Development of Protein-Protein Docking Methodology and Benchmarking Environment

By
Ivan Anishchanka

Submitted to the graduate degree program in Computational Biology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Ilya Vakser, Chairperson

_____
Petras Kundrotas, Co-chair

_____
Eric Deeds

_____
Wonpil Im

_____
John Karanicolas

_____
Krzysztof Kuczera

Date Defended: 05/02/2016

The Dissertation Committee for Ivan Anishchanka
certifies that this is the approved version of the following dissertation:

Development of Protein-Protein Docking Methodology and
Benchmarking Environment

_____
Ilya Vakser, Chairperson

_____
Petras Kundrotas, Co-chair

Date approved: 05/02/2016

Abstract

Structural characterization of proteins is essential for understanding life processes at the molecular level. However, only a fraction of known proteins have experimentally determined structures. That fraction is even smaller for protein-protein complexes. Thus, structural modeling of protein-protein interactions (docking) primarily has to rely on modeled structures of the individual proteins, which typically are less accurate than the experimentally determined ones. Such "double" modeling is the Grand Challenge of structural reconstruction of interactome. Yet it remains so far largely untested in a systematic way. This work presents development of comprehensive docking benchmark sets of protein models, and systematic validation of state-of-the-art docking methodologies on these sets. Thorough analysis of template-based and template-free docking performance reveals that even highly inaccurate protein models yield meaningful docking predictions. The results show that the existing docking methodologies can be successfully applied to protein models with a broad range of structural accuracy; the template-based docking is much less sensitive to inaccuracies of protein models than the free docking; and docking can be successfully applied to entire proteomes where most proteins are models of different accuracy.

Acknowledgements

I am grateful for the opportunity to be a PhD student at the KU Center for Computational Biology, working with such brilliant and creative people. I wish to sincerely thank these individuals without whom this thesis would not have been possible.

First and foremost, I would like to thank my advisor, Dr. Ilya Vakser, for his supportive guidance, expert advice and a wonderful environment for studying and doing research in his lab. For me, he is an eminent example of the clarity of thought and the aptitude for revealing and paying attention to the most important and fundamental aspects of scientific problems. I would also like to express my deep gratitude to Dr. Petras Kundrotas for constant support and involvement. The immense amount of time we spent thoroughly discussing every detail of my project was invaluable for crystallizing out the ideas presented in this thesis. I am also thankful to Vakser lab members, Taras Dauzhenka, Saveliy Belkin, Varsha Badal, and Madhurima Das, for the valuable discussions and making my life at KU much more interesting.

I am profoundly thankful to Dr. Alexander Tuzikov. Under his guidance, in a friendly, creative and respectful atmosphere I got introduced to scientific research, and bioinformatics in particular. This experience influenced me greatly in choosing the path of a scientific investigator.

Finally, I would like to thank my friends and family for their participation in my graduate studies at KU. I thank my parents Vladimir and Irina for their unconditional help and support, always encouraging me to do my best in all my undertakings and showing their pride and interest in my work. I am also extremely thankful to my brother Sergey for always being my best and most supportive friend, able to bring back the joy and optimism

even in the seemingly most desperate situations. He is my etalon of a desperate, moderately mad scientist whom I would like to take after. And last, but by far not least, I would like to thank my wife Vasilina, for her active part as an unprejudiced critic of my writings, careful listener of presentations and humorous commenter, all in one, bringing in life and fresh perspective into everyday routine and saving me from going mad due to overworking. You have constantly been my foremost source of inspiration and I am grateful to you for the happiness and joy that you bring to my life.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Protein interactome

Protein-protein interactions (PPI) drive many cellular processes. Structural characterization of PPI is important for better understanding of these processes and for our ability to manipulate them. Genome sequencing has determined a massive amount of protein sequences. At the same time, the number of corresponding 3D structures is far lagging, due to the limitations of the experimental techniques for protein structure determination. This gap is supposed to be bridged by computational approaches, using experimentally determined structures as templates to model related proteins. Analysis of the rapidly growing Protein Data Bank (PDB) suggests that the protein structure space is continuous and close to complete, which provides an opportunity to model a large part of the "protein universe" (1-3).

When it comes to protein interactions, high-throughput experimental techniques (two-hybrid analysis, mass spectrometry, etc.) provide data for recreating interaction networks (interactomes) for many organisms and/or biochemical pathways. To understand the mechanisms of these interactions, it is essential to have the structures of the protein-protein complexes. However, the fraction of experimentally determined PPI structures is even smaller than that of the individual proteins, due to the massive amount of protein interactions that is significantly larger than the number of individual proteins, and a much greater difficulty of crystallizing protein-protein complexes. Experimental techniques for structure determination of PPI have limited capabilities. The X-ray crystallography, the

major source of today's knowledge on atomic-level structures of PPI, accounts only for 26% of known PPI in *E. coli* and 6.7% in human (4). Thus, the structure of most known protein interactions has to be determined by computational methods for PPI modeling (protein docking) (5).

1.2 Computational methods for structural modeling of PPI

Current computational methods for structural modeling of PPI (docking) generally belong to two major categories: (a) free (or *ab initio*) docking, where relative positions of the two proteins are systematically sampled and, generally, no information other than the structure of the two proteins, is assumed to be known *a priori*; and (b) template-based docking, where the prediction is made according to sequence or structure similarity of the target proteins to the ones in co-crystallized complexes (6-9).

1.2.1 Template-free approach

Docking historically started with the *ab initio* methods based on physical potentials (primarily, van der Waals interactions (10)). In a vast class of template-free docking methods (11) initial sampling of the conformational space is done by correlation techniques, where both proteins are projected onto a spatial grid and their shape complementarity is rapidly evaluated using Fast Fourier Transformation (FFT) (12). In this approach, protein structures are treated as rigid bodies. This basic docking algorithm (12) is currently increasingly supplemented by knowledge-based approaches (e.g. statistical potentials (13, 14), constraints-driven docking (15), etc.). Most of the existing free-docking servers (PIPER (13), HADDOCK (15), GRAMM-X (16), ZDOCK (17), etc.) employ

constraint-based *ab initio* approach and have an established record of successful practical applications. However, despite this success, free docking methods have serious limitations, mostly due to the large size of the search space and structural flexibility upon the complex formation.

1.2.2 Template-based docking

Following a long-standing pattern in individual protein structure prediction, PPI modeling is increasingly employing template-based methods. Several groups (18-21) working on sequence-similarity based PPI modeling have concluded that this methodology yields accurate PPI models, given suitable templates. Currently available experimental PPI structures provide sequence-based templates for ~ 15% of all known PPI (22). The template pool for PPI modeling can be significantly expanded (up to additional 45% for some interactomes) by exploiting structural similarity between protein complexes (4). In particular, it has been shown that valid templates for PPI modeling by structure alignment can be found for almost all known PPI that involve proteins for which the structure is known or can be built by homology (templates are available for the homology modeling of a significant part of the individual proteins (1)). Proteins with dissimilar sequences and function can bind in a similar way (23-26) and the structure similarity was exploited in detection of active sites for small ligands (see reviews (27, 28) and references therein), protein-protein binding sites (24, 26, 29, 30), druggable hot spots in PPI interfaces (31), and in predicting the fact of protein interactions (26, 32, 33). Currently, the structural similarity methodology for PPI modeling is becoming increasingly popular (5).

1.2.3 Accounting for conformational changes upon complex formation

A serious obstacle to the docking of protein structures is the conformational changes upon complex formation (34). Whereas the ultra-low resolution docking may be applicable to cases with large inaccuracies (35), the problem is explicitly addressed by docking methods that allow structure flexibility (36). Rigid-body moves with side-chains repacking is sufficient if proteins undergo moderate conformational changes upon binding (37-39). For difficult targets, backbone flexibility can be accounted for by low-frequency normal mode analysis (40-43), backbone perturbations using the fold-tree-based method (44), and semi flexible refinement of interface residues in torsion angle space followed by Cartesian dynamics refinement in explicit solvent (45).

1.3 Benchmarking of docking algorithms

1.3.1 CAPRI experiment

The ability of protein docking methods to predict the structure of a protein–protein complex has been regularly assessed since 2001 in the community-wide experiment on Critical Assessment of Predicted Interactions, CAPRI (46). Originally, the participating groups performed blind structure predictions for protein–protein complexes given the structure of the component proteins in the unbound form. CAPRI assessors compare the submitted predictions to the unpublished X-ray structures of the complexes, offering an objective evaluation of the existing docking approaches. In recent CAPRI rounds the participants were not always provided with the experimental structures of the interacting proteins but were increasingly asked to model the component proteins prior to the docking. This

experience has already shown that in many cases the models of individual proteins provide sufficient structural details for successful docking predictions (47) even in the "twilight zone" of sequence similarity (48).

1.3.1 Docking benchmarks

To facilitate the development of protein docking algorithms specialized benchmarks of protein-protein complexes have been developed (49-52). Such benchmarks are composed of docking test cases, for which the structures of the complex and of both unbound components are available in PDB. Special care is usually taken to make the benchmarks non-redundant and to include non-obligate interactions only. Testing of docking algorithms on these sets shows their performance in bound or, which is more realistic, unbound docking. Latest docking benchmarks include several hundred complexes of different types (51, 52) and provide a common ground for systematic analysis and comparison of existing docking algorithms.

1.4 Scope and outline of the dissertation

The current research on structural characterization of PPI networks suggests the wide use of protein models, as opposed to their experimentally determined structures. The implication for docking is that it has to be applicable to protein models of limited accuracy. Inevitably, there is a widespread skepticism about the validity of such "double modeling." Therefore, comprehensive benchmark studies on specialized sets of modeled structures are needed to reveal the applicability of docking techniques to protein models, and to pave the way for further advancement of protein-protein docking methodologies.

The primary scope of this work is revealing the predictive power of state-of-the-art docking approaches applied to modeled protein structures. Chapters 2 and 3 present two unique benchmark sets of protein models specifically developed to provide a framework for testing the tolerance limits of current template-free and template-based docking algorithms to local structural inaccuracies in the target proteins. Chapter 4 advances the template-based docking approach developed previously in our lab (4, 53) by a carefully curated non-redundant library of templates. In Chapter 5, the updated template-based protocol, along with the template-free docking approach (35), are systematically tested on the above benchmarks, showing that the existing docking methodologies are applicable to protein models, even in case of low protein structure accuracy. Chapter 6 investigates an important problem of quality of the rigid-body docking prediction. It shows that although the template-based docking, unlike the free docking, is not based on the surface complementarity paradigm, the resulting steric clashes are similar to those in the free docking, due to the generally higher quality of predictions, and thus potentially can be refined by similar protocols.

Chapter 2 is a reprint of Anishchenko I, Kundrotas PJ, Tuzikov AV, Vakser IA. Protein models: The Grand Challenge of protein docking. *Proteins.* 2014; 82: 278–287. Supporting information for this Chapter is in Appendix A.

Chapter 3 is a reprint of Anishchenko I, Kundrotas PJ, Tuzikov AV, Vakser IA. Protein models docking benchmark 2. *Proteins.* 2015; 83: 891–897. Supporting information for this Chapter is in Appendix B.

Chapter 4 is a reprint of Anishchenko I, Kundrotas PJ, Tuzikov AV, Vakser IA. Structural templates for comparative protein docking. *Proteins.* 2015;83:1563–1570.

Chapter 5 is a reprint of Anishchenko I, Kundrotas PJ, Vakser IA. Modeling complexes of modeled proteins. 2016. *Submitted.* Supporting information for this Chapter is in Appendix C.

Chapter 6 is a reprint of Anishchenko I, Kundrotas PJ, Vakser IA. Structural quality of unrefined models in protein docking. 2016. *To be submitted.*

Chapter 2

Protein models: The Grand Challenge of protein docking

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

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

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

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

Proteins. 2014; 82: 278–287

2.1 Introduction

Genome sequencing efforts have determined a massive amount of protein sequences. At the same time, the number of corresponding 3D structures is far lagging, due to the limitations of the experimental techniques for protein structure determination. This gap is supposed to be bridged by computational approaches, using experimentally determined structures as templates to model related proteins. The rapidly growing PDB provides an opportunity to model a large part of the 'protein universe' (1-3). When it comes to protein-protein interactions (PPI), high-throughput experimental techniques (two-hybrid analysis, mass spectroscopy, etc.) provide data for recreating interaction networks for many organisms and/or biochemical pathways. To understand the mechanisms of these interactions, it is essential to have the structures of the protein-protein complexes. However, the fraction of experimentally determined PPI structures is even smaller than that for the individual proteins, due to a larger number of interactions than the number of individual proteins, and a greater difficulty of crystallizing protein-protein complexes.

Computational methods for structural modeling of PPI (docking) historically started with *ab initio* methods based on physical potentials (primarily, van der Waals interactions (10)), currently increasingly supplemented by knowledge-based approaches (e.g. statistical potentials (13, 14), constraints-driven docking (15), etc.). Following a long-standing pattern in individual protein structure prediction, PPI modeling is increasingly employing template-based methods. Efforts of several groups (18-21) working on sequence-similarity based PPI modeling have concluded that this methodology yields accurate PPI models, given suitable templates. The template pool for PPI modeling can be significantly expanded by exploiting structural similarity between protein complexes (4). The structural similarity

methodology for PPI modeling is becoming increasingly popular (5).

These efforts have paved the way to large-scale structural PPI modeling (5). However, the majority of structures to be docked in such studies will themselves be models of limited accuracy. Thus, to directly address the widespread skepticism about the meaningfulness of such 'double modeling,' comprehensive benchmark studies on a carefully selected set of model structures are needed (5). Sets of protein models ('decoys') are used in structural studies of individual proteins (54, 55) and small ligand-receptor interactions (56). However, the existing protein-protein benchmark sets (57, 58), are restricted to the X-ray structures, which are generally not representative of the potentially limited accuracy of protein models.

In our previous study on the applicability of low-resolution template-free protein-protein docking to modeled structures (59), a representative nonredundant set of cocrystallized protein-protein complexes was used to build an array of models of each protein in the set. A procedure was developed to generate the models with RMSD of 1, 2, 3, …, 10 Å from the crystal structure, by repacking of the secondary structure elements. Because of the limited availability of the templates for individual proteins, such templates were not utilized in the procedure. Thus, the resulting 'simulated models' of the proteins, while reflecting the general structural accuracy of the homology models, were not necessarily structurally similar to those.

A much greater current availability of the templates provides an opportunity to generate a new benchmark set of models, explicitly utilizing the actual homology models of the proteins, and thus providing a more adequate benchmarking resource. This article presents a set of structures with several levels of controlled inaccuracy, which mimic high-

throughput homology models. The distortions are 1, 2, …, 6 Å C^{α} RMSD from the X-ray structures of proteins in the DOCKGROUND benchmark set (49, 57). The models were generated by a combination of homology modeling (HM), simulated annealing (SA), and Nudged Elastic Band (NEB) method (60, 61). The sets and the accompanying data provide a comprehensive resource for the development of docking methodology for modeled proteins.

2.2 Methods

The set of complexes is a tool for benchmarking the performance of docking procedures on protein models. Docking programs take the 3D structures of two separate proteins as an input and predict the structure of their complex. To evaluate the prediction, the structure of the correct (X-ray) complex should be available. Thus, the benchmark set consists of models of the individual proteins (not models of complexes) generated from the corresponding structures in co-crystallized complexes. The binary complexes from DOCKGROUND were split into two chains, and models were built independently for each of the monomers.

From the initial set of 100 protein complexes (DOCKGROUND benchmark 3) we excluded 37 complexes with multi-chain interactors. Six models were built for each of the remaining 126 single proteins (63 complexes) within the pre-set accuracy limits (±0.2 Å from 1, 2, …, 6 Å), resulting in 126 x 6 = 756 models in the final set. Our previous study indicated that proteins with RMSD > 6 Å, typically, to a significant extent lose structural recognition characteristics at the binding sites. Thus, 6 Å was used as the upper limit in this study.

Each protein sequence in the dataset was first subjected to single-template homology modeling procedure with the corresponding native structure excluded from the template pool. Templates for the homology models were identified by aligning profile of the target sequence against profiles of all non-redundant sequences in PDB using Needleman-Wunsch dynamic programing algorithm (62) with affine gap penalty (63) as implemented in our in-house program (21). Sequence profiles were extracted from position-specific scoring matrices obtained by 5-iteration PSI-BLAST (64) search against non-redundant sequence database with the substitution matrix BLOSUM62 (65). Alignments of identical sequences from the same organism were excluded from consideration. The model structures were built by the NEST program from JACKAL package (66) with default parameters. Assignment of proteins' secondary structures was by DSSP (67). The HM resulted in ~10,000 full-length models, out of which 290 satisfied our accuracy criteria (38% of the intended 756 structures in the model set).

The remaining 466 models were generated using NEB method (60, 61), in which a low-energy pathway between two protein conformations is approximated by a series of images of the molecule, with the endpoint images fixed in space. All atoms of each image are connected to the corresponding atoms of the previous and next images by virtual elastic 'springs' that keep the image from sliding down the energy landscape onto adjacent images. The NEB pathway was represented by 16 images including endpoints. The first 8 frames were copies of the starting point, whereas the last 8 were copies of the end structure. Pathway minimization by a combination of heating and equilibration, as in SA, but applied to the entire multi-image system, generated structures with RMSD between the end-point RMSD values. The procedure started with heating of the system from 0 to 300 K within

20 *ps* with the spring constant between images k_{NEB} = 1 *kcal/mol* (stage 1). To increase the linkage between images, three short (10 *ps*) molecular dynamics runs were performed with k_{NEB} = 5, 10 and 50 *kcal/mol* (stage 2), and the last value was used during all subsequent steps. The system was then heated from 300 to 400 *K*, and from 400 to 500 *K* and then cooled from 500 to 300 *K* (stage 3). Each heating and cooling run was conducted within 50 *ps* interval and followed by 50 *ps* molecular dynamics equilibration run. Finally, the system was cooled from 300 to 0 *K* within 12 ps (stage 4). Langevin thermostat with collision frequency 1000 ps^{-1} was used for temperature coupling in all NEB calculations, and a simple leapfrog integrator was exploited to propagate the dynamics. The generalized Born implicit solvent model (68) was used in all computations. The non-bonded cut-off distance was set at 12 Å. During initial heating and simulated annealing stages in 300 – 500 *K* temperature range the time step of 0.5 *fs* was utilized; otherwise 1.0 *fs* value was chosen. The NEB calculations were performed by the program *sander.MPI* from the *Amber 10* package (69) with Amber ff03 force-field (70).

The models with RMSD within the set limits were selected for further consideration. Otherwise the NEB procedure was repeated with new end-points selected from the intermediate structures of the previous trajectory. As the starting point of the NEB trajectory, we used the homology model with the closest RMSD below the intended accuracy level, or the native structure of the protein. For the final point of the NEB trajectory, we chose the homology model with the closest RMSD above the intended accuracy level. If such model was not available, the structure was generated by SA from the starting-point homology model (this was the case for 55 monomers in our dataset). We did not use just SA for model generation because the absence of the NEB ‘springs’ makes

it difficult to control the distortion level of the final structure, and also causes considerable distortions of the secondary structure elements at high annealing temperatures.

2.3 Results and discussion

The outline of the procedure is shown in **Figure 2-1** for 2hle, chain A. The initial HM yielded two models. The first one (with 2.98 Å RMSD) was built using chain A of 1kgy with sequence identity 43.1%. The template for the second model (with 4.99 Å RMSD) was chain A of 1nuk with 42.6% sequence identity. The remaining four models were generated by three NEB runs. The start and the end points of the first NEB trajectory were the X-ray structure and the homology model with 2.98 Å RMSD, correspondingly. This NEB run yielded models with 0.90 and 1.95 Å RMSD. The starting point for the second NEB trajectory was the homology model with 3.51 Å RMSD built using chain B of 1shw (sequence identity 42.5%). The final trajectory point was the 4.99 Å model. This run yielded the model with 4.04 Å RMSD. The third NEB run had 4.99 Å model, as the starting point. The end point (7.50 Å RMSD) was generated from this model by SA with the annealing temperature 500 *K* and heating, equilibration and cooling times 100, 300 and 100 *ps*, correspondingly. The run produced the model with 6.02 Å RMSD. As seen in **Figure 2-1**, all models have β-strands and a globular structure, characteristic to the native structure. The 6 Å model has most of the strands with the dihedral angles twisted out of the exact β-strand range and thus displayed as loops. Free terminal fragment, observed in the native structure (**Figure 2-1**), gradually disappears with increasing distortion. Such fragments may introduce a bias for shape-complementarity-based docking procedures, and thus were removed from the structures.

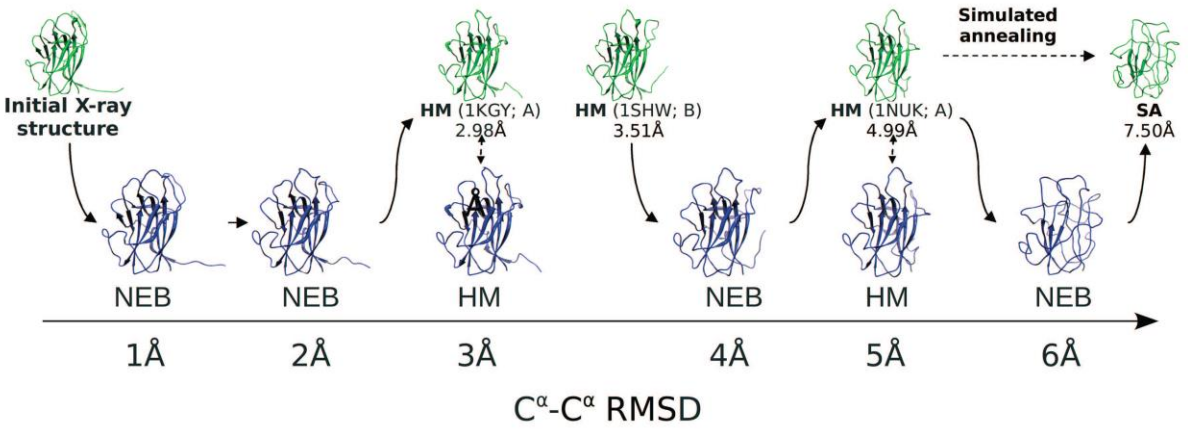

Figure 2-1: *Model-generating procedure*. Models for 2hle, chain A, are generated by the Nudged Elastic Band technique (NEB), homology modeling (HM), and simulated annealing (SA). The base structures (green) are used for building the final models (blue). For homology models, the templates are shown in parentheses, along with the corresponding C^{α} RMSD values. Solid arrows show the path obtained by the NEB procedure connecting two fixed end-points with the intermediate structures at intended accuracy levels.

2.3.1 Assessment of models

Protein models may have inaccuracies, in principle, anywhere in the structure. Thus, to avoid bias in docking benchmarking, models should have distortions distributed along the polypeptide chain. Thus, we considered distribution of distances between C^{α} atoms of corresponding residues in the model and the native structure, as shown in **Figure 2-2** for the chain A of 1r8s. This protein has significant conformational change upon binding, such that the secondary structure patterns in bound and unbound states are different (**Figure 2-2**C). The interface consists of residues 28 through 68 (the residue numbers are from the bound structures) and contains a double-stranded β-sheet and an α-helix (**Figure 2-2**A). Residues 28 – 35 form a loop that enters the binding cleft of the interacting protein. In the

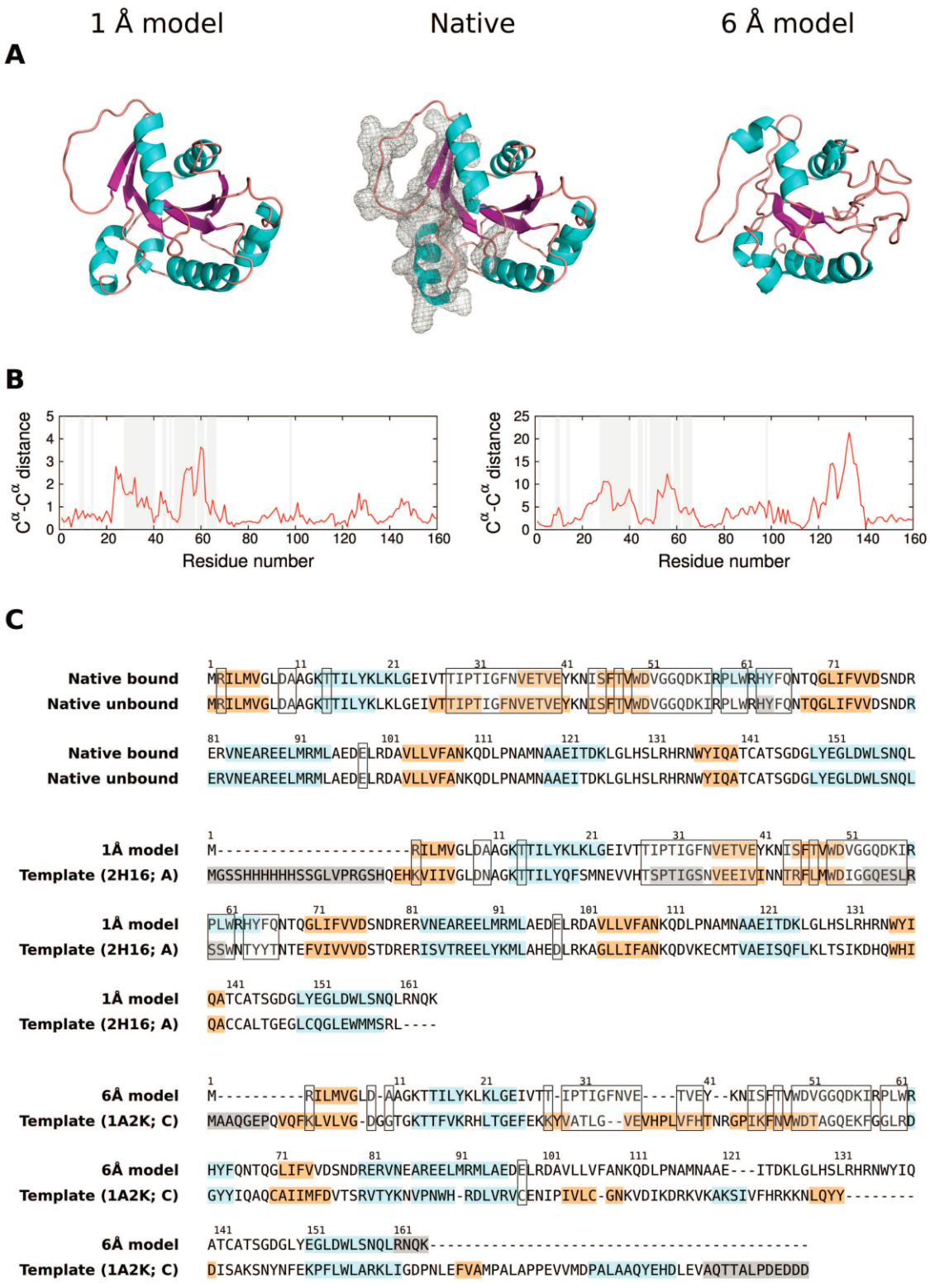

Figure 2-2: *Assessment of model quality*. (A) 1 Å and 6 Å models of 1r8s, chain A, are shown with the native structure, along with (B) distributions of C^{α}–C^{α} distances between the native and the

model structures. (C) Secondary structure patterns of bound (1r8s, chain A) and unbound (1rrf, chain A) states, along with the sequence alignments of 1 Å and 6 Å models with their corresponding templates, show α-helices in cyan and β-strands in orange. Residues in the SEQRES section of the PDB files, which are missing in the ATOM section, are in gray. The interface is shown by the mesh surface in the native 3D structure (A), by the shaded regions (B) and by transparent boxes in the alignments (C).

unbound protein, this loop is assembled in a β-strand forming a β-sheet with two neighboring interface β-strands (residues 36 – 40 and 45 – 50). The short interface α-helix (residues 58 – 64), visible in the bound structure, becomes a loop in the unbound structure. The identified templates resembled mostly the unbound protein and, consequently, the resulting homology models were primarily distorted in the interface area, as can be seen in the 6 Å model in **Figure 2-2**A (shadowed areas in **Figure 2-2**B). The peak in the C^{α}–C^{α} diagram for this model, between residues 120 – 140, is caused by insertions and deletions in the alignment with the template (chain C of 1a2k; **Figure 2-2**C). Out of 46 homology models in the 6 ± 0.2 Å RMSD range, only three were considered for the final set. The other 43 models were rejected after visual inspection of superimposed models and the native structure within the complex. The interface loop in these models either had substantial clashes with the partner protein or deviated from the X-ray structure such that it did not enter the binding site. The final selected model has RMSD 6.01 Å, the closest one to 6 Å.

