.42	6.80E-02	2.81E-02	1.10E-02	mainly coil	+
<i>k</i>	1.24	0.88	0.93	6.15E-02	4.67E-02	2.39E-02	N-cap α	-
<i>l</i>	0.81	0.86	0.84	8.65E-02	4.39E-02	5.89E-03	N-cap α	--
<i>m</i>	0.15	0.61	0.78	9.01E-26	1.10E-13	1.00E-06	α	---
<i>n</i>	0.18	0.19	0.62	2.02E-02	8.58E-05	9.46E-04	C-cap α	---
<i>o</i>	0.40	0.56	0.52	3.64E-02	9.82E-03	2.86E-06	C-cap α	---
<i>p</i>	1.06	0.71	0.74	1.25E-01	2.66E-02	1.64E-03	C-cap α to N-cap β	--

Table. S2: Protein blocks (PBs) distribution in short medium and long IDLs:

Propensity values of all 16 PBs to occur in the Short, Medium and Long IDLs. The preferred Structural alphabets (PB) (Propensity value > 1.0) are marked in bold. The next three columns show the binomial probabilities of finding a given structural alphabet in the Short, Medium and Long linkers respectively. The expected probabilities are computed from background distribution of protein blocks in the entire dataset. The bold values represent significance at $P < 0.01$. The next column represents the gross secondary structural features corresponding to the given protein block. The last column represents the overall observed pattern for each amino-acid in all the three groups. The '+' and the '-' symbols represent significant over representation and under representation at a $P < 0.05$.

Supplementary Table. S3

Group (IDL)	μ	ρ	$Var(\chi)$	Watson test statistic	P value
Short	-26.49	0.2790	0.720	0.0241	n.s.
Medium	-43.78	0.1562	0.843	0.0278	n.s.
Long	-69.76	0.2696	0.730	0.0431	n.s.

Table. S3: Statistics of distribution of χ in three groups of IDLs: Table showing the circular mean (μ), mean resultant vector (ρ), circular variance $Var(\chi)$, test statistic and the corresponding P value for Watson's goodness fit test for the von Mises distribution. n.s. indicates non-significance at a $\alpha = 0.05$; indicating that the angles follow von Mises distributions.

Supplementary Table. S4

1A0IA	1DDIA	1GKZA	1JY1A	1N4KA	1QCSA	1SZNA	1V0WA	1Y4WA	2BMWA
1A62A	1DYQA	1GSMA	1K1SA	1N67A	1QHTA	1T1UA	1V4AA	1Y6IA	2BVYA
1A6QA	1DZFA	1GTKA	1K7IA	1NE6A	1QR0A	1T6CA	1V4SA	1YDX	2BW0A
1A76A	1E0CA	1GV2A	1K87A	1NE9A	1QS2A	1T7VA	1VCTA	1YGEA	2C1IA
1A8DA	1E43A	1H09A	1KGSA	1NG2A	1QSAA	1T9HA	1VEMA	1YI9A	2C9AA
1A8PA	1EFYA	1H2WA	1KHBA	1NI3A	1QZZA	1TKEA	1VI7A	1YIOA	2CG7A
1AF7A	1ELVA	1H4UA	1KHIA	1NIJA	1R2JA	1TMOA	1VINA	1YKSA	2CIWA
1B24A	1EN2A	1H6TA	1KJWA	1NKRA	1R4XA	1TUAA	1VLIA	1YQYA	2CVEA
1B9WA	1ENFA	1H8LA	1KKHA	1NL1A	1R5LA	1U04A	1VLOA	1YRWA	2CX1A
1BCOA	1EQFA	1HP1A	1KL9A	1NM8A	1R6XA	1U1JA	1VLYA	1YTQA	2D3IA
1BDGA	1ET9A	1HSKA	1KS9A	1NMLA	1RC9A	1U3DA	1VPDA	1YVRA	2D3NA
1BIKA	1EU1A	1HT6A	1KSKA	1NR0A	1RH1A	1U5PA	1VQZA	1YVUA	2D5BA
1BLFA	1EWFA	1HVXA	1KV9A	1NTYA	1RHS A	1UA7A	1W1OA	1Z2MA	2EIFA
1BU8A	1F20A	1HX0A	1KZLA	1O70A	1RIQA	1UARA	1WD3A	1Z3XA	2FCBA
1BUPA	1F2QA	1IG8A	1L8QA	1OI7A	1RL6A	1UASA	1WF3A	1Z6FA	2FXUA
1C2AA	1F5NA	1IH7A	1LBUA	1OLLA	1RLRA	1UCTA	1WJ9A	1Z77A	2G3RA
1C40A	1F97A	1IHGA	1LCFA	1OWLA	1RP1A	1UD2A	1WLFA	1ZARA	2GKEA
1C96A	1FDRA	1IN4A	1LJ8A	1OXJA	1RRKA	1UEKA	1WMDA	1ZGHA	2GNOA
1CA1A	1FKMA	1IOWA	1LOXA	1P2FA	1RVKA	1UFAA	1WOSA	1ZHVA	2GUYA
1CCZA	1FNDA	1IPAA	1LR7A	1P4XA	1RZ4A	1UGNA	1WV3A	1ZSQA	2HBJA
1CDYA	1FNLA	1IV8A	1LS1A	1P77A	1S2MA	1UHAA	1WXQA	1ZSWA	2I1QA
1CFBA	1FTSA	1J09A	1LSLA	1PGSA	1S35A	1UMKA	1WZAA	1ZY9A	2IBBA
1CIDA	1FVIA	1J5YA	1LY2A	1PIEA	1S5JA	1UNSA	1X38A	2AAAA	2J07A
1CNFA	1G1TA	1JAEA	1M15A	1PIIA	1S6YA	1UOKA	1X6OA	2B20A	2NAPA
1CRZA	1G4RA	1JAKA	1M53A	1PJRA	1SATA	1UT9A	1XC3A	2B3XA	2OLRA
1CX4A	1G5AA	1JB9A	1M9IA	1Q1CA	1SFEA	1UWVA	1XHBA	2B4VA	2SLIA
1D5RA	1G94A	1JBWA	1MD8A	1Q46A	1SQGA	1UWYA	1XOVA	2B78A	3SEBA
1D9XA	1G9KA	1JCFA	1MIXA	1Q7HA	1SQWA	1UX6A	1XTIA	2BIBA	3TSSA
1DCQA	1GIRA	1JU3A	1MXGA	1Q8IA	1SVBA	1UXYA	1Y02A	2BJQA	5EAUA

