HHS Public Access

Author manuscript

Proteins. Author manuscript; available in PMC 2016 September 01.

Published in final edited form as:

Proteins. 2015 September ; 83(9): 1563–1570. doi:10.1002/prot.24736.

Structural templates for comparative protein docking

Ivan Anishchenko[1,2], **Petras J. Kundrotas**[1,*], **Alexander V. Tuzikov**[2], and **Ilya A. Vakser**[1,3,*]

[1]Center for Bioinformatics, The University of Kansas, Lawrence, Kansas 66047, USA

[2]United Institute of Informatics Problems, National Academy of Sciences, 220012 Minsk, Belarus

[3]Department of Molecular Biosciences, The University of Kansas, Lawrence, Kansas 66045, USA

Abstract

Structural characterization of protein-protein interactions is important for understanding life processes. Because of the inherent limitations of experimental techniques, such characterization requires computational approaches. Along with the traditional protein-protein docking (free search for a match between two proteins), comparative (template-based) modeling of protein-protein complexes has been gaining popularity. Its development puts an emphasis on full and partial structural similarity between the target protein monomers and the protein-protein complexes previously determined by experimental techniques (templates). The template-based docking relies on the quality and diversity of the template set. We present a carefully curated, non-redundant library of templates containing 4,950 full structures of binary complexes and 5,936 protein-protein interfaces extracted from the full structures at 12Å distance cut-off. Redundancy in the libraries was removed by clustering the PDB structures based on structural similarity. The value of the clustering threshold was determined from the analysis of the clusters and the docking performance on a benchmark set. High structural quality of the interfaces in the template and validation sets was achieved by automated procedures and manual curation. The library is included in the Dockground resource for molecular recognition studies at http://dockground.bioinformatics.ku.edu.

Keywords

protein recognition; protein modeling; benchmark sets; structure prediction; protein interactions

INTRODUCTION

Proteins often function by interacting with other proteins. Thus structural characterization of protein-protein interactions is important for understanding life processes. Due to the inherent limitations of experimental techniques, computational approaches are needed for such characterization. Following current paradigm and terminology in modeling of individual proteins, structural modeling of protein-protein complexes (docking) can be roughly divided into: (*i*) free docking, where sampling of the binding modes is performed with no regard to the possible existence of similar experimentally determined complex structures (templates),

*Corresponding authors: Ilya A. Vakser and Petras J. Kundrotas, Center for Bioinformatics, The University of Kansas, 2030 Becker Drive, Lawrence, Kansas 66047; tel: (785) 864-1057, fax: (785) 864-5558, vakser@ku.edu and pkundro@ku.edu.

and (*ii*) template-based docking, where such similar complexes determine docking predictions.

Free docking methods were initially developed as *ab initio* approaches based on the physical potentials (primarily, van der Waals interactions),[1] currently increasingly supplemented by the knowledge-based approaches (statistical potentials, docking constraints, etc.).[1,2] Despite their reasonable success, free docking methods have shown serious limitations, mostly due to the large size of the search space and structural flexibility upon the complex formation.

The template-based modeling of protein complexes relies on target/template relationships based on sequence,[3] sequence/structure (threading) and structure similarity,[1-7] with the latter showing a great promise in terms of availability of the templates.[8] The docking problem assumes *a priori* knowledge of the structures of the participating proteins. Thus, the docking templates may be found by structure (rather than sequence) alignment of the target monomers to the full structures of co-crystallized complexes. Evolutionary conserved surface patches may yield similar binding modes for otherwise dissimilar proteins[9,10] implying that docking can also be performed by the structure alignment of the target proteins with the interface parts of the co-crystallized complexes.

The key element in successful structure alignment application is the quality (diversity, non-redundancy and completeness of PDB structure) of the template libraries (generic or specific sets of 3D structures of binary complexes and/or their interfaces). Simply selecting all pairwise protein-protein complexes from PDB would produce the complete set of currently known structures. However, such “brute force” set will have many identical or highly similar complexes and some complex types will be overrepresented. The set will also contain erroneous, low-quality and biologically irrelevant structures.[11,12] Thus, groups working on structure alignment docking typically generate their own template libraries by filtering PDB in order to retain only the relevant interactions. A genome-wide study[13] utilized a library of ~30,000 full structures of template complexes extracted from PDB and PQS[14] databases with the intention (due to the termination of PQS) to switch to the PISA server.[15] The PRISM docking protocol,[16] where protein complexes are modeled by structure alignment of the interface regions, used a library of 8,205 protein-protein interfaces that represent unique interface architectures.[17] Classification of interfaces into biological and those due to crystal packing, obligatory and nonobligatory was done by NOXclass procedure[18] and structural comparison of the initial 49,512 interfaces was performed by the geometric hashing with subsequent hierarchical clustering. A more recent study by the same group introduces a new library of 22,605 entries which is suggested for interface-based structure alignment docking.[19] The interfaces in this set were extracted from the full structures using effective distance cut-off ~ 10 Å, while an earlier study[20] indicated that the maximum success rate in interface structure alignment docking is achieved when template interfaces are extracted with a larger, 12 Å distance cut-off.

In this paper we describe our most recent sophisticated set of templates that addresses many drawbacks of the existing sets. We extract the interfaces at the optimal distance cut-off and cluster full structures and interfaces separately using various thresholds for structural similarity. Resulting datasets of full structures and interfaces, available in the Dockground

resource http://dockground.bioinformatics.ku.edu, were generated using clustering threshold determined by the performance of the docking protocols.

METHODS

The methods used in this study involved procedures for the structure quality control of protein-protein interfaces, clustering of the complexes and interfaces, and optimization of the clustering parameters based on the performance of the template libraries in the docking runs.

Chain inter-penetration

Complexes from the initial set (see Results) were checked for inter-penetration of chains by an automated procedure. For each residue of a protein in a complex, all atoms of the other protein within 6Å distance were selected. An imaginary line through C and N backbone atoms of the residue and two half planes joined by this line were drawn. By rotating the half planes around this axis, the maximal sector free of atoms of the second protein was determined. If the corresponding angle between the planes was < 90°, the residue was considered buried. Complexes with two consecutive buried residues in any of the chains were excluded from the set.

