Supplementary Information

for

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

Ramachandra M. Bhaskara[1], Alexandre G. de Brevern[2,3,4,5] and Narayanaswamy Srinivasan[1]

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

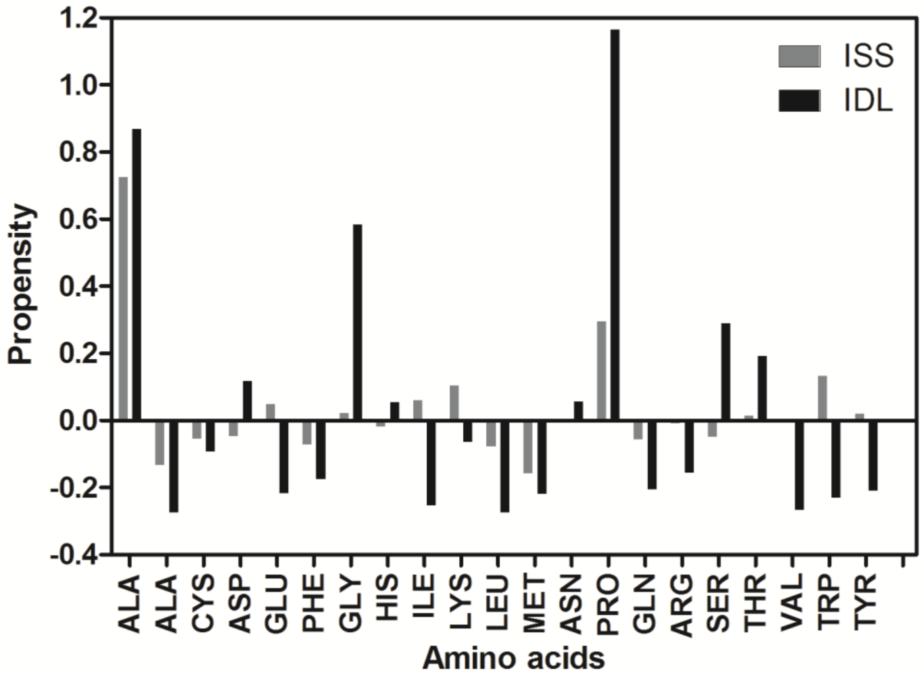
