Structures of distantly related interacting protein homologs are less divergent than non-interacting homologs

Nagarajan Naveenkumar[1,2], Vasam Manjveekar Prabantu[1], Sneha Vishwanath[1], Ramanathan Sowdhamini[2] (iD) and Narayanaswamy Srinivasan[1]

1 Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India
2 National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India

Homologous proteins can display high structural variation due to evolutionary divergence at low sequence identity. This classical inverse relationship between sequence identity and structural similarity, established many years ago, has remained true between homologous proteins of known structure over time. However, a large number of heteromeric proteins also exist in the structural data bank, where the interacting subunits belong to the same fold and maintain low sequence identity between themselves. It is not clear if there is any selection pressure to deviate from the inverse sequence–structure relationship for such interacting distant homologs, in comparison to pairs of homologs which are not known to interact. We examined 12,824 fold pairs of interacting homologs of known structure, which includes both heteromers and multi-domain proteins. These were compared with monomeric proteins, resulting in 26,082 fold pairs as a dataset of non-interacting homologous systems. Interacting homologs were found to retain higher structural similarity than non-interacting homologs at diminishing sequence identity in a statistically significant manner. Interacting homologs are more similar in their 3D structures than non-interacting homologs and have a preference towards symmetric association. There appears to be a structural constraint between remote homologs due to this commitment.

Protein evolution is characterized by sequence and structural differences between homologous proteins. Relationship between sequence similarity and structural divergence between homologous proteins was first described by Chothia and Lesk [1]. They analyzed 32 homologous pairs of proteins of known structure and derived an inverse relationship between sequence identity and root mean square deviation (RMSD). This landmark work made a major impact especially in the areas of protein structure prediction and modeling.

Apart from the relationship between sequence similarity and structural divergence, it has been shown that the structural homologs can share low sequence identity [2,3]. In their analysis, of the 32 homologous pairs, there are six protein pairs with low sequence identity (less than 35%) and low structural difference (RMSD value less than 1.5 Å) [1]. This observation indicates that some homologous proteins retain highly similar 3D structures even at low sequence identity. With the availability of many protein structures, the

1

homology between proteins is more confidently ascertained as a result of structural similarity and by consideration of their functions [4]. The classic work by Chothia and Lesk has been revisited by many further works [5,6] which have laid the foundation to study all areas of protein structure prediction, inverse folding and many other. They have contributed to improving our understanding of evolution of protein structures [7–10].

These previous analyses were largely confined to homologous proteins which do not interact with each other, i.e., homologous proteins from different organisms that are not known to interact with each other, or non-interacting paralogs were considered. In this study, we have performed analysis for a special class of homologous protein systems, where the homologous protein modules are known to interact with each other as heteromers and interacting homologous domains in multi-domain proteins. Interacting protein systems are known to co-evolve in order to maintain their interactions [11–13]. We aim to explore if co-evolution of interacting homologs influence extent of similarity in their three dimensional structures [14,15]. We investigated this aspect using a dataset of 3D structures of complexes of interacting remote homologs and compared it with non-interacting homologs of known 3D structures from PDB [16].

Materials and methods

Three different datasets, corresponding to non-interacting homologs, heterodimers with homologous subunits and domain repeats of multi-domain proteins were created for our analysis. All the three datasets constitute proteins of known crystal structure determined at 3 Å or better resolution, and are non-redundant at 95% sequence identity (i.e., no two entries in the dataset have more than 95% sequence identity). These are discussed further below.

Dataset of non-interacting homologs

We considered homologous protein structures that exist as monomer in the asymmetric or biological unit as deposited in the Protein Data Bank; these monomeric homologs are from the same organism as well as from different organisms and are not known to interact with each other. The evolutionary relatedness between the monomers are identified based on their placement in the structural hierarchy at the fold level by SCOPe database.

The non-interacting homologs dataset used in the current work consists of 26,082 pairs (provided in File S1) of homologs corresponding to 2126 monomeric proteins of known structure from PDB.