- i. Resolution better than 3.0 Å
- ii. R-free less than or equal to 35%
- iii. Monomeric proteins in both the asymmetric and biological units
- iv. Proteins having continuous SCOP domain boundaries (Version 1.75)
- v. Non-redundant at 30.0% (A cluster information at 20.0% redundancy is also provided in Table S5)

Table. S4: Dataset of multi-domain proteins (two-domain proteins) with known structure ($n = 290$). These proteins were mined from the PDB (Berman, Kleywegt, Nakamura & Markley 2012). The following criteria were taken into account in the filtering process. The chain information for each PDB is provided by the 5th character.

Supplementary Table S5

1E43A	1VI7A	1UT9A	1LJ8A	1PIEA	1H09A	1Hska	1UEKA	1RZ4A	1RL6A
1HX0A	1Q46A	1KHBA	1RRKA	1R6XA	1WXQA	1IN4A	1EQFA	1GSMA	1WLFA
3SEBA	1EN2A	1X38A	1W10A	2B20A	1JCFA	2CIWA	1ZARA	1MIXA	1G1TA
1NKRA	1IH7A	1D9XA	1J09A	2C1IA	1NE9A	1O70A	1P77A	1CFBA	1Z2MA
1LCFA	2B3XA	1F5NA	1EWFA	1BUPA	1UX6A	1WF3A	1FVIA	1KZLA	1KHIA
1U3DA	1YGEA	1JU3A	1A8DA	2B78A	1A0IA	1BCOA	1IPAA	1YIOA	1Q7HA
1IG8A	1C96A	2BIBA	1PIIA	1DDIA	1N67A	1FTSA	1A8PA	1ET9A	1X6OA
1SATA	1RLRA	2OLRA	2BVYA	1NI3A	1ZSWA	1XC3A	2C9AA	1VCTA	1NL1A
1JB9A	1U1JA	5EAUA	1XHBA	1CA1A	1VQZA	1YI9A	1VINA	1Z77A	1ZHVA
1RHSA	1S5JA	1YVRA	1RIQA	1RVKA	1KKHA	1GTKA	1P4XA	1QCSA	2EIFA
1FNLA	2NAPA	1Y4WA	1F20A	1YDXA	1L8QA	2GNOA	1R5LA	3TSSA	1LY2A
1TMOA	1H2WA	1ZY9A	1WMDA	1IHGA	1NIJA	1KJWA	1H4UA	1TUAA	1A62A
1T1UA	1QHTA	1HP1A	1JY1A	1A6QA	1GKZA	1N4KA	1FDRA	1WJ9A	1C2AA
1UOKA	1Q8IA	1VEMA	1YKSA	1UASA	1NMLA	1H6TA	1Q1CA	1WV3A	2G3RA
2AAAA	1IV8A	1K87A	1SQGA	2FXUA	1A76A	1KS9A	1KSKA	1YTQA	1LSLA
1BU8A	1U04A	1YQYA	2B4VA	1M15A	1XOVA	1YRWA	1QR0A	2CX1A	1BIKA
1GIRA	2SLIA	1ZSQA	1SZNA	1Z6FA	1FKMA	1LS1A	1TKEA	1CDYA	1GV2A
1UWYA	1YVUA	1C4OA	1V4AA	1G4RA	1VLYA	1T9HA	1RC9A	1CIDA	1B9WA
1XTIA	1KV9A	1UFAA	1JBWA	1R2JA	1PGSA	1VPDA	1ZGHA	1NG2A	1Y02A
1WOSA	1M9IA	2D5BA	1S6YA	1EFYA	2I1QA	1DCQA	1LBUA	1SQWA	2CG7A
1MD8A	1LOXA	1JAKA	1UWVA	1VLIA	2BW0A	1AF7A	1DZFA	1CCZA	1LR7A
1CX4A	1G5AA	1V0WA	1HT6A	1K1SA	1D5RA	2GKEA	1S35A	1B24A	
1UMKA	1PJRA	1RH1A	1CRZA	1QZZA	1IOWA	1R4XA	1U5PA	1OXJA	
1Z3XA	1QSAA	1WZAA	1SVBA	1UXYA	1T6CA	1T7VA	2IBBA	1J5YA	
1P2FA	1NR0A	1WD3A	2HBJA	2BJQA	1NTYA	1OI7A	1F97A	1SFEA	