The binding site distortion (albeit a smaller one) is also observed in the 1 Å model, which was obtained from the NEB trajectory with the native X-ray structure and 3.13 Å homology model as the start and the end points, respectively. Both peaks in the C^{α}–C^{α} distance distribution (left hand panel in **Figure 2-2**B) for this model are caused by

crystallographically unresolved regions in the template 2h16 (gray regions in **Figure 2-2**C), which caused these parts of the model to be built *ab initio* (and thus with lower accuracy).

The peaks in the C^{α} distance distributions, corresponding to non-aligned residues, were observed in all models. Such relatively big local distortions are characteristic in homology models and cannot be completely avoided. On the other hand, it was shown previously that due to the stronger conservation of protein-protein binding sites, alignments of the interface sequence fragments tend to contain fewer gaps compared to the rest of alignments (71). Thus, for further consideration we chose models with the least pronounced peaks and, thus, the lowest level of distortion in the binding region. Finally, all candidate models were visually inspected to exclude those with large distorted parts, corresponding to structural segments built *ab initio*, due to big alignment gaps, or structural defects in the template PDB files.

In the majority of cases, to build the low-energy path between two protein conformations, homology models of the same protein were used as the endpoints for NEB. The intermediate NEB structures should inevitably reproduce (some) structural properties of the endpoints. However, we realize that such correspondence is not strict and may depend on the similarity between the endpoints. In this sense, NEB models are not exactly homology models but "homology-like" models.

In our analysis, we investigated the effect of potentially mis-charged residues on the structure deformations in our models. Our benchmark set has to contain plausible (typical) homology-like models, but not necessarily high-quality ones. The initial set of models was obtained by single-template HM using NEST to mimic high-throughput real-case scenario.

The program uses the default parameters and does not allow user control of the charge state of individual residues. The inaccuracies in conformations of individual residues obtained by NEST should follow those inherent in homology models. However, this may not hold for the NEB models. Comparison of C^{α} deviations in homology and NEB structures (see 6 Å models in **Figure A-1**A) showed that the histidines in NEB models are on average more distorted than in homology models. This difference is statistically significant according to two-sample Kolmogorov-Smirnov (K-S) test at 95% confidence level.

However, such analysis cannot unambiguously answer the question whether these differences are caused by improperly set charges or the modeling procedure itself. To better understand the results for histidines, we performed the same analysis for the other 19 amino acids. In most cases (92.5%) the K-S test showed statistically significant differences between homology and NEB models (**Figure A-1**B). At the same time, the distortions in histidines were similar to the average distortions in all other types of residues in NEB models (6 Å models are shown in **Figure A-2**), confirmed by the K-S test. Thus, the modeling procedure itself (NEB) is likely the main source of the distortions.

All models were also evaluated based on C^{α} RMSD values for the interface residues alone (**Figure 2-3**). The interface residues in each of the 126 proteins in the set were extracted at 6 Å cut-off from the corresponding X-ray structures of the complexes, and superimposed with the equivalent residues in the models. The results in **Figure 2-3** show that distortions at the interfaces are smaller than in full structures, although variations in RMSD are high. The correlation coefficient between C^{α} RMSD of the entire structure and the interface is 0.72, which is statistically significant.

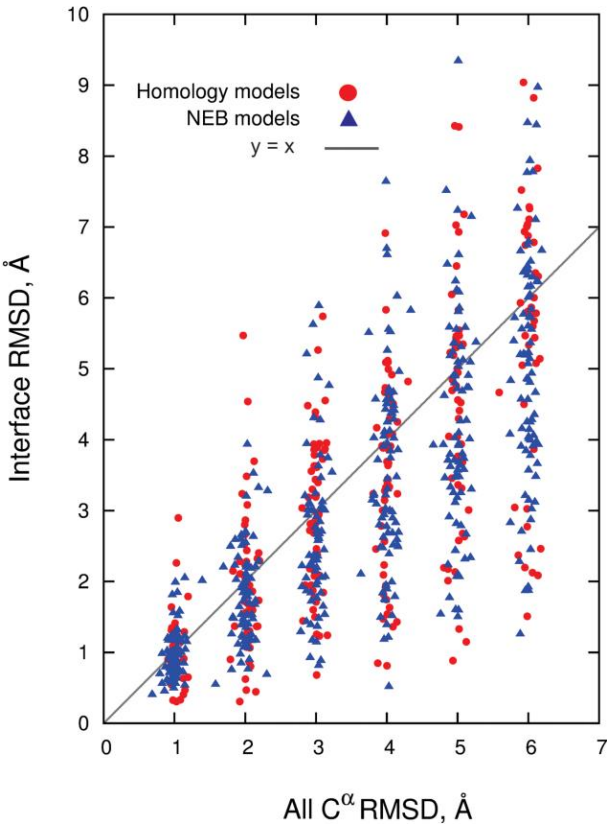

Figure 2-3: *C^{α} RMSD of the entire structure vs. interface.*

Current modeling approaches are assessed in the Critical Assessment of Structure Prediction (CASP) (72). To show that our final models are similar to those that could be obtained in real-case scenario, we compared our models with the latest CASP results in terms of correlation between overall RMSD and the global distance test GDT_TS score (73) (the score is a major criterion in CASP for accessing model quality). The GDT algorithm reflects both local and global structural distortions by several superimpositions with different cut-off values. At each cutoff, the procedure finds superimposition that maximizes the number of C^{α}–C^{α} pairs within the cutoff. If some distortions are tolerated at large cutoffs, they should still appear at smaller ones. The similarity of correlation for both

datasets (**Figure 2-4**) indicates appropriateness of our procedures for generating structures resembling the real-case scenario protein models. The data in **Figure 2-4** also show that each RMSD range contains models of different quality (wide distribution of GDT_TS scores) pointing to the overall representativeness of the set. More even distributions of C^{α}–C^{α} distances usually correspond to the lower values of the GDT_TS score.

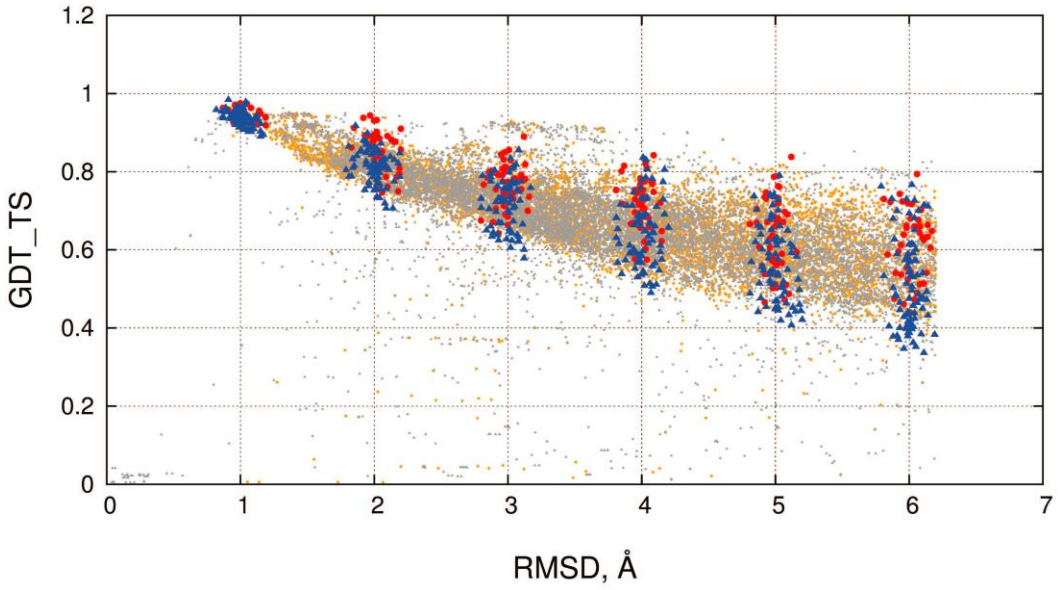

Figure 2-4: *Comparison of the quality of the distorted protein structures with CASP predictions*. CASP server predictions are in orange and human predictions are in gray. Models built in this study are in red (homology) and blue (NEB).

2.3.2 Web interface

The benchmark set of protein models for 63 binary complexes is available within the DOCKGROUND resource at http://dockground.bioinformatics.ku.edu/MODEL/request.php (**Figure 2-5**). The first four columns of the table contain brief information on the

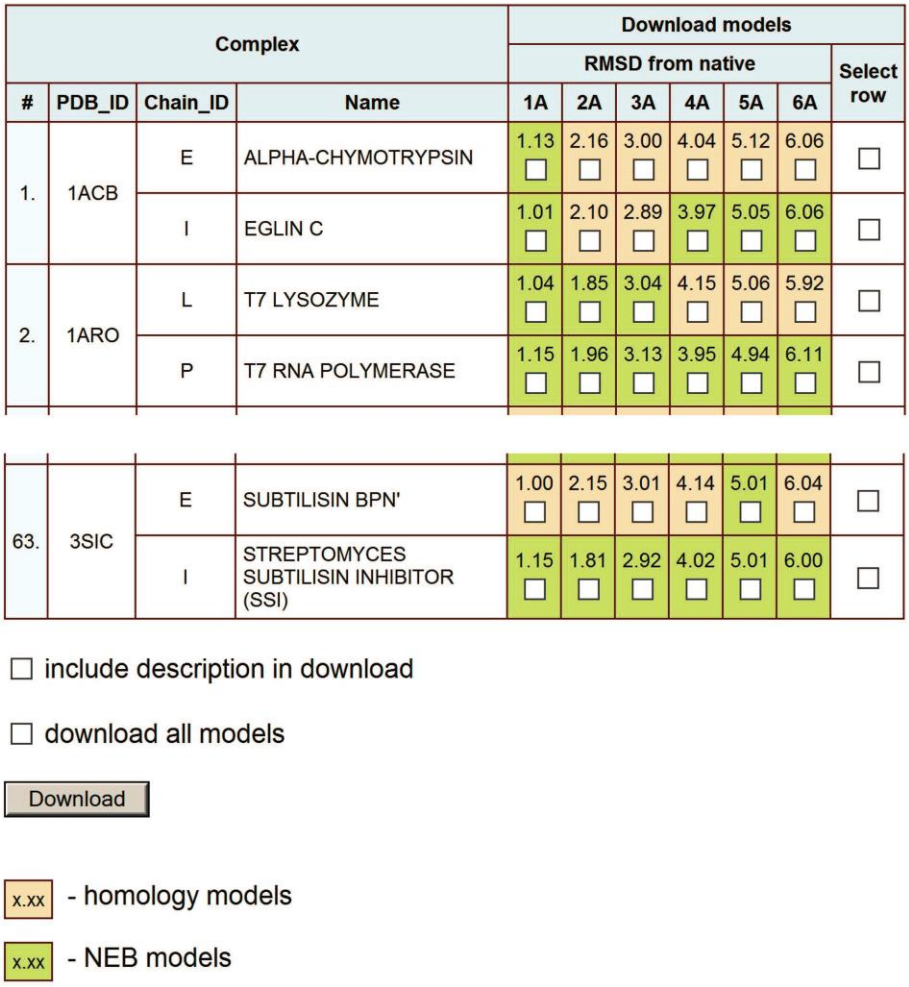

Figure 2-5: *Web interface for the benchmark set of protein models.*

complexes, followed by six columns with exact RMSD values for the generated models, along with checkboxes to select the models for customized download. Cells are colored according to the model type: orange for the homology models and green for the NEB models. An option to select all six models for a particular protein chain is provided in the last column. The *'download all models'* box downloads the entire benchmark set. The selected models are downloaded as a single ZIP file containing PDB files of the models.

The ATOM section of the model files contains only residues in the initial X-ray structure, but the entire sequence of the chain is included in the SEQRES section. Brief information on the model (the model type, HM or NEB, RMSD and GDT_TS values, templates for homology models or end-points for the NEB trajectory) is in the REMARK section.

If the box '*include description in download*' is checked, each PDB file is accompanied by a PDF file with a detailed description of the model. The PDF file includes a description of proteins used as templates for modeling as well as extensive data on the results of the model analysis. The file (example in **Figure 2-6**) contains images of superimposed native X-ray and modeled structures, information on the model type (HM or NEB), RMSD and GDT_TS values, data on the initial X-ray structure and the template used for homology modeling, target/template sequence alignment, secondary structure elements, start and end points for the low-energy path in NEB models, C^{α}–C^{α} distances for superimposed native and model structures, distribution of C^{α}–C^{α} distances for superimposed structures along the protein sequence, BLOSUM62 values for the amino acid sequence of the model, graphical representation of the secondary structure elements distribution along the protein sequence, distribution of C^{α}–C^{α} distances for superimposed native and model structures in projections onto the principal axes of the molecule, visual representation of the GDT_TS test results, and the location of interface residues.

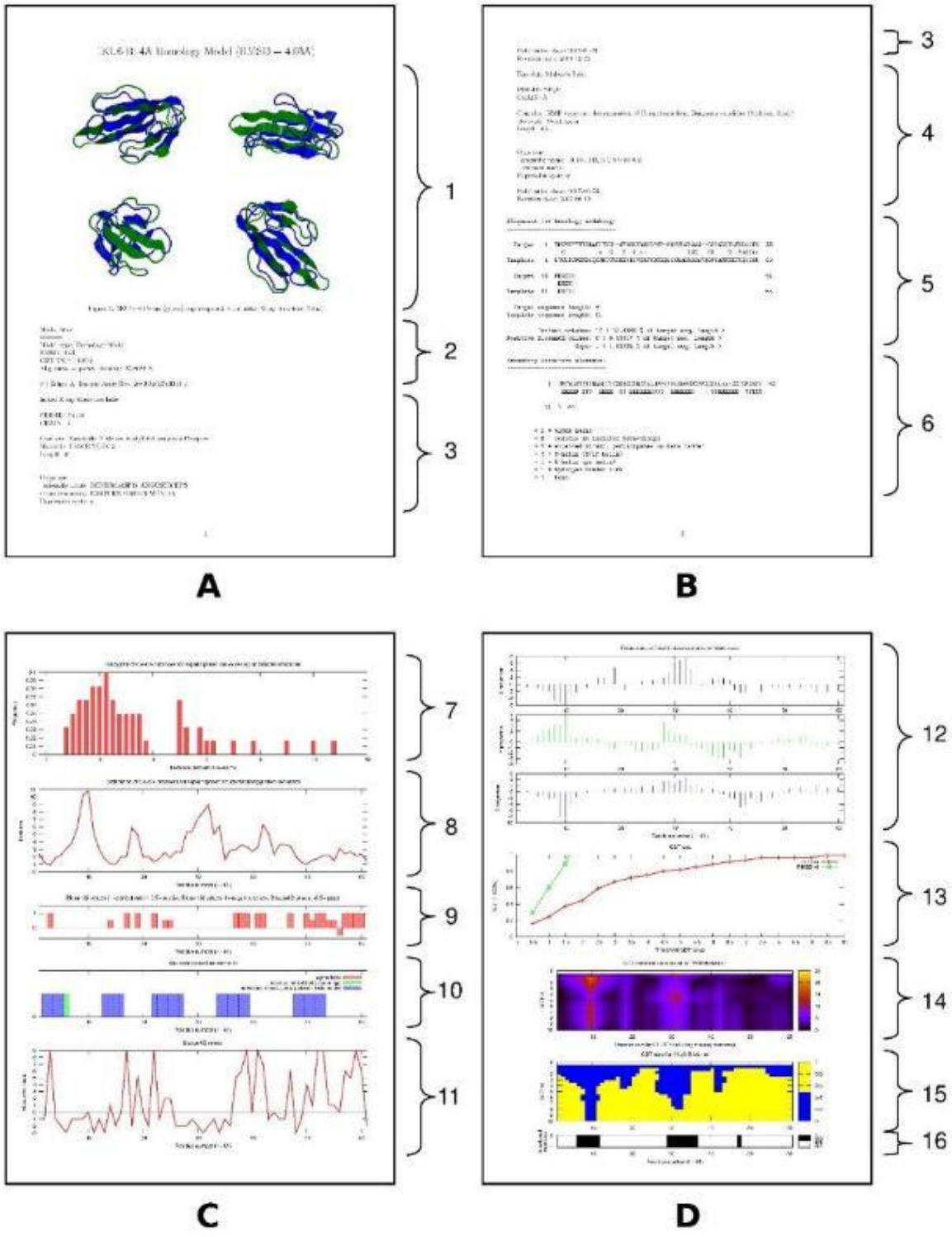

Figure 2-6: *Information on the complex from the accompanying downloadable file*. The 4 Å homology model of 1ku6, chain B, is characterized by: (1) images of the superimposed native X-ray and modeled structures; (2) information on the model type (HM or NEB), RMSD and GDT_TS

values; data on the initial X-ray structure (3) and the template (4) used in homology modeling, both retrieved from PDB; (5) target/template sequence alignment; (6) secondary structure elements in the model structure as defined by DSSP (in sections 4 – 6, PDF files for NEB models contain information on both proteins, which were used as start and end points of the low-energy path); (7) histogram of C^{α}–C^{α} distances for superimposed native (X-ray) and modeled structures; (8) distribution of C^{α}–C^{α} distances for superimposed structures along the protein sequence; (9, 11) BLOSUM62 values for the amino acid sequence of the model from the alignment (5) (sections 9, 11 are provided for HM models only); (10) graphical representation of the secondary structure elements distribution (6) along the protein sequence; (12) distribution of C^{α}–C^{α} distances for superimposed native and model structures along the protein sequence (8) in projections onto the principal axes of the molecule; (13-15) visual representation of the GDT_TS test results; (16) location of the interface residues within the protein sequence

2.4 Conclusions and future directions

The docking approaches often have to rely on modeled rather than experimentally-determined structures of the interactors. Structures of modeled proteins are typically less accurate than the ones determined by X-ray crystallography or NMR. Thus the utility of approaches to dock these structures should be assessed by thorough benchmarking specifically designed for protein models. To be credible, such benchmarking has to be based on carefully curated sets of structures with levels of distortion typical for the modeled proteins. This paper presents such a suite of models based on the benchmark set of the X-ray structures from the DOCKGROUND resource (http://dockground.bioinformatics.ku.edu) by a combination of homology modeling and Nudged Elastic Band method. For each monomer, six models were generated with pre-defined C^{α} RMSD from the native structure (1, 2, … , 6 Å). The sets and the accompanying data provide a comprehensive resource for the development of docking methodology for modeled proteins.

Our future research will focus on two major directions. First, a larger, more

representative set of protein models, based on the bound DOCKGROUND benchmark will consist of several hundreds of protein-protein complexes, with corresponding arrays of models, as opposed to 63 in the current set, which is based on the much smaller DOCKGROUND unbound benchmark. We will also explore alternative methods for model generation (e.g. threading combined with refinement trajectories), which may potentially provide a larger percentage of actual models, and decrease or eliminate the fraction of the artificially generated intermediate distorted structures. Second, we will systematically benchmark the template free and template based docking methods to determine their applicability to modeled proteins of various accuracies. The results obtained on the smaller set presented in this paper will allow comparison of the models docking to the docking of unbound X-ray structures (traditional benchmark of docking methodologies), whereas the results on the larger set will assure greater statistical significance. This will also facilitate the development of the docking approaches adequately accommodating the limited accuracy of the protein models.

Chapter 3

Protein models docking benchmark 2

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

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

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

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

Proteins. 2015; 83: 891–897

3.1 Introduction

Protein-protein interactions play a central role in life processes at the molecular level. The structural characterization of these interactions is essential for our ability to understand these processes and to utilize this knowledge in biology and medicine. Limitations of experimental techniques to determine the structure of protein-protein complexes leave the vast majority of these complexes to be determined by computational modeling. The modeling is also important for revealing the mechanisms of protein association. The protein-protein docking problem is one of the focal points of activity in computational structural biology. The three-dimensional structure of a protein-protein complex, generally, is more difficult to determine experimentally than the structure of an individual protein. Adequate computational techniques to model protein interactions are important because of the growing number of known protein structures, particularly in the context of structural genomics. The rapidly growing Protein Data Bank (PDB) provides templates for modeling of a large part of the proteome (1, 74), where individual proteins can be docked by template-free or template-based techniques (5, 6, 75-78).

However, sensitivity of the docking methods to the inherent inaccuracies of protein models, as opposed to the experimentally determined high-resolution structures, remains largely untested, primarily due to the absence of appropriate benchmark set(s). Structures in such a set should have pre-defined inaccuracy levels and, at the same time, resemble actual protein models in terms of structural motifs/packing. The set should also be large enough to ensure statistical reliability of the benchmarking results. Traditionally, the existing protein-protein benchmark sets contained, only X-ray structures (49, 58). An earlier study on low-resolution free docking of protein models utilized simulated (not

actual) protein models – artificially distorted structures with limited similarity to homology models (59).

Recently we presented a set of protein models (79) based on 63 binary protein-protein complexes from the DOCKGROUND resource (49), which have experimentally resolved unbound structures for both interactors. This allowed comparison to the "classical" problem of docking unbound crystallographically determined structures. However, only 38% of structures in the dataset were true homology models and the rest was generated by the Nudged Elastic Band (NEB) algorithm (60, 61). In this article, we report a new, > 2.5 times larger set of protein models with six levels of accuracy. All structures were built by the I-TASSER modeling package (80, 81) without any additional procedure for generating intermediate structures. Thus, the new set contains a much larger number of complexes, all of them *bona fide* models, providing an objective, statistically significant benchmark for systematic testing protein-protein docking approaches on modeled structures.

3.2 Methods

3.2.1 Selection of X-ray structures

We used the built-in engine of the DOCKGROUND resource (57) (available at http://dockground.bioinformatics.ku.edu) to generate the initial set of binary hetero complexes with moderate and high resolution (3.5 Å and better) crystallographically determined structures and a well-defined interface (≥ 250 $Å^2$ of buried solvent accessible surface area per chain, and ≥ 10 interface residues in each chain). Redundancy was

removed by the sequence identity threshold of 30% between a pair of chains. Complexes with a protein containing < 3 secondary structure elements were excluded. For computational efficiency of the subsequent modeling, we also purged complexes with monomers of substantially different sizes. In addition to the computational aspect, the level of structural accuracy characterized by the full structure RMSD depends on the size of the protein: models of shorter proteins may be significantly more distorted in terms of the secondary structure content, whereas models of longer proteins may have significantly larger local deviations. Thus, we set the maximum ratio of the protein sizes to 3, eliminating ~25% of structures from the pool of complexes (see **Figure B-1**). Finally, the set was visually inspected to remove complexes with coiled coil interfaces and those with interwoven chains. The cleaned set subjected to the modeling procedure contained 293 binary complexes.

3.2.2. Modeling procedure

The flowchart of the protocol for the model generation is shown in **Figure 3-1**. Sequences extracted from SEQRES tag of the selected PDB files were submitted to the stand-alone I-TASSER 1.0 suite of programs (80, 81). To ensure varying levels of model accuracy, the package was run several times with different cut-off values for the sequence identity between target and putative templates. We varied this parameter from 1 to 0.2 with 0.1 step plus the additional value of 0.25 introduced to diversify models built at sequence identity levels close to the threshold of homology detection (82). Even if the native structure was selected as the top-ranked template at the threading stage, it was further subjected to the

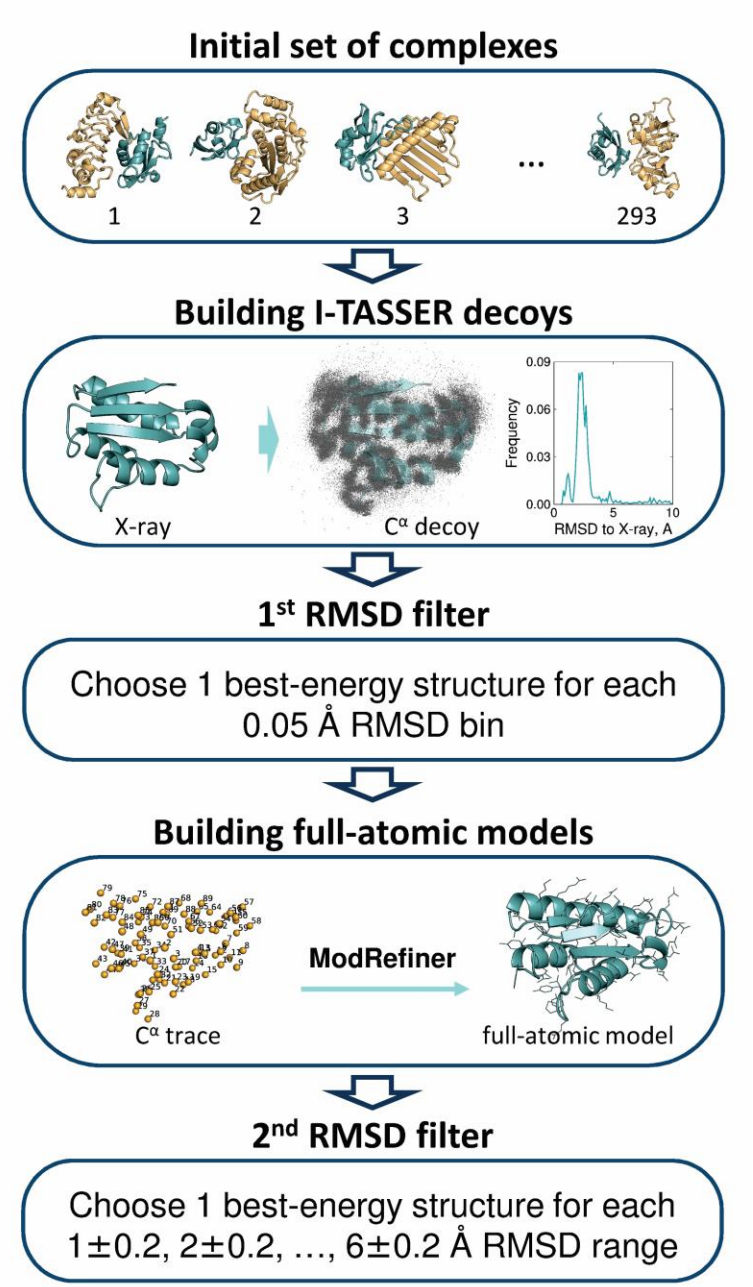

Figure 3-1: *Flowchart of the model generating procedure*.

structural assembly (along with other high-ranked templates), and subsequent model refinement (see Ref. (80) for a detailed description of I-TASSER protocol). This introduced structural variations into the final models even at the cut-off value 1.