Dataset of interacting homologs

Heterodimers

We obtained 3D structures of protein assemblies from PDB and considered all pairs of interacting subunits. In our analysis, we refer them as heterodimers although they could be a part of multimeric protein assembly. We chose those pairs of subunits which are interacting and are likely evolutionarily related. The interacting protein subunits were ensured to correspond to the same structural fold by SCOPe definition [17]. Protein–protein interface residues are identified using Protein Interaction Calculator [18]. The subunits are considered to be interacting only, when the geometric mean of number of interface residues from both subunits are greater than five. This dataset consists of 12,639 interacting sub-units or heterodimers (provided in File S2) with interacting homologous subunits corresponding to 1875 PDB entries.

Domain repeats of multi-domain proteins

The third dataset consists of homologous domains which are evolutionarily related and interacting in a muti-domain protein. These are two-domain proteins from PDB, where both the domains in the protein chain correspond to same structural fold by SCOPe definition. Hence, these are PDB entries where two structurally similar repeat-domains with significant inter-domain interactions. Spatial interaction between the two domains was confirmed using the same criteria as used for heterodimers. This dataset finally composed of 185 domain pairs (provided in File S3) corresponding to 185 PDB entries.

Both the datasets described above correspond to datasets of interacting proteins of known structure.

Structural similarity measures

Structural superposition and structure-based sequence identity have been carried out using TM-align algorithm [19], whereas the structural difference between two superimposed protein structures has been measured by Structural Distance Metric (SDM) [20]. SRMS is a converted similarity measure of RMSD and a convenient representation that scales between 0 and 1. It is calculated as $1 - \text{RMSD}/5$ Å. Pairwise fractional topological equivalence (PFTE) is the ratio of the number of equivalences to the total number of residues in the smaller protein. PYMOL has been used for graphics visualization and presentation of 3D structures of proteins in the analysis [21].

$$\text{SDM} = -100\log[(w_1 * \text{SRMS}) + (w_2 * \text{PFTE})]$$

$$\text{SRMS} = 1 - \text{r.m.s.d in } \mathring{A}/5\mathring{A}$$

$$\text{PFTE} = \frac{\text{No. of equivalent C}\alpha \text{ atoms}}{\text{No. of residues in the smallest protein}}$$

$$w_1 = \frac{[(1-\text{SRMS}) + (1-\text{PFTE})]}{2}$$

$$w_2 = \frac{(\text{SRMS} + \text{PFTE})}{2}$$

The structural asymmetry in the heterodimers, for topologically equivalent C-alpha atoms, was calculated using a method proposed for homodimers earlier [22].

Statistical test and plotting

All the plotting, curve fitting, and statistical tests were performed using numpy, scipy, and pylab packages from PYTHON(Python software foundation, Beaverton, OR, USA).