Clustering of complexes and interfaces

Pairwise structural alignment of all complexes and interfaces in the initial set (see Results) was performed by MM-align[21] (an offshoot of the TM-align program[22] specifically designed for comparison of protein complex structures) with TM-score[23] normalized by the length of the larger complex. The TM-scores were further used to construct an undirected graph, with nodes representing individual complexes (interfaces) and edges reflecting their similarity. Two vertices in the graph were connected by an edge if the corresponding TM-score was not less than a specified threshold value TM_T, which varied in the course of computations. The resulting graph was split into clusters by a two-stage procedure. At the first stage, the graph was divided into connected components by the breadth-first search algorithm.[24] The minimum cut in a graph G was defined as the minimum number of edges $k(G)$, which needs to be removed to disconnect the graph into two (connected) components. A component with n nodes ($n > 1$) was considered highly connected if the condition $k(G) > n/2$ is satisfied.[25] The basic clustering algorithm by Hartuv and Shamir[25] uses the Stoer-Wagner *mincut* algorithm[26] to iteratively split a graph into subgraphs until they become highly connected. These highly connected subgraphs are induced subgraphs of the original graph and represent clusters. Several heuristics[25] were also applied to speed-up computations and to enhance the quality of clusters by adopting nodes, which became separated after the direct application of the basic, non-enhanced approach by Hartuv and Shamir.[25] The enhanced algorithm was applied to every connected component, which did not represent a complete graph. The clustering procedure was implemented as a standalone C++ program; the *igraph* library (http://igraph.sourceforge.net/index.html) was used to handle operations on the graphs.

Validation set of protein-protein complexes

Docking performance on a validation set of complexes was used to determine the clustering threshold. To generate the validation set, the initial list of structures was taken from Dockground at 30% sequence identity cut-off. We selected only moderate- and high-resolution X-ray structures (resolution ≤ 3.5 Å) with a well-defined interface (mean accessible surface area buried by each chain ≥ 250 $Å^2$, and ≥ 10 residues at an interface in each chain). Complexes with a protein containing < 3 secondary structure elements were excluded from consideration, as well as complexes with monomers of substantially different size, where one protein is three or more times lager than the other (according to the number of residues). Finally, the set was visually inspected to clean out coiled-coil complexes (to decrease the modeling noise, since the alignment of any helix in a target to such a template has high TM-score) and complexes with interwoven chains.

Docking protocol

We used the template-based docking protocol similar to the one developed previously in our lab.[20,27] The procedure performs spatial rearrangement of 3D structures of two target proteins (treated as rigid bodies) to match either the entire monomers of the co-crystallized complexes (from the full-structure template library) or their interfaces only (from the interface template library). Structural alignment of proteins was performed by TM-align.[22] The resulting pool of putative matches was filtered to retain only significant matches with TM-scores of both alignments > 0.4. Models were scored by the average TM-score of both alignments. When the docking protocol was run in the benchmarking mode, the self-matches were avoided by excluding templates with both TM-scores > 0.9. Assessment of resulting models was done in terms of C^{α} ligand RMSD with receptors optimally superimposed. This RMSD definition was chosen, as opposed to the slightly different one used in CAPRI[28] (superimposition of the native interface residues in the native and the modeled complexes), for consistency with our previous studies.[8,12]

RESULTS AND DISCUSSIONS

Initial set of structures

We built two separate libraries, one consisting of the full two-chain structures and the other of the interface fragments. The flowchart of the generation process is in Figure 1. The initial pool of the X-ray structures with resolution ≤ 3.5 Å and buried interface area ≥ 250 $Å^2$ per chain was extracted for both libraries from the Dockground co-crystallized protein-protein complexes.[11] We imposed an additional constraint that interfaces should consist of at least ten residues in each chain. At the point of computation, the Dockground version was based on December 2012 PDB release. Protein complexes in Dockground are derived from the PDB Biological Unit files. Thus our set likely consists of biologically functional complexes, although some false positives are inevitable.[12] Each complex was further checked for inter-penetration of chains by an automated procedure developed for this task (see Methods) and complexes like the one shown in Figure 2 were removed (284 entries). This resulted in 12,134 structurally redundant complexes. Interfaces were extracted from these complexes using 12Å distance cut-off between heavy atoms of residues belonging to different chains. The extracted interfaces were clustered and analyzed in terms of structural connectivity and

docking performance in order to choose the clustering parameters. Full structures were further filtered by an additional requirement that at least three regular (> 4 residues each) secondary structure elements be present in each interacting protein. The secondary structure elements (α-helices and/or β-strands) were detected by the DSSP tool.[29] The resulting reduced set of 11,774 complexes was subjected to the clustering and analysis procedures, same as the interfaces.

Connectivity of the structural space of protein-protein complexes

To eliminate structural redundancy, the intermediate sets of 11,774 complexes and 12,134 interfaces had to be clustered by some measure of structural similarity. In this study, for such a measure we used TM-score.[23] TM-score (ranging from 0 to 1) is produced by the TM-align routine,[22] which was previously successfully employed in the template-based docking,[8,12,20,27,30] although other programs for the structural alignment with their own structural similarity scores were utilized by others.[31,32]

For efficient clustering, it is useful to understand how similar complexes and interfaces are connected in the structural space. We analyzed similarity graphs built at different threshold values of TM-score (TM_T, see Methods) in terms of the size of the connected components (initial, first-approximation clusters with some missing edges between the nodes) and the clustering coefficient (the probability of neighbors of a given node to be connected between themselves[33]). As seen in the main panels of Figure 3, a substantial fraction of connected components belongs to either isolated nodes (53 – 68 % of complexes and 56 – 72 % of interfaces, depending on TM_T) or pairs of connected nodes (14 – 16 % of complexes and 13 – 16 % of interfaces) and cannot be split further. Interestingly, this property is persistent within a broad TM_T range.