The first modeling stage produced on average ~ 10^4–10^5 intermediate C^α models per protein. These models were grouped based on their C^α residual mean square deviation (RMSD) to the native X-ray structure using RMSD window 0.05 Å starting from 0 Å. The structure with the lowest value of I-TASSER internal energy was selected as representative for each group. To obtain the final full-atom structures, the representative models were submitted to the ModRefiner program (part of the I-TASSER software suite) (83). The C^α RMSD between full-atom models and the native structures were re-calculated and the models within the RMSD intervals 1±0.2 Å, 2±0.2 Å, …, 6±0.2 Å were selected. If several models of the same protein had RMSD in the same interval, the model with the lowest energy, according to ModRefiner, was selected. The procedure generated 3266 models for the initial set (92.8 % of the total 293×6×2 intended models). The final benchmark set was compiled from the complexes with *both* proteins having models in all six RMSD intervals (165 complexes).

3.2.3 Analysis of model structures

The relative content of the secondary structure elements in a structure was calculated as the number of residues in α-helices and β-strands in a model divided by the corresponding number in the native structure. The secondary structure residues were identified by the DSSP program (67, 84). For the analysis of the interface accuracy, models where superimposed onto corresponding X-ray structures by minimizing all C^α RMSD (85, 86), and the model/X-ray RMSD of the residues at the interface in the co-crystallized complex was calculated.

3.3 Results and discussion

The new benchmark is a significant and qualitative improvement over the previously released set 1 (79). It contains (a) a much larger number of complexes, which is important for a statistical significance of the benchmarking, and (b) all complexes in the set are true models, which is essential for the benchmarking authenticity. Based on the benchmark structures, we estimated the highest accuracy of the predicted complexes, according to CAPRI criteria.

3.3.1 Comparison with the previous benchmark set

Model benchmark 2 is significantly larger than benchmark 1. The set of models presented in this paper contains 165 complexes vs. 63 complexes in the previous set (79). Thus, the benchmarking results based on this set will be statistically more reliable (while the previous models set allows a direct comparison with the docking of unbound X-ray structures). The difference in the initial choice of complexes for the two sets (bound and unbound DOCKGROUND parts for the new and the old sets, respectively) caused a small overlap between the sets (only two complexes are shared by the sets: 1oph, chains A and B, and 2a5t, chains A and B). Because of the difference in the final model selection, the models from the new set tend to have slightly smaller TM-scores when aligned to the corresponding X-ray structure, compared to the models from the previous set (**Figure B-2**). In the previous set, preference was given to models with a more uniform distribution of distortions along the protein chain. Thus, more residues were involved in the alignment, resulting in higher TM-scores. No such filter was used to compile the new set, which is more adequate to the real case scenario of modeling/docking.

All structures in the new benchmark set are true models. After the first stage (**Figure 3-1**), the modeling protocol generated ~550 models on average per X-ray structure for each of the six RMSD bins. These models were statistically almost uniformly distributed over all accuracy levels. For each particular protein, however, structural diversity of the models depends heavily on the availability and the spectrum of the PDB templates for that protein. To build homology-like models for the RMSD values not covered by the template pool, in our previous study (79) we utilized NEB procedure (60, 61). In terms of GDT_TS score (72, 73), the NEB models were similar to the models submitted to round IX of CASP (79). However, the analysis of the NEB structures ("simulated" models) revealed that their local characteristics (deviations of C^{α} coordinates in models from the X-ray structures) are different from those observed in the real models. **Figure 3-2** shows distributions of the relative secondary structure content (see

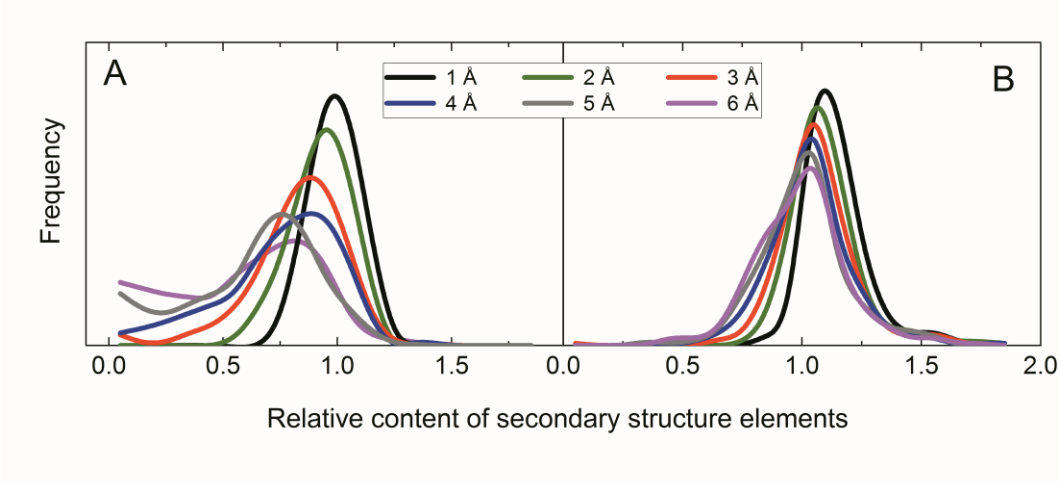

Figure 3-2: *Relative content of the secondary structure elements in protein models of different accuracy*. The plots for the old (A) and the new (B) sets show distribution of the number of residues in α-helices and β-strands in a model divided by the corresponding number in the native structure. The curves were smoothed using Savitzky-Golay method in the Origin 2015 software package.

3.2 Methods) for the models in the previous and the new sets. In the previous set (**Figure 3-2**A), the distribution peak shifts to the left and the standard deviation increases with the increase of models' inaccuracy. This indicates the reduction of the secondary structure content. The 1 Å RMSD models are closest to the native structures (corresponding distribution has its maximum close to 1, with small standard deviation). The models from the new set (**Figure 3-2**B) have more consistent distributions, with less spread in both the averages and the standard deviations. A small shift of model distributions to the right from the X-ray distribution indicates that secondary structure elements in models tend to be longer than in the native X-ray structures. This is likely to be inherent to the I-TASSER algorithm, which by design puts an emphasis on the secondary structure elements during model refinement. **Figure 3-3**A shows an example: 6 Å RMSD models of 1oph, chain B generated by the NEB and I-TASSER. It clearly shows that the secondary structures are substantially distorted in the NEB model, whereas well preserved in the I-TASSER model. The secondary structure content of the I-TASSER model is also close to the native X-ray structure as demonstrated by the highlighted portions of the sequence alignments in **Figure 3-3**B. We are not aware of any computational technique that can reliably simulate intermediate protein structures with real (e.g. homology, threading, etc.) model-like properties. Thus, we did not use any procedures to generate/simulate intermediate structures with set RMSD values between the I-TASSER decoys. Consequently, only the reduced number of 165 complexes, which have all six models for both proteins generated by the same true modeling procedure, was included in the final set.

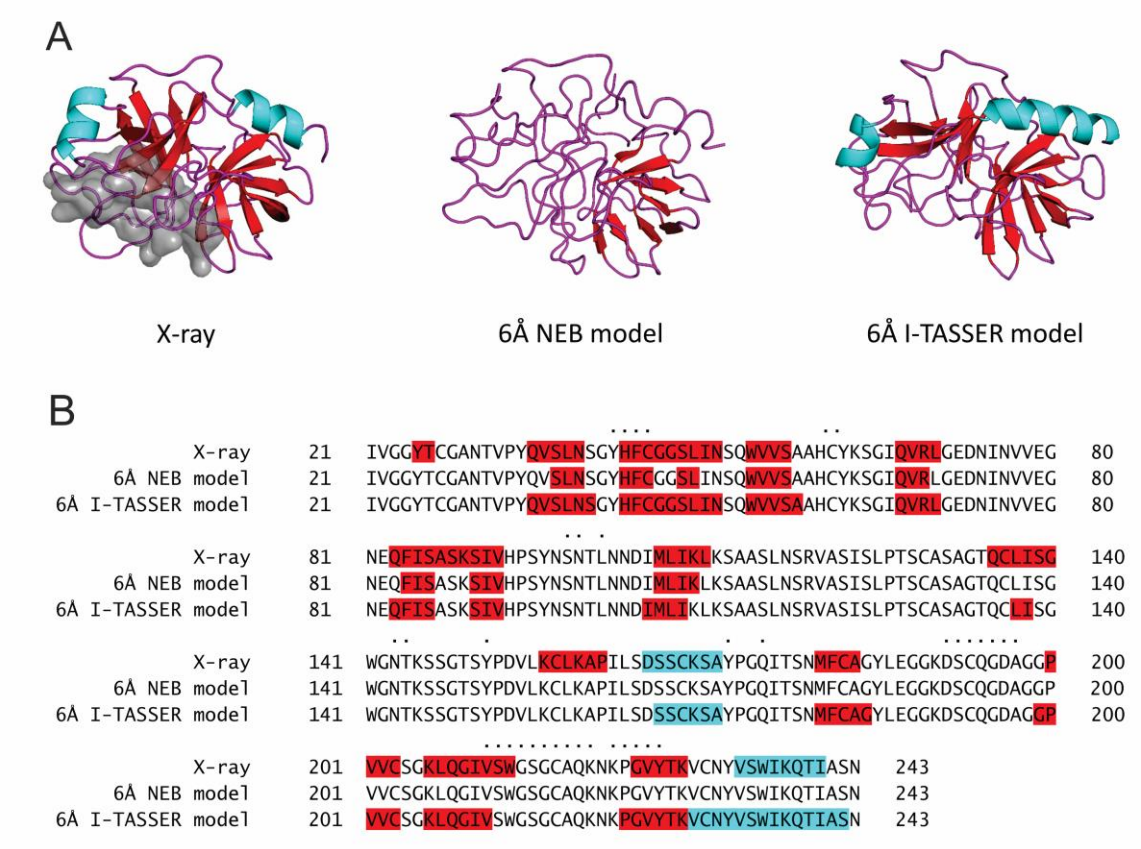

Figure 3-3*: Comparison of X-ray, NEB and I-TASSER structures*. Secondary structure content for the PDB entry 1oph, chain B; α-helices and β-strands are in cyan and red, respectively. The interface is shown by gray surface (A) and dots (B). The secondary structure is substantially distorted in the NEB model, whereas well preserved and close to the X-ray structure in the I-TASSER model.

3.3.2 Accuracy limits for the docking predictions

Although the majority of models preserve their global fold (TM-scores to X-ray > 0.5), their local structural distortion can be substantial. For 42% of modeled structures, RMSD of the interface residues is larger than the RMSD of the entire structure. At the same time, for approximately the same number of models, interfaces are far more accurate than the entire model (**Figure 3-4**). Also, average interface RMSD (open circles in **Figure 3-4**)

resemble very closely all C^{α} RMSD, indicating that, on average, interfaces are as distorted as the full structure. Interface RMSD values calculated from the alignments generated by TM-score (87) were very similar to the ones in **Figure 3-4** (data not shown).

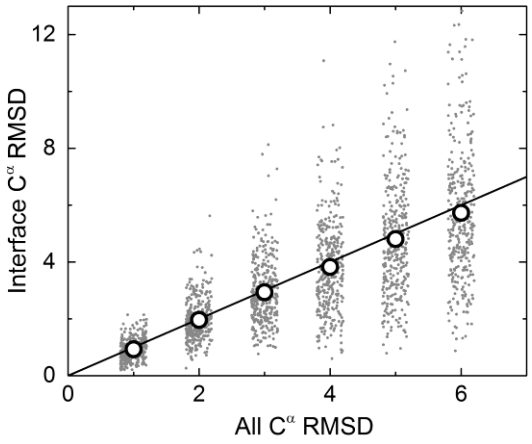

Figure 3-4: *Correlation of interface and full structure accuracy of the protein models*. The y=x line is for reference. Open circles are average interface RMSD of all models at each level of full structure accuracy.

These local structural variations limit the accuracy of docking. To qualitatively estimate that limit, we superimposed two models of the complex monomers (for simplicity, we used pair of the models with the same model-to-native RMSD) onto corresponding X-ray structures, by minimizing C^{α}–C^{α} RMSD (85, 86) (henceforth referred to as "ideal" model complexes). Thus, for each X-ray complex in the set we obtained six model structures, the quality of which was further assessed by CAPRI criteria (36) (except clashes). The vast majority of complexes built with models of ≤ 4 Å RMSD are of high and medium accuracy (**Figure 3-5**), with only seven complexes falling into the incorrect

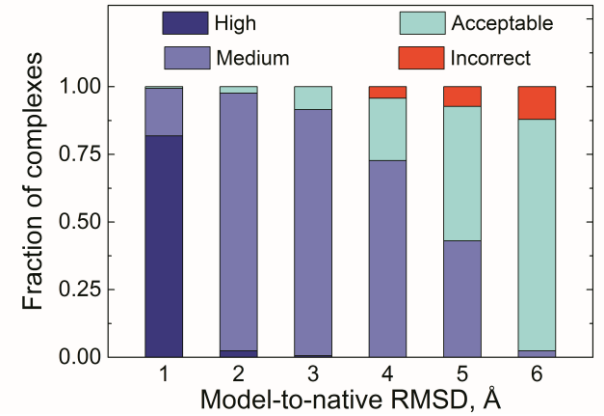

Figure 3-5: *Quality of model-model complexes according to CAPRI criteria*.

category. Models of lower accuracy produce complexes predominantly of acceptable accuracy (82 and 141 for 5 and 6 Å RMSD, respectively). Only 12 (5 Å models) and 20 (6 Å models) complexes were incorrect, mainly due to the rearrangement in packing of the interface loop(s), which leads to the distortion of the native contacts (the fraction of correctly predicted contacts drops below 10% in the incorrect models) although ligand RMSD remains < 10 Å (or interface RMSD < 4 Å). The results (**Figure 3-5** weakly depend on how a model and the X-ray structures are aligned. The interface C^{α} RMSD values for model/native structure superposition by TM-score and by RMSD minimization were similar (**Figure B-3**). For example, when alignment was performed by TM-score, the number of models in each CAPRI category changed by ~10% (**Figure B-4**).

In the real-case modeling scenario, prior to docking one would not know what protein residues belong to the interface. The whole paradigm of docking is to predict these residues (along with their contacts). Thus, all C^{α} RMSD is an appropriate measure of model's accuracy. However, we also analyzed the quality of the "ideal" model complexes in terms of CAPRI criteria, as it relates to the interface RMSD (**Figure B-5**). Complexes

of high and medium accuracy could be built from the higher accuracy protein models (1 – 4 Å global RMSD), whereas lower accuracy models (6 Å global RMSD) produced few medium accuracy complexes for small interface RMSD (1 – 3 Å).

We have also evaluated the quality of the "ideal" model complexes generated from all protein models at set accuracy levels. All complexes within each accuracy bin were either in the same or in, at most, two adjacent CAPRI quality categories. Thus, the selected model structures in our set are representative for the entire model pool (an example of the results for two complexes is in **Figure B-6**).

3.3.3 Set content and availability

The 165 complexes in the benchmark set originate from a variety of organisms (**Figure 3-6**), which ensures representativeness of the results obtained using this set. The set is available in the DOCKGROUND resource (**Figure 3-7**) as a single zip archive. The archive contains the text file with the list of monomers in the set, the README file with the explanation of nomenclature for the file and folder names, and 330 folders, one for each monomer in the set. Each folder contains six PDB formatted files of the monomer models along with the original PDB structure. Residue numbers correspond to SEQRES section of the original PDB file.

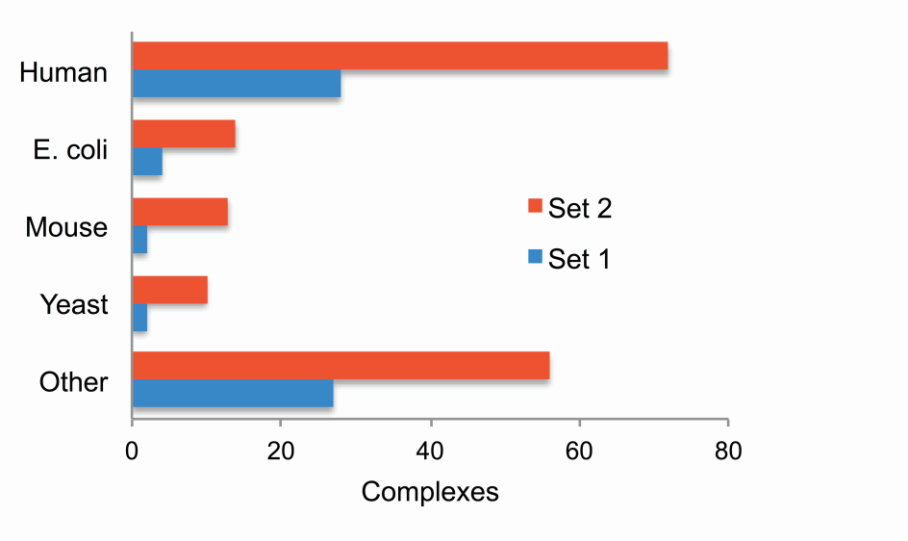

Figure 3-6: *Source organisms for complexes in the previous and the new benchmark sets*. Four most highly populated organisms are shown.

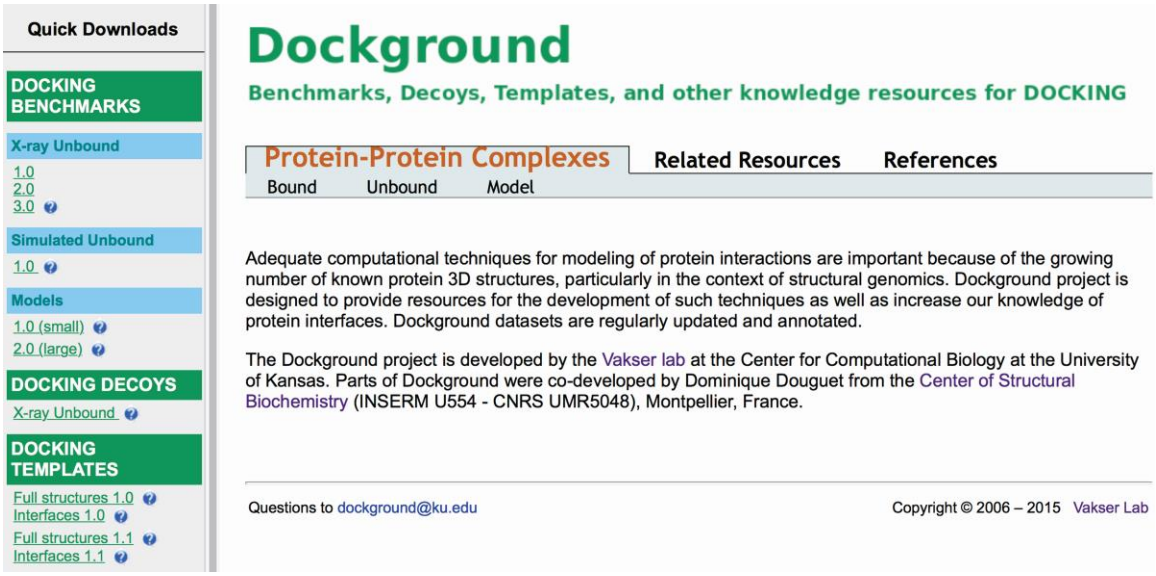

Figure 3-7: *Dockground resource for protein recognition studies*.

Chapter 4

Structural templates for comparative protein docking

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

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

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

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

Proteins. 2015;83:1563–1570

4.1 Introduction

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

Free docking methods were initially developed as *ab initio* approaches based on the physical potentials (primarily, van der Waals interactions) (6), currently increasingly supplemented by the knowledge-based approaches (statistical potentials, docking constraints, etc.) (5, 6). Despite their reasonable success, free docking methods have shown serious limitations, mostly due to the large size of the search space and structural flexibility upon the complex formation.

The template-based modeling of protein complexes relies on target/template relationships based on sequence (75), sequence/structure (threading) and structure similarity (5, 6, 75-78, 88), with the latter showing a great promise in terms of availability of the templates (4). The docking problem assumes *a priori* knowledge of the structures of the participating proteins. Thus, the docking templates may be found by structure (rather than sequence) alignment of the target monomers to the full structures of co-crystallized complexes. Evolutionary conserved surface patches may yield similar binding modes for

otherwise dissimilar proteins (23, 26) implying that docking can also be performed by the structure alignment of the target proteins with the interface parts of the co-crystallized complexes.

The key element in successful structure alignment application is the quality (diversity, non-redundancy and completeness of PDB structure) of the template libraries (generic or specific sets of 3D structures of binary complexes and/or their interfaces). Simply selecting all pairwise protein-protein complexes from PDB would produce the complete set of currently known structures. However, such "brute force" set will have many identical or highly similar complexes and some complex types will be overrepresented. The set will also contain erroneous, low-quality and biologically irrelevant structures (57, 89). Thus, groups working on structure alignment docking typically generate their own template libraries by filtering PDB in order to retain only the relevant interactions. A genome-wide study (90) utilized a library of ~30,000 full structures of template complexes extracted from PDB and PQS (91) databases with the intention (due to the termination of PQS) to switch to the PISA server (92). The PRISM docking protocol (33), where protein complexes are modeled by structure alignment of the interface regions, used a library of 8,205 protein-protein interfaces that represent unique interface architectures (93). Classification of interfaces into biological and those due to crystal packing, obligatory and nonobligatory was done by NOXclass procedure (94) and structural comparison of the initial 49,512 interfaces was performed by the geometric hashing with subsequent hierarchical clustering. A more recent study by the same group introduces a new library of 22,605 entries which is suggested for interface-based structure alignment docking (95). The interfaces in this set were extracted from the full structures using effective distance

cut-off ~ 10 Å, while an earlier study (53) indicated that the maximum success rate in interface structure alignment docking is achieved when template interfaces are extracted with a larger, 12 Å distance cut-off.

In this paper we describe our most recent sophisticated set of templates that addresses many drawbacks of the existing sets. We extract the interfaces at the optimal distance cut-off and cluster full structures and interfaces separately using various thresholds for structural similarity. Resulting datasets of full structures and interfaces, available in the DOCKGROUND resource http://dockground.bioinformatics.ku.edu, were generated using clustering threshold determined by the performance of the docking protocols.

4.2 Methods

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

4.2.1 Chain inter-penetration

Complexes from the initial set (see 4.3 Results and discussions) were checked for inter-penetration of chains by an automated procedure. For each residue of a protein in a complex, all atoms of the other protein within 6 Å distance were selected. An imaginary line through C and N backbone atoms of the residue and two half planes joined by this line were drawn. By rotating the half planes around this axis, the maximal sector free of atoms

of the second protein was determined. If the corresponding angle between the planes was < 90º, the residue was considered buried. Complexes with two consecutive buried residues in any of the chains were excluded from the set.

4.2.2 Clustering of complexes and interfaces

Pairwise structural alignment of all complexes and interfaces in the initial set (see 4.3 Results and discussions) was performed by MM-align (96) (an offshoot of the TM-align program (97) specifically designed for comparison of protein complex structures) with TM-score (87) normalized by the length of the larger complex. The TM-scores were further used to construct an undirected graph, with nodes representing individual complexes (interfaces) and edges reflecting their similarity. Two vertices in the graph were connected by an edge if the corresponding TM-score was not less than a specified threshold value TM_T, which varied in the course of computations. The resulting graph was split into clusters by a two-stage procedure. At the first stage, the graph was divided into connected components by the breadth-first search algorithm (98). The minimum cut in a graph G was defined as the minimum number of edges $k(G)$, which needs to be removed to disconnect the graph into two (connected) components. A component with n nodes ($n > 1$) was considered highly connected if the condition $k(G) > n/2$ is satisfied (99). The basic clustering algorithm by Hartuv and Shamir (99) uses the Stoer-Wagner *mincut* algorithm (100) to iteratively split a graph into subgraphs until they become highly connected. These highly connected subgraphs are induced subgraphs of the original graph and represent clusters. Several heuristics (99) were also applied to speed-up computations and to enhance the quality of clusters by adopting nodes, which became separated after the direct

application of the basic, non-enhanced approach by Hartuv and Shamir (99). The enhanced algorithm was applied to every connected component, which did not represent a complete graph. The clustering procedure was implemented as a standalone C++ program; the *igraph* library (http://igraph.sourceforge.net/index.html) was used to handle operations on the graphs.

4.2.3 Validation set of protein-protein complexes

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

4.2.4 Docking protocol

We used the template-based docking protocol similar to the one developed previously in our lab (53, 101). The procedure performs spatial rearrangement of 3D structures of two target proteins (treated as rigid bodies) to match either the entire monomers of the co-

crystallized complexes (from the full-structure template library) or their interfaces only (from the interface template library). Structural alignment of proteins was performed by TM-align (97). The resulting pool of putative matches was filtered to retain only significant matches with TM-scores of both alignments > 0.4. Models were scored by the average TM-score of both alignments. When the docking protocol was run in the benchmarking mode, the self-matches were avoided by excluding templates with both TM-scores > 0.9. Assessment of resulting models was done in terms of C^{α} ligand RMSD with receptors optimally superimposed. This RMSD definition was chosen, as opposed to the slightly different one used in CAPRI (36) (superimposition of the native interface residues in the native and the modeled complexes), for consistency with the previous studies from our lab (4, 89).

4.3 Results and discussions

4.3.1 Initial set of structures

We built two separate libraries, one consisting of the full two-chain structures and the other of the interface fragments. The flowchart of the generation process is in **Figure 4-1**. The initial pool of the X-ray structures with resolution $\leq$ 3.5 Å and buried interface area $\geq$ 250 $Å^2$ per chain was extracted for both libraries from the DOCKGROUND co-crystallized protein-protein complexes (57). We imposed an additional constraint that interfaces should consist of at least ten residues in each chain. At the point of computation, the DOCKGROUND version was based on December 2012 PDB release. Protein complexes in DOCKGROUND

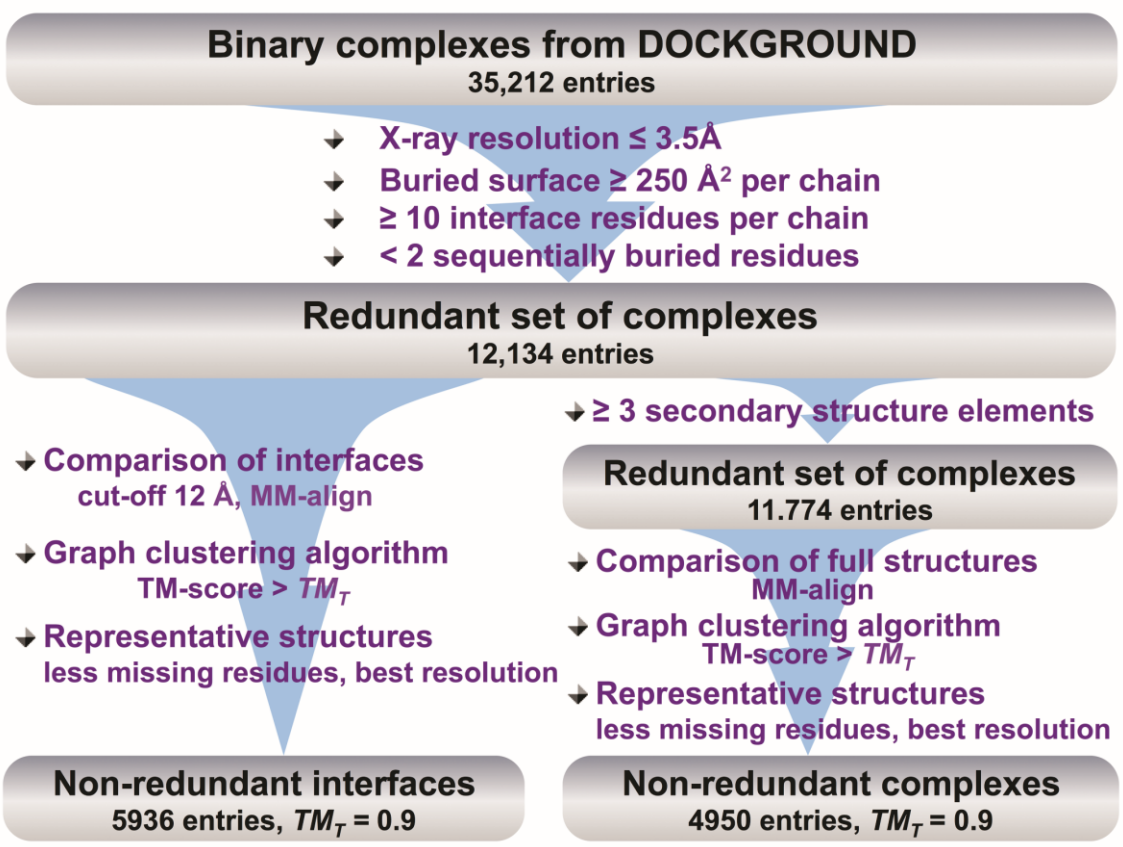

Figure 4-1: *Flowchart of algorithm for generation of full-structure and interface template libraries*.

are derived from the PDB Biological Unit files. Thus our set likely consists of biologically functional complexes, although some false positives are inevitable (89). Each complex was further checked for inter-penetration of chains by an automated procedure developed for this task (see 4.2 Methods) and complexes like the one shown in **Figure 4-2** were removed (284 entries). This resulted in 12,134 structurally redundant complexes. Interfaces were extracted from these complexes using 12 Å distance cut-off between heavy atoms of residues belonging to different chains. The extracted interfaces were clustered and analyzed in terms of structural connectivity and docking performance in order to choose

the clustering parameters. Full structures were further filtered by an additional requirement that at least three regular (> 4 residues each) secondary structure elements be present in each interacting protein. The secondary structure elements (α-helices and/or β-strands) were detected by the DSSP tool (67). The resulting reduced set of 11,774 complexes was subjected to the clustering and analysis procedures, same as the interfaces.