Results and discussion

The datasets comprise of 26,082 pairs (2126 PDB entries) of non-interacting homologs, 12,639 pairs (1875 PDB entries) corresponding to interacting homologous subunits of heterodimers and 185 entries corresponding to 185 interacting homologous domains from multi-domain protein structures. The structures in every pair in each of these datasets were superimposed on each other followed by calculating their RMSD, SDM and sequence identity. Figure 1 shows the distribution between sequence identity and SDM score for each of the three datasets. Figure 1A scatter plot which corresponds to monomeric proteins (non-interacting homologs), that spans all values of sequence identity. Figure 1B,C shows the scatterplots for heterodimers with homologous subunits (interacting homologs) and interacting two-domain proteins
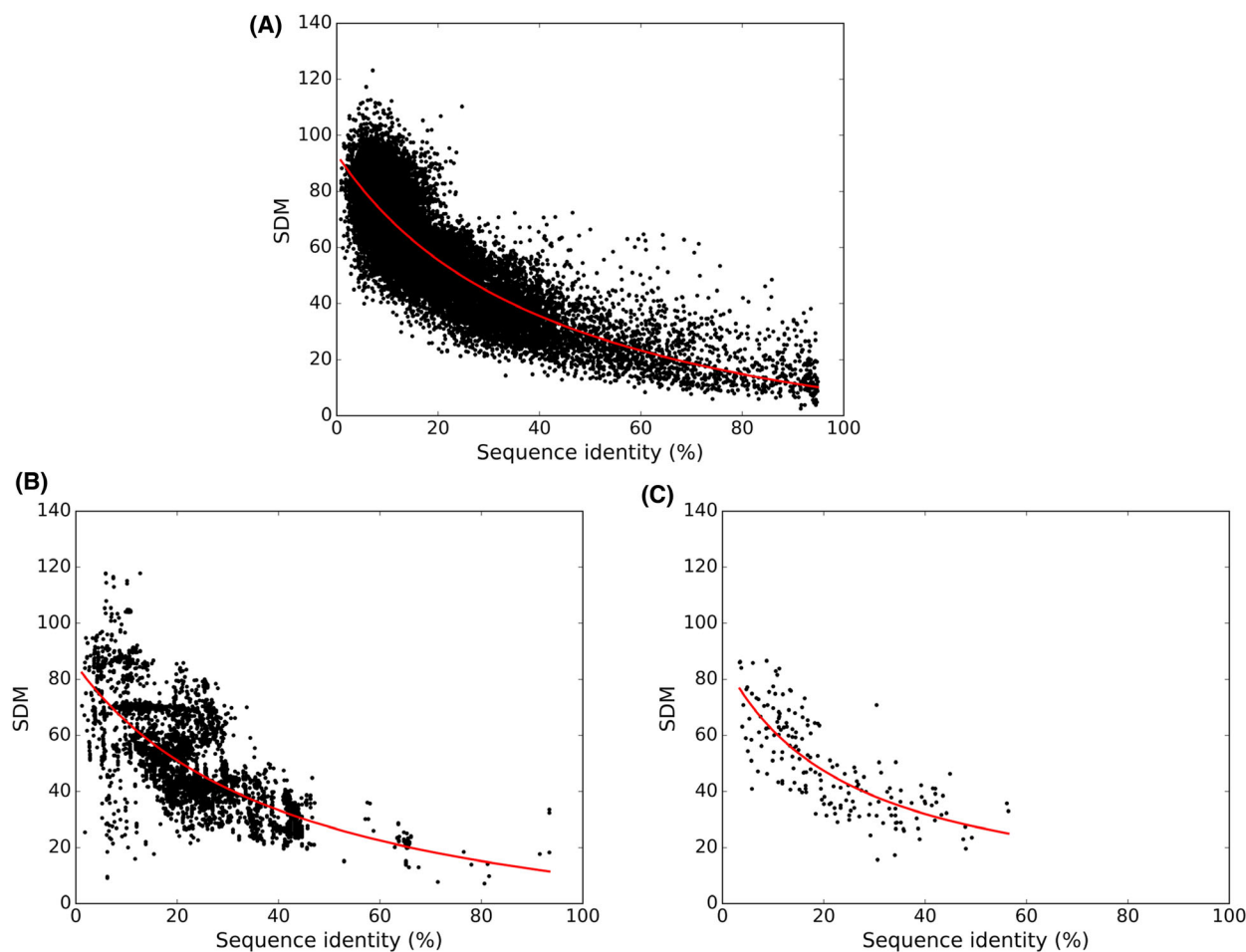

Fig. 1. Relationship between sequence similarity and structural deviation. Scatter plot with sequence identity along *X*-axis and SDM score along *Y*-axis. (A) Representing non-interacting homologous pairs of monomeric proteins, (B) Scatter plot of interacting homologous subunits of heterodimers and (C) Scatter plot of interacting homologous domains of two-domain repeats. The best fit line in red color shows the trend of the distribution.