In terms of clusters, the clustering coefficient can be viewed as a measure of the extent to which the groups of connected nodes in a graph are close to the complete graphs (or ideal clusters), in which every pair of nodes is connected by an edge. Inserts in Figure 3 show the clustering coefficient of graphs for full structures (panel A) and interfaces (panel B) in the full TM_T range from 0.0 to 1.0. Due to the random matches of short structural fragments, TM-score seldom gets very close to 0. Thus, at low TM_T, there are edges in the graph between almost all nodes making the graph close to complete and resulting in high clustering coefficient of almost 1. The similarity graphs then will be close to complete graphs comprising almost entire set of complexes/interfaces (left sides of inserts in Figure 3). When TM_T increases, the clustering coefficient decreases dramatically and has a minimum at $TM_T = 0.27$ for the full structures and $TM_T = 0.28$ for the interfaces, which is consistent with a previous estimate of the average TM-score for random match 0.17.[23] With further increase of TM_T, the statistical significance of a structural match increases as well. Starting from $TM_T \sim 0.5$ (the lowest TM-score for proteins with similar folds[22,34]), the clustering coefficient stops growing and remains unchanged (~0.98 for both full structures and interfaces) up to $TM_T \sim 0.9$ for full complexes and $TM_T \sim 0.8$ for interfaces. High values of the clustering coefficient within such TM_T ranges suggest that the graph nodes are clustered in almost optimal way. The decrease in the clustering coefficient for $TM_T > 0.9$ stems from small structural differences (especially in the loops) often present in different

PDB files for otherwise identical or very similar (in terms of sequences) proteins. Due to the smaller size of interfaces, TM-score between pair of interfaces, on average, is smaller than TM-score between corresponding pair of the full structures (Figure 4) and thus the clustering coefficient for the interfaces starts to drop closer to $TM_T \sim 0.8$.

Importantly, even at the highest value of $TM_M = 1.0$, 991 (8%) complexes and 594 (5%) interfaces are removed from the corresponding libraries. This shows that a blind selection of all pairwise complexes from PDB would result in a library with a considerable number of identical entries.

Analysis of clusters

We utilized a clustering approach, which, first, divides the similarity graph into "loosely" connected components and then further splits them into tightly connected clusters (see Methods). Figure 5 shows how the number of connected components N_{CC} and the number of resulting clusters N_C varies with TM_T. Since the majority of the connected components cannot be split further (as shown in Figure 3), N_C is only slightly larger than N_{CC} for most values of TM_T. The relative increase in the number of clusters is ~ 5% for $TM_T = 0.6 – 0.9$ for both full complexes and interfaces (green lines in Figure 5). This correlates with the high values of clustering coefficient in these TM_T ranges (Figure 3).

Finally, we checked the quality of the resulting clusters at different TM_T by calculating TM-scores between members of each cluster in order to detect pairs of nodes within a cluster that lack an edge (TM-score < TM_T). We found that only ~ 3% of the final clusters have pairs of dissimilar complexes (or interfaces) with TM-score < TM_T (circles in Figure 6). The clustering algorithm we employed allows the final clusters to have as little as 50% of edges (compared to complete graphs).[25] However, the analysis of the actual clusters showed that the fraction of dissimilar pairs in the vast majority of clusters is < 30%, with the mean value close to 10% (box-and-whiskers plots in Figure 6). The quality of clusters remains roughly the same within the full TM_T range, with minor variations at $TM_T < 0.6$ for full complexes (Figure 6A) and $TM_T > 0.9$ for both complexes and interfaces (Figures 6 A and B). Sequence identity, in general, follows the same trend, i.e. ~70% of the clusters have the minimal sequence identity between the members > 90%. However, in some extreme cases (e.g., the cluster of 49 complexes from RNA polymerase), there are cluster members with sequence identities ≤ 30%.

Template libraries in docking: selecting optimal parameters

The success of docking depends heavily on the diversity of the template library. On the other hand, the running time of the template-based docking is directly proportional to the size of the template set. Thus, an optimal template library should be large enough to maximize the docking success rate, but should not contain excessive entries, which only marginally improve the performance. This approach to the optimal library is different from the one used to compile PRISM interface library,[19] where optimization was performed according to the quality of the resulting clusters. For practical docking purposes, our choice was rather to optimize the performance of the modeling of complexes based on our templates, through benchmarking.

We tested 26 full-structure and 26 interface libraries, generated at TM_T ranging from 0.50 to 1.00 with 0.02 step, on a non-redundant set of 293 hetero complexes (see Methods). Success rate was defined as a ratio of targets, for which at least one model had interface C^{α} ligand RMSD < 5 Å, to the total number of targets.[8,12] To exclude the influence of the scoring scheme, we calculated success rate for the entire set of models, although results for the top ten models, ranked by the average TM-score, were also obtained (not shown separately due to qualitative similarity to the all-models results). Templates that were similar to a particular target (both TM-scores for a target-template pair exceed 0.9) were left out from the consideration. Such exclusion of similar structures leads to success rates higher than reported in a recent benchmark study[35] where the main focus was on docking in the "twilight zone" of low target/template similarity (sequence identities between target and template < 30%).

Results of the test are shown in Figure 7. As one can expect, more entries in the template library (higher TM_T values) lead to higher success rates of the docking. Such monotonic behavior holds for almost entire TM_T range from 0.5 to ~0.9. For $TM_T > 0.9$ the success rate is largely saturated. A slight increase in the success rates at $TM_T > 0.9$ is an artifact of our procedure. While TM-scores used in the clustering are obtained by the MM-align for complexes, TM-scores for exclusion were produced by TM-align for separate monomers. At certain TM_T (especially at values close to the similarity criteria), a cluster may have a representative, identified as the similar structure (thus excluding entire cluster from consideration) and other structures with one out of two TM-scores of TM-align slightly less than the similarity criteria (this could have TM-score of the MM-align exceeding TM_T). These structures at higher TM_T can split into a separate cluster with representative identified as non-similar and thus yielding good-quality model (in total, there are seven such cases in the full-structure set and one case in the interface set).

The differences in the success rates of the full structure and the interface-based alignments (Figure 7) can be explained by different TM-score values for the full structures and the corresponding interfaces. In docking, the templates with both TM-scores > 0.9 were excluded from consideration. In full structure-based docking, the TM-score was calculated based on the alignment of two full proteins, whereas in interface-based docking the TM-score was obtained by aligning target proteins with the template interfaces. According to Figure 4, the latter TM-score should be generally lower than the former. Thus, some templates, excluded as self-matches in full structure-based docking (both TM-scores > 0.9) still represented suitable interface templates.

The number of clusters (and, consequently, computational time) starts growing exponentially at $TM_T > 0.9$ (Figure 5). This, along with the results in Figure 7, suggests that the optimal library ought to be generated at $TM_T = 0.9$.

