DOCKGROUND Membrane Protein-Protein Set

Ian Kotthoff, Petras J. Kundrotas* and Ilya A. Vakser*

*Corresponding authors: Ilya A. Vakser and Petras J. Kundrotas, Computational Biology Program, The University of Kansas, 2030 Becker Drive, Lawrence, Kansas 66045; Tel: (785) 864-1057, Fax: (785) 864-5558, Email: vakser@ku.edu and pkundro@ku.edu

Running title: Membrane protein-protein set

Keywords: Protein interactions; protein modeling; protein docking

ABSTRACT

Membrane proteins are significantly underrepresented in Protein Data Bank despite their essential role in cellular mechanisms and the major progress in experimental protein structure determination. Thus, computational approaches are especially valuable in the case of membrane proteins and their assemblies. The main focus in developing structure prediction techniques has been on soluble proteins, in part due to much greater availability of the structural data. Currently, structure prediction of protein complexes (protein docking) is a well-developed field of study. However, the generic protein docking approaches are not optimal for the membrane proteins because of the differences in physicochemical environment and the spatial constraints imposed by the membranes. Thus, docking of the membrane proteins requires specialized computational methods. Development and benchmarking of the membrane protein docking approaches has to be based on high-quality sets of membrane protein complexes. In this study we present a new dataset of 456 non-redundant alpha helical binary interfaces. The set is significantly larger and more representative than the previously developed sets. In the future, it will become the basis for the development of docking and scoring benchmarks, similar to the ones for soluble proteins in the DOCKGROUND resource http://dockground.compbio.ku.edu.

Introduction

Membrane proteins account for a large part (up to 25%) of the human proteome. These proteins individually and in association with other membrane proteins, perform a wide range of functions, such as transporting nutrients, maintaining electrochemical gradients, cell-cell signaling, and structural support [1]. Recent advances in cryogenic electron microscopy have made it possible to determine the structure of increasingly large number of membrane proteins [2, 3]. However, they are still significantly underrepresented among structures in the Protein Data Bank [2, 4]. Experimental determination of the 3D structures of protein-protein complexes, in general, is more difficult than that of the individual proteins, compounding the difficulty of determining protein structures in the membrane. Thus, computational methods for prediction of protein-protein complexes (protein docking) are essential for structural characterization of protein-protein interactions in the membranes. The membrane environment constrains protein-protein interactions by limiting protein insertion angles and depths [5, 6]. Thus, the dimensionality of the docking space in membranes is less than that for the soluble protein-protein complexes. In soluble proteins, a coarse-grained representation determined by the global fold often suffices for a meaningful prediction [7]. However, the recognition factors in membrane proteins are smaller in scale than those in the soluble protein-protein complexes. Thus, docking of the membrane proteins require atomic-level accuracy [8, 9].

Because of the combination of structural and physicochemical characteristics of the membrane proteins that distinguish them from the soluble ones and the specifics of the membrane environment, docking methodologies developed for the soluble proteins are not optimal for the membrane proteins [10]. Thus, specialized computational methods for docking of the membrane proteins have to be developed. In order to accomplish that, one needs high-quality datasets of membrane protein-protein complexes, necessary for the development and benchmarking of such methods. The sets have to be large enough to ensure statistical reliability of the results. Existing sets of membrane protein-protein complexes, contain a relatively small numbers of entries [5]. Koukos et al. describe a complex set of 37 transmembrane targets [11]. The Memdock benchmark consists of 65 target alpha helical transmembrane complexes [10]. In this paper we present a new dataset of 456 non-redundant alpha helical binary interfaces, as the foundation for the future development of the comprehensive resource for structural studies of membrane protein-protein complexes.

