Structures Of Remotely Related Interacting Protein Homologues Are Less Divergent Than Non-Interacting Homologues

Nagarajan Naveenkumar

Indian Institute of Science

Vasam Manjveekar Prabantu

Indian Institute of Science

Sneha Vishwanath

University of Cambridge

Ramanathan Sowdhamini

National Centre for Biological Sciences

Narayanaswamy Srinivasan (✉ iiscsrinivasan@gmail.com)

Indian Institute of Science

Research Article

Keywords: Interacting homologues, Protein-protein interaction, Sequence identity, Structural similarity, Monomer, Heteromer, Two domain repeat, Multi-domain protein and Asymmetry.

Posted Date: February 24th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1375647/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Homologous proteins could display high structural variations felt 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, 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 homologues, in comparison to pairs of homologues which are not known to interact.

Results: We examined 12,824 fold pairs of interacting homologues of known structure, which includes both heteromers and multi-domain proteins. These were compared with monomeric proteins, resulting 26,082 fold pairs as a dataset of non-interacting homologous systems. Our analysis suggests that interacting homologues retain higher structural similarity than non-interacting homologues at diminishing sequence identity in a statistically significant manner.

Conclusion: Interacting homologues are more similar in their 3D structures than non-interacting homologues and has a preference towards symmetric association. There appears to be a structural constraint between remote homologues due to this commitment.

Introduction

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 homologues can share low sequence identity [2, 3]. In their analysis, out of 32 homologous pairs, there are six protein pairs with low sequence identity (less than 35%) and low structural difference (RMSD value less than 1.5Å) [1]. This observation indicates that there are homologous proteins retain highly similar 3-D structures even at low sequence identity. With the availability of many protein structures, the 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 and 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 in 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 homologues influence extent of similarity in their three dimensional structures [14, 15]. We investigated this aspect using a dataset of 3-D structures of complexes of interacting remote homologues and compared it with non-interacting homologues of known 3-D structures from PDB [16].

Results And Discussion

The datasets comprise of 26,082 pairs (2,126 PDB entries) of non-interacting homologues, 12,639 pairs (1,875 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 homologues), contain entries that spans all values of sequence identity. Figure 1B and 1C show the scatterplots for heterodimers with homologous subunits (interacting homologues) and interacting two-domain proteins with interaction between homologous domains (interacting homologues) respectively. The number of entries higher than 40% sequence identity is meager (Fig. 1C). 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 homologues. The best fit line highlighted in red color in Figs. 1A, 1B and 1C indicate the trend of the distribution obtained by fitting the data in the equation of $(a/(x + b) + c)$.

Owing to sparse data in the two datasets of interacting homologues, we combined the datasets of interacting homologues and compared the distribution with that of non-interacting homologues. Figure 2A indicates that the best fit line for interacting homologues tend to be lower than the non-interacting homologues 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 4,858 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 homologues and shown the distribution in Fig. 2B, along with statistical significance provided in Table 1. It can be clearly observed that the mean values of SDM scores for interacting homologues is lower than non-interacting homologues for much of the sequence identity range. This difference is statistically significant (Table 1).

This suggests that structural similarity between the interacting homologues is generally higher than the structural similarity between non-interacting homologues.

Table 1
The statistical significance in terms of P-value for every 5% sequence identity range of the SDM score distribution between interacting and non-interacting homologues.

Sequence identity range (%)	p-value
≤ 5	7.65E-05
5 < SI ≤ 10	< 2.6E-16
10 < SI ≤ 15	< 2.6E-16
15 < SI ≤ 20	< 2.6E-16
20 < SI ≤ 25	< 2.6E-16
25 < SI ≤ 30	< 2.6E-16
30 < SI ≤ 35	< 2.6E-16
35 < SI ≤ 40	< 2.6E-16

In the current work, although the interacting homologues belong to the same fold, they cannot form a perfectly symmetric complex, since the subunits are chemically different owing to low sequence identity between them. This phenomenon of inherent asymmetry was observed by us earlier [17]. 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 [17]. 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.

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 (please 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 homologues). As expected, the structural deviation is shown to decrease with increase in sequence identity. The structural deviation between interacting homologues is lower than structural deviation between non-interacting homologues for a given sequence identity range. This implies that interacting homologues are more similar in their 3D structures than non-interacting homologues. We also report in this paper that such interacting homologues prefer to retain symmetrical association. Therefore, there could be an underlying structural constraint for interacting homologues 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 homologues, would hopefully enable future computational methods and approaches for modeling higher order structures and assemblies involving association between homologous chains.

Materials And Methods

Three different datasets, corresponding to non-interacting homologues, 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 3Å 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 homologues

We considered homologous protein structures that are known to exist as monomer; these monomeric homologues 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 homologues dataset used in the current work consists of 26082 pairs (provided in *additional_file_1.txt*) of homologues corresponding to 2126 monomeric proteins of known structure from PDB.

Dataset of Interacting homologues

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[18]. Protein-protein interface residues are identified using Protein Interaction Calculator[19]. 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 *additional_file_2.txt*) 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 multi-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 *additional_file_3.txt*) 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 [20], whereas the structural difference between two superimposed protein structures has been measured by SDM Structure distance metric [21]. Pymol has been used for graphics visualization and presentation of 3D structures of proteins in the analysis[22].