Availability of the template and the benchmark sets

A representative complex from each cluster at $TM_T = 0.9$ was selected based on the best resolution and the smallest number of missing residues. The resulting sets of 4,950 full-structure complexes and 5,936 interfaces (representing ~40% of folds in SCOP[36] and in a more recent ECOD database, http://prodata.swmed.edu/ecod) are available on the Web

within our Dockground resource at http://dockground.bioinformatics.ku.edu, under “docking templates” tab. The sets are downloadable as zip archives (one for full structures and the other for the interfaces) each containing folders “templates,” “targets” and “info.” The folder “templates” contains two PDB-formatted files of atomic coordinates per library entry. The files are named by the original PDB file, from which the entry was extracted, as follows:

$$[XXXX][M_1][CH_1][M_2][CH_2]\_N,$$

where [*XXXX*] is the 4-symbol PDB code, [M_1] and [M_2] are the model numbers, [CH_2] and [CH_2] are the chain identifiers for the first and the second component of the library entry, and N = 1 or 2 identifies the component. Separation of library entries into two files makes it easier to use the set in the docking programs. However, simple joining of the two files (e.g., with *cat* command in Linux) will produce the complex (interface) structure without geometrical clashes and distinct chain identifiers. The folder “targets” in both full-structure and interface archives consists of 2×293 similarly named PDB-formatted files for the full structures of validation set used in this study. The folder “info” contains two text files per structure in the validation set (named as the files in the “target” folder, but with the extension .txt) with information on all meaningful structural alignments (TM-scores > 0.4) of the target files to full-structures or interfaces of the template set. The folder also contains a text file with information on the resulting models. In validation, some target complexes (64 for full structure and 33 for interface templates) had at least one model with interface RMSD > 5 Å and the minimal of the TM-scores of the components > 0.8, indicating high similarity to a wrong template. The “difficult_targets.txt” files in the “info” folders contain the list of such targets.

The sets can be used either for modeling of unknown protein complexes of interest by full or interface structural alignment (using only structures in the “templates” folder) or for benchmarking of new modeling techniques (using both “target” and “template” folders and comparing results with the data in the “info folder”).

Acknowledgments

This study was supported by NIH grant R01GM074255 and NSF grant DBI1262621. Calculations were conducted in part on ITTC computer cluster at The University of Kansas.

References

1. Vakser IA. Protein-protein docking: From interaction to interactome. Biophys J. 2014; 107:1785–1793. [PubMed: 25418159]
2. Vakser IA. Low-resolution structural modeling of protein interactome. Curr Opin Struct Biol. 2013; 23:198–205. [PubMed: 23294579]
3. Aloy P, Pichaud M, Russell RB. Protein complexes: Structure prediction challenges for the 21st century. Curr Opin Struct Biol. 2005; 15:15–22. [PubMed: 15718128]
4. Szilagyi A, Zhang Y. Template-based structure modeling of protein–protein interactions. Curr Opin Struct Biol. 2014; 24:10–23. [PubMed: 24721449]
5. Dey F, Zhang QC, Petrey D, Honig B. Toward a “structural BLAST”: Using structural relationships to infer function. Protein Sci. 2013; 22:359–366. [PubMed: 23349097]

6. Kuzu G, Keskin O, Gursoy A, Nussinov R. Constructing structural networks of signaling pathways on the proteome scale. Curr Opin Struct Biol. 2012; 22:367–377. [PubMed: 22575757]
7. Szilagyi A, Grimm V, Arakaki AK, Skolnick J. Prediction of physical protein-protein interactions. Phys Biol. 2005; 2:S1–S16. [PubMed: 16204844]
8. Kundrotas PJ, Zhu Z, Janin J, Vakser IA. Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci USA. 2012; 109:9438–9441. [PubMed: 22645367]
9. Keskin O, Nussinov R. Favorable scaffolds: Proteins with different sequence, structure and function may associate in similar ways. Protein Eng. 2005; 18:11–24.
10. Zhang QC, Petrey D, Norel R, Honig BH. Protein interface conservation across structure space. Proc Natl Acad Sci USA. 2010; 107:10896–10901. [PubMed: 20534496]
11. Douguet D, Chen HC, Tovchigrechko A, Vakser IA. DOCKGROUND resource for studying protein-protein interfaces. Bioinformatics. 2006; 22:2612–2618. [PubMed: 16928732]
12. Kundrotas PJ, Vakser IA, Janin J. Structural templates for modeling homodimers. Protein Sci. 2013; 22:1655–1663. [PubMed: 23996787]
13. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature. 2012; 490:556–560. [PubMed: 23023127]
14. Henrick K, Thornton JM. PQS: A protein quaternary structure file server. Trends Biochem Sci. 1998; 23:358–361. [PubMed: 9787643]
15. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol. 2007; 372:774–797. [PubMed: 17681537]
16. Tuncbag N, Gursoy A, Nussinov R, Keskin O. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nat Protoc. 2011; 6:1341–1354. [PubMed: 21886100]
17. Tuncbag N, Gursoy A, Guney E, Nussinov R, Keskin O. Architectures and functional coverage of protein–protein interfaces. J Mol Biol. 2008; 381:785–802. [PubMed: 18620705]
18. Zhu H, Domingues FS, Sommer I, Lengauer T. NOXclass: Prediction of protein-protein interaction types. BMC Bioinformatics. 2006; 7:27. [PubMed: 16423290]
19. Cukuroglu E, Gursoy A, Nussinov R, Keskin O. Non-redundant unique interface structures as templates for modeling protein interactions. PloS One. 2014; 9:e86738. [PubMed: 24475173]
20. Sinha R, Kundrotas PJ, Vakser IA. Protein docking by the interface structure similarity: How much structure is needed? PloS One. 2012; 7:e31349. [PubMed: 22348074]
21. Mukherjee S, Zhang Y. MM-align: A quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucleic Acids Res. 2009; 37:e83. [PubMed: 19443443]
22. Zhang Y, Skolnick J. TM-align: A protein structure alignment algorithm based on the TM-score. Nucl Acid Res. 2005; 33:2302–2309.
23. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004; 57:702–710. [PubMed: 15476259]
24. Cormen, TH.; Leiserson, CE.; Rivest, RL.; Stein, C. Introduction to Algorithms. The MIT Press; 2009. p. 1312
25. Hartuv E, Shamir R. A clustering algorithm based on graph connectivity. Inform Process Lett. 2000; 76:175–181.
26. Stoer M, Wagner F. A simple min-cut algorithm. J Acm. 1997; 44:585–591.
27. Sinha R, Kundrotas PJ, Vakser IA. Docking by structural similarity at protein-protein interfaces. Proteins. 2010; 78:3235–3241. [PubMed: 20715056]
28. Lensink MF, Wodak SJ. Docking, scoring, and affinity prediction in CAPRI. Proteins. 2013; 81:2082–2095. [PubMed: 24115211]
29. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–2637. [PubMed: 6667333]
30. Kundrotas PJ, Vakser IA. Global and local structural similarity in protein-protein complexes: Implications for template-based docking. Proteins. 2013; 81:2137–2142. [PubMed: 23946125]

31. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A. PRISM: Protein interactions by structural matching. Nucleic Acids Res. 2005; 33:W331–W336. [PubMed: 15991339]
32. Petrey D, Honig B. GRASP2: Visualization, surface properties, and electrostatics of macromolecular structures and sequences. Methods Enzymol. 2003; 374:492–509. [PubMed: 14696386]
33. Watts DJ, Strogatz SH. Collective dynamics of ‘small-world’ networks. Nature. 1998; 393:440–442. [PubMed: 9623998]
34. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics. 2010; 26:889–895. [PubMed: 20164152]
35. Negroni J, Mosca R, Aloy P. Assessing the applicability of template-based protein docking in the twilight zone. Structure. 2014; 22:1356–1362. [PubMed: 25156427]
36. 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; 247:536–540. [PubMed: 7723011]

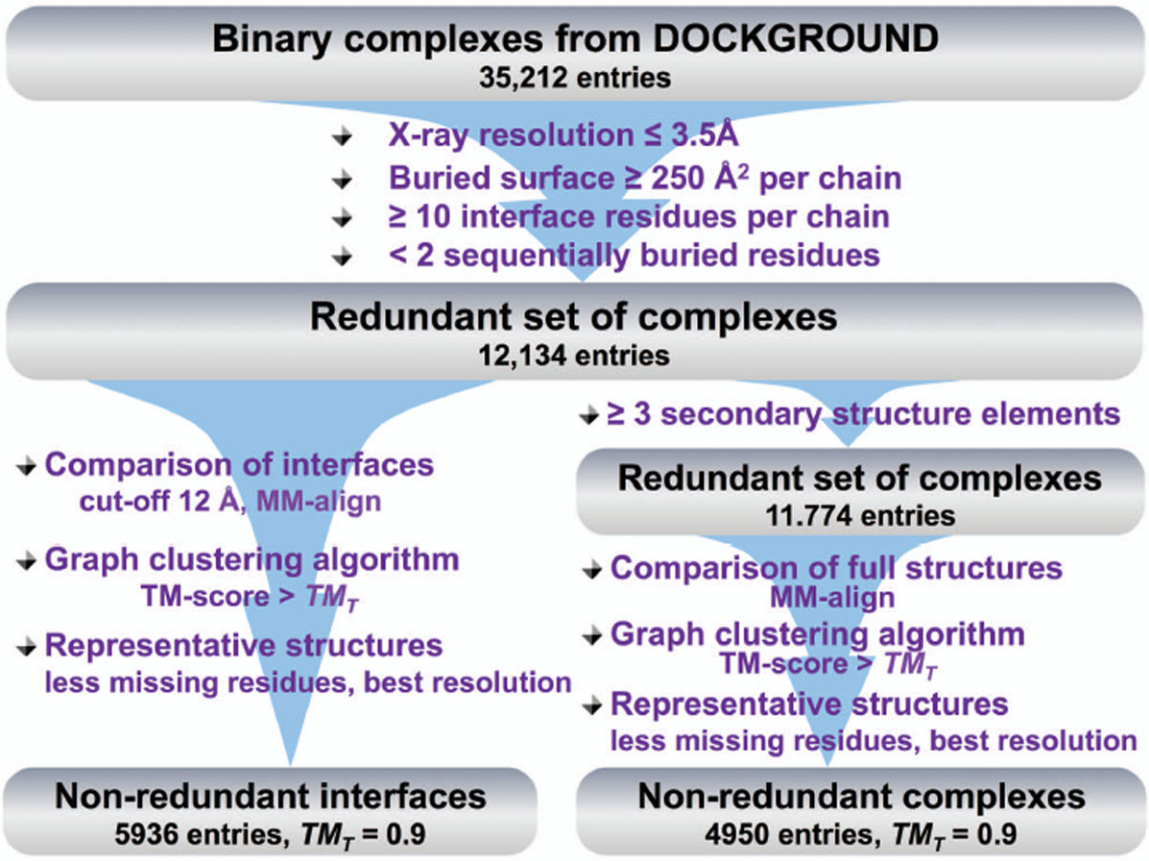

Figure 1.
Flowchart of algorithm for generation of full-structure and interface template libraries.