Results and Discussion

Generation of Dataset

Initial PDB biounit structures used in this study were downloaded from the Orientation of Proteins in Membrane's alpha helical transmembrane database [12]. This database contains both the structure of the protein and the computationally determined membrane. At the time of retrieval (October 2019), the dataset contained 4,359 alpha helical and 530 beta-barrel protein structures. Beta-barrel membrane proteins are found almost exclusively in Gram-negative bacteria, mitochondria and chloroplasts [13]. Because of that, the number of such structures is relatively small. Beta-barrel proteins are also structurally distinct form the helical membrane embedded proteins. Thus, beta-barrel structures require development of different docking approaches and, consequently, specialized datasets for their development and benchmarking. Therefore, we restricted our set to alpha-helical proteins only. After all monomeric proteins were filtered out, 3,359 entries remained. They were further split by the protein chain. To keep only the transmembrane part of the protein, the extramembrane parts of the structures were deleted. All binary interactions formed between any two chains were considered forl dataset entry. Thus, one PDB structure could yield several interacting pairs. To characterize the interface size, we used FreeSASA [14] to calculate solvent accessible solvent area (SASA) buried upon protein binding (for that purpose, treating the membrane proteins like the soluble ones). Two chains were considered interacting if their buried SASA was > 250 Å^2 per chain. This resulted in 7,964 pairwise combinations of the transmembrane segments.

To remove redundancy in the protein set, one can consider sequence-based criteria. However, aligning and calculating identity for fragmented sequences of the transmembrane parts is not a trivial and straightforward task. Thus, we chose to remove the redundancy at the level of combined tertiary and quaternary structures. For that, we determined all-against-all TM-scores produced by MM-align [15] (hereafter referred to as MM-score), an offshoot of the structural alignment program align, specifically designed for aligning multi-chain structures. The dataset was clustered by Highly Connected Subgraphs method [16] with various MM-score cutoffs ranging from 0.4 to 1.0. The clustering threshold was optimized by analyzing the number of resulting clusters and the fraction of singletons (clusters with one element). The optimal value of the MM-score clustering cutoff can be selected downwards starting from the point with a significant decrease in the number of clusters with smaller clustering cutoff (MM-score < 0.7 in Figure 1A). Another consideration for choosing the clustering threshold is the minimal number of the singletons (Figure 1B). Based on these two considerations, we selected the optimal clustering cutoff at MM-score 0.6, which yielded 456 clusters. The largest cluster contained 851 interfaces, 153 clusters were singletons and 48 clusters contained two interfaces. Representative structures from the clusters for inclusion into final dataset were those with the best structure resolution. If two or more representative structures had the same resolution, the one with the least missing residues was selected. An example of a cluster and its representative is shown in Figure 2. To determine the level of similarity in each cluster, we calculated RMSD values between cluster members. The RMSD ranged from 4.6 Å to 0 Å with the average 1.1 Å.

In a real case docking scenario, the bound structures of the proteins would not be available. Thus, we investigated the availability of experimentally determined unbound structures corresponding to the structures in our set. Psi-blast was used to locate unbound chains at sequence identity cutoff 0.6 and coverage of the alignment 75%. The search did not find any complexes for which both chains had an unbound structure. Thus, the focus of developing more adequate benchmark sets for docking of the membrane proteins should be on simulated/modeled unbound structures [17-19]. The transmembrane parts of the proteins in our dataset are straight or kinked helices. For them, simulation of the unbound structures by modeling would result in the side-chain repacking and, possibly, in minor changes in the helix-helix angles, which is an easy case for protein-protein docking according to the commonly accepted classification of unbound structures. Given the essential lack of the experimental data on the membrane bound/unbound protein-protein complexes, such task will require effort beyond our current report.