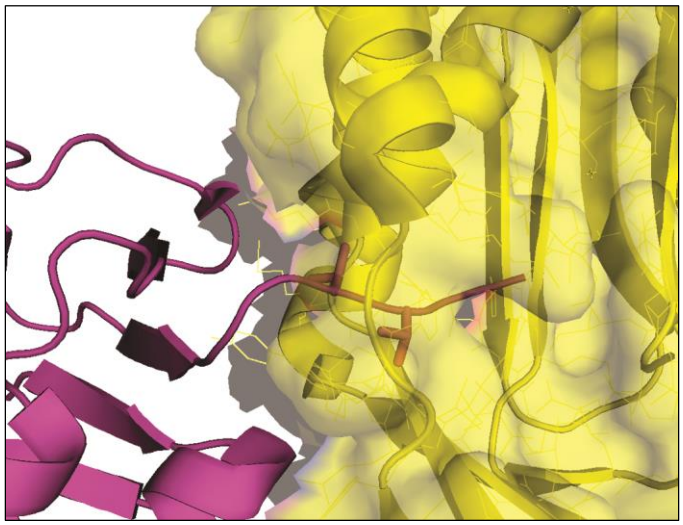

Figure 4-2: *Example of a "bad" complex*. Chains D and E from 1fma are shown in magenta and yellow, respectively, with penetrating chain removed by the automated procedure described in the text. Buried Val and Thr residues at the C-terminal of chain F identified by the procedure are shown as sticks (the last two residues at the terminus are Gly).

4.3.2 Connectivity of the structural space of protein-protein complexes

To eliminate structural redundancy, the intermediate sets of 11,774 complexes and 12,134 interfaces had to be clustered by some measure of structural similarity. In this study, for such a measure we used TM-score (87). TM-score (ranging from 0 to 1) is produced by the TM-align routine (97), which was previously successfully employed in the template-based

docking (4, 53, 89, 101, 102), although other programs for the structural alignment with their own structural similarity scores were utilized by others (103, 104).

For efficient clustering, it is useful to understand how similar complexes and interfaces are connected in the structural space. We analyzed similarity graphs built at different threshold values of TM-score (TM_T, see 4.2 Methods) in terms of the size of the connected components (initial, first-approximation clusters with some missing edges between the nodes) and the clustering coefficient (the probability of neighbors of a given node to be connected between themselves (105)). As seen in the main panels of **Figure 4-3**, a substantial fraction of connected components belongs to either isolated nodes

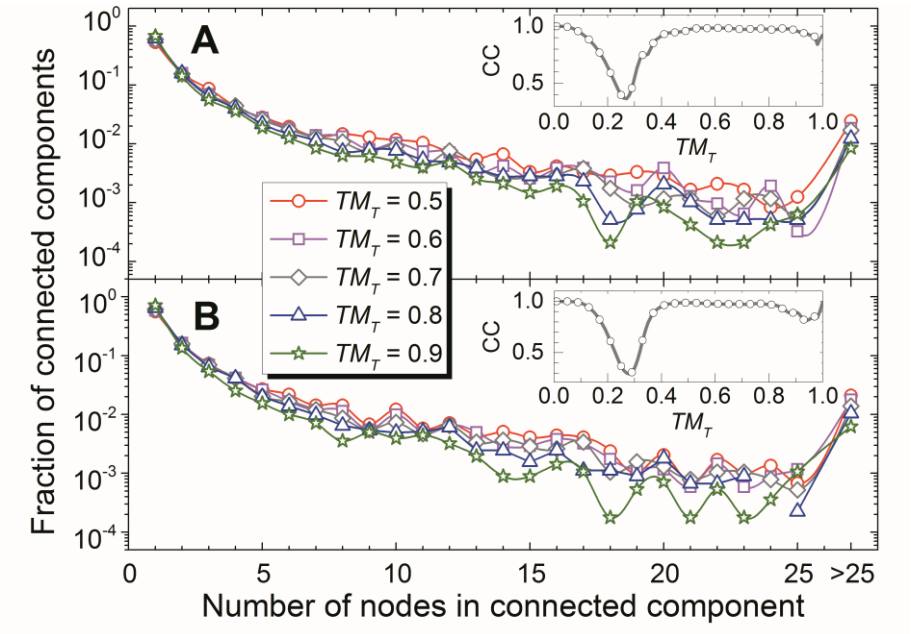

Figure 4-3: *Properties of similarity graphs*. (A) Protein-protein complexes and (B) protein-protein interfaces. The main panels show distributions of connected component size at different thresholds of the clustering TM-score (TM_T). The inserts display dependence of clustering coefficient CC on TM_T.

(53 – 68 % of complexes and 56 – 72 % of interfaces, depending on TM_T) or pairs of connected nodes (14 – 16 % of complexes and 13 – 16 % of interfaces) and cannot be split further. Interestingly, this property is persistent within a broad TM_T range.

In terms of clusters, the clustering coefficient can be viewed as a measure of the extent to which the groups of connected nodes in a graph are close to the complete graphs (or ideal clusters), in which every pair of nodes is connected by an edge. Inserts in **Figure 4-3** show the clustering coefficient of graphs for full structures (panel A) and interfaces (panel B) in the full TM_T range from 0.0 to 1.0. Due to the random matches of short structural fragments, TM-score seldom gets very close to 0. Thus, at low TM_T, there are edges in the graph between almost all nodes making the graph close to complete and resulting in high clustering coefficient of almost 1. The similarity graphs then will be close to complete graphs comprising almost entire set of complexes/interfaces (left sides of inserts in **Figure 4-3**). When TM_T increases, the clustering coefficient decreases dramatically and has a minimum at $TM_T = 0.27$ for the full structures and $TM_T = 0.28$ for the interfaces, which is consistent with a previous estimate of the average TM-score for random match 0.17 (87). With further increase of TM_T, the statistical significance of a structural match increases as well. Starting from $TM_T \sim 0.5$ (the lowest TM-score for proteins with similar folds (97, 106)), the clustering coefficient stops growing and remains unchanged (~0.98 for both full structures and interfaces) up to $TM_T \sim 0.9$ for full complexes and $TM_T \sim 0.8$ for interfaces. High values of the clustering coefficient within such TM_T ranges suggest that the graph nodes are clustered in almost optimal way. The decrease in the clustering coefficient for $TM_T > 0.9$ stems from small structural differences (especially in the loops) often present in different PDB files for otherwise identical or very similar (in

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

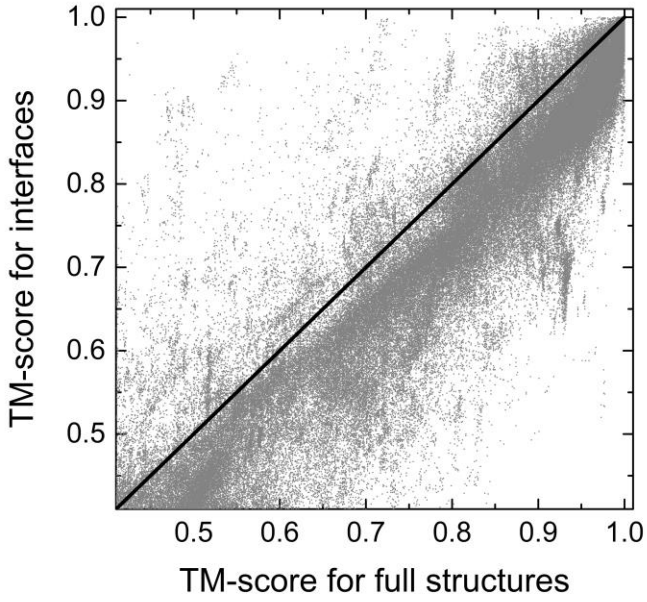

Figure 4-4: *Correlation of protein-protein and interface-interface TM-scores.*

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

4.3.3 Analysis of clusters

We utilized a clustering approach, which, first, divides the similarity graph into "loosely" connected components and then further splits them into tightly connected clusters (see 4.2 Methods). **Figure 4-5** shows how the number of connected components N_{CC} and the number of resulting clusters N_C varies with TM_T. Since the majority of the connected components cannot be split further (as shown in **Figure 4-3**), N_C is only slightly larger than N_{CC} for most values of TM_T. The relative increase in the number of clusters is ~ 5% for TM_T = 0.6 – 0.9 for both full complexes and interfaces (green lines in **Figure 4-5**). This correlates with the high values of clustering coefficient in these TM_T ranges (**Figure 4-3**).

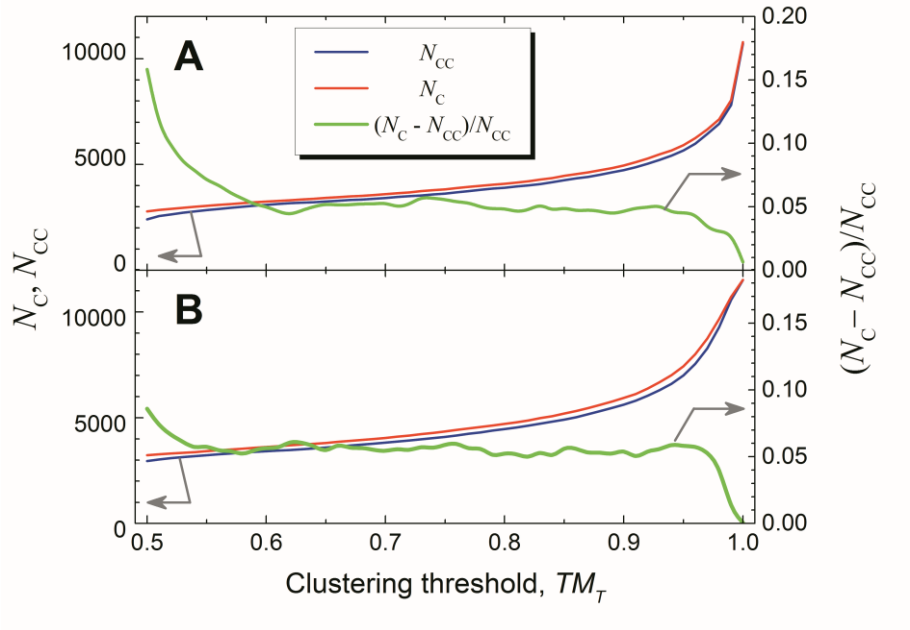

Figure 4-5: *Number of connected components and clusters as a function of clustering threshold*. (A) Protein-protein complexes, and (B) protein-protein interfaces. N_{CC} is the number of connected components, and N_C is the number of clusters. Green lines (scaled to the right-hand axes) show the relative increase in the number of connected graph parts after splitting the connected components into tightly connected clusters.

Finally, we checked the quality of the resulting clusters at different TM_T by calculating TM-scores between members of each cluster in order to detect pairs of nodes within a cluster that lack an edge (TM-score < TM_T). We found that only ~ 3% of the final clusters have pairs of dissimilar complexes (or interfaces) with TM-score < TM_T (circles in **Figure 4-6**). The clustering algorithm we employed allows the final clusters to have as little as 50% of edges (compared to complete graphs) (99). However, the analysis of the actual clusters showed that the fraction of dissimilar pairs in the vast majority of clusters is < 30%, with the mean value close to 10% (box-and-whiskers plots in **Figure 4-6**). The quality of clusters remains roughly the same within the full TM_T range, with minor

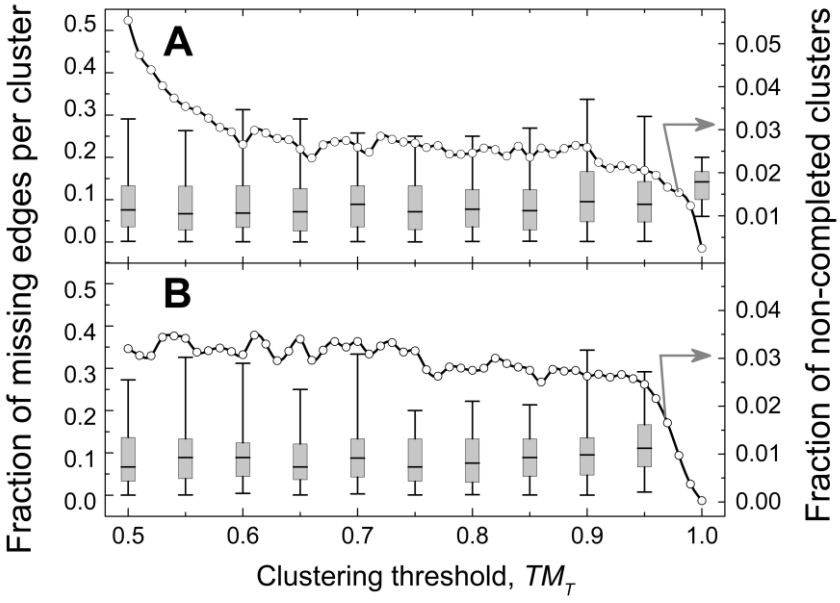

Figure 4-6: *Quality of clusters at different clustering thresholds*. (A) Protein-protein complexes, and (B) protein-protein interfaces. Distributions of missing edges per cluster are shown as box-and-whickers plots with horizontal lines for minimal, maximal and median values in the distributions and boxes containing second and third quartiles of data. The circles (scaled to the right hand axis) show how the fraction of clusters, which are not complete sub-graphs of the initial similarity graphs, depends on TM_T.

variations at $TM_T < 0.6$ for full complexes (**Figure 4-6**A) and $TM_T > 0.9$ for both complexes and interfaces (**Figure 4-6**A and B). Sequence identity, in general, follows the same trend, i.e. ~70% of the clusters have the minimal sequence identity between the members > 90%. However, in some extreme cases (e.g., the cluster of 49 complexes from RNA polymerase), there are cluster members with sequence identities ≤ 30%.

4.3.4 Template libraries in docking: selecting optimal parameters

The success of docking depends heavily on the diversity of the template library. On the other hand, the running time of the template-based docking is directly proportional to the size of the template set. Thus, an optimal template library should be large enough to maximize the docking success rate, but should not contain excessive entries, which only marginally improve the performance. This approach to the optimal library is different from the one used to compile PRISM interface library (95), where optimization was performed according to the quality of the resulting clusters. For practical docking purposes, our choice was rather to optimize the performance of the modeling of complexes based on our templates, through benchmarking.

We tested 26 full-structure and 26 interface libraries, generated at TM_T ranging from 0.50 to 1.00 with 0.02 step, on a non-redundant set of 293 hetero complexes (see 4.2 Methods). Success rate was defined as a ratio of targets, for which at least one model had interface C^{α} ligand RMSD < 5 Å, to the total number of targets (4, 89). To exclude the influence of the scoring scheme, we calculated success rate for the entire set of models, although results for the top ten models, ranked by the average TM-score, were also obtained (not shown separately due to qualitative similarity to the all-models results).

Templates that were similar to a particular target (both TM-scores for a target-template pair exceed 0.9) were left out from the consideration. Such exclusion of similar structures leads to success rates higher than reported in a recent benchmark study (107) where the main focus was on docking in the "twilight zone" of low target/template similarity (sequence identities between target and template < 30%).

Results of the test are shown in **Figure 4-7**. As one can expect, more entries in the template library (higher TM_T values) lead to higher success rates of the docking. Such monotonic behavior holds for almost entire TM_T range from 0.5 to ~0.9. For $TM_T > 0.9$ the success rate is largely saturated. A slight increase in the success rates at $TM_T > 0.9$ is an artifact of our procedure. While TM-scores used in the clustering are obtained by the MM-align for complexes, TM-scores for exclusion were produced by TM-align for separate monomers. At certain TM_T (especially at values close to the similarity criteria), a cluster

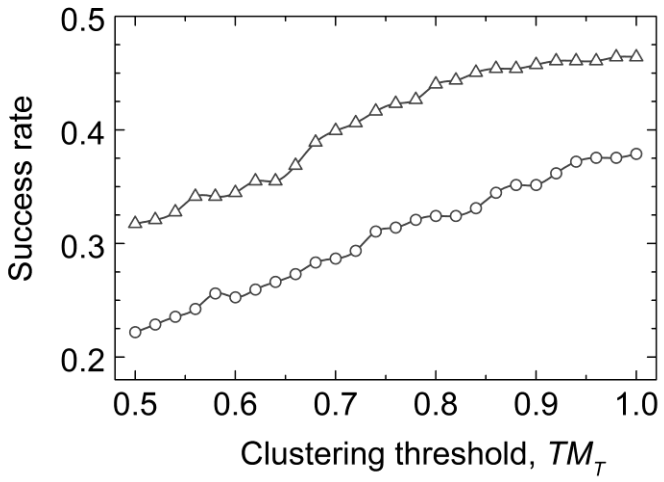

Figure 4-7: *Performance of structure alignment at different clustering thresholds*. Full-structure (circles) and interface (triangles) libraries were generated at different threshold values. Success rates were calculated for the entire pool of structures excluding templates similar to the target (see text).

may have a representative, identified as the similar structure (thus excluding entire cluster from consideration) and other structures with one out of two TM-scores of TM-align slightly less than the similarity criteria (this could have TM-score of the MM-align exceeding TM_T). These structures at higher TM_T can split into a separate cluster with representative identified as non-similar and thus yielding good-quality model (in total, there are seven such cases in the full-structure set and one case in the interface set).

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

The number of clusters (and, consequently, computational time) starts growing exponentially at TM_T > 0.9 (**Figure 4-5**). This, along with the results in **Figure 4-7**, suggests that the optimal library ought to be generated at TM_T = 0.9.

4.3.5 Availability of the template and the benchmark sets

A representative complex from each cluster at TM_T = 0.9 was selected based on the best resolution and the smallest number of missing residues. The resulting sets of 4,950 full-structure complexes and 5,936 interfaces (representing ~40% of folds in SCOP (108) and

in a more recent ECOD database, http://prodata.swmed.edu/ecod) are available on the Web within the DOCKGROUND resource at http://dockground.bioinformatics.ku.edu, under "docking templates" tab. The sets are downloadable as zip archives (one for full structures and the other for the interfaces) each containing folders "templates," "targets" and "info." The folder "templates" contains two PDB-formatted files of atomic coordinates per library entry. The files are named by the original PDB file, from which the entry was extracted, as follows:

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

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

interface RMSD > 5 Å and the minimal of the TM-scores of the components > 0.8, indicating high similarity to a wrong template. The “difficult_targets.txt” files in the “info” folders contain the list of such targets.

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

Chapter 5

Modeling complexes of modeled proteins

Ivan Anishchenko[1], Petras J. Kundrotas[1], Ilya A. Vakser[1,2]

[1]Center for Computational Biology and [2]Department of Molecular Biosciences,
The University of Kansas, Lawrence, Kansas 66047, USA

Submitted

5.1 Introduction

Protein-protein interactions (PPI) drive many cellular processes. Structural characterization of PPI is important for better understanding of these processes and for our ability to manipulate them. Experimental techniques for structure determination of PPI have limited capabilities. The X-ray crystallography, the major source of today's knowledge on atomic-level structures of PPI, accounts only for 26% of known PPI in *E. coli* and 6.7% in human (4). Thus, the structure of most known protein interactions has to be determined by computational methods for PPI modeling (protein docking) (5).

Modern protein docking methods generally belong to two major categories: (a) free docking, where relative positions of the two proteins are systematically sampled and, generally, no information other than the structure of the two proteins, is assumed to be known *a priori*; and (b) template-based docking, where the prediction is made according to sequence or structure similarity of the target proteins to the ones in co-crystallized complexes (6-9). Although the co-crystallized protein-protein structures are still few, our earlier study (4) showed that valid templates for PPI modeling by structure alignment can be found for almost all known PPI that involve proteins for which the structure is known or can be built by homology (templates are available for the homology modeling of a significant part of the individual proteins (1)). A serious obstacle to the docking of protein structures is the conformational changes upon complex formation (34). Whereas the ultra-low resolution docking may be applicable to cases with large inaccuracies (35), the problem is explicitly addressed by docking methods, which allow structure flexibility (36). The community-wide experiment on Critical Assessment of Predicted Interactions, CAPRI (36, 109) offers an objective comparative evaluation of existing docking approaches.

Most proteins in interactome are themselves models of limited accuracy (6). An important question asked by protein modelers (110), and biological researchers in general, is: what kind of structural information can be obtained from the docking of protein models and what is the reliability of such predictions? Protein models were shown to have significant utility in protein-ligand interactions and characterization of functional sites (71, 111, 112). Protein-protein docking of models by information driven approach was validated on a set of CAPRI targets (47). High-resolution free docking was recently tested on a diverse set of protein models to reveal that meaningful predictions can be obtained even for models with significant distortions, although at significantly lower success rates (113). The systematic benchmarking on arrays of protein models at different accuracy levels was performed by the ultra-low resolution approach (59). The results showed that such docking determines the gross structural features of the complex for a significant portion of protein models, including highly inaccurate ones. However, because of limited availability of templates for modeling of individual proteins, the study was based on "simulated models" of the proteins, which reflected the general structural accuracy of the homology models, but were not necessarily structurally similar to those. The study also was restricted to the ultra-low resolution free docking (35), effectively predicting the binding sites only.

In this paper, we address the problem of models' utility in protein docking using our recent benchmark sets of actual protein models (79, 114). The quality of free and template-based docking predictions built from these models was thoroughly assessed to reveal the tolerance limits of docking to structural inaccuracies of protein models. The predictive power of currently available rigid-body and flexible docking approaches is

similar (36). Thus in this study we used basic rigid-body approaches, developed in our group, that would clearly reveal the general similarities and differences in free and template-based docking performance depending on the modeling accuracy of the interacting proteins.

The results show that the existing docking methodologies can be successfully applied to protein models with a broad range of structural accuracy; the template-based docking is much less sensitive to inaccuracies of protein models than the free docking; and docking can be successfully applied to entire proteomes where most proteins are models of different accuracy.

5.2 Methods

5.2.1 Benchmark sets of protein models

The sensitivity of docking protocols to the inherent inaccuracies of protein models was tested on two specialized and carefully curated benchmark sets (79, 114). Both sets are similar by design and contain binary protein-protein complexes with each monomer represented by six models with increasing levels of inaccuracy (model-to-native C^{α} RMSD within 1 ± 0.2 Å, 2 ± 0.2 Å, … 6 ± 0.2 Å intervals). The first, smaller set of 63 complexes (Benchmark 1) is based on the unbound docking benchmark set 3 of X-ray structures from the DOCKGROUND resource (http://dockground.compbio.ku.edu). It was built by a combination of homology modeling by NEST (66), simulated annealing (SA) and Nudged Elastic Band method (NEB) (60, 61) as implemented in Amber10 package (69). About one third of the structures in the set are homology models, and the rest are generated by NEB

and SA (simulated models). Benchmark 1 also contains X-ray unbound structures of both interactors, which allows comparison of models docking to the traditional unbound docking. The second, larger set of models (Benchmark 2) is based on the bound DOCKGROUND part (57) and thus lacks the unbound X-ray structures (114). However, the number of complexes in the Benchmark 2 is significantly larger than in the Benchmark 1 (165 vs. 63), which should increase the statistical reliability of the benchmarking. Also, importantly, all models in the Benchmark 2 are bona fide models, generated by I-TASSER (80, 81), thus adequately reflecting the reality of modeling in the real case scenario.

5.2.2 Docking protocols

The free docking was performed by the FFT (Fast Fourier Transform) program GRAMM (12, 115) at low resolution, with 3.5 Å grid step and 10° angular interval. Top 100,000 matches were scored by the Miyazawa-Jernigan statistical potential (116) and clustered.

The template-based docking was performed by full structure alignment protocol (101), using template library (117) of 4,950 co-crystallized binary complexes from DOCKGROUND (57). Target proteins were structurally aligned to the template monomers by TM-align (97). The resulting models (target/template TM-score > 0.4 only) were scored by the average of the two TM-scores (87).

5.2.3 Metrics for docking accuracy

The accuracy of the predicted model-model complex combines the accuracy of the docking with the accuracy of the monomers modeling. Thus the docking assessment in this case is more complicated than in traditional docking of the X-ray structures.

To quantify the difference between docking modes, we calculated the fraction of shared residue contacts. Each configuration i of the protein-protein complex is characterized by a set S_i of N_i pairwise contacts

$$S_i = \{(a,b)_1, (a,b)_2, \ldots (a,b)_{N_i}\}, \tag{5-1}$$

where (a, b) is a pair of residues a of the receptor (larger protein in the complex) and b of the ligand (smaller protein in the complex) interacting across the interface. The similarity between configurations i and j, FSC_{ij} (fraction of shared contacts), can be calculated as the Jaccard index of the two sets S_i and S_j

$$\mathrm{FSC}_{ij} = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}. \tag{5-2}$$

As opposed to ligand RMSD (RMSD between ligands in two docking modes with receptors superimposed), the FSC_{ij} between similar docking modes does not have substantial variation from complex to complex (**Figure C-1**). The fraction of native contacts (f_{nat}), in CAPRI definition (118), cannot be directly used for pairwise comparison of model-model docking predictions because of the required reference set of the native interface residues/contacts, which varies in different docking models. In this regard, FSC_{ij} (Eq. 5-2) can be considered a modification of f_{nat}, such that the number of shared contacts is normalized by the number contacts in either of the two models, making the score symmetric ($FSC_{ij} = FSC_{ji}$).

5.2.4 Assessing docking predictions by CAPRI criteria

The docking predictions were assigned to four accuracy categories (high, medium, acceptable, incorrect) according to the CAPRI criteria (118) (**Table C-1**). A docked model-model complex was compared to a reference complex built by superimposition of two protein models with the same model-to-native RMSD onto the corresponding monomers from the native X-ray structure (114). Such "ideal" model-model complexes provide an estimate of the highest level of accuracy, which can be achieved in the rigid-body docking of protein models. The co-crystallized X-ray structures were also used as the reference structures.