with interaction between homologous domains (inter-acting homologs), respectively. Overall, only 854 entries out of 12,824 entries retain sequence identity higher than 40% sequence identity. We have 11 such entries in domain repeat-dataset (11 out of 185 entries) (Fig. 1C) and 843 entries in heterodimers (843 out of 12 639 entries; Fig. 1B). Low frequency of occurrence of domain repeats with high sequence identity has also been noticed in previous studies [2,3]. Overall, inverse relationship between sequence identity and structural divergence is seen in all the three datasets that was also observed in Chothia-Lesk analysis for a smaller dataset of non-interacting homologs. The best fit line highlighted in red color in Fig. 1A–C indicates the trend of the distribution obtained by fitting the data in the equation of $(a/(x + b) + c)$. The parameters for three datasets are mentioned in Table 1. This equation was used to capture the hyperbolic inverse relation between SDM and sequence identity.

Owing to sparse data in the two datasets of interacting homologs, we combined the datasets of interacting homologs and compared the distribution with that of non-interacting homologs. Figure 2A indicates that the best fit line for interacting homologs tend to be lower than the non-interacting homologs especially at low sequence identity ranges. Further, we compared the two datasets specifically at the low sequence identity range of 0–40%. The data corresponding to 0–40% sequence identity range gives rise to 24,021 numbers of non-interacting homologous pairs and 4858 numbers of interacting homologous pairs. We calculated the mean and standard deviation of SDM scores for every 5% sequence identity bin, for both non-interacting and interacting homologs and shown the distribution in Fig. 2B, along with statistical significance provided in Table 2. The difference in the distribution of the points for each interval between the interacting and non-interacting homologs were checked for using the Kolmogorov–Smirnov test and the difference between the median and the variance were statistically confirmed using Mann–Whitney U test. It can be clearly observed that the mean values of SDM scores for interacting homologs is lower than non-interacting homologs for much of the sequence identity range, especially at the sequence identity less than 25%. This suggests that structural similarity between the interacting homologs is generally higher than the structural similarity between non-interacting homologs, at sequence identity less than 25%. It was interesting to observe that the interacting domains have higher structural deviation than the non-interacting domains in the sequence identity range of 25–30%.

KIn the current work, although the interacting homologs belong to the same fold, they cannot form a perfectly symmetric complex, since the subunits are

Table 1. The values of the parameters a, b, and c for each of the dataset.

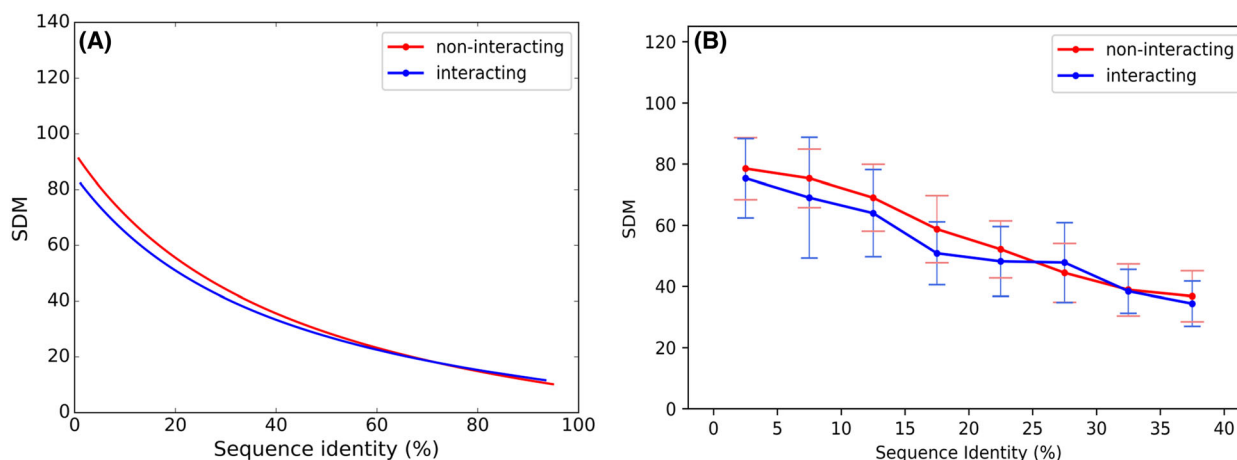| Dataset	a	b	c
Heteromer	4651.99	43.13	−22.61
Domain	2254.55	24.99	−2.71
Non-interacting	5507.15	44.95	−29.24