Analysis of Dataset

Membrane environment imposes restrictions on helix insertion angles. Also, the helices in the membrane can be straight or have a kink. We analyzed differences/similarities in the arrangements of helices belonging to the same or different chains. We calculated the angles between all pairs of interacting helices in the final dataset. The distribution of the angles was analyzed separately for the pairs of intra- and inter-chain helices. Angles were calculated between vectors connecting N- and C-termini of a transmembrane helix. The vectors were drawn by performing a linear regression through all C^{α} atoms of the helix. To assign the vector unambiguously, we excluded short helices (those consisting of less than two turns or eight residues). The length of a helix was defined by a continuous stretch of eight or more alpha-helical residues as determined by DSSP [20, 21]. With such definition, angles < 90° indicated parallel helices, and those > 90° - the antiparallel ones. We used two alternative distance cutoffs to determine whether a pair of helices is interacting: any C^{α} atom of one helix to any C^{α} atom of the other helix (i) < 6 Å and (ii) < 12 Å (an empirical value based on maximizing docking success rates for soluble proteins [22]). Distributions of the helix-helix angles for both cutoffs (Figure 3) are practically indistinguishable. Thus, here we discuss the results obtained with the 6 Å cutoff only.

Significant part of the dataset (96 non-redundant entries) contains kinked helices where one or two non-helical residues were present between longer stretches of the alpha-helical residues (≥ 8 residues). For such cases, angles were considered separately between vectors drawn through each part of the kinked helix (Figure 4). This resulted in 1,270 pairs of interacting kinked and 5,725 pairs of non-kinked helices. Distribution of interacting angles for such pairs are shown in Figure 3, along with the distributions of interaction angles between 5,783 of kinked and 16,009 pairs of non-kinked interacting helices, belonging to the same polypeptide chain.

The inter-chain helix-helix interactions occurred more frequently between parallel helices, whereas the intra-chain interactions were more commonly formed by the anti-parallel helices (sequentially adjacent helices are more likely to interact). Distributions for the kinked helices showed preference for 20-25° kink angles (in ~50° and ~140° peaks of the interacting kinked helices distributions). The kinked helices accounted for most interactions angles close to 90° indicating the membrane environment pressure for parallel arrangement of long helices.

The membrane protein-protein set is incorporated in the DOCKGROUND resource for protein recognition studies http://dockground.compbio.ku.edu, in its Bound protein-protein part. The membrane set page (Figure 5) allows download of the entire set, or the individual complexes, along with their visual analysis.

Conclusions and Future Work

Membrane proteins play an essential role in cellular mechanisms. Despite that and the major progress in experimental structure determination, they are still significantly underrepresented in Protein Data Bank. While computational approaches to protein structure determination are important in general, they are especially valuable in the case of membrane proteins and protein-protein assemblies. Due to a number of reasons, not the least of which is much greater availability of structural data, the main focus of structure prediction techniques has been on soluble proteins.

Structure prediction of protein-protein complexes is a well-developed field of study. However, because of the differences in physicochemical environment in the membranes and the spatial constraints of the membranes, the generic protein-protein docking approaches are not optimal for the membrane proteins. Thus, specialized computational methods for docking of the membrane proteins must be developed. Development and benchmarking of such methods requires high-quality datasets of membrane protein-protein complexes. In this study, we presented a new dataset of interacting alpha helical transmembrane protein segments extracted from 456 binary interactions. To reduce the ambiguity in the selection criteria, the redundancy in the dataset was removed at the structural rather than sequence level. The dataset is significantly larger and more representative than previously developed sets of transmembrane proteins.

In the future, this set will become the basis for the development of docking and scoring benchmarks, similar to the ones developed for soluble proteins in the DOCKGROUND resource. The sets will contain simulated unbound and modeled structures of the monomers (docking benchmark sets) and docking decoys (scoring benchmark sets) containing correct (near native) and incorrect predictions (decoys) for the development of scoring procedures.

Availability

The dataset is available online on the DOCKGROUND resource webpage:
http://dockground.compbio.ku.edu.