URL: http://pauling.mbu.iisc.ernet.in/

This document contains the following

Figures S1-S6

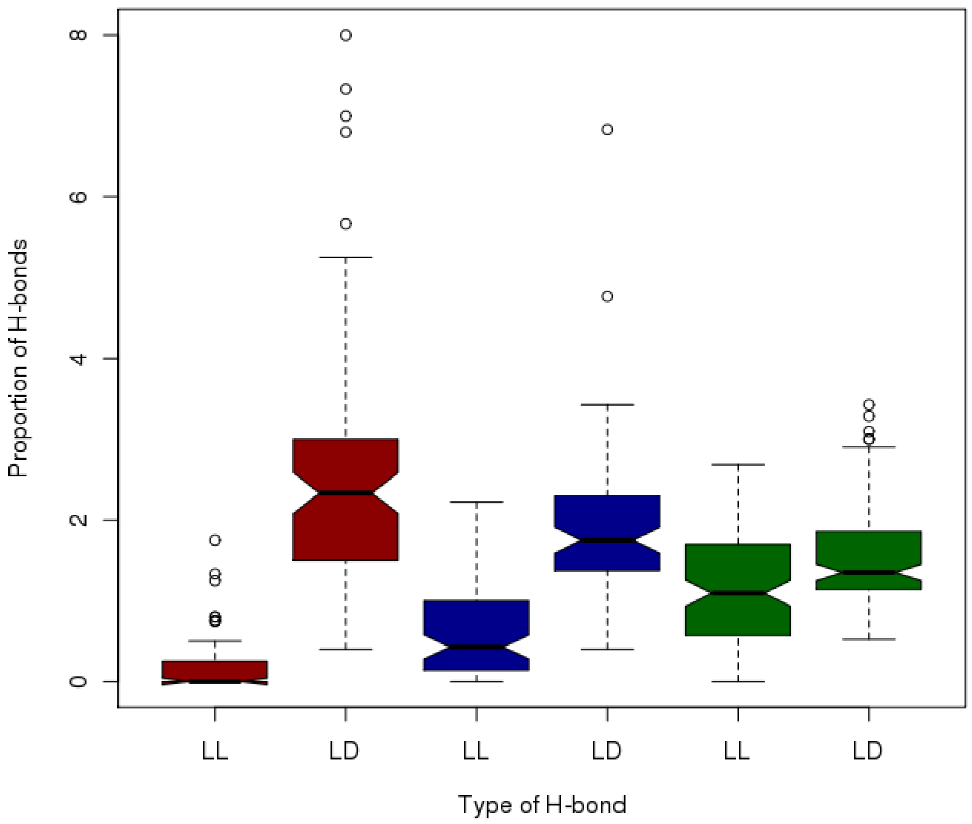
Tables S1-S4

Supplemental methods



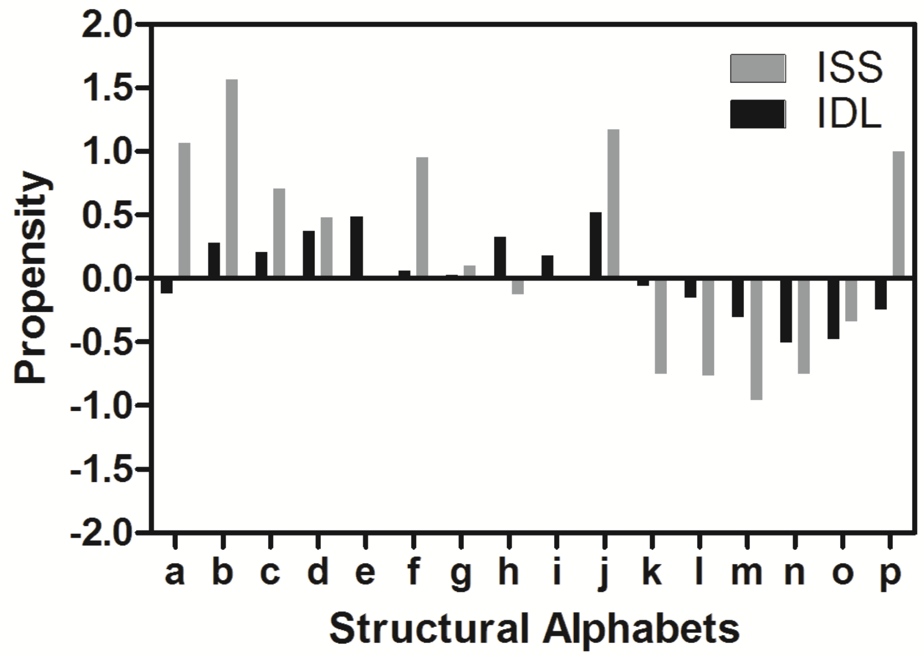
Supplementary Figure S1

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.