Table. S5: Dataset of multi-domain proteins (two-domain proteins) with known structure at 20.0% redundancy cut-off. ($n = 246$). Blastclust algorithm was used to cluster the set of sequences at 20%. Majority of the ($> 75\%$) sequences were not redundant at 20%. Information on ~50 proteins is lost as they were between the sequence identities between 20 and 30%.

Supplementary Table S6

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>	<i>i</i>	<i>j</i>	<i>k</i>	<i>l</i>	<i>m</i>	<i>n</i>	<i>o</i>	<i>p</i>	-
<i>a</i>	2.78	-0.35	0.30	-0.57	-1.74	-1.13	-0.09	-2.14	-0.31	-0.70	-3.18	-1.41	-3.27	-2.22	-0.80	-0.72	-4.00
<i>b</i>	-0.35	2.78	-0.78	-1.20	-0.81	-2.26	-0.66	0.09	0.71	-0.76	-0.46	-0.09	-3.89	-1.51	-1.01	-0.26	-4.00
<i>c</i>	0.30	-0.78	2.03	-0.38	-1.64	-1.18	0.50	-2.10	-1.48	-1.43	-3.20	-3.83	-2.78	-1.83	-0.64	0.00	-4.00
<i>d</i>	-0.57	-1.20	-0.38	1.41	-0.18	-1.06	-1.88	-1.50	-1.23	-0.89	-3.37	-3.29	-7.99	-5.17	-3.56	-2.89	-4.00
<i>e</i>	-1.74	-0.81	-1.64	-0.18	3.04	0.70	0.82	0.60	-2.01	-0.23	-0.77	-2.90	-5.83	-2.24	-3.09	-3.59	-4.00
<i>f</i>	-1.13	-2.26	-1.18	-1.06	0.70	2.44	-0.68	-0.49	-2.89	0.00	-1.22	-0.77	-2.62	-3.41	-2.43	-2.42	-4.00
<i>g</i>	-0.09	-0.66	0.50	-1.88	0.82	-0.68	3.27	-1.03	-0.42	-1.16	-2.34	-0.95	-0.85	0.30	0.49	1.04	-4.00
<i>h</i>	-2.14	0.09	-2.10	-1.50	0.60	-0.49	-1.03	3.36	-0.47	1.10	0.71	-0.79	-3.67	-2.00	-0.13	-2.48	-4.00
<i>i</i>	-0.31	0.71	-1.48	-1.23	-2.01	-2.89	-0.42	-0.47	3.70	1.01	-0.67	-0.01	-3.93	-2.09	-1.55	0.71	-4.00
<i>j</i>	-0.70	-0.76	-1.43	-0.89	-0.23	0.00	-1.16	1.10	1.01	3.96	0.40	0.02	-2.17	-1.78	-1.01	-0.38	-4.00
<i>k</i>	-3.18	-0.46	-3.20	-3.37	-0.77	-1.22	-2.34	0.71	-0.67	0.40	2.77	-0.63	-1.65	-1.63	-3.01	-2.03	-4.00
<i>l</i>	-1.41	-0.09	-3.83	-3.29	-2.90	-0.77	-0.95	-0.79	-0.01	0.02	-0.63	2.64	-1.08	-0.34	-0.44	-2.06	-4.00
<i>m</i>	-3.27	-3.89	-2.78	-7.99	-5.83	-2.62	-0.85	-3.67	-3.93	-2.17	-1.65	-1.08	1.11	-0.58	-1.31	-1.11	-4.00
<i>n</i>	-2.22	-1.51	-1.83	-5.17	-2.24	-3.41	0.30	-2.00	-2.09	-1.78	-1.63	-0.34	-0.58	3.58	0.04	0.53	-4.00
<i>o</i>	-0.80	-1.01	-0.64	-3.56	-3.09	-2.43	0.49	-0.13	-1.55	-1.01	-3.01	-0.44	-1.31	0.04	3.43	-0.01	-4.00
<i>p</i>	-0.72	-0.26	0.00	-2.89	-3.59	-2.42	1.04	-2.48	0.71	-0.38	-2.03	-2.06	-1.11	0.53	-0.01	2.93	-4.00
-	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	-4.00	1.00