Figures

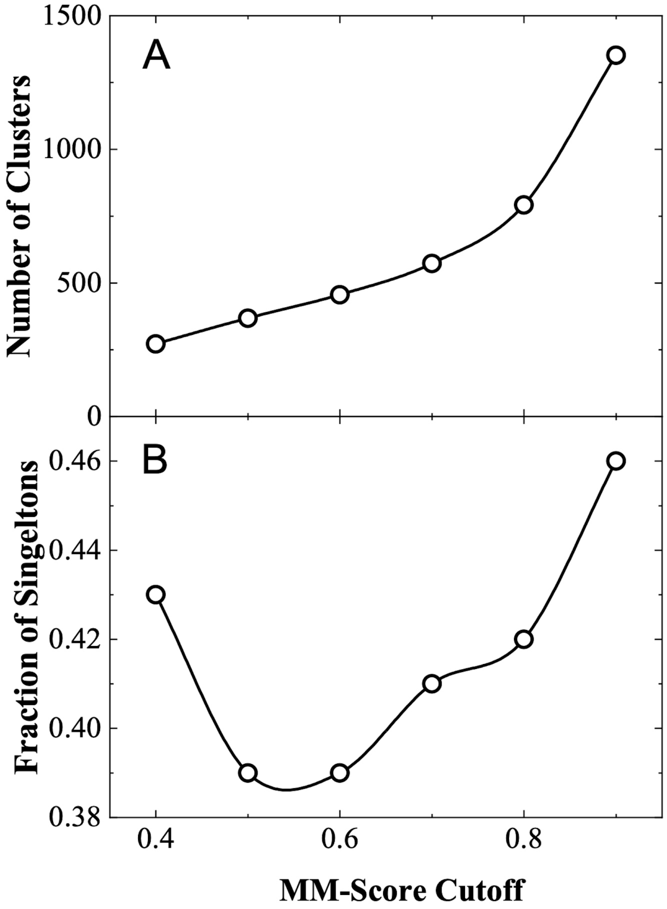


Figure 1. *Analysis of structure clusters at different clustering cutoffs.* The number of clusters produced at each clustering cutoff (A) grows monotonously, providing no clear indication of the optimal clustering cutoff. The frequency of single complex clusters has a distinct minimum, suggesting MM-score 0.6 as the optimal cutoff.

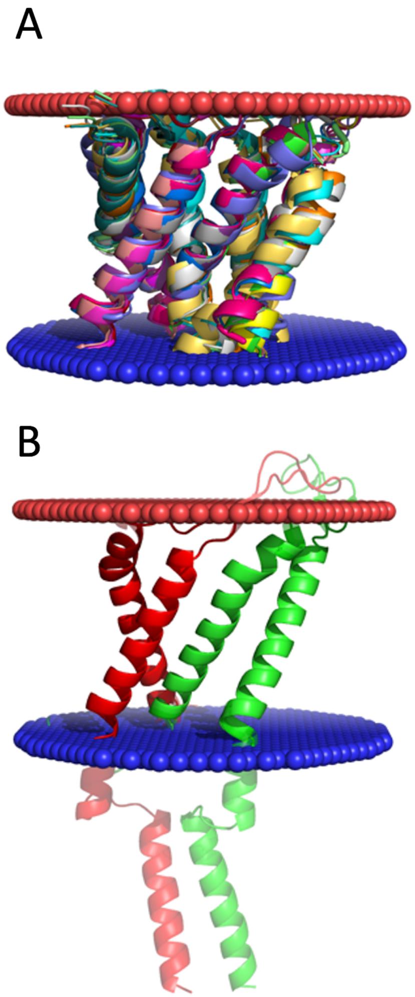


Figure 2*. Example of the structure cluster.* (A) The cluster consists of 15 binary interactions between transmembrane portions of the protein structures, each having two anti-parallel helices. All proteins in the cluster, from which these interfaces were extracted, were GO annotated as part of an ion channel. (B) The cluster's representative structure contains transmembrane segments of chains A (red) and B (green) from 2wcd with extramembrane parts (not included in the dataset) blurred for clarity. The extra- and intra-cellular sides of the membrane are in red and blue, respectively. The cluster has an average RMSD 1.64 Å. The most distant cluster members are chains A and B of PDB 6ctd and chains C and D of PDB 4y7k with RMSD 3.12 Å.