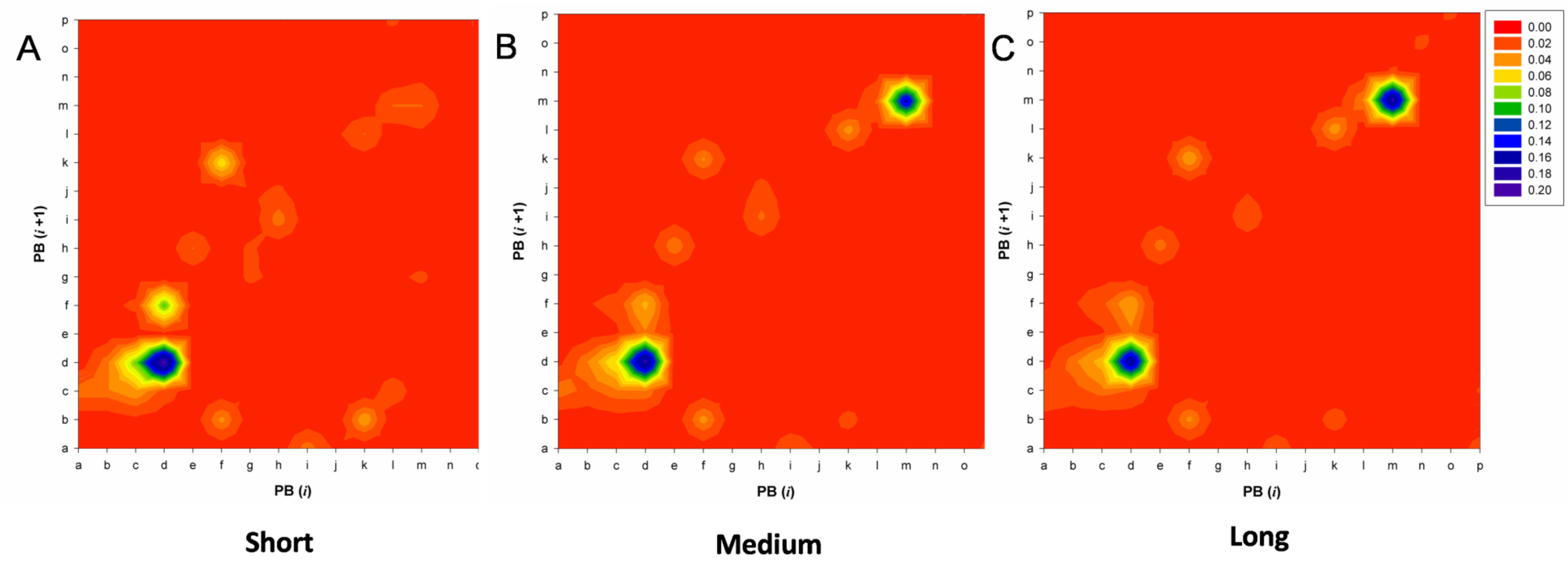
Supplementary Figure S2

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. 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.



