

Modeling Protein Assemblies in the Proteome

by

Güray Kuzu

A Thesis Submitted to
the Graduate School of Sciences and Engineering
in Fulfillment of the Requirement for
the Degree of

Doctor of Philosophy
in
Computational Sciences and Engineering

Koç University

October, 2014

to my dad...

ABSTRACT

Most (if not all) proteins function when associated in multimolecular assemblies. An important aim of structural biology is to attain the structures of protein assemblies at the atomic scale. Experimentally, structures are increasingly available but many are still missing or incomplete. Some experimental methods provide high resolution data for small proteins and some provide low resolution data for large proteins. Computational approaches can help bridge this resolution gap. They are needed to determine structural data of multi-molecular protein assemblies at atomic scale. Existing computational methods have made substantial progress toward this aim; however, current approaches are still limited. Some involve manual adjustment of experimental data; some are automated docking methods, which are computationally expensive and not applicable to large-scale proteome studies; still others exploit the symmetry of the complexes, thus they are not applicable to non-symmetrical complexes. Our study aims to take steps toward overcoming these limitations. We have developed a strategy to construct protein assemblies computationally based on binary interactions predicted by a motif-based protein interaction prediction tool, PRISM (PRotein Interactions by Structural Matching). PRISM predicts pair-wise interactions; here we take a step toward multimolecular assemblies, which reflects the more prevalent cellular scenarios. This method is able to construct homo-/hetero-complexes and symmetric/asymmetric complexes without a limitation on the number of components, considers conformational changes and is applicable to large-scale studies. We modeled a benchmark set of various protein assemblies starting from the unbound forms and obtained successful predictions (0.5 - 5.6 Å). Moreover, we exploit different conformations of the proteins available in the Protein Data Bank (PDB) to consider protein flexibility in modeling protein assemblies. We could increase the success in prediction of binary interactions from 27% to 67% and obtained higher accuracy in modeling protein assemblies by exploiting alternative conformations. Furthermore, we modified our method to exploit

electron microscopy (EM) density maps to eliminate improper structures. Filtering structures through EM data prevents assembly construction based on wrong structures and saves computational time. We successfully modeled protein assemblies using EM data, most of them with RMSD less than 5 Å and correlation in density maps are close to or higher than 0.8. Comparing our results with other methods' showed higher accuracy in our interface predictions. We present the methods, illustrate their results, and highlight the current limitations.

ÖZET

Proteinlerin çoğu (belki de hepsi) çok moleküllü yapılar halinde işlevlerini yerine getirir. Yapısal biyolojinin önemli bir amacı da protein tümleşiklerinin yapılarını atom boyutunda tanımlayabilmektir. Deneysel olarak elde edilen mevcut protein yapıları artıyor olsa da birçoğu halen bilinmiyor ya da eksik olarak biliniyor. Bazı deneysel yöntemler küçük proteinler için yüksek çözünürlükte, bazıları da büyük proteinler için düşük çözünürlükte veri sağlıyor. Hesaplamalı yaklaşımlar, bu çözünürlük ayrımını birleştirmek için yardım edebilirler. Bu teknikler çok moleküllü protein tümleşiklerinin yapılarını atom seviyesinde belirlemek için gereklidir. Mevcut hesaplamalı yöntemler bu amaç doğrultusunda önemli katkılarda bulunmuştur, ancak hala geçerlilikleri sınırlıdır. Bazılarında deneysel veriye göre ayarlama yapılması gereklidir, bazıları çok hesaplama gerektiren ve büyük boyutlardaki proteom çalışmalarında kullanılamayacak protein yerleştirme yöntemleridir, diğerleri de tümleşik yapılardaki simetriği temel aldığından simetrik olmayanlarda kullanılamaz. Bu çalışmadaki amacımız bu sınırlamaları aşmaktır. Protein tümleşiklerini, motifleri temel alarak protein etkileşimlerini tahmin eden bir araç (PRISM: Yapısal eşleme ile protein etkileşimleri) kullanarak elde ettiğimiz ikili etkileşimler üzerinden hesaplamalı olarak kurduk. PRISM ikili etkileşimleri tahmin eder; biz ise burada hücre içinde daha çok görülen bir olguyu, çok moleküllü tümleşikleri çalıştık. Bu yöntem, aynı ya da farklı proteinlerden oluşan, simetrik ya da simetrik olmayan tümleşikleri, bileşen sayısında sınırlama olmadan ve yapısal değişiklikleri dikkate alarak kurabilir ve büyük boyutlardaki çalışmalarda kullanılabilir. Çeşitli protein tümleşiklerinden hazırladığımız değerlendirme kümesindeki örnekleri, proteinlerin bağlanmamış hallerinden başlayarak başarılı bir şekilde tahmin ettik (0.5 - 5.6 Å). Ayrıca, protein tümleşiklerini modellerken proteinlerin bükülgenliğini hesaba katmak için Protein Veri Bankasındaki (PDB) farklı yapılarından yararlandık. Bu sayede, ikili etkileşimlerin tahminindeki başarıımızı %27'den %67'ye çıkardık ve protein tümleşiklerinin modellenmesinde daha doğru

sonular elde ettik. Bunlara ek olarak, yntemimizi elektron mikroskobu (EM) yoėunluk haritasını kullanarak uygun olmayan yapıları eleyecek ekilde deėiřtirdik. Yapıları EM verisinden geirmek yanlış yapılarla protein tmleřiklerin kurulmasını nler ve fazla hesaplamalardan kurtarır. EM verilerini kullanarak protein tmleřiklerini başarılı bir ekilde modelledik; sonuların biroėunda ortalama karekk uzaklık 5 Å'dan az ve yoėunluk haritaları ile baėıntıları 0.8'e yakın ya da daha byk. Diėer yntemlerle karřılařtırdıėımızda kendi yntemimizin protein ara yz tahmininde daha doėru sonular elde ettiėini bulduk. Bu alıřmada yntemlerimizi tanımlar, sonularımızı sunar ve mevcut sınırlamalarımızı ortaya koyuyoruz.

ACKNOWLEDGEMENTS

I would like to present my sincere thanks to my advisors Attila Gürsoy and Ozlem Keskin Özkaya. The whole story started when they accepted me to their group. We have done great jobs together and published several papers; maybe my name is in the first place but their names are at the end, for sure. Without their support, guidance and patience, none of them could be done.

I would like to thank Ruth Nussinov, who is an amazing professor. Her broad experience and deep knowledge greatly contributed to this project. I would like to thank also my thesis committee, Burak Erman, Mehmet Sayar, Nathan Lack, Türkan Haliloğlu and Elif Özkırımlı Ölmez, especially for their precious time. It was a great pleasure and honor to have them in the jury.

I acknowledge the financial support of the Scientific and Technological Research Council of Turkey (TÜBİTAK) and facilities of Koç University during my doctorate study.

I thank my officemates, all members of the COSBI family. I would like to thank three of them especially; I had a nice and easy start with the presence of Nurcan Tunçbağ in my first months; I am very thankful to Engin Çukuroğlu for his help in technical problems and scientific discussions we had together; it was always nice to have a computer geek in the group, like Alper Başpınar, while working computationally.

I also thank my dear friends, Mustafa Buzkan and Önder Onaran, and my housemates, Engin Çukuroğlu, Utku Boz and Barış Çağlar for the joy and happy times. I would like to give my special thanks to my family and Elif Muku for their endless support, patience and love during these four years.

TABLE OF CONTENTS

Chapter 1 INTRODUCTION.....	15
Chapter 2 LITERATURE REVIEW.....	18
2.1 Physical and Chemical Aspects of Protein-Protein Interactions.....	18
2.1.1 Interaction Occurs Through the Interface.....	18
2.1.2 Structural Features of Interfaces.....	21
2.1.3 Interfaces are Conserved Throughout the Evolution.....	22
2.1.4 Some Residues are Energetically More Important in the Interface: Hot Spots	24
2.2 Identification of Protein-Protein Interactions.....	24
2.2.1 Computational Methods to Predict Structure of Binary Interactions	26
2.2.2 Computational Methods to Model Protein Assemblies.....	31
2.2.3 Electron Microscopy Density Map is a Helpful Guide	34
Chapter 3 MATERIALS AND METHODS	38
3.1 PRISM: PRotein Interactions by Structural Matching.....	38
3.1.1 Template Dataset	40
3.1.2 Target Dataset.....	41
3.1.3 Prediction of Protein-Protein Interaction.....	41
3.2 Assembly Construction	43
3.2.1 Constructing Protein Assemblies Based on Binary PRISM Predictions.....	43
3.2.2 Protein Assembly Benchmark and Evaluation of the Predictions.....	47
3.3 Exploiting Alternative Conformations to Improve Predictions	50
3.3.1 PRISM Predictions with Different Conformations and the Docking Benchmark Dataset.....	52
3.3.2 Eliminating Bias by Discarding Information on the Bound Form	57
3.4 Constructing Protein Assemblies Based on EM Density Map.....	60
3.4.1 Case Studies and the Benchmark Dataset in Modeling Protein Assemblies..	60
Chapter 4 RESULTS OF MODELING PROTEIN ASSEMBLIES.....	66

4.1 Reconstruction of Assemblies.....	66
4.2 Construction of Assemblies from Unbound Protein Structures.....	68
4.3 Modeling Complexes of MEK, KSR, ERK and Raf Proteins.....	70
4.4 Limitations in Constructing Protein Assemblies.....	73
4.5 Concluding Remarks.....	74
Chapter 5 RESULTS OF EXPLOITING CONFORMATIONAL ENSEMBLES.....	76
5.1 Exploiting Conformational Ensembles in PPIs.....	77
5.1.1 PPI Prediction Results	77
5.1.2 95% Sequence Homology was Used to Enlarge the Target Set More	86
5.1.3 A Case Study: ERK Interactions	93
5.2 Exploiting Conformational Ensembles in Modeling Protein Assemblies.....	101
5.3 Concluding Remarks.....	103
Chapter 6 RESULTS OF EXPLOITING ELECTRON MICROSCOPY DATA.....	106
6.1 Results.....	108
6.1.1 Case Study 1. HIV Protein gp120 – Membrane Protein CD4 – Antibody 17b Complex	108
6.1.2 Case Study 2. HIV Protein gp120 – Membrane Protein CD4 – Antibody m36 Complex	110
6.1.3 Benchmark.....	110
6.2 Concluding Remarks.....	120
Chapter 7 CONCLUSION.....	124
BIBLIOGRAPHY.....	127
VITA.....	155
APPENDIX.....	156

LIST OF FIGURES

Figure 2.1 A two-chain interface.	20
Figure 2.2 Blind docking and knowledge-based approaches	28
Figure 2.3 Cn and Dn symmetries	33
Figure 2.4 The EM density map and the PDB structure of HIV protein interaction	35
Figure 2.5 EMDB 5020 in different contour levels	36
Figure 3.1 PRISM algorithm.	39
Figure 3.2 Assembly construction based on PRISM binary predictions.	45
Figure 3.3 Exploiting different conformations to enlarge target set.	51
Figure 3.4 An example of allosteric effect on binding – Case 17.	54
Figure 3.5 An example of conformational change upon binding – Case 12ii.	55
Figure 3.6 PRISM predictions with standard and enlarged target datasets.	56
Figure 3.7 Prediction using or not using the structures of bound forms.....	59
Figure 3.8 Flowchart of assembly construction using EM density map.....	62
Figure 4.1 Energy score vs. RMSD of predictions.	67
Figure 4.2 Models of complexes of MEK, KSR, ERK and Raf proteins.	72
Figure 5.1 Distribution of predictions based on energy and IS-score.	91
Figure 5.2 MAPK signaling pathway given in the KEGG database.	96
Figure 5.3 Interactions of ERK protein as given in the KEGG database and predicted by PRISM.	97
Figure 5.4 Predicted interactions of MEK1-MEK2.....	98
Figure 5.5 Trimer structure of ERK, MEK1 and MEK2.	100
Figure 5.6 Interacting conformations of ERK interacting proteins.	101
Figure 6.1 Predicted structure docked in the EM density map.	107
Figure 6.2 HIV interactions.	109
Figure 6.3 Benchmark assemblies docked into the EM maps.	113
Figure 6.4 Results of segmentation by Segger	118
Figure 6.5 Models obtained by using Multifit and Segger.	119

LIST OF TABLES

Table 2.1 Advantages and limitations of experimental techniques in identification protein structures.....	26
Table 3.1 FiberDock energies of interfaces in the benchmark	44
Table 3.1 FiberDock energies of interfaces in the benchmark	46
Table 3.2 Structural features of the benchmark proteins	49
Table 3.3 Bound and unbound forms of benchmark proteins.....	50
Table 3.4 FiberDock energies of the bound forms given in the benchmark dataset.....	53
Table 3.5 Alternative structures as the substitute of the bound form	58
Table 3.6 Changes in template organization.....	58
Table 3.7 Scores obtained and templates used in homology modeling	61
Table 3.8 Benchmark dataset of modeling protein assemblies.....	64
Table 4.1 Results for construction of protein assemblies starting from their components	68
Table 4.2 Results for the construction of protein assembly starting from the unbound forms of the components.....	69
Table 5.1 Energy values of PRISM results obtained using standard template set.....	78
Table 5.2 Favorable PRISM predictions of difficult cases using standard template set and standard target set.....	79
Table 5.3 Favorable PRISM predictions of difficult cases using standard template set and enlarged target set	80
Table 5.4 Energies, IS-scores and contact fractions of PRISM predictions obtained using standard template set and enlarged target set	81
Table 5.5 Energies, IS-scores and contact fractions of PRISM predictions obtained using standard template set and standard target set	82
Table 5.6 Energy values of PRISM results obtained using rearranged templates and target sets	83
Table 5.7 Favorable predictions using rearranged templates and standard target set.....	83