Fig. 2. Inverse relationship of sequence similarity and structural difference of interacting and non-interacting homologous systems. Statistical difference between distributions for non-interacting and interacting homologous systems with sequence identity along *X*-axis and SDM score along *Y*-axis. The red color indicates the non-interacting homologs and blue color indicates the interacting homologs. (A) Best fit lines representing the trend of data distribution for non-interacting homologs and interacting homologs. (B) Mean (indicated by dots) and standard deviation (indicated by bars) of SDM scores for every 5% of sequence identity ranging between 0% and 40%.

Table 2. The statistical significance in terms of *P*-value for every 5% sequence identity range of the SDM score distribution between interacting and non-interacting homologs.

| Sequence identity range (%)	KS-test (*P*-value)	Mann–Whitney *U* test (*P*-value)
≤ 5	7.65E-05	0.0692
5 < SI ≤ 10	< 2.6E-16	3.02E-12
10 < SI ≤ 15	< 2.6E-16	< 2.6E-16
15 < SI ≤ 20	< 2.6E-16	< 2.6E-16
20 < SI ≤ 25	< 2.6E-16	< 2.6E-16
25 < SI ≤ 30	< 2.6E-16	5.5E-04
30 < SI ≤ 35	< 2.6E-16	0.3007
35 < SI ≤ 40	< 2.6E-16	6.5E-04

chemically different owing to low sequence identity between them. This phenomenon of inherent asymmetry was observed by us earlier [22]. We had devised a measure referred as 'asymmetric score' that calculates the extent of deviation from perfect symmetry for a pair of subunits in heterodimers, i.e., for a complex generated by crystallographic symmetry axis, the asymmetry score is zero. This method was proposed earlier for homodimers [22]. We have now adapted it for heterodimers, where only the topologically equivalent C-alpha atoms are considered for calculating asymmetry score to measure global asymmetry. Our earlier studies had shown inherent asymmetry plays a vital role in structure and function of the homodimers [22]. We performed similar study for few cases of heterodimers, case 1: Ku heterodimer (DNA bound [PDB ID: 1JEY] and Unbound [PDB ID: 1JEQ]) and case 2:

Endosomal adaptor protein (p14)/MEK-binding partner 1 (Mp1) [PDB ID: 1VEU] heterodimer. The first case is not part of our dataset but was chosen for the presentation since it interestingly exhibits asymmetry upon DNA binding.

Ku heterodimer (Ku70 and Ku80 subunits) binds to DNA double-strand breaks and facilitates non-homologous end joining pathway of DNA repair [23]. The Ku70 and Ku80 subunits share a three domain topology comprising of an amino-terminal α/β domain, central β-Barrel domain and a helical C-terminal region [23]. These three different domain topologies are structurally similar, but collectively contribute to deviation from symmetry due to conformational differences between the subunits. The heterodimeric subunits form a preformed ring-like structure, encircling the free-end of the DNA element in a sequence non-specific manner [23] (Fig. S1A). Comparison of the DNA-bound form with its unbound form indicates that this complex turns asymmetric upon DNA binding. The unbound form is assigned a gross asymmetry score of 1.93 (Fig. S1B), whereas the DNA-bound form acquires gross asymmetry score of 3.32 indicating that the asymmetry is required to perform its function.

Endosomal adaptor protein (p14)/MEK-binding partner 1 (Mp1) heterodimer is an endosomal adaptor/scaffold complex which regulates mitogen-activated protein kinase (MAPK) signaling. Together, they form a tight dimer (with a Kd of 12.8 nм [24]; Fig. S2A). With an apparent symmetric association, this complex also possess an inherent asymmetry due to difference in their sequences (low sequence identity of 12.5%).