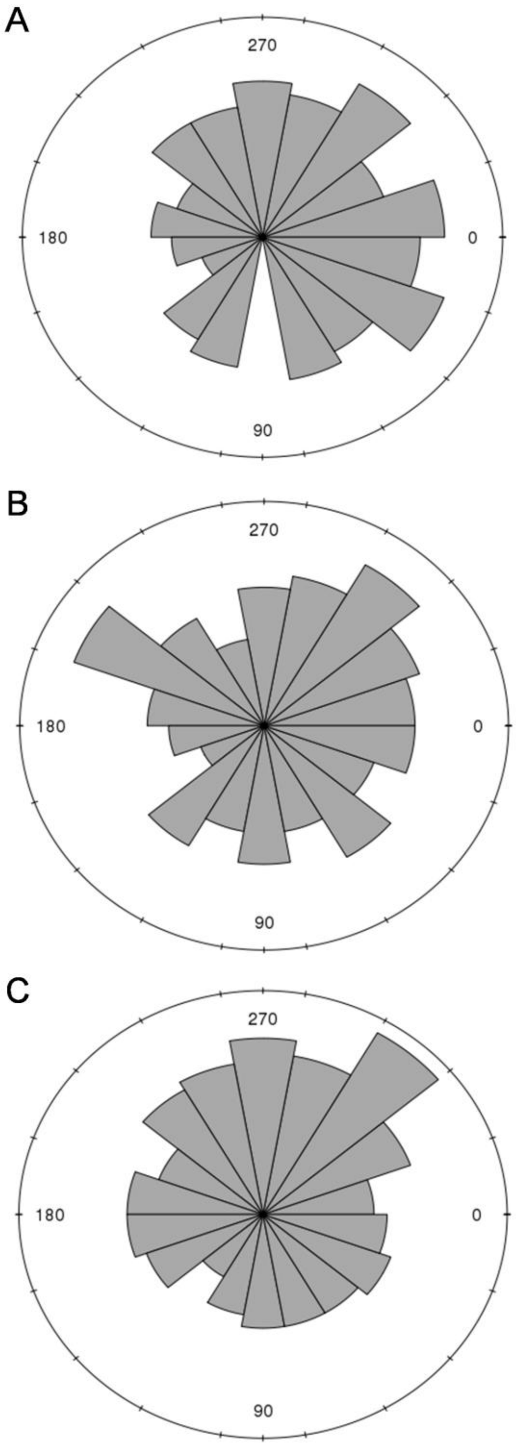
Supplementary Figure S3

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.



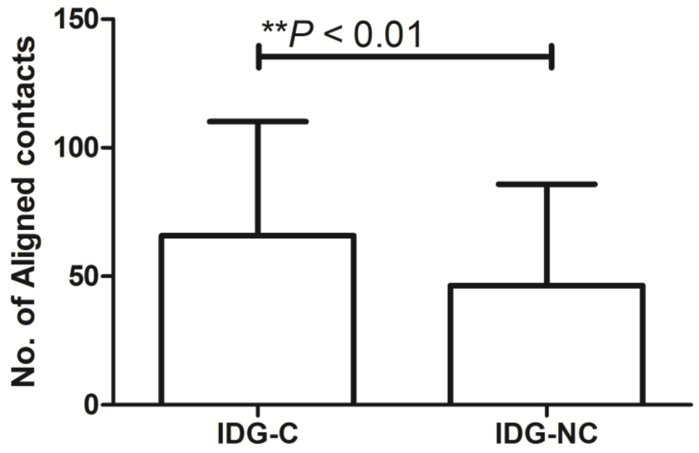
Supplementary Figure S4

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 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).

Supplementary Table. S1

	Propensities			Binomial Probabilities			
Amino acid	**Short**	**Medium**	**Long**	**Short**	**Medium**	**Long**	**Remarks**
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

	Propensities			Binomial Probabilities				
PBs	Short	Medium	Long	Short	Medium	Long	Gross Str.	Remarks
a	0.97	0.94	0.85	1.21E-01	7.01E-02	1.14E-02	N-cap β	-
b	**1.52**	**1.18**	**1.28**	1.86E-02	3.31E-02	**2.85E-04**	N-cap β	+ +
c	**1.45**	**1.27**	**1.16**	**6.06E-03**	**3.14E-03**	**1.19E-03**	N-cap β	+ + +
d	**1.60**	**1.47**	**1.32**	**1.38E-06**	**2.18E-10**	**2.63E-15**	β	+ + +
e	0.89	**1.73**	**1.48**	1.55E-01	**9.46E-04**	**7.15E-05**	C-cap β	+ +
f	**1.58**	**1.03**	**1.01**	**4.10E-03**	5.46E-02	3.07E-02	C-cap β	+
g	**2.72**	0.77	0.92	**2.97E-03**	1.05E-01	6.36E-02	mainly coil	+
h	**1.66**	**1.55**	**1.23**	3.57E-02	**6.26E-03**	1.04E-02	mainly coil	+/-
i	**1.50**	**1.37**	**1.09**	9.05E-02	4.21E-02	4.89E-02	mainly coil	+
j	**1.98**	**1.69**	**1.42**	6.80E-02	2.81E-02	1.10E-02	mainly coil	+
k	**1.24**	0.88	**0.93**	6.15E-02	4.67E-02	2.39E-02	N-cap α	-
l	0.81	0.86	0.84	8.65E-02	4.39E-02	**5.89E-03**	N-cap α	- -
m	0.15	0.61	0.78	**9.01E-26**	**1.10E-13**	**1.00E-06**	α	- - -
n	0.18	0.19	0.62	2.02E-02	**8.58E-05**	**9.46E-04**	C-cap α	- - -
o	0.40	0.56	0.52	3.64E-02	**9.82E-03**	**2.86E-06**	C-cap α	- - -
p	**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 α = 0.05; indicating that the angles follow von Mises distributions.