Table 5.8 Favorable predictions using rearranged templates and enlarged target set	84
Table 5.9 Energies, IS-scores and contact fractions of PRISM predictions obtained using rearranged templates and standard target set	84
Table 5.10 Energies, IS-scores and contact fractions of PRISM predictions obtained using rearranged templates and enlarged target set	85
Table 5.11 Favorable PRISM predictions of difficult cases obtained using additional conformations found by 95% sequence homology	87
Table 5.12 Energies, IS-scores and contact fractions of PRISM predictions obtained using additional conformations found by 95% sequence homology	87
Table 5.13 The success of PRISM predictions	90
Table 5.14 Successful PRISM predictions in energetically favorable PRISM predictions	90
Table 5.15 Biological/non-biological interaction test results of predictions with favorable energies but lower IS-scores	92
Table 5.16 Different conformations of ERK interacting proteins	94
Table 5.17 Energetically most favorable PRISM predictions of the query ERK interacting proteins	94
Table 5.18 Energetically most favorable PRISM predictions obtained using alternative conformations of the ERK interacting proteins	95
Table 5.19 Alternative conformations found from the PDB	102
Table 5.20 Results for the construction of protein assemblies starting from alternative conformations of the unbound forms.....	103
Table 6.1 Assembly construction results of the benchmark dataset	112
Table 6.2 Assembly construction results of three methods in our benchmark set.....	117
Table 6.3 Assembly construction results in the Multifit benchmark set	120

NOMENCLATURE

a	Alpha
b	Beta
f_{nat}	Native fraction
$f_{\text{non-nat}}$	Non-native fraction
iRMSD	Interface, root mean square distance
CAPRI	Critical assessment of predicted interactions
CDD	Conserved domain database
EM	Electron microscopy
EMDB	Electron microscopy data bank
FFT	Fast Fourier transform
FRET	Fluorescence resonance energy transfer
HIV	Human immunodeficiency virus
KSR	Kinase suppressor of Ras
MAPK	Mitogen-activated protein kinase
MS	Mass spectrometry
NMA	Normal mode analysis
NMR	Nuclear magnetic resonance
PDB	Protein data bank
PPI	Protein-protein interaction
PRISM	Protein interactions by structural matching
RMSD	Root mean square distance
RP-MS	Radical probe-mass spectrometry
RSA	Relative surface accessibility
SAXS	Small angle x-ray scattering
SCOP	Structural classification of proteins
VMD	Visual molecular dynamics

Chapter 1

INTRODUCTION

Proteins function through interactions; and protein-protein interactions (PPIs) play a crucial role in all biological processes. An interaction with another protein physically perturbs the structure; and the perturbation (or, signal) propagates in the cell [1]. To understand how signals propagate and how regulation takes place, we need to know the interactions of the proteins. Most (if not all) proteins function when associated in multimolecular assemblies. Attaining the structures of protein assemblies at the atomic scale is an important aim of structural biology. Experimentally identified structures are increasing; yet, the known structural data is still limited, and the available structures have high resolution for the small proteins and low resolution for the large ones. Computational approaches can help to verify experimental observations, predict new ones and also assist to bridge the resolution gap between high and low resolution scales of the experimental results.

Each computational method to model protein assemblies has its own limitations and drawbacks. Main goal is to obtain the most accurate predictions; however, applicability in a large scale and coverage of protein flexibility are the significative factors here. The search of all possible relative orientations and the appropriate conformations of the proteins bring huge computational cost; on the other hand, treating structures as rigid bodies disregards protein flexibility and leads to not very accurate predictions. Another important problem is to choose the ‘best’ prediction among the solution set obtained using the prediction method. Computational methods are needed especially in the absence of the experimentally verified structures; therefore, they should provide one structure as the solution on their own. Experimental data can be exploited for the selection of the solution and the verification of the structure. The reciprocal assistance of experimental and computational approaches assists to obtain more accurate structure

of protein assemblies, therefore, to understand biological mechanism and cellular processes in the organisms.

This dissertation presents a novel method to construct multi-molecular protein assemblies and overcome the limitations of current methods; integrative methods need curation and manual adjustment of experimental data, *ab initio* docking methods are computationally expensive and not applicable to large-scale proteome studies, and some methods exploit the symmetry of the complexes and not applicable to non-symmetrical complexes. Our method can model symmetric/asymmetric, cyclic/acyclic and hetero-/homo-complexes without the limitation of subunit numbers. We exploit different conformation of the proteins and experimental data like electron microscopy maps to obtain better predictions. These subjects are covered in separate chapters; the outline of this dissertation is as follows:

In Chapter 2, we present an extended review focusing on structural aspects of protein interactions. The chapter covers physical and chemical aspects of the PPIs, structural and evolutionary perspective of protein interfaces and their features, and the importance of energetically significant residues, hot spots.

In Chapter 3, we present the materials and methods. First we present the binary interaction prediction method which our assembly construction method is based on and then introduce the algorithm of the assembly construction. After that we explain the method to find alternative conformations of the query proteins and next we introduce another method to exploit electron microscopy density maps. We also present the benchmark sets and the evaluation of the predictions.

In Chapter 4, we model the multi-molecular protein assemblies. We test our method on a benchmark set and show the results, then illustrate on a case study. After that we present its limitations and discuss it.

Chapter 5 covers how exploiting conformational ensembles helps in modeling protein-protein interactions and protein assemblies. Alternative conformations bring protein flexibility in predictions of knowledge-based methods. We first test on a docking benchmark set and show protein interaction predictions are improved by exploiting different conformations. Then we show on a case study that alternative conformations enrich protein interaction networks. After that the method is applied in modeling protein assemblies.

Chapter 6 covers how exploiting experimental data improves predictions in modeling protein assemblies. Electron microscopy can filter structures predicted by the computational methods and select the best model among the solution set. We exploit electron microscopy maps at each addition of a protein to the assembly, which eliminates improper models at each step and saves computational time. We test in benchmark sets and compare it with other methods.

At the end of this dissertation we present a final chapter discussing the results with a major conclusion of this study and explaining future directions.

Chapter 2

LITERATURE REVIEW

In this chapter, we review protein interactions from physical and chemical aspects, and methods in identification of protein-protein interactions and protein assemblies. In the first section, we represent how proteins interact physically, chemical interaction at the interface, conservation of residues and hot spot residues. In the second section, we review experimental and computational methods to determine binary interactions and protein assemblies; then, we introduce electron microscopy data and the methods using it in the construction of protein complexes.

2.1 Physical and Chemical Aspects of Protein-Protein Interactions

Biological reactions occur through the proteins; and the interactions among them play a crucial role. Protein interactions are biophysical phenomena. Although the medium is full of numerous molecules at different sizes, proteins to be interacted find each other and the interaction occurs. Proteins are interacting through their surfaces with the help of the shape and biochemical complementarity regarding the flexibility of the molecules and the environmental conditions [2-6]. Hence, protein interactions are special reactions rather than random processes. In this section, we first describe the physical and chemical aspects of protein-protein interactions, features of interfaces and introduce the residues evolutionary and energetically important.

2.1.1 Interaction Occurs Through the Interface

The region where the two molecules are contacting is called the binding site, or considering both sides, the interface (**Figure 2.1**). Understanding the protein structures is very important to identify these regions. If the structures of contacting proteins are known, it is quite easy to determine the interface. Interfacial residues are usually found

by calculation of close residues from two sides based on the distance in the three dimensional space [7-9] or with the help of accessible surface area calculations [9-11].

Many studies investigate properties of interactions to enlighten protein interaction phenomena [4, 5, 7, 10-22]. The stability of the interaction is provided with binding forces; hydrogen bonds, salt bridges, electrostatic interactions and hydrophobic attractions [23, 24]. Disulphide bonds are also but rarely seen between binding proteins. These attractions determine also the specificity of the interaction [24]. For example, obligate complexes rely usually on hydrophobic attractions and transient complexes on hydrogen bonds and salt bridges [4, 10, 12, 24] although obligate interfaces are not more hydrophobic than transient ones [25, 26]. Understanding such details of the complexes is crucial to understand biological processes; many signaling events, which are known to be disease-related, are regulated by transient protein-protein interactions [27, 28].

Considering the composition of the residues, we know that they usually differ between obligate and transient, or homo- and hetero-complexes [4, 10, 12, 24, 29-33]. Hydrophobic residues like Methionine, Phenylalanine, Proline, Alanine, Aspartic acid and Leucine (but not Isoleucine and Glycine) are commonly found in homodimer interfaces; whereas hydrophilic residues like Tryptophan, Cysteine, Histidine, Glutamine, Asparagine, Tyrosine and Serine (but not Threonine) are dominant in heterodimer interfaces [12, 33]. Interfaces are structurally similar to the cores of globular proteins [34], and many studies have shown that protein binding process is very similar to the protein folding [17-19, 35]. The structural shape of interfaces is planar and well packed, but also differs with respect to type of the interaction [10, 36]. For example, relatively large surface areas are observed in homodimer complexes compared to heterodimer complexes [12, 33].

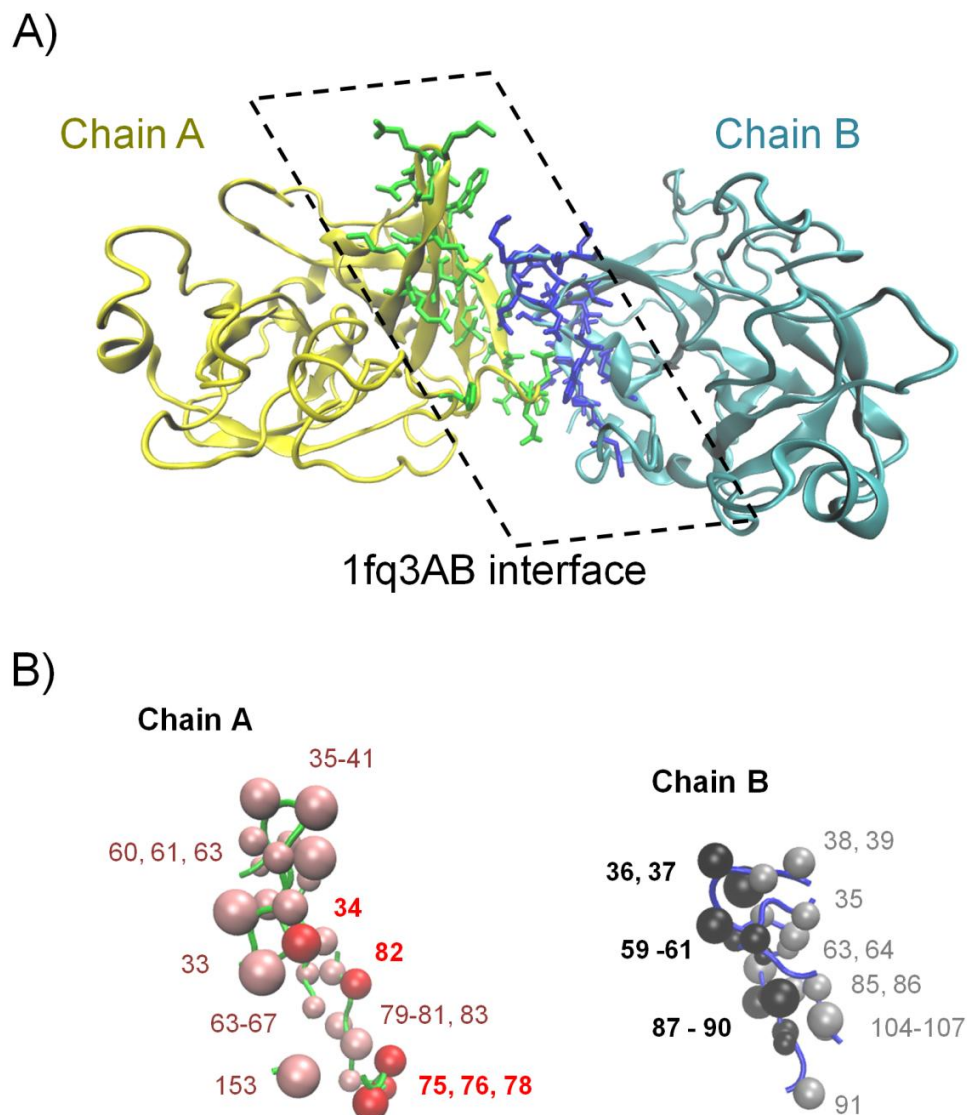


Figure 2.1 A two-chain interface [37].

A) An example of a two-chain interface (Protein Data Bank-PDB identifier: 1fq3. Chain A is shown in yellow and chain B is in cyan. Green and blue residues represent contacting residues which interact across the interface and the nearby residues. B) The interface consists of bits and pieces of each of the chains, and some isolated residues. The chain A side of the interface consists of 5 contacting (red) and 24 nearby (pink) residues. There are 9 contacting (black) and 17 nearby (gray) residues in the chain B interface.