5.2.5 Assessing template-based docking predictions

In addition to the CAPRI criteria and TM-score, we assessed the quality of the template-based docking using FSC-score, defined similarly to FSC_{ij} (Eq. 5-2)

$$\text{FSC-score} = \frac{|S_{\text{templ}} \cap S_{\text{model}}|}{|S_{\text{templ}} \cup S_{\text{model}}|}, \tag{5-3}$$

where S_{templ} and S_{model} are contact sets in the template and in the model built from that template, respectively. However, as opposed to FSC_{ij} the FSC-score needs an additional rule for finding contacts shared by the two complexes with different monomers. We considered the template contacts (a_{templ}, b_{templ}) and the model contacts (a_{model}, b_{model}) shared if in the alignments used to build the model, residues a_{templ} and b_{templ} are aligned to the

residues a_{model} and b_{model}, correspondingly. Equation 5-3 then can be rewritten in a simpler form

$$\text{FSC-score} = \frac{N_{\text{shared}}}{N_{\text{templ}} + N_{\text{model}} - N_{\text{shared}}}, \tag{5-4}$$

where $N_{\text{templ}} = |S_{\text{templ}}|$ and $N_{\text{model}} = |S_{\text{model}}|$ are the total number of contacts in the template and the model, respectively. Almost all models with FSC-score ≤ 0.05 are incorrect (**Figure C-2**) and thus were excluded from further consideration. Such simple filtering not only eliminated > 50% of bad predictions, but also ensures that any template-based docking prediction has a certain amount of contacts. Unlike the CAPRI criteria, the FSC-score can be used for the assessment of the docking models in the real-case modeling scenario when the reference native structure is not available.

5.3 Results and Discussions

5.3.1 Detection of near-native solutions

Protein interactions are driven by a funnel-like energy landscape, with the native structure of the complex inside the funnel (119). Thus the success of docking depends directly on the ability to detect the funnel. Since the energy landscape is a function of atomic coordinates, the landscapes of inherently inaccurate protein models differs from the landscapes of the corresponding X-ray structures. Thus the question is: whether the funnels can still be detected in the case of models. We addressed this question by analyzing spatial

distributions of top 1000 free and all template-based docking predictions for each model accuracy level for Benchmark 2 (**Figure 5-1**).

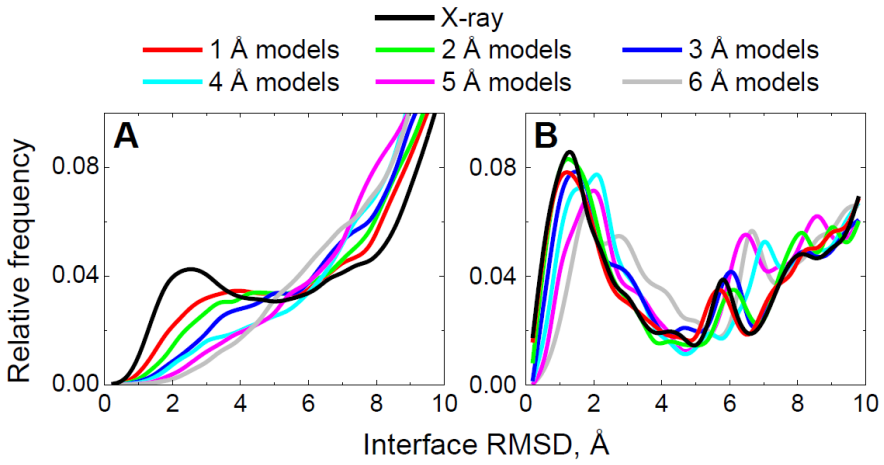

Figure 5-1: *Distribution of near-native and false-positive docking matches according to the accuracy of protein models*. Top 1000 free docking (A) and all template-based docking (B) predictions, for each of the 165 complexes from the Models Docking Benchmark 2, at each at the six accuracy levels, were compared to the corresponding "ideal" complexes (see 5.2 Methods) in terms of I-RMSD. In the docking of the X-ray structures, comparisons were made to the corresponding native X-ray structures.

The bimodal distribution of interface RMSD between docking predictions and corresponding reference complex indicates detection of the funnels by the free (119) and the template-based (120) docking. As expected, the native peak is clearly observed if bound X-ray structures are docked by both free and template-based methods (black lines in **Figure 5-1**). With the decrease of models' accuracy, the peak in the free docking results diffuses and is no longer detectable for models with distortions ≥ 4 Å RMSD (**Figure 5-1**A). The near-native cluster of the free docking solutions, corresponding to this peak, decays exponentially (**Figure C-3**) due to large structural distortions at interface regions in the

dataset (RMSD between interface C^{α} atoms of model and the native structures for about half of the models is larger than RMSD calculated over all C^{α} atoms, see inset in **Figure C-3**).

The template-based docking yields I-RMSD distributions with the distinct peak, corresponding to the near-native solutions, at all levels of monomer accuracy (**Figure 5-1**B). Unlike the free docking, which is based on protein surface complementarity (and as a consequence, is sensitive to the local structural distortions), the template-based algorithm accounts for the entire protein fold. Thus the observed bimodality reflects the link between protein structure and function, which implies similar binding modes of structurally related proteins.

Because of the high sensitivity to the local structural inaccuracies, the success rate of free docking decreases much more rapidly with the increasing level of model inaccuracy, compared to the template-based docking (dark gray and hatched bars in **Figure 5-2**). Interestingly, for some complexes, free docking yielded good predictions for the models, but not for the X-ray structures (hatched bars in **Figure 5-2**A), due to the degree of noise inherent to free docking. The template-based docking has almost no such cases (the hatched parts are hardly distinguishable in **Figure 5-2**B), which is related to the high degree of template conservation.

In the above analysis, all docking predictions of models were compared to the corresponding "ideal" complexes (see 5.2 Methods). If the native bound conformations were used instead, docking success rates decrease slightly along with the quality of models assessed by the CAPRI criteria, but the observed trends remain the same (**Figure C-4**).

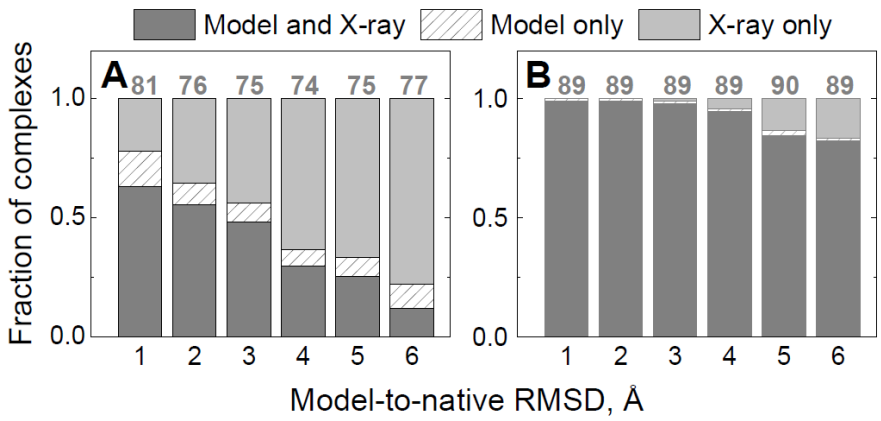

Figure 5-2: *Docking success rates for protein models compared to the success rates for X-ray structures*. Successfully predicted complexes (those for which at least one acceptable or better quality prediction is among top 10 docking poses), in free docking (left hand panel) and template-based docking (right hand panel) are in dark gray. Complexes with successful predictions by X-ray docking only are in light gray. Complexes with successful predictions by model docking only are in hatched bars. The quality of the model docking was accessed relative to the "ideal" complexes (see 5.2 Methods). The data are normalized by the total numbers of complexes in all three categories shown on top of the bars.

5.3.2 Stability of the solutions space

In addition to the analysis of top predictions, built into the traditional "success rates" metrics (**Figure 5-2**), analysis of a much broader range of predictions adds to the assessment of the docking quality. Docking with consistent hits near the correct prediction is more reliable than the one where the hits are widely dispersed.

In template-based docking, the number of predictions is limited by the number of detected templates, which varies from zero to several hundred. Most good (acceptable or better quality) model and X-ray docking predictions were built on the same templates (**Figure 5-3**A). Despite large local distortions (inset in **Figure C-3**), global folds of the

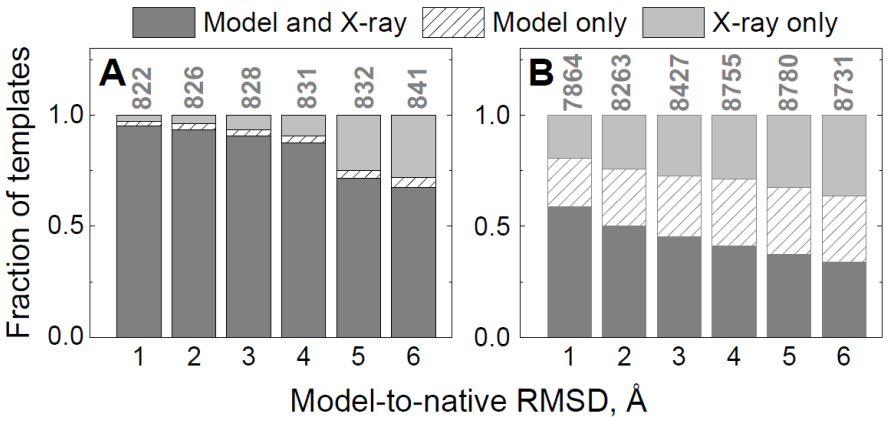

Figure 5-3: *Conservation of templates in template-based docking of models*. Dark gray bars show templates common for the docking predictions of X-ray structures and docking predictions of the corresponding models. Light gray bars show templates for the docking of X-ray structures predictions only. Hatched bars show templates for the docking of models predictions only. Data for good (acceptable or higher quality) predictions (A), and incorrect predictions (B) is normalized by the total number of templates shown on top of the bars.

native structures are preserved in the majority of the models in the benchmark set (inset in **Figure C-6**). Most templates yielding good models in X-ray template-based docking have target/template TM-scores > 0.6 (**Figure C-5**). Distortions in models, albeit reducing target-template structural similarity (distributions for good models in **Figure C-5** shift to the left as distortions in monomers increase), are, in most cases, not sufficient to bring the model under the detection threshold (TM-score 0.4). Thus, if a template yields a good X-ray docking prediction (typically with high TM-score), there is a high probability that the same template would be selected in the model docking, yielding a good docking prediction as well, although often in a lower quality category. Significant drop in the template-based docking performance is correlated with the loss of native folds at large inaccuracy levels (**Figure C-6**).

Templates for incorrect docking predictions usually share less structural similarity to the target, resulting in TM-scores closer to the detection threshold, already seen for the X-ray-template pairs (**Figure C-5**). Consequently, even a slight TM-score reduction for the model/template pairs, makes ~20% (1 Å models) to 36 % (6 Å models) of such templates not detectable (light gray bars in **Figure 5-3**B). Interestingly, a significant amount of templates (22% for the 1 Å models to 39% for the 6 Å models) is detected only in the model docking (hatched bars in **Figure 5-3**B). 95% of those templates have model-template TM-score 0.40 – 0.53, which means that their detection in the model docking is due to a small increase in the TM-score above the detection threshold due to "favorable" local structural variations in the models.

Shared templates (dark gray bars in **Figure 5-3**) yield model docking predictions with patterns of interface residue contacts similar to those in the X-ray predictions, irrespective of the monomers' accuracy (**Figure 5-4**). Local structural inaccuracies in the models of interacting proteins make some contacts disappear, or result in new contacts (average f_{nat} values corresponding to 1 Å and 6 Å models in **Figure 5-4**B are 0.78 and 0.44 respectively, implying loss of 22 and 56% of native contacts). Still, predominantly non-zero FSC_{ij} values suggest preservation of the docked monomers position with the increasing model inaccuracy. Such trend is similar for both good (acceptable and higher quality) and bad (incorrect) predictions with slightly less pronounced effect for the latter (**Figure C-7**), and with a fraction of the bad predictions (1.6% for the 1 Å models, to 6.1% for the 6 Å models) losing native contacts completely (minor peak in distributions at ~0, in **Figure C-7**).

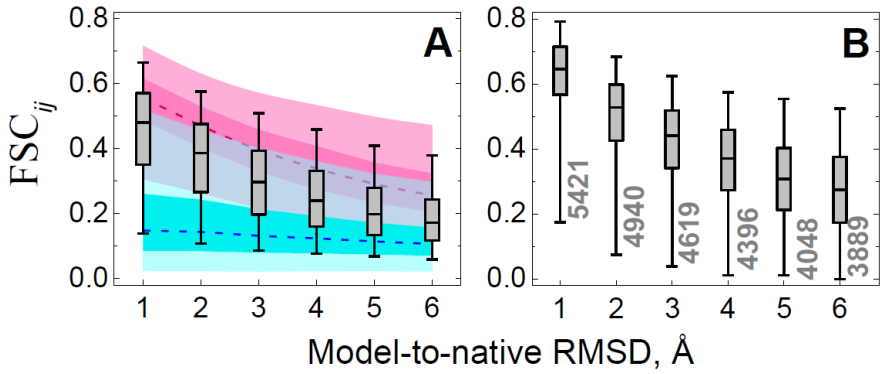

Figure 5-4: *Comparison of free and template-based docking of models predictions with the docking of X-ray structures predictions in terms of fraction of shared contacts*. For each level of model accuracy and each complex in the set, docking of X-ray structure prediction with the maximum fraction of shared contacts FSC_{ij} (Eq. 5-2) was used for comparison with each of the top 1000 free docking of models predictions. The resulting 165×1000 FSC_{ij} scores were plotted as gray box-and-whisker diagrams, separately for each distortion level (A). Box areas and whiskers contain 25 – 75% and 5 – 95% of data, respectively (outliers not shown). Lower bounds (blue) were estimated using 1000 randomly selected matches from top 100,000 free docking of models predictions. Upper bounds (red) were evaluated on a 1000-matches subset among 100,000 free docking of models predictions with the maximum FSC_{ij} similarity to the top 1000 docking of X-ray structures predictions. Darker and lighter areas of the upper and lower bounds correspond to boxes and whiskers respectively, and the dashed lines indicate medians. For the template-based docking (B), only pairs of model and X-ray predictions that share the same template (dark gray bars in **Figure 5-3**, and numbers at the whiskers in this figure) were considered. Upper and lower limits for the template-based docking were not estimated due to statistically insufficient number of docking predictions.

The templates' conservation and the models they produce are illustrated in **Figure 5-5** for the two variable domains (chains L and H) in F_V fragment of the anti-dansyl monoclonal antibody 1dlf. Immunoglobulins are widely represented in PDB, thus template-based docking of this structure results in a large pool of ~600 models. A tight

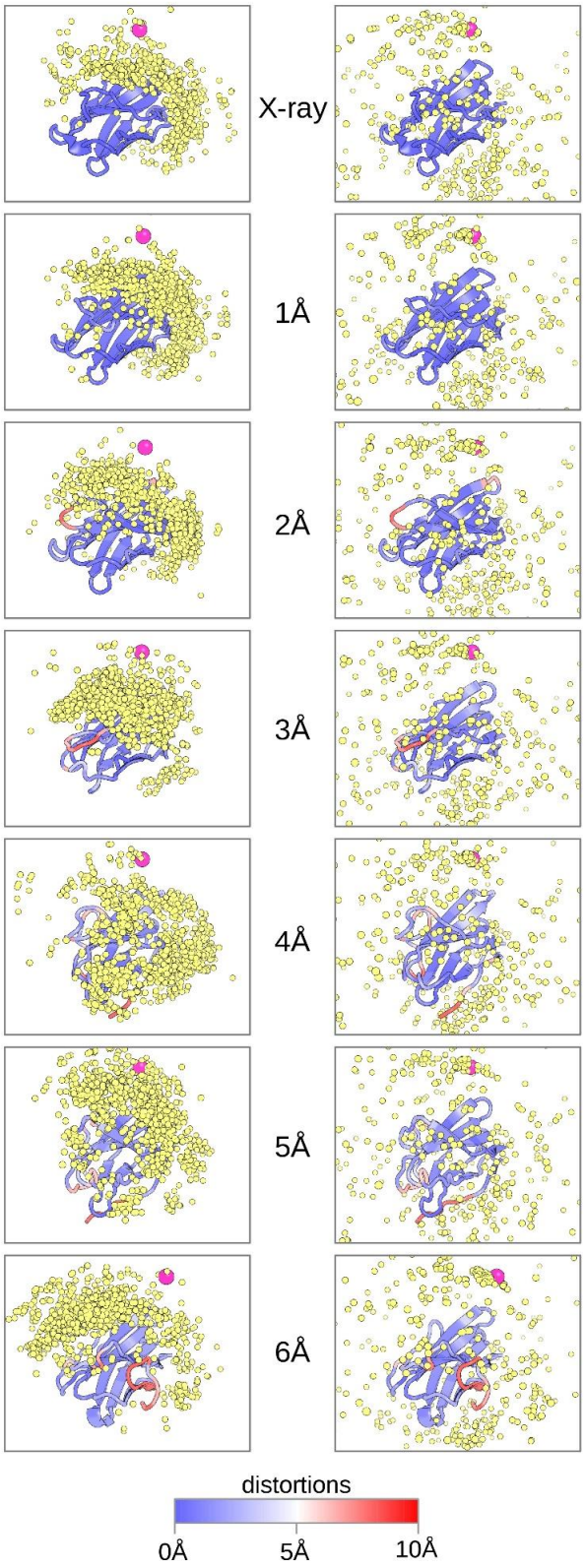

Figure 5-5: *Example of clustering in free and template-based docking*. Co-crystallized structures of the 1dlf chains H and L, along with their models at six levels of accuracy from the Benchmark 2 were used in free (left-hand panel) and template-based (right-hand panel) docking. Top 1000 free and all template-based predictions are shown. Predicted matches are shown by yellow spheres, corresponding to the ligand (L chain) native interface center of mass. Magenta sphere corresponds to the native interface. The receptor structure (H chain) shown in cartoon is color-coded to reveal the location of distortions and their level. Distortions are measured as C^{α}–C^{α} distances calculated from RMSD-based superposition of the model onto the corresponding monomer from the co-crystallized complex.

cluster of near-native solutions is preserved at all accuracy levels, whereas non-native matches have essentially random pattern with only a fraction of models shared between all accuracy levels.

In free docking, the pool of initial models is much larger that in the template-based docking. Thus only top 1000 solutions were selected for scoring and clustering. Since templates are not utilized in this method, we used a different approach to analyze the stability and conservation of the docking solutions.

Connectivity properties of similarity graphs constructed from top or randomly selected 1000 predictions were almost independent of the level of monomer distortion, albeit with substantial differences between these two groups of graphs (**Figure C-8**). Pairwise comparison of distributions of cluster sizes for six accuracy levels and the X-ray structures for each complex (in total, $\binom{7}{2} \times 165 = 3465$ comparisons) indicated that only ~16 % of distribution pairs can be considered significantly different (comparison was done using two-sided Mann-Whitney *U* test (121) at 0.05 significance level). This implies that the number and the size of clusters in top 1000 predictions do not vary significantly with the distortion level as well, albeit with some preference for the clusters originating from more distorted protein models to become less populated (169 distribution pairs with clusters growing in size when distortion level increases, versus 389 opposite cases, as was identified by the one-sided Mann-Whitney *U* test).

However, in terms of the fraction of shared contacts, the fee docking of models differs from the "native ensemble" (top 1000 X-ray free docking predictions) to a significantly larger extent, then in template-based docking (**Figure 5-4**) The divergence of the model predictions from the native ensemble increases with the increase of the model

inaccuracy (**Figure 5-4**A). Same trend is also observed for the upper bound of the average similarity (**Figure 5-4**A, red) indicating that even in the best case scenario local distortions in monomers allow only partial recovery of the residue contacts in the X-ray predictions. On the other hand, randomly selected docking models (**Figure 5-4**A, blue) share considerably less similarity to the native ensemble than the top 1000 predictions (box-and-whiskers in **Figure 5-4**A) for all model accuracy levels, implying preservation of some contacts from the native ensemble in all model predictions. Thus, local distortions in monomers substantially reduce the number of near-native solutions – **Figure C-3**). However the clustering pattern remains almost unchanged (**Figure C-8**).

A clustering example of top 1000 free docking matches is shown in **Figure 5-5** by the same variable fragment of the anti-dansyl monoclonal antibody. Most predictions are aggregated in the proximity of the large groove in the receptor (preserved in all models), formed by a concave β-sheet. The pool of the docking solutions in all cases is obviously non-random and some degree of similarity can be observed between docking of the co-crystallized X-ray structures and models.

5.3.3 Template-based or free: which is preferable?

Protein docking methodologies are usually tested on unbound protein X-ray structures, with the challenge to accommodate the conformational difference from the bound protein. In this study, we challenged the docking programs much further since our protein models are, on average, significantly more different from the native bound structures, than the unbound X-ray structures. In the widely used protein-protein unbound X-ray docking Benchmark 4 (58) only 24 out of 176 complexes (14%) are considered difficult, with I-

RMSD in unbound-bound superposition > 2.2 Å. In comparison, in our Models Benchmarks 1 and 2, 71% and 65% of complexes, respectively, are that different from the bound X-ray structure. Thus, the unbound X-ray structures are easier to dock than the protein models (**Figure C-9**).

In free docking, the conformational deviation of the unbound X-ray structures can be mitigated by the low-resolution approach (115) (albeit at the loss of atomic details in the docked complexes). Naturally, the low-resolution approach should help in the docking of protein models as well. Indeed, while high-resolution docking outperforms the low-resolution one on the X-ray structures and on models with small RMSDs to the native structures (**Figure 5-6**), starting from 2 Å RMSD (which roughly corresponds to the transition from "easy" to "difficult" unbound docking) the low-resolution docking systematically has higher success rate. Nevertheless, both low- and high-resolution docking have steeper decline of success rates with the increase of models' inaccuracy than the template-based docking (**Figure 5-6**). In our implementation, target-template similarity is assessed for the global fold, which determines the robustness of the docking solutions with respect to the local structural distortions in the protein models.

Interestingly, success rates of free and template-based dockings saturate differently with the increase of the number of considered top solutions (**Figure 5-7**). Rapid saturation of the template-based success rates indicates that the scoring scheme (see 5.2 Methods) almost always finds the correct template among top 10 detected templates (only 8 complexes have their good models ranks reduced to the top 1000 predictions at 6 Å accuracy). Moreover, ~60% of the complexes retain a good model, although often in the

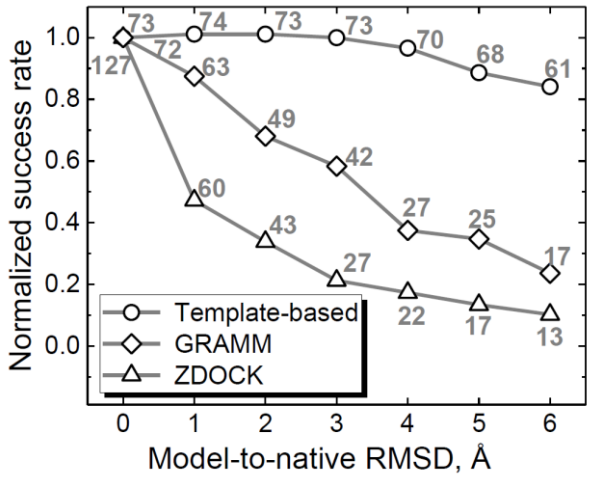

Figure 5-6: *Normalized success rates for the template-based and free docking*. The free docking at low resolution was performed by GRAMM, and at high resolution by ZDOCK 3.0.2 (122). The complex was predicted successfully if one out of top 10 predictions was correct (acceptable, medium or high quality). All success rates are normalized by the ones for the co-crystallized X-ray structures. The numbers above the data points show the absolute number of successful docking outcomes (out of 165 complexes in Benchmark 2).

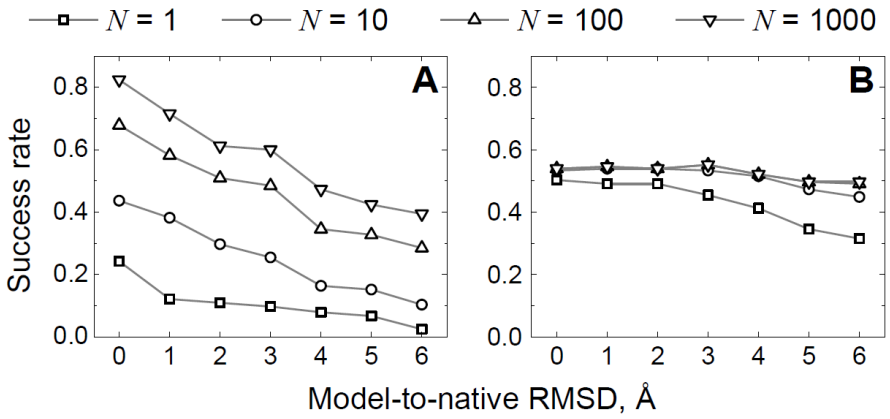

Figure 5-7: *Docking success rates for different number of top solutions*. The successful prediction was defined as one correct structure (acceptable, medium or high quality) in top N predictions. The rates are shown for free (A) and template-based docking (B).

lower quality category (larger *N* in the top *N* criterion), at the top of the list for all accuracy levels. Contrary to that, the free docking success rate consistently increases if more predictions are selected for the final analysis indicating a large room for improvement in the scoring of the initial scan stage models.

5.4 Conclusions

We conducted systematic benchmark studies of template-based and free protein-protein docking methodologies on comprehensive sets of monomer protein models with a full array of accuracy levels. The results unambiguously show that the existing docking methodologies are applicable to protein models, even in case of relatively low protein structure accuracy. The template-based docking is significantly less sensitive to the distortions in protein models compared to the free docking. The template-based methodology yields model-model complexes with high degree of similarity to the docking predictions of the native X-ray structures, and its success rate is almost independent of the accuracy level (at the tested range). The results suggest that for protein models the use of template-based docking is preferable provided a good template can be found. The scoring scheme based on similarities of global folds reliably finds such good templates. However for some complexes (e.g., those with alternative binding modes or in the twilight zone of target/template similarity) such scoring may not lead to a correct solution. Thus, further improvement of the scoring would be useful in order to increase confidence in model-model docking.

Free docking is essential for a number of important applications, including detection of transient complexes (123), modeling of protein association (124), and such.

With the increase of the distortions in monomer models, the free docking performance significantly deteriorates. Still, the low-resolution in free docking provides a degree of tolerance to local model distortions (success rates are non-negligible even at 4–6 Å distortion). However, to increase docking reliability, the free docking scoring needs much greater improvement than the scoring for template-based predictions.

The scoring for both template-based and free docking can be complemented by various constraints (e.g. automated literature search (125), evolutionary inferred residue-residue contacts (126, 127), chemical shifts (128), etc.). With the continuous growth of publicly available information on protein interactions, the utility of such constrains will be increasing, expanding our abilities to reliably model protein interactome.

Chapter 6

Structural quality of unrefined models in protein docking

Ivan Anishchenko[1], Petras J. Kundrotas[1], Ilya A. Vakser[1,2]

[1]Center for Computational Biology and [2]Department of Molecular Biosciences,
The University of Kansas, Lawrence, Kansas 66047, USA

To be submitted

6.1 Introduction

Structural characterization of protein-protein interactions is essential for our ability to understand and manipulate biomolecular processes. Structures of protein-protein complexes are more difficult to determine experimentally than the structures of the individual proteins. Moreover, proteins potentially participate in multiple protein-protein interactions, making the number of protein-protein prediction targets much larger than that of the individual proteins. Thus, only a fraction of known protein-protein interactions has experimentally resolved structures (4). Modeling in essential for generating such structures, as well as for learning the principles of molecular recognition and structure/function relationships. Prediction of protein-protein structures (protein docking) aims at determining the spatial arrangement of the target proteins within the complex, given the known structure (experimental or modeled) of the individual proteins.

