MODELING PROTEIN INTERACTIONS THROUGH STRUCTURE ALIGNMENT

By

ROHITA SINHA

Submitted to the graduate degree program in Bioinformatics
and the Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctorate of Philosophy

_____
Chairperson Ilya A. Vakser, PhD

_____
Krzysztof Kuczera, PhD

_____
Eric Deeds, PhD

_____
Kyle Camarda, PhD

_____
Mark Richter, PhD

_____
Gerald Henry Lushington, PhD

Date Defended:

The Dissertation Committee for Rohita Sinha
certifies that this is the approved version of the following thesis:

MODELING PROTEIN INTERACTIONS THROUGH STRUCTURE ALIGNMENT

_____

Chairperson Ilya A. Vakser, PhD

Date approved:

Abstract

Rapid accumulation of the experimental data on protein-protein complexes drives the paradigm shift in protein docking from "traditional" template free approaches to template based techniques. Homology docking algorithms based on sequence similarity between target and template complexes can account for ~ 20% of known protein-protein interactions. When homologous templates for the target complex are not available, but the structure of the target monomers is known, docking through structural alignment may provide an adequate solution. Such an algorithm was developed based on the structural comparison of monomers to co-crystallized interfaces. A library of the interfaces was generated from the biological units. The success of the structure alignment of the interfaces depends on the way the interface is defined in terms of its structural content. We performed a systematic large-scale study to find the optimal definition/size of the interface for the structure alignment-based docking applications. The performance was the best when the interface was defined with a distance cutoff of 12 Å. The structure alignment protocol was validated, for both full and partial alignment, on the DOCKGROUND benchmark sets. Both protocols performed equally for higher-accuracy models (*i*-RMSD $\leq$ 5 Å). Overall, the partial structure alignment yielded more acceptable models than the full structure alignment (86 acceptable models were provided by partial structure alignment only, compared to 31 by full structure alignment only). Most templates identified by the partial structure alignment had very low sequence identity to targets and such templates were hard to detect by sequence-based methods. Detailed analysis of the models obtained for 372 test cases concluded that templates for higher-accuracy

models often shared not only local but also global structural similarity with the targets. However, interface similarity even in these cases was more prominent, reflected in more accurate models yielded by partial structure alignment. Conservation of protein-protein interfaces was observed in very diverse proteins. For example, target complexes shared interface structural similarity not only with hetero- and homo-complexes but also, in few cases, with crystal packing interfaces. The results indicate that the structure alignment techniques provide a much needed addition to the docking arsenal, with the combined structure alignment and template free docking success rate significantly surpassing that of the free docking alone.

Acknowledgements

First and foremost, I want to thank my advisor Ilya A. Vakser. It has been an honor to be his first Ph.D. student at KU. I appreciate all his contributions of time, thoughts, and funding to make my Ph.D. experience productive and full of learning. His natural calm, even during tough times, and enthusiasm towards research was motivational for me. I am also thankful for the freedom he provided to put new ideas during several rounds of the worldwide blind test (CAPRI), making them a great source of learning.

The members of the Vakser group have contributed immensely to my personal and professional time at KU. I would like to place my special gratitude to Petras J. Kundrotas, who had always been like a mentor and a source of inspiration at every step of my PhD. I would also like to thank other lab members: Andrey Tovchigrechko, Anatoly Ruvinsky, Zhengwei Zhu, Liu Shiyong, Ying Gao, Jagtar Hunjan, Tatsiana Kirys and Mallika Veeramalai.

Lastly, I would like to thank my family and friends for all their love and encouragement. For my parents whose blessings provided me the strength to pursue my studies at a place half a world away from my home and most of all, for my loving and supporting brother Siddharth and my wife Shweta. Thank you.

Rohita Sinha
University of Kansas

Table of Contents

List of Tables

List of Figures

List of Acronyms

ASA:	Accessible Surface Area
CAPRI:	Critical Assessment of Prediction of Interactions
CASP:	Critical Assessment of Protein Structure Prediction
Da:	Dalton
DG99:	DOCKGROUND benchmark set (99 unbound-unbound cases)
DG372:	DOCKGROUND benchmark set (372 bound-bound cases)
EM:	Electron Microscopy
FFT:	Fast Fourier Transform
FSA:	Full Structure Alignment
i-RMSD:	Interface-Root Mean Square Deviation
NP:	Nondeterministic Polynomial Time
NPC:	Nuclear Pore Complex
PDB:	Protein Data Bank
PPI:	Protein-Protein Interaction
PSA:	Partial Structure Alignment
PSI:	Protein Structure Initiative
RMSD:	Root Mean Square Deviation
SS:	Secondary Structure

CHAPTER 1: INTRODUCTION

Most proteins are made of more than one polypeptide chain [1]. Among these proteins, many, if not all, tend to interact with other proteins to form binary or higher order complexes responsible for an array of cellular processes. Genome-wide studies of several organisms have found that most proteins are part of multi-molecular assemblies [2-4] and alterations in protein interactions can lead to diseases [5]. Protein-protein interactions are important to the biological processes such as cellular regulation, signal transduction, etc. Thus, the study of principles governing protein-protein interactions (PPIs) along with structural details of protein complexes is essential for defining the cellular network of proteins and development of new drugs.

The interest in PPIs is as old as our ability to measure the weight of biological macromolecules, such as proteins. Pioneering work by Svedberg, determining the molecular weights of biomolecules, led to the realization that proteins in solution exist as aggregates of subunits and this state can be altered by changing the pH of the solution. His experiments with the ultracentrifuge defied the contemporary belief that hemoglobin is a single molecule of molecular weight 67000 daltons (Da), and described it as an aggregate of four subunits in the solution with molecular weight ~ 16000 Da for each subunit [6, 7]. Works of Svedberg have drawn attention to the fact that proteins have a tendency to interact and the interactions can be transient in nature. However, these studies failed to give any lead to the biochemical importance of subunit interactions.

Biochemical importance of protein quaternary structure was showcased in 1960, by Changeux, Gerhart, and Monod [8-11]. Their study of "allosteric interactions" and experiments on L-threonine deaminase showed that the functional forms of the proteins can be aggregates of non-active subunits. They further elucidated that association of substrates to protein subunits can change their inter-subunit interactions and relative conformations. Similar results were obtained for hemoglobin, where binding of oxygen leads to ~ 19% reduction in the distances between the heme molecules.

These and other studies led to the realization that cellular control mechanisms and regulation of enzyme activities are influenced by protein subunit interactions, which generated a widespread interest in protein interaction mechanisms and their quaternary structures.

1.1 Classification of protein-protein complexes

Development of experimental techniques detecting PPIs and the structures of protein assemblies has greatly increased our understanding of protein complexes. The increase of the number of protein complex structures in the Protein Data Bank (PDB) [12, 13] allows statistically significant analysis of the properties of protein complexes.

Systematic studies of the nature of protein complexes and the diversity of their interfaces place protein interactions into several different classes [14]. A multi-subunit protein may have identical or non-identical subunits (polypeptide chains). An "oligomer" is a multi-subunit protein with a definite number of subunits, whereas a

"polymer" is defined as a collection of an indefinite number of subunits. The subunits of oligomeric proteins are called "protomers", and a protomer consists of either a single polypeptide chain or multiple polypeptide chains. The extent of interactions between protomers is observed to correlate with their expression profiles (Figure 1.1).

Protein complexes can be classified on the basis of the following properties:

A- Nature of protomers

In an oligomeric protein, if the protomers are identical in nature then the complex is known as "homo-oligomer", otherwise called "hetero-oligomer". In the case of homo-oligomers, when protomers interact through identical surface patches the mode of interaction is defined as "isologous", otherwise termed as "heterologous" [11].

B- Stability of individual protomer

Protein complexes can be classified either as "obligate" or "non-obligate" according to the stability of their protomers. In an "obligate complex" protomers are co-expressed and do not exist as independent structures *in vivo*. However, protomers in "non-obligate complexes" exist independently in their folded functional forms and interact to carry out their functions. Non-obligate complexes are often hetero-oligomeric in nature and perceived to have weak transient interactions. However, they have diverse affinities and localization (Figure 1.1). For example, non-obligate interactions such as antibody-antigen have subunits with different locations of origin but show strong binding affinity [14].

C- Lifetime of a complex

Protein complexes have different lifetimes in the cellular environment. Depending on its lifespan, a protein complex is either described as "permanent" or "transient". Permanent complexes are stable *in vivo* whereas transient complexes dissociate to their individual protomers after a short-lived interaction. Few transient complexes are considered strong because they need a molecular trigger to switch their oligomeric states. For example, the heterotrimeric guanosine triphosphate (GTP)-binding protein dissociates into the Gα and Gβγ subunits upon GTP binding, but forms a stable trimer with bound guanosine diphosphate (GDP) [15].

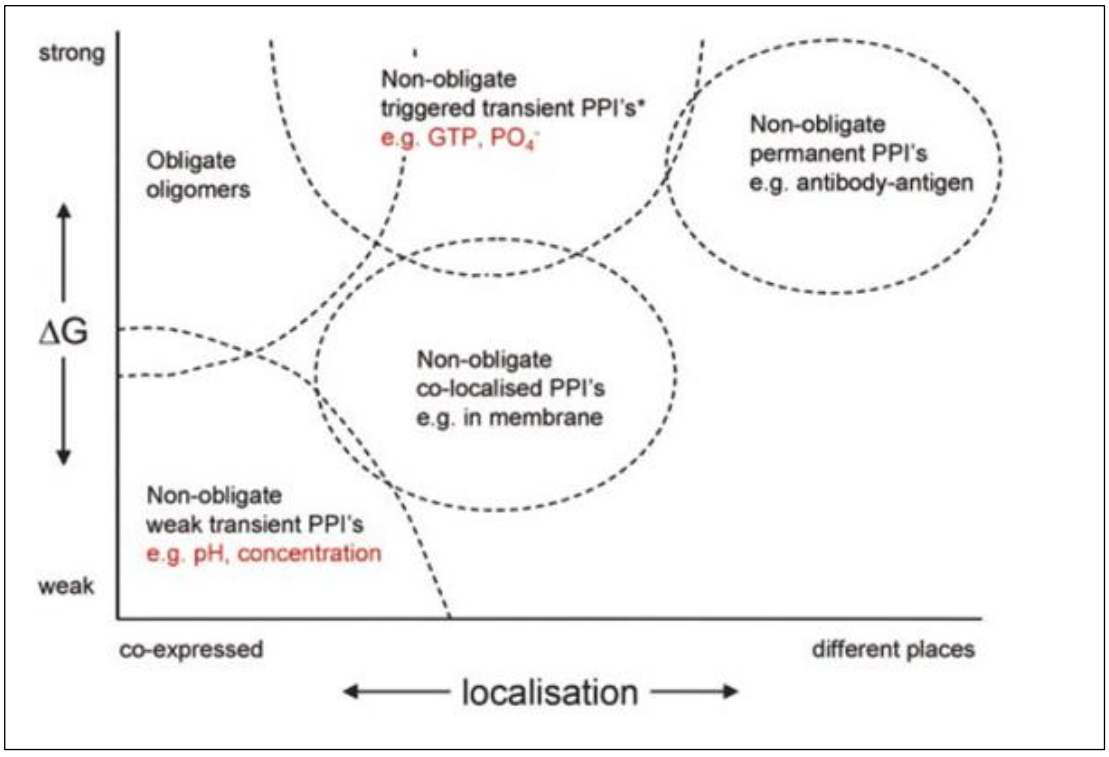

Figure 1.1: Characterization of protein interactions on the basis of the localizations and binding strengths. The obligate oligomers are always strongly attached but the non-obligate complexes show diverse binding strengths. Figure is obtained from [14].

Transient interactions play a significant role in the cellular regulatory system [16]. Their structures are hard to solve by X-ray crystallography; therefore, computational methods are often necessary for their characterization. Transient interactions affect the cellular regulations in the following ways:

- Transition of oligomeric state provides an allosteric control over the protein activity.
- A transient switch from monomer to dimer turns on the protein function. For example, lambda phage cro repressor (DNA-binding protein) is only active in their dimeric state.
- A transient interaction may lead to chemical modifications or exchange reactions, e.g. enzyme-substrate and electron transfer.
- Proteins may undergo a transient phase of aggregation to generate the concentration gradient.

Physiological conditions and environment change continuously inside the cell and play an important role in the control of transient interactions. The pH or ionic strength, concentration of proteins and other regulatory effector molecules (ions, chemical compounds) are regulated by the cell to control the oligomeric equilibrium of proteins.

1.2 Techniques to study protein-protein complexes

Most proteins *in vivo*, exist either as stable complexes or interact transiently with other proteins to perform metabolic and regulatory activities. Following are the common methods to study protein complexes.

1.2.1 Detection of protein-protein interactions

Proteins interact with other proteins while carrying out their cellular functions. The PPI networks are very large and it is estimated that a single protein interacts with ~ 10 other proteins [17-19]. Therefore, it is important to detect protein interaction partners prior to the systematic structure elucidation of the protein complexes. Detection of PPIs requires high-throughput experimental as well as computational methods [20-22] to detect all possible PPIs.

1.2.1.1 Experimental methods

Common experimental techniques for discriminating between the interacting and non-interacting protein pairs are affinity chromatography, affinity blotting, immunoprecipitation, cross-linking, and yeast two-hybrid. PPI data obtained through experimental techniques are stored in databases like DIP [23] and BIND [24]. These experiments have a significant number of false positive predictions and require additional experiments to confirm the results.

1.2.1.2 Computational methods

Experimental techniques providing PPI data are labor intensive and have a high share of false positive predictions [25-27]. Computational methods that detect PPIs complement and validate the experimental studies [28]. A study by Dandekhar [29] shows that for the 75% of co-localized gene pairs there are physical interactions between the encoded products. Proteins can be identified as functionally related if they share a similar phylogenetic profile [30]. Proteins with co-crystallized structures are an important resource for the prediction of new protein interactions. Protein pairs that are homologous to the co-crystallized proteins tend to interact similarly provided the interacting residues on the interface are conserved [30-33]. A few studies calculate the statistical probability of interaction for a given pair of domains, to predict PPIs [34-36]. To recognize new PPIs, conserved but short signature segments taking part in the interactions were derived from the experimentally defined protein interaction pairs through Support Vector Machine (SVM) techniques [37, 38]. The program PIPE defines proteins as interacting if they have a set of short polypeptide fragments that have been observed in known interacting protein pairs [39]. These common sets of protein fragments are assumed to be responsible for the interactions.

1.2.2 Describing the structures of protein-protein complexes

1.2.2.1 Experimental methods

X-ray crystallography is the most widely used technique to provide the structural details of protein complexes. The second most common method for studying protein structures is Nuclear Magnetic Resonance (NMR). It provides valuable

information on the dynamics of the proteins. Macromolecules larger than 100KDa are difficult to analyze using NMR, and NMR also requires large quantities of samples for the analysis. Electron microscopy (EM) provides a low resolution image of protein molecules and the resolution ranges between 5-15Å. Thus, to provide a reasonable atomic model of a protein complex, EM requires high-resolution structures of the subunits of the complex to fit the low resolution image.

1.2.2.2 Computational methods

Despite advances in experimental methods, the total number of co-crystallized complexes is still very low compared to the known PPIs. Therefore, there is a need for the development of methods to surmount the limitations of experimental techniques. With the rapid advancement in the computing power, computational methods modeling structures of protein complexes offer an adequate solution and complement experimental methods.

Computational methods of modeling protein complexes accept either sequences or structures of the subunits as input with the aim of producing an atomic model of the complex. Computational approaches predicting protein-protein complexes can be classified into the following major categories:

(A) Free modeling

(B) Template based modeling

(C) Hybrid approaches

A- Free modeling

The “Free modeling” category in Critical Assessment of protein Structure Prediction (CASP), a blind test for modeling structures of individual proteins, contains targets for which there are no templates available. Such targets are considered “new folds” [40]. Similarly, in computational modeling of protein-protein complexes, where the procedure does not depend on the presence of co-crystallized complexes (templates) such approaches are considered “Free modeling”.

Protein-protein docking methods came into the picture with the early works of Greer & Bush [41] and Wodak & Janin [42]. These studies were bound-bound docking experiments based on a simple surface complementarity search. Since then protein-protein docking has come a long way in terms of algorithms and scoring functions. Present docking methods still face the challenge of conformational changes upon complex formation. Existing “free modeling” protocols can be placed in one of the following types:

(1) Rigid body docking

(2) Flexible docking

1- Rigid body docking

Rigid body docking is defined as a docking protocol, which does not take into account the conformational changes in target proteins during the docking process. Such procedures work well for the bound-bound targets but their predictive power for the unbound protein structures is limited.

With the growth of the number of co-crystallized protein complexes in PDB it has been revealed that PPIs involve a varied degree of conformational changes. Protein-protein docking benchmark sets [43-45] represent the diversity of protein complexes and show that for > 50% of the complexes, the all-atom root mean square deviation (RMSD) between bound and unbound forms is < 2.0 Å. This is an indication that docking techniques which account for minor conformational changes can be reasonably successful.

The cubic grid model, proposed by Jiang & Kim [46], provides a low resolution representation of proteins. It has the softness necessary to accommodate minor conformational changes of proteins. Similar models are still relevant for rigid body docking and applied in docking programs, such as GRAMM [47], ZDOCK [48], etc.

A typical rigid body docking algorithm has two main steps:

1.A- Global search

The algorithm generates millions of binding modes for a pair of proteins. In the case of "free docking" there are six degrees of freedom (three translations and three rotations). Coverage of such a huge search space in a time efficient manner is essential for practical applications of docking methods.

Techniques like correlation by Fast Fourier Transform (FFT) [49] have made the coverage of protein-protein conformation space a feasible task. Such algorithms calculate protein surface cross correlations with proteins projected onto a grid. Monte-

Carlo, simulated annealing [50], and genetic algorithm [51] are alternative approaches to docking. They start with a random orientation and attempt to minimize the energy of the system. Simulated annealing allows selection of higher energy orientations based on certain probability, helping to avoid local minima. To minimize the search time and explore protein surface complementarity, "geometric hashing" is applied. Designed for matching three-dimensional objects, geometric hashing is an efficient docking approach. It also works with low resolution representation of proteins and therefore accommodates the minor conformational changes [52, 53].

1.B- Scoring

Protein-protein interfaces are not simple enough to apply only shape complementarity to discriminate between binding and non-binding patches. Numerous binding modes generated through the above search algorithms require additional parameters to bring the best model to the top. Most existing docking procedures apply various scoring parameters to rank predicted models. An efficient and accurate scoring function is essential for the practical application of a docking experiment. A free docking procedure generally applies physics-based energy functions to calculate the interaction energy of the protein molecules. Different types of force fields with various contributing factors are used to score the predicted complexes. Commonly used scoring functions may involve electrostatic interactions based on the Poisson-Boltzmann equation for the electrostatic energy contribution. To simplify the computation, only Poisson's equation can be applied [54, 55]. Other major parameters are hydrophobic interactions, hydrogen bonds and van der Waals interactions [56].

2- Flexible docking

Flexible docking methods take into account the conformational changes in protein molecules. Flexible docking is required due to two main reasons. First, proteins are flexible molecules and change their conformations while interacting with other proteins. The degree of flexibility ranges from small side-chain movements [57, 58] to big domain shifts [59]. When these conformational changes are relatively large (>2.0 Å), rigid body docking tends to fail. Second, with the advances in computational structural biology there are reasonably accurate models for the proteins when the experimental structures are not available. Such models may have certain degrees of conformational deviations from their bound as well as unbound forms. Thus, protein docking methods require incorporation of the structural flexibility.

Flexibility of the main chain is accounted for either by allowing movement during minimization or by docking an ensemble of protein conformations which are either generated computationally or obtained by NMR [60-62].

High resolution modeling of a protein complex requires an accurate sampling of side-chain conformations at the protein interface. There are studies reflecting improvement in docking predictions with the incorporation of the side chain flexibility [63-65].

B- Template based modeling

Large scale experimental efforts initiated by second generation structural genomics, focus on protein complexes. Examples of such efforts are SPINE2Complex

and 3D Repertoire. SPINE2 (http://www.spine2.eu/SPINE2/) focuses on complexes in signaling pathways linking immunology, neurobiology and cancer. 3D Repertoire (ended in Jan 2010) focused on protein complexes from yeast proteome. Such experimental efforts along with Protein Structure Initiative (PSI) in the US, led to the exponential growth of PDB data in terms of heteromeric complexes [66].

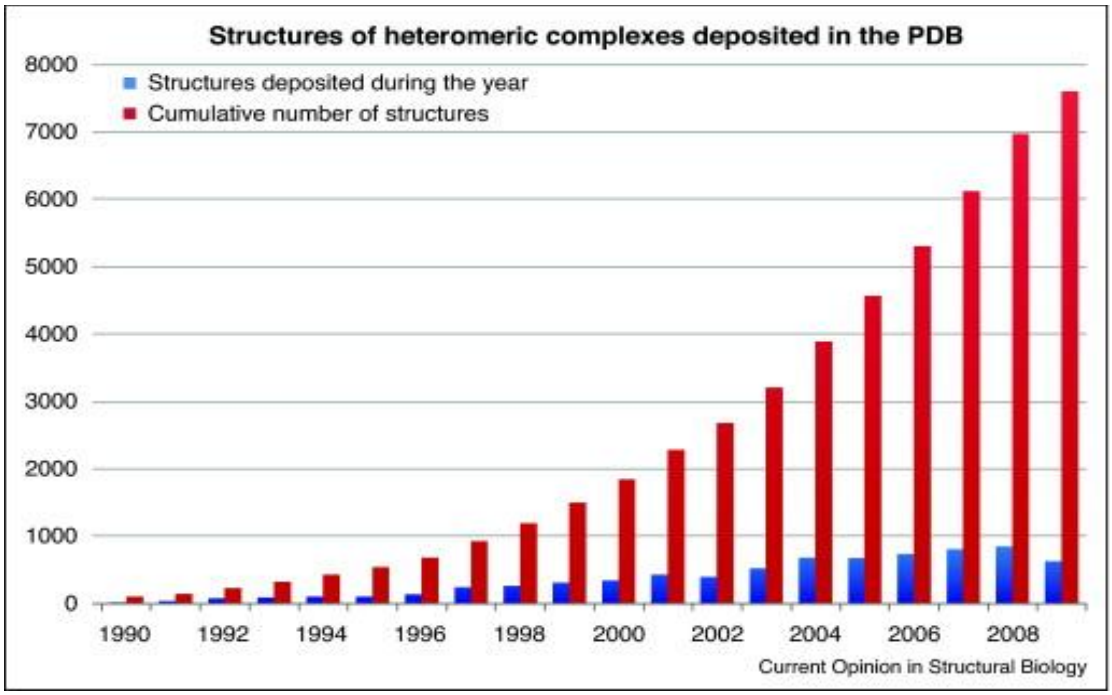

Figure 1.2: Growth of heteromeric protein complexes in PDB. Figure is obtained from [66].

Template based methods are defined as modeling of protein complexes on the basis of existing co-crystallized structures of proteins. Increase of the numbers of protein complexes in PDB (Figure 1.2) encourages extending the template based modeling paradigm from single chain structure prediction to the protein complex modeling. Homology modeling requires a certain degree of sequence identity to transfer the structural information from template to target molecule. An early work of

Aloy & Russell [67] demonstrated that the domains sharing > 30% of sequence identity interact similarly. However, the study also found that the similarities of folds between the proteins do not ensure similar interactions.

In continuation of the above work, protein complexes from the yeast proteome (102 protein complexes) were subjected to homology modeling [68]. Low resolution EM data were used for the cross validation of the models. Templates were primarily selected through sequence homology. In the absence of homology, complexes sharing similar folds with target components were used as templates. Out of 102 cases, nearly complete models were generated for 42 protein complexes.

Similarly, Davis et al. [69] modeled ~ 1250 higher order protein complexes from yeast. Target domains were aligned to the template proteins and interfaces were scored by statistical potentials. For higher order complexes, proteins with more than two domains were taken as templates and predicted complexes were merged if they contained different domains of a single protein. Predictions were validated against the DIP [23] and BIND [24] datasets and successfully validated structures were deposited into MODBASE [70]. This study was different from Aloy's [68] in terms of the template source PIBASE [71], and performed the structural alignment of the targets to the template structures instead of the comparative modeling.

With increasing evidence that protein binding patches are more conserved than the global folds of the proteins [72], structural similarities with binding patches were detected and applied to model new protein complexes [73]. It showed reasonable success on a benchmark set of 59 complexes. Prediction of PPIs through structural

similarity of protein interfaces, has increased the focus on geometric properties of protein binding sites [74]. Alloy & Russell [75] calculated the upper limit of the types of quaternary structures as ~ 10,000 types of protein complex structures. Skolnick & Gao [76] concluded in their study that interface structural space is ~ 80% complete.

Comparative modeling of protein complexes faces the challenge of limited availability of the templates. To extend the template space, M-TASSER applied multimeric threading to detect remotely related templates [77]. The procedure input is protein sequences which are individually modeled through threading and then subjected to iterative threading in the dimers library. The method was tested to predict the quaternary structures on a set of ~250 dimers. About 80% of the dimer interactions were correctly predicted with an impressive RMSD average of 1.3 Å. Similarly, profile based alignment was applied to detect the remotely related template sequences [78] performing better than PSI-BLAST [79] detection of templates. General protocol of template based modeling of protein complexes is summarized in Figure 1.3.