2.1.2 Structural Features of Interfaces

Proteins interact biophysically. Therefore, fundamental determinants in protein interactions are the shape and physicochemical complementarity [2, 4, 38, 39]. Protein surfaces are full of pockets, crevices and indentations [40-43]. Some pockets which usually exist before binding are filled with the complementary protein like the key and lock model, when they associate during the interaction [44]. This is also valid for protein-ligand interactions. Size and shape of pockets are principally considered to design molecules or peptides like drugs. Therefore, structural information of the protein has a great importance in molecular interactions. The number of small pockets on the protein surface is not very few [7, 10, 12]; on the contrary, much more than expected before. Besides, a cavity that does not pre-exist can be formed upon interacting with a small molecule [45]. In the lack of cavities, it becomes more difficult to inhibit a PPI with a small molecule due to the flat large surface area.

In addition to the shape and size of the cavity, physicochemical properties of residues as well as their distribution in the cavity are important for the molecular interaction. Thus, biochemical and electronic complementarity are also important.

Another parameter to characterize the interface is amino acid frequency [31, 46, 47]. The amino acid propensity of the interface is similar to interior of the protein [14, 48]. The interface is usually rich of aromatic and hydrophilic residues, like Cysteine, Tyrosine, Phenylalanine and Tryptophan [14, 49]. Positively charged Arginine and Histidine also prefer to be at interface compared to both surface and core [50]. Threonine, Proline, Lysine, Glutamic Acid and Alanine are least commonly found in the interface [51]. Ofra and Rost [31] have predicted the type of the interaction with 63-100% accuracy rate just by using amino acid composition and residue-contact preferences. In another study, frequency of amino acids in transient and obligate complexes is analyzed. It is found that Glycine is more frequently seen in transient interfaces rather than the surfaces; whereas Glycine frequency is the same in interface

and on surface of obligate complexes [49]. In the same study, pairwise contact preferences of the amino acid types are also investigated. Cystine-Cystine shows the highest preference of all possible contacts. This can be due to its ability to form disulfide bond. It is also found that this pair is more frequent in obligate interfaces than in transient interfaces. Moreover, acidic and basic amino acids are observed to be contacted mostly with other types of amino acids, but rarely with other amino acids with similar physicochemical properties. However, Histidine is found as an exception. It prefers to contact with itself but not with acidic amino acids. Furthermore, nonpolar amino acids prefer to contact with other nonpolar amino acids.

The other steric property of protein-protein interfaces is that they generally have an accessible surface area in the range of 1200-2000 Å² [4, 52]. This property is commonly used to distinguish biological interfaces from non-biological ones. Non-biological interfaces are formed between proteins which do not come together and contact each other actually in their physiological states [53]. Non-biological interactions are also called as crystal packing. These proteins are experimentally crystallized together; but it is mostly due to the enforcement by the crystallographic packing environment in the experiment. Non-biological interactions mostly have smaller accessible surface areas, which are about 400-600 Å² [4, 54].

2.1.3 Interfaces are Conserved Throughout the Evolution

Interacting proteins find each other specifically and selectively in a crowded medium. Therefore, proteins should identify each other through complementary interaction sites. This requires the conservation of the same (or similar) residues in the right orientation for both partners of the interaction [49]. In order to preserve the interaction, residues in the active or binding sides of proteins resist undergoing a change through the evolution [55, 56]. It is proved in many studies that amino acids in interfaces are more conserved than on rest of the protein surfaces [49, 57-65]. Conservation analysis of residues can be used to predict protein binding sites. The

analyses can be based on the sequence [66-68], or the structure [65, 69, 70], or both sequence and the structure [66, 71, 72]. When obligate and transient complexes are compared, obligate complexes are found to have more conserved interfaces than the transient complexes [49, 64]. This supports that proteins of obligate complexes co-evolve, or proteins of transient complexes have higher ability to adapt to a change in their partners [64]. Moreover, analyses in the contact preferences of residues show that, residues prefer to contact with the other residues having a similar conservation grade [49]. In other words; highly conserved residues prefer to contact with highly conserved residues, and much variable residues prefer to contact with variable residues on the other side. Conserved residues are not randomly distributed and thus, some residues are more crucial than the others for the interaction [73]. If a contacting residue pair has an important role for the interaction, the residues are both conserved; if not, the residues are variable at the same grade [49].

The structures of interfaces are more conserved than the global structures of the proteins [74-76]. Even if their global structures and functions are different, proteins can interact through interfaces with similar architectures [7, 77]. Structures of the interfaces have been clustered based on the domain or the whole structure. The Conserved Domain Database (CDD) provides protein domains conserved in evolution, and they are extracted via multiple sequence alignments [78, 79]. Schroeder and coworkers [80] have found 6,000 distinct types of interfaces by clustering domain interfaces. Sali and his group [81] have shown that proteins in the same SCOP families have similar binding architecture. Aloy and coworkers [82] have classified domain-based interactions of known three-dimensional structure and developed 3did web server. PPIClust provides clusters of similar 3D interface patterns in protein complexes [83]. Gao and skolnick have found that structural space of protein-protein interfaces is close to complete and clustered interfaces of dimers into roughly 1,000 distinct types [84].

2.1.4 Some Residues are Energetically More Important in the Interface: Hot Spots

All contacting residues do not equally contribute to the binding energy. Regardless of the size of the binding site, only a few residues are responsible for the majority of the total binding energy [73, 85-87]. These residues are called “hot spots”. Hot spots can be identified experimentally by Alanine Scanning Mutagenesis [59, 85]. In this method, each residue is mutated to an Alanine; and if a significant drop is observed in the energy due to the mutation, the residue is assigned as a hot spot. Hot spots can have different physicochemical properties; for example, they can be hydrophobic or polar [45, 64, 88]. Arginine, Tyrosine and Tryptophan are frequently found as hot spots; whereas, Leucine, Serine, Threonine, Valine [86, 89] are rarely. Methionine is rarely [86] or frequently [64] found as hot spot by different studies.

Hot spots are buried and tightly packed in the three dimensional (3D) space [73]. They are located near the center of the interface and away from the solvent [86, 90]. However, they are usually found in discontinuous, discrete highly packed regions in interfaces. These clusters are called “hot regions” [73]. Furthermore, there is a strong correlation between structurally conserved residues and hot spots [64, 73, 91-93]. There are many computational studies based on their physical, biological and/or evolutionary features for the hot spot prediction. These approaches analyze combinations of hot spot features, like conservation, physicochemical properties, residue propensity, sequence profiles, accessible surface area, and contribution to binding energy [67, 71, 94-100]. Hot spots are also significant for drug design [101]. Since they are responsible for key contact potentials of the interactions, drugs are aimed to target these residues [90, 102-105].

2.2 Identification of Protein-Protein Interactions

Proteins function through interactions; and PPIs play a crucial role in all biological processes. An interaction with another protein physically perturbs the structure; and the perturbation (or, signal) propagates in the cell [1]. To understand how signals propagate

and how regulation takes place, we need to know the interactions of the proteins. Among the first steps toward this goal is the identification of PPIs. Experimentally, identification of pairwise PPIs has been addressed by techniques such as yeast two-hybrid system [106], phage display [107], protein arrays [108], and affinity purification [109], and databases like DIP [110], MINT [111], BIND [112], BioGrid [113], and IntAct [114], have usefully compiled these data. Based on these, protein interaction networks (PINs) have been constructed [115, 116], on a pathway [117-119] or proteome scale [120-122]. However, those high-throughput experiments can result in many false positives [123, 124]. In addition, numerous computational methods are used to identify protein interactions using sequence: Chou and Cai predicted PPIs in a hybridization space [125]; Shen et al. used n-gram method in prediction [126]; Dong and Zhou combined four kinds of basic building blocks, n-grams, patterns, motifs and binary profiles [127]; Guo et al. used a support vector machine approach [128]; Shia et al. predicted using correlation coefficient and high-quality interaction dataset [129]; Xia et al. used meta predictor to identify PPIs [130]. Despite all these efforts, 10% of the human interactome is known [131]. However, structural data is more trustworthy. It provides not only ‘which proteins interact’ but also ‘how proteins interact’. The structure of interacting proteins show which residues or atoms are in contact through the interactions, the interface residues and the binding site of the proteins. Structures of protein complexes can be obtained by several experimental techniques, such as X-ray crystallography [132, 133], nuclear magnetic resonance (NMR) spectroscopy [134], electron microscopy (EM) [135], Fluorescence Resonance Energy Transfer (FRET) spectroscopy [136], and small angle X-ray scattering (SAXS) [137], albeit at different resolutions. High resolution data is available in the Protein Data Bank (PDB) [138] (over 103,000 structures as of September 2014); however, it is incomplete and does not cover all structures. In addition, it may contain non-biological crystallographic packing interactions [139].

Experimental determination of protein complexes can be difficult. Despite many experimental methods, shortcomings abound (**Table 2.1**): X-ray crystallization is applicable only to molecules that can be cloned, crystallized and purified in large quantities, and often crystals suitable for structural determination of protein assemblies cannot be obtained [133]. NMR is limited to relatively small molecular sizes [140]. Cryo-EM, cryo-electron tomography, FRET spectroscopy and SAXS are more suitable for structural determination of large molecules and assemblies, but their resolution is not atomic scale [135, 141]. These limitations can be addressed by complementary experimental and computational techniques.

Table 2.1 Advantages and limitations of experimental techniques in identification protein structures

Experimental techniques	Advantages	Limitations
X-ray crystallization	provide high resolution data	crystallization problems, limited in size
NMR	provide high resolution data	limited in size
EM, FRET, SAXS	applicable for large proteins	provide low resolution data

2.2.1 Computational Methods to Predict Structure of Binary Interactions

Computational approaches can help to verify experimental observations and predict new ones. The aim is to predict PPIs through different approaches, such as using the information of the already known interactions, mimicking the natural mechanisms, following the structural and physicochemical complementarity or considering energetic interactions of atoms. Since the prediction is done computationally, it is cheaper than performing experiments, and thus, applicable in larger scales. Another aim is to overcome limitations of the experimental methods and determine the interactions that cannot be identified by experiments. Some computational methods exploit experimental data to improve their performance, which is a good example of complementarity of

experimental and computational approaches. The critical assessment of predicted interactions (CAPRI) [142] demonstrates the performance of docking strategies, and help in following their progress. It is an ongoing series of events in which researchers try to predict PPIs which have been recently determined experimentally. The main computational PPI prediction protocols are blind docking and knowledge-based methodologies (**Figure 2.2**). Blind docking strategies consider possible structural combinations of query proteins to find the ‘best’, native bound state. They search for the optimum combination, mostly energetically, of the appropriate conformations of the proteins interacting through correct sites. Thus, the simultaneous search is for conformations of the proteins and their relative orientations. For this comprehensive search, methods are developed based on different techniques. ZDock [143], PIPER [144] and GRAMM-X [145] use the Fast Fourier Transform (FFT) approaches in the global search for the binding mode of proteins. A layer of grid points that surround proteins are defined, and then FFT is used to optimize grid-based shape complementarity, physicochemical or electrostatic complementarity and/or desolvation. RosettaDOCK [146] is generally preferred in the high-resolution refinement step. ATTRACT [147] and FiberDock [148] use normal mode analysis (NMA) to sample conformational backbone variability. NMA introduces protein flexibility as a linear combination of discrete modes, and a coarse-grained protein model or a Monte Carlo simulation is combined to search for the energetically optimum conformations. Some methods utilize different experimental data to increase their accuracy. PatchDock [149] uses the entire surface; however, if experimental constraints are available they can limit the search. Its transformational search is based on local feature matching combined with the structural complementarity search. MolFit [150] and ATTRACT consider experimentally determined interface residues. Those residues assist to determine right orientations while searching the physicochemical and structural complementarity. ZDock and pyDock [151] block non-interface residues, which reduces the solution set and eliminates false positive predictions. HADDOCK [152] utilizes experimental data, mainly NMR, to extract interface information including contacts and relative

orientations. Using the experimental data, it determines both interacting and non-interacting residues (active and passive residues) and restricts the docking considering those residues. PROXIMO [153] and MultiFit [154] use radical probe mass spectrometry (RP-MS) and EM data in docking, respectively. These methods are developed for specific experimental methods. RP-MS probe the structure of proteins and their interaction with other macromolecules by treating with high fluxes of radicals, and PROXIMO identifies interacting proteins based on RP-MS results. Multifit simultaneously optimizes the fit of the model into an EM density map and the structural complementarity between proteins. It is applicable also on multi-molecular protein assemblies (**section 2.3**).