Modeling protein-protein complexes at atomic resolution with all degrees of freedom taken into account in global docking scan is computationally prohibitive. Thus, most docking programs perform initial search using simplified representation of protein structures. For example, in a large class of *ab initio* (or template-free) docking methods (6) the initial rigid-body search for surface complementarity is performed by correlation using Fast Fourier Transformation (FFT) on digitized protein images (12). The inter-penetrations of the protein structures in the FFT-based methods are explicitly penalized by proper weighting of the grid points corresponding to the protein's interior with respect to the surface regions (12). In practical docking, the target proteins are either experimentally determined unbound structures, with conformations different from those within the complex, or computational models, often of limited accuracy. Thus the rigid-body docking

has to have some tolerance to the steric clashes. This may lead to non-physical overlaps between atoms in the resulting models of the complex. To remove the clashes, rigid-body moves with the side-chains repacking may be sufficient for proteins with moderate conformational changes upon binding (37-39). For difficult targets, backbone flexibility can be accounted for by low-frequency normal mode analysis (40-43), backbone perturbations using the fold-tree-based method (44), and semi flexible refinement of interface residues in torsion angle space followed by Cartesian dynamics refinement in explicit solvent (45).

Another class of docking methods belongs to the template-based category, exploiting structural similarity between the target proteins and the existing protein-protein complexes in Protein Data Bank (PDB) (129) (*templates*) (4, 33, 90). The initial models of the complex are generated by structural superimposition of the target proteins onto the co-crystallized proteins in the template. This procedure, as opposed to the *ab initio* methods, does not explicitly constrain structural penetrations. Thus the initial template-based models may have severe structural overlaps. Despite increasing popularity of the template-based docking, a systematic analysis of the clashes, which can be used in development of procedures for their removal, is lacking. In this paper, we compare and analyze the extent of clashes in unrefined template-based and template-free docking models. The results show that, contrary to the common expectation, in acceptable and better quality docking models, the clashes in template-based docking are comparable to those in free docking, due to the overall higher quality of the template-based docking predictions. This suggests that the free docking refinement protocols can in principle be applied to the template-based docking predictions.

6.2 Methods

Ab initio (template-free) docking was performed by the rigid-body FFT protocol as implemented in GRAMM (12, 35). Top 100,000 predictions from the scan stage with 3.5 Å spatial grid step, and 10° angular step were rescored by Miyazawa-Jernigan (MJ3h) statistical potential (116). 1,000 matches with the lowest MJ3h energy were retained for further analysis.

The template-based docking protocol, developed previously in our lab (53, 101) utilizes experimentally determined structures of protein-protein complexes (*templates*) for full structure alignment (FSA) to the target proteins by TM-align (97). The algorithm performs a systematic search for best templates in the full-structure template library (117) composed of 4,950 co-crystallized binary complexes from DOCKGROUND (57). Models with any of the two TM-scores < 0.4 and the fraction of contacts shared by the target and the template < 0.05 were previously shown to be unreliable (4, 130), and thus were removed from the final pool of predictions.

The free and template-based docking was performed on the unbound set 3 (49) and model set 2 (114) from DOCKGROUND (http://dockground.compbio.ku.edu). The unbound set consists of 102 protein-protein complexes and the unbound structures of each protein. The set of protein models is composed of 165 binary protein-protein complexes with each monomer represented by six models with increasing levels of inaccuracy (model-to-native C^{α} RMSD 1 ± 0.2 Å, 2 ± 0.2 Å,…, 6 ± 0.2 Å), and the co-crystallized bound structure of each complex as reference.

6.3 Results and discussions

Typically, protein docking starts with simplified low-resolution representation of protein structures, and the increase of the resolution at the subsequent refinement (5). This simplification is based on the existence of low-resolution recognition (35, 131-135), which allow prediction of the gross structural features of protein-protein complexes, even from highly simplified protein structures. In terms of the intermolecular energy landscape, the low-resolution recognition reflects the existence of the intermolecular energy funnel (119), which guides the two interacting proteins along the binding pathway. Various methods of measuring the size of the docking funnel (119, 136, 137) are consistent in their estimates for its upper bound of ~10 Å ligand RMSD (L-RMSD). This is also consistent with the range of the electrostatic and desolvation energies in protein-protein complexes (137). From the docking perspective, it is unlikely that an initial model outside the docking funnel can be further refined towards a near-native prediction. Thus, in this study, we focus on the docking predictions that can be potentially refined.

To quantify the amount of clashes in the docking models, an intersection of the van der Waals volumes, ΔV_{vdw}, of the two interacting proteins was calculated for their projection onto a cubic grid with the 1.0 Å step (van der Waals radii values according to Ref. (138)). To obtain a quantity, which is independent of the interface size, we normalized ΔV_{vdw} by the average solvent-accessible surface area buried upon complex formation (this quantity hereafter referred to as *average penetration*, d_{av})

$$d_{\text{av}} = \frac{\Delta V_{\text{vdw}}}{\left(\Delta\text{SASA}_{AB} + \Delta\text{SASA}_{BA}\right)/2}, \tag{6-5}$$

where ΔSASA_{ij} is solvent-accessible surface area of protein i screened by protein j. We measured the severity of clashes in docking models by calculating the *maximal penetration*, defined as follows. For every point $\vec{x}$ on $\Delta\text{SASA}_{\text{AB}}$, the closest point $\vec{y}$ on $\Delta\text{SASA}_{\text{BA}}$ is determined, and the maximum of these distances represents the maximal penetration of the two proteins

$$d_{\max} = \max_{\vec{x} \in \text{SASA}_{AB}} \left(\min_{\vec{y} \in \text{SASA}_{BA}} \left\| \vec{x} - \vec{y} \right\| \right) - 2.8\,\text{Å} \qquad (6\text{-}6)$$

Since the solvent-accessible surface can be considered as the molecular surface "inflated" by the radius of a water molecule (1.4 Å), a correction of 2×1.4 Å is introduced in Eq. 6-6 to eliminate the effect of mutual penetrations of water shells. Rapid calculation of solvent-accessible surfaces was achieved by Le Grand and Merz algorithm (139) and the use of *k-d* trees (140) for quick retrieval of spatially adjacent atom pairs.

The unbound structures of 102 protein-protein complexes from the DOCKGROUND benchmark set (49) were docked by the template-free GRAMM (12, 35) and template-based FSA (53, 101) protocols (see 6.2 Methods). Models of acceptable and higher quality (according to CAPRI criteria (118)) were retained in both protocols, resulting in the pools of 2,513 and 134 models for the GRAMM and FSA predictions, correspondingly. Despite different paradigms of the two methodologies (shape complementarity in GRAMM and structural similarity in FSA), most models had clashes with comparable average and maximal penetrations, with only a minor increase of clashes in the FSA predictions (**Figure 6-1**). The composition of clashes (side-chains vs. backbones) has a similar trend (**Figure 6-2**).

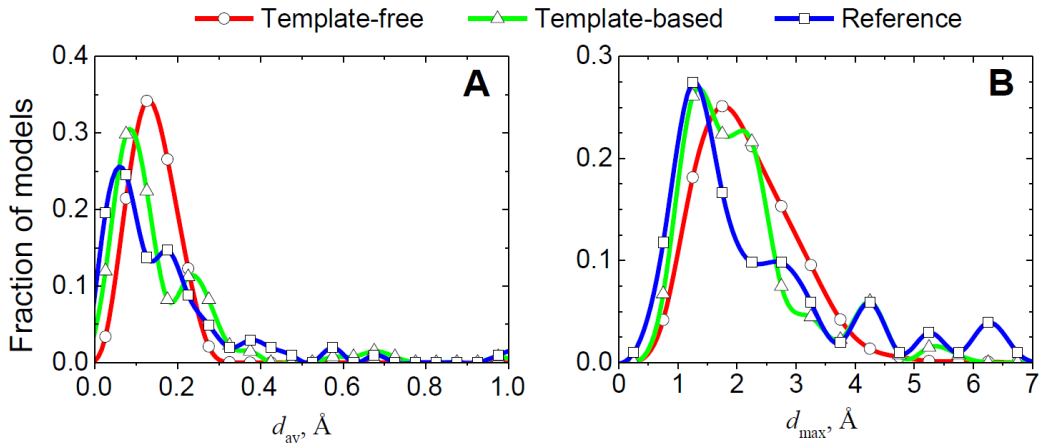

Figure 6-1: *Clashes in docking of unbound proteins*. For 102 complexes in DOCKGROUND Benchmark 3, 2513 template-free by GRAMM, and 134 template-based by FSA docking models of acceptable and higher quality were assessed by average (A) and maximum (B) penetrations, calculated from Eqs. 6-5 and 6-6, respectively. Reference is the distribution of clashes in the 102 reference complexes obtained by superimposition of the two unbound protein structures onto corresponding proteins in the co-crystallized complex.

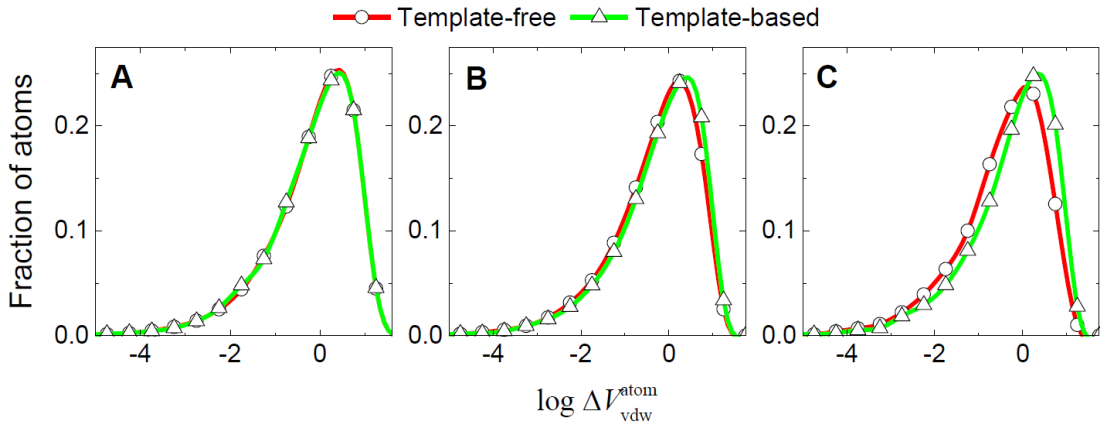

Figure 6-2: *Side chain and backbone clashes in docking of unbound proteins*. Volumes of intersections ΔV_{vdw}^{atom} (Å^3) were calculated for each pair of overlapping atoms, based on their radii and the interatomic distance. The shown distributions are obtained for 201,422 and 12,827 pairs of

side-chain atoms (A), 191,700 and 14,937 pairs of backbone and side-chain atoms (B), and for 28,982 and 4,466 pairs of backbone atoms (C) in docking models generated by free and template-based protocols, respectively.

Most GRAMM models are of acceptable quality (1,967 out of 2,513), whereas most FSA predictions are of high and medium quality (56 and 49 out of 134, correspondingly). For the GRAMM models the amount of clashes is almost independent of the docking quality while, less accurate docking predictions by FSA have more clashes than the more accurate ones (**Figure 6-3**). The largest discrepancies in the amounts of clashes between GRAMM and FSA predictions are observed for the models of acceptable quality (**Figure 6-3**C; an example in **Figure 6-4**).

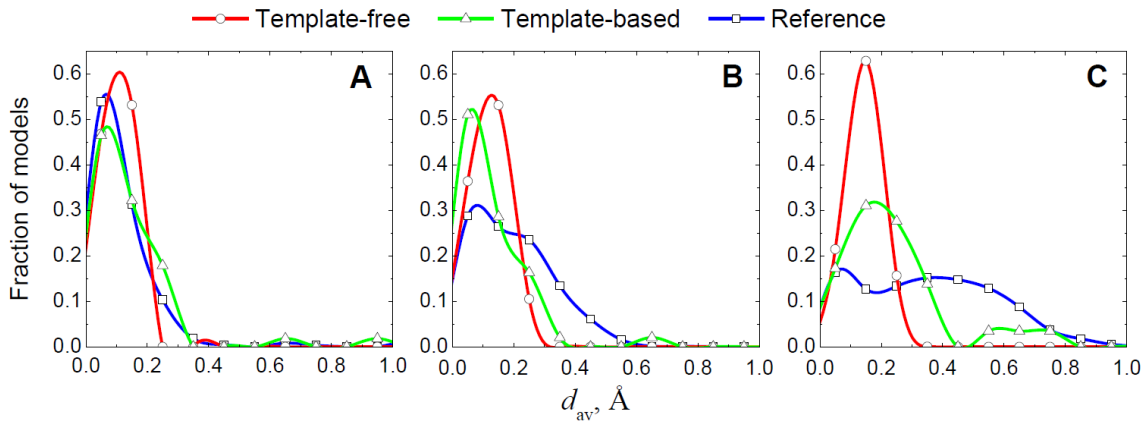

Figure 6-3: *Clashes in docking of different quality*. Distributions of average penetrations (Eq. 6-5) are shown separately for high (A), medium (B) and acceptable (C) quality models (according to CAPRI criteria). Plots are obtained for 32, 514, 1967 free and 56, 48, 29 template-based high, medium, and acceptable quality models, respectively. The reference distributions were obtained from the analysis of clashes in random models. For each target with at least one free or template-based prediction within a certain quality category, ten random models (one for targets with acceptable free models) of the same quality were generated (see text and caption to **Figure 6-5**).

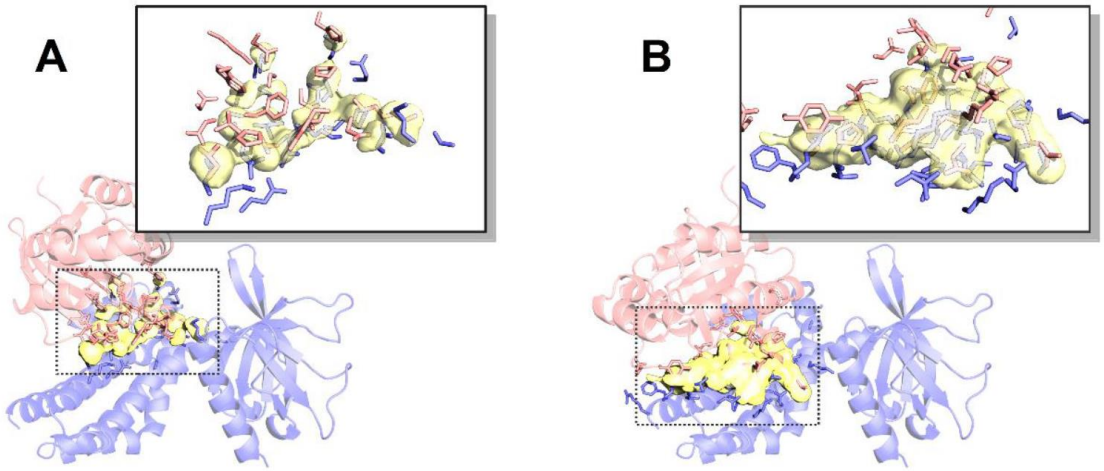

Figure 6-4: *Example of clashes in acceptable quality docking by free (A) and template-based (B) protocols*. Unbound structures corresponding to 2nz8, chains A and B, from DOCKGROUND Benchmark 3 were used. The unbound structure 1mh1, chain A, is in blue, and the unbound structure 1nty, chain A is in red. Overlapping van der Waals volumes are in yellow. The interface side-chains selected at 3 Å cut-off are in sticks. Average, d_{av}, and maximum, d_{max}, penetrations are 0.15 Å and 1.62 Å for the free and 0.58 Å and 3.80 Å for the template-based predictions, respectively.

To estimate the amount of clashes in models of a given accuracy, we generated ten random models of a protein-protein complex for each complex yielding acceptable or higher quality docking predictions (**Figure 6-5**). The amount of clashes in the random models decreases with the increase of the docking quality (thin lines in **Figure 6-3**). The random docking models have a larger amount of clashes than both GRAMM and FSA models. In all quality categories, the clashes in FSA models are closer to the clashes in the random docking models than to the clashes in the GRAMM docking, which inherently include the penalty for the clashes (**Figure 6-3**).

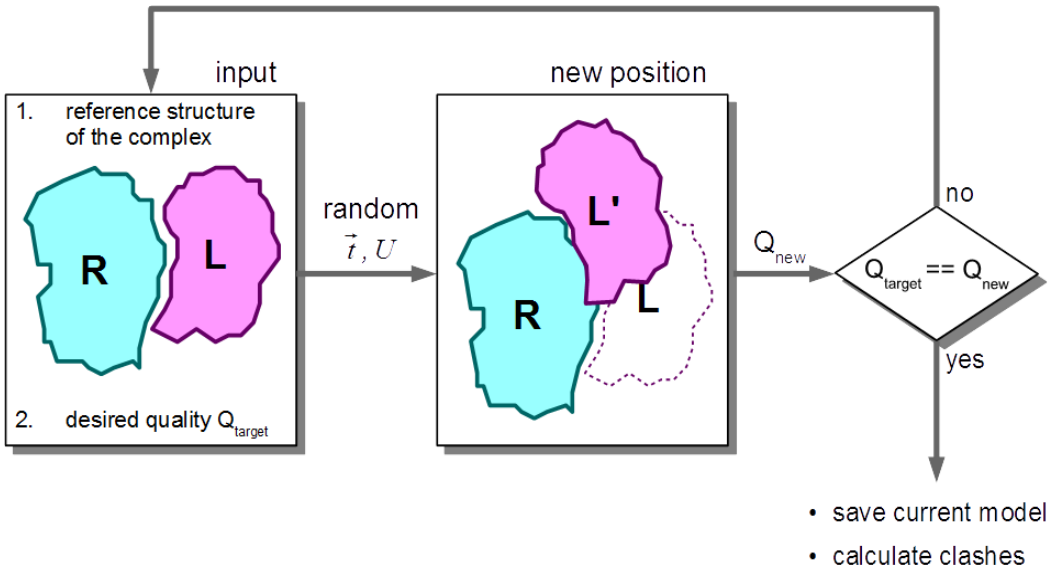

Figure 6-5: *Flowchart of random model generation.* Given two proteins in their reference positions (overlapped with the co-crystallized monomers), and the intended quality Q_{target} (high, medium, or acceptable), the procedure repeatedly generates a model by randomly translating (translation vector $\vec{t}$) and rotating (rotation matrix *U*) the ligand L with respect to the receptor R. At each trial, the quality Q_{new} of the complex RL' is calculated. The procedure is repeated until the model with the intended quality is obtained.

Conditions that define each CAPRI quality category restrict the receptor-ligand configuration space to a small area near the native state of the complex, resulting in the upper limit for the clashes (d_{av} and $d_{max} \leq$ 1.2 Å and 7.0 Å, respectively, in any of the docking models analyzed). Thus, clashes in a near-native prediction produced by *any rigid-body* docking method are inherently restricted to this limit. Therefore, a minimization procedure capable of removing clashes from the random models should be sufficient for the most docking predictions as well.

In the structural reconstruction of protein-protein interaction networks, most docked complexes would consist of individual protein models (5). Deviations of such

models from the native structures could significantly exceed the structural variations observed in the proteins upon binding (e.g., the average interface C^{α} RMSD between bound and unbound conformations in the Benchmark 5 (51) is ~1.4 Å). Thus, we also analyzed clashes in the docking predictions generated from our benchmark set of protein models (114) (**Figure 6-6**). The decrease in protein structural accuracy yields increasing amounts of clashes in FSA docking. Although the FSA docking success rates are weakly dependent on the proteins accuracy, the docking models of highly distorted protein models are mainly

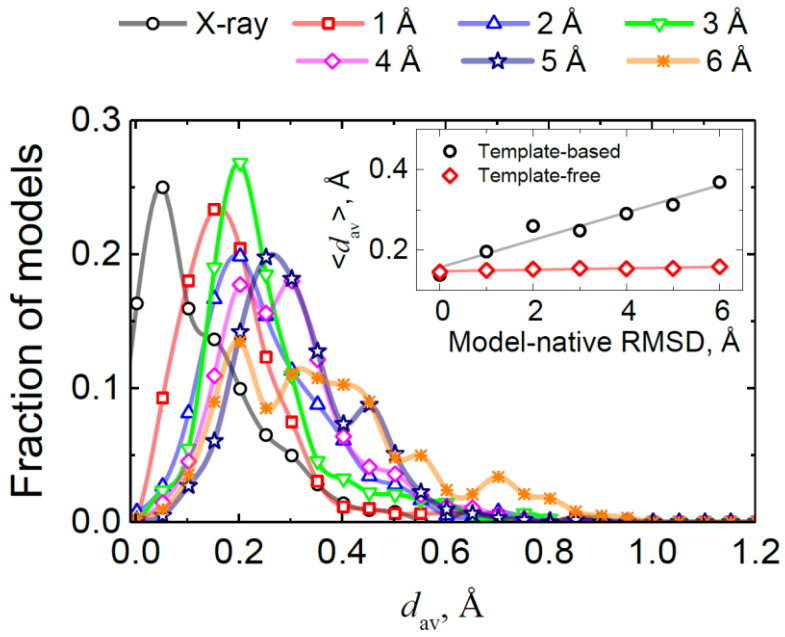

Figure 6-6: *Clashes in docking of modeled proteins*. Protein models are from 165 complexes in the DOCKGROUND model set 2. Distributions of average penetrations, d_{av} (Eq. 6-5), in the template-based docking predictions of acceptable and higher quality are shown separately for each accuracy level of protein models (1 to 6 Å RMSD from the corresponding native structures). For the reference, the plot shows d_{av} distribution of docking predictions from the co-crystallized bound proteins. The inset shows the mean values of the main panel distributions along with corresponding mean values of d_{av} distributions in free docking of the same set of modeled proteins.

in the acceptable quality category (130) and the FSA docking models of acceptable quality are, in general, characterized by the larger amounts of clashes (**Figure 6-3**). Due to the free docking paradigm that penalizes clashes, GRAMM yields docking predictions, on average, with a constant amount of clashes, regardless of the monomer's accuracy (inset in **Figure 6-6**).

6.4 Conclusions and future directions

Without the explicit constraints on the structural penetration, the template-based docking models resemble the random models and are more likely to have clashes than the free docking. However, because of the generally higher quality of the template-based predictions, the clashes in the free and template-based docking are overall similar. Thus approaches to structural refinement of the docking predictions developed for the free docking, should in principle be applicable to the template-based docking. In our future studies we plan a comparative evaluation of the refinement protocols on the free and template-based docking output.

Conclusions

Protein models have been considered a significant obstacle to docking because of their inherent inaccuracy, which may vary in a wide range, depending on the availability of modeling templates. Specialized benchmarking framework developed in this work allowed us to show that even highly inaccurate protein models can result in meaningful docking predictions. The template-based docking methodology was found to be much more tolerant to the structural inaccuracies in individual proteins, with only a moderate decrease in success rates for highly distorted protein models (5–6 Å RMSD), compared to the "bound" docking of their co-crystallized conformers. These results, along with the previous efforts of several groups (4, 33, 90), justify the use of template-based docking methodology as a primary tool for structural reconstruction of PPI networks.

Besides the benchmarking results, this work further advances the template-based techniques by introducing sophisticated template sets and novel scoring approaches. It investigates the quality of predictions in the template-based and free docking, showing unexpected similarity in the template-based and free docking output, with important implications for the refinement of template-based docking predictions. The reported development of the docking methodologies will be utilized in the genome-wide modeling of protein interactions, within the GWIDD project (22), and in similar efforts to structurally characterize molecular processes in living systems.

Bibliography

1. Levitt M (2009) Nature of the protein universe. *Proc. Natl. Acad. Sci. USA* 106:11079–11084.
2. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, & Skolnick J (2006) On the origin and highly likely completeness of single-domain protein structures. *Proc. Natl. Acad. Sci. USA* 103:2605–2610.
3. Skolnick J, Arakaki AK, Lee SY, & Brylinski M (2009) The continuity of protein structure space is an intrinsic property of proteins. *Proc. Natl. Acad. Sci. USA* 106:15690–15695.
4. Kundrotas PJ, Zhu Z, Janin J, & Vakser IA (2012) Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl. Acad. Sci. USA* 109:9438–9441.
5. Vakser IA (2013) Low-resolution structural modeling of protein interactome. *Curr. Opin. Struct. Biol.* 23:198-205.
6. Vakser IA (2014) Protein-protein docking: From interaction to interactome. *Biophys. J.* 107:1785-1793.
7. Mosca R, Pons T, Ceol A, Valencia A, & Aloy P (2013) Towards a detailed atlas of protein–protein interactions. *Curr. Opin. Struct. Biol.* 23:929–940.
8. Petrey D & Honig B (2014) Structural bioinformatics of the interactome. *Ann. Rev. Bioph.* 43:193-210.
9. Petrey D, *et al.* (2015) Template-based prediction of protein function. *Curr. Opin. Struct. Biol.* 32:33-38.
10. Vakser IA & Kundrotas P (2008) Predicting 3D structures of protein-protein complexes. *Curr. Pharm. Biotech.* 9:57-66.
11. Huang SY (2015) Exploring the potential of global protein-protein docking: an overview and critical assessment of current programs for automatic ab initio docking. *Drug Discov. Today* 20:969-977.

12. Katchalski-Katzir E, *et al.* (1992) Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA* 89:2195-2199.
13. Kozakov D, Brenke R, Comeau SR, & Vajda S (2006) PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins* 65:392-406.
14. Mintseris J, *et al.* (2007) Integrating statistical pair potentials into protein complex prediction. *Proteins* 69:511–520.
15. de Vries SJ, van Dijk M, & Bonvin AMJJ (2010) The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.* 5:883-897.
16. Tovchigrechko A & Vakser IA (2006) GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res.* 34:W310-W314.
17. Hwang H, Vreven T, Pierce BG, Hung JH, & Weng ZP (2010) Performance of ZDOCK and ZRANK in CAPRI rounds 13-19. *Proteins* 78:3104-3110.
18. Pieper U, *et al.* (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* 32:D217-222.
19. Aloy P & Russell RB (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. USA* 99:5896–5901.
20. Grimm V, Zhang Y, & Skolnick J (2006) Benchmarking of dimeric threading and structure refinement. *Proteins* 63:457-465.
21. Kundrotas PJ, Lensink MF, & Alexov E (2008) Homology-based modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles. *Int. J. Biol. Macromol.* 43:198-208.
22. Kundrotas PJ, Zhu ZW, & Vakser IA (2010) GWIDD: Genome-wide protein docking database. *Nucleic Acids Res.* 38:D513-D517.
23. Keskin O & Nussinov R (2005) Favorable scaffolds: Proteins with different sequence, structure and function may associate in similar ways. *Protein Eng.* 18:11-24.
24. Keskin O & Nussinov R (2007) Similar binding sites and different partners: Implications to shared proteins in cellular pathways. *Structure* 15:341-354.

25. Petrey D, Fischer M, & Honig B (2009) Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proc. Natl. Acad. Sci. USA* 106:17377-17382.
26. Zhang QC, Petrey D, Norel R, & Honig BH (2010) Protein interface conservation across structure space. *Proc. Natl. Acad. Sci. USA* 107:10896–10901.
27. Cavasotto CN & Phatak SS (2009) Homology modeling in drug discovery: current trends and applications. *Drug Discov. Today* 14:676-683.
28. Hasegawa H & Holm L (2009) Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.* 19:341-348.
29. Gunther S, May P, Hoppe A, Frommel C, & Preissner R (2007) Docking without docking: ISEARCH-Prediction of interactions using known interfaces. *Proteins* 69:839-844.
30. Mitchell EM, Artymiuk PJ, Rice DW, & Willett P (1990) Use of techniques derived from graph-theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* 212:151-166.
31. Kozakov D*, et al.* (2011) Structural conservation of druggable hot spots in protein-protein interfaces. *Proc. Natl. Acad. Sci. USA* 108:13528-13533.
32. Aytuna AS, Gursoy A, & Keskin O (2005) Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics* 21:2850-2855.
33. Tuncbag N, Gursoy A, Nussinov R, & Keskin O (2011) Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat. Protoc.* 6:1341-1354.
34. Andrusier N, Mashiach E, Nussinov R, & Wolfson HJ (2008) Principles of flexible protein–protein docking. *Proteins* 73:271–289.
35. Vakser IA, Matar OG, & Lam CF (1999) A systematic study of low-resolution recognition in protein-protein complexes. *Proc. Natl. Acad. Sci. USA* 96:8477-8482.
36. Lensink MF & Wodak SJ (2013) Docking, scoring, and affinity prediction in CAPRI. *Proteins* 81:2082-2095.