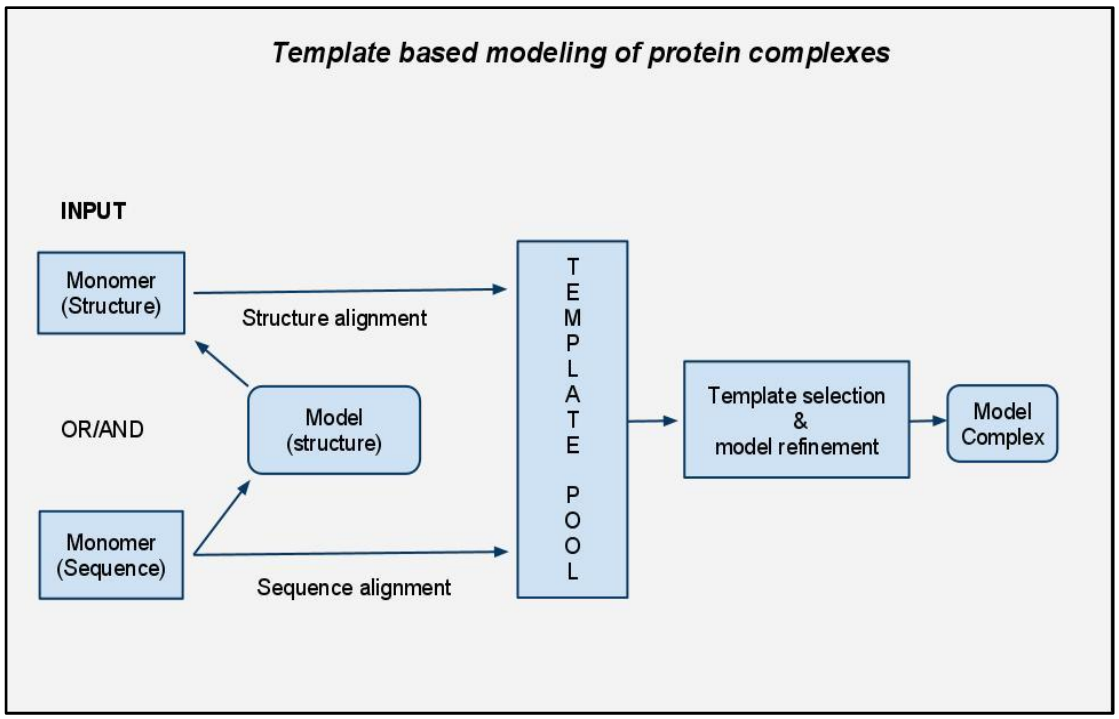

Figure 1.3: A generalized diagram of template based modeling of protein complexes. Input is either sequence or structure of the target proteins. Targets are aligned to template complexes through sequence or structure alignment, and a template showing significant alignment is used to model new protein complex.

C- Hybrid approaches

Experimental methods providing high resolution structural data, due to their intrinsic limitations, cannot cover the protein interaction network. On the other hand, computational methods have their own challenges, such as an enormous degrees of freedom, limited template pool, etc. A natural approach would be the use of experimental data (other than binding modes of the co-crystallized structures) as constraints to drive the computational modeling procedures. Such approaches have seen many successes in the recent past [80]. The following are cases in which experimental data was applied to assist computational modeling.

C.1- Modeling higher order complexes

A combination of biophysical data with computational approaches has helped in modeling macromolecular assemblies like nuclear pore complex (NPC), RNA polymerase II and ribosome. NPC is a 50 MDa macromolecule with ~ 30 subunits and a total of 450 proteins (Figure 1.4). To solve the structure, experimental data was translated into spatial constraints and the energy function was generated and optimized to maximize the compliance with constraints [81]. Since most of the biochemical mechanisms are carried out through large protein assemblies, their successful modeling improves our understanding of cellular machinery [82].

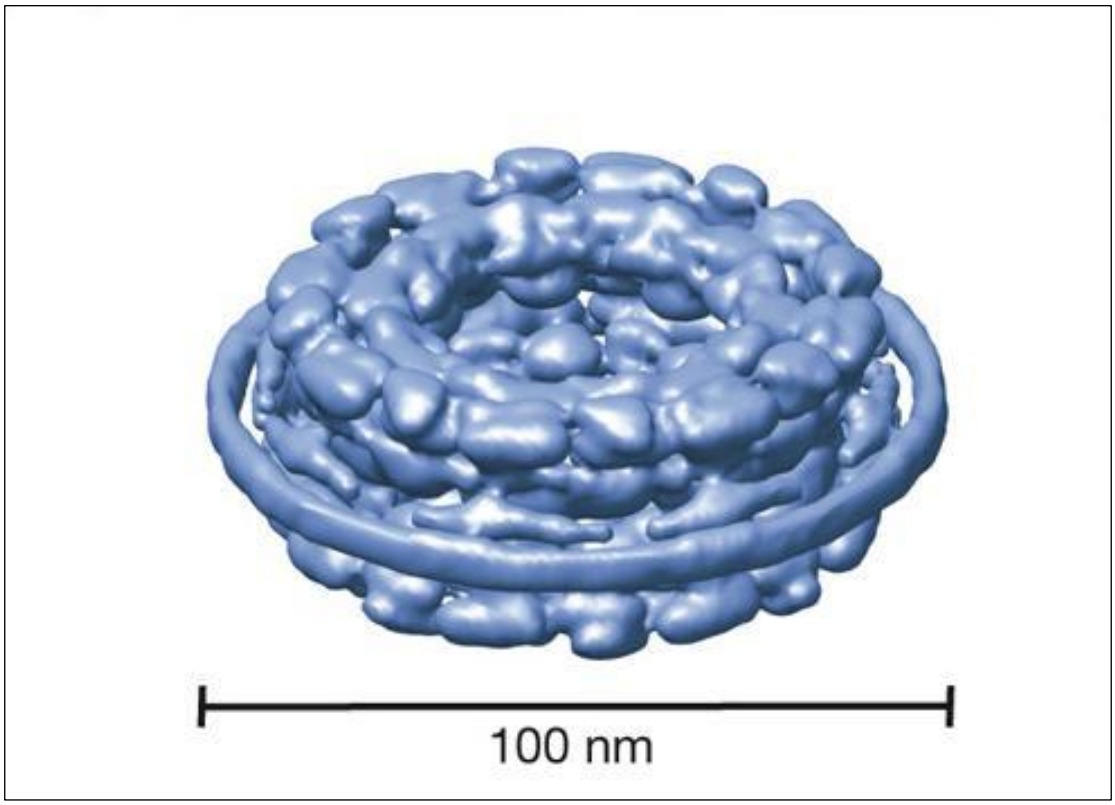

Figure 1.4: A low resolution image of Nuclear Pore Complex (NPC). Figure is obtained from [81]

C.2- Statistical potentials

Practical implementation of the Boltzmann distribution law allows one to derive residue pair potentials. Statistical data is obtained from solved structures of protein complexes. Statistical potentials are important because they implicitly take care of thermodynamics and solvation effects. Potentials derived for residue-residue contacts can be applied at the scan stage (the initial docking stage performed with computationally inexpensive scoring functions such as shape complementarity). Boltzmann distribution for a specific pair of residues is represented as:

$$P^{i}_{(A-B)} = \frac{e^{-\frac{E^{i}_{(A-B)}}{kT}}}{Z} \quad (1.1)$$

$$Z = \sum_{i=1}^{N} e^{-\frac{E^{i}_{(A-B)}}{kT}} \quad (1.2)$$

A-B - residue pair at a specific distance

$E^{i}_{(A-B)}$ - energy of the i^{th} state, related to residue pair (A-B) at a specific distance

k - Boltzmann constant: $1.38\ 10^{-23}$ J/K

T - absolute temperature

N - total number of energy states

$P^{i}_{(A-B)}$ – the probability of the i^{th} state

Z - Partition function

Equation 1.1 can be inversed and solved to the following form:

$$\Delta E^{i}_{(A-B)} = -kT \ln \frac{P^{i}_{(A-B)}}{P_{(A-B)}} \quad (1.3)$$

$\Delta E^{i}_{(A-B)}$ - energy contribution of the i^{th} energy state in the total energy of the system.

$P_{(A-B)}$ - the probability of the reference state.

Equation 1.3 provides energy contribution of residue pair (A-B) to the overall interaction energy of the system. The residue pair interaction data is extracted from known co-crystallized structures.

C.3- Docking with constraints

Protein complexes can be modeled incorporating experimental data (other than binding modes of the co-crystallized structures) to the free docking protocols with the aim of either restricting the global search space or filtering docking predictions. HADDOCK [83], a data driven docking protocol, uses multiple types of biochemical and biophysical data such as site directed mutagenesis, NMR (chemical shift, Residual Dipolar Couplings), mass-spectroscopy and computational interface predictions to guide the conformational search. Other programs like GRAMM-X [84], Zdock [48], PyDOCK [85, 86] and PatchDock [87] can filter their results based on experimental constraints. Multifit [88] uses EM data to fit the docking output.

In summary, computational methods are vital for the study of PPIs. Parallel to the maturing free docking methodologies, there are efforts to develop template based modeling techniques. It is evident that the success of the template based approach is dependent on the richness of the template pool. Along with PDB there are additional repositories providing information of the template structures; secondary databases, such as DOCKGROUND [89] and Protein Quaternary Structure (PQS) [90] contain

structural information on the biological units. As per PQS statistics, there are a significant number of protein complex structures to evaluate the modeling abilities of template based methods on the genomic scale (Table 1.1).

Table 1.1: The number of biological units in PQS.

Oligomer size	Number of generated oligomers[a]	Number of homo-oligomers	Number of hetero-oligomers
Monomer/complex	22514		
Dimer	18708	13974	4734
Trimer	4055	1922	2133
Tetramer	6495	4205	2351
Pentamer	459	213	246
Hexamer	2019	1257	762
Heptamer	103	49	54
Octamer	865	508	357
Nanomer	95	11	84
Decamer	171	98	73
Undecamer	28	18	10
Dodecamer	511	233	278
Tetradecamer	52	37	15
Hexadecamer	101	18	83
Octadecamer	27	7	20

[a]Biological units available for each class of oligomers.

Data is obtained from [90].

1.3 Research presented in this thesis

Typical free docking methods suffer from the following limitations:

(1) They are largely dependent on the surface complementarity, which makes them sensitive to the structural details of the target proteins. Conformational changes and modeled structures pose a great challenge to these protocols.

(2) Scoring functions for ranking the predicted models often fail to rank the near native predictions to the top.

(3) Additional experimental information or constraints to add confidence to the predictions are required.

The limitations make way for the development of template based methods, which have an edge over the free docking.

This thesis presents the study of the application of template based modeling to predict new protein complexes through structural alignment of target and template proteins. It also demonstrates the applicability of structural alignment methods to genome-wide high-throughput docking experiments.

The importance of template based modeling of protein interactions grows with the increasing number of solved co-crystallized protein structures. Unlike free docking, template based docking is relatively less sensitive to the structural details of the target proteins and has an evolutionary basis for the predictions. Therefore, it provides a greater degree of confidence in the predictions.

Since the docking problem assumes *a priori* knowledge of the structures of the participating proteins, templates may be found by structural (rather than sequence) alignment of the target monomers and the co-crystallized complexes. This thesis establishes structure alignment protocol as a method ready to be applied on the genome-wide scale to model new protein complexes.

The work presented in this thesis is broadly divided into three parts. In the first part a structural definition of the protein interface is obtained. It determines the optimum distance cutoff to define the interfaces for structural alignment. In the second part, the ability of the interface structure alignment method to model new protein complexes is tested. The results demonstrate that the success of the structural alignment method increases the ability to go beyond the template space covered by sequence based prediction methods. Further, the structure alignment method complements the free docking protocol and provides a significantly higher number of near native models. Previously structure alignment (global structural match) was applied to predict PPIs and protein complexes' structures [69, 91]. However, for the first time we benchmark its ability to provide acceptable models of protein complexes. The third part of the work describes the pros and cons of aligning global folds vs. the alignment of interfaces. It shows the extent of structural conservation across the protein-protein complexes and its impact on the applicability of full structure alignment (FSA) and partial structure alignment (PSA) methods.

This study improves the ability to model new protein complexes and to better understand the role of structural alignment in modeling the networks of protein-protein complexes.

References

1. Janin, J., R.P. Bahadur, and P. Chakrabarti, *Protein-protein interaction and quaternary structure.* Q Rev Biophys, 2008. 41:133-80.

2. Li, S., C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.O. Vidalain, J.D. Han, A. Chesneau, T. Hao, D.S. Goldberg, N. Li, M. Martinez, J.F. Rual, P. Lamesch, L. Xu, M. Tewari, S.L. Wong, L.V. Zhang, G.F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H.W. Gabel, A. Elewa, B. Baumgartner, D.J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S.E. Mango, W.M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K.C. Gunsalus, J.W. Harper, M.E. Cusick, F.P. Roth, D.E. Hill, and M. Vidal, *A map of the interactome network of the metazoan C. elegans.* Science, 2004. 303:540-43.

3. Krogan, N.J., G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A.P. Tikuisis, T. Punna, J.M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M.D. Robinson, A. Paccanaro, J.E. Bray, A. Sheung, B. Beattie, D.P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M.M. Canete, J. Vlasblom, S. Wu, C. Orsi, S.R. Collins, S. Chandran, R. Haw, J.J. Rilstone, K. Gandi, N.J. Thompson, G. Musso, P. St Onge, S. Ghanny, M.H. Lam, G. Butland, A.M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O'Shea, J.S. Weissman, C.J. Ingles, T.R. Hughes, J. Parkinson, M. Gerstein, S.J. Wodak, A. Emili, and J.F. Greenblatt, *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.* Nature, 2006. 440:637-43.

4. Giot, L., J.S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y.L. Hao, C.E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C.A. Stanyon, R.L. Finley, Jr., K.P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R.A. Shimkets, M.P. McKenna, J. Chant, and J.M. Rothberg, *A protein interaction map of Drosophila melanogaster.* Science, 2003. 302:1727-36.

5. Al-Khoury, R. and B. Coulombe, *Defining protein interactions that regulate disease progression.* Expert Opin Ther Targets, 2009. 13:13-17.

6. Svedberg, T., *Mass and Size of Protein Molecules.* Nature, 1929. 123:871.

7. Svedberg, T. and R. Fahraeus, *A new method for the determination of the molecular weight of the proteins.* J. Am. Chem. Soc., 1926. 48:430-438.

8. Changeux, J.P., *Allosteric proteins: from regulatory enzymes to receptors--personal recollections.* Bioessays, 1993. 15:625-34.

9. Gerhart, J.C. and A.B. Pardee, *The enzymology of control by feedback inhibition.* J Biol Chem, 1962. 237:891-96.

10. Monod, J., J.P. Changeux, and F. Jacob, *Allosteric proteins and cellular control systems.* J Mol Biol, 1963. 6:306-29.

11. Monod, J., J. Wyman, and J.P. Changeux, *On the nature of allosteric transitions: A plausible model.* J Mol Biol, 1965. 12:88-118.

12. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, *The Protein Data Bank.* Nucleic Acids Res, 2000. 28:235-42.

13. Berman, H., K. Henrick, H. Nakamura, and J.L. Markley, *The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data.* Nucleic Acids Res, 2007. 35:D301-3.

14. Nooren, I.M. and J.M. Thornton, *Diversity of protein-protein interactions.* EMBO J, 2003. 22:3486-92.

15. Digby, G.J., R.M. Lober, P.R. Sethi, and N.A. Lambert, *Some G protein heterotrimers physically dissociate in living cells.* Proc Natl Acad Sci U S A, 2006. 103:17789-94.

16. Nooren, I.M. and J.M. Thornton, *Structural characterisation and functional significance of transient protein-protein interactions.* J Mol Biol, 2003. 325:991-1018.

17. Kumar, A. and M. Snyder, *Protein complexes take the bait.* Nature, 2002. 415:123-4.

18. Russell, R.B., F. Alber, P. Aloy, F.P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali, *A structural perspective on protein-protein interactions.* Curr Opin Struct Biol, 2004. 14:313-24.

19. von Mering, C., R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork, *Comparative assessment of large-scale data sets of protein-protein interactions.* Nature, 2002. 417:399-403.

20. Phizicky, E.M. and S. Fields, *Protein-protein interactions: methods for detection and analysis.* Microbiol Rev, 1995. 59:94-123.

21. Berggard, T., S. Linse, and P. James, *Methods for the detection and analysis of protein-protein interactions.* Proteomics, 2007. 7:2833-42.

22. Ng, S.K. and S.H. Tan, *Discovering protein-protein interactions.* J Bioinform Comput Biol, 2004. 1:711-41.

23. Xenarios, I., D.W. Rice, L. Salwinski, M.K. Baron, E.M. Marcotte, and D. Eisenberg, *DIP: the database of interacting proteins.* Nucleic Acids Res, 2000. 28:289-91.

24. Bader, G.D., I. Donaldson, C. Wolting, B.F. Ouellette, T. Pawson, and C.W. Hogue, *BIND--The Biomolecular Interaction Network Database.* Nucleic Acids Res, 2001. 29:242-45.

25. Aloy, P. and R.B. Russell, *The third dimension for protein interactions and complexes.* Trends Biochem Sci, 2002. 27:633-38.

26. Franzot, G. and O. Carugo, *Computational approaches to protein-protein interaction.* J Struct Funct Genomics, 2003. 4:245-55.

27. Marcotte, E.M., M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg, *Detecting protein function and protein-protein interactions from genome sequences.* Science, 1999. 285:751-53.

28. Pitre, S., M. Alamgir, J.R. Green, M. Dumontier, F. Dehne, and A. Golshani, *Computational methods for predicting protein-protein interactions.* Adv Biochem Eng Biotechnol, 2008. 110:247-67.

29. Dandekar, T., B. Snel, M. Huynen, and P. Bork, *Conservation of gene order: a fingerprint of proteins that physically interact.* Trends Biochem Sci, 1998. 23:324-8.

30. Aloy, P. and R.B. Russell, *Interrogating protein interaction networks through structural biology.* Proc Natl Acad Sci U S A, 2002. 99:5896-901.

31. Aloy, P. and R.B. Russell, *InterPreTS: protein interaction prediction through tertiary structure.* Bioinformatics, 2003. 19:161-62.

32. Lu, L., H. Lu, and J. Skolnick, *MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading.* Proteins, 2002. 49:350-64.

33. Ogmen, U., O. Keskin, A.S. Aytuna, R. Nussinov, and A. Gursoy, *PRISM: protein interactions by structural matching.* Nucleic Acids Res, 2005. 33:W331-36.

34. Deng, M., S. Mehta, F. Sun, and T. Chen, *Inferring domain-domain interactions from protein-protein interactions.* Genome Res, 2002. 12:1540-48.

35. Han, D.S., H.S. Kim, W.H. Jang, S.D. Lee, and J.K. Suh, *PreSPI: a domain combination based prediction system for protein-protein interaction.* Nucleic Acids Res, 2004. 32:6312-20.

36. Han, D.S., H.S. Kim, W.H. Jang, S.D. Lee, and J.K. Suh, *PreSPI: design and implementation of protein-protein interaction prediction service system.* Genome Inform, 2004. 15:171-80.

37. Bock, J.R. and D.A. Gough, *Predicting protein--protein interactions from primary structure.* Bioinformatics, 2001. 17:455-60.

38. Martin, S., D. Roe, and J.L. Faulon, *Predicting protein-protein interactions using signature products.* Bioinformatics, 2005. 21:218-26.

39. Pitre, S., F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Jessulat, N. Krogan, X. Luo, and A. Golshani, *PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs.* BMC Bioinformatics, 2006. 7:1-15.

40. Moult, J., K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard, and A. Tramontano, *Critical assessment of methods of protein structure prediction-Round VII.* Proteins-Structure Function and Bioinformatics, 2007. 69 Suppl 8:3-9.

41. Greer, J. and B.L. Bush, *Macromolecular shape and surface maps by solvent exclusion.* Proc Natl Acad Sci U S A, 1978. 75:303-7.

42. Wodak, S.J. and J. Janin, *Computer analysis of protein-protein interaction.* J Mol Biol, 1978. 124:323-42.

43. Hwang, H., T. Vreven, J. Janin, and Z. Weng, *Protein-protein docking benchmark version 4.0.* Proteins, 2010. 78:3111-14.

44. Chen, R., J. Mintseris, J. Janin, and Z. Weng, *A protein-protein docking benchmark.* Proteins, 2003. 52:88-91.

45. Gao, Y., D. Douguet, A. Tovchigrechko, and I.A. Vakser, *DOCKGROUND system of databases for protein recognition studies: unbound structures for docking.* Proteins, 2007. 69:845-51.

46. Jiang, F. and S.H. Kim, *"Soft docking": matching of molecular surface cubes.* J Mol Biol, 1991. 219:79-102.

47. Vakser, I.A., *Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex.* Proteins, 1997. Suppl 1:226-30.

48. Chen, R., L. Li, and Z. Weng, *ZDOCK: an initial-stage protein-docking algorithm.* Proteins, 2003. 52:80-7.

49. Katchalski-Katzir, E., I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, and I.A. Vakser, *Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.* Proc Natl Acad Sci U S A, 1992. 89:2195-99.

50. Bernauer, J., J. Aze, J. Janin, and A. Poupon, *A new protein-protein docking scoring function based on interface residue properties.* Bioinformatics, 2007. 23:555-62.

51. Totrov, M. and R. Abagyan, *Detailed ab initio prediction of lysozyme-antibody complex with 1.6 A accuracy.* Nat Struct Biol, 1994. 1:259-63.

52. Schneidman-Duhovny, D., Y. Inbar, V. Polak, M. Shatsky, I. Halperin, H. Benyamini, A. Barzilai, O. Dror, N. Haspel, R. Nussinov, and H.J. Wolfson, *Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking.* Proteins, 2003. 52:107-12.

53. Inbar, Y., H. Benyamini, R. Nussinov, and H.J. Wolfson, *Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies.* Phys Biol, 2005. 2:S156-65.

54. Honig, B. and A. Nicholls, *Classical electrostatics in biology and chemistry.* Science, 1995. 268:1144-49.

55. Mandell, J.G., V.A. Roberts, M.E. Pique, V. Kotlovyi, J.C. Mitchell, E. Nelson, I. Tsigelny, and L.F. Ten Eyck, *Protein docking using continuum electrostatics and geometric fit.* Protein Eng, 2001. 14:105-13.

56. Janin, J., *The kinetics of protein-protein recognition.* Proteins-Structure Function and Bioinformatics, 1997. 28:153-61.

57. Dey, S., A. Pal, P. Chakrabarti, and J. Janin, *The subunit interfaces of weakly associated homodimeric proteins.* J Mol Biol, 2010. 398:146-60.

58. Ruvinsky, A.M., T. Kirys, A.V. Tuzikov, and I.A. Vakser, *Side-Chain Conformational Changes upon Protein-Protein Association.* J Mol Biol, 2011. 408:356-65.

59. Goh, C.S., D. Milburn, and M. Gerstein, *Conformational changes associated with protein-protein interactions.* Curr Opin Struct Biol, 2004. 14:104-9.

60. Shatsky, M., R. Nussinov, and H.J. Wolfson, *FlexProt: alignment of flexible protein structures without a predefinition of hinge regions.* J Comput Biol, 2004. 11:83-106.

61. Emekli, U., D. Schneidman-Duhovny, H.J. Wolfson, R. Nussinov, and T. Haliloglu, *HingeProt: automated prediction of hinges in protein structures.* Proteins, 2008. 70:1219-27.

62. Schneidman-Duhovny, D., R. Nussinov, and H.J. Wolfson, *Automatic prediction of protein interactions with large scale motion.* Proteins, 2007. 69:764-73.

63. Schueler-Furman, O., C. Wang, and D. Baker, *Progress in protein-protein docking: atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility.* Proteins-Structure Function and Bioinformatics, 2005. 60:187-94.

64. Fernandez-Recio, J., M. Totrov, and R. Abagyan, *ICM-DISCO docking by global energy optimization with fully flexible side-chains.* Proteins-Structure Function and Bioinformatics, 2003. 52:113-17.

65. Camacho, C.J., *Modeling side-chains using molecular dynamics improve recognition of binding region in CAPRI targets.* Proteins-Structure Function and Bioinformatics, 2005. 60:245-51.

66. Stein, A., R. Mosca, and P. Aloy, *Three-dimensional modeling of protein interactions and complexes is going 'omics.* Curr Opin Struct Biol, 2011. 21:200-8.

67. Aloy, P., H. Ceulemans, A. Stark, and R.B. Russell, *The relationship between sequence and interaction divergence in proteins.* J Mol Biol, 2003. 332:989-98.

68. Aloy, P., B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A.C. Gavin, P. Bork, G. Superti-Furga, L. Serrano, and R.B. Russell, *Structure-based assembly of protein complexes in yeast.* Science, 2004. 303:2026-29.

69. Davis, F.P., H. Braberg, M.Y. Shen, U. Pieper, A. Sali, and M.S. Madhusudhan, *Protein complex compositions predicted by structural similarity.* Nucleic Acids Res, 2006. 34:2943-52.

70. Pieper, U., N. Eswar, H. Braberg, M.S. Madhusudhan, F.P. Davis, A.C. Stuart, N. Mirkovic, A. Rossi, M.A. Marti-Renom, A. Fiser, B. Webb, D. Greenblatt, C.C. Huang, T.E. Ferrin, and A. Sali, *MODBASE, a database of annotated comparative protein structure models, and associated resources.* Nucleic Acids Res, 2004. 32:D217-22.

71. Davis, F.P. and A. Sali, *PIBASE: a comprehensive database of structurally defined protein interfaces.* Bioinformatics, 2005. 21:1901-7.

72. Henschel, A., W.K. Kim, and M. Schroeder, *Equivalent binding sites reveal convergently evolved interaction motifs.* Bioinformatics, 2006. 22:550-55.

73. Gunther, S., P. May, A. Hoppe, C. Frommel, and R. Preissner, *Docking without docking: ISEARCH--prediction of interactions using known interfaces.* Proteins, 2007. 69:839-44.

74. Kim, W.K., A. Henschel, C. Winter, and M. Schroeder, *The many faces of protein-protein interactions: A compendium of interface geometry.* PLoS Comput Biol, 2006. 2:1151-64.

75. Aloy, P. and R.B. Russell, *Ten thousand interactions for the molecular biologist.* Nat Biotechnol, 2004. 22:1317-21.