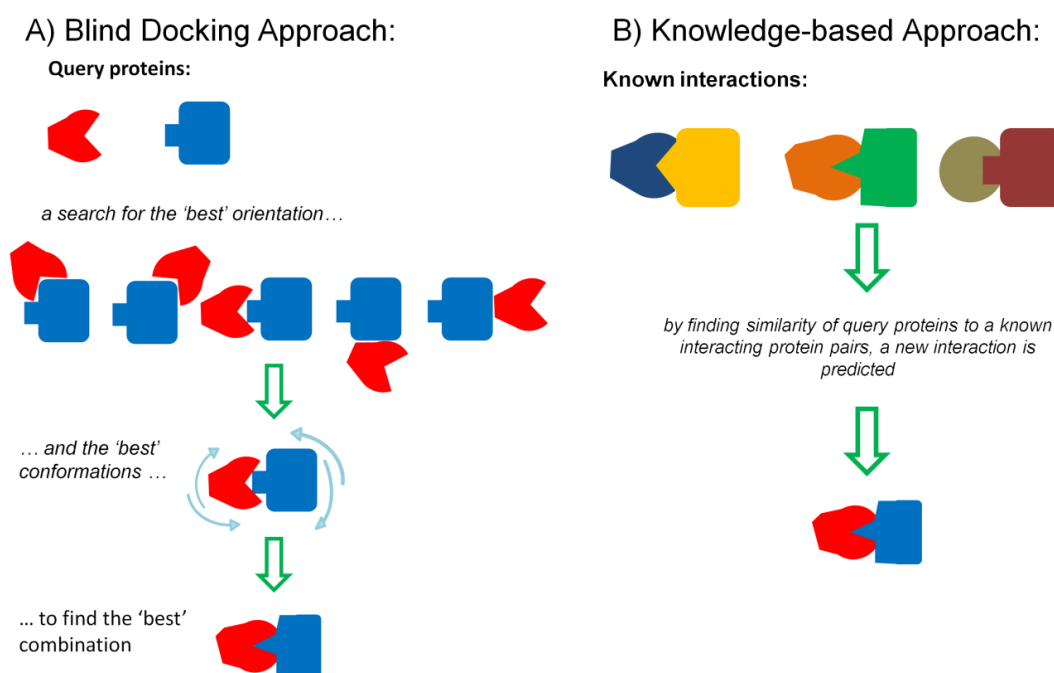


Figure 2.2 Blind docking and knowledge-based approaches

A) Blind docking approaches consider possible orientations and conformations of the proteins and search for the 'best' combinations of those. B) Knowledge-based approaches search for the similarity of query proteins to the known interacting pairs and predict new interactions.

Predicting PPIs by docking is computationally very expensive. Protein size affects the computation time, which makes docking of larger proteins much slower [155]. Moreover, scoring functions in the docking algorithms are still not optimal to predict which proteins interact and how they interact on the interactome scale [156, 157]. Consequently, docking approaches are not appropriate for large-scale studies [158, 159]. Data derived from known interactions can restrict the solution space in knowledge-based approaches. Homology-based methods can be powerful in predicting interactions (even for unstructured proteins) as shown by the pioneering work of Aloy and Russel [160], who also created a web server (InterPreTS [161]) which scores the predictions based on empirical potentials derived from known interactions. After identifying contacting residues in a known interaction, it is determined if the interactions are preserved through the empirical potentials by considering homologues of both interacting proteins. More recently, Kundrotas et al. [162] have also predicted the structures of interacting proteins based on sequence homology (HOMBACOP). The improvement comes from the alignment of the interface residues; they put extra weight on aligning interface residues in the dynamical programming, and increased gap penalties for the interfacial segments. The GWIDD database [163] usefully provides experimental and homology-based models, and WSsas [164] presents functional residues in structural homologs. IBIS [165, 166] analyzes and predicts interaction partners by considering both structure and sequence conservation and indicates locations of binding sites in proteins by means of their homologous proteins. Multiprospector [167] utilizes multimeric threading. It extends the threading approach to multimeric threading and considers both monomers and chains of complexes, and then computes the interaction energy based on the statistical interfacial pair potentials. Finally, M-Tasser takes a step further and predicts protein quaternary structure from sequence via a hierarchical method and also considers certain protein flexibility [168].

Domain information has also been utilized in protein interaction prediction. Shoemaker et al. [169] showed that members of domain families can dock in the same way. Davis et al. [170] matched the overall domains and scored predictions by statistical potentials derived from the known interactions. Aloy and coworkers [171] combined the overall structural fold and sequence similarities and developed the 3DID web server [172] which identifies domain-based interactions.

In interface-based approaches, an interaction between query proteins can be achieved fairly reliably if the structure of a similar interface is available, based on the similarity of query protein surfaces to interfaces of known interacting proteins [88]. This concept is similar to homology modeling for structure prediction. Using known data lowers substantially the computational time, making PPI prediction on a large-scale possible. PRotein Interactions by Structural Matching (PRISM) [88, 173, 174] is the pioneer interface-based PPI prediction method, which is applicable to structural interactome data. ISearch [175] is based on a similar concept but has a template set of domain-domain interactions. Sinha et al. [176] use a template-based docking independent of sequence homology and consider local structural alignments rather than global structural alignments.

Comparison of the knowledge based methods is not straightforward because they use different benchmark datasets. Nevertheless, their individual performance shows their success. In the work of Aloy and Russel [160], 59 of 2590 predicted interactions are experimentally verified. Multiprospector [167] predicts 36 homodimers and 15 heterodimers correctly among 40 homodimers and 15 heterodimers. In the work of Davis et al. [170], 270 binary and 8 multi complexes are experimentally verified among 3387 binary and 1234 multi complexes predicted. ISearch [175] can predict a model for 45 of 59 cases. Sinha et al. [176] correctly predict 56% of 372 complexes in their bound forms. PRISM can predict 87 of 88 cases of a docking benchmark dataset [177].

2.2.2 Computational Methods to Model Protein Assemblies

In vivo, the overwhelming majority of the proteins function when they are part not only of binary interactions, but of multimolecular assemblies. Protein assemblies are responsible for the vast majority (or all) of the processes in the cell [178]; the RNA polymerase transcription machinery [179], the ribosome, translation engine [180], chaperonins [181], and the proteasome, protein degradation machine [182], provide some examples. Computational methods are available to model the structures of protein assemblies. Even though they also face inherent shortcomings and hurdles, they can assist experimental techniques. Integrative structural determination methods combine experimental results from different sources with computational constraints, models and theory. EM maps are combined with atomic structures of single protein components [183-189], or atomic models [190]; and binary interaction data of the protein components of a complex obtained by affinity purification and mass spectrometry (MS) are integrated with comparative modeling [191]. Structures of complexes are modeled based on NMR-derived data [192, 193]. Eventually, several structures of protein complexes, like the nuclear pore complex [184], eukaryotic ribosome [194], human RNA polymerase II [195], AAA-ATPase/20S core particle subcomplex of the 26S proteasome [196] and histone methyltransferase complex Set1C from yeast [197] have been predicted using integrative methods. Despite their immense potential power, to date, these methods have been limited to a small set of proteins, especially to the most highly studied ones; this is because of the need for a broad range of different types of data, manual adjustment and data curation.

Ab initio docking approaches have been used for the prediction of structures of protein complexes. These methods are developed as a modification of binary interaction predictions and some utilize different types of experimental data to increase their accuracy. M-ZDOCK [198] constructs Cn symmetric multimers (**Figure 2.3**) based on the idea of ZDOCK [143, 199] which blocks non-interface residues in docking and can use experimental data to filter the solutions. As mentioned before, PatchDock [149,

200] finds solutions based on the shape complementarity and can use experimental data to detect binding site; SymmDock [149, 201] restricts the search to symmetric cyclic transformations and constructs homo-complexes with cyclic symmetry. Another useful docking tool is HADDOCK [202] which use utilizes a variety of experimental data, mainly NMR, same as in its binary predicitions. Six subunit complexes can be constructed, and the method has been tested on symmetrical cases. However expensive computation of the *ab initio* docking is a barrier for large-scale protein complex predictions. Computationally, modeling of multi-molecular assemblies from the structures of their monomeric components is challenging, because of the large number of possible combinations of the components [203].

Some studies have focused on the symmetry of the components of the complex. Eisenstein et al. [204] constructed the symmetrical structure of the helical protein coat of tobacco mosaic virus. Later, a similar approach was used to assemble cyclic (C_n) and dihedral (D_n) symmetrical structures [205, 206] (**Figure 2.3**). Comeau and Camacho [207] also predicted C_n and D_n symmetrical structures. In addition, they assembled oligomers starting from dimers. Schneidman-Duhovny et al. [201] developed a protocol to construct cyclic symmetrical structures, and Huang et al. [208] could dock C_2 symmetrical dimers. Andre et al. [209] developed a protocol to predict symmetrical assemblies starting from the structure or the sequence of a single subunit. Imposing symmetry constraints in the protocol limits the space of the predictions, making it unsuitable for the prediction of non-symmetrical protein complexes. Non-symmetrical complexes have not been studied as much as the symmetrical ones. Inbar et al. [203] developed a protocol for the construction of hetero multimolecular protein assemblies. In this multimolecular assembly protocol, CombDock, subunits are considered as “puzzle pieces” and the native complex as “puzzle solution”. CombDock considers all pairwise dockings and combinatorially builds the final assembly. Finding the right combination is computationally hard (NP-hard) [203], therefore, CombDock uses a heuristic based on greedy construction of subassemblies. The protocol has successfully

re-constructed a protein complex from its components. However, computing all pairwise dockings (N units, $N(N-1)/2$ pairwise sets of docking configurations) still presents challenges for both computation time, and in particular, may miss solutions where the complexes are less as dimers but gain stability in the larger assembly.

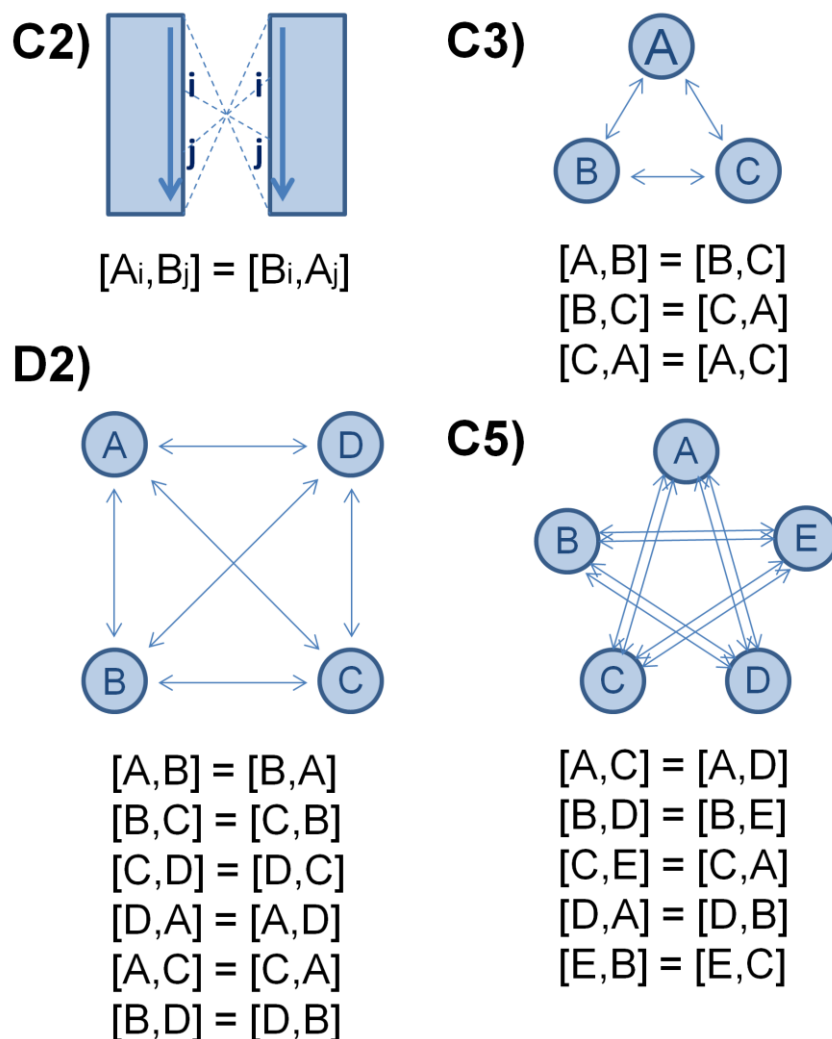


Figure 2.3 C_n and D_n symmetries

Examples of C_2 , D_2 , C_3 and C_5 are given. Symmetrical groups are given in square brackets.

2.2.3 Electron Microscopy Density Map is a Helpful Guide

EM is a preferred method to identify structures of protein assemblies and has come forward among others, FRET and SAXS, with the recent advances in its technology. It produces volumetric density maps [210-212] at progressively higher resolutions, down to 4-6 Å [213-215] (**Figure 2.4**). EM does not fix structures and proteins can be observed in their native environment, unlike in X-ray crystallography, which can lead to functionally irrelevant conformational changes in the formation of crystals. EM density maps are available in the EMDataBank (EMDB) [216] (over 2500 maps as of September, 2014). They are similar to clouds; denser regions indicate higher probability in observing atoms. Different contour levels show regions with different density (**Figure 2.5**). A contour level to show the ‘best’ image is assigned for each density map. EM density maps provide data allowing visualization of the whole complex but not how the proteins in the complex are oriented nor the interaction details. Since their resolutions are not high enough, an unambiguous model cannot be constructed based on just EM density maps. Structural data at higher resolution is used to interpret EM maps. EM density maps supported with high resolution data assist to understand the structure more accurately, therefore, decipher the function of the protein assembly, identify conformational changes and their regions, and determine regions of proteins whose structures are uncharacterized or can be characterized under certain conditions. Computational methods aim to fill that gap and bridge the resolution scale. Computational methods can exploit and complement EM data (i) to fit assembly subunits into their density maps, and (ii) to filter docking solutions obtained computationally.

