A DATASET OF PROTEIN-PROTEIN INTERFACES FOR DRUG REPURPOSING

Zeynep Abalı[1], Attila Gursoy[2], and Ozlem Keskin[3]

[1] Computational Science and Engineering Program, Koc University, 34450, Istanbul, Turkey
[2] Computer Engineering, Koc University, 34450, Istanbul, Turkey
[3] Chemical and Biological Engineering, Koc University, 34450, Istanbul, Turkey
phone: + (90) 537 277 52 40, email: zabali16@ku.edu.tr

ABSTRACT

Proteins have crucial functions in many processes, ranging from structural building blocks to signaling or acting as catalysts to store and transport different molecules. Most functions of proteins are not conducted in isolation by a monomer (a single chain of protein structure) but with the interaction of multiple partners. Understanding the structural architecture of protein interfaces is one of the key challenges in explaining how proteins interact and function. Considering the crucial functions of proteins, protein-protein interfaces can be important targets for drug discovery and repurposing studies; this is only possible if the structural data available can be utilized in a complete, biologically correct, and technically accessible manner. Protein Data Bank (PDB) is an invaluable resource for protein structures, but it is highly redundant in nature. In this study, we analyzed interface structures from PDB; we compared protein-protein interfaces with respect to amino acid sequences of their chains, and we compared sequentially unique interfaces structurally to form a unique set of interface representatives as a dataset for further studies.

1. BACKGROUND

Proteins interact with each other through regions called interfaces. Even though the number of protein structures determined each year is increasing rapidly, in 2004, it is estimated that there should be approximately 10,000 structural protein interactions [1]. This suggests that the interaction structures of proteins are highly conserved, and different proteins interact with each other in similar ways. Knowing unique types of protein interactions or knowing the protein interactions that occur in the same way have important implications. For example, a drug that binds and inhibits the action of a certain molecule may also be a drug candidate for a different disease, with the same structure of interaction, even though the proteins that take part in the interaction are different.

2. METHODS

Figure 1 shows an overview of the methods.

2.1 Identifying Interface Structures from PDB

All interface structures from PDB [2] are identified using the following criteria [3]:

- Each interface chain should have at least 5 residues
- Residues that are within Van der Waals (VDW) radii + 0.5Å are contacting residues
- Any residue with a Cα atom within 6 Å distance from any atom of a contacting residue is a nearby residue

2.2 Structural Comparison of Sequentially Similar Interfaces

Sequences of all chains that include an interface chain are clustered using MMSeqs2 [4]. If two interfaces are coming from sequentially similar chains, they are grouped together into sequential similarity groups.
Interfaces within each sequential similarity group were compared structurally using iAlign [5]. A threshold of 0.311 (corresponding to a p-value of $1x10^{-5}$) IS-score is used to identify structurally similar interfaces. Agglomerative hierarchical clustering is used for forming sequentially and structurally unique clusters.

2.3 Structural Comparison of Sequentially and Structurally Unique Interfaces

Representatives from sequentially and structurally unique clusters and remaining interface structures that are not in sequential similarity groups are structurally compared using iAlign. Complete-linkage hierarchical clustering is used with a threshold value of 0.279 IS-score (p-value of $1x10^{-4)}$ for the final structural clusters.

3. RESULTS

As of March 2021, PDB had 176,570 protein structures available. Within these structures, we identified 771,246 two-chain structures. Further analysis of two-chain structures showed that there are 534,203 interface structures that meet the interface criteria.
Grouping of interfaces according to their sequence similarity produced 74,248 sequentially unique dimer structures, where 37,422 of which include two or more interface structures. Structural comparison of interfaces from sequentially similar dimer groups resulted in 83,239 sequentially and structurally unique interfaces.
2,336,894,731 structural comparisons are performed to compare all sequentially and structurally unique interfaces with each other. Clustering of similar interfaces resulted in 97,143 structurally unique representative interface structures.