37. Andrusier N, Nussinov R, & Wolfson HJ (2007) FireDock: fast interaction refinement in molecular docking. *Proteins* 69:139-159.
38. Gray JJ, *et al.* (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* 331:281-299.
39. Moghadasi M, *et al.* (2015) The impact of side-chain packing on protein docking refinement. *J. Chem. Inf. Model.* 55:872-881.
40. May A & Zacharias M (2008) Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. *Proteins* 70:794-809.
41. Venkatraman V & Ritchie DW (2012) Flexible protein docking refinement using pose-dependent normal mode analysis. *Proteins* 80:2262-2274.
42. Mashiach E, Nussinov R, & Wolfson HJ (2010) FiberDock: Flexible induced-fit backbone refinement in molecular docking. *Proteins* 78:1503-1519.
43. Moal IH & Bates PA (2010) SwarmDock and the use of normal modes in protein-protein docking. *Int. J. Mol. Sci.* 11:3623–3648.
44. Wang C, Bradley P, & Baker D (2007) Protein-protein docking with backbone flexibility. *J. Mol. Biol.* 373:503-519.
45. de Vries SJ, *et al.* (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins* 69:726-733.
46. Janin J, *et al.* (2003) CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins* 52:2-9.
47. Rodrigues, J.P.G.L.M., *et al.* (2013) Defining the limits of homology modeling in information-driven protein docking. *Proteins* 81:2119–2128.
48. Rost B (1999) Twilight zone of protein sequence alignments. *Protein Eng.* 12:85-94.
49. Gao Y, Douguet D, Tovchigrechko A, & Vakser IA (2007) DOCKGROUND system of databases for protein recognition studies: Unbound structures for docking. *Proteins* 69:845-851.
50. Chen R, Mintseris J, Janin J, & Weng ZP (2003) A protein-protein docking benchmark. *Protein. Struct. Funct. Genet.* 52:88-91.

51. Vreven T, *et al.* (2015) Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. *J. Mol. Biol.* 427:3031-3041.
52. Kru F, Korff G, Elghobashi-Meinhardt N, & Knapp EW (2015) ProPairs: a data set for protein-protein docking. *J. Chem. Inf. Model.* 55:1495-1507.
53. Sinha R, Kundrotas PJ, & Vakser IA (2012) Protein docking by the interface structure similarity: How much structure is needed? *PloS One* 7:e31349.
54. Samudrala R & Levitt M (2000) Decoys 'R' Us: A database of incorrect conformations to improve protein structure prediction. *Protein Sci.* 9:1399-1401.
55. Carbajo D & Tramontano A (2012) A resource for benchmarking the usefulness of protein structure models. *BMC Bioinformatics* 13:188.
56. Brylinski M & Skolnick J (2010) Q-DockLHM: Low-resolution refinement for ligand comparative modeling. *J. Comput. Chem.* 31:1093–1105.
57. Douguet D, Chen HC, Tovchigrechko A, & Vakser IA (2006) DOCKGROUND resource for studying protein-protein interfaces. *Bioinformatics* 22:2612–2618.
58. Hwang H, Vreven T, Janin J, & Weng Z (2010) Protein–protein docking benchmark version 4.0. *Proteins* 78:3111–3114.
59. Tovchigrechko A, Wells CA, & Vakser IA (2002) Docking of protein models. *Protein Sci.* 11:1888-1896.
60. Elber R & Karplus M (1987) A method for determining reaction paths in large molecules - application to myoglobin. *Chem. Phys. Lett.* 139:375-380.
61. Chu JW, Trout BL, & Brooks BR (2003) A super-linear minimization scheme for the nudged elastic band method. *J. Chem. Phys.* 119:12708-12717.
62. Needleman S & Wunsch CD (1970) A general method applicable to search for similarities in amino acid sequence of two proteins. *J. Mol. Biol.* 48:443-453.
63. Gotoh O (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.* 162:705-708.
64. Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of database programs. *Nucleic Acids Res.* 25:3389–3402.
65. Henikoff S & Henikoff JG (1993) Performance evaluation of amino acid substitution matrices. *Proteins* 17:49-61.

66. Petrey D, *et al.* (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins* 53:430-435.
67. Kabsch W & Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577-2637.
68. Onufriev A, Bashford D, & Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins* 55:383-394.
69. Case DA, *et al.* (2008) AMBER 10 (University of California, San Francisco).
70. Duan Y, *et al.* (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* 24:1999-2012.
71. Kundrotas PJ & Vakser IA (2010) Accuracy of protein-protein binding sites in high-throughput template-based modeling. *PLoS Comp. Biol.* 6:e1000727.
72. Moult J, Fidelis K, Kryshtafovych A, & Tramontano A (2011) Critical assessment of methods of protein structure prediction (CASP)–Round IX. *Proteins* 79 (Suppl 10):1-5.
73. Zemla A (2003) LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* 31:3370-3374.
74. Schwede T (2013) Protein modeling: What happened to the ‘‘protein structure gap’’? *Structure* 21:1531-1540.
75. Aloy P, Pichaud M, & Russell RB (2005) Protein complexes: Structure prediction challenges for the 21st century *Curr. Opin. Struct. Biol.* 15:15-22.
76. Szilagyi A & Zhang Y (2014) Template-based structure modeling of protein–protein interactions. *Curr. Opin. Struct. Biol.* 24:10–23.
77. Kuzu G, Keskin O, Gursoy A, & Nussinov R (2012) Constructing structural networks of signaling pathways on the proteome scale. *Curr. Opin. Struct. Biol.* 22:367-377.
78. Dey F, Zhang QC, Petrey D, & Honig B (2013) Toward a ‘‘structural BLAST’’: Using structural relationships to infer function. *Protein Sci.* 22:359-366.

79. Anishchenko I, Kundrotas PJ, Tuzikov AV, & Vakser IA (2014) Protein models: The Grand Challenge of protein docking. *Proteins* 82:278–287.
80. Roy A, Kucukural A, & Zhang Y (2010) I-TASSER: A unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5:725-738.
81. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9:40.
82. Chung SY & Subbiah S (1996) A structural explanation for the twilight zone of protein sequence homology. *Structure* 4:1123-1127.
83. Xu D & Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys. J.* 101:2525-2534.
84. Joosten RP, *et al.* (2011) A series of PDB related databases for everyday needs. *Nucleic Acids Res.* 39:D411-D419.
85. Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. *Acta Cryst. A* 32:922-923.
86. Kabsch W (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst. A* 34:827-828.
87. Zhang Y & Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. *Proteins* 57:702-710.
88. Szilagyi A, Grimm V, Arakaki AK, & Skolnick J (2005) Prediction of physical protein-protein interactions. *Phys. Biol.* 2:S1-S16.
89. Kundrotas PJ, Vakser IA, & Janin J (2013) Structural templates for modeling homodimers. *Protein Sci.* 22:1655-1663.
90. Zhang QC, *et al.* (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* 490:556-560.
91. Henrick K & Thornton JM (1998) PQS: A protein quaternary structure file server. *Trends Biochem. Sci.* 23:358-361.
92. Krissinel E & Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* 372:774–797.
93. Tuncbag N, Gursoy A, Guney E, Nussinov R, & Keskin O (2008) Architectures and functional coverage of protein–protein interfaces. *J. Mol. Biol.* 381:785-802.

94. Zhu H, Domingues FS, Sommer I, & Lengauer T (2006) NOXclass: Prediction of protein-protein interaction types. *BMC Bioinformatics* 7:27.
95. Cukuroglu E, Gursoy A, Nussinov R, & Keskin O (2014) Non-redundant unique interface structures as templates for modeling protein interactions. *PloS One* 9:e86738.
96. Mukherjee S & Zhang Y (2009) MM-align: A quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. *Nucleic Acids Res.* 37:e83.
97. Zhang Y & Skolnick J (2005) TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 33:2302-2309.
98. Cormen TH, Leiserson CE, Rivest RL, & Stein C (2009) *Introduction to Algorithms* (The MIT Press) Third Edition Ed p 1312.
99. Hartuv E & Shamir R (2000) A clustering algorithm based on graph connectivity. *Inform. Process. Lett.* 76:175-181.
100. Stoer M & Wagner F (1997) A simple min-cut algorithm. *J. ACM* 44:585-591.
101. Sinha R, Kundrotas PJ, & Vakser IA (2010) Docking by structural similarity at protein-protein interfaces. *Proteins* 78:3235-3241.
102. Kundrotas PJ & Vakser IA (2013) Global and local structural similarity in protein-protein complexes: Implications for template-based docking. *Proteins* 81:2137–2142.
103. Ogmen U, Keskin O, Aytuna AS, Nussinov R, & Gursoy A (2005) PRISM: Protein interactions by structural matching. *Nucleic Acids Res.* 33:W331-W336.
104. Petrey D & Honig B (2003) GRASP2: Visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol.* 374:492-509.
105. Watts DJ & Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393:440-442.
106. Xu J & Zhang Y (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* 26:889-895.
107. Negroni J, Mosca R, & Aloy P (2014) Assessing the applicability of template-based protein docking in the twilight zone. *Structure* 22:1356–1362.

108. Murzin AG, Brenner SE, Hubbard T, & Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536-540.

109. Vajda S, Vakser IA, Sternberg MJE, & Janin J (2002) Meeting report: Modeling of protein interactions in genomes. *Proteins* 47:444-446.

110. Moult J, Fidelis K, Kryshtafovych A, Schwede T, & Tramontano A (2014) Critical assessment of methods of protein structure prediction (CASP) — round X. *Proteins* 82 (Suppl 2):1-6.

111. Skolnick J, Zhou H, & Gao M (2013) Are predicted protein structures of any value for binding site prediction and virtual ligand screening? *Curr. Opin. Struct. Biol.* 23:191–197.

112. Zhao J, Dundas J, Kachalo S, Ouyang Z, & Liang J (2011) Accuracy of functional surfaces on comparatively modeled protein structures. *J. Struct. Funct. Genomics* 12:97-107.

113. Maheshwari S & Brylinski M (2015) Predicted binding site information improves model ranking in protein docking using experimental and computer-generated target structures. *BMC Struct. Biol.* 15:23.

114. Anishchenko I, Kundrotas PJ, Tuzikov AV, & Vakser IA (2015) Protein models docking benchmark 2. *Proteins* 83:891-897.

115. Vakser IA (1995) Protein docking for low-resolution structures. *Protein Eng.* 8:371-377.

116. Miyazawa S & Jernigan RL (1999) Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins* 34:49-68.

117. Anishchenko I, Kundrotas PJ, Tuzikov AV, & Vakser IA (2015) Structural templates for comparative protein docking. *Proteins* 83:1563–1570.

118. Mendez R, Leplae R, De Maria L, & Wodak SJ (2003) Assessment of blind predictions of protein–protein interactions: Current status of docking methods. *Proteins* 52:51-67.

119. Tovchigrechko A & Vakser IA (2001) How common is the funnel-like energy landscape in protein-protein interactions? *Protein Sci.* 10:1572-1583.

120. Kundrotas PJ & Vakser IA (2013) Protein-protein alternative binding modes do not overlap. *Protein Sci.* 22:1141-1145.
121. Mann HB & Whitney DR (1947) On a test of whether one of 2 random variables is stochastically larger than the other. *Ann. Math. Stat.* 18:50-60.
122. Pierce BG, Hourai Y, & Weng Z (2011) Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PloS One* 6:e24657.
123. Kozakov D, *et al.* (2014) Encounter complexes and dimensionality reduction in protein–protein association. *eLife* 3:e01370.
124. Zhou HX & Bates PA (2013) Modeling protein association mechanisms and kinetics. *Curr. Opin. Struct. Biol.* 23:887–893.
125. Badal VD, Kundrotas PJ, & Vakser IA (2015) Text mining for protein docking. *PLoS Comp. Biol.* 11:e1004630.
126. Hopf TA, *et al.* (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 3:e0343.
127. Ovchinnikov S, Kamisetty H, & Baker D (2014) Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* 3:e02030.
128. Stratmann D, Boelens R, & Bonvin AMJJ (2011) Quantitative use of chemical shifts for the modeling of protein complexes. *Proteins* 79:2662–2670.
129. Berman HM, *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28:235-242.
130. Anishchenko I, Kundrotas PJ, & Vakser IA (2016) Modeling complexes of modeled proteins. *To be published*.
131. Lasker K, Sali A, & Wolfson HJ (2010) Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. *Proteins* 78:3205-3211.
132. Nicola G & Vakser IA (2007) A simple shape characteristic of protein-protein recognition. *Bioinformatics* 23:789-792.
133. Vacha R & Frenkel D (2011) Relation between molecular shape and the morphology of self-assembling aggregates: a simulation study. *Biophys. J.* 101:1432-1439.

134. Vakser IA (1996) Main-chain complementarity in protein-protein recognition. *Protein Eng.* 9:741-744.

135. Zhang Q, Sanner M, & Olson AJ (2009) Shape complementarity of protein-protein complexes at multiple resolutions. *Proteins* 75:453-467.

136. Hunjan J, Tovchigrechko A, Gao Y, & Vakser IA (2008) The size of the intermolecular energy funnel in protein-protein interactions. *Proteins* 72:344-352.

137. Kozakov D, Clodfelter KH, Vajda S, & Camacho CJ (2005) Optimal clustering for detecting near-native conformations in protein docking. *Biophys. J.* 89:867-875.

138. Tsai J, Taylor R, Chothia C, & Gerstein M (1999) The packing density in proteins: Standard radii and volumes. *J. Mol. Biol.* 290:253-266.

139. Le Grand SM & Merz KM (1993) Rapid approximation to molecular surface area via the use of Boolean logic and look-up tables. *J. Comput. Chem.* 14:349-352.

140. Bentley JL (1975) Multidimensional binary search trees used for associative searching. *ACM Commun.* 18:509-517.

Appendix A

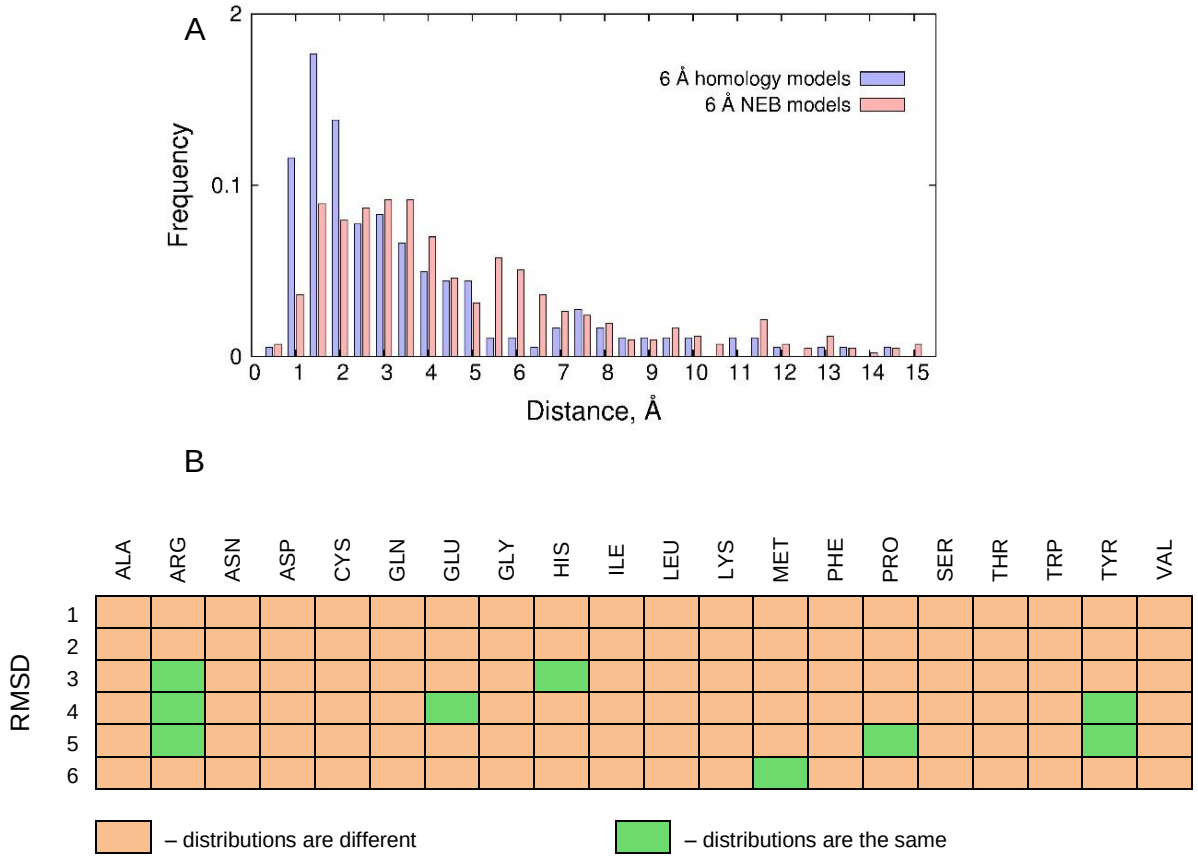

Figure A-1: *Deviation of C^{α} positions in protein models*. The deviation is calculated from the corresponding X-ray structure, with models and the X-ray structure superimposed by minimizing RMSD. (A) Histidines in 6 Å models. (B) Comparison of deviation distributions in homology and NEB models.

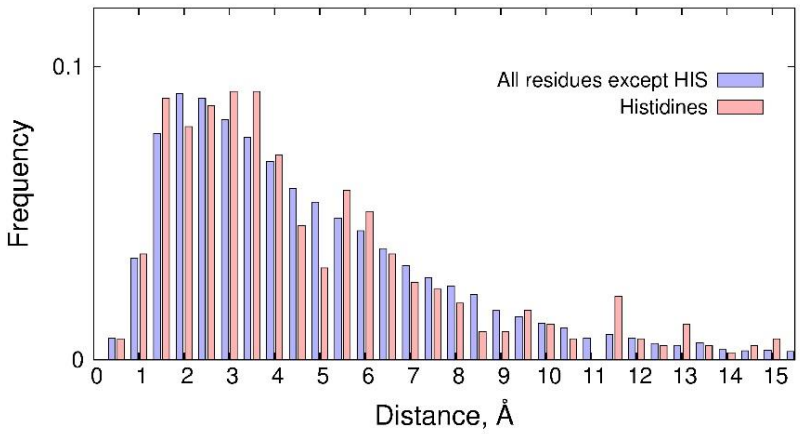

Figure A-2: *Deviation of C^{α} positions in histidines and all other residues in 6 Å NEB models.*

Appendix B

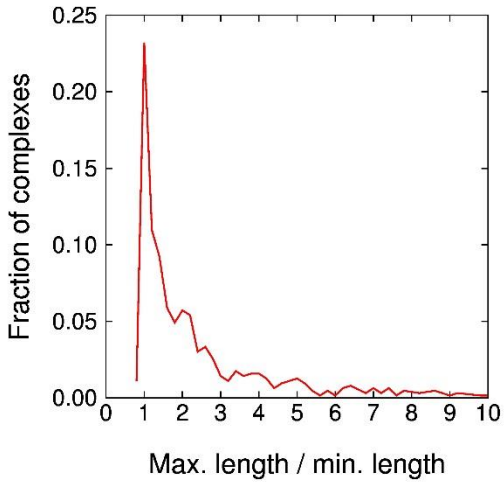

Figure B-1: Distribution of protein sizes in a set of 629 binary complexes initially selected from Dockground. The ratio for a complex is the number of residues in the longer protein divided by the number of residues in the shorter one.

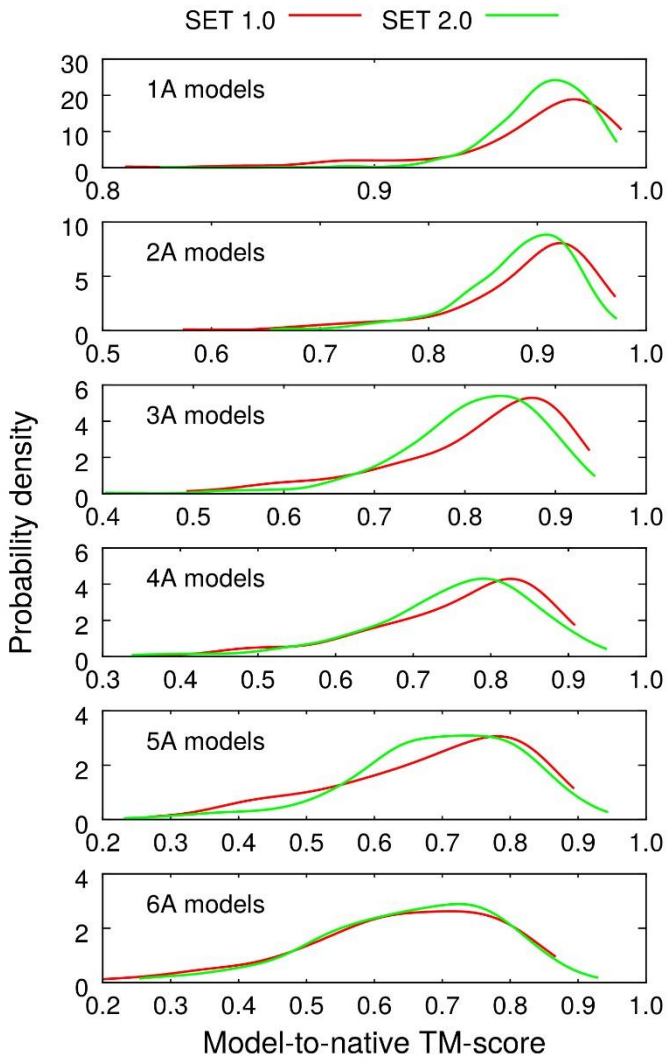

Figure B-2: *Distributions of TM-scores between protein models and the native structures*. In selection of the final models for Set 1.0, the preference was given to those with a more uniform distribution of distortions along the protein chain. Thus, more residues were involved in the alignment, resulting in higher TM-scores. No such filter was used for Set 2.0, more adequately reflecting the real case scenario in modeling/docking. Kernel density estimation technique (GNUPLOT program) was used to smooth the calculated frequencies. All plots are normalized so that the area under the curve is equal to 1.

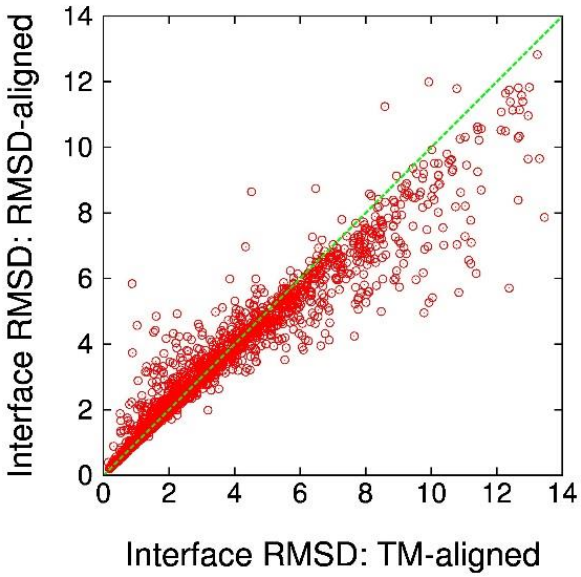

Figure B-3: *Interface C^{α} RMSD of model/native structure superposition by TM-score and RMSD minimization.*

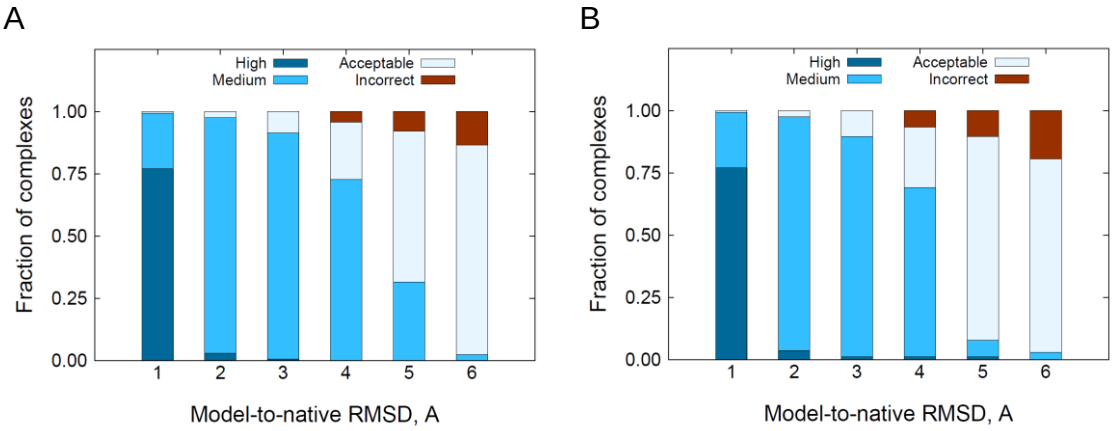

Figure B-4: *Quality of model-model complexes according to CAPRI criteria*. The superposition with the native structure performed by (A) RMSD minimization, and (B) TM-score.

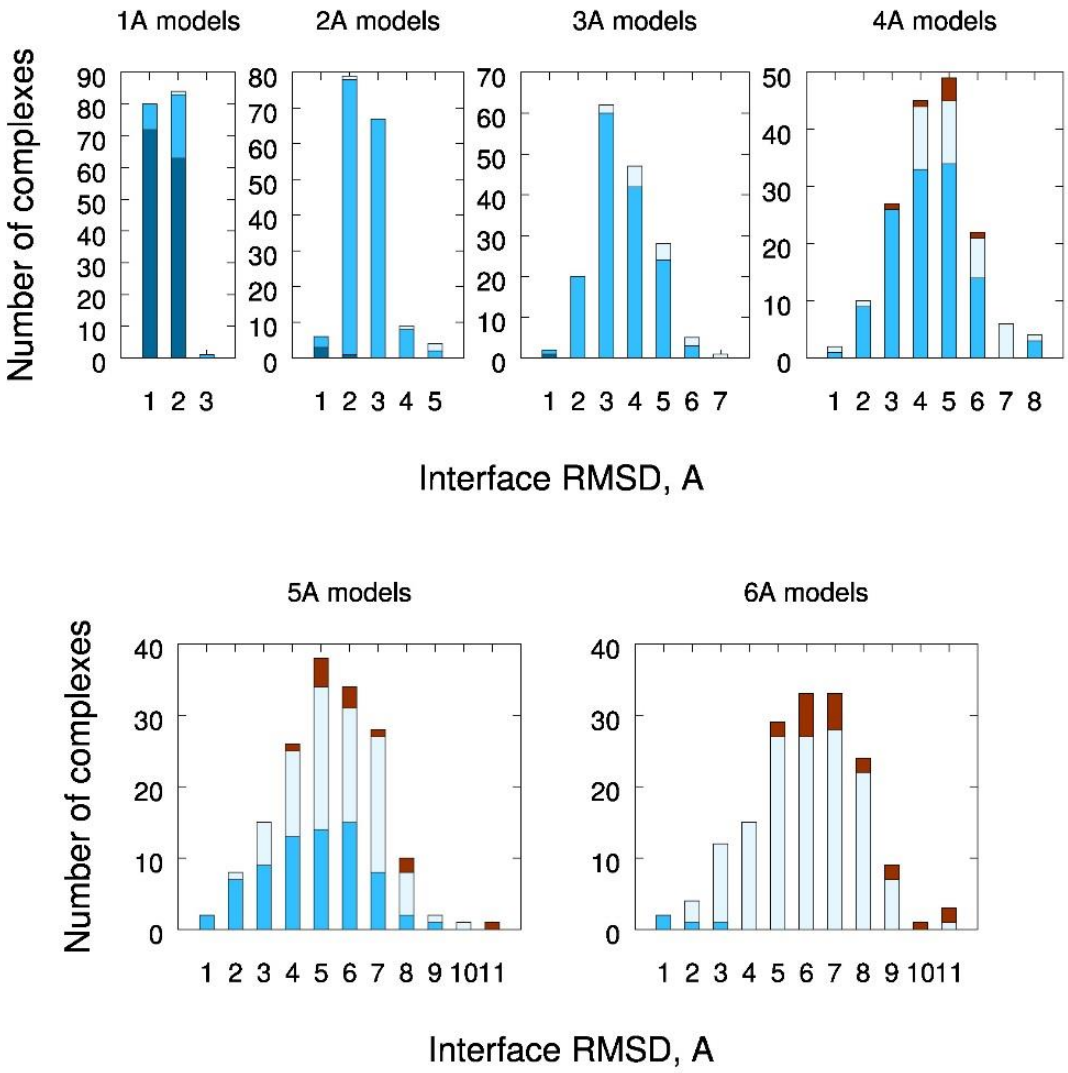

Figure B-5: *Quality of model-model complexes according to CAPRI criteria as a function of interface RMSD*. The sum of all bars in each panel is 165 – the total number of complexes.