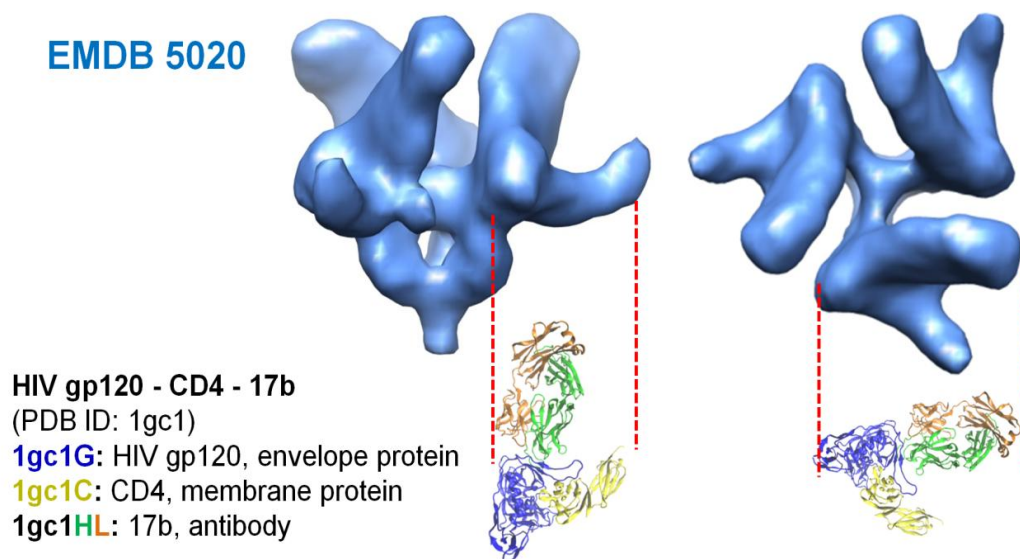


Figure 2.4 The EM density map and the PDB structure of HIV protein interaction

Side and top views of the EM density map and the PDB structure of the complex of envelope protein HIV gp120, membrane protein CD4 and antibody 17b is given. The density map covers three HIV gp120-CD4-17b complexes (PDB ID: 1gc1) and HIV gp41. Size comparison of the density map and the PDB structure is shown by red dashed lines.

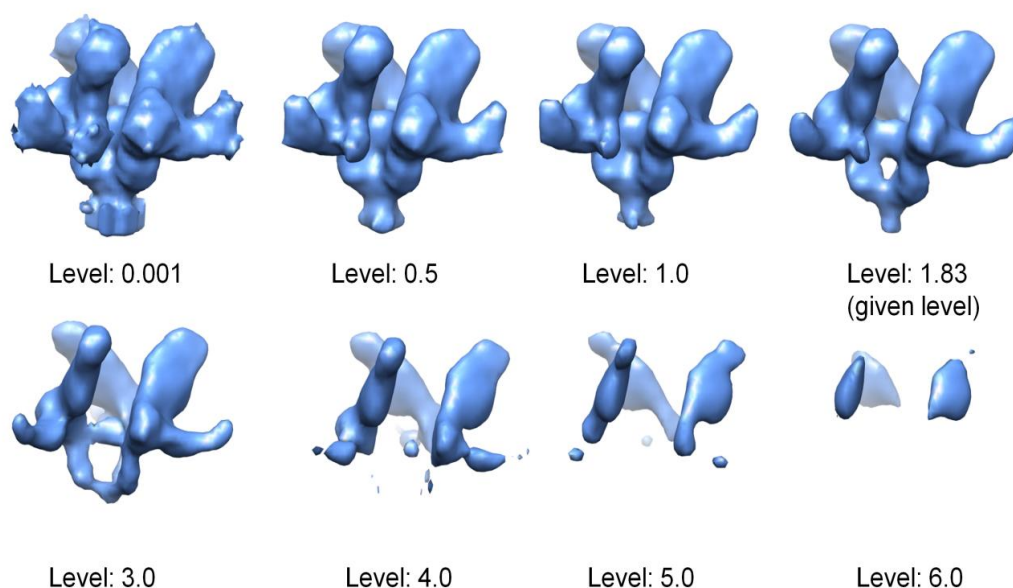


Figure 2.5 EMDB 5020 in different contour levels

Contour level is the threshold value for the density of EM maps. Higher levels are for the denser regions. Given level is 1.83 for EMDB 5020.

Computational methods, such as ADP_EM [217], Foldhunter [218], URO/UROX [219, 220], DockEM [221], EMfit [222, 223] and Situs [224], perform exhaustive search to fit protein structures into EM density maps. They perform rigid-body docking. Situs is also capable of flexible docking [225]. Some methods are specialized on particular features; FRM [226] and gEMfitter [227] use fast Fourier transform to fit structures into the density map, Gorgon [228] considers secondary structure matching and 3SOM [229] is based on surface overlap maximization. Some other methods do segmentation of EM density maps using different approaches, like level sets (VolRover) [230], elastic networks (hENM) [231], watershed [232] and watershed/scale-space filtering (Segger) [233, 234]. After the segmentation process, subunits can be docked individually into the segments obtained. MultiFit [154, 188] takes a step further and

does multiple docking. It fits multiple protein structures simultaneously based on molecular docking and molecular fitting approaches. Another important tool is UCSF Chimera [235], which is used for interactive visualization and analysis of molecular structures and related data, including density maps. Structures can be fitted into density maps automatically. Chimera has plug-in versions of Multifit and Segger. However, methods with the priority on EM density maps or to dock protein structures into EM density maps can provide models with false positive interfaces, structures in which proteins are not contacting or have many clashes, therefore, can lead to improper assembly structures.

The second strategy is to filter docking solutions using EM density maps. Computational docking methods [149, 198, 201-203, 236] construct protein assembly models by exploiting high resolution structural data. They produce a solution set, which includes different conformations of the assembly and false positives. It is difficult to select a specific model as the solution without using experimental data. EM density maps can help to evaluate models and select the ‘best’.

Chapter 3

MATERIALS AND METHODS

In this chapter, we present the materials and methods we used in the study. We first present the binary interaction prediction method, PRISM. Then, we introduce our method to construct multi-molecular protein assemblies. Next, we explain how we exploit alternative conformations of the proteins to consider protein flexibility in our predictions. After that, we introduce a modified version of the method to model protein assemblies, which exploit EM density maps. Here we also present the benchmark sets which we tested our methods on and evaluation of our predictions

3.1 PRISM: PRotein Interactions by Structural Matching

Our method to model multi-molecular protein assemblies is based on binary interaction prediction. We exploit a template based protein interaction prediction tool, PRISM [88, 173, 174] to predict binary interactions through structural motif search, and use these predictions to construct protein assemblies. It is based on the observation that proteins tend to interact via recurring motifs, regardless of the global similarity of the structures of the chains [237, 238]. Here, we first present the binary interaction prediction method, PRISM.

PRISM attempts to predict protein-protein interactions based on structural similarity of the proteins to the complementary sides of a known interface. If it is known that there is an interaction between proteins A and B, and protein A' is structurally similar to protein A and protein B' is structurally similar to protein B, it is claimed that A' and B' may interact with each other [88]. PRISM considers a potential binary interaction by querying whether target interfaces structurally and evolutionarily complement each other in a way similar to template interfaces. Then, by using FiberDock, flexible

refinement of docking solution candidates is performed by optimizing the side-chain orientations. Binding energy is also calculated for the refined structures. To carry out such a protocol, the first step involves the availability of target structures and generation of template datasets. A flowchart summarizing the prediction algorithm is given in **Figure 3.1**.

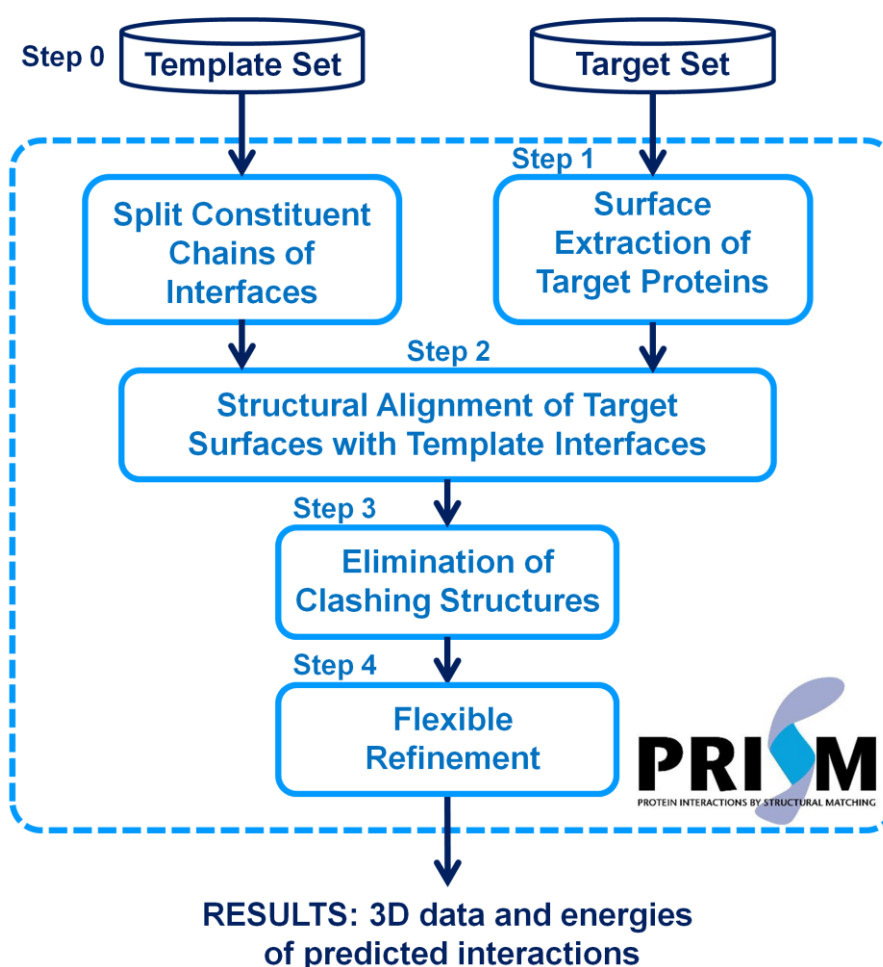


Figure 3.1 PRISM algorithm [236].

Steps 1-4 (flow through blue boxes) provide the PRISM flowchart. Inputs are template and target datasets and outputs are 3D structures of predicted binary interactions and their energies. Step 0 is the template organization, step 1 is surface extraction of target proteins, step 2 is the structural alignment process, step 3 is elimination of clashing structures, and step 4 is the flexible refinement process.

3.1.1 Template Dataset

All interfaces of two chain protein complexes available in the PDB [138] were extracted. Interfaces consist of interacting residues between two chains and neighboring residues. Neighboring residues are in the spatial vicinity of interacting residues and constitute the scaffold of the interface. Two residues from two different chains are considered as interacting if they are at a distance smaller than the sum of van der Waals radii plus a threshold of 0.5 Å. In addition, a non-interacting residue whose C α is closer than 6.0 Å to the C α of any interacting residue is marked as a neighboring residue. In order to obtain a non-redundant dataset, 49,512 two-chain interfaces (as of February 2006) extracted in the first step were clustered structurally following an iterative all-against-all structural comparison in a sequence order-independent way [7, 239]. 8,205 clusters were obtained. Interface members of each cluster are structurally similar to the representative interface. A cluster should contain at least five non-homologous sequences.

The template interface can be constructed in several ways: one can use (i) all representatives (8,205) of the interfaces, or (ii) a subset of the representatives, for example, the hetero-dimeric protein interfaces (1,036), or the non-obligate protein interfaces (158 interfaces). The results strictly depend on the template set. If there is no similar motif in the template set, the algorithm cannot find any similarity between the target protein and template structures; thus a potential interaction for target proteins cannot be predicted. Therefore, choosing the right template set for the target proteins is very important. User can also use his own template set, but the data of the structures in template set should be added in PDB format. Although it may seem as a disadvantage that outcome is a function of the template set, the algorithm finds reliable results in a short computation time if such motifs are available in the template set. The type of reduction of the template set is determined with respect to characteristics of the query

molecules. Computational hot spots are found by using the HotPoint web server [240]. HotPoint web server is used to find computational hotspots. The PDB code of the input protein should be entered or PDB files of the protein can be loaded. The interacting chains are given and the distance threshold to fetch the interface residues can be chosen as default value, which is summation of van der Waals radii of two atoms plus 0.5 Å, or a value defined by the user. On the result page, contacting residues are displayed with their features (residue number, residue name, chain that the residue belongs to, corresponding relative accessibility surfaces area values in monomer and complex forms, a score for its potential to be a hotspot, result of the prediction: hotspot or not). Moreover, interface file in PDB format and hotspot prediction result file as well as a link for visualization of the interactive 3D model are available on the result page. PRISM then searches for a potential interaction by comparing the surfaces of target proteins to the partners of known template interfaces while accounting for evolutionary conservation.

3.1.2 Target Dataset

Proteins in a target dataset are searched for a potential interaction. Target proteins should have structural data in the PDB. However, artificial proteins can be searched for a potential interaction if its structural data can be added in PDB format. Target set should not contain any DNA or RNA structures, since these kinds of structures are not computed for interaction prediction. The data of query proteins are extracted from the PDB. Multimeric proteins are split into their monomers, and homologous chains are counted only once. The surfaces of the molecules are extracted by using the NACCESS program (described in the next section).