76. Gao, M. and J. Skolnick, *Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected.* Proc Natl Acad Sci U S A, 2010. 107:22517-22.

77. Chen, H. and J. Skolnick, *M-TASSER: an algorithm for protein quaternary structure prediction.* Biophys J, 2008. 94:918-28.

78. Kundrotas, P.J., M.F. Lensink, and E. Alexov, *Homology-based modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles.* Int J Biol Macromol, 2008. 43:198-208.

79. Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. 25:3389-402.

80. de Vries, S.J., A.D. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, T. Wassenaar, and A.M. Bonvin, *HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets.* Proteins-Structure Function and Bioinformatics, 2007. 69:726-33.

81. Alber, F., S. Dokudovskaya, L.M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B.T. Chait, A. Sali, and M.P.

Rout, *The molecular architecture of the nuclear pore complex.* Nature, 2007. 450:695-701.

82. Lasker, K., J.L. Phillips, D. Russel, J. Velazquez-Muriel, D. Schneidman-Duhovny, E. Tjioe, B. Webb, A. Schlessinger, and A. Sali, *Integrative structure modeling of macromolecular assemblies from proteomics data.* Mol Cell Proteomics, 2010. 9:1689-702.

83. Dominguez, C., R. Boelens, and A.M. Bonvin, *HADDOCK: a protein-protein docking approach based on biochemical or biophysical information.* J Am Chem Soc, 2003. 125:1731-37.

84. Tovchigrechko, A. and I.A. Vakser, *GRAMM-X public web server for protein-protein docking.* Nucleic Acids Res, 2006. 34:W310-14.

85. Pons, C., A. Solernou, L. Perez-Cano, S. Grosdidier, and J. Fernandez-Recio, *Optimization of pyDock for the new CAPRI challenges: Docking of homology-based models, domain-domain assembly and protein-RNA binding.* Proteins, 2010. 78:3182-88.

86. Cheng, T.M., T.L. Blundell, and J. Fernandez-Recio, *pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking.* Proteins, 2007. 68:503-15.

87. Schneidman-Duhovny, D., Y. Inbar, R. Nussinov, and H.J. Wolfson, *PatchDock and SymmDock: servers for rigid and symmetric docking.* Nucleic Acids Res, 2005. 33:W363-67.

88. Lasker, K., M. Topf, A. Sali, and H.J. Wolfson, *Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly.* J Mol Biol, 2009. 388:180-94.

89. Douguet, D., H.C. Chen, A. Tovchigrechko, and I.A. Vakser, *DOCKGROUND resource for studying protein-protein interfaces.* Bioinformatics, 2006. 22:2612-18.

90. Henrick, K. and J.M. Thornton, *PQS: a protein quaternary structure file server*. Trends Biochem Sci, 1998. 23:358-61.

91. Doolittle, J.M. and S.M. Gomez, *Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens.* Virol J, 2010. 7:1-15.

CHAPTER 2: ALGORITHMS AND RESOURCES

2.1 Protein structure alignment

2.1.1 Structure alignment protocol

We use TM-align [1] as the structural alignment method. The procedure reflects the degree of structural similarity through TM-score [2]. TM-align performs a fast and exhaustive search to find the optimum alignment of two given protein structures and the alignment with the highest TM-score is the final output. Since alignment of the structures is a nondeterministic polynomial time hard (NP-hard) problem, TM-align takes different start points and systematically maximizes the TM-score to find the best alignment.

TM-align performs alignment of Cα atoms and thus is independent of the rotameric states of the side chains. Since it is mainly the side chains that change their conformation during binding [3], the Cα alignment solves the problem of minor conformational differences between the template (unbound) and target (bound) proteins.

TM-align takes several initial alignments and the initial alignments are obtained through the following methods:

(1) Dynamic programming, where residues are represented by their secondary structure (SS) elements. The score matrix is a binary matrix (1, 0). Aligned residues with identical SS elements score 1, otherwise 0.

(2) Gapless threading of the smaller protein against the larger protein. Alignment with the best TM-score is selected.

(3) Dynamic programming is used to obtain the best alignment. The scoring matrix is a combination of the SS matrix and the matrix used in gapless threading.

(4) The optimum alignment of the fragmented proteins, e.g. protein interfaces. In such cases only the largest fragment of the smaller protein is considered for threading.

Once an initial alignment is obtained, iterative dynamic programming is applied to obtain the optimum structure alignment. The TM-score matrix is used as the scoring matrix during iterations of dynamic programming.

2.1.2 Measuring degree of structural similarity

RMSD is a traditional measure of the structural similarity between two proteins. Despite being intuitive in nature, RMSD is sensitive to the degree of alignment or the alignment coverage. A target-template alignment with 2 Å RMSD and 50% alignment coverage provides a poorer template than an alignment with 3 Å RMSD and 80% alignment coverage [2].

Another problem in scoring the structural similarity is the dependence on protein size for randomly related proteins. It is observed that proteins with smaller sizes can generate a significantly higher score in the alignment. TM-score is designed to tackle the above problems. The TM-score for an aligned pair of proteins is defined as:

$$\text{TM-score} = \text{Max}\left[{}^{1}/_{L_N}\sum_{i=1}^{L_T} 1/1 + \left(\frac{d_i}{d_{0(L_{min})}}\right)^2\right] \quad (2.1)$$

L_N - length of the target protein

L_T - length of aligned residues

d_i - distance between the i^{th} aligned residues.

L_{min} - length of the smaller protein

The equation to calculate d_0 is optimized to the following form:

$$d_{0(Lmin)} = 1.24\sqrt[3]{L_{min} - 15} - 1.8 \quad (2.2)$$

In the case of RMSD, residues with a poor or high degree of structural alignment are both averaged with the same weight, whereas in TM-score the degree of contribution changes with the quality of alignment.

The value of $d_{0(Lmin)}$ (Equation 2.2) is very efficient in differentiating random alignments with good quality alignments. The d_0 values of 5 and ($1.24\sqrt[3]{Lmin - 15} - 1.8$) are compared in Figure 2.1. For $d_0 = 5$ the TM-score is dependent on the length of proteins, whereas the modified equation (Equation 2.2) restricts the TM-score to 0.17 for the random alignments irrespective to the length of proteins.

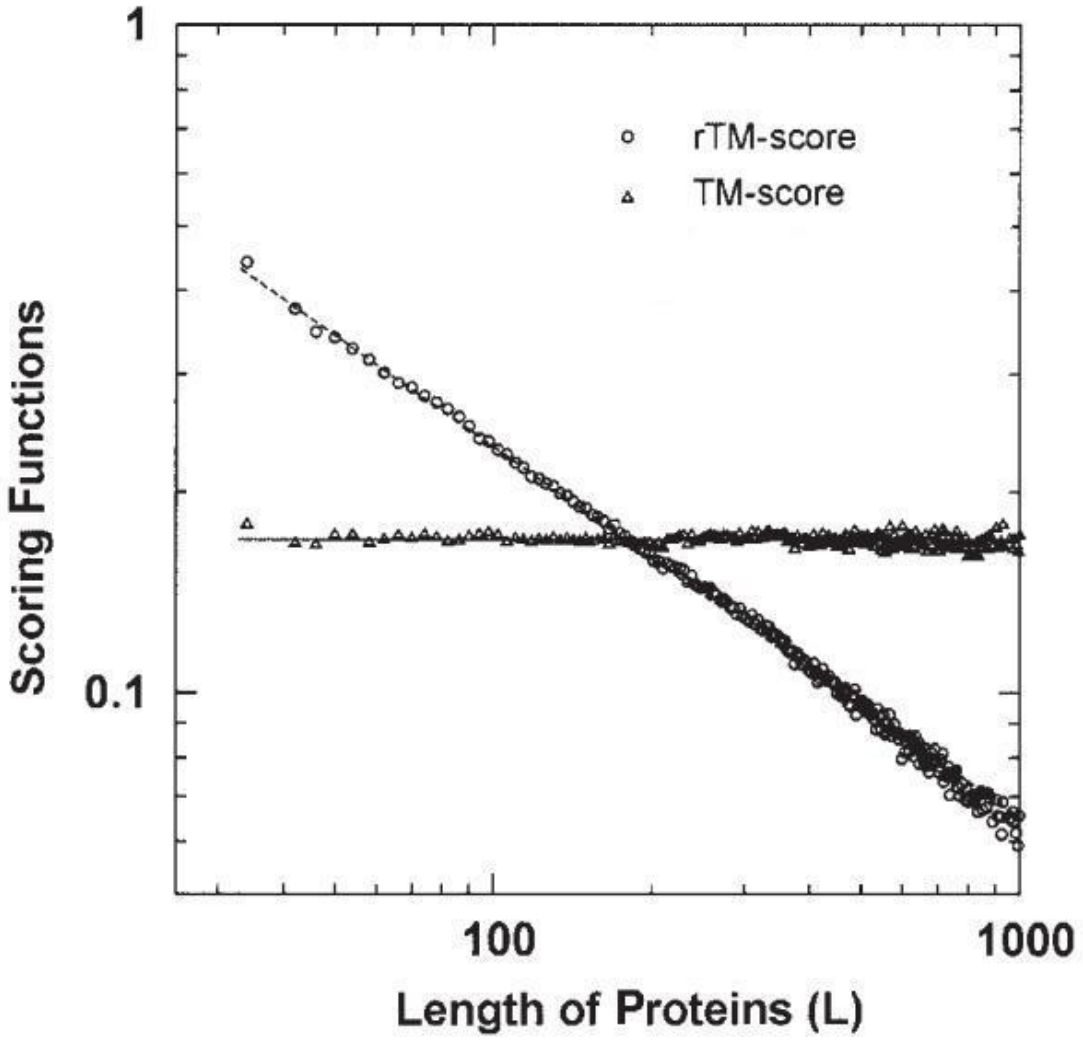

Figure 2.1: Performance of TM-score for different values of d_0. Scoring functions with raw value of d_0=5 (rTM-score) and $d_{0(Lmin)}$= $\mathbf{1.24\sqrt[3]{L_{min}-15}-1.8}$ (TM-score) are compared. The raw score is not able to discriminate between the random and good structural matches and it depends on the length of proteins. Figure is obtained from [2].

TM-scores of structural alignments range between 0 and 1. While a score ≥ 0.5 signifies the fold similarity between the target and template protein, an alignment score ≤ 0.17 is regarded as random alignment. Cutoff values defining degree of structural similarity are empirically derived.

2.2 Generation of template library using DOCKGROUND

We selected biological units as the source of templates which helped us to increase the diversity of templates. Asymmetric units, the conventionally deposited structures in PDB, are the smallest subunit of a protein crystal lattice that can be transformed to generate the unit cells of the protein crystal (however, asymmetric units do not necessarily correspond to the biologically functional forms). Along with PDB, there are other resources which offer biological units of proteins with second degree of annotations: ProtBuD [4], PQS [5] and DOCKGROUND [6, 7].

DOCKGROUND uses the symmetry operations suggested by the structure authors to generate the biological units. For such a method it is hard to discriminate between the real functional units and the crystal packing. In our case we decided to use biological units since we did not want to miss any template from the pool.

We generated libraries of interfaces where interface definition is based on the distance between any atoms across the interface. The X-ray resolution of the template structures has to be < 3 Å, structures have to come from at least a dimeric biological unit, and the sequence identity between different complexes has to be < 90%. The selection resulted in 11,932 complexes. The interface backbone atoms of the selected complexes were extracted and stored in the libraries of interfaces. Interface residue is defined as the one having at least one atom within a certain distance (varied from 6 to 16 Å) of any atom of the other protein in the complex.

2.3 Structure prediction protocol

As stated above we use TM-align to align the target proteins with the template proteins from our library. Not all alignments lead to the prediction of models. Figure 2.2 describes the flow of the template selection protocol, which tends to select the alignment with a certain degree of significance (defined in the next section).

The docking program is implemented in C and requires five command line arguments (receptor.pdb, ligand.pdb, path of the template library, alignment protocol FSA/PSA and number of top ranked model files as output). It makes its first call to function surface( ) which runs DSSP [8] and returns surface residues of target files in PDB format to the working folder. The second call goes to the TM-align program, which runs for each template in the library and returns the TM-score, alignment length, aligned residues and a transformation matrix. For each template, function surf( ) is called to decide the significance of the alignment. If the alignment is significant, function wrt( ) writes the information (template name, TM-scores, transformation matrix) to the output file. If the alignment is not significant, wrt( ) writes to the "log" file, describing the reasons of the failure. Then functions rtMAT( ) and rtPDB( ) are called to read the transformation matrix from the output file and generate the model complex file in the PDB format.

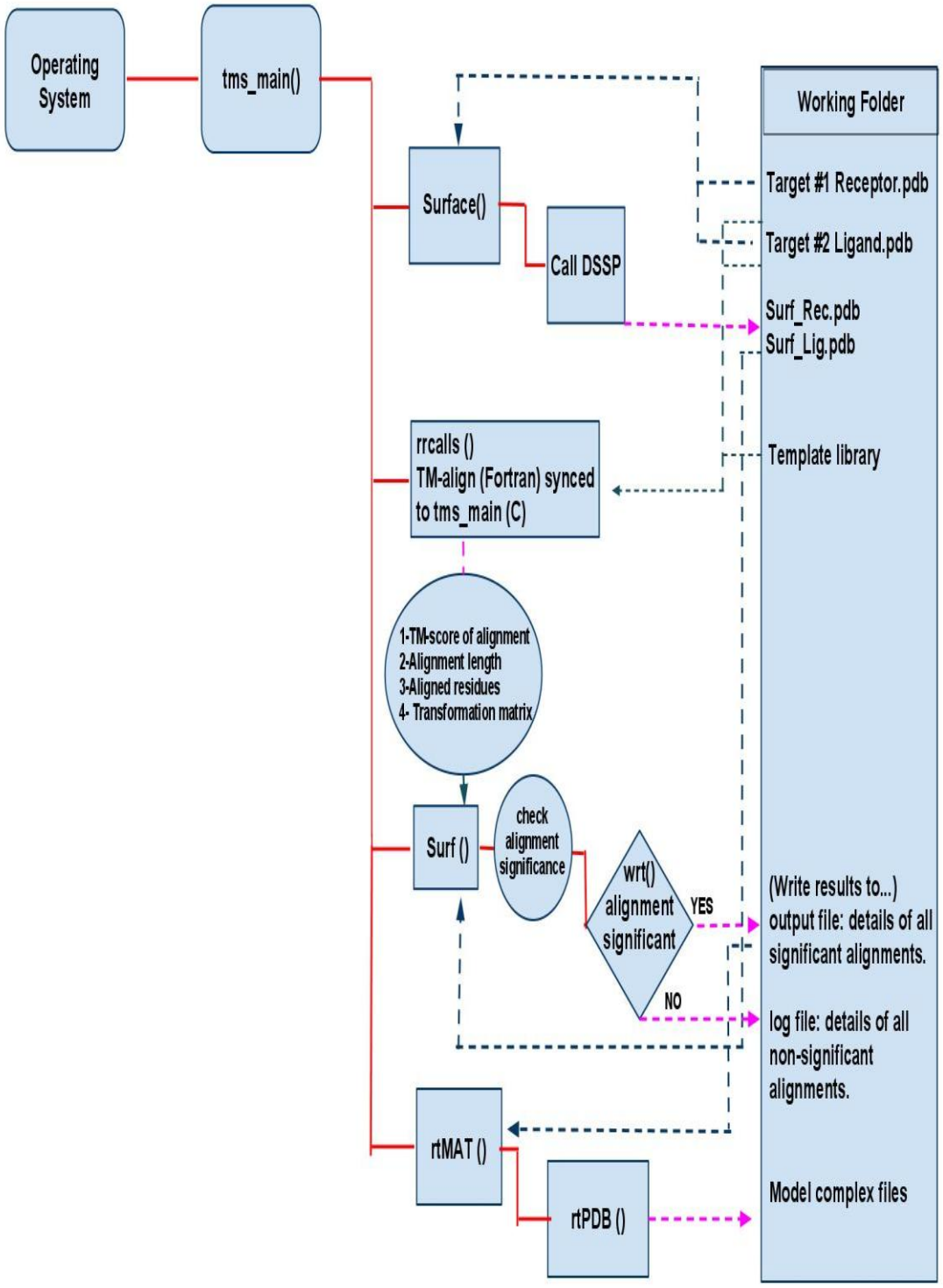

Figure 2.2: Flowchart of structure alignment and model prediction protocol.

2.4 Significance of the alignment

Structure alignment protocols tend to produce a model for each template in the library, so it is essential to discard the random alignments between the target and template proteins and retain only good quality matches. TM-score is adequate in characterizing degree of structural similarity but provides no information on the location of alignment (surface or core of target proteins). To avoid the structural clashes in the model complexes, alignments involving a significant amount of surface residues are selected for further processing. Following are the criteria used to call an alignment "significant".

An alignment is defined as significant when: (*i*) TM-score of at least one of the alignments is $\geq$ 0.4, (*ii*) at least 50% of the aligned residues (for both receptor and ligand) are on the protein surface, and (*iii*) at least 40% of the interface residues are aligned to target proteins.

2.5 Assessing the quality of model complexes

A significant alignment of template and target molecule structures, results in a putative model for the target protein complex. While benchmarking, it is essential to assess the quality of the models by comparing them to an already solved native complex. The quality of the resulting models are assessed by RMSD between ligand interface Cα atoms in the model and in the native complex (*i*-RMSD), based on the optimal alignment of the receptor structures (the larger molecules). The distance threshold for the interface residues in the *i*-RMSD calculations is 6 Å.

Analysis of the intermolecular energy funnels [9] suggests that the models with *i*-RMSD up to 8-10 Å can be locally minimized/refined to the near native structures. Therefore, in the present work a model with *i*-RMSD < 10 Å is considered acceptable.

The rank of a model complex is based on the sum of the alignment scores (TM-score) of the target monomers and template components.

2.6 Classification of the models

The resulting models are classified based on the parameters of the structural alignments between the target and the template monomers (Table 2.1). The alignments are performed on the entire structures of both the target and the template, rather than on the interface fragments used to generate the model. If the model is redundant with the template (Table 2.1) then it is considered as a self-match and not counted in the docking success rate (not evaluated).

Table 2.1: Classification of models.

Model class	TM-score	Alignment coverage, %	Sequence identity, %[a]
Redundant	0.9 – 1	80 – 100	95 – 100
Structural homolog	0.5 – 0.9	80 – 100	–
Partial structural homolog	0.5 – 0.9	0 – 80	–
Non-homolog	< 0.5	–	–

[a]Sequence identity by TM-align corresponding to the optimal structural alignment of proteins.

To compare the structure alignment methods with homology modeling, sequence identities between the template and target proteins are determined. The model complexes are classified on the basis of difficulty for homology modeling to detect the corresponding template: easy (sequence identities of both target-template pairs > 40%), medium (sequence identity of at least one target-template pair from 20% to 40%), and difficult (sequence identity of at least one target-template pair < 20%). The sequence alignments are performed using ClustalW [10].

2.7 Characterizing surface residues on the target proteins

We use the DSSP program to define the surface residues of the target proteins. It defines the surface residues on the basis of their accessible surface area (ASA). DSSP uses Lee & Richard's method [11] to find the ASA.

2.8 Benchmark sets used in the study

To validate the docking, we used the DOCKGROUND benchmark set containing 99 protein-protein complexes (27 enzyme-inhibitor, 6 antibody-antigen, 2 cytokine or hormone/receptors, and 64 other complexes), for which both monomers have both bound and unbound structures available (referred as DG99). To enhance statistical reliability of the results we also used an extended set of 372 non-redundant two chain bound complexes at 30% sequence identity level, extracted from DOCKGROUND (referred as DG372).

References

1. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Res, 2005. 33:2302-9.

2. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins-Structure Function and Bioinformatics, 2004. 57:702-10.

3. Betts, M.J. and M.J. Sternberg, *An analysis of conformational changes on protein-protein association: implications for predictive docking.* Protein Eng, 1999. 12:271-83.

4. Xu, Q., A. Canutescu, Z. Obradovic, and R.L. Dunbrack, Jr., *ProtBuD: a database of biological unit structures of protein families and superfamilies.* Bioinformatics, 2006. 22:2876-82.

5. Henrick, K. and J.M. Thornton, *PQS: a protein quaternary structure file server.* Trends Biochem Sci, 1998. 23:358-61.

6. Douguet, D., H.C. Chen, A. Tovchigrechko, and I.A. Vakser, *DOCKGROUND resource for studying protein-protein interfaces.* Bioinformatics, 2006. 22:2612-8.

7. Gao, Y., D. Douguet, A. Tovchigrechko, and I.A. Vakser, *DOCKGROUND system of databases for protein recognition studies: unbound structures for docking.* Proteins-Structure Function and Bioinformatics, 2007. 69:845-51.

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

9. Hunjan, J., A. Tovchigrechko, Y. Gao, and I.A. Vakser, *The size of the intermolecular energy funnel in protein-protein interactions.* Proteins-Structure Function and Bioinformatics, 2008. 72:344-52.

10. Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins, *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. 23:2947-8.

11. Lee, B. and F.M. Richards, *The interpretation of protein structures: estimation of static accessibility.* Journal of Molecular Biology, 1971. 55:379-400.

CHAPTER 3: PROTEIN DOCKING BY THE INTERFACE STRUCTURE SIMILARITY: HOW MUCH STRUCTURE IS NEEDED?

3.1 Research summary

Methodology described in this chapter is based on the structure alignment using protein interfaces as templates. The success of the approach by definition hinges on the way the interface is defined in terms of its structural content. A number of definitions of the interfaces are most often based on the change in solvent accessible surface area upon binding or on various types of distance cutoffs across the interface. Varying definitions significantly influence the size and the composition of the interfaces, thus having a major effect on the interface alignment. This chapter describes a systematic large-scale study to find the optimal definition/size of the interfaces for the structure alignment-based docking applications [1].

3.1.1 Structural description of protein interfaces

Defining interfaces for structural alignment based on the residues in direct physical contact only may lead to wrong results due to the loss of significant structural details at the interface. On the other hand, large distance cutoffs may impair the ability to find local structural similarity at the interface due to the presence of large non-interface parts (in the extreme case, the entire protein structure). Thus, selection of the cutoff distance for the interface definition in the context of the structural alignment can be considered as an optimization.

To find the optimal distance, we used five interface libraries with different

values of the distance: 6 Å, 8 Å, 10 Å, 12 Å and 16 Å (see Chapter 2 for details). Figure 3.1 shows an example of interface fragments in the 1bp3 complex corresponding to different cutoff distances. One can clearly see the gradual appearance of the secondary structure elements as the cutoff value increases. The interface of the first protein in the complex (blue ribbons in Figure 3.1) largely consists of two α-helices (residues G161–S184 and H18–Y28) interacting with β-sheet (β-strands W272-V279 and D291–V297) and loop fragments (residues Y240–M248, K385-W391, L202–I209 and P329–E366) from the second protein (red ribbons in Figure 3.1). However, the fragment from the 6 Å library (Figure 3.1A) contains only a short fragment (residues D171–I179) of one of the α-helices and the β-sheet structure of the second component is indiscernible with only short fragments (S270-T274 and E292-Y294) visible. Such representation is clearly inadequate for the successful structural alignment that involves secondary structure elements. The fragment from the 8 Å library (Figure 3.1B) has the longer α-helix (D171- R183) in the first protein and a visible β-sheet-like structure in the second component, but the second α-helix of the first protein still remains obscure. The fragment from the 10 Å library (Figure 3.1C) already shows one full α-helix of the first protein and the complete β-sheet structure of the second protein. Yet, the second α-helix from the first protein (residues Q22-D26) is only partially visible. Only the fragment from the 12 Å library reveals the complete structural details of the interface (Figure 3.1D). Further increasing the distance leads to the inclusion of significant non-interface parts of protein structure (the effect already seen in Figure 3.1C and D). A similar trend was observed in other interface library entries.

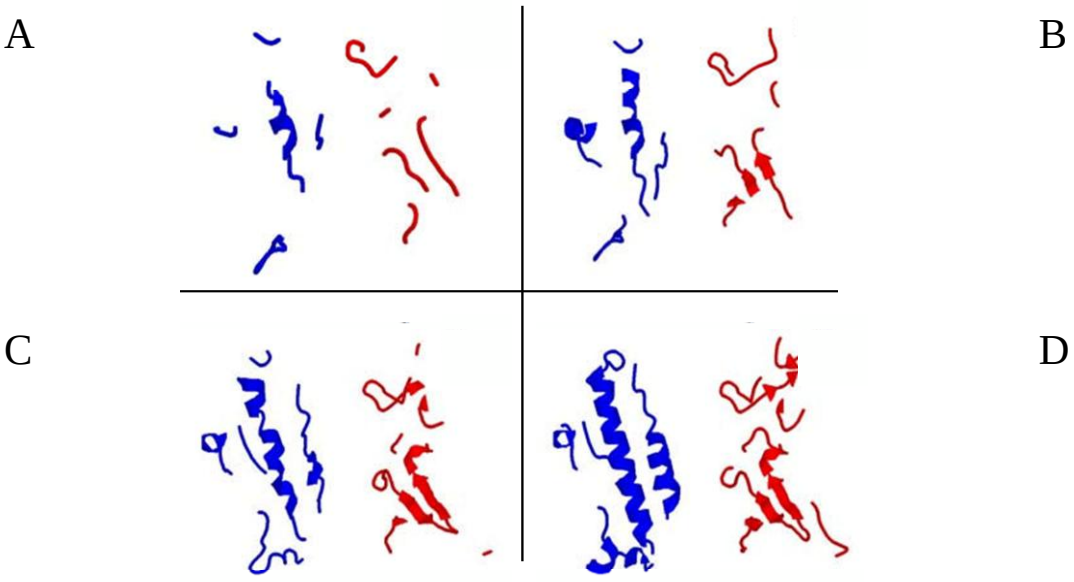

Figure 3.1: Example of interface fragments corresponding to different cutoff values. Fragments of 1bp3 complex were extracted using interface cutoffs: (A) 6 Å, (B) 8 Å, (C) 10 Å, and (D) 12 Å. Ligand (the smaller protein in the complex) is in blue and Receptor (the larger protein in the complex) is in red.