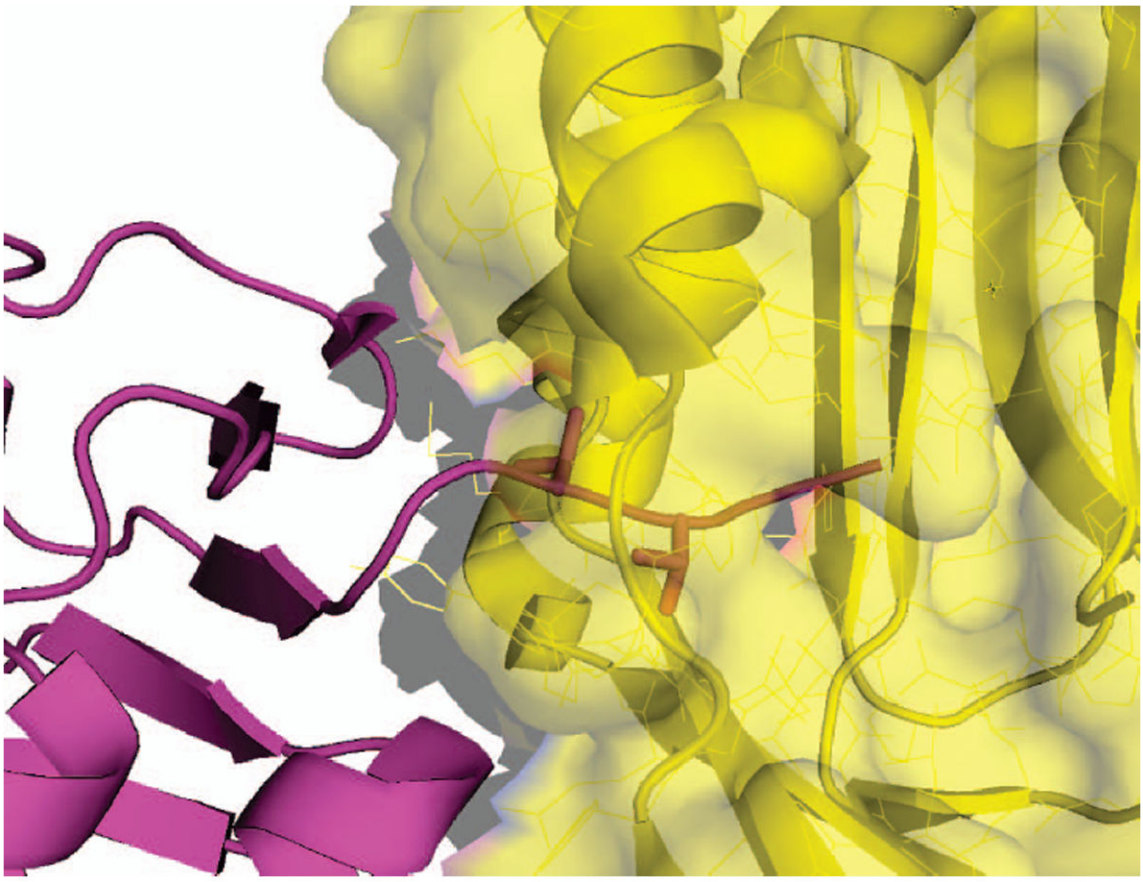

Figure 2.
Example of a “bad” complex. Chains D and E from 1fma are shown in magenta and yellow, respectively, with penetrating chain removed by the automated procedure described in the text. Buried Val and Thr residues at the C-terminal of chain F identified by the procedure are shown as sticks (the last two residues at the terminus are Gly).

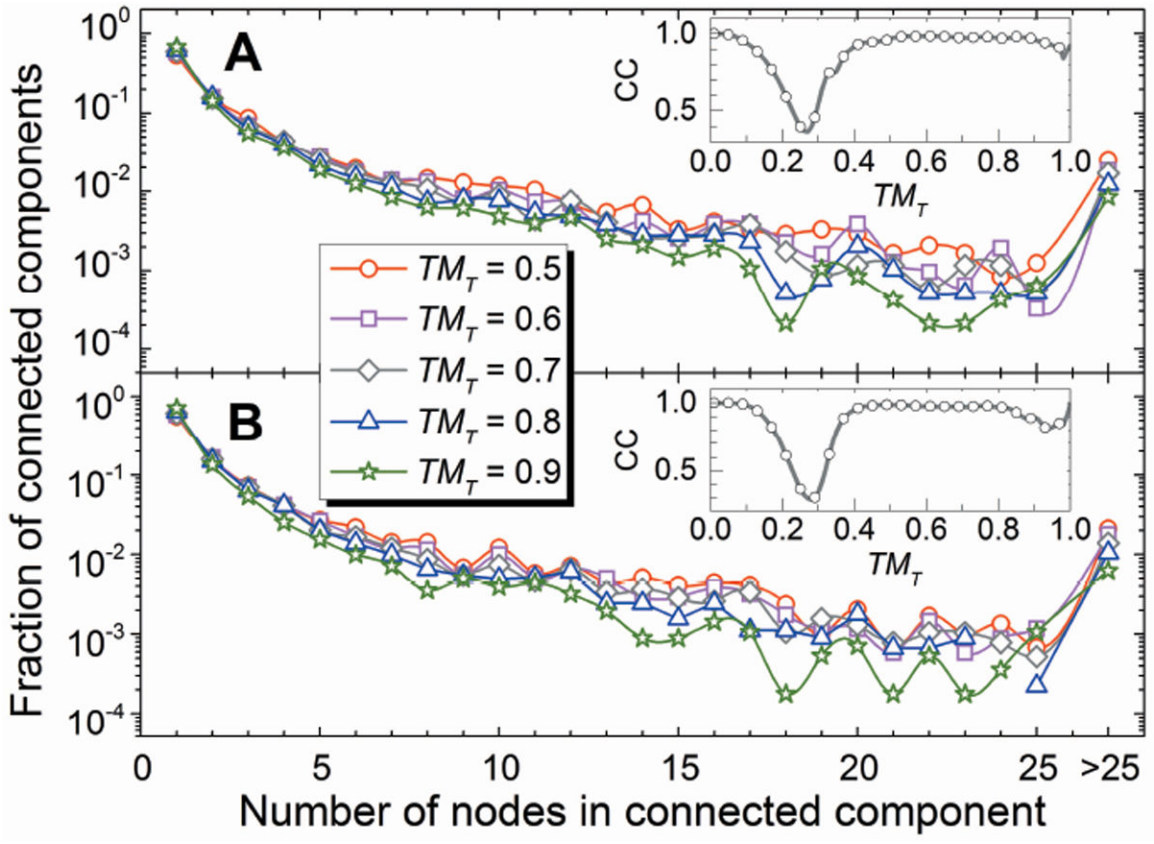

Figure 3.

Properties of similarity graphs. (A) Protein-protein complexes and (B) protein-protein interfaces. The main panels show distributions of connected component size at different thresholds of the clustering TM-score (TM_T). The inserts display dependence of clustering coefficient CC on TM_T.