3.1.3 Prediction of Protein-Protein Interaction

PRISM suggests a possible interaction between two target proteins A' and B', if protein A' shares structural similarity with one side of template interface I, which is

extracted from a known interaction between protein A and protein B, and protein B' is structurally similar to the other side of the interface I.

The surfaces of target proteins are extracted using the NACCESS program [241]. NACCESS finds the accessible surface area by rolling a solvent probe on the given molecule. The radius of the solvent probe is chosen as 1.4 Å. NACCESS calculates the relative surface accessibilities (RSAs) of residues, which are the percent accessibility with respect to the accessibility of the residue type X in an extended ALA-X-ALA tripeptide [12]. Residues whose RSA values are greater than 15% are considered as surface residues. 'Nearby' residues are then added to the surface shell as described above, but the threshold value is chosen as 5.0 Å. Structural similarity between target and template interfaces is assessed using MultiProt [242, 243]. MultiProt aligns the target surface with each complementary partner of the representative template interfaces and determines the common geometrical cores between structures. MultiProt's output is the 10 best alignments for sub-structural matching of a target protein surface with a template interface. Target surfaces should geometrically match with 50% of the residues of the template chains if the template chains contain at most 50 residues. This matching threshold is 30% for the larger template chains. In addition, at least one conserved hot spot should be correctly matched between the template interface and the target surface. If a target protein has no hotspot, the algorithm cannot find a potential interaction for this target protein. It is expected that target proteins with any interface size have at least one hotspot. Moreover, at least 5 pairs of matched residues from each side of the template interface should be against each other in order to guarantee the correct matching for the left and right partners. If structures of two proteins are similar to each side of a template interface, in other words, one target protein has a similar structure to one side of a template interface and another target protein has a similarity to the other side of the same template interface; it is expected that parts of the template interface that are similar to structures of each target protein can match with each other. There should be at least 5 pairs of matched residues from each side of the template

interface against each other in order to associate these two target proteins as having a potential interaction.

Target proteins which pass the alignment process and match with the partners of the same template interface are next checked if it is physically possible for them to constitute a complex. If the Ca atom of a residue from one partner is at a distance closer than 3 Å to the Ca atom of a residue from the complementary partner, these two residues are considered as clashing. A threshold of 5 clashes makes the interaction physically impossible.

Finally, FiberDock [244] is used for flexible refinement of the predicted complexes and for calculation of the energy of the interaction. Steric clashes of side-chains due to their orientations are solved via conformational adjustment of the side-chains and the binding energy of the final transformed structures is calculated. In the process of optimizing the predicted protein complex, hydrogen atoms of molecules are also considered and the orientation of the clashing interface residues is adjusted according to the repulsive van der Waals forces. Then, FiberDock calculates binding energies. However, if the solution cannot converge, the global energy cannot be computed. FiberDock ranks the docked solutions by the calculated energies. Hence, FiberDock checks if a potential interaction estimated by PRISM is favorable in terms of global energy.

3.2 Assembly Construction

3.2.1 Constructing Protein Assemblies Based on Binary PRISM Predictions

Modeling protein assemblies is performed based on the binary interaction predictions obtained by PRISM. The construction of the protein assemblies based on PRISM predictions is illustrated in **Figure 3.2**. In the prediction of binary interactions, query proteins are submitted as the target set and the template set can be chosen according to the type of interactions which are being searched. It can include interface

templates related to a certain pathway (like ubiquitination [245] or apoptosis [246]), a certain template interface group (interfaces of obligate or non-obligate interactions [247]) or all template interfaces [248]. If there is no information about the assembly, interactions of the query proteins can be searched using the whole template set. Since we construct assemblies based on binary interactions, we need to set a threshold energy value for binary interactions. **Table 3.1** shows FiberDock energies of the interfaces in the benchmark. One pair of homologous chains is given in the table. The highest energy value is -13.74. Here, we considered results with at most -10 energy units as biologically favorable. The user can set another threshold energy value. We processed non-redundant biologically favorable results in the assembly construction. Only one of the solutions for the same proteins and with the same energy value was selected for further processing to eliminate repetitious computation.

Table 3.1 FiberDock energies of interfaces in the benchmark

Protein 1	Protein 2	Energy unit
2e86A	2e86B	-244.51
1eerA	1eerB	-129.35
1gp2A	1gp2B	-95.30
1gp2B	1gp2G	-242.47
1adoA	1adoB	-75.63
1akjA	1akjB	-96.57
1akjA	1akjD	-37.86
1akjD	1akjE	-93.63
1z0kA	1z0kB	-106.41
1z0kA	1z0kC	-42.56
1z0kC	1z0kD	-114.20
1b0cA	1b0cB	-13.74
1wnrA	1wnrB	-78.38

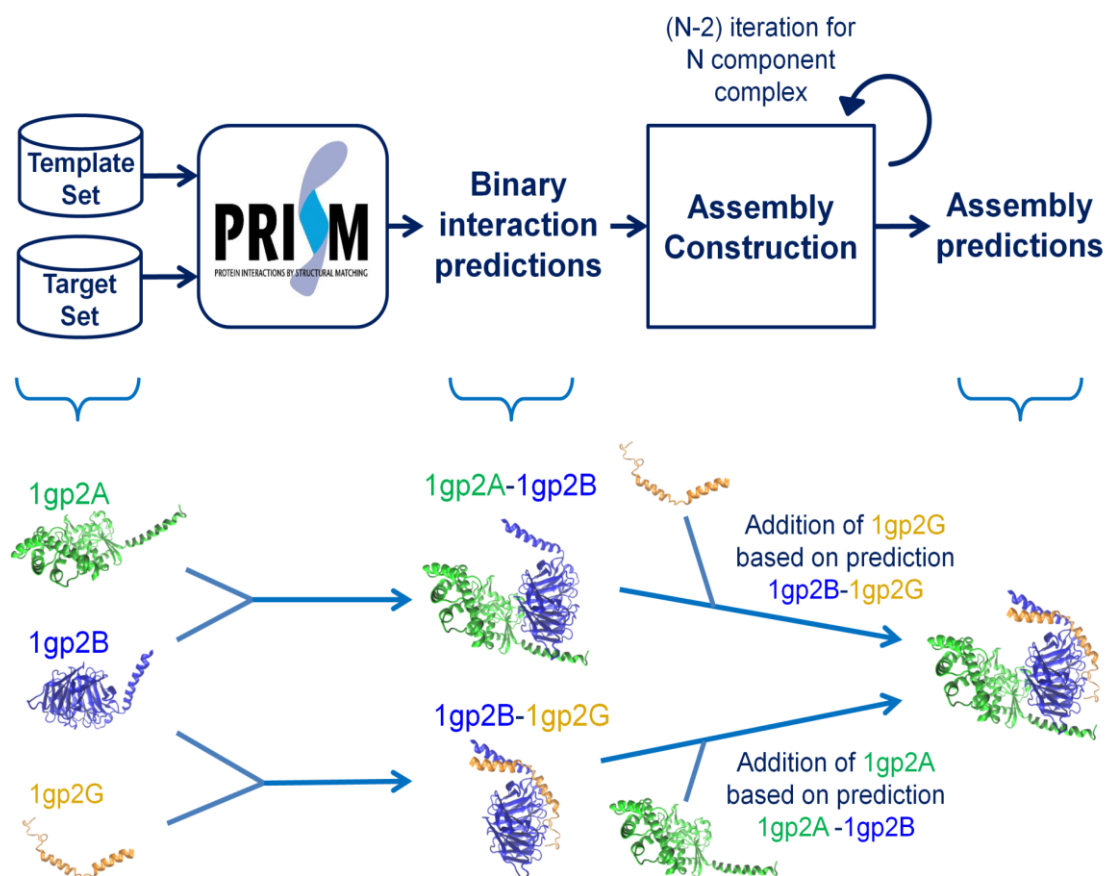


Figure 3.2 Assembly construction based on PRISM binary predictions [236].

First PRISM predicts binary interactions among target proteins. Then, assembly is constructed in an iterative process. $N-2$ iterations are needed to construct an N component assembly. Proteins are added one by one at each iteration based on PRISM binary predictions. There the protein to be added is transformed according to the PRISM prediction and the flexible refinement step of PRISM is run for the addition. The protein is added if the energy calculated is favorable.

Table 3.2 FiberDock energies of interfaces in the benchmark

Protein 1	Protein 2	Energy unit
2e86A	2e86B	-244.51
1eerA	1eerB	-129.35
1gp2A	1gp2B	-95.30
1gp2B	1gp2G	-242.47
1adoA	1adoB	-75.63
1akjA	1akjB	-96.57
1akjA	1akjD	-37.86
1akjD	1akjE	-93.63
1z0kA	1z0kB	-106.41
1z0kA	1z0kC	-42.56
1z0kC	1z0kD	-114.20
1b0cA	1b0cB	-13.74
1wnrA	1wnrB	-78.38

Assembly construction starts with an interacting protein pair, which is a PRISM prediction. In the first iteration, another protein is bound to one of these interacting proteins based on the corresponding predicted PRISM interaction between the protein to be added and the one in the first interaction. First, the protein to be added is transformed next to the sub-assembly structure (as in step 3 of PRISM, **Figure 3.1**) and flexible refinement is done for this candidate interaction (as in step 4 of PRISM, **Figure 3.1**). The candidate protein can be a new protein or one of the proteins in the first pair. The assembly construction process is carried out starting with each non-redundant biologically favorable PRISM interaction and each candidate protein is assessed whether it can be added based on the interactions predicted by PRISM. All possible combinations are considered. Addition of a protein can give as many solutions as the number of non-redundant biologically favorable predictions. To shorten the computation time, some specific interactions, e.g. the ones with lowest energy values,

can be processed. However, there is no guarantee that the assembly will be constructed based on the biologically most favorable predictions.

For an N component assembly, the addition process is performed N-2 times (since it starts with a binary interaction). At each iteration, non-redundant biologically favorable predictions are filtered. The protein is added to the sub-assembly if the interaction has energy below the cut-off value of -10 energy unit. The process is aborted before the assembly reaches the specified number of components if another protein cannot be added to the sub-assembly with a sufficiently low energy value. The solution set can have similar structures. We clustered assembly results based on their structural similarity. Alignment and RMSD calculations are performed using the VMD (Visual Molecular Dynamics) tool [249], version 1.9.1. RMSD threshold is taken as 3.0 Å, considering backbone heavy atoms (N, C α , C, O) of all residues. We chose the structure with the lowest energy score as the representative of the cluster.

3.2.2 Protein Assembly Benchmark and Evaluation of the Predictions

We prepared a benchmark of the protein assembly structures from the PDB. The benchmark includes 8 structures (**Table 3.2**): three 3-chain and three 4-chain assemblies with different number of homologous chains and one assembly of 5-, and 7-chain assemblies. Assemblies were selected in different sizes, ranging between 290 to 1,452 residues in total, and subunits range between 58-363 residues. Asymmetric/symmetric and homo-/hetero-complexes were experimentally obtained in different resolutions, ranging between 1.50-2.90 Å, and cover small proteins and four main SCOP (Structural Classification of Proteins [250]) classes: all alpha proteins, all beta proteins, alpha and beta proteins (a/b), alpha and beta proteins (a+b). Similarity of chain sequences against other chain sequences is between 0.3-32.6% (the average is 11.1% and the median is 9.8%) and identity is between 0.3-18.5% (the average is 6.7% and the median is 5.7%). The benchmark includes symmetrical cyclic structures (PDB IDs: 2e86, 1b0c and 1wnr), which are the most challenging structures for our method, since it is difficult to

add the last protein and complete the cyclic structure based on binary interactions. Unbound forms of the proteins and their structural difference to bound forms are given in **Table 3.3**. Predictions are evaluated based on structural similarity to the PDB structure and the energy score. Structural similarity is measured by the RMSD values calculated for backbone atoms (N, C α , C, O) of all residues. The energy value of an assembly is the summation of energy values calculated for the addition of each protein.