REFERENCES

1. Aloy, P. and R.B. Russell, *Ten thousand interactions for the molecular biologist.* Nat Biotechnol, 2004. **22**(10): p. 1317-21.
2. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
3. Cukuroglu, E., et al., *Non-redundant unique interface structures as templates for modeling protein interactions.* PLoS One, 2014. **9**(1): p. e86738.
4. Steinegger, M. and J. Soding, *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.* Nat Biotechnol, 2017. **35**(11): p. 1026-1028.
5. Gao, M. and J. Skolnick, *iAlign: a method for the structural comparison of protein-protein interfaces.* Bioinformatics, 2010. **26**(18): p. 2259-65.

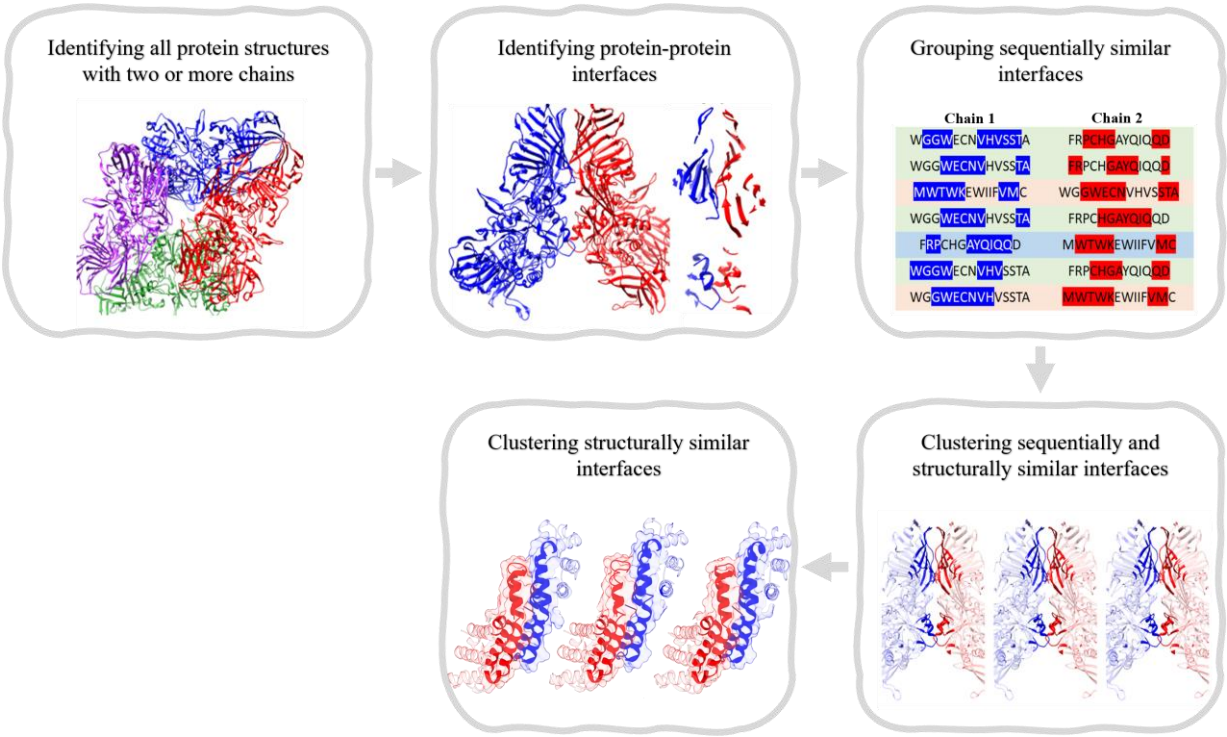

Figure 1 An overview of the methods is provided here. We first identified all protein structures with two or more chains and the interface structures between them. After identifying the interfaces and dimers that form the interfaces, we compared the dimers sequentially and interfaces structurally to create a set of representative interfaces from the available structural protein data.