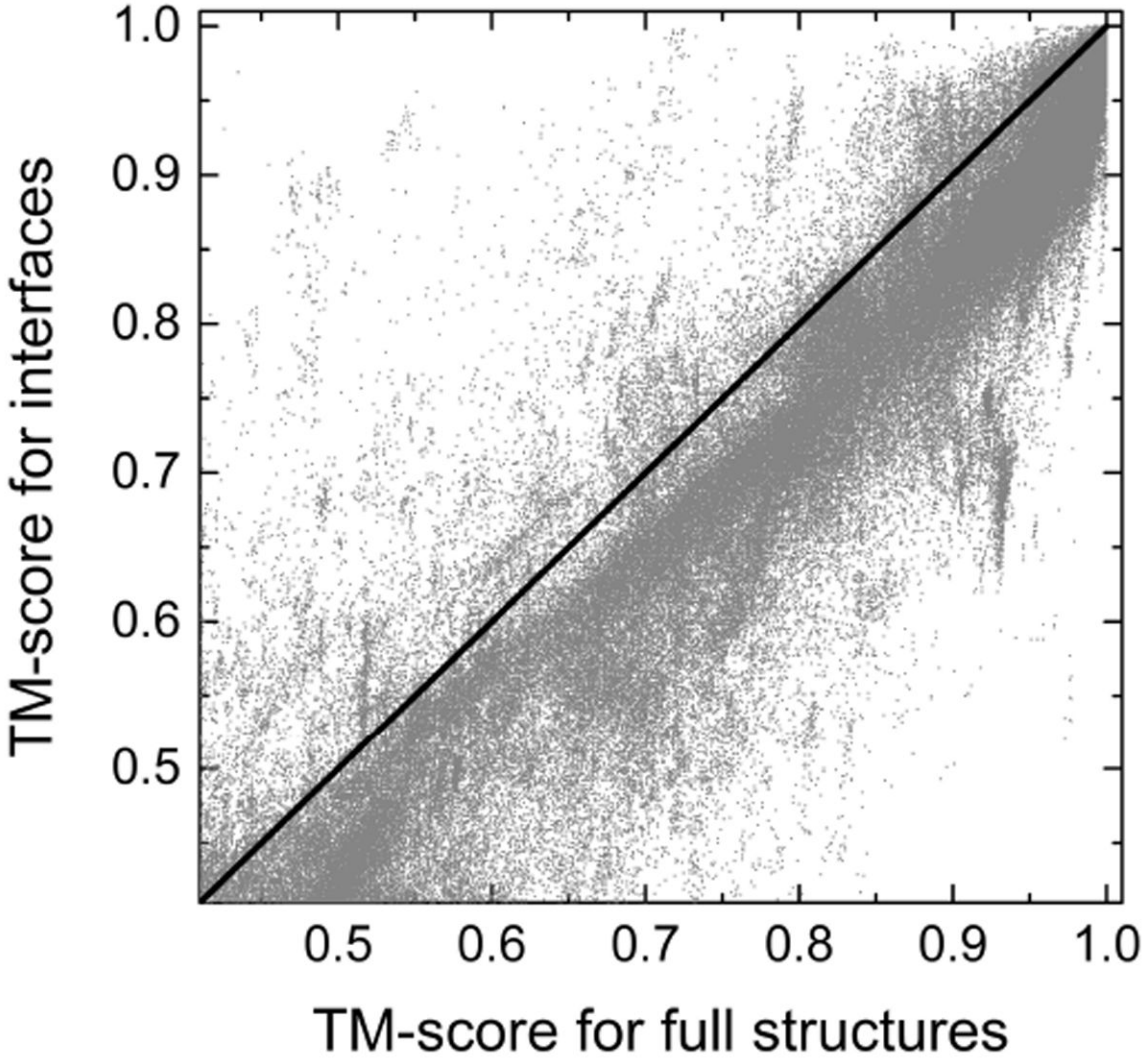

Figure 4.
Correlation of protein-protein and interface-interface TM-scores.

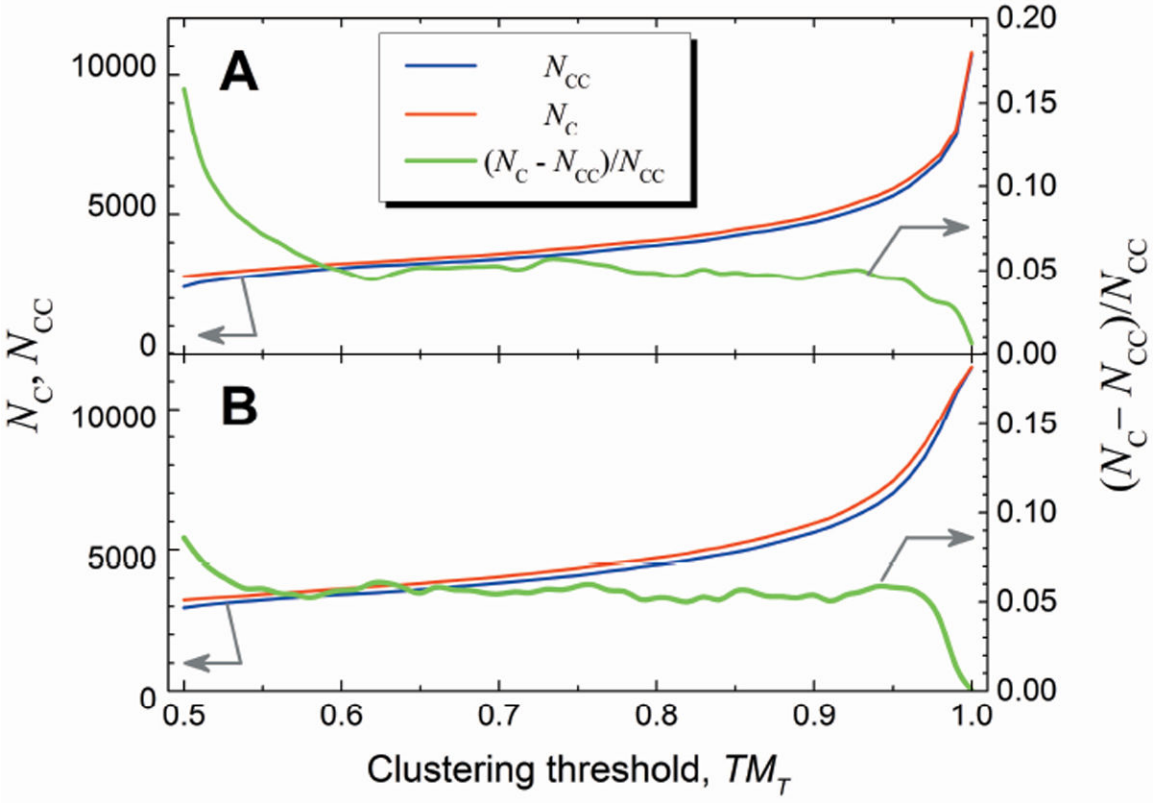

Figure 5.

Number of connected components and clusters as a function of clustering threshold. (A) Protein-protein complexes, and (B) protein-protein interfaces. N_{CC} is the number of connected components, and N_C is the number of clusters. Green lines (scaled to the right-hand axes) show the relative increase in the number of connected graph parts after splitting the connected components into tightly connected clusters.