Table 3.3 Structural features of the benchmark proteins

PDB ID	Protein Name	No. Chains	Homologous Chains	No. Residues in chains	Total Residue	Resolution (Å)	Structural type	SCOP class
2e86	Copper-containing nitrite reductase trimer	3	3	3x337	1011	1.50	Symmetric, homocomplex	All beta proteins
1eer	Erythropoietin complexed with its receptor	3	2 +1	2x227 +166	620	1.90	Asymmetric, heterocomplex	All beta proteins /All alpha proteins
1gp2	G protein heterotrimer: Gi alpha 1-beta 1-gamma 2	3	1 +1 +1	353 +340 +71	764	2.30	Asymmetric, heterocomplex	Alpha and beta proteins (a/b) /All beta proteins /All alpha proteins
1ado	Aldolase tetramer	4	4	4x363	1452	1.90	Asymmetric, homocomplex	Alpha and beta proteins (a/b)
1z0k	Rab GTPase complexed with rabenosyn-5	4	2 +2	2x172 +2x69	482	1.92	Asymmetric, heterocomplex	Alpha and beta proteins (a/b)/ All alpha proteins
1akj	MHC Class I glycoprotein HLA-A2 with T-cell coreceptor CD8	4	1 +1 +2	276 +99 +2x120	615	2.65	Asymmetric, heterocomplex	All beta proteins /Alpha and beta proteins (a+b)
1b0c	Pancreatic trypsin inhibitor	5	5	5x58	290	2.80	Symmetric, homocomplex	Small proteins
1wnr	10 kDa chaperonin	7	7	7x94	658	2.90	Symmetric, homocomplex	All beta proteins

Table 3.4 Bound and unbound forms of benchmark proteins

Bound Form	Unbound Form	RMSD (Å)
2e86A	1et5A	0.43
1eerA	1buyA	3.12
1eerBC	1ernA	3.04
1gp2A	1giaA	7.61
1gp2B	1tbgABCD	1.51
1gp2G	1tbgEFGH	5.31
1adoABCD	1aldA	2.37
1akjA	2clrA	1.21
1akjB	2clrB	1.20
1akjDE	1cd8A	0.77
1z0kAC	2bmeA	1.50
1z0kBD	1yzmA	4.63
1b0cABCDE	9ptiA	1.45
1wnrABCDEFG	3nx6A	0.98

3.3 Exploiting Alternative Conformations to Improve Predictions

Here, we exploit alternative conformations of the query proteins to consider protein flexibility in our predictions. Different conformations of target proteins are found in the PDB. First, chains of structures which have the same sequence as the query protein are detected (**Figure 3.3B**). 100% (and then 95%) FASTA sequence homology between the molecules is considered. Then, the molecular structures are compared using MultiProt [251] (**Figure 3.3C**). If MultiProt matches the candidate structure with less than 90% of the query structure and the root-mean-square-deviation (RMSD) value between the matched residues of the two structures is more than 2.0 Å, the candidate structure is considered as a different conformation of the query protein. Different conformations of the target proteins are added to the target set (**Figure 3.3D**); and PRISM is run with the enlarged target set to test if the prediction improves.

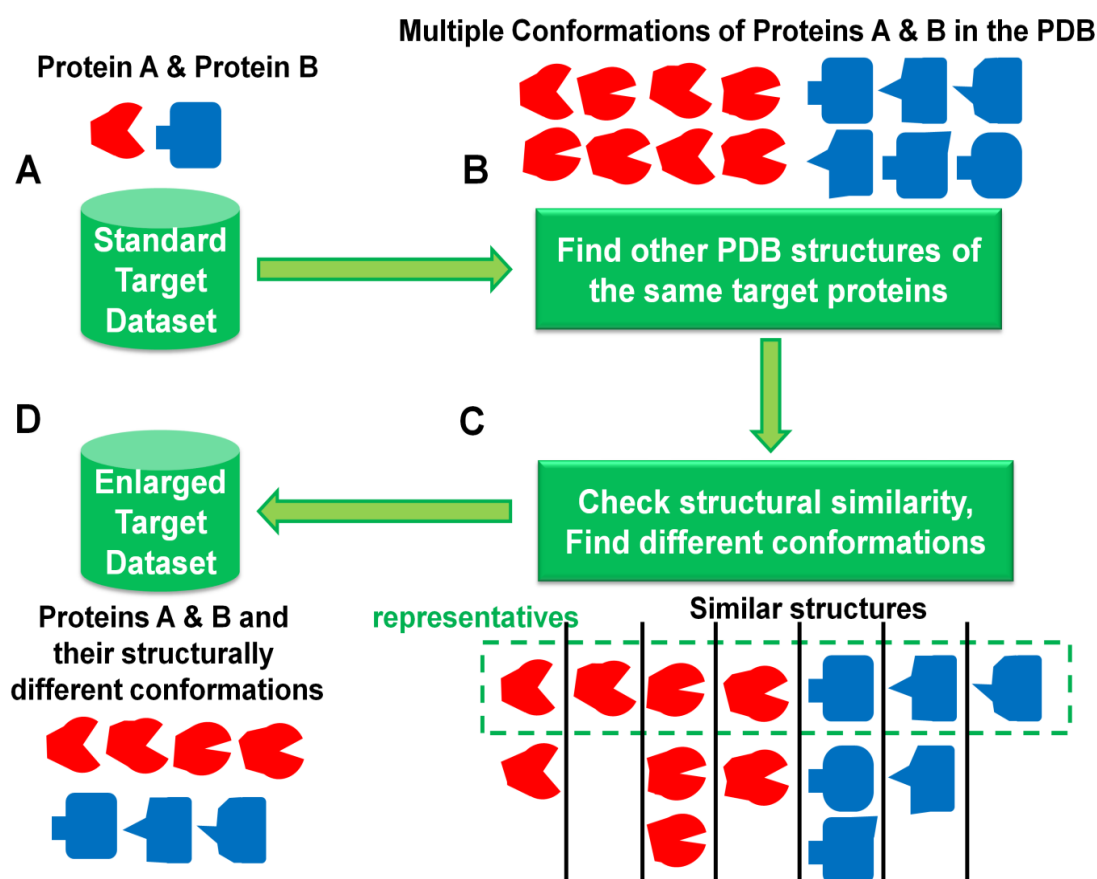


Figure 3.3 Exploiting different conformations to enlarge target set [252].

A) The standard target set includes one conformation of each protein. B) The PDB structures of the standard target proteins are found using sequence homology. C) Conformations are clustered using structural alignment. Structurally similar conformations are given in the same column. Representatives of the clusters are alternative conformations of the standard target proteins. D) The target dataset is enlarged with the different conformations of the standard target proteins.

3.3.1 PRISM Predictions with Different Conformations and the Docking Benchmark Dataset

We tested the method on ‘difficult’ cases of a docking benchmark dataset [253]. The benchmark set provides the bound state of the two given molecules in each case. A list of all interactions studied for 30 ‘difficult’ cases is given in **Table A.1**. In some cases, structures are multimeric and there is more than one binary interaction. In these cases (cases 1, 2, 11, 12 and 18), each binary interaction was considered separately. The structural difference between the bound and unbound forms determines the case-type of the interaction: ‘difficult’ cases of the benchmark dataset include structures with interface root-mean-square distance (iRMSD) larger than 2.2 Å [253]. **Figures 3.4 and 3.5** provide examples of allostery [254] and the conformational changes upon binding. First, PRISM was run with the target set containing only the two molecules from the benchmark to predict their bound state (**Figure 3.6A**). Second, different conformations of the two molecules were found (**Table A.2**; more than one chain name indicates that the molecule is similar to each chain) and added to the target set (**Figure 4.1**). Then, PRISM was run with the enlarged target set (where this is repeated for each case in the docking benchmark) (**Figure 3.6B**). The energy value cut-off, -10 energy unit, was used to determine favorable predictions. **Table 3.4** provides the FiberDock energies of the bound forms for the benchmark dataset. All interfaces except 2hmiCB and 1h1vAG (28 out of 30 interactions) have low energies (below -10 energy unit). Based on our previous studies [173, 245-247, 255], we set the cut-off value as -10. The success of a prediction was evaluated based on IS-score [256], which is a metric to evaluate protein-protein interaction predictions. The IS-score does not consider the equivalence of target and template residues [256], unlike the Critical Assessment of PRedicted Interactions (CAPRI) criteria [257] which are based on RMSD and native contact fractions. These could present problems because PDB structures may lack residues or domains (due to e.g. conformational disorder), and alternative conformations of the query proteins may differ in length. Predicted models are considered as ‘near native’ if the IS-score > 0.17, ‘acceptable’ if the IS-score is between 0.12 and 0.17, and ‘incorrect’ if the IS-score <

0.12. We also presented the native (f_{nat}) and the non-native ($f_{\text{non-nat}}$) contact fractions; f_{nat} : is the number of contacts correctly predicted divided by the number of the native contacts, $f_{\text{non-nat}}$: the number of contacts incorrectly predicted divided by the number of the predicted contacts. IS-score detects substructure similarity and characterizes the similarity of the interfaces of the complex. The alignment exploits iAlign [258] and IS-score compares only matched residues. Zero contact value indicates the model and the native complex have no good match and the model is incorrect. In addition, we present the performance of PRISM as the ratio of successful predictions which were evaluated by IS-score (IS-score > 0.12) to the energetically favorable PRISM predictions (energy < -10 energy unit). The percentage of favorable predictions of PRISM that successfully match the bound forms is given for the benchmark dataset.

Table 3.5 FiberDock energies of the bound forms given in the benchmark dataset

Case	Bound forms	Energy (energy unit)	Case	Bound forms	Energy (energy unit)
1i	1e4kA:C	-23.42	12ii	1fakL:T	-57.60
1ii	1e4kB:C	-27.34	13	1h1vA:G	+3.92
2i	2hmiC:B	-0.64	14	1ibrA:B	-144.40
2ii	2hmiD:B	-37.85	15	1iraY:X	-86.78
3	1f6mA:C	-48.71	16	1jk9A:B	-46.51
4	1fq1A:B	-50.90	17	1jmoA:H	-170.48
5	1pxvA:C	-110.27	18i	1jzdA:C	-21.03
6	1zliA:B	-57.35	18ii	1jzdB:C	-49.11
7	2o3bA:B	-21.60	19	1r8sA:E	-140.42
8	1atnA:D	-107.31	20	1y64A:B	-49.86
9	1bkdR:S	-96.87	21	1zm4A:B	-27.33
10	1de4A:C	-74.58	22	2c0lA:B	-62.85
11i	1eerA:B	-129.35	23	2i9bA:E	-87.64
11ii	1eerA:C	-83.53	24	2idoA:B	-124.11
12i	1fakH:T	-21.83	25	2ot3B:A	-116.03

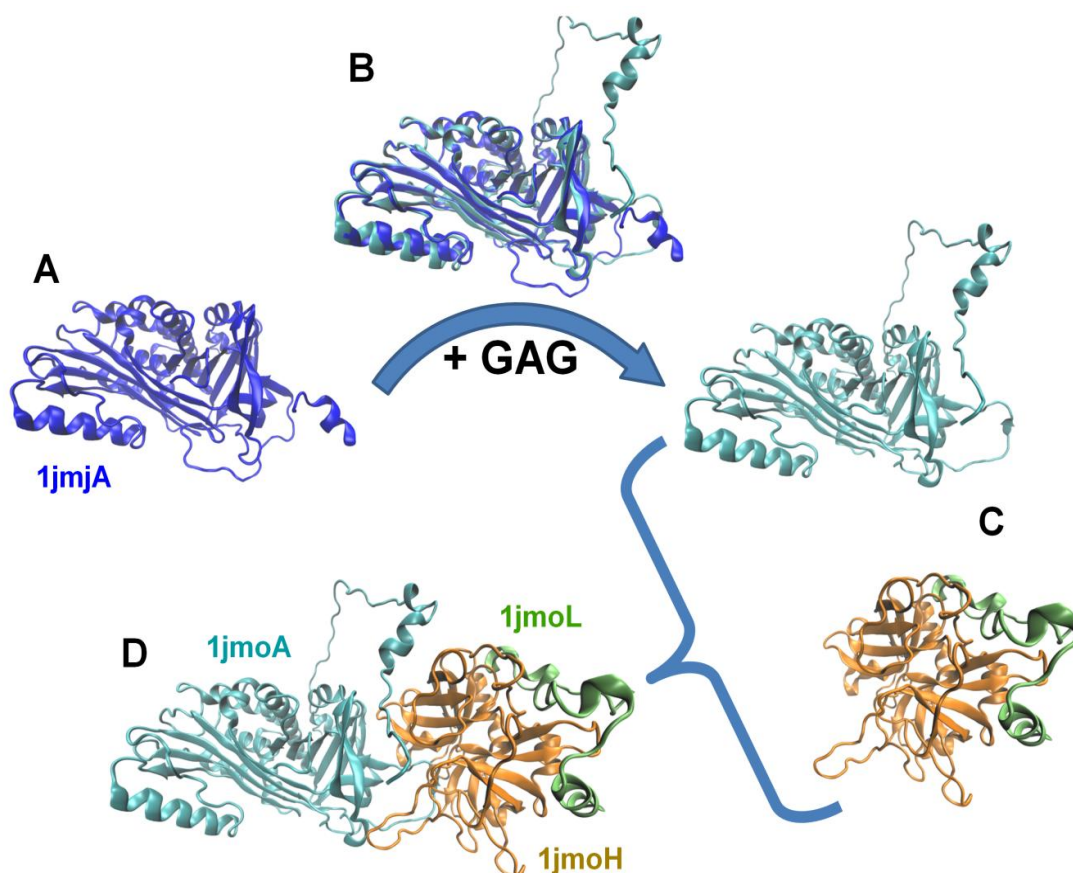


Figure 3.4 An example of allosteric effect on binding – Case 17 [252].

A) Unbound form of heparin cofactor II (blue, 1jmjA). B) Allosteric effect of glycosaminoglycan (GAG) leads to conformational change in heparin cofactor II. Unbound (blue) and bound (cyan) forms of heparin cofactor II are given together. C) Bound forms of heparin cofactor II (cyan: 1jmoA) and thrombin (orange: heavy chain, 1jmoH; green: light chain, 1jmoL) are shown separately. D) Interaction of heparin cofactor II and thrombin.