Fig. 3. Histogram of asymmetry scores of self-interacting heterodimers. Values are shown in bins for every 5 Å by bars (along *X*-axis) and their relative frequency of number of entries (along *Y*-axis). Three illustrative dimeric structures are shown corresponding to three ranges (PDB codes and asymmetry scores are provided).
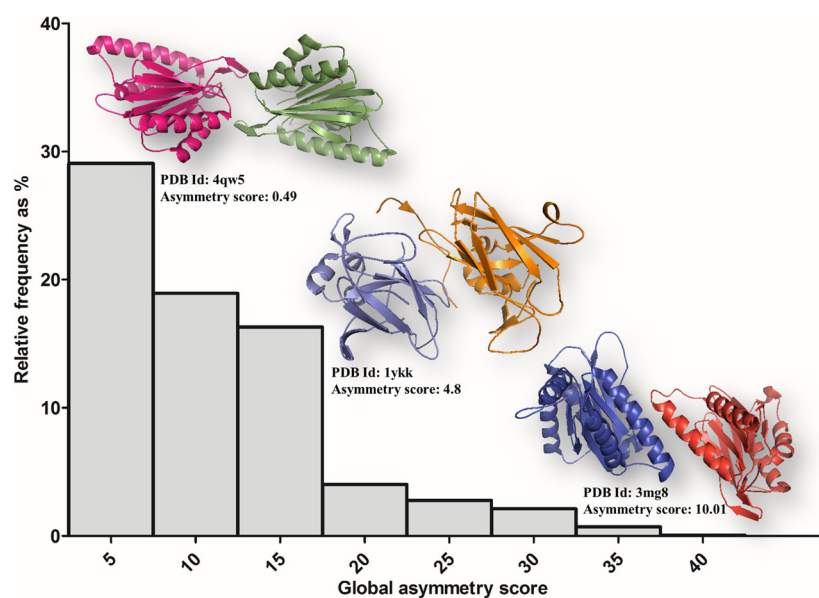
The structural asymmetry (gross asymmetry score is 1.64) in a complex is achieved by conformation and orientation of the interacting protein domains. The superposition of p14 and MP1 in (Fig. S2B) shows they are structurally similar (RMSD is 2.60) with conformational differences in the loop regions. The asymmetric loop as highlighted in blue color is functionally important and is required to target p14 to late endosomes [24].

Figure 3 shows the distribution of asymmetry scores for the heterodimers used in the dataset, which clearly shows that the frequency distribution is highest for dimers with low asymmetry scores. The inserts in Fig. 3 shows the 3D structures of three heterodimers characterized by different asymmetry scores of 0.49, 4.8 and 10.01 for illustration. These observations indicate preference towards low asymmetry in heterodimers with homologous subunits. One of the requirements for the low asymmetry is high similarity in tertiary structures of the interacting subunits, which were reflected as low SDM scores as well (see previous section).

Conclusion

Using available 3D structures and analysis of three types of datasets from PDB, we have re-examined the relationship between sequence identity and structure divergence between homologous proteins, with a commitment to engage in interactions (interacting homologs). As expected, the structural deviation is shown to decrease with increase in sequence identity. The structural deviation between interacting homologs is lower than structural deviation between non-interacting homologs for a given sequence identity range. This implies that interacting homologs are more similar in their 3D structures than non-interacting homologs. We also report in this article that such interacting homologs prefer to retain symmetrical association. Therefore, there could be an underlying structural constraint for interacting homologs to retain their overall tertiary structural similarity between them and symmetry in quaternary structure, even at low sequence identity. The inferences drawn from our analyses of remotely related interacting homologs, would hopefully enable future computational methods and approaches for modeling higher order structures and assemblies involving association between homologous chains.

Acknowledgements

Conflict of interest

The authors declare no conflict of interest.

Data availability statement

All data generated or analyzed during this study are included as additional files.Data accessibility

Author contributions

NS and RS conceived the idea for the work. NN and VMP have equal contribution to the work. NN has compiled the manuscript. SV has contributed in formulating the dataset and manuscript preparation.

References

1 Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J*. 1986;**5**:823–6.