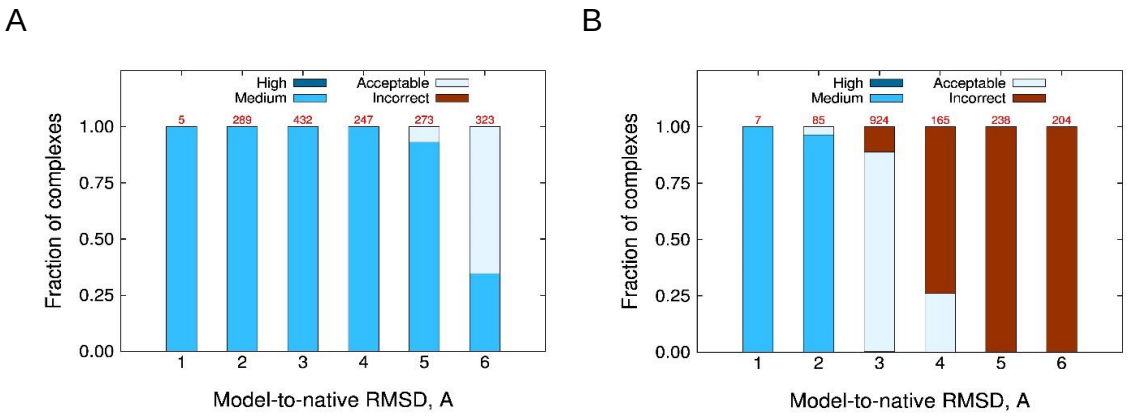

Figure B-6: *Quality of the "ideal" complexes built from all models at certain accuracy level*. The examples are (A) 1oph and (B) 2a5t. At the top of the bins are total numbers of monomer models at this accuracy level.

Appendix C

Table C-1: *Docking accuracy according to CAPRI criteria*

Quality category	Condition
High	f_{nat} [(1)] ≥ 0.5 *and* (L-RMSD[(2)] ≤ 1.0 Å *or* I-RMSD[(3)] ≤ 1.0 Å)
Medium	f_{nat} ≥ 0.3 *and* (1.0 < L-RMSD ≤ 5.0 Å *or* 1.0 < I-RMSD ≤ 2.0 Å)
Acceptable	f_{nat} ≥ 0.1 *and* (5.0 < L-RMSD ≤ 10.0 Å *or* 2.0 < I-RMSD ≤ 4.0 Å)
Incorrect	f_{nat} < 0.1 *and* (L-RMSD > 10.0 Å *and* I-RMSD > 4.0 Å)

[(1)] Fraction of predicted native residue–residue contacts

[(2)] C^{α} ligand RMSD when receptors are optimally aligned

[(3)] Interface C^{α} RMSD calculated over the set of native interface residues after a structural superposition of these residues

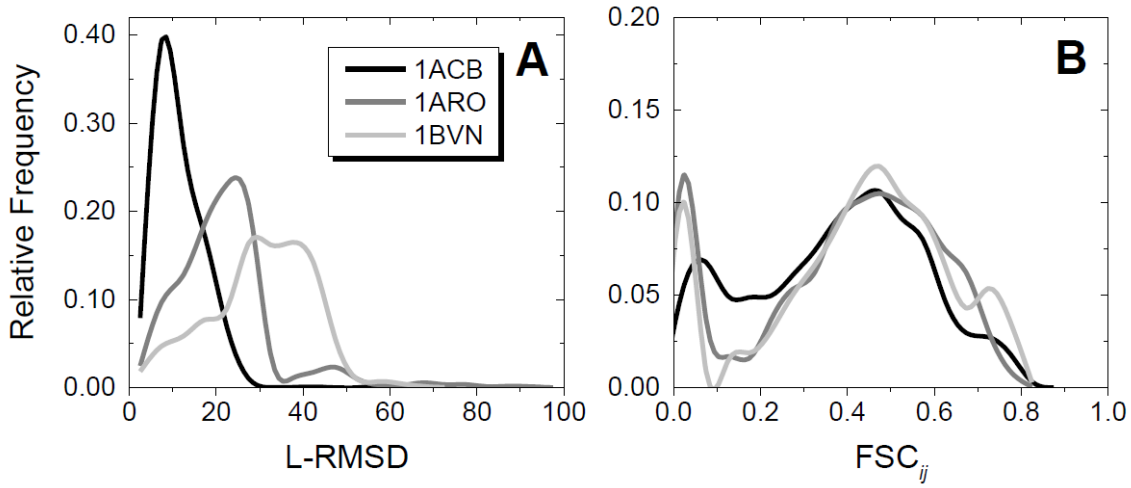

Figure C-1: *Similar docking modes represented by different metrics*. The data shows results of pairwise comparison of free docking top 1000 solutions for three protein-protein complexes from the Benchmark 1 in terms of (A) ligand RMSD and (B) fraction of shared contacts (Eq. 5-2, main text). The docked ligands in each pair of predicted configurations had to satisfy conditions $|\vec{t}| <$ 15 Å and $\cos^{-1}\frac{\mathrm{tr}(U)-1}{2} < \frac{\pi}{2}$, where $\vec{t}$ and U are translation vector and 3×3 rotation matrix, respectively, needed to obtain ligand position in one configuration from the ligand position in the other configuration. The unbound structures of proteins in the three complexes were used in docking. As opposed to ligand RMSD, the FSC_{ij} between similar docking modes does not have substantial variation from complex to complex.

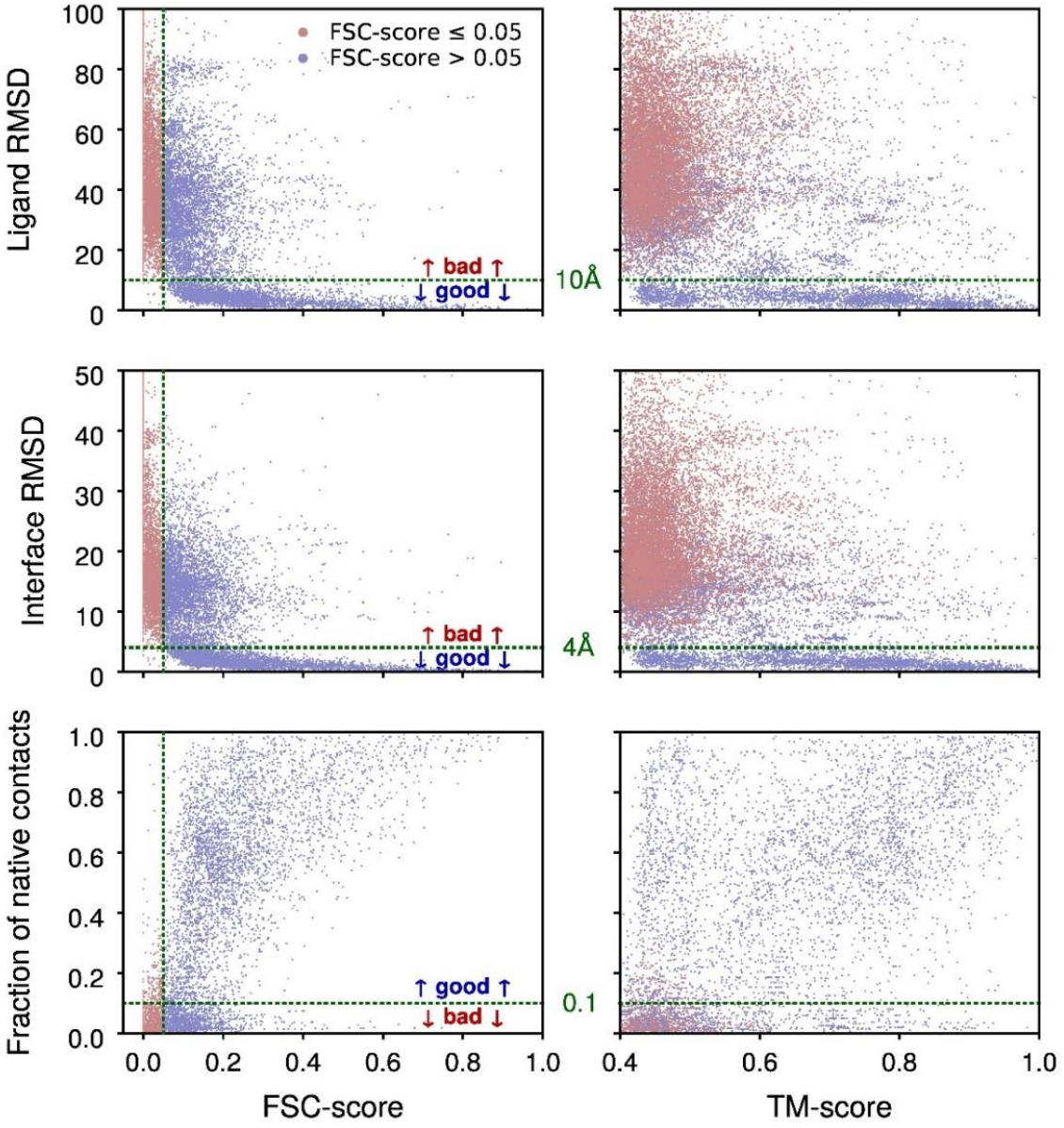

Figure C-2: *Filtering of the template-based docking solutions with the FSC-score*. The data was obtained on a subset of 807 hetero complexes from the full-structure template library of 4,950 co-crystallized binary complexes from DOCKGROUND. Each complex in the subset was re-docked by the template-based docking using all remaining 806 structures as templates. For all resulting models with TM-score > 0.4 (see 5.2 Methods), three components of the CAPRI criteria (ligand RMSD, interface RMSD and fraction of native contacts) were plotted against FSC-scores (left-hand panels) and TM-scores (right-hand panels). Green horizontal lines separate correct solutions (acceptable, medium or high quality predictions) from the incorrect ones. Almost all models (99.8% or 12,499

out of 12,524) with the FSC-score ≤ 0.05 (vertical green lines on the left-hand panels) fall into the incorrect CAPRI category. Those models constitute 61% of all bad solutions. Such differences between good and bad models are not captured by the original scoring with the TM-score (right-hand panels).

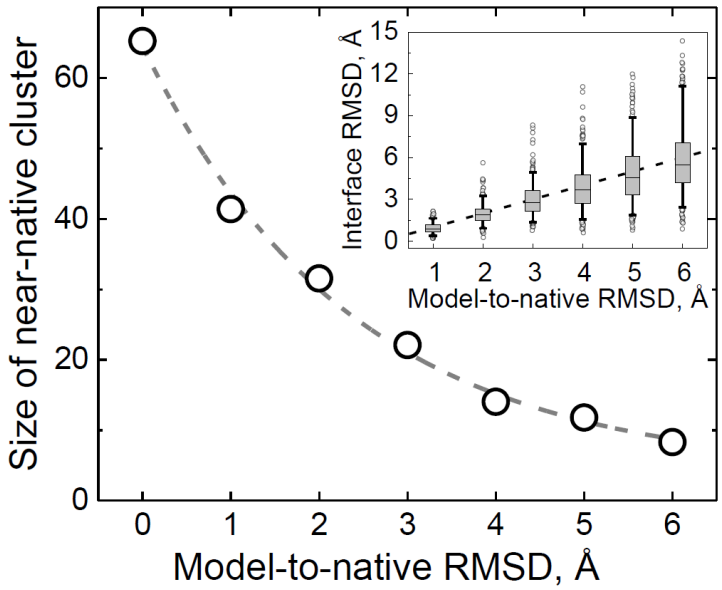

Figure C-3: *Blurring of near-native clusters produced by free docking at increasing levels of models' inaccuracy*. Near-native (acceptable, medium and high quality) docking solutions among top 1000 free docking predictions were counted for each of 165 complexes in the Benchmark 2 at each level of monomer distortion ("0" level means native X-ray structures). The counts were averaged over dataset complexes separately at each distortion level and the average numbers were plotted as function of monomer accuracy (x-axis). The dashed line is an exponential decay function fitted to the data points. Inset shows box-and-whiskers diagrams of distribution of RMSD values calculated between C^{α} atoms of interface residues in model and native monomers for 2×165 proteins from the Benchmark 2 at six levels of monomer accuracy. Box areas and whiskers contain 25 – 75 % and 5 – 95 % of data, respectively. Dashed line in the inset is $y = x$ function.

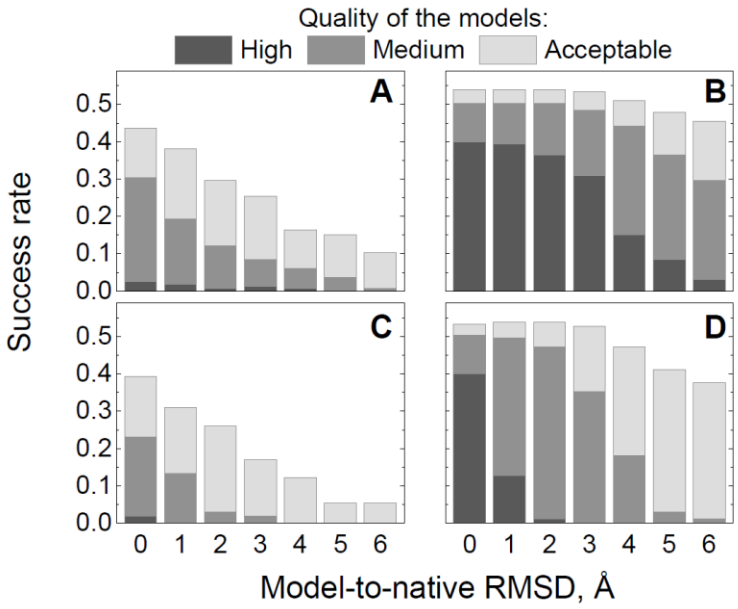

Figure C-4: *Docking success rates assessed by CAPRI criteria*. Left-hand panels show free docking and the right-hand ones show template-based docking. Two types of reference complexes were used (see 5.2 Methods): 'ideal model' (upper panels) and the native co-crystallized X-ray structure (lower panels).

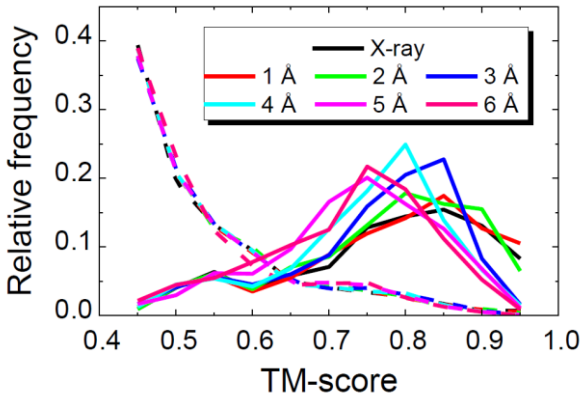

Figure C-5: *Target-template similarities in the template-based models built from monomers of different accuracy*. For each level of monomer accuracy, all template-based predictions resulting from docking of 165 complexes from the Benchmark 2 were combined and assessed by the CAPRI criteria. TM-scores between target and templates were then calculated separately for good (acceptable and better quality, solid lines) and incorrect (dashed lines) models and plotted as histograms (0.05 TM-score window) normalized by the total number of predictions in each category (~1,500 for good predictions and ~12,000 for incorrect ones).

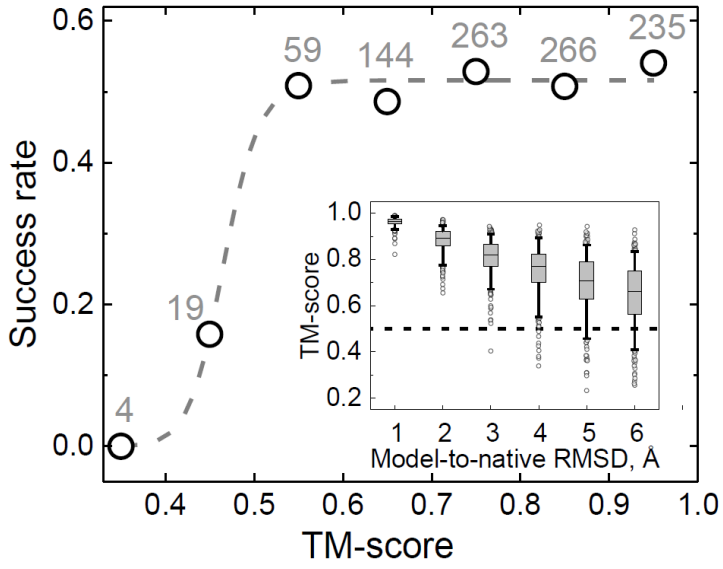

Figure C-6: *Correlation between success rates of the template-based docking and structural distortions of the protein models expressed in terms of TM-score between model and native X-ray structures*. The 165×6 model-model complexes (two models with the same accuracy level originating from the same native complex) from the Benchmark 2 were divided into seven groups with average TM-score (of two TM-scores for separate alignments of both monomers in the model-model complex to corresponding native structures) < 0.40, [0.4,0.5), [0.5,0.6), [0.6,0.7), [0.7,0.8), [0.8,0.9) and [0.9,1.0]. Within each group, success rate for the top 10 predictions was calculated and plotted as function of medium TM-score for each group. Total number of complexes in each TM-score range is shown above the data points. The dashed line is the Hill function fitted to the data points. Variations in TM-scores of the protein models from the Benchmark 2 at six levels of distortions are shown in the inset as box-and-whiskers diagrams. Box areas and whiskers contain 25 – 75 % and 5 – 95 % of data, respectively. Dashed line in the inset shows TM-score 0.5 threshold, above which a pair of proteins are likely to have the same fold (Xu & Zhang, Bioinformatics, 2010, 26:889-95).

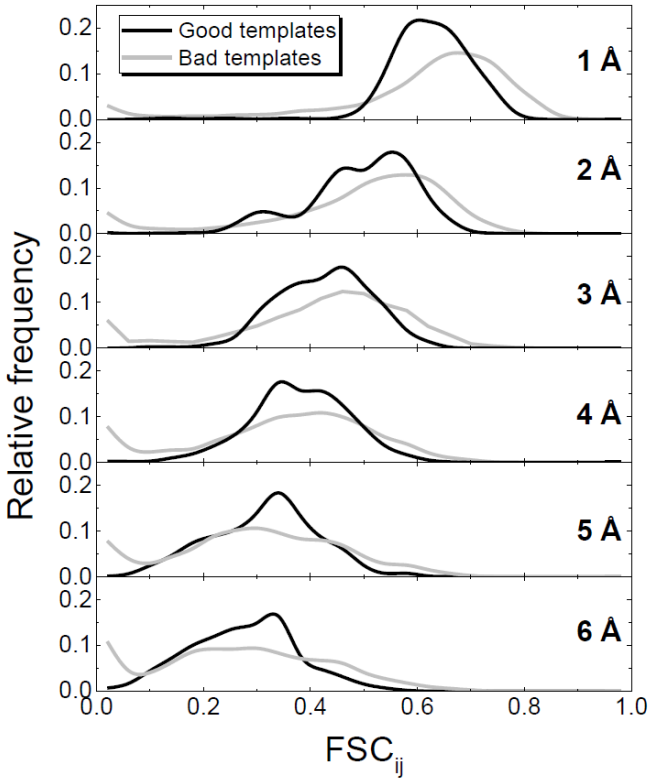

Figure C-7: *Similarities between model-model and X-ray-X-ray template-based predictions originating from the same template*. Based on the template-based docking of bound X-ray structures from the Benchmark 2, all templates were subdivided into good (yielding acceptable or better quality models), or bad (otherwise) templates. The docking models built from distorted protein structures were compared to the corresponding X-ray-based models in terms of fraction of shared contacts, FSC_{ij} (Eq. 5-2, main text). With the increasing distortions in monomers (vertically stacked plots), similarity between docking models decreases (distributions shift to the left), but non-zero similarity values suggest the conservation of the general binding positions. Such trend is true for both good and bad templates, although for a small fraction of bad models a complete loss of similarity is observed (minor peak in distributions at ~0).

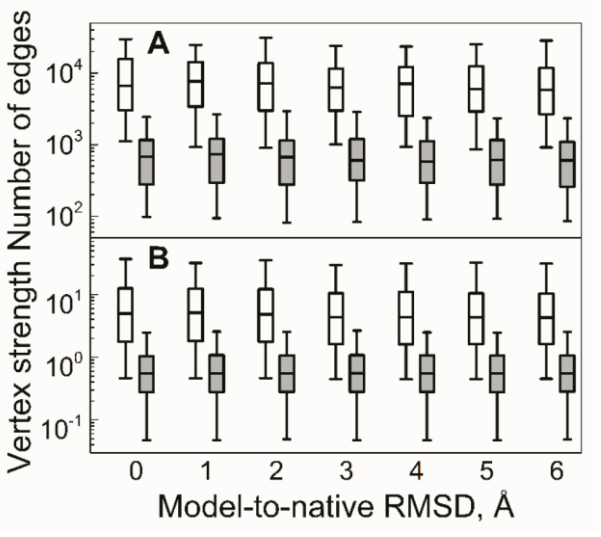

Figure C-8: *Comparison of graph properties for top-ranked and random free docking predictions at different levels of monomer distortions*. The top 1000 predictions are shown by open boxes, and the random ones by gray boxes. Similarity graphs for the docking predictions were built for each of 165 complexes from the Benchmark 2 at each distortion level (see 5.2 Methods). The box-and-whiskers distributions are for the number of edges in a graph (A), and for vertex strengths (B) in all 165 graphs at each distortion level. The vertex strength was calculated as sum of weights (FSC_{ij} values) of all edges originating from a given vertex. Box areas and whiskers contain 25 – 75% and 5 – 95% of data, respectively (outliers not shown).

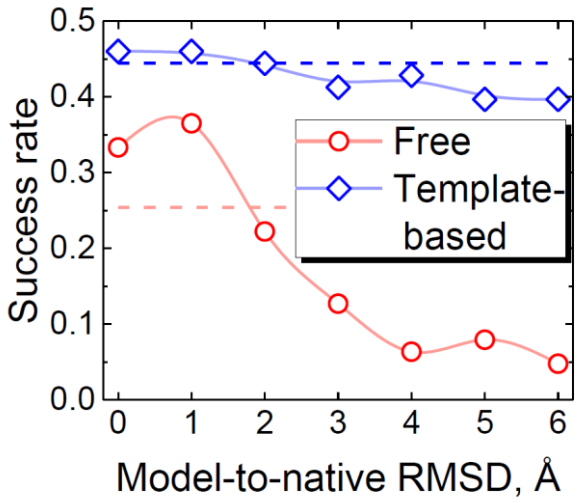

Figure C-9: *Docking of models vs. docking of unbound X-ray structures*. Free and template-based docking success (defined as one correct prediction in top 10, see main text) rates for 63 complexes from the Models Docking Benchmark 1 are shown as function of monomers' accuracy. Dashed lines show the performance of the two methods on the corresponding unbound X-ray structures. Average C^{α} RMSD between bound and unbound conformations of proteins in the Benchmark 1 is 1.36 Å.

Appendix D

List of publications

Reprinted in this thesis

1. **Anishchenko I**, Kundrotas PJ, Tuzikov AV, Vakser IA. Protein models: The Grand Challenge of protein docking. *Proteins.* 2014; 82: 278–287.
2. **Anishchenko I**, Kundrotas PJ, Tuzikov AV, Vakser IA. Protein models docking benchmark 2. *Proteins.* 2015; 83: 891–897.
3. **Anishchenko I**, Kundrotas PJ, Tuzikov AV, Vakser IA. Structural templates for comparative protein docking. *Proteins.* 2015; 83: 1563–1570.
4. **Anishchenko I**, Kundrotas PJ, Vakser IA. Modeling complexes of modeled proteins. 2016. *Submitted*
5. **Anishchenko I**, Kundrotas PJ, Vakser IA. Structural quality of unrefined models in protein docking. 2016. *To be submitted*

Other related work

1. Lensink MF, Velankar S, Kryshtafovych A, Huang SY, Schneidman-Duhovny D, Sali A, Segura J, Fernandez-Fuentes N, Viswanath A, Elber R, Grudinin S, Popov P, Neveu E, Lee H, Baek M, Park S, Heo L, Lee GR, Seok C, Qin S, Zhou HX, Ritchie DW, Maigret B, Devignes MD, Ghoorah A, Torchala M, Chaleil RAG, Bates PA, Ben-Zeev E, Eisenstein M, Negi SS, Weng Z, Vreven T, Pierce BG,

Borrman TM, Yu J, Ochsenbein F, Guerois R, Vangone A, Rodrigues JPGLM, van Zundert G, Nellen M, Xue L, Karaca E, Melquiond ASJ, Visscher K, Kastritis PL, Bonvin AMJJ, Xu X, Qiu L, Yan C, Li J, Ma Z, Cheng J, Zou X, Shen Y, Peterson LX, Kim HR, Roy A, Han X, Esquivel-Rodriguez J, Kihara D, Yu X, Bruce NJ, Fuller JC, Wade RC, **Anishchenko I**, Kundrotas PJ, Vakser IA, Imai K, Yamada K, Oda T, Nakamura T, Tomii K, Pallara C, Romero-Durana M, Jiménez-García B, Moal IH, Férnandez-Recio J, Joung JY, Kim JY, Joo K, Lee J, Kozakov D, Vajda S, Mottarella S, Hall DR, Beglov D, Mamonov A, Xia B, Bohnuud T, Del Carpio DA, Ichiishi E, Marze N, Kuroda D, Burman SSR, Gray JJ, Chermak E, Cavallo L, Oliva R, Tovchigrechko A, Wodak SJ. Prediction of homo- and hetero-protein complexes by ab-initio and template-based docking: a CASP-CAPRI experiment. 2016. *Proteins. Accepted*

2. Hadarovich A, Kundrotas PJ, **Anishchenko I**, Tuzikov AV, Vakser IA. GO-score: a new functional ontology-based measure for comparative protein docking. 2016. *To be submitted*
3. **Anishchenko I**, Kundrotas PJ, Vakser IA. Novel statistical contact potentials for structure modeling of proteins and protein-protein complexes. 2016. *To be submitted*