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

$$SRMS = 1 - r.m.s.d \text{ in } \text{\AA}/3.5$$

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

$$w_1 = \frac{[(1 - SRMS) + (1 - PFTE)]}{2} \quad w_2 = \frac{(SRMS + PFTE)}{2}$$

The structural asymmetry in the heterodimers, for topologically equivalent C-alpha atoms, was calculated using a method proposed for homodimers earlier [17].

List Of Abbreviations

PDB - Protein Data Bank

RMSE - Root mean square deviation

SDM - Structure distance metric

Declarations

Ethical approval and consent to participate

Not applicable

Availability of data and materials

All data generated or analyzed during this study are included in additional files

Contributor Information

Narayanaswamy Srinivasan, Email: iiscsrinivasan@gmail.com

Ramanathan Sowdhamini, Email: mini@ncbs.res.in

Naveenkumar Nagarajan, Email: vasanthccmb@gmail.com

Vasam Manjveekar Prabantu, Email: vasamp@iisc.ac.in

Sneha Vishwanath, Email: sv446@cam.ac.uk

Declarations

Consent for publications

Competing interests

There are no competing interests

Funding

Research from NS group is supported by the Department of Science and Technology (DST), University grants commission (UGC), Department of Biotechnology (DBT), Government of India.

Author's 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.

Acknowledgements

NS and RS's JC Bose grant (SERB) and Bioinformatics Centre grants (DBT) are acknowledged. RS acknowledges funding and support provided by JC Bose Fellowship (JBR/2021/000006) from Science and Engineering Research Board, India and Bioinformatics Centre Grant funded by Department of Biotechnology, India (BT/PR40187/BTIS/137/9/2021).

References

1. Chothia C, Lesk AM: The relation between the divergence of sequence and structure in proteins. *The EMBO journal* 1986, 5(4):823–826.
2. Sudha G, Naveenkumar N, Srinivasan N: Evolutionary and structural analyses of heterodimeric proteins composed of subunits with same fold. *Proteins* 2015, 83(10):1766–1786.
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):878–881.
4. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C: SCOP: a structural classification of proteins database. *Nucleic acids research* 2000, 28(1):257–259.
5. Argos P: Computer analysis of protein structure. *Methods in enzymology* 1990, 182:751–776.
6. Orengo CA, Flores TP, Taylor WR, Thornton JM: Identification and classification of protein fold families. *Protein engineering* 1993, 6(5):485–500.
7. Fiser A, Sali A: Modeller: generation and refinement of homology-based protein structure models. *Methods in enzymology* 2003, 374:461–491.
8. Godzik A, Kolinski A, Skolnick J: De novo and inverse folding predictions of protein structure and dynamics. *Journal of computer-aided molecular design* 1993, 7(4):397–438.
9. Marks DS, Hopf TA, Sander C: Protein structure prediction from sequence variation. *Nature biotechnology* 2012, 30(11):1072–1080.
10. Sali A, Blundell TL: Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology* 1993, 234(3):779–815.
11. Perica T, Chothia C, Teichmann SA: Evolution of oligomeric state through geometric coupling of protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America* 2012, 109(21):8127–8132.
12. Teichmann SA: The constraints protein-protein interactions place on sequence divergence. *Journal of molecular biology* 2002, 324(3):399–407.
13. Mintseris J, Weng Z: Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* 2005, 102(31):10930–10935.
14. Park J, Lappe M, Teichmann SA: Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *Journal of molecular biology* 2001, 307(3):929–938.

15. Marsh JA, Teichmann SA: Structure, dynamics, assembly, and evolution of protein complexes. *Annual review of biochemistry* 2015, 84:551–575.
16. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. *Nucleic acids research* 2000, 28(1):235–242.
17. Swapna LS, Srikeerthana K, Srinivasan N: Extent of structural asymmetry in homodimeric proteins: prevalence and relevance. *PloS one* 2012, 7(5):e36688.
18. Murzin AG, Brenner SE, Hubbard T, Chothia C: SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology* 1995, 247(4):536–540.
19. Tina KG, Bhadra R, Srinivasan N: PIC: Protein Interactions Calculator. *Nucleic acids research* 2007, 35(Web Server issue):W473-476.
20. Zhang Y, Skolnick J: TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 2005, 33(7):2302–2309.
21. Johnson MS, Sutcliffe MJ, Blundell TL: Molecular anatomy: phyletic relationships derived from three-dimensional structures of proteins. *Journal of molecular evolution* 1990, 30(1):43–59.
22. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

Figures

Figure 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.

Figure 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 homologues and blue color indicates the interacting homologues. A) Best fit lines representing the trend of data distribution for non-interacting homologues and interacting homologues. B) Mean (indicated by dots) and standard deviation (indicated by bars) of SDM scores for every 5% of sequence identity ranging between 0-40%.

Figure 3

Histogram of asymmetry scores of self-interacting heterodimers

Values are shown in bins for every 5Å 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).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [additionalfile1.txt](#)
- [additionalfile2.txt](#)
- [additionalfile3.txt](#)