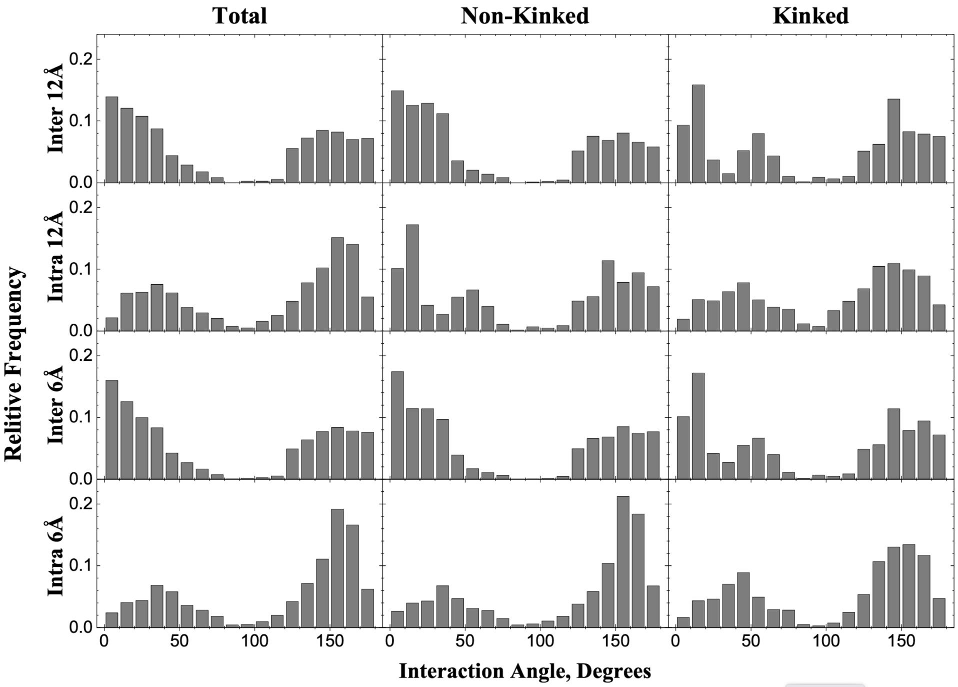


Figure 3. *Distribution of angles between interacting intra- and inter-chain helices.*