Table. S6: Protein Block substitution matrix used to compute PB-substitution scores. Block substitution scores derived for protein block alignments at 62% identity. These scores are equivalent to the BLOSUM62 scores derived for amino-acids, but represent the ability of local conformational substitutions and preference in the homologous proteins of known structure. The diagonal elements represent the relative proportion of conformational conservations. The last column and the last row represent the conformational substitution scores at amino-acid insertions and deletions. For more details on the scores computations please refer to Joseph A.P. *et al.*, Biochimie (Joseph, Srinivasan & de Brevern 2011).

Supplemental Methods:

Propensity calculations and distribution statistics for amino acids, PBs and di-PBs.

Amino-acid propensities for all the 20 amino-acids to occur in the IDL segments and ISS segments were computed as the ratio of frequency of the i th amino acid in the IDL/ISS segment and the frequency of the same in the entire protein.

$$P_{i,AA} = \frac{f_{i,IDL,AA}}{f_{i,tot,AA}}$$

PB propensities for all the 16 PBs are also computed similarly.

$$P_{i,PB} = \frac{f_{i,IDL,PB}}{f_{i,tot,PB}}$$

di-PB propensities for all PB ij pairs were also computed.

$$P_{ij,PB} = \frac{f_{ij,IDL,PB}}{f_{ij,tot,PB}}$$

Propensities were rescaled while plotting so that the preferred amino-acids/PBs/di-PBs are greater than 0.

The frequencies for all the amino-acids/PBs and diPBs were compared to the background frequencies by χ^2 tests. The statistical estimates for over and under-representations of amino-acids/PBs and di-PBs were computed by computing the binomial tests. The background frequencies were used as the expected probabilities (See Tables S1 and S2).

Supplementary Results

IDLs are distinct from inter-secondary structural elements

Amino-acid distributions of IDLs are distinct.

The distribution of amino acids for short ($\chi^2 = 19.4$; $df = 19$; $P = 0.43$) and long IDLs ($\chi^2 = 23.7$; $df = 19$; $P = 0.21$) are not very different from the rest of the protein, while the medium sized IDLs are different from the rest of the protein showed significant deviations from this distribution ($\chi^2 = 33.2$; $df = 19$; $P = 0.022$). Binomial tests for each amino acid from the short, medium and long IDLs showed that the Gly, His, Ile, Lys and Pro are consistently highly represented in all the three groups of IDLs (See Table S1).

Conformational preferences of IDLs.

PBs *b*, *c* and *d* (corresponding to the β sheet cores and N-ter β) and PBs *h*, *i*, *j* and *k* (corresponding mainly to coil) are predominantly found. Although PBs *l*, *m* and *n* (α -helical segments) are found, they are significantly under represented (See Table S2). We observed that PBs *g* and *k* are relatively more specific to short IDLs. Their frequencies are far higher in comparison to medium and long IDLs (3.5% and 6.8% respectively). The frequency of PB *d* is higher in all the three groups of IDLs.

Medium and Long IDLs form more stable secondary structures.

The ability for stable secondary structure formation of the linker regions were assessed by measuring the H-bonds patterns of the linker regions. The short linkers have more proportion of H-bonds, per linker residue with the rest of the domains (LD) in comparison to the medium and long sized linkers (Fig. S2). The medium and longer linkers on the other hand have a good proportion of H-bonds within the linker segments (LL) itself.

Supplementary References

- Berezovsky, I. N. & Trifonov, E. N. (2001). Van der Waals locks: loop-n-lock structure of globular proteins. *J Mol Biol* 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.
- Gao, M. & Skolnick, J. (2010). iAlign: a method for the structural comparison of protein-protein interfaces. *Bioinformatics* 26: 2259-2265.
- Joseph, A. P., Srinivasan, N. & de Brevern, A. G. (2011). Improvement of protein structure comparison using a structural alphabet. *Biochimie* 93: 1434-1445.
- Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S. & Overington, J. P. (1998). JOY: protein sequence-structure representation and analysis. *Bioinformatics* 14: 617-623.
- Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res* 33: W382-388.