3.1.2 Structural alignment with interfaces

The modeling procedure (see details in Chapter 2) is applied to the libraries with different cutoff values. The Cα-only alignment was performed by TM-align [2]. For comparison, we also carried out structure alignment for several targets by another popular program SKA [3] and found no essential differences in the resulting models.

Structural deficiencies in the fragments from smaller cutoff libraries are reflected in the lower TM-score values for the alignments between such fragments and the target structures, thus substantially reducing the rank of the correct models. For example, 1bp3 complex (interface shown in Figure 3.1) is structurally homologous to a target complex 3hhr (TM-scores 0.8 and 0.7 for structural alignments of entire

1bp3 and 3hhr receptors and ligands, respectively, with corresponding sequence identities 31% and 66%). However, the 1bp3 interface fragment from the 6 Å library did not generate any models for the 3hhr target due to TM-scores that were below the statistical significance threshold (0.15 and 0.2 for the receptors and ligands, correspondingly). On the other hand, models generated using 1bp3 fragments from the 8 Å, 10 Å, 12 Å and 16 Å libraries had RMSD between ligand interface Cα atoms in the model and in the native complex (*i*-RMSD) 4.18 Å, 4.22 Å, 4.22 Å and 4.3 Å correspondingly. However, the 8 Å library model was ranked 42 among all 8 Å library models generated for this target, whereas model the ranked 1 had *i*-RMSD = 38.0 Å. Only models built using interface libraries with adequate structural details (10 Å, 12 Å and 16 Å libraries) were ranked 1 by the TM-score. Interestingly, a similar trend holds even for highly similar proteins. For example, the 1eay template complex is very similar to the target complex 1a0o (TM-scores 0.8 and 0.9 for structural alignments of the entire 1a0o and 1eay receptors and ligands respectively, with corresponding sequence identities 96% and 100%). However, 1eay interface fragments from the 6Å library could not generate statistically significant alignments for the 1a0o target (TM-scores 0.35 and 0.07). Models generated using the 1eay fragments from 8 Å, 10 Å, 12 Å and 16 Å libraries had *i*-RMSD = 1.5 Å, 1.7 Å, 2.0 Å and 2.2 Å, respectively. However 8 Å and 10 Å library models were ranked 818 and 35 respectively, whereas the 12 Å and 16 Å library models were ranked 5 and 1. Thus 12 Å and 16 Å libraries provided correct models for the 1a0o target within top 10 predictions. The *i*-RMSD values for the 12 Å and 16 Å libraries' models were similar to RMSD between the entire structures of bound 1eay and unbound 1a0o complexes (2.2 Å).

Relatively poor ranking of models from the small cutoff libraries was due to the fact that the small fragments lacking well defined secondary structure elements can be aligned to a random place in the target structure (thus generating models with high TM-score but large *i*-RMSD). At the same time, alignment of such fragments of a bound protein to the unbound target interface may have a significantly lower TM-score. This is especially true if there is a significant conformational change between bound and unbound structures. As shown in Figure 3.1, the distance of 12 Å and above provides full structural details of the interfaces. Thus, it reduces the possibility of the "good" random alignment and enhances the TM-score of the correct alignment by increasing parts of well aligned interface areas.

3.1.3 Modeling success rates for different interface libraries

To validate the docking, DG99 set was used [4] (see description in Chapter 2). The models were generated and evaluated using our five interface libraries. The results presented in Figure 3.2 are the success rates defined as a percentage of target complexes for which at least one model within a certain pool (top 10, top 100, and all models generated for the target) has *i*-RMSD ≤ 5, 8, and 10 Å. The *i*-RMSD ≤ 5 Å is comparable to the criteria for discriminating acceptable-quality models of protein-protein complexes in CAPRI [5]. Models with *i*-RMSD < 10 Å are considered acceptable in the present study.

The data in Figure 3.2 shows that the success rates for the 10 Å, 12 Å and 16 Å libraries are significantly higher than those for the 6 Å and 8 Å libraries (see discussion above). The 12 Å library models consistently had high success rates. In the

case of relaxed acceptance criteria for 16 Å library docking, the matches with *i*-RMSD ≤ 10 Å were in top 10 predictions, whereas models from the 12 Å library had ranks significantly worse than 10. This was the case for the 1he8 docking using 16 Å (model ranked 4 with *i*-RMSD 6.3 Å) and 12 Å (model ranked 19 with *i*-RMSD 6.0 Å) template fragments from 1k8r, and for the 2g45 docking using 16 Å template fragments from 1nbf (model ranked 4 with *i*-RMSD 9.5 Å) and 12 Å template fragments from 1tgz (model ranked 74 with *i*-RMSD 9.7 Å).

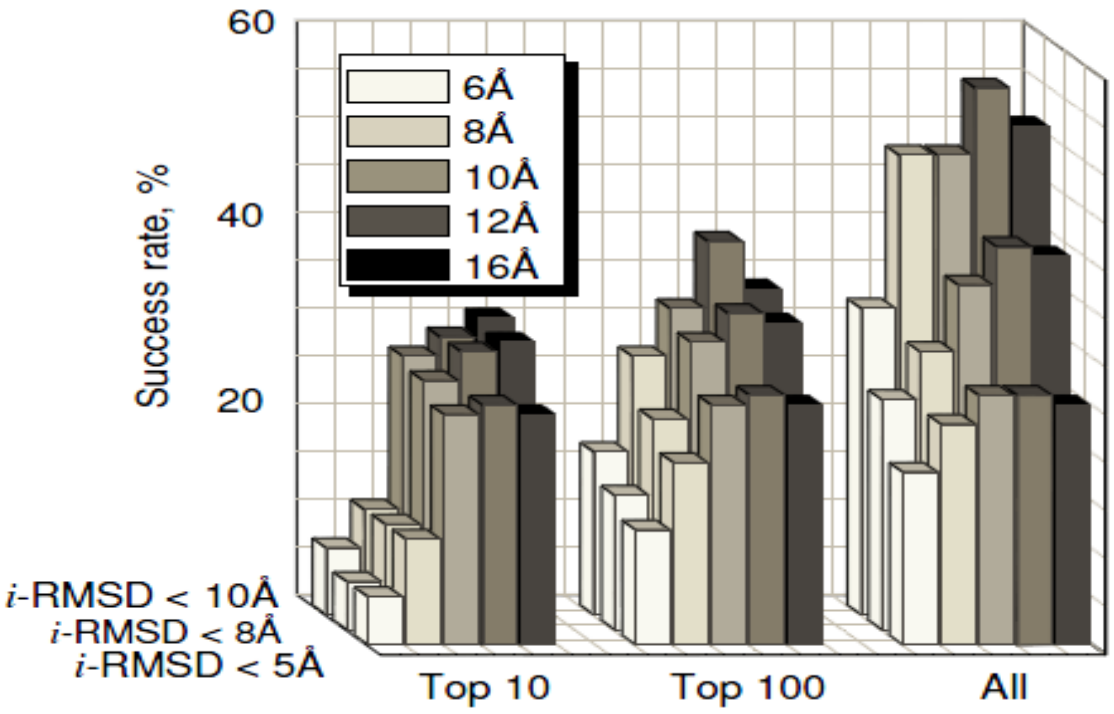

Figure 3.2: Docking success rates for different interface libraries. The docking was performed on the DG99 benchmark set. The success rate is defined as percentage of target complexes for which at least one match is in the top 10, top 100, and in all matches generated for the target and has *i*-RMSD ≤ 5, 8, and 10 Å. The results are shown for 6, 8, 10, 12, and 16 Å interface libraries [1] (see the text for details).

For some targets, the 16 Å library was unable to generate an acceptable model while the 12 Å library (smaller fragments) succeeded. An example of such a case is shown in Figure 3.3 where models for the ligand in 3sic were generated using ligand fragments from 1oyv. As the figure shows, the structures of 3sic and 1oyv ligands have dissimilar folds (TM-score for the alignment of the entire ligand structures is 0.7 with overall sequence identity 66%). The 3sic ligand is a trypsin inhibitor with the "classic" binding loop (residues E67-D76, marked 1 in Figure 3.3D). The secondary structure elements closest to this loop are α-helix and β-sheet (marked 2 and 3 in Figure 3.3D). The 12 Å library fragment from the 1ovy ligand (red ribbons in Figure 3.3C) contains a α-helix-like loop (residues T88-G93), which aligns well with the α-helix in the 3sic ligand (Figure 3.3A). The orientation of two other binding loops in the 1oyv ligand relative to this α-helix-like loop is similar to the relative orientations of the binding loop and α-helix in the 3sic ligand, yielding an accurate model for the 3sic target (*i*-RMSD 1.1 Å with rank 3). The 1oyv fragment from the 16 Å library (red ribbons in Figure 3.3E) contains a significant part of the non-interface β-sheet, which aligns with the β-sheet in the 3sic ligand (Figure 3.3B). Since orientations of these β-sheets relative to the binding site are different for the 3sic and 1oyv ligands, the resulting model has significantly larger *i*-RMSD = 7.0 Å. The model was not acceptable because more than 50% of the structural alignment contains non-surface residues of the target protein (this criterion is required to insure that the interface fragments do not align with the core of proteins producing random output). Increasing the distance cutoff defining the interface eventually leads to the inclusion of the entire monomer structures, thus transforming partial structural alignment into full structure alignment.

The detailed comparison of the partial (interface only) and the full protein structure alignment is discussed in the next two chapters. In the context of this chapter it is worth mentioning that the overall success rates there follow essentially the same trend as shown in Figure 3.2 for the 12 Å and 16 Å libraries, i.e. tend to decrease for the full-structure alignment models, especially with relaxed model acceptance criteria (larger *i*-RMSD and less demanding top ranking). Generally, the partial and the full structural alignments are applicable to different types of target/template similarity.

General utility of the docking approaches requires applicability to experimentally determined as well as modeled structures of monomers of limited accuracy, especially in large-scale (e.g., genome-wide) modeling of protein networks. Such approaches have to be fast (high-throughput) and tolerant to significant structural inaccuracies of the monomers [6]. Overall, the 12 Å cutoff appears to be optimal for the relaxed model acceptance criteria needed for docking of modeled structures. It also provides faster alignment than the one with larger cutoffs. Thus it is well suited for the high-throughput structural modeling of protein-protein complexes in large PPI networks.

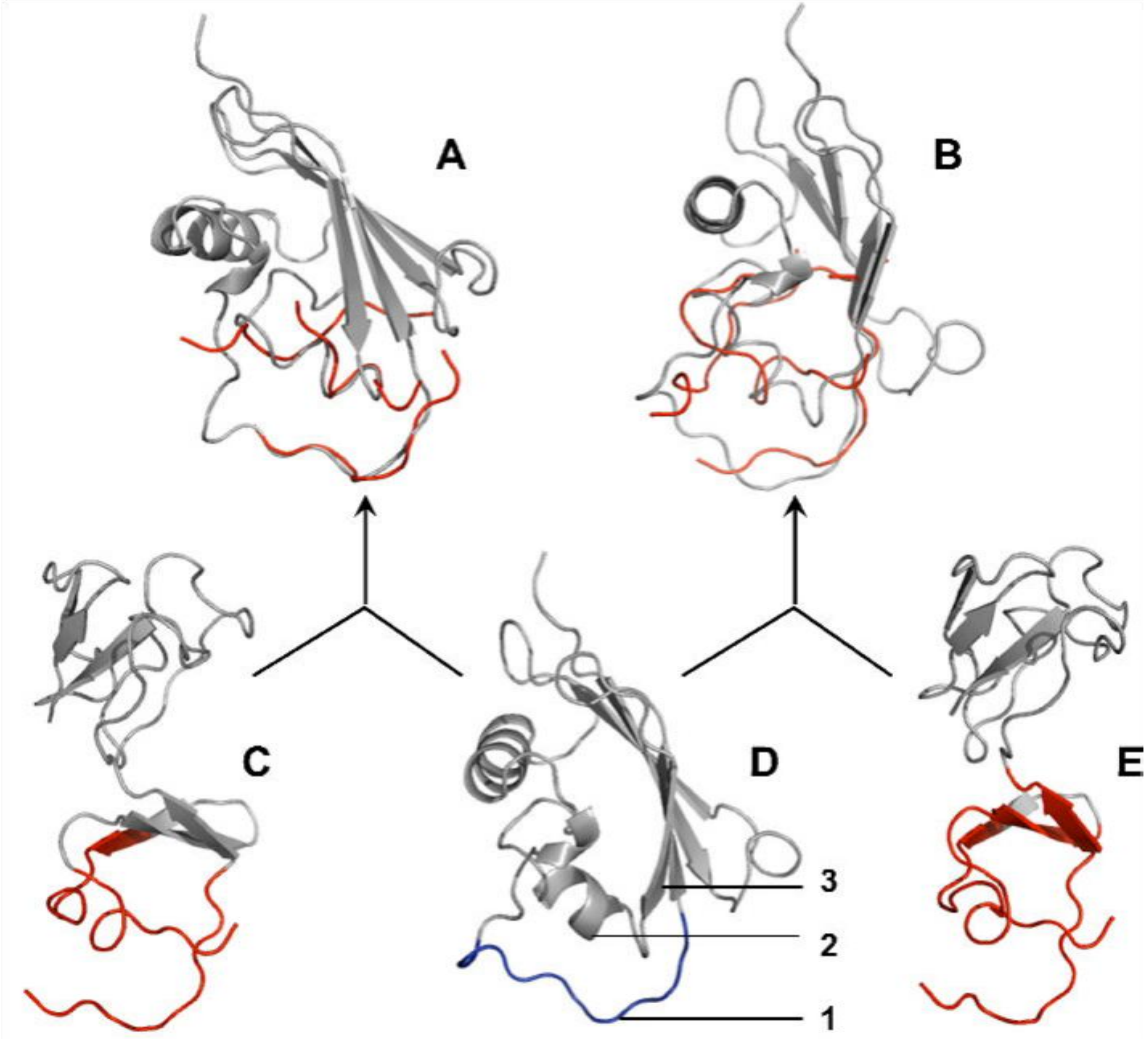

Figure 3.3: Example of docking based on 12 Å and 16 Å interface libraries. 3sic ligand (gray ribbons in A, B, D) was aligned with fragments of 1oyv ligand (red) extracted using 12 Å (A) and 16 Å (B) interface cutoffs. For comparison, the entire structure of 1oyv ligand is shown with 12 Å (C) and 16 Å (E) fragments (red). The entire structure of 3sic ligand with the loop participating in binding (blue) is shown in D. The binding loop in 3sic ligand is marked 1, and the α-helix and the β-sheet closest to this loop are marked 2 and 3, respectively.

References

1. Sinha, R., P.J. Kundrotas, and I.A. Vakser, *Protein docking by the interface structure similarity:How much structure is needed?* BMC Structural Biology 2011. Submitted.

2. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Research, 2005. 33:2302-9.

3. Petrey, D., Z. Xiang, C.L. Tang, L. Xie, M. Gimpelev, T. Mitros, C.S. Soto, S. Goldsmith-Fischman, A. Kernytsky, A. Schlessinger, I.Y. Koh, E. Alexov, and B. Honig, *Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling.* Proteins-Structure Function and Bioinformatics, 2003. 53 Suppl 6:430-5.

4. Gao, Y., D. Douguet, A. Tovchigrechko, and I.A. Vakser, *DOCKGROUND system of databases for protein recognition studies: unbound structures for docking.* Proteins-Structure Function and Bioinformatics, 2007. 69:845-51.

5. Lensink, M.F. and S.J. Wodak, *Docking and scoring protein interactions: CAPRI 2009.* Proteins-Structure Function and Bioinformatics, 2010. 78:3073-84.

6. Sali, A., R. Glaeser, T. Earnest, and W. Baumeister, *From words to literature in structural proteomics.* Nature, 2003. 422:216-25.

CHAPTER 4: DOCKING BY STRUCTURAL SIMILARITY AT PROTEIN-PROTEIN INTERFACES

4.1 Research summary

This chapter addresses the issues related to the development of docking through structure alignment. Structural similarity of proteins at varying degrees (global or interface) can be extrapolated to the similarity in their binding modes. Thus, the true potential of the structural alignment methods can be established through benchmarking the protocol at both local as well as global scales of structural similarity (FSA and PSA). At the same time a high-throughput application of the structure alignment method would ride on its ability to detect the templates hard to detect by sequence based methods (e.g. homology docking) which account for only a fraction of known PPIs.

In order to take into account the above, a systematic benchmarking and analysis of the interface alignment was performed on both DG99 and DG372 benchmark sets [1]. The performance was compared with FSA. The ability of the structure alignment method was assessed to extend the template space beyond detectable sequence similarity. Additionally, the present work also explored the idea of supplementing free docking protocol with the structure alignment method and measured their collective coverage of protein-protein complexes present in the benchmark sets [2].

4.1.1 Benchmarking global and local structural alignment methods

Both protocols (FSA and PSA) are systematically evaluated on the DG99 and DG372 benchmark sets. There are two categories of predicted models: (i) higher-accuracy models (*i*-RMSD ≤ 5Å) and (ii) lower-accuracy models (*i*-RMSD between 5-10 Å). Performances of both protocols are summarized in Table 4.1.

Both alignment protocols performed about equally well on both datasets for the higher-accuracy models. Significant parts of the datasets (42% and 56% of targets in the DG99 and DG372 datasets, respectively) had the best models produced by both protocols within the same accuracy range. The majority of the best FSA and PSA higher-accuracy models were built using the same template (Table 4.1, numbers in parenthesis for the common models). Thus, local structural similarity at the interfaces of target and template complexes is often accompanied by the global structural similarity between target and template monomers. However, a significant part of both datasets has the best model built by only one of the protocols.

In summary, the results show that the partial and the full structural alignment methods are complementary to each other and their combination significantly expands the number of identified templates for protein docking.

Table 4.1: Comparison of Full and Partial structure alignment.

Model *i*-RMSD	Number of targets modeled by					
	both PSA and FSA[a]		PSA only[b]		FSA only[b]	
	DG99	DG372	DG99	DG372	DG99	DG372
0 – 5 Å	26 (26)	130 (125)	0	13 (11)	2 (0)	15 (14)
5 – 10 Å	10 (4)	38 (2)	14	73	5	16

[a]Number of targets for which the best models produced by both partial structure alignment using the 12Å library (PSA) and full-structure alignment (FSA) protocols using the same (number in parentheses) or different templates have *i*-RMSD in a given accuracy range.

[b]Number of targets for which the best model produced by one of the protocols (PSA or FSA) has *i*-RMSD value in a given accuracy range, whereas the other protocol either yielded the best model (based on the same or different template) with *i*-RMSD value in a lower-accuracy range (number in parentheses) or failed to produce any statistically significant structure alignment for one or both target monomers.

4.1.2 Modeling protein complexes with "Partial Structure Alignment"

Out of 100 targets for both datasets for which the best model at all accuracy levels was built by PSA only, significant sequence identity (> 20%) between one pair of target-template monomers was observed in just 14 cases. An example is shown in Figure 4.1A for the target complex of bovine chymotrypsin with eglin C and the template complex of pig trypsin with its inhibitor. The receptors of both complexes have similar conformation (RMSD of aligned structures only 0.9Å) with 45% sequence identity. On the other hand, the ligands have only 5% sequence identity and are so structurally different that FSA did not produce a statistically significant model for this template (TM-score [3] of the global ligand alignment < 0.2). However, both

ligands share similar trypsin inhibitor-like loops that make up the entire ligand binding interface. Thus, in this case PSA produced an accurate model with *i*-RMSD = 1.3Å.

The remaining 86 PSA-only targets had sequence identity with the identified templates < 20% for both monomer pairs. An example is shown in Figure 4.1B for a PSA model of the complex between human cyclophilin and snRNP proteins built using an interface fragment between two chains (out of 4 identical chains in the asymmetric unit) of human transcription factor. The interface fragments used to build the model consisted of 71 and 89 residues for the template monomers, but the common structural motif (two short β-strands highlighted in magenta and red, Figure 4.1B) consists of only 4 residues for both the target and the template. Despite the significant difference in the shape of these β-strands, the PSA model has *i*-RMSD = 4.9Å. The overall structures of the target and the template are very different (with sequence identities 5% and 4% between receptors and ligands, respectively) and the FSA model for this target with the same template has *i*-RMSD = 37.0Å (*i*-RMSD = 6.8Å using a different template).

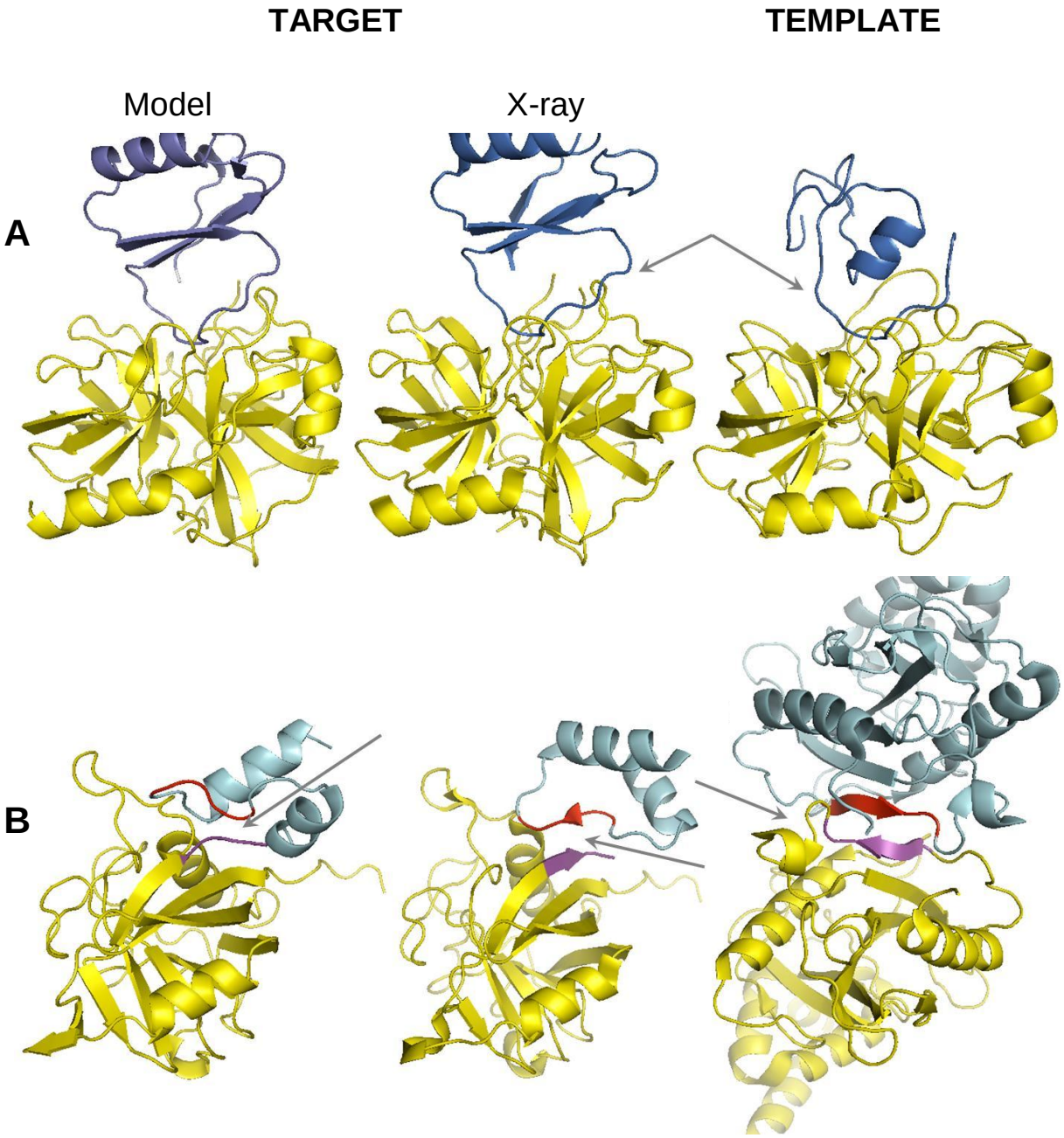

Figure 4.1: Examples of docking results by partial structural alignment. (A) Non-homologous ligands: target 1acb, chains E and I, and template 1ldt, chains T and L; match *i*-RMSD = 1.3Å. (B) Non-homologous receptors and ligands: target 1mzw, chains A and B, and template 1m1l, chains B and C; match *i*-RMSD = 4.9Å. Structural elements responsible for the alignment are in magenta and red and/or are indicated by arrows.

.

4.1.3 Performance of the model ranking scheme

Protein docking procedures need adequate scoring functions for the predicted matches. Here we did an analysis of the performance of our ranking scheme (see Chapter 2) for both FSA and PSA protocols. The results (see Supplementary data Table S1-S4) showed that for lower-accuracy models, the scoring function tends to assign low ranks to the near-native predictions generated by either PSA or FSA. Lower-accuracy models often have structural similarity only between interfaces of the target and the template, thus decreasing TM-scores of the entire monomer alignments (if any such alignment is found at all). At the same time FSA may find a template complex where one of the monomers is similar to the target monomer (TM-score close to 1.0), but binds a dissimilar protein at another binding site. This enhances the aggregate TM-score, bringing the incorrect model to the top of the prediction pool. A similar reason causes low ranking of the PSA models. In addition there are many small interface fragments in the template library which may align well (high TM-score) to non-interface parts of the target complex, thus decreasing the rank of the near-native PSA models even further than the corresponding FSA models. However, the situation is significantly different for higher-accuracy models, where not only the interfaces of the target and the template complexes are similar but often also the entire structures. Out of 143 targets, for which the best PSA models had *i*-RMSD < 5 Å, 108 predictions were ranked 1, and only 5 had rank below 1000. Among the 145 best FSA models with *i*-RMSD < 5 Å, 116 had rank 1, and no models were ranked below 1000. For 130 targets both protocols yielded the best models with *i*-RMSD < 5 Å and 125 of those models were built using the same template (same-template models). For 102 of those

targets, the best model was ranked 1 by both protocols. For the remaining 23 same-template models, ranking by PSA and FSA was the same in 5 cases, 10 PSA models had better ranking, and 8 FSA models had better ranking. Out of 5 common targets with different templates for the best PSA and FSA models, in one case (target 1f5q, chains A and B) the best model was ranked 1 by both protocols, in two cases PSA ranking was better, and in two cases FSA ranking was better. Thus, for ranking such models both methods perform equally well and placed the best models at the top of the prediction pool.