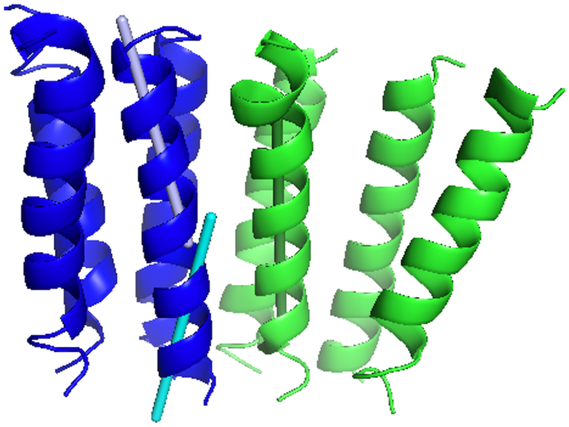
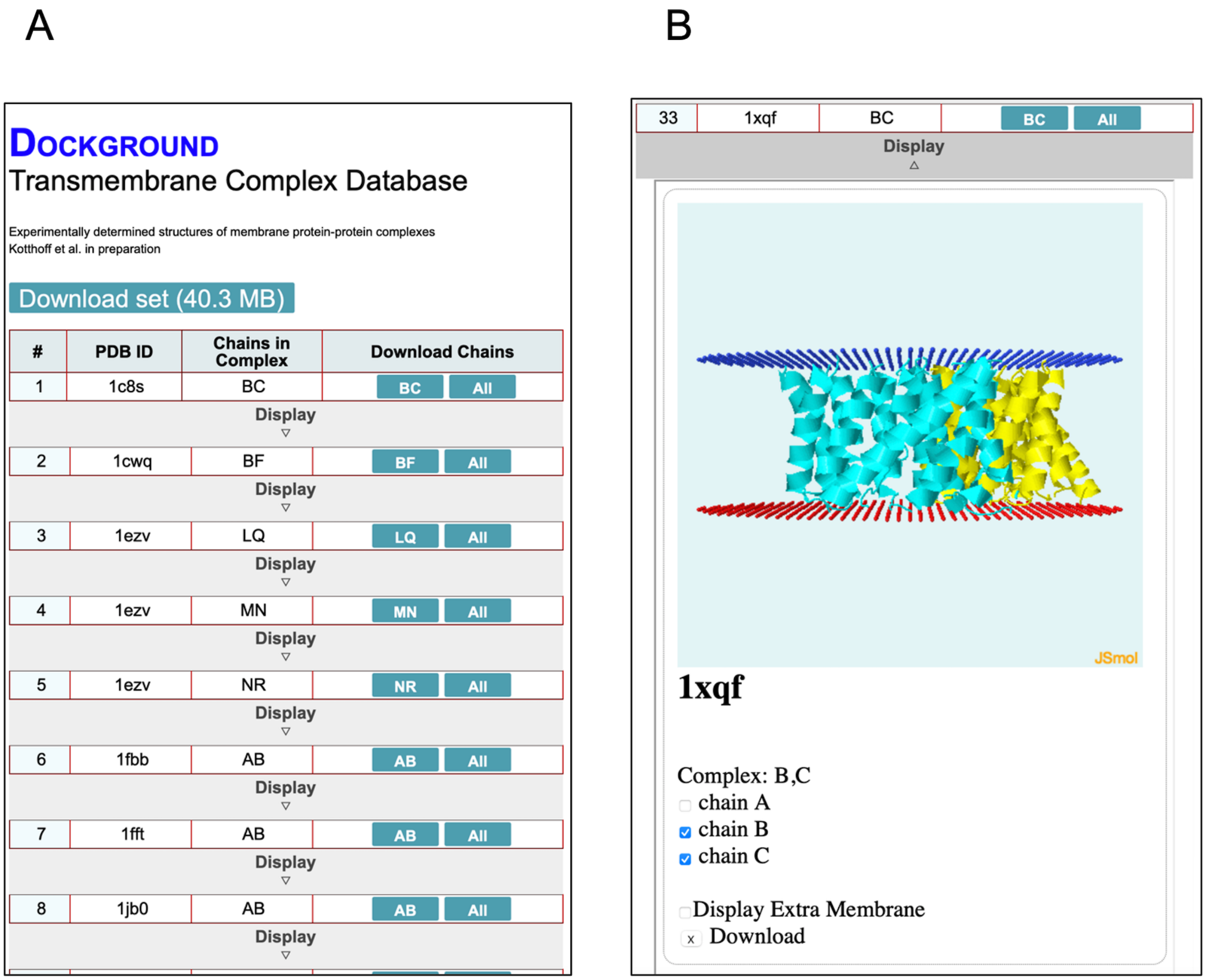


Figure 4. *Example of a kinked helix.* The transmembrane segments of chain A are in blue, and of chain B are in green (PDB structure 2xq4). Chain A contains a kinked helix with two direction vectors (gray and cyan) used separately in calculation of the interaction angles for that helix.



DOCKGROUND

Transmembrane Complex Database

Experimentally determined structures of membrane protein-protein complexes
Kotthoff et al. in preparation

Download set (40.3 MB)

#	PDB ID	Chains in Complex	Download Chains
1	1c8s	BC	BC All
		Display ▽	
2	1cwq	BF	BF All
		Display ▽	
3	1ezv	LQ	LQ All
		Display ▽	
4	1ezv	MN	MN All
		Display ▽	
5	1ezv	NR	NR All
		Display ▽	
6	1fbb	AB	AB All
		Display ▽	
7	1fft	AB	AB All
		Display ▽	
8	1jb0	AB	AB All
		Display ▽	

33	1xqf	BC	BC All
		Display △	

1xqf

Complex: B,C
☐ chain A
☑ chain B
☑ chain C

☐ Display Extra Membrane
x Download

Figure 5. *DOCKGROUND webpage for the membrane protein-protein set*. (A) The list of complexes for download as a whole or as individual complexes. (B) Visualization of a particular complex.