Supplementary Table. S4

1A0I	1DDI	1GKZ	1JY1	1N4K	1QCS	1SZN	1V0W	1Y4W	2BMW
1A62	1DYQ	1GSM	1K1S	1N67	1QHT	1T1U	1V4A	1Y6I	2BVY
1A6Q	1DZF	1GTK	1K7I	1NE6	1QR0	1T6C	1V4S	1YDX	2BW0
1A76	1E0C	1GV2	1K87	1NE9	1QS2	1T7V	1VCT	1YGE	2C1I
1A8D	1E43	1H09	1KGS	1NG2	1QSA	1T9H	1VEM	1YI9	2C9A
1A8P	1EFY	1H2W	1KHB	1NI3	1QZZ	1TKE	1VI7	1YIO	2CG7
1AF7	1ELV	1H4U	1KHI	1NIJ	1R2J	1TMO	1VIN	1YKS	2CIW
1B24	1EN2	1H6T	1KJW	1NKR	1R4X	1TUA	1VLI	1YQY	2CVE
1B9W	1ENF	1H8L	1KKH	1NL1	1R5L	1U04	1VLO	1YRW	2CX1
1BCO	1EQF	1HP1	1KL9	1NM8	1R6X	1U1J	1VLY	1YTQ	2D3I
1BDG	1ET9	1HSK	1KS9	1NML	1RC9	1U3D	1VPD	1YVR	2D3N
1BIK	1EU1	1HT6	1KSK	1NR0	1RH1	1U5P	1VQZ	1YVU	2D5B
1BLF	1EWF	1HVX	1KV9	1NTY	1RHS	1UA7	1W1O	1Z2M	2EIF
1BU8	1F20	1HX0	1KZL	1O70	1RIQ	1UAR	1WD3	1Z3X	2FCB
1BUP	1F2Q	1IG8	1L8Q	1OI7	1RL6	1UAS	1WF3	1Z6F	2FXU
1C2A	1F5N	1IH7	1LBU	1OLL	1RLR	1UCT	1WJ9	1Z77	2G3R
1C4O	1F97	1IHG	1LCF	1OWL	1RP1	1UD2	1WLF	1ZAR	2GKE
1C96	1FDR	1IN4	1LJ8	1OXJ	1RRK	1UEK	1WMD	1ZGH	2GNO
1CA1	1FKM	1IOW	1LOX	1P2F	1RVK	1UFA	1WOS	1ZHV	2GUY
1CCZ	1FND	1IPA	1LR7	1P4X	1RZ4	1UGN	1WV3	1ZSQ	2HBJ
1CDY	1FNL	1IV8	1LS1	1P77	1S2M	1UHA	1WXQ	1ZSW	2I1Q
1CFB	1FTS	1J09	1LSL	1PGS	1S35	1UMK	1WZA	1ZY9	2IBB
1CID	1FVI	1J5Y	1LY2	1PIE	1S5J	1UNS	1X38	2AAA	2J07
1CNF	1G1T	1JAE	1M15	1PII	1S6Y	1UOK	1X6O	2B20	2NAP
1CRZ	1G4R	1JAK	1M53	1PJR	1SAT	1UT9	1XC3	2B3X	2OLR
1CX4	1G5A	1JB9	1M9I	1Q1C	1SFE	1UWV	1XHB	2B4V	2SLI
1D5R	1G94	1JBW	1MD8	1Q46	1SQG	1UWY	1XOV	2B78	3SEB
1D9X	1G9K	1JCF	1MIX	1Q7H	1SQW	1UX6	1XTI	2BIB	3TSS
1DCQ	1GIR	1JU3	1MXG	1Q8I	1SVB	1UXY	1Y02	2BJQ	5EAU

Table. S4: Dataset of multi-domain proteins (two-domain proteins) with known structure (*n* =290).

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).