4.1.4 Structure and sequence homology

Structure alignment procedures are computationally demanding (although to a lesser extent than sophisticated multi-template modeling of individual proteins). Thus, for high-throughput structural modeling where computational speed is essential, it is necessary to understand how many of the structural alignment models can be obtained by a computationally less expensive homology docking approach. For this purpose, we performed the sequence based analysis of target-template proteins when acceptable models were produced (see Supplementary data Table S1-S4).

Distribution of higher-accuracy models at different levels of the homology docking complexity (Figure 4.2) showed that the easy cases make up a small part (9.4%) of DG372 dataset, whereas the majority of the models are medium (13.7%) and difficult (19.4%) cases. Interestingly, in a significant number of medium (22 models) and difficult (16 models) cases, the target and the template complexes corresponded to multi-binding proteins, where the same (or similar, with sequence

identity > 70%) protein binds dissimilar partners (with sequence identities corresponding to medium or difficult cases for the homology modeling) at the same binding site.

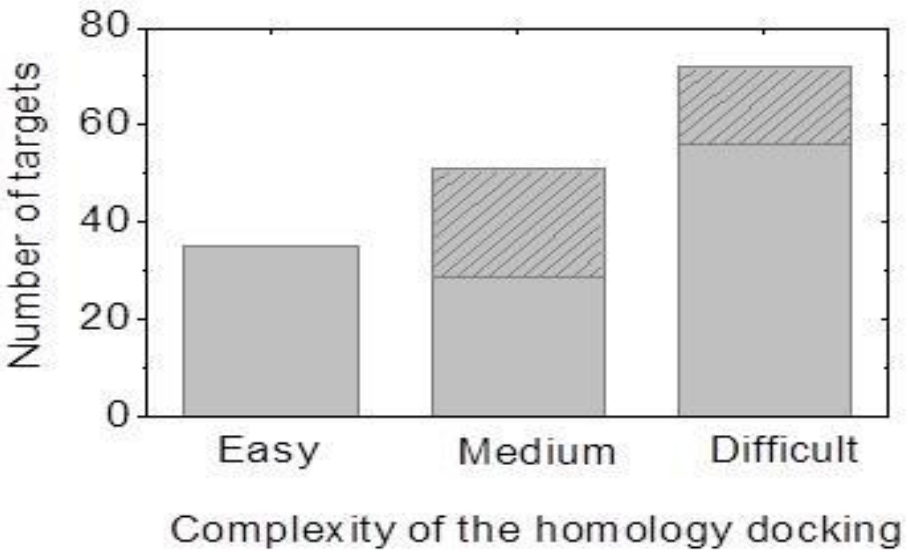

Figure 4.2: Success of structure alignment in terms of complexity for homology modeling. Numbers of targets in the DG372 dataset with higher-accuracy FSA and/or PSA models are shown for different levels of complexity for the homology docking. Dashed regions in the bars correspond to the number of targets with high sequence identity (larger than 70%) between one sequence pair.

Out of 127 lower-accuracy models, only 2 were of medium difficulty: (i) FSA model (6.9 Å *i*-RMSD) for the target 1fle (chains E and I) with the template 1eja (chains A and B) with the sequence identities 39% and 25% between receptors and ligands, correspondingly (note that PSA model for the same target with 5.6 Å *i*-RMSD was built using another template, chains A and I of the 1tx6 complex, with even lower

sequence identities, 39% and 15%, for receptors and ligands); and (ii) FSA (7.3 Å *i*-RMSD) and PSA (5.8 Å *i*-RMSD) models for the target 1g3n (chains A and C) with the template 1f5q (chains A and B) with sequence identities 45% and 22% for the receptors and ligands. All other FSA and PSA lower-accuracy models were difficult cases for the homology docking, with sequence identities as low as 2% in some cases. However TM-scores even for such low sequence identities indicate significant structural similarity between the target and the template (see Supplementary data Tables S1 and S2).

4.1.5 Comparison to free docking

As shown above, the structural alignment is a useful tool in finding templates hardly detectable by fast sequence based methods. On the other hand, it is important to understand where the structural alignment stands with respect to the well-established and widely used free docking techniques. Since the docking techniques are usually tested on the set of unbound structures, we compared the performance of PSA and the free docking GRAMM-X server [4] on the DG99 unbound set.

GRAMM-X is a protein-protein docking web-server derived from original GRAMM [5]. It performs FFT based global search followed by refinement and rescoring through multiple knowledge-based potentials.

The results are shown in Figure 4.3. A significant part of the targets successfully docked by GRAMM-X was modeled by PSA as well, in the case of both higher- and lower-accuracy models (60% and 71% of all successful free docking

models for higher- and lower-accuracy models, respectively). In turn, PSA produced 14 higher-accuracy and 4 lower-accuracy models for targets where GRAMM-X failed in any acceptable-accuracy docking.

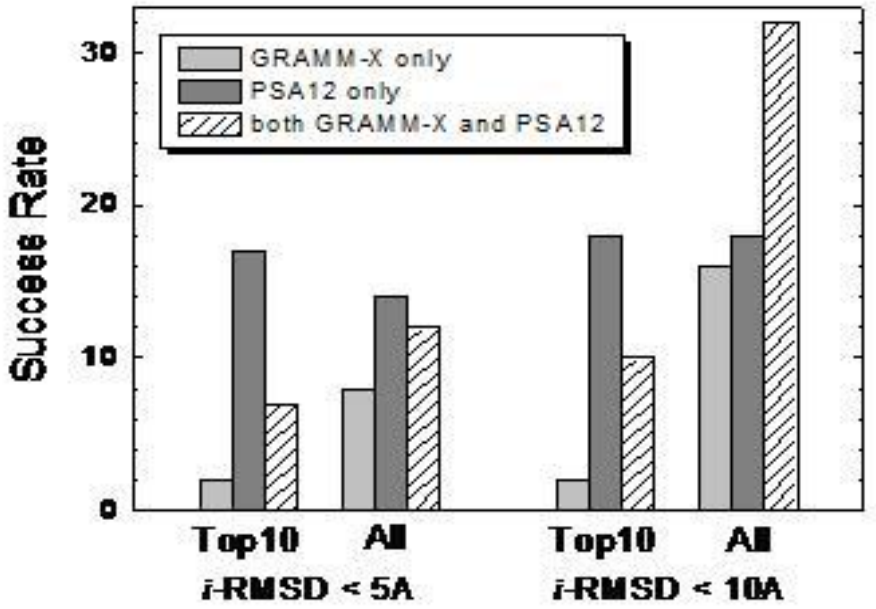

Figure 4.3: Comparison of the success rates in template-based and free docking. The success rates are defined as the percentage of targets in DG99 unbound dataset for which higher-accuracy only (*i*-RMSD < 5Å) and all acceptable (*i*-RMSD < 10Å) models were produced by free docking only (GRAMM-X), template-based only (PSA), and both.

The structure alignment approach was also tested on previous Critical Assessment of Prediction of Interactions (CAPRI) [6] targets, with limited success, which is in sharp contrast with the significantly higher success rate for the docking benchmark sets. The obvious reason is that the CAPRI targets are usually hand-picked to avoid, with few exceptions, close homologies with co-crystallized complexes (needed as templates for structural alignment). However, for a typical biological

problem, the existence of homologous co-crystallized complexes, of course, is not to be avoided but welcomed. Thus, in this respect the docking benchmarks, which do not preclude the increasingly available co-crystallized homologous complexes, are more representative of the ‘real world’ biology.

The structural alignment algorithm is generally more reliable than the free docking methodology. Its utility is increasing with more structural templates being determined by crystallography and NMR. Thus the emerging docking strategy should involve a search for available docking templates prior to the free docking modeling. This paradigm is especially valid in genome-wide high-throughput modeling, where most structures of the monomers will be models with structural accuracy lower than that obtained by the X-ray/NMR.

References

1. Gao, Y., D. Douguet, A. Tovchigrechko, and I.A. Vakser, *DOCKGROUND system of databases for protein recognition studies: Unbound structures for docking.* Proteins, 2007. 69:845-851.

2. Sinha, R., P.J. Kundrotas, and I.A. Vakser, *Docking by structural similarity at protein-protein interfaces.* Proteins, 2010. 78:3235-41.

3. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins-Structure Function and Bioinformatics, 2004. 57:702-710.

4. Tovchigrechko, A. and I.A. Vakser, *GRAMM-X public web server for protein–protein docking.* Nucleic Acids Res., 2006. 34:W310-W314.

5. Vakser, I.A., *Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex.* Proteins, 1997. Suppl 1:226-30.

6. Janin, J., *Assessing predictions of protein-protein interaction: the CAPRI experiment.* Protein Sci, 2005. 14:278-83.

CHAPTER 5: GLOBAL AND LOCAL STRUCTURAL SIMILARITY IN PROTEIN-PROTEIN COMPLEXES

5.1 Research summary

Chapter 4 described our efforts to benchmark structure alignment protocol on the scale of both local as well as global fold similarity (FSA and PSA). It showed that both protocols provide a significant degree of success in modeling protein complexes.

Comparable successes of FSA and PSA protocols for higher-accuracy models and higher success of PSA in modeling lower-accuracy complexes raises the challenge to determine the extent of structural conservation in the protein-protein complexes. Thus, the goal of this chapter is to understand how frequently interface similarity of two proteins is *not* extended to their global fold similarity.

Here we addressed this fundamental issue by modeling 372 protein complexes by full and partial structural alignment and analyzing the results in terms of the degree of structural similarity between the target and the template complexes and its impact on the quality of the model complexes [1].

Model complexes were classified into the following three categories:

(1) Complexes with both full and local structure similarities

(2) Complexes with only local structure similarity

(3) Complexes with only full structure similarity

5.1.1 Complexes with both full and local structure similarities

We compared models for 372 protein complexes (see Chapter 2 for structure generation protocol and test set) built by PSA with the corresponding models obtained by the FSA. The comparison is summarized in Chapter 4, Table 4.1.

For significant parts of the dataset (126 targets or 34%) the structural similarity between the target and the template is not only substantial for the interface but also for the entire structure. However, most of the PSA models, belonging to this group, have systematically lower *i*-RMSD values compared to the corresponding FSA models (see Supplementary data Table S3 and S4). In total, there are 92 such models, out of which 17 have *i*-RMSD differences > 1 Å. Only in 19 cases FSA model has a lower *i*-RMSD compared to the corresponding PSA model (in 4 cases the differences are > 1 Å). This implies that structures of the protein-protein interfaces tend to be more conserved compared to the rest of the proteins, which correlates with the previous observations of higher sequence conservation at the protein-protein interfaces [2-4]. As discussed in Chapter 4, the majority of these models are either medium or difficult cases for sequence based methods.

The advantage of PSA is discussed here through the following two examples: The first example is illustrated in Figure 5.1 for the models of subtilisin BPN from *Bacilus amyloliquefaciens* complex with synthetic protein (chains L and R from 3sic), modeled on subtilisin Carlsberg from *Bacilus lichaniformis* complex with ovomucoid protein from *Meleagris gallopavo* (chains R and L from 1r0r). Both subtilisins have similar global structures with high sequence identity (70%). Thus their FSA and PSA

alignments are similar too (Figure 5.1A and B). However, the aligned sequences of the inhibitors have only 12% sequence identity. Only the "classic" inhibitor loops are similar, whereas the rest of the structures are quite different (yellow and magenta ribbons in Figure 5.1C). Thus, PSA correctly aligns the interface parts of the target and the template (Figure 5.1D) yielding an accurate model with only 0.9 Å *i*-RMSD. FSA seeks to find the minimal distance between all Cα atoms of the target and the template. Thus the alignment of the interface loops becomes less accurate (Figure 5.1C) and resulting model has 4.9 Å *i*-RMSD.

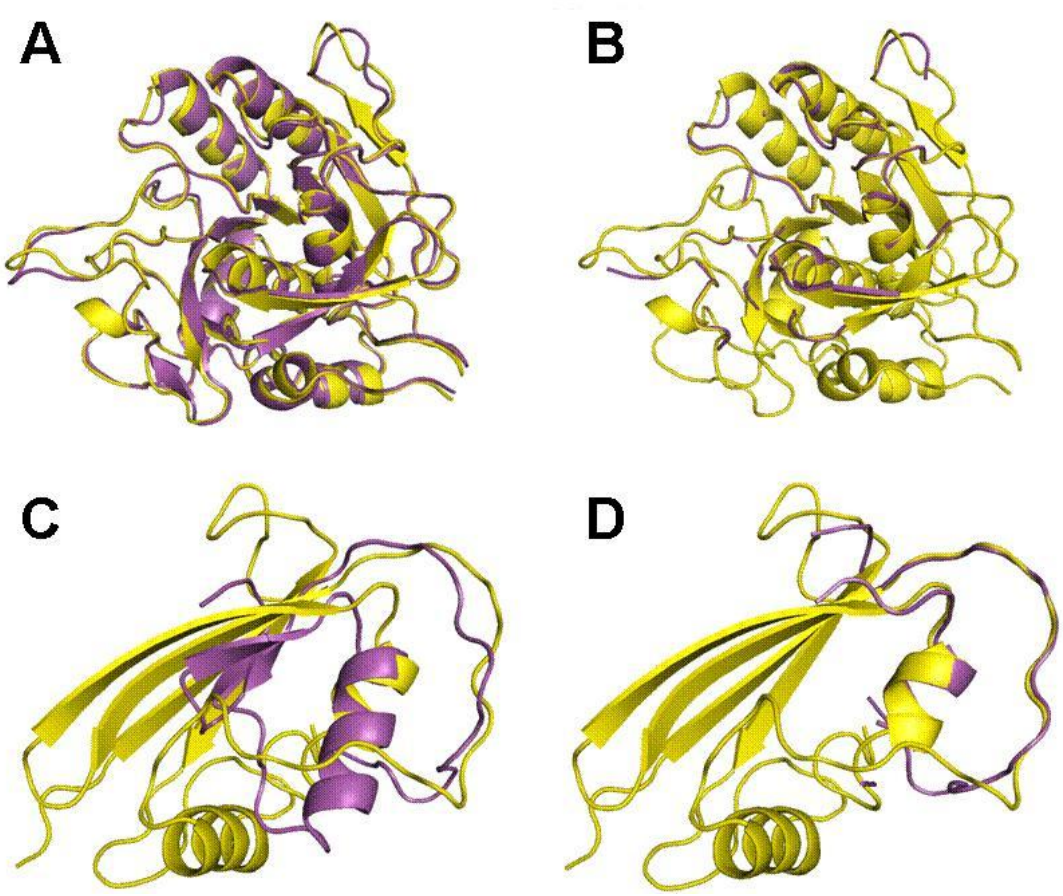

Figure 5.1: Example (#1) of the local alignment more accurate than the full alignment. FSA (A and C) and PSA (B and D) alignments between target 3sic (in yellow) and template 1r0r (in magenta) complexes. The alignments of the receptors (chains E of the 3sic and 1r0r) are shown in A and B, and the alignments of the ligands (chain I) are shown in C and D.

The second example is illustrated in Figure 5.2 for the models of human signaling complex (chains B and A from 1ki1), built on another human signaling complex (chains A and B from 2nz8). Ligands of both the target and the template share near identical overall structure with high 78% sequence identity (Figure 5.2C). Receptors of both the target and the template have clearly distinguishable two-domain structures, with only one of the domains participating in the binding. The structures of separate domains are very similar (although with low 18% sequence identity), but their

orientation in the target and the template is different (yellow and magenta ribbons in Figure 5.2A). Thus FSA yielded a model with 5.0 Å *i*-RMSD. PSA correctly aligned the interface parts of the target and the template (Figure 5.2B) producing a model with 0.6 Å *i*-RMSD. However, such extreme cases are not very common in our dataset; they were observed only in 5 targets with higher-accuracy models.

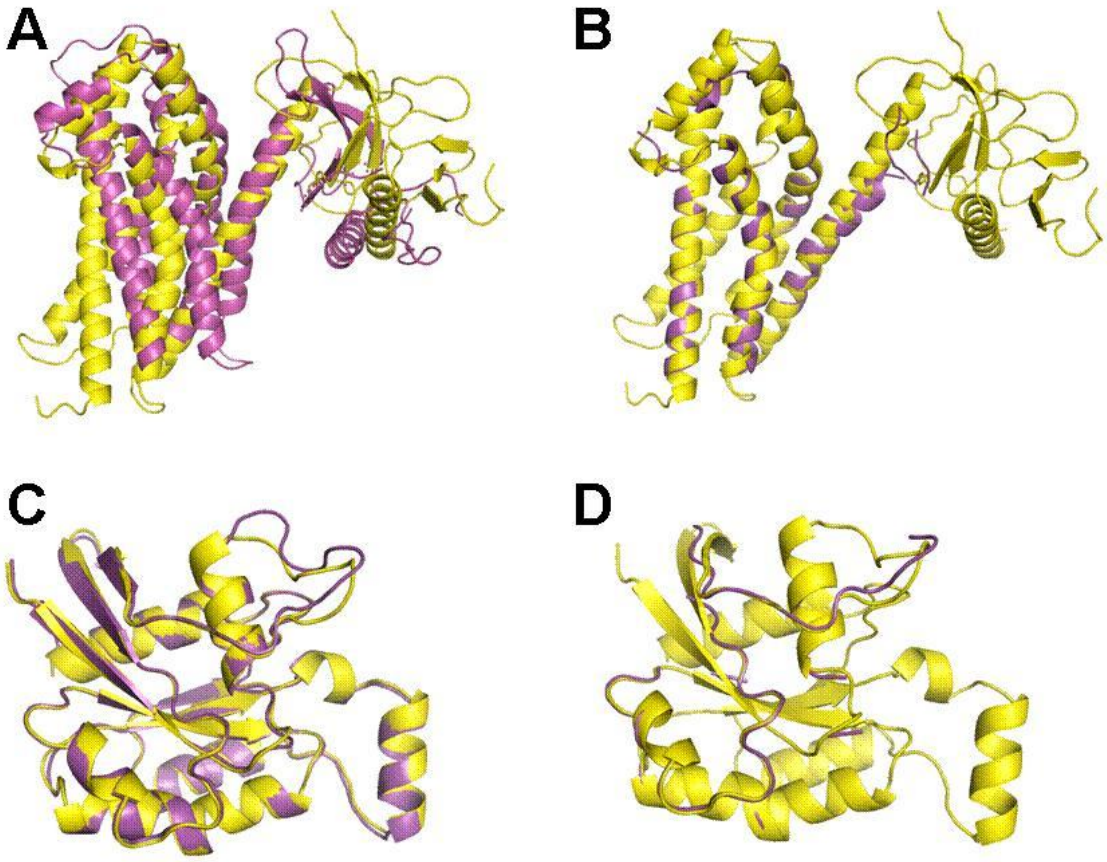

Figure 5.2: Example (#2) of local alignment more accurate than the full alignment. FSA (A and C) and PSA (B and D) alignments between target 1ki1 (in yellow) and template 2nz8 (in magenta) complexes. The alignments of the receptors (chains B of the 1ki1 and 2nz8) are displayed in A and B, and the alignment of ligands (chains A) are shown in C and D.

Similar structures of one of the target and the template monomers accompanied by dissimilar structures of the other monomers are a common feature of all higher-accuracy PSA models. Thus, if it is known that a protein binds different proteins at the same binding site (e.g., above enzyme-inhibitor complexes), the PSA is a better alternative.

5.1.2 Complexes with only local structure similarity

For the targets with lower-accuracy models (5 Å < *i* -RMSD ≤ 10 Å) the interface-only conservation was even more prominent. PSA produced models for a significant part of the dataset (73 PSA-only targets, 19.6%) while FSA failed to yield any model of reasonable accuracy. Similar structural fragments may involve a small part of the interface, as illustrated by the PSA model (Figure 5.3A) of mice protein signaling complex (1vet) built based on interfacial fragments between two chains of RUVA protein from *E. coli* (4otc, Figure 5.3C). The interface fragments used to build the model consist of 45 and 53 residues for template monomers however; the common structural motif consists of two short β-strands (in magenta and red in Figure 5.3). The shape of these β-strands differs slightly in the target and the template X-ray structures (Figure 5.3B and C), thus the PSA model has 6.0 Å *i*-RMSD (Figure 5.3A) due to the wrong tilt of the ligand. The overall structures of the target and template are so different (with sequence identities 4% and 3% between receptors and ligands, respectively) that FSA failed to produce any statistically significant models for this target.

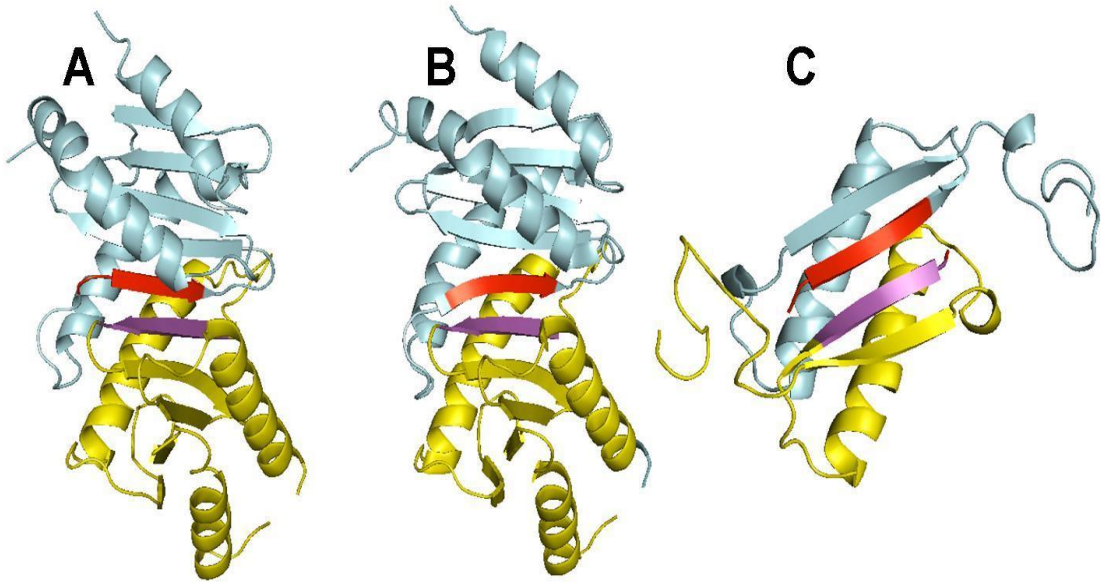

Figure 5.3: Local alignment on a small part of the interface. (A) Model and (B) X-ray structure of the target complex (1vet, chains A and B), and (C) X-ray structure of the template complex (4otc, chains B and C). Receptors are in yellow and ligands are in blue. Parts of the structures responsible for a near native PSA model are shown for receptors (in magenta) and for ligands (in red).

Interestingly, the majority of the PSA-only targets (67 targets) were modeled using homo-dimeric template complexes, primarily from different organisms. Only one template for higher-accuracy models and three templates for lower-accuracy models were from the same species. Three templates for lower-accuracy models shared a common organism with the target for one of the monomers. In 14 cases (two higher-accuracy and twelve lower-accuracy models) the interfaces of the homo-dimeric templates were present only in biological units built from the asymmetric units (often a single protein chain) in the PDB entries using translational/rotational matrices (in all cases templates are from the different organisms). Moreover,

sometimes a biological interface was modeled using similarity with the crystal packing interface as shown in Figure 5.4 for the complex of colicin E3 with its immunity protein (Figure 5.4B). PSA yielded the best model for this complex based on the X-ray structure of colicin E3 homo-dimer (Figure 5.4C). The biological function of colicin is to kill excess *E. coli* cells by binding and cleaving the enemy cell DNA. To prevent the host cell suicide the colicins form complexes with their immunity proteins inhibiting the DNA binding site [5]. In either case colicins do not exist *in vivo* as homo-dimers. The colicin E3 and its immunity protein are quite dissimilar (19% sequence identity and TM-score for the alignment of entire structures < 0.2). Thus FSA failed to produce a statistically significant model while PSA produced a lower-accuracy model with 7.3 Å *i*-RMSD (Figure 5.4A).

Because of the absence of unambiguous criteria for distinguishing biological and crystallographic interfaces it is hard to provide the exact number of such cases. In general, the results correlate with the conclusions of the recent study [6] that only localized regions on protein-protein interfaces are conserved among structural neighbors.

5.1.3 Complexes with only full structure similarity

A significant part of the dataset (31 targets or 8.3%) was modeled by the FSA protocol only (see Chapter 4, Table 4.1). Analysis of those models revealed three main causes for the worse PSA performance (or its complete failure). The first reason is related to differences in length of interface loop(s) connecting the otherwise similar interface β-strands in the target and the template (in total, 7 such cases in the dataset).