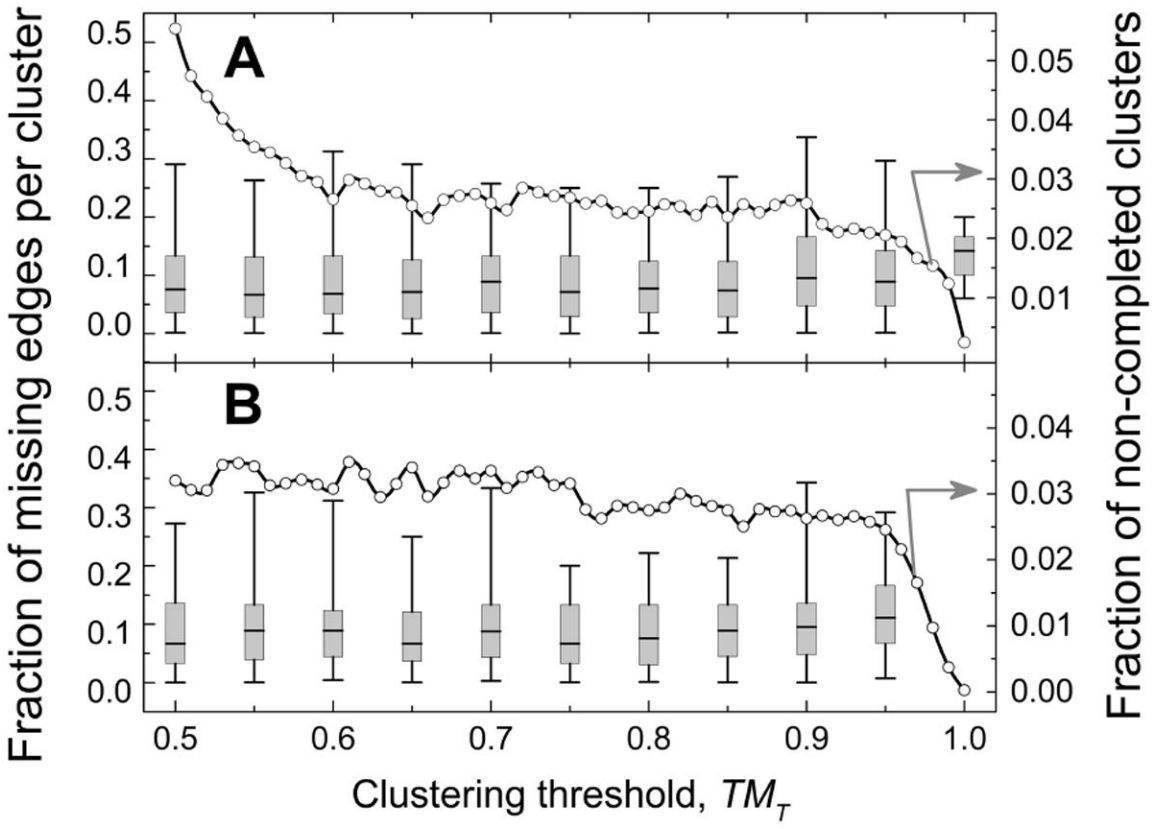

Figure 6.

Quality of clusters at different clustering thresholds. (A) Protein-protein complexes, and (B) protein-protein interfaces. Distributions of missing edges per cluster are shown as box-and-whickers plots with horizontal lines for minimal, maximal and median values in the distributions and boxes containing second and third quartiles of data. The circles (scaled to the right hand axis) show how the fraction of clusters, which are not complete sub-graphs of the initial similarity graphs, depends on TM_T.

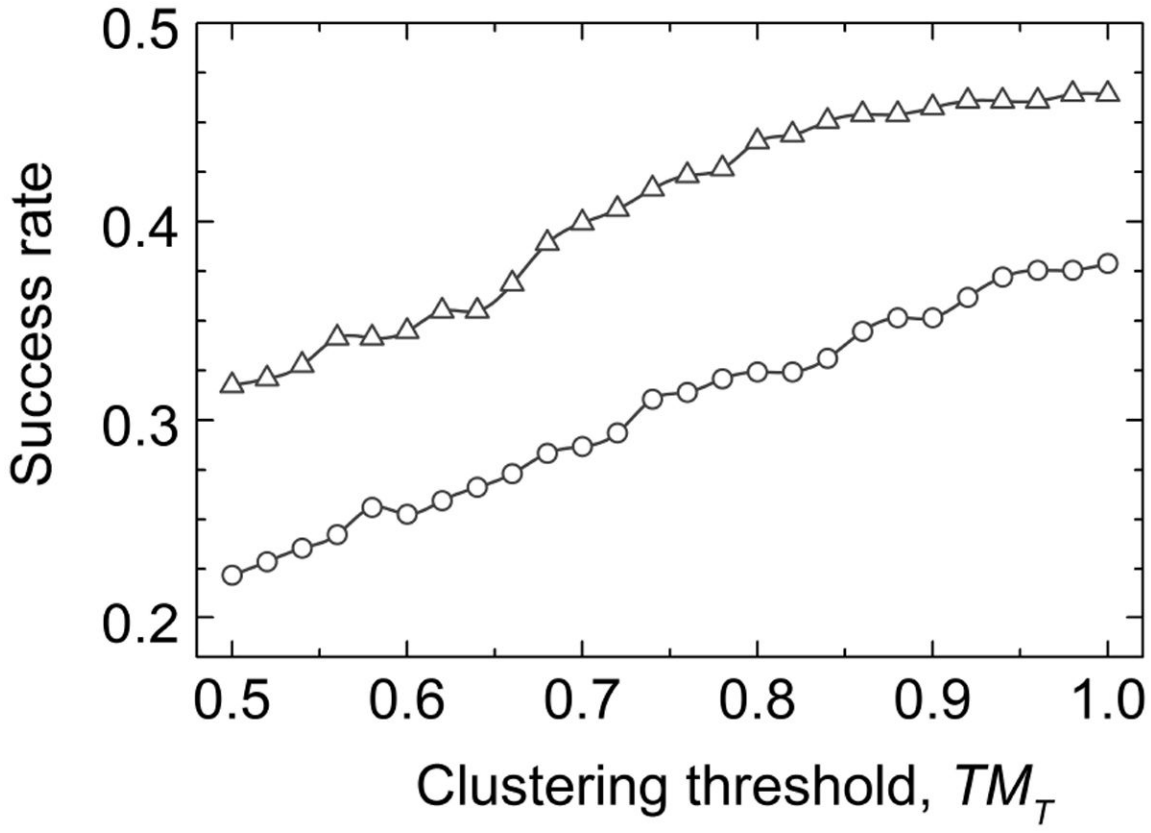

Figure 7.
Performance of structure alignment at different clustering thresholds. Full-structure (circles) and interface (triangles) libraries were generated at different threshold values. Success rates were calculated for the entire pool of structures excluding templates similar to the target (see text).