References

1. Muller MP, Jiang T, Sun C, Lihan M, Pant S, Mahinthichaichan P, et al. Characterization of lipid-protein interactions and lipid-mediated modulation of membrane protein function through molecular simulation. Chem Revews. 2019;119:6086-161.
2. Tsirigos KD, Govindarajan S, Bassot C, Vastermark A, Lamb J, Shu N, et al. Topology of membrane proteins - predictions, limitations and variations. Curr Opin Struct Biol. 2018;50:9-17.
3. Li F, Egea PF, Vecchio AJ, Asial I, Gupta M, Paulino J, et al. Highlighting membrane protein structure and function: A celebration of the Protein Data Bank. J Biol Chem. 2021;296:100557.
4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28:235-42.
5. Kaczor AA, Selent J, Sanz F, Pastor M. Modeling complexes of transmembrane proteins: Systematic analysis of protein-protein docking tools. Mol Inf. 2013;32:717-33.
6. Kozma D, Simon I, Tusnady GE. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucl Acid Res. 2013;41:D524-D9.
7. Vakser IA, Matar OG, Lam CF. A systematic study of low-resolution recognition in protein-protein complexes. Proc Natl Acad Sci USA. 1999;96:8477-82.
8. Vakser IA. Challenges in protein docking. Curr Opin Struct Biol. 2020;64:160-5.
9. Vakser IA. Long-distance potentials: An approach to the multiple-minima problem in ligand-receptor interaction. Protein Eng. 1996;9:37-41.
10. Hurwitz N, Schneidman-Duhovny D, Wolfson HJ. Memdock: An alpha-helical membrane protein docking algorithm. Bioinformatics. 2016;32:2444-50.
11. Koukos PI, Faro I, van Noort CW, Bonvin AMJJ. A membrane protein complex docking benchmark. J Mol Biol. 2018;430:5246-56.

12. Lomize AL, Pogozheva I, Mosberg HI. Large-scale computational analysis of protein arrangement in the lipid bilayer. Biophys J. 2011;100:492-.

13. Fairman JW, Noinaj N, Buchanan SK. The structural biology of beta-barrel membrane proteins: A summary of recent reports. Curr Opin Struct Biol. 2011;21:523-31.

14. Mitternacht S. FreeSASA: An open source C library for solvent accessible surface area calculations. F1000Research. 2016;5:189.

15. Mukherjee S, Zhang Y. MM-align: A quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucl Acids Res. 2009;37:e83.

16. Hartuv E, Shamir R. A clustering algorithm based on graph connectivity. Inform Process Lett. 2000;76:175-81.

17. Kirys T, Ruvinsky AM, Singla D, Tuzikov AV, Kundrotas PJ, Vakser IA. Simulated unbound structures for benchmarking of protein docking in the DOCKGROUND resource. BMC Bioinformatics. 2015;16:243.

18. Anishchenko I, Kundrotas PJ, Vakser IA. Modeling complexes of modeled proteins. Proteins. 2017;85:470–8. doi: 10.1002/prot.25183.

19. Singh A, Dauzhenka T, Kundrotas PJ, Sternberg MJE, Vakser IA. Application of docking methodologies to modeled proteins. Proteins. 2020;88:1180-8. doi: 10.1002/prot.25889.

20. Touw WG, Baakman C, Black J, te Beek TAH, Kriger E, Joosten RP, et al. A series of PDB-related databanks for everyday needs. Nucl Acids Res. 2015;43:D364-D8.

21. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577-637.

22. Sinha R, Kundrotas PJ, Vakser IA. Protein docking by the interface structure similarity: How much structure is needed? PloS One. 2012;7:e31349.