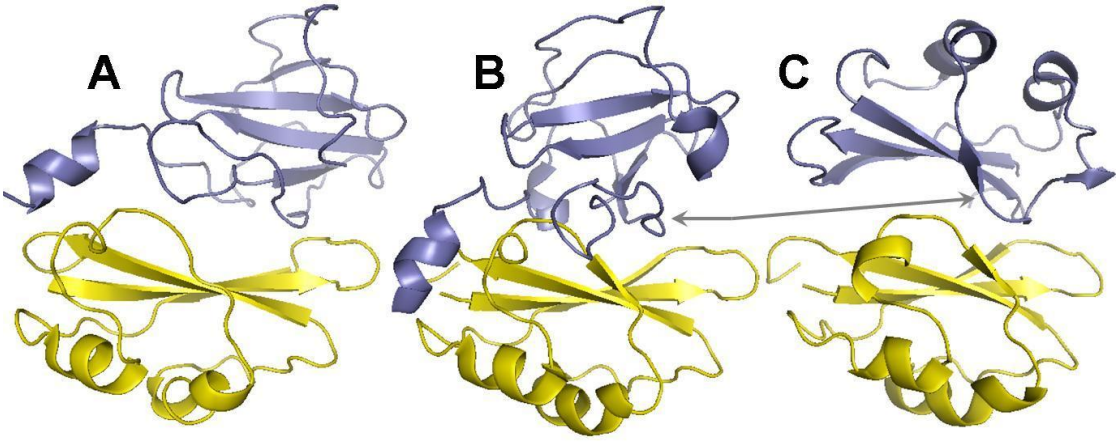

Figure 5.4: Local alignment on a crystal packing interface. (A) Model and (B) X-ray structures of the target complex (1e44, chains A and B), and (C) X-ray structure of template complex (3eip, chains A and B). Receptors are in yellow and ligands are in blue. Arrows indicate parts of the structures responsible for the near-native PSA model.

This leads to a shift in the alignment of the structural fragments. Thus PSA, while still capable of building a near-native model based on the same or different template, yields a model in the lower-accuracy range compared to the FSA model, where the entire structure ensures the alignment of correct parts of the interface β-strands. Figure 5.5 shows an example of target 1itb (ligand complex with human interleukin-1 beta) and template 1cvs (ligand complex with human fibroblast growth factor 2). Overall the ligand structure of the target (yellow and magenta ribbons in Figure 5.5B) and the template (gray and white ribbons in Figure 5.5A) are quite similar. Thus FSA protocol correctly aligns the full structures (Figure 5.5D) yielding the best model with 4.8 Å *i*-RMSD. Both ligands belong to the cytokine superfamily in SCOP [7] classification. However sequence identity between the ligands and receptors is 15% and 14% respectively, which makes it a difficult case for homology modeling. The main difference is in the length of the interface loop connecting two β-strands that are

partially at the interface (magenta and white ribbons in Figure 5.5 for target and template, respectively). This loop is longer and the interface part of the β-strand is shorter in the target structure. Thus PSA aligns the wrong loop and strands parts (Figure 5.5C), generating the best model with 7.3 Å *i*-RMSD.

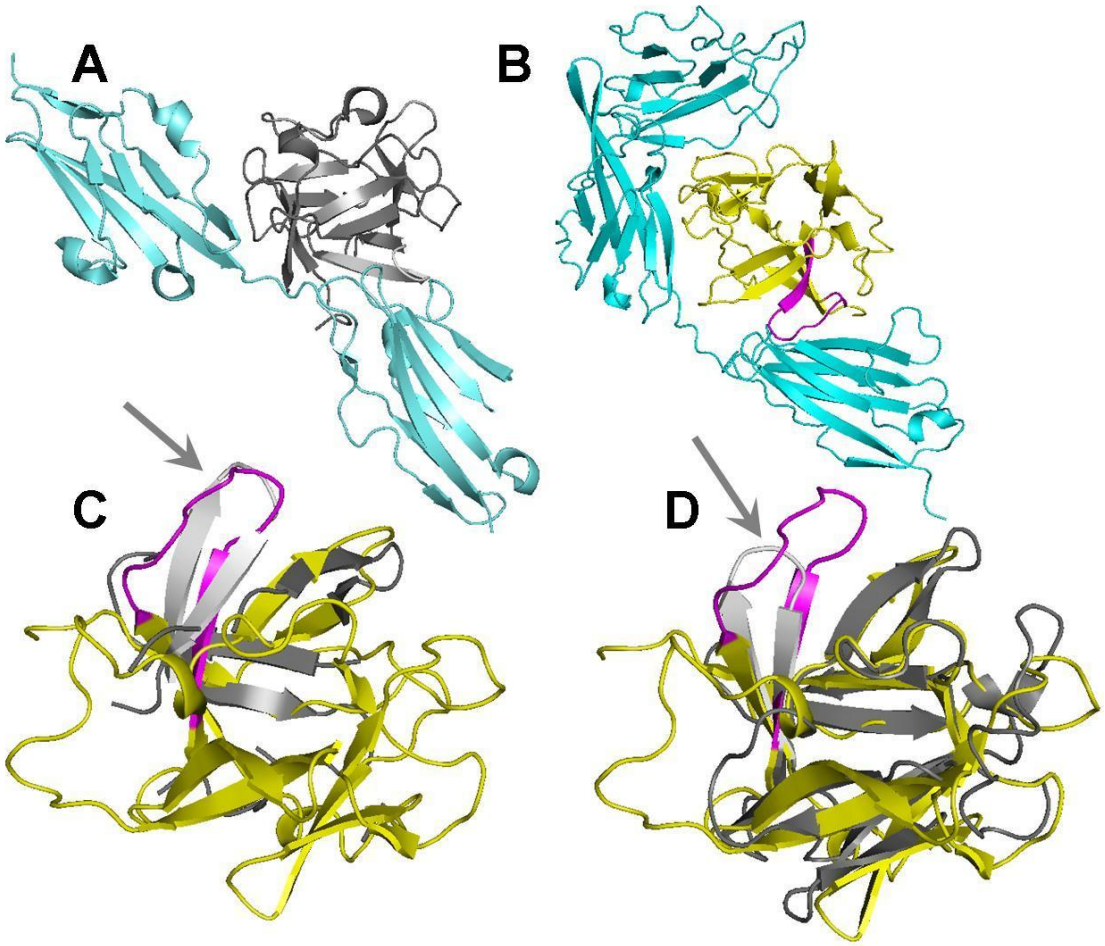

Figure 5.5: Example (#1) of the full alignment more accurate than the local alignment. (A) The X-ray structures of template (1cvs, chains A and C) and (B) the target (1itb) complexes, along with (C) PSA and (D) FSA alignments of the target ligand. The receptors are in cyan while ligands for the target and template are in gray and yellow, respectively. Arrows indicate parts of ligand β-strands essential for the model building, highlighted in magenta and white for the target and template.

The second source for the PSA failure stems from the presence of the four-helix bundle structure motif in the target and the template monomers where only parts of the helices participate in binding. In such cases the interface helix fragments from the template are aligned to a random place on the target helices resulting in a wrong model, whereas the FSA protocol correctly aligns the entire helix bundles. Figure 5.6 illustrates such a case of target 1f6f (ligand complex with *Ovis aries* placental lactogen Figure 5.6B) and template 1pvh (ligand complex with human leukemia inhibitor factor, Figure 5.6A). Both monomers have α-helical structures and belong to the same long-chain cytokines SCOP family with the sequence identity between them only 7%. The overall structure of these monomers is very similar (see the superimposed structures in Figure 5.6F), resulting in the best FSA model (Figure 5.6C) with 4.5 Å *i*-RMSD. However, PSA aligns the interface parts of the template helices (white ribbons in Figure 5.6) to non-interfacial parts of the target helices (magenta and yellow ribbons in Figure 5.6) producing an incorrect model with 15.0 Å *i*-RMSD (Figure 5.6E). PSA was capable of producing the best model with 7.8 Å *i*-RMSD based on another template structure (2aux).

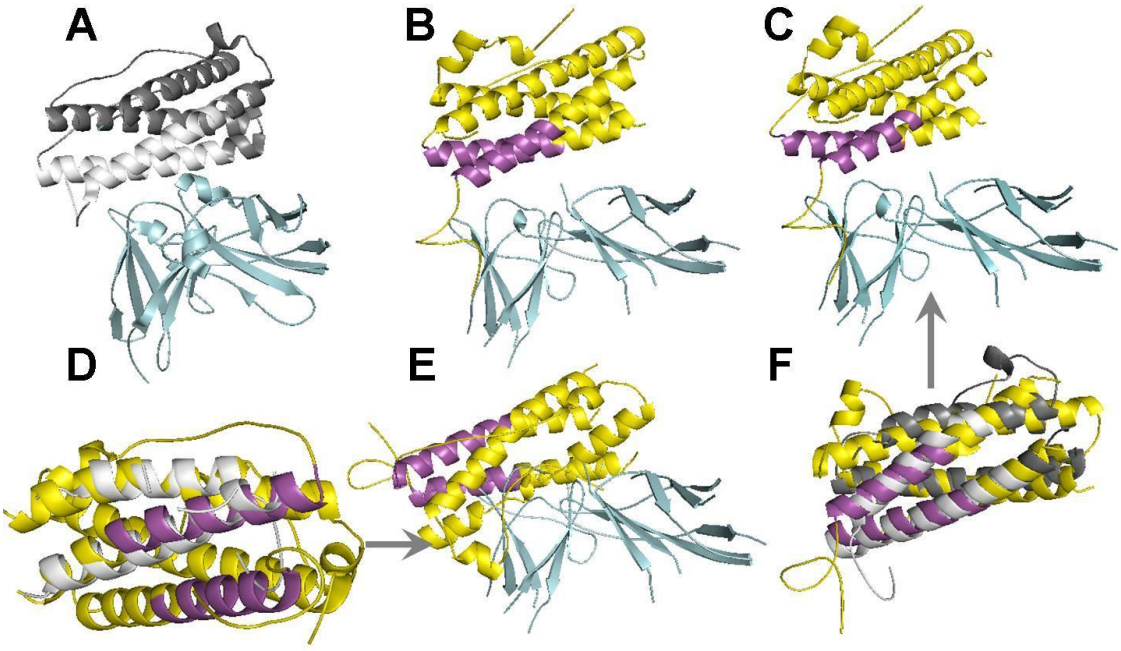

Figure 5.6: Example (#2) of the full alignment more accurate than the local alignment. (A) X-ray structures of the template (1pvh) and (B) target (1f6f) complexes along with (C) FSA model of the target complex and (F) FSA-alignment of the target ligand. (D) PSA-alignment of the target ligand with (E) the PSA model of the target complex. The receptors are in cyan while ligands for the target and template are in gray and yellow, respectively. Interfacial parts of ligand helices are highlighted for the target (in magenta) and template (in white).

In the third group of the FSA-only targets, there is a *local* structural similarity between the target and the template *away* from the interface. These similar pieces are not large enough to produce higher-accuracy FSA models, but sufficient to dominate FSA alignments, thus correctly orienting the target monomers. The sequence identities between the target and the template monomers in all such cases were < 10%, implying that such templates are hardly detectable by ordinary sequence-homology algorithms. Due to the absence of structural similarities between the target and the template interfaces, PSA yields the near-native model with substantially higher *i*-RMSD or no

near-native model at all. An example is shown in Figure 5.7 for the complex of Colicin D with its immunity protein (chains A and B in 1v74, Figure 5.7A). FSA produces a near-native model with 5.8 Å *i*-RMSD. The model was based on the alignments (Figure 5.7B and C) with the monomers from the template complex Colicin E5 with its immunity protein (chains A and B in 2vhz, Figure 5.7D). As one can see, despite the biological function similarity of the target and the template, their overall structures, including interfaces, are quite dissimilar with low target-template sequence identities (9% and 7%, for the receptors and ligands, respectively). However, the same mutual orientations of non-interface helices and part of a β-strand (shown by arrows in Figure 5.7) in the target and the template yielded the near-native FSA model.

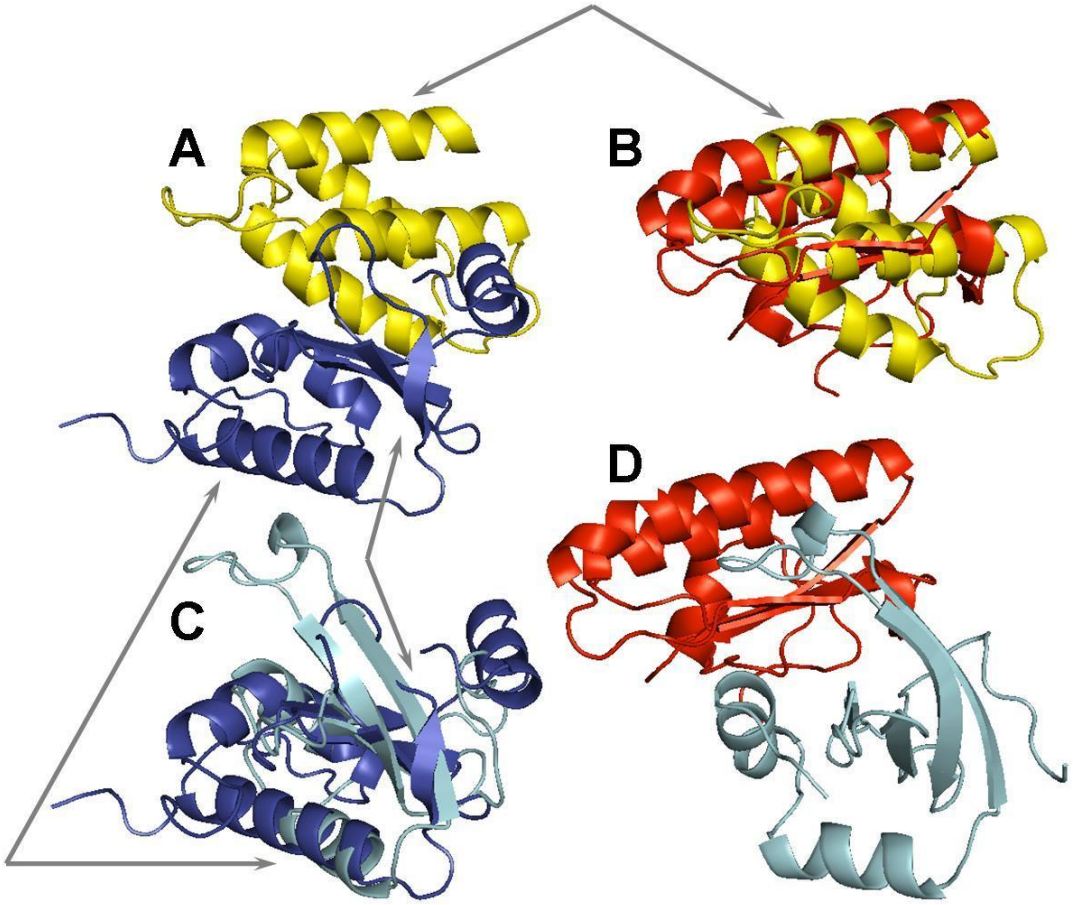

Figure 5.7: Example (#3) of the full alignment more accurate than the local alignment. (A) X-ray structures of the target (1v74, chains A and B) and (D) template (2fhz, chains A and B) complexes, along with (B) FSA alignment for the ligands and (C) receptors. Arrows indicate parts of the target monomers essential for the near-native FSA-model.

References

1. Sinha, R., P.J. Kundrotas, and I.A. Vakser, *Global and local structural similarity in protein-protein complexes.* To be submitted.

2. Kundrotas, P.J. and I.A. Vakser, *Accuracy of protein-protein binding sites in high-throughput template-based modeling.* PLoS Comput Biol, 2010. 6:e1000727.

3. Keskin, O. and R. Nussinov, *Similar binding sites and different partners: Implications to shared proteins in cellular pathways.* Structure, 2007. 15:341-354.

4. Pils, B., R.R. Copley, and J. Schultz, *Variation in structural location and amino acid conservation of functional sites in protein domain families.* Bmc Bioinformatics, 2005. 6:-.

5. Keeble, A.H., L.A. Joachimiak, M.J. Mate, N. Meenan, N. Kirkpatrick, D. Baker, and C. Kleanthous, *Experimental and computational analyses of the energetic basis for dual recognition of immunity proteins by colicin endonucleases.* Journal of Molecular Biology, 2008. 379:745-759.

6. Zhang, Q.C., D. Petrey, R. Norel, and B.H. Honig, *Protein interface conservation across structure space.* Proc Natl Acad Sci U S A, 2010. 107:10896-901.

7. Murzin, A.G., S.E. Brenner, T. Hubbard, and C. Chothia, *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* Journal of Molecular Biology, 1995. 247:536-40.

CHAPTER 6: CONCLUSIONS

A systematic study of the docking methodology based on the structural alignment of protein interfaces was performed to determine the optimal size of the structure in the alignment. The results showed that structural areas corresponding to cutoff values ≤ 10 Å across the interface inadequately represented structural details of the interfaces. The use of such areas in the modeling significantly reduced docking success rates. Increasing the cutoff beyond 12 Å did not significantly increase the success rate for higher-accuracy models and decreased the success rate for lower-accuracy models. While larger structural segments (full structures at the extreme) could provide better alignment for some of the complexes, the modeling time for aligning larger fragments increased. Thus the 12 Å cutoff appears to be optimal overall for the interface alignment-based docking and the best choice for the large-scale (e.g., on the scale of the entire genome) applications to protein interaction networks. Such systems contain only a limited number of experimentally determined monomer structures and by necessity are populated by monomer models of limited accuracy obtained by high-throughput computational techniques. Thus these monomer models require relaxed docking acceptance criteria (i-RMSD ≤ 10 Å) where the 12 Å cutoff provides the best results.

Template-based protein-protein docking was performed by taking advantage of the structural similarity between template and target proteins at different scales (global and local). A library of 11,932 interfaces was generated from the biological units derived from the PDB, and used as a template resource to model new complexes.

Protein-protein interfaces were defined on the basis of the optimum distance cutoff (12 Å) obtained from the first part of the work. The structure alignment protocol was validated on the DOCKGROUND benchmark sets (DG99 and DG372). Results showed that the templates for higher-accuracy models often share not only local but also global structural similarity with the targets, regardless of the degree of sequence identity between the target-template. However, the templates for lower-accuracy models typically had only local structural similarity with the target structures. Overall, the PSA approach yielded more accurate models than the FSA. Most of the templates identified by the PSA had low sequence identity with the target, which makes them hard to detect by sequence-based methods. Thus the application of structural alignment appears to perform better than typical docking protocols in producing acceptable near-native models and shows a significantly high success for the DOCKGROUND benchmark sets. Evidently, the structure alignment method expands the template space beyond the easily detectable sequence similarity range.

Trends obtained from the second part of the work elucidated a greater correspondence between FSA and PSA protocols in providing higher-accuracy models but the same trend did not continue in lower-accuracy models. A high-throughput implementation of structural similarity protocols (both global and local) at genome wide scale requires a clear demarcation of their individual applicability. The third part of the thesis addressed this issue by understanding the extent of structural conservation in protein-protein complexes.

Application of structure alignment method on the statistically significant test set (DG372) sheds light on the following facts: For a majority of higher-accuracy PSA only models only one component of the template shared global structural similarity with the target protein while the other component had dissimilar global fold and significantly lower sequence identity with the corresponding target protein. Thus, if it is known that a protein in question binds different proteins at a single binding site (like many enzyme-inhibitor complexes) the PSA is a better alternative. Interestingly the majority of the lower-accuracy models through PSA were modeled using homo-dimers as templates and insignificant sequence and structural similarities (at global scale) were observed between homo-dimeric templates and target proteins. This suggests that the majority of the space of interface geometries is probably covered by homo-oligomers.

The results presented in this thesis conclude that the structure alignment techniques significantly improve the predictive power of computational techniques modeling protein interactions, drastically expanding template space. Many target template pairs identified by the structural alignment are from distant organisms and perform diverse functions, again suggesting that conservation of structural elements in biological macromolecules is related to physical properties of individual atoms rather than to "generic" properties of larger atom groups. The utility of the approach is increasing with the greater availability of the docking templates - co-crystallized protein complexes. With the growing abundance of the computationally modeled protein subunits the future of the structure alignment methods would depend on their ability to accommodate the structural inaccuracies present in the monomers modeled

in silico. Thus, in future, the structure alignment methods are required to be developed and benchmarked to work with computationally modeled proteins.

SUPPLEMENTAL TABLE S1. Lower accuracy (5 Å < i-RMSD < 10 Å) models built by the PSA12 protocol.

	BEST LOWER-ACCURACY MODEL (among all predictions)										TOP MODEL (Nr.1, if different from the best model)						BEST RANKED MODEL (if different from the best and top models)						
	Target[1]		Template[1]		Rank	*i*-RMSD, Å	TMScore[3]		Seq ID, %		Template[1]		*i*-RMSD, Å	TMScore[3]		Seq-ID, %			Rank	*i*-RMSD, Å	TMScore[3]		Seq-ID, %	
							R	L	R	L				R	L	R	L				R	L	R	L
	Targets, for which models were built by both PSA12 and FSA protocols																							
1	1cf7	BA	1yj9	VS	2314	5.6	46	23	12	15	1yac	BA	35.0	71	83	7	6	1kcf	AB	452	6.4	52	38	5	5
2	1clv	AI	1bvn	PT	17	9.6	80	20	53	14	1viw	AB	11.4	97	33	96	5								
3	1cxz	AB	1lwu	AL	266	9.0	55	52	7	5	1tu3	AF	22.0	77	88	27	12								
4	1ebd	BC	1pm9	AA[2]	2433	9.2	40	32	9	6	1gk4	EC	33.0	76	76	3	13	1b4f	EC	1881	9.7	53	22	4	12
5	1f02	IT	1fe6	BA	2670	8.5	20	41	5	10	1yac	AB	44.0	60	63	10	9								
6	1f93	BE	1sjc	CA	1524	6.9	40	20	7	2	1xwr	AC	16.0	74	54	11	11	2g37	AB	1230	7.7	35	46	7	4
7	1fle	EI	1tx6	AI	26	5.6	90	22	39	15	1h9h	EI	17.2	88	47	39	34								
8	1g3n	AC	1f5q	AB	8	5.8	60	78	46	22	1h1s	CB	36.5	78	88	47	19	2f2c	BA	3	6.6	68	84	92	30
9	1gc1	GC	2avu	FE	1153	9.5	48	22	9	12	2j01	VN	32.5	25	95	7	11								
10	1gcQ	BC	1ekj	CA	428	8.6	50	20	9	8	2fpd	AC	23.5	85	80	24	24	1igq	CA	147	8.9	36	53	13	13
11	1h2s	AB	1zmo	CA	2809	5.1	41	41	7	11	1yac	AB	18.2	87	77	10	8	2b05	AD	613	7.8	51	51	10	4
12	1i8l	AC	1qp1	BC	88	9.5	43	63	12	18	1dqt	BA	48.5	48	85	11	64	1bre	CD	41	9.8	51	63	12	18
13	1im3	AD	1sct	CB	2465	9.4	40	23	11	14	1k8d	AB	41.0	97	54	70	14								
14	1ktk	EA	1x1o	AB	1980	9.8	22	44	13	8	1hez	AE	23.0	76	59	27	7								
15	1m27	AC	2f86	DF	723	8.8	25	41	12	8	1d1z	BA	42.0	94	32	90	10	1e509	QR	605	9.3	42	27	10	7
16	1nt2	BA	1yp2	BC	2111	8.3	36	40	7	8	1xpp	AC	40.6	73	86	7	11								
17	1nw9	BA	1htq	FS	412	9.9	44	39	8	5	1xwr	CA	43.0	66	76	5	10								
18	1puf	AB	1wvq	BB[2]	470	9.7	41	48	6	11	1yac	BA	40.3	76	84	6	9								
19	1qo3	AC	2j6e	BH	2184	8.3	48	26	15	8	1zs4	DA	29.3	74	67	5	9								
20	1s1q	AB	1pzn	FG	1926	8.2	42	21	9	8	2c7n	AB	30.5	44	87	8	96	1hkx	LN	1613	9.3	45	23	11	8
21	1s6v	AB	1dpb	AA[2]	1301	7.0	41	36	12	8	2oqg	CD	39.4	67	72	9	7								
22	1sgf	GB	1ezs	CB	4	9.5	91	34	41	10	1ylc	AB	31.0	90	38	41	10								
23	1spp	AB	2uy7	AB	637	7.8	28	41	11	13	1gam	AB	41.0	50	53	16	14	2uy6	AB	584	8.6	29	41	11	14
24	1sq2	LN	1p5v	AB	1487	9.7	20	42	11	15	1gpq	CB	31.9	95	28	98	10								
25	1t0f	AC	1mh9	AA[2]	1087	6.7	40	33	13	7	1xwr	CA	24.7	61	76	6	14								
26	1tdq	AB	1t5r	DB	1115	9.2	40	26	14	7	1rld	21	18.4	36	82	6	25								
27	1th1	AC	1hds	DB	1334	9.8	40	30	8	13	1zs4	DA	77.6	73	63	3	10								
28	1us7	AB	2gel	AA[2]	1511	9.6	35	47	12	12	2ccn	AB	28.5	70	93	3	3								
29	1uzx	AB	1usy	AC	897	9.8	33	43	7	7	2c7n	AB	28.7	45	95	8	100								
30	1w1w	CG	2jeb	AA[2]	2863	9.8	52	18	15	5	1yac	AB	37.1	75	72	11	5								
31	1x3w	AB	1tlf	AB	591	8.7	41	44	12	4	1yac	BA	27.7	80	52	9	7	1lwu	LA	299	9.0	45	47	10	10
32	1z3e	AB	1xg5	BC	2279	9.2	18	41	9	7	1xwr	CA	24.2	67	76	15	9								
33	2a01	AD	1n8j	JD	2102	9.8	48	20	11	11	1n4x	LH	50.8	35	81	8	44								
34	2a19	BA	1ltx	RA	2786	9.1	43	22	9	7	2bni	AD	38.1	73	74	4	5								
35	2a5d	BA	2h5x	BC	2440	5.6	23	46	12	11	1r4a	HD	28.1	41	91	55	6								
36	2asS	BC	1kmi	YZ	1730	8.2	40	21	9	5	1noe	HF	25.2	62	75	7	9								
37	2mta	CA	2j6e	IB	497	9.3	30	41	10	12	7pcy	AA[2]	43.2	27	73	14	22								
38	3ygs	CP	1uj2	BB[2]	862	8.1	40	41	11	12	1xwr	CA	18.0	65	78	14	12	1sqx	EK	273	9.0	54	38	10	13

SUPPLEMENTAL TABLE S1 (contd.)