2 Sudha G, Naveenkumar N, Srinivasan N. Evolutionary and structural analyses of heterodimeric proteins composed of subunits with same fold. *Proteins*. 2015;**83** (10):1766–86.

3 Wright CF, Teichmann SA, Clarke J, Dobson CM. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature*. 2005;**438**(7069):878–81.

4 Lo Conte L, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. *Nucleic Acids Res*. 2000;**28**(1):257–9.

5 Argos P. Computer analysis of protein structure. *Methods Enzymol*. 1990;**182**:751–76.

6 Orengo CA, Flores TP, Taylor WR, Thornton JM. Identification and classification of protein fold families. *Protein Eng Des Sel*. 1993;**6**(5):485–500.

7 Fiser A, Šali A. MODELLER: generation and refinement of homology-based protein structure models. *Methods Enzymol*. 2003;**374**:461–91.

8 Godzik A, Kolinski A, Skolnick J. De novo and inverse folding predictions of protein structure and dynamics. *J Comput Aided Mol Des*. 1993;**7**(4):397–438.

9 Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. *Nat Biotechnol*. 2012;**30**(11):1072–80.

10 Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993;**234**(3):779–815.

11 Perica T, Chothia C, Teichmann SA. Evolution of oligomeric state through geometric coupling of protein interfaces. *Proc Natl Acad Sci USA*. 2012;**109**(21):8127–32.

12 Teichmann SA. The constraints protein-protein interactions place on sequence divergence. *J Mol Biol*. 2002;**324**(3):399–407.

13 Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci USA*. 2005;**102**(31):10930–5.

14 Park J, Lappe M, Teichmann SA. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J Mol Biol*. 2001;**307**(3):929–38.

15 Marsh JA, Teichmann SA. Structure, dynamics, assembly, and evolution of protein complexes. *Annu Rev Biochem*. 2015;**84**:551–75.

16 Berman HM et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;**28**(1):235–42.

17 Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995;**247**(4):536–40.

18 Tina KG, Bhadra R, Srinivasan N. PIC: protein interactions calculator. *Nucleic Acids Res*. 2007;**35**:W473–6.

19 Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. 2005;**33**(7):2302–9.

20 Johnson MS, Sutcliffe MJ, Blundell TL. Molecular anatomy: phyletic relationships derived from three-dimensional structures of proteins. *J Mol Evol*. 1990;**30**(1):43–59.

21 DeLano WL. Pymol: an open-source molecular graphics tool, {CCP4} Newsl. *Protein Crystallogr*. 2002;**40**:1–8.

22 Swapna LS, Srikeerthana K, Srinivasan N. Extent of structural asymmetry in homodimeric proteins: prevalence and relevance. *PLoS ONE*. 2012;**7**(5):e36688.

23 Walker JR, Corpina RA, Goldberg J. Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. *Nature*. 2001;**412**(6847):607–14.

24 Kurzbauer R, Teis D, de Araujo ME, Maurer-Stroh S, Eisenhaber F, Bourenkov GP, et al. Crystal structure of the p14/MP1 scaffolding complex: how a twin couple attaches mitogen-activated protein kinase signaling to late endosomes. *Proc Natl Acad Sci USA*. 2004;**101**(30):10984–9.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

File S1. Dataset of non-interacting homologs: Monomeric protein pairs.

File S2. Dataset of Interacting homologs: Heterodimers.

File S3. Dataset of Interacting homologs: Domain repeats of multi-domain proteins.

Fig. S1. Ku heterodimeric complex. (A) Ku heterodimer (Ku70 and Ku80 subunits) bound to DNA [PDB ID: 1JEY]. (B) DNA-unbound form of Ku heterodimer [PDB ID: 1JEQ].

Fig. S2. p14/MP1 scaffolding complex. (A) p14/MP1 heterodimer [PDB ID: 1VEY], interface residues involved in dimerization are shown in sticks. (B) Superposition of p14 and MP1 protein domains. The loop region highlighted in blue is involved in targeting p14 to late endosome.