	BEST LOWER-ACCURACY MODEL (among all predictions)										TOP MODEL (Nr.1, if different from the best model)							BEST RANKED MODEL (if different from the best and top models)							
	Target[1]	Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq ID, %		Template[1]	*i*-RMSD, Å	TMScore[3]		Seq-ID, %		Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq-ID, %				
					R	L	R	L			R	L	R	L				R	L	R	L			
	Targets, for which models were built by the PSA12 protocol only																							
1	1a9n AB	1ekj BD	184	6.9	40	52	10	10	1oia AB	32.4	32	94	69	10										
2	1ais BA	2g3m EB	2646	9.0	40	32	5	5	2nrn AC	31.0	71	65	4	4	1jd2 BF	411	9.6	77	16	10	11			
3	1avw AB	1h9i EI	21	9.6	99	27	99	7	2iln AI	15.5	98	40	82	6										
4	1b0n AB	1u2e CA	2390	5.9	24	44	8	3	1xpp CA	31.7	91	76	11	6	2jb0 AA[2]	2254	6.5	33	40	12	6			
5	1bvn PT	1viw AB	3	8.9	94	25	51	7	1xv8 AB	16.5	96	24	85	4										
6	1c4z AD	1gmj CD	133	9.1	51	45	4	6	1yac BA	41.0	74	66	11	11										
7	1dev AB	1we3 SR	1885	9.2	42	17	9	8	1zs4 DA	24.1	65	61	10	12										
8	1dml AB	1cuk AA[2]	1290	5.8	33	47	12	5	1zs4 AD	39.6	74	61	7	10										
9	1dx5 MI	1g6g AB	311	9.9	41	31	12	12	1tab EI	48.0	92	36	37	13										
10	1e44 BA	3eip AB	1	7.3	20	87	17	99																
11	1efn AB	1j8f AB	1277	6.2	22	40	3	7	1xpp AC	27.2	40	87	9	15	2fpd CA	39	7.9	83	20	22	9			
12	1fqJ CA	1gkr AC	3305	7.2	14	45	3	10	2ode DC	21.3	30	95	7	74	1htq FS	2210	8.3	17	50	2	9			
13	1g73 AC	2o26 AX	3097	9.9	40	22	6	4	1n0e HF	21.1	61	77	8	10										
14	1ghq AB	1s7o AC	115	7.8	61	21	4	9	1qsj BA	58.5	95	30	75	9										
15	1gl4 AB	1hkx HK	817	7.2	45	22	11	15	1fve AC	26.0	50	74	11	13	1h6k XC	795	9.6	42	26	6	3			
16	1go4 BF	1ezs BA	3104	8.8	43	19	9	12	1xpp AC	37.1	74	89	10	8										
17	1gpw AB	1w5l BA	527	8.9	43	45	5	3	1xpp AC	42.0	50	87	8	10										
18	1gzs AB	2ot3 BA	212	7.4	72	32	26	11	1yac AB	42.7	76	83	12	9										
19	1h6k AX	2gel AA[2]	1250	7.4	49	30	6	8	1yac BA	17.3	86	56	7	10										
20	1h9d AB	2b6e AD	77	7.9	45	46	8	14	2j01 NV	37.4	91	24	12	10										
21	1hx1 AB	1sqx EK	676	9.8	58	36	9	4	2ccf AB	34.7	62	93	2	8										
22	1iyj BA	1t0s BB[2]	458	7.8	48	26	10	3	1zs4 DA	34.2	54	64	2	12										
23	1jtd AB	1nzy AC	277	8.7	57	22	11	12	2g2u AB	22.2	99	39	68	11										
24	1k8r AB	1uad AC	18	9.5	91	30	51	9	1xpp CA	31.8	83	68	13	16										
25	1kg0 BC	2b01 AB	454	10.0	48	32	10	9	1xpp CA	23.9	84	60	7	9										
26	1ksh AB	1xq4 AC	2103	5.5	23	48	10	7	1yac AB	34.3	70	73	10	10	2g6z BC	1969	8.4	46	27	14	16			
27	1ktz BA	1rm6 CF	67	8.8	49	32	9	12	2h62 DA	33.3	53	73	30	17										
28	1kzy CA	2ac0 DA	2378	6.5	56	50	12	97	1xpp AC	30.1	79	89	10	8	1ngk LK	2254	8.7	20	40	9	11			
29	1ldj AB	1xvj BA	1437	8.8	40	33	4	6	1xwr CA	33.1	73	61	2	11										
30	1ltx AR	1hdu EB	2726	8.2	48	22	6	4	1nf4 AO	37.1	73	65	4	6	2he0 BA	1999	9.7	42	33	7	8			
31	1m2v BA	1wsp CB	2446	9.1	29	40	3	2	1m2o CA	25.1	54	86	96	16										
32	1m9f AD	2h15 BA	689	6.3	37	45	9	11	1xwr AC	16.0	69	75	7	8										
33	1mbx AC	1ofh GA	2455	7.1	42	22	11	5	1xpp AC	43.0	92	68	14	10										
34	1mr1 AD	1v2z AA[2]	1698	8.9	33	40	11	9	1n0e FH	33.3	72	56	10	9										
35	1mvf AE	1nbe DB	374	7.4	43	33	11	8	2uzi HL	22.2	88	35	62	9										
36	1n0w AB	1z0k BA	1110	7.7	43	27	9	5	1n0e EG	23.5	69	60	10	5										
37	1nmu AB	1nlx LG	2835	8.6	40	23	6	6	1xpp AC	46.0	68	90	5	10	2o97 BA	746	9.8	45	40	4	13			
38	1nq1 AB	1htq FS	1266	6.6	42	14	10	2	1exb AE	40.1	45	63	10	10										
39	1oc0 AB	1ou8 AA[2]	1845	5.3	42	17	8	5	1xpp CA	12.4	89	59	6	6	1uqr BH	775	7.1	42	27	6	6			
40	1ol5 AB	1q95 LI	1430	9.6	40	32	8	4	1yac AB	28.3	72	64	12	3										
41	1ory AB	1oo0 AB	3134	6.3	48	20	16	7	1xwr CA	30.1	78	78	8	6	1q3q CA	929	6.4	46	39	6	1			

SUPPLEMENTAL TABLE S1 (contd.)

	BEST LOWER-ACCURACY MODEL (among all predictions)										TOP MODEL (Nr.1, if different from the best model)							BEST RANKED MODEL (if different from the best and top models)						
	Target[1]		Template[1]		Rank	*i*-RMSD, Å	TMScore[3]		Seq ID, %		Template[1]		*i*-RMSD, Å	TMScore[3]		Seq-ID, %		Template[1]		Rank	*i*-RMSD, Å	TMScore[3]		Seq-ID, %	
							R	L	R	L				R	L	R	L					R	L	R	L
	Targets, for which models were built by the PSA12 protocol only																								
42	1oxb	AB	1yg8	EK	904	9.0	45	42	11	13	1gk4	EC	33.3	83	66	10	8	2dge	DB	677	9.8	51	40	13	13
43	1p8v	AC	1n0e	BH	11	7.3	58	67	9	8	1yac	AB	38.3	69	67	12	9								
44	1qav	BA	1uzv	BC	1045	5.5	48	23	14	16	2v1w	AA[2]	23.1	77	86	22	27	2bv4	BB[2]	1037	7.2	46	25	14	12
45	1r8s	AE	1r4a	DH	874	6.9	44	42	57	6	1n0e	EC	38.2	60	86	11	11								
46	1rp3	AB	1eoi	AB	1134	9.7	32	52	13	8	1yac	BA	40.0	78	70	7	11								
47	1rzr	CT	1jr3	ED	1242	7.8	44	35	10	6	1kkl	II[2]	24.9	40	88	80	4								
48	1syx	AB	2d9q	AB	2068	8.0	42	22	11	7	2oqg	CD	32.0	66	69	10	11	2jeo	AA[2]	2023	8.4	43	22	8	6
49	1t9g	DS	2uwj	EF	265	7.0	33	67	4	2	2cce	BA	25.3	86	55	2	4								
50	1th8	AB	1nva	BA	2753	7.9	32	43	6	5	1xpp	CA	34.2	87	70	12	14								
51	1txq	AB	2fpd	BA	666	7.3	58	15	13	15	1gk4	CE	23.2	44	87	8	11	1gu4	BA	135	7.9	16	79	9	10
52	1u0s	YA	1h8e	CG	2441	6.9	19	42	5	9	1zs4	AD	29.5	66	68	10	11								
53	1usu	AB	1yj9	21	3241	7.8	48	42	3	7	1xwr	CA	40.3	68	75	4	6								
54	1uw4	BA	1wvi	DA	2470	7.7	40	31	12	8	1zs4	DA	31.8	77	68	5	12	1dm0	DE	2434	9.7	41	30	6	9
55	1v5i	AB	1wo8	DE	1879	7.2	46	21	10	9	1yvw	AC	19.5	70	60	6	11								
56	1vet	AB	4otc	BC	783	6.0	40	44	8	11	1n0e	EG	25.1	75	63	13	12								
57	1xg2	AB	2f16	RP	1268	9.5	22	55	12	8	1xwr	AC	40.0	82	70	5	9								
58	1z2c	BA	1k8r	BA	86	9.4	32	83	30	7	1cxz	BA	44.1	56	86	93	6								
59	1z92	AB	1fe6	DC	2173	9.3	43	15	9	6	1zs4	DA	43.0	67	45	12	9								
60	1zbd	AB	2h61	HB	1372	6.5	41	38	13	7	1tu3	AF	22.4	81	61	33	9								
61	1zbx	AB	1l0n	JK	855	9.4	40	29	7	12	2oqg	CD	31.8	63	54	12	14								
62	2a41	AC	1geg	CB	614	9.7	50	34	12	4	1ma9	BA	35.2	98	35	96	2								
63	2a5y	BA	2jaq	AB	2036	8.2	41	35	8	10	1yac	BA	32.4	82	64	7	9								
64	2ajf	AE	2nys	AB	336	8.5	51	32	4	14	1tu3	FA	27.5	78	37	2	8								
65	2atq	AB	1up8	DA	1291	7.9	47	31	11	6	1n0e	FH	38.0	80	53	4	10								
66	2aw2	AB	1t61	AA[2]	377	7.3	45	25	9	7	2j01	NV	13.9	95	27	12	10								
67	2bfx	AD	1ooy	AA[2]	2760	7.1	41	13	10	2	2np8	BA	30.9	96	41	22	6	1bi8	AB	780	7.4	55	25	20	6
68	2bh1	AX	2ao9	GB	2341	9.3	22	42	8	11	1xpp	CA	34.7	83	65	5	10	1d1z	DC	1909	9.5	41	28	10	10
69	2bkk	AB	1mal	BA	1028	6.3	40	40	13	10	1blx	AB	44.9	52	86	15	33								
70	2btf	AP	1vlh	ED	1040	9.2	50	31	10	9	1n0e	AC	38.0	78	57	1	1								
71	2c5D	AD	2cov	FG	215	6.8	35	48	6	10	1bre	BE	26.0	1	37	4	8	1swu	CB	366	9.8	46	34	6	9
72	2eY4	AC	2iia	AA[2]	1131	9.5	21	43	7	11	2hvy	AC	40.0	98	22	96	15								
73	3fap	AB	1odb	CD	1263	6.9	25	45	13	12	1zs4	AD	30.8	65	74	7	10	1gqm	GH	790	8.8	27	47	12	9

[1] PDB code followed by IDs (as in PDB file) of the receptor (R) and ligand (L) chains in the complex.

[2] Interface of a biological unit complex constructed from the transformation matrix of the given chain, provided in the PDB file.

[3] Multiplied by 100.

SUPPLEMENTAL TABLE S2. Lower accuracy (5 Å < i-RMSD < 10 Å) models built by the FSA protocol.

	BEST MODEL (model with lowest i-RMSD among all predictions)										TOP MODEL (Nr.1, if different from the best model)						BEST RANKED MODEL (if different from the best and top models)							
	Target[1]	Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq ID, %		Template[1]	*i*-RMSD, Å	TMScore[3]		Seq-ID, %		Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq-ID, %		
					R	L	R	L			R	L	R	L				R	L	R	L	
	Targets, for which models were built by both PSA12 and FSA protocols																							
1	1cf7 BA	2oob AB	436	6.7	44	42	10	20	2acj AD	14.6	71	70	11	14										
2	1clv AI	1z0j BA	202	9.8	40	28	2	5	1xv8 AB	25.7	98	28	51	3										
3	1cxz AB	1ykh BA	1323	8.4	22	58	12	12	1i4d DA	30.9	95	73	57	13	1zva AA	308	9.8	33	68	7	12			
4	1ebd BC	2ib0 BA	738	9.0	43	36	6	8	1xdi AB	20.4	95	41	25	3										
5	1f02 IT	1fe6 BA	2414	8.5	19	41	5	10	2nrn CA	69.8	37	67	3	12	2bt2 AE	1390	9.5	28	46	7	7			
6	1f93 BE	1xiw DA	1380	6.9	41	28	10	4	1ru0 AA	14.4	97	44	66	7	1oia BA	428	8.9	43	42	14	9			
7	1fle EI	1eja AB	32	6.9	95	27	39	25	1h9i EI	19.9	95	55	5	9										
8	1g3n AC	1f5q AB	7	7.3	82	85	45	22	2f2c BA	7.8	83	91	92	30	2f2c BA	1	7.8	83	91	92	30			
9	1gc1 GC	1ppj JE	1379	8.8	43	27	2	11	2yxm AA	47.9	40	77	6	11										
10	1gcQ BC	1igq DE	152	8.5	46	57	13	10	1bu1 AD	17.7	87	87	32	21	2fpd AB	94	8.6	86	80	24	24			
11	1h2s AB	1yg2 AA	1798	5.1	35	51	14	4	2nrn CA	53.8	78	84	2	6	1a2x BA	298	8.7	75	38	2	6			
12	1i81 AC	1fyt DE	183	9.9	47	68	13	8	1dqt BA	49.2	60	89	10	64										
13	1im3 AD	1ws8 BA	1485	9.8	40	43	6	8	1q94 AB	41.0	98	61	91	14										
14	1ktk EA	1dqt BA	270	9.7	68	40	12	11	1igc HA	68.3	82	63	25	5										
15	1m27 AC	2gyz AA	484	8.6	34	44	12	6	1x27 CB	28.8	82	88	14	18										
16	1nt2 BA	1lb1 AB	2175	9.6	26	46	10	13	2nnw AB	37.0	67	95	24	41	1h3l AB	965	9.7	48	33	6	12			
17	1nw9 BA	2j8e BA	83	8.4	60	28	9	13	1xb0 CB	37.9	37	97	10	83										
18	1puf AB	1tu3 AF	848	8.6	31	42	8	9	1le8 AB	56.5	85	74	20	27										
19	1qo3 AC	2j8o AB	1940	8.3	41	26	12	12	1im3 AD	19.1	93	33	70	13										
20	1s1q AB	2gu9 BA	740	9.2	35	41	10	16	2gmi BC	41.8	62	95	13	93										
21	1s6v AB	2v1s EB	743	7.8	42	28	7	12	2bcn CB	22.3	99	97	98	97										
22	1sgf GB	1gl1 CK	4	9.4	89	41	34	12	2f3c EI	44.2	95	37	41	7										
23	1spp AB	2co7 AB	808	7.8	44	40	8	10	1szb AB	27.7	69	73	11	12	1p5u AB	742	8.8	40	46	13	8			
24	1sq2 LN	1u3h HE	197	9.8	23	79	11	26	1jtp LA	16.5	96	80	94	22										
25	1t0f AC	2o1k BB	653	5.8	43	20	3	18	2bni AD	18.4	61	30	3	12										
26	1tdq AB	1k9i HB	11	9.5	29	86	7	30	2msb BA	54.6	32	91	9	23										
27	1th1 AC	1m1e AB	5	9.3	97	25	96	9	1i7x CD	11.8	96	36	97	13										
28	1us7 AB	1xwr DB	1035	8.5	30	44	10	7	2hy6 BA	25.1	76	89	3	3										
29	1uzx AB	2gmi CB	401	9.3	37	41	8	12	2nvu CJ	49.1	59	90	15	56										
30	1w1w CG	1cz3 AB	1949	8.1	43	34	10	8	1aie AA	35.2	68	56	3	8										
31	1x3w AB	1t33 BA	1508	8.4	33	42	10	7	2f4m AB	13.1	86	50	32	29										
32	1z3e AB	1fbq BA	635	8.4	35	43	9	11	1dxs AA	28.9	34	67	10	13										
33	2a01 AD	1oqd KA	1600	8.9	31	42	3	10	1uwx AH	55.6	36	88	7	26										
34	2a19 BA	1w2i BA	1441	9.2	42	32	8	13	2nrn AD	45.0	71	77	4	4										
35	2a5d BA	2hlk AA	3305	7.6	26	40	11	10	1r4a HD	52.4	35	96	6	55	1pzm BA	2608	9.6	24	47	11	13			
36	2asS BC	2cov EG	1262	8.8	25	43	7	10	1buh AB	46.5	28	94	11	99										
37	2mta CA	1k5j ED	822	9.3	26	43	12	14	7pcy AA	17.1	26	75	14	22										
38	3ygs CP	2gsc BC	134	9.0	46	42	10	16	2nsn AA	27.2	65	78	19	17										

SUPPLEMENTAL TABLE S2 (contd.)

	BEST MODEL (model with lowest i-RMSD among all predictions)										TOP MODEL (Nr.1, if different from the best model)							BEST RANKED MODEL (if different from the best and top models)						
	Target[1]		Template[1]		Rank	*i*-RMSD, Å	TMScore[3]		Seq ID, %		Template[1]		*i*-RMSD, Å	TMScore[3]		Seq-ID, %		Template[1]		Rank	*i*-RMSD, Å	TMScore[3]		Seq-ID, %	
							R	L	R	L				R	L	R	L					R	L	R	L
	Targets, for which models were built by the FSA protocol only																								
1	1nvu	SR	1vs1	CA	2501	8.4	25	52	10	11	2gzh	BA	43.9	57	82	3	33	2b1f	CA	32	9.2	75	47	2	5
2	1g4u	SR	1xar	BA	2977	8.6	49	30	8	10	1tu3	FA	19.9	53	90	2	22	1ayi	AA	1379	9.9	47	38	5	12
3	1t6b	XY	1q5y	DB	2013	8.6	35	42	3	9	1uex	AC	44.5	34	84	2	12								
4	1bzq	AL	1oa2	DE	1059	6.1	32	40	11	14	9rsa	AB	27.1	98	28	99	8	1i8l	CA	384	8.1	22	62	11	8
5	1p9m	CB	1bp3	BA	9	6.1	73	70	20	12	1ilr	AB	33.3	60	85	15	25								
6	1f02	IT	3p2p	BA	1267	7.8	27	48	6	19	2ovc	AA	20.2	35	68	4	90								
7	1f3v	BA	1nlx	KG	1607	7.0	30	42	11	14	1flk	AA	54.3	94	30	44	8	1ru0	AB	574	9.7	41	40	10	7
8	1f6m	AC	1vk0	FE	2327	7.1	28	44	8	10	1nsw	BC	38.5	37	90	8	45								
9	1gvn	BA	2a1b	DE	2273	7.5	40	36	7	13	2nrn	AB	27.1	60	73	3	8	2hmq	BA	473	8.6	35	55	10	10
10	1ltx	AR	1fqj	BC	1078	7.7	34	40	4	2	2cee	BA	47.1	78	50	2	2								
11	1r4a	AE	1th8	AA	590	7.0	37	53	12	9	2nz8	AB	23.5	80	54	15	5	1uo2	AB	559	8.6	45	47	4	6
12	1uad	AC	1igc	AH	140	5.9	40	56	7	9	2uzi	RL	40.6	96	53	50	11								
13	1v74	AB	2fhz	BA	135	5.8	52	41	16	13	2hy6	GA	27.9	57	56	8	10								
14	1wq1	XA	1wnf	BA	3359	7.5	43	26	10	12	2gzd	BC	37.0	90	58	34	2								
15	2auh	AB	2cov	IH	1685	5.8	40	28	4	9	2ivs	BA	33.8	88	41	35	3	1yoz	AB	280	8.7	49	44	7	7
16	2g45	AA	1efn	BD	578	5.4	32	40	14	12	2hd5	AB	31.5	32	95	6	90								

[1] PDB code followed by IDs (as in PDB file) of the receptor (R) and ligand (L) chains in the complex.

[2] Interface of a biological unit complex constructed from the transformation matrix of the given chain, provided in the PDB file.

[3] Multiplied by 100.

SUPPLEMENTAL TABLE S3. Higher accuracy (i-RMSD < 5 Å) models built by the PSA12 protocol.

	BEST MODEL (model with lowest *i*-RMSD among all predictions)									TOP MODEL (Nr.1, if different from the best model)							BEST RANKED MODEL (if different from the best and top models)							
	Target[1]	Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq ID, %			Template[1]	*i*-RMSD, Å	TMScore[3]		Seq-ID, %			Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq-ID, %	
					R	L	R	L				R	L	R	L					R	L	R	L	
	Targets, for which models were built by both PSA12 and FSA protocols																							
1	1agr AE	2ode AB	1	0.6	99	96	85	54																
2	1axi BA	1bp3 BA	1	1.8	86	92	31	90																
3	1ay7 AB	1b27 AD	1	0.9	65	94	24	97																
4	1b34 AB	1h64 VW	1	1.0	87	78	20	23																
5	1bh9 BA	1b67 AB	1	1.5	86	83	17	7																
6	1blx AB	1g3n AB	2	0.9	82	94	96	44	1bi8	DC	1.1	81	98	86	85									
7	1bnd AB	1b98 AM	3	0.4	88	87	99	56	1btg		0.6	90	92	52	57									
8	1brs AD	1ay7 AB	1	0.9	25	96	24	95																
9	1bui AC	1bml AC	1	2.8	97	50	93	6																
10	1c1y AB	1k8r AB	2	3.1	89	47	56	12	1n0e	HF	33.4	61	75	12	10									
11	1c9p AB	1ezs DA	2	0.9	98	27	80	11	1h9i	EI	12.8	99	28	99	18									
12	1cd9 BA	2d9q BA	1	1.8	90	95	47	98																
13	1ci6 AB	2oqq BA	32	0.9	79	86	25	25	2hy6	GA	20.0	89	88	14	17	1io4	A,B	31	1.6	82	83	7	16	
14	1cse EI	1to2 EI	1	2.5	97	80	68	33																
15	1d3b AB	1h64 VW	1	0.5	81	90	22	20																
16	1d4x AG	1h1v AG	1	0.9	98	62	92	14																
17	1d6r AI	1tx6 AI	1	0.7	99	57	82	23																
18	1df9 BC	2iln BI	1	1.7	49	86	12	48																
19	1dF9 AC	2iln AI	1	2.1	52	86	12	49																
20	1dfj EI	1z7x ZY	1	1.4	96	94	70	76																
21	1dhk AB	1viw AB	1	1.0	95	95	53	98																
22	1dkf BA	1g5y AD	1	1.4	90	96	26	89																
23	1dtd AB	2abz BE	1	0.7	95	98	64	95																
24	1eai BD	2tld EI	34	3.8	82	27	38	10	1h9h	EI	18.6	87	37	39	16	1acb	E,I	24	6.0	89	25	39	16	
25	1f5q AB	1g3n AC	1	2.3	71	81	45	23																
26	1f6f BA	1a22 BA	1	0.7	81	76	30	22																
27	1fbv AC	2c2v SB	1	2.3	68	75	4	30																
28	1ffg AB	1eay AC	2	2.6	67	72	56	13	1yac	BA	15.5	79	60	11	10									
29	1fm9 AD	1dkf CD	2	1.1	97	85	91	21	1uhl	AB	1.9	97	90	86	22									
30	1foe AB	1kil BA	1	0.8	83	96	21	71																
31	1fqj AB	1agr AE	1	0.7	94	95	70	34																
32	1fr2 BA	1mz8 BA	1	1.9	93	53	67	56																
33	1fs1 BA	2p1m AB	1	0.7	71	76	34	3																
34	1g3N AB	1blx AB	1	0.9	84	90	96	44																
35	1gl0 EI	2f91 AB	1	0.6	87	84	37	72																
36	1gl1 AI	2f91 AB	1	1.0	86	62	37	50																
37	1gpq AD	1uuz AD	1	0.7	93	99	21	99																
38	1h59 AB	1wqj IB	1	0.9	92	65	84	32																
39	1h9h EI	1brc EI	1	0.7	98	32	80	14																

SUPPLEMENTAL TABLE S3 (contd.)

	BEST MODEL (model with lowest i-RMSD among all predictions)									TOP MODEL (Nr.1, if different from the best model)							BEST RANKED MODEL (if different from the best and top models)							
	Target[1]	Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq ID, %			Template[1]	*i*-RMSD, Å	TMScore[3]		Seq-ID, %			Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq-ID, %	
					R	L	R	L				R	L	R	L					R	L	R	L	
	Targets, for which models were built by both PSA12 and FSA protocols																							
40	1hl6 BA	2hyi AB	1	0.5	95	93	86	54																
41	1i7w CD	2gl7 AB	1	1.5	98	43	99	15																
42	1iil EA	1e0o DC	1	2.0	58	95	96	52																
43	1ira YX	1itb BA	1	2.4	78	80	99	25																
44	1jat AB	2c2v BC	1	1.3	91	91	66	47																
45	1jch AB	2b5u AB	1	0.3	96	98	98	98																
46	1jdh AB	1g3j AB	1	2.7	95	40	97	52																
47	1jiw PI	1smp AI	1	1.0	93	86	53	36																
48	1jk9 AB	1hl5 AB	1	0.4	87	89	9	54																
49	1jow BA	1f5q AB	3	3.0	71	80	44	20	2uue	AB	3.2	81	82	45	17									
50	1jtg AB	2g2u AB	1	0.6	98	98	67	99																
51	1jw9 BD	1zdu 12	1	1.9	95	66	45	22																
52	1k93 AD	1yrt AB	1	1.5	64	89	18	48																
53	1kac AB	2j1k AQ	1	3.0	76	94	23	92																
54	1kgy AE	2hle AB	1	2.2	80	93	41	95																
55	1ki1 BA	2nz8 AB	1	0.6	77	95	18	71																
56	1klf BA	2uy6 BA	1	1.4	72	83	11	30																
57	1kps AB	2ggr AB	1	0.9	96	90	98	80																
58	1ku6 AB	1fss AB	1	0.8	97	93	59	99																
59	1kz7 AB	2nz8 BA	1	0.5	89	95	33	68																
60	1l6x AB	1fc2 DC	1	0.8	94	75	95	47																
61	1l7v AC	2nq2 AC	16	3.0	68	69	27	19	1xpp	CA	36.5	86	70	5	9									
62	1lpb BA	1eth AB	1	0.4	97	93	85	97																
63	1m1e AB	1i7x CD	11	1.7	99	36	95	8	1yac	BA	40.7	78	78	8	4									
64	1mq8 AB	1t0p BA	1	1.0	84	89	16	92																
65	1nbf AD	2hd5 AB	1	0.5	88	96	23	98																
66	1nex BA	1fs1 CD	1	1.2	69	85	4	47																
67	1nf3 AC	2ov2 AI	1	1.6	96	54	64	8																
68	1nun BA	1cvs AC	1	1.5	79	92	67	33																
69	1o6s AB	2omw AB	1	0.6	99	98	98	89																
70	1oey JA	2npt AD	1	4.3	70	62	14	10																
71	1ofh GA	1g4a EC	1	2.9	79	67	72	79																
72	1ohz AB	2b59 AB	4	4.5	78	55	16	9	2ccl	AB	17.5	99	89	93	87									
73	1oiu AB	2f2c BA	2	2.1	82	75	44	17	1h27	CB	41.0	96	99	96	99									
74	1oo0 AB	1p27 AB	1	0.5	95	97	90	73																
75	1oqd AK	1xu2 DT	1	1.6	84	83	35	90																
76	1oqe AK	1xu1 AR	1	1.8	85	48	36	21																
77	1oyv BI	1r0r EI	1	1.0	98	37	99	16																
78	1p5v AB	2co7 BA	1	3.0	82	64	43	17																
79	1p9M AB	1i1r AB	1	2.6	87	73	99	24																

SUPPLEMENTAL TABLE S3 (contd.)

	BEST MODEL (model with lowest i-RMSD among all predictions)										TOP MODEL (Nr.1, if different from the best model)							BEST RANKED MODEL (if different from the best and top models)							
	Target[1]	Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq ID, %		Template[1]		*i*-RMSD, Å	TMScore[3]		Seq-ID, %		Template[1]		Rank	*i*-RMSD, Å	TMScore[3]		Seq-ID, %	
					R	L	R	L				R	L	R	L					R	L	R	L
	Targets, for which models were built by both PSA12 and FSA protocols																							
80	1ppf EI	1tgs ZI	1	1.1	87	31	31	25																
81	1pqz AB	1de4 AB	1	1.2	73	96	21	69																
82	1r0r EI	1mee AI	4	2.3	97	30	70	15	1xwr	CA	41.3	69	67	9	5									
83	1r1k AD	1uhl AB	1	1.2	90	84	45	35																
84	1s4y BA	2goo AC	1	1.3	91	94	41	60																
85	1sgp EI	2f3c HR	1	2.2	61	74	17	33																
86	1shw BA	2hle BA	1	3.2	81	81	27	41																
87	1stf EI	1yvb AI	1	1.2	89	70	36	16																
88	1sv0 AC	1lky BA	1	2.4	78	92	41	27																
89	1t6g AC	2b42 AB	1	1.2	98	89	99	43																
90	1taf AB	1a7w AA	1	1.0	82	85	22	21																
91	1tbr KS	1cgi EI	1	1.8	87	72	35	19																
92	1toc BR	1tbr HR	5	4.0	96	33	99	17	1bth	HP	45.0	91	46	88	14									
93	1tt5 AB	1y8q BA	1	1.3	90	84	16	22																
94	1tx6 AI	2iln AI	1	1.3	98	79	82	27																
95	1uea AB	2e2d AC	1	2.0	90	75	60	40																
96	1ugh EI	1uug AB	1	0.2	97	97	54	99																
97	1uuz AD	1gpq AD	1	0.8	63	94	21	99																
98	1vg0 AB	2bcg GY	1	0.9	91	76	23	32																
99	1w98 AB	2uue AB	1	1.4	88	89	98	22																
100	1wmh AB	2npt AD	2	1.4	75	68	16	23	1xpp	AC	23.1	71	81	10	14									
101	1wr6 AE	1yd8 HU	1	1.4	80	85	91	98																
102	1wrd AB	1yd8 HU	1	2.6	83	87	25	96																
103	1wyw AB	2d07 AB	1	1.2	93	94	95	48																
104	1x86 AB	1kil BA	1	1.1	91	92	26	53																
105	1xb2 AB	1efu CD	1	1.2	85	75	53	23																
106	1xdk BA	1dkf AD	1	1.6	96	97	28	26																
107	1xk4 BD	1odb AA	1	1.2	89	89	39	45																
108	1xou BA	1kql AB	126	3.0	68	69	17	7	1xpp	AC	24.0	84	89	2	4	2cce	A,B	23	7.5	80	79	3	10	
109	1xu1 AR	1oqd IQ	1	1.3	87	53	35	27																
110	1yvb AI	1nb3 AI	2	1.2	85	64	37	11	1stf	EI	1.3	89	65	38	16									
111	1z0j AB	1z0k AB	1	1.7	88	82	38	36																
112	1zlh AB	1zli AB	1	0.6	96	97	45	99																
113	2a5t AB	1ii5 AA	1	3.2	81	80	14	14																
114	2apo AB	2aus CD	1	0.5	95	72	56	60																
115	2ass BA	2e31 CD	1	1.2	59	39	94	15																
116	2ayo AB	1nbf AD	1	1.4	84	99	15	99																
117	2b59 AB	1ohz AB	1	4.5	63	58	17	9																
118	2bkr AB	1tgz AB	3	3.1	84	74	20	16	1tgz	AB	3.1	84	74	20	16									
119	2c1m AB	2c1t AC	1	1.7	91	68	47	30																

SUPPLEMENTAL TABLE S3 (contd.)

	BEST MODEL (model with lowest i-RMSD among all predictions)									TOP MODEL (Nr.1, if different from the best model)							BEST RANKED MODEL (if different from the best and top model)							
	Target[1]	Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq ID, %			Template[1]	*i*-RMSD, Å	TMScore[3]		Seq-ID, %			Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq-ID, %	
					R	L	R	L				R	L	R	L					R	L	R	L
	Targets, for which models were built by both PSA12 and FSA protocols																							
120	2c2v HT	1fbv CA	18	2.4	71	45	30	5		2oqg DC	45.0	57	68	14	11									
121	2ckh AB	1euv AB	2	0.9	90	87	29	46		1tgz AB	1.3	97	94	59	51									
122	2ey4 AE	2aus CD	1	1.0	98	84	85	92																
123	2fi4 EI	2ra3 EI	1	0.9	98	99	98	74																
124	2goo AC	1s4y BA	1	1.3	93	91	41	61																
125	2sni EI	2tec EI	1	2.5	97	83	40	35																
126	3hhr CA	1cd9 BA	14	4.8	54	60	15	13		1a22 BA	27.7	86	96	92	95									
127	3sic EI	1r0r EI	1	0.9	98	41	70	12																
128	4cpa AI	2abz BE	1	1.0	98	27	95	20																
129	4htc HI	1hrt HI	1	0.3	94	92	88	84																
130	4sgb EI	1ldt TL	38	2.1	61	27	15	17		1noe CA	18.1	57	42	13	6		1eai A,C	25	4.0	56	34	17	18	
	Targets, for which models were built by the PSA12 protocol only																							
1	1mzw AB	1m1l BC	2120	4.9	29	42	10	4		1yac BA	21.1	63	85	9	4		1uo2 B,A	810	7.2	40	38	3	17	
2	1pq1 AB	1c3v BA	3321	4.8	32	42	10	3		2nrn BD	32.0	74	82	4	11		2d8d B,A	202	7.0	45	74	10	14	
3	1mox AC	1nq1 AB	1	2.0	71	61	80	39																
4	1or7 AC	1i4d DA	2936	4.6	30	40	15	7		1xwr AC	35.7	83	71	13	8		2gsc C,B	1138	6.5	44	39	10	11	
5	1nkp AB	1gd2 EF	92	1.6	84	76	14	12		1jnm AB	8.9	88	92	13	15		2o1k B,A	45	2.8	87	81	12	8	
6	1qa9 AB	1aar AB	151	4.1	46	40	18	11		1u2h AA	42.6	74	76	12	10		1nap C,A	143	8.9	46	41	11	8	
7	1acb EI	1ldt TL	84	1.3	80	30	41	9		1yac BA	46.0	81	65	10	7		1ppf E,I	63	5.7	87	26	29	12	
8	1h41 AD	2uue AB	45	3.2	72	47	55	10		1n0e EG	31.8	83	60	12	8									
9	1dp5 AB	1e0r BB	2210	4.2	32	51	7	4		1xpp CA	39.0	85	68	6	3		1sb8 A,A	1421	5.7	27	61	12	2	
10	1fcc BD	1ve5 AB	1552	4.7	31	40	10	5		1uwx HA	24.1	80	90	94	21		1h9d A,B	1029	7.1	48	27	11	9	
11	1fyh AB	1y6m LR	258	4.7	37	45	11	17		1ekj DB	34.6	59	58	10	12									
12	1xdt TR	1sbf AA	660	5.0	41	32	8	4		1mdt BA	52.7	96	29	98	2		1yl5 A,B	538	9.4	41	33	9	5	
13	1dow AB	1zw0 BA	291	4.3	59	51	7	9		1xpp CA	33.3	88	80	9	10		1nkd A,A	193	6.6	62	56	7	10	

[1] PDB code followed by IDs (as in PDB file) of the receptor (R) and ligand (L) chains in the complex.

[2] Interface of a biological unit complex constructed from the transformation matrix of the given chain, provided in the PDB file.

[3] Multiplied by 100.

SUPPLEMENTAL TABLE S4. Higher accuracy (i-RMSD < 5 Å) models built by the FSA protocol.

	BEST MODEL (model with lowest *i*-RMSD among all predictions)										TOP MODEL (Nr.1, if different from the best model)							BEST RANKED MODEL (if different from the best and top models)							
	Target[1]		Template[1]		Rank	*i*-RMSD, Å	TMScore[3]		Seq ID, %		Template[1]		*i*-RMSD, Å	TMScore[3]		Seq-ID, %		Template[1]		Rank	*i*-RMSD, Å	TMScore[3]		Seq-ID, %	
							R	L	R	L				R	L	R	L					R	L	R	L
	Targets, for which models were built by both PSA12 and FSA protocols																								
1	1agr	AE	2ode	AB	1	0.5	99	97	85	54															
2	1axi	BA	1bp3	BA	1	2.4	86	92	31	90															
3	1ay7	AB	1b27	AD	1	1.0	62	96	24	97															
4	1b34	AB	1h64	VW	1	1.8	81	83	20	23															
5	1bh9	BA	1b67	AB	1	1.7	87	83	17	7															
6	1blx	AB	1g3n	AB	2	1.3	92	91	96	44	1bi8	DC	1.2	92	98	86	85								
7	1bnd	AB	1b98	AM	3	0.6	86	87	99	58	1btg	BC	0.7	86	93	52	57								
8	1brs	AD	1ay7	AB	1	0.8	62	97	24	95															
9	1bui	AC	1bml	AC	1	3.3	94	68	96	6															
10	1c1y	AB	1k8r	AB	1	4.1	96	59	56	12															
11	1c9p	AB	1ezs	DA	1	1.7	99	28	80	11															
12	1cd9	BA	2d9q	BA	1	1.8	94	95	47	98															
13	1ci6	AB	2oqq	BA	105	1.0	79	86	25	25	1io4	AB	16.5	87	90	21	67	2ccn	BB[2]	40	8.2	84	87	14	17
14	1cse	EI	1to2	EI	1	2.8	99	78	68	33															
15	1d3b	AB	1h64	VW	1	0.9	86	84	22	20															
16	1d4x	AG	1h1v	AG	1	1.1	98	79	92	14															
17	1d6r	AI	1tx6	AI	1	0.9	99	64	82	23															
18	1df9	BC	2iln	BI	1	2.2	63	83	12	48															
19	1dF9	AC	2iln	AI	1	2.8	62	83	12	49															
20	1dfj	EI	1z7x	ZY	1	1.7	96	96	67	76															
21	1dhk	AB	1viw	AB	1	1.0	98	98	51	98															
22	1dkf	BA	1g5y	AD	1	1.5	76	81	26	89															
23	1dtd	AB	2abz	BE	1	0.7	98	97	64	95															
24	1eai	BD	1tbr	KS	120	2.1	90	23	31	17	1h9h	EI	18.6	92	38	39	16								
25	1f5q	AB	2f2c	BA	1	3.2	87	84	45	23															
26	1f6f	BA	1a22	BA	1	1.0	78	78	30	22															
27	1fbv	AC	2c2v	SB	1	2.9	60	86	4	30															
28	1ffg	AB	1eay	AC	1	2.8	84	82	56	13															
29	1fm9	AD	1dkf	CD	8	2.0	91	85	90	21	1uh1	AB	2.8	97	90	86	22								
30	1foe	AB	1kil	BA	1	0.9	60	98	21	71															
31	1fqj	AB	1agr	AE	1	1.0	98	94	70	34															
32	1fr2	BA	1mz8	BA	1	2.9	96	85	67	56															
33	1fs1	BA	2p1m	AB	1	1.0	70	73	34	3															
34	1g3N	AB	1blx	AB	1	1.3	92	91	96	44															
35	1gl0	EI	2f91	AB	1	0.8	90	84	37	72															
36	1gl1	AI	2f91	AB	1	1.2	89	62	37	50															
37	1gpq	AD	1uuz	AD	1	2.4	74	97	21	99															
38	1h59	AB	1wqj	IB	1	0.9	87	92	84	32															
39	1h9h	EI	1brc	EI	10	0.8	99	31	80	14	1dx5	MI	29.3	92	45	38	10	2f3c	EI	3	1.1	99	34	80	27

SUPPLEMENTAL TABLE S4 (contd.)

	BEST MODEL (model with lowest i-RMSD among all predictions)								TOP MODEL (Nr.1, if different from the best model)						BEST RANKED MODEL (if different from the best and top models)						
	Target[1]	Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq ID, %		Template[1]	*i*-RMSD, Å	TMScore[3]		Seq-ID, %		Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq-ID, %	
					R	L	R	L			R	L	R	L				R	L	R	L
	Targets, for which models were built by both PSA12 and FSA protocols																				
40	1hl6 BA	2hyi AB	1	0.5	93	95	86	54													
41	1i7w CD	2gl7 AB	1	1.6	99	42	99	15													
42	1iil EA	1e0o DC	1	2.0	55	96	96	52													
43	1ira YX	1itb BA	1	2.0	83	87	99	25													
44	1jat AB	2c2v BC	1	1.4	93	97	66	47													
45	1jch AB	2b5u AB	1	0.2	96	98	98	98													
46	1jdh AB	1g3j AB	1	2.8	96	39	97	52													
47	1jiw PI	1smp AI	1	1.4	93	88	53	36													
48	1jk9 AB	1hl5 AB	1	0.8	80	93	9	54													
49	1jow BA	1f5q AB	7	3.3	87	83	44	20	1g3n AC	9.2	85	91	94	30							
50	1jtg AB	2g2u AB	1	0.7	98	99	67	99													
51	1jw9 BD	1zud 12	1	2.4	96	74	45	22													
52	1k93 AD	1yrt AB	1	3.6	80	90	18	48													
53	1kac AB	2j1k AQ	1	4.0	83	97	23	92													
54	1kgy AE	2hle AB	1	2.4	90	96	41	95													
55	1ki1 BA	2nz8 AB	1	5.0	58	98	18	71													
56	1klf BA	2uy6 BA	1	2.4	67	87	11	30													
57	1kps AB	2ggr AB	1	1.4	98	94	98	80													
58	1ku6 AB	1fss AB	1	0.8	98	94	59	99													
59	1kz7 AB	2nz8 BA	1	0.7	90	95	33	68													
60	1l6x AB	1fc2 DC	1	1.0	96	82	95	47													
61	1l7v AC	2nq2 AC	1	3.0	84	77	27	19													
62	1lpb BA	1eth AB	1	1.1	98	97	85	97													
63	1m1e AB	1i7x CD	1	1.4	97	36	95	8													
64	1mq8 AB	1t0p BA	1	1.0	84	89	16	92													
65	1nbf AD	2hd5 AB	1	0.7	83	97	23	98													
66	1nex BA	1fs1 CD	1	1.3	73	87	4	47													
67	1nf3 AC	2ov2 AI	1	1.6	98	54	64	8													
68	1nun BA	1cvs AC	1	2.6	81	94	67	33													
69	1o6s AB	2omw AB	1	0.6	99	98	98	89													
70	1oey JA	2npt AD	15	4.5	61	72	14	10	1wmh AB	6.2	70	73	17	10							
71	1ofh GA	1g4a EC	1	2.9	80	67	72	79													
72	1ohz AB	2b59 AB	2	4.3	78	55	16	9	2ccl AB	17.5	99	97	93	87							
73	1oiu AB	2f2c BA	2	2.0	90	81	44	17	1h27 CB	43.6	98	99	96	99							
74	1oo0 AB	1p27 AB	1	0.3	95	97	90	73													
75	1oqd AK	1xu2 DT	1	1.4	91	83	35	90													
76	1oqe AK	1xu1 AR	1	1.4	91	39	36	21													
77	1oyv BI	1r0r EI	1	1.1	98	38	99	16													
78	1p5v AB	2co7 BA	1	3.7	89	64	41	17													
79	1p9M AB	1i1r AB	1	2.3	95	85	99	24													

SUPPLEMENTAL TABLE S4 (contd.)

	BEST MODEL (model with lowest i-RMSD among all predictions)									TOP MODEL (Nr.1, if different from the best model)							BEST RANKED MODEL (if different from the best and top models)							
	Target[1]	Template[1]	Rank	*i* - RMSD, Å	TMScore[3]		Seq ID, %			Template[1]	*i* - RMSD, Å	TMScore[3]		Seq-ID, %			Template[1]	Rank	*i* - RMSD, Å	TMScore[3]		Seq-ID, %	
					R	L	R	L				R	L	R	L					R	L	R	L
	Targets, for which models were built by both PSA12 and FSA protocols																							
80	1ppf EI	1tgs ZI	1	1.4	90	72	31	25																
81	1pqz AB	1de4 AB	10	1.9	74	97	21	69	1kjv	AB	2.0	77	97	23	85									
82	1r0r EI	2sic EI	1	1.0	99	52	70	12																
83	1r1k AD	1uhl AB	1	1.6	86	93	45	35																
84	1s4y BA	2goo AC	1	1.3	76	96	41	60																
85	1sgp EI	2f3c HR	1	2.3	71	74	17	33																
86	1shw BA	2hle BA	1	3.2	79	91	27	41																
87	1stf EI	1yvb AI	1	2.1	90	65	36	16																
88	1sv0 AC	1lky BA	1	2.3	87	83	41	27																
89	1t6g AC	2b42 AB	1	1.7	97	89	97	43																
90	1taf AB	1a7w AA[2]	1	1.0	82	88	22	21																
91	1tbr KS	1cgi EI	1	1.9	88	67	35	19																
92	1toc BR	1tbr HR	22	4.1	98	32	99	17	1tfx	AC	45.6	95	71	38	15									
93	1tt5 AB	1y8q BA	4	1.6	90	77	16	22	1jw9	BB[2]	2.1	87	90	9	15									
94	1tx6 AI	2iln AI	1	1.5	99	67	82	27																
95	1uea AB	2e2d AC	1	2.3	92	85	60	40																
96	1ugh EI	1uug AB	1	0.8	95	96	54	99																
97	1uuz AD	1gpq AD	1	2.7	74	97	21	99																
98	1vg0 AB	2bcg GY	1	1.7	85	89	23	32																
99	1w98 AB	2uue AB	1	1.6	94	81	98	22																
100	1wmh AB	2npt AD	1	1.8	82	72	16	23																
101	1wr6 AE	1yd8 HU	1	1.7	86	95	91	98																
102	1wrd AB	1yd8 HU	1	3.0	85	95	25	96																
103	1wyw AB	2d07 AB	1	2.0	94	95	95	48																
104	1x86 AB	1kil BA	1	2.6	73	96	26	53																
105	1xb2 AB	1efu CD	1	2.2	82	73	56	23																
106	1xdk BA	1dkf AD	1	0.5	95	92	28	26																
107	1xk4 BD	1odb AA[2]	1	1.2	92	93	38	45																
108	1xou BA	1uo2 AB	20	1.3	76	88	9	9	2hy6	BA	8.4	86	84	9	10									
109	1xu1 AR	1oqd IQ	1	3.7	91	55	35	27																
110	1yvb AI	1nb3 AI	2	1.9	89	64	37	11	1stf	EI	2.0	89	65	36	16									
111	1z0j AB	1z0k AB	1	1.9	95	85	38	36																
112	1zlh AB	1zli AB	1	0.9	97	97	45	99																
113	2a5t AB	1ii5 AA[2]	1	3.7	87	85	14	14																
114	2apo AB	2aus CD	1	1.4	95	72	56	60																
115	2ass BA	2e31 CD	1	3.1	35	93	94	15																
116	2ayo AB	1nbf AD	1	1.7	78	99	15	99																
117	2b59 AB	1ohz AB	1	4.9	78	55	17	9																
118	2bkr AB	1tgz AB	1	3.8	84	74	20	16																
119	2c1m AB	2c1t AC	1	3.4	91	68	47	30																

SUPPLEMENTAL TABLE S4 (contd.)

	BEST MODEL (model with lowest i-RMSD among all predictions)									TOP MODEL (Nr.1, if different from the best model)							BEST RANKED MODEL (if different from the best and top models)							
	Target[1]	Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq ID, %			Template[1]	*i*-RMSD, Å	TMScore[3]		Seq-ID, %			Template[1]	Rank	*i*-RMSD, Å	TMScore[3]		Seq-ID, %	
					R	L	R	L				R	L	R	L					R	L	R	L	
	Targets, for which models were built by both PSA12 and FSA protocols																							
120	2c2v HT	1fbv CA	1	3.8	85	57	30	5																
121	2ckh AB	1euv AB	4	0.9	89	87	29	46	1tgz AB	1.4	96	94	60	52										
122	2ey4 AE	2aus CD	1	1.4	97	84	85	92																
123	2fi4 EI	2ra3 EI	1	0.3	98	92	98	74																
124	2goo AC	1s4y BA	1	1.3	76	96	41	61																
125	2sni EI	2tec EI	1	2.7	94	82	40	35																
126	3hhr CA	1cd9 BA	14	4.4	70	65	15	13	1a22 BA	27.6	91	95	92	95	2d9q BA	3	5.8	75	66	15	13		
127	3sic EI	1r0r EI	1	4.9	99	36	70	12																
128	4cpa AI	2abz BE	1	0.5	98	28	95	20																
129	4htc HI	1hrt HI	1	0.3	94	92	88	84																
130	4sgb EI	1eai BD	60	4.0	70	33	17	18	4pro BA	29.9	91	38	30	7										
	Targets, for which models were built by the FSA protocol only																							
1	1a2x AB	1b8z BA	996	4.5	31	63	9	9	2nrn AD	43.5	50	88	5	14	1iv5 AB	800	5.1	36	64	10	8		
2	1eer BA	1bp3 BA	1	4.3	75	58	17	15	1bp3 BA	4.3	75	58	17	15										
3	1eeR CA	1cd9 DC	3	4.4	74	58	16	13	1a22 BA	28.3	73	59	17	14										
4	1f6F CA	1pvh AB	14	4.5	79	65	23	14	2d9q BA	6.9	83	70	13	12										
5	1i1r AB	2d9q BA	1	4.2	79	76	27	16	2d9q BA	4.2	79	76	27	16										
6	1qbk BC	1ibr BA	1	4.2	74	96	11	89	1ibr BA	4.2	74	96	11	89										
7	1iar BA	1cd9 DC	1	4.6	67	67	16	10	1cd9 DC	4.6	67	67	16	10										
8	1itb BA	1cvs DB	2	4.8	46	79	14	16	1ev2 GC	5.2	48	79	13	15										
9	1lfd BA	1k8r AB	1	3.4	98	61	99	14	1k8r AB	3.4	98	61	99	14										
10	1j2j BA	2h7v AC	1	4.8	83	63	16	3	2h7v AC	4.8	83	63	16	3										
11	1pk1 AB	1sv0 CA	2	3.7	73	61	15	17	1lky AB	5.8	74	64	20	24										
12	1pvh AB	1cd9 BA	1	4.3	79	72	25	11	1cd9 BA	4.3	79	72	25	11										
13	1xd3 AB	2bkr AB	3	4.4	38	93	13	55	2c7n AB	17.4	35	97	99	5										
14	2b5i BA	1cd9 DC	14	4.2	65	60	17	11	1eer CA	7.0	66	63	16	13										
15	2b5I CA	1bp3 BA	15	3.7	76	61	15	11	1a22 BA	10.6	71	69	15	12	1eer BA	11	5.1	76	63	16	13			

[1] PDB code followed by IDs (as in PDB file) of the receptor (R) and ligand (L) chains in the complex.

[2] Interface of a biological unit complex constructed from the transformation matrix of the given chain, provided in the PDB file.

[3] Multiplied by 100.