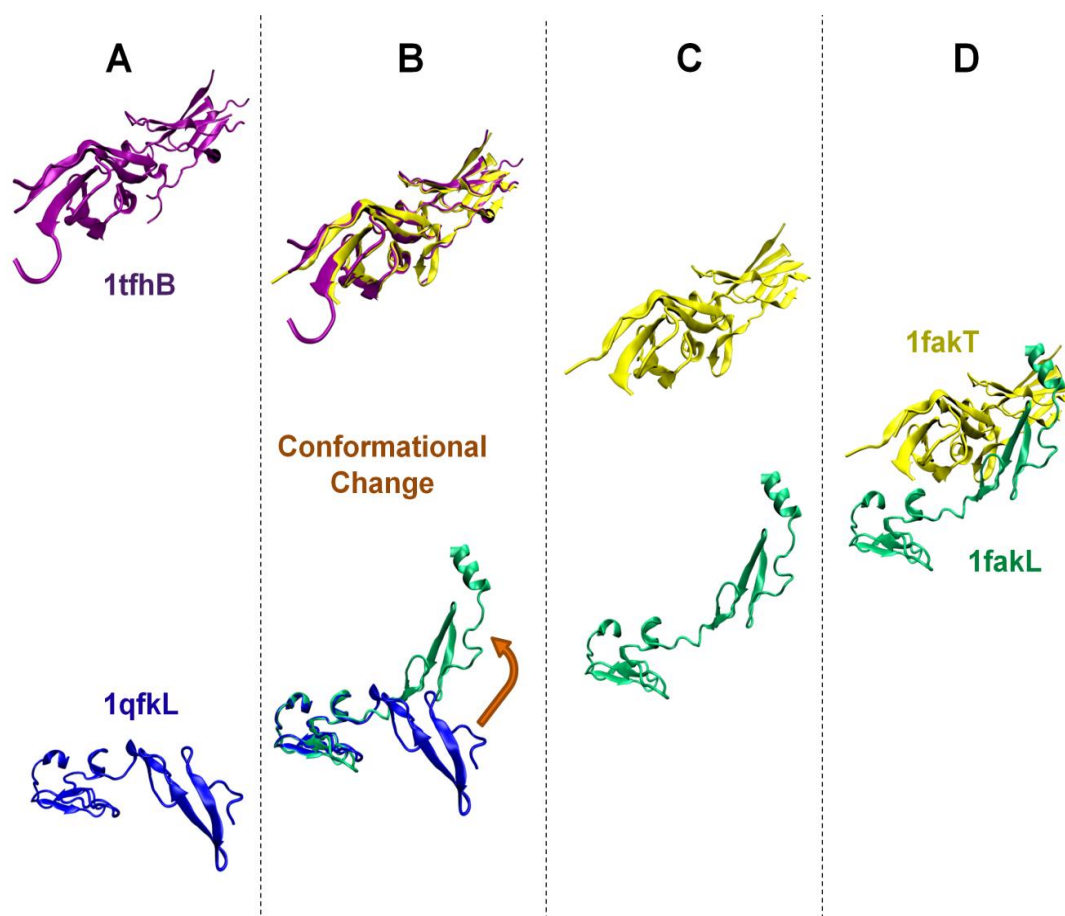


Figure 3.5 An example of conformational change upon binding – Case 12ii [252].

A) Unbound forms of soluble tissue factor (purple, 1tfhB) and blood coagulation factor VIIA (blue: 1qfkL). B) Conformational changes are shown, unbound and bound forms of the soluble tissue factor are given in purple and yellow, and unbound and bound forms of the soluble tissue factor are given in blue and green, respectively. Large conformational change of the blood coagulation factor VIIA is indicated via an arrow. C) Bound forms are shown separately. D) Interaction of soluble tissue factor (1fakT) and blood coagulation factor VIIA (1fakL).

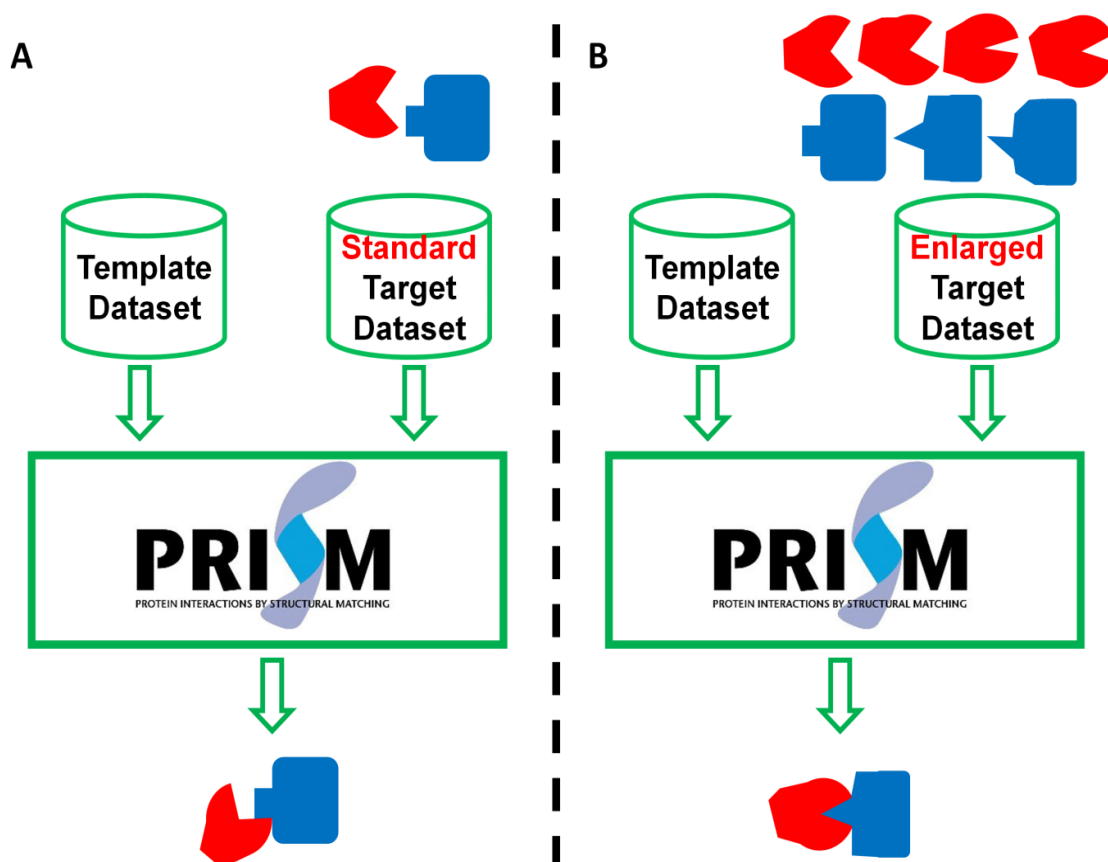


Figure 3.6 PRISM predictions with standard and enlarged target datasets [252].

A) PRISM is run using standard target dataset. No favorable interaction can be found for the given conformations of the proteins. B) PRISM is run using enlarged target dataset. A favorable interaction is found with one conformation of each protein.

3.3.2 Eliminating Bias by Discarding Information on the Bound Form

Bound forms of the proteins exist in the PDB. These forms can appear as alternative structures of the unbound forms or the template interface. For example, in interaction 14, the bound form given in the benchmark dataset is 1ibrAB. 1ibrA and 1ibrC are found as alternative structures of the unbound form 1qg4A, and 1ibrB is for 1f59A (**Table A.2**). Moreover, the template 1ibrAB is used to predict this interaction. To eliminate bias, the bound forms were discarded from the target and template datasets (**Figure 3.7**). First, PDB entries of the bound forms were discarded from the PDB list; then alternative forms were found. These new alternative target structures are given in **Table 3.5**. Second, if a template interface was the interface of a bound form, another interface from the same template cluster was chosen as the representative. If the template cluster had no other member, this template was eliminated. The changes in the template organization are shown in **Table 3.6**. Since the bound forms are given also as the unbound forms in cases 2i (molecule: 2hmiC) and 2ii (molecule: 2hmiD) in the benchmark dataset, these cases were not considered and the remaining 28 interactions were processed under this scenario.

Table 3.6 Alternative structures as the substitute of the bound form

Difficult Case	Query Molecule	Alternative Target Structure
4	1fq1A	-
6	1zliB	1zlhB
9	1bkdS	-
12i	1fakT	1danT
12ii	1fakL	2puqL
14	1ibrB	-
14	1ibrAC	1wa5A
15	1iraY	-
16	1jk9B	1ej8A
17	1jmoA	-
19	1r8sAE	-
20	1y64B	-
22	2c0lB	-
23	2i9bAB	3bt1A
23	2i9bCD	2fd6A
23	2i9bEGH	-
24	2idoB	-
25	2ot3B	-

Table 3.7 Changes in template organization

Difficult Case	Original Template Representative	Substitute Template Representative
3	1f6mAC	-
4	1fq1AB	-
5	1pxvAC	1y4hAC
11i	1eerAB	1cn4BC
11ii	1eerAC	1cn4AC
14	1ibrAB	-
18i	1jzdAC	-
19	1r8sAE	1re0AB
20	1y64AB	-

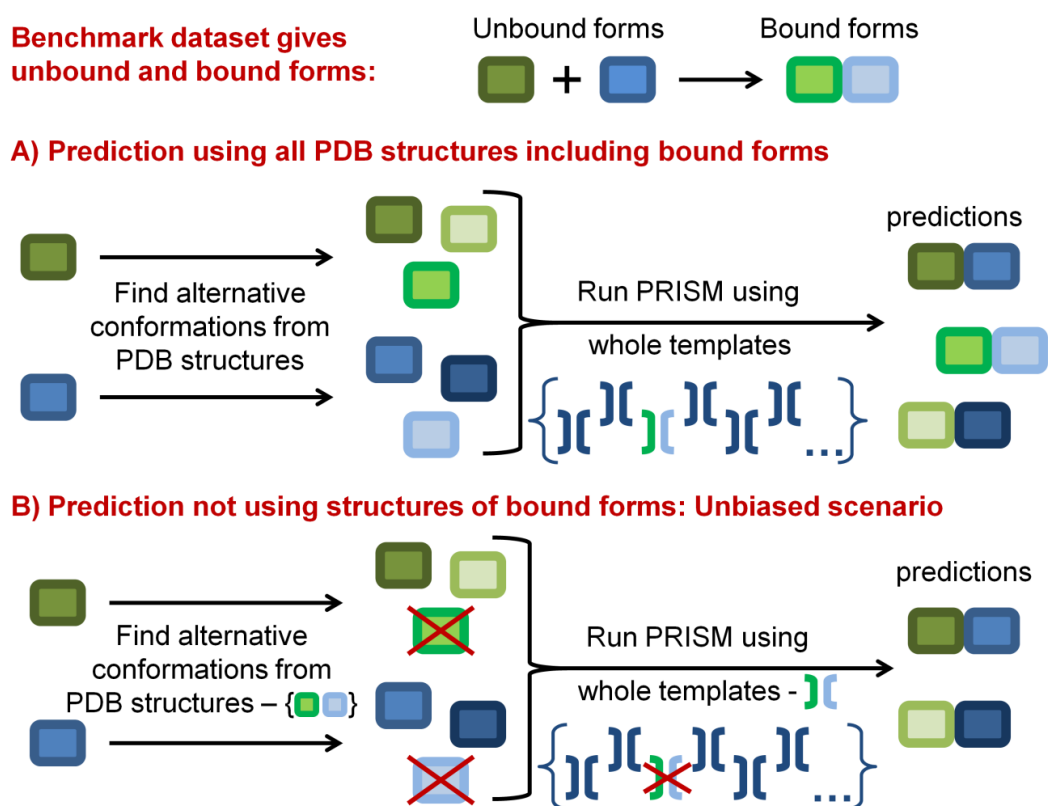


Figure 3.7 Prediction using or not using the structures of bound forms [252].

The benchmark dataset gives the unbound and the bound forms of the molecules. Different color tones represent different conformations of proteins shown in green and blue. The green-cyan template represents the interface of the bound forms. We followed two scenarios in PRISM predictions. A) First, we found alternative conformations from all PDB structures and run PRISM using all templates. B) Second, we eliminated bound forms from the PDB structures and then found alternative conformations. In addition, if a template interface was the interface of a bound form, another interface from the same template cluster was chosen as the representative. If the template cluster had no other member, this template was eliminated.

3.4 Constructing Protein Assemblies Based on EM Density Map

Here, we modified the protein assembly modeling method to exploit EM density maps. **Figure 3.8** represents the new approach; steps 1-4 are the PRISM algorithm, covered in **section 3.1**, and we modeled protein assemblies by adding iteration between steps 3 and 4, covered in **section 3.2**. We modified the algorithm by adding another step, docking structures into the EM density map, as the fifth. In the fifth step, after PRISM predicts binary interactions, assemblies are constructed based on these interactions. Sub-complexes are filtered via docking into the EM density map using the Situs tool [224] (Step-5 in **Figure 3.8**). The user needs to enter the approximate number of residues present in the density map for a proper docking via Situs. Structures that can fit into the EM density map pass this filter. N-unit complexes are constructed in N-2 iterations of the last two steps of PRISM and docking into EM density map (Steps 3-5 in **Figure 3.8**). All sub-complexes are processed in the next iteration, and a query protein is added at each iteration based on the binary interactions predicted by PRISM. The method gives N-unit complexes that fit into the EM density map. Since we aim to construct an assembly structure which the EM map indicates, we consider the interrelationship between the density map and the structure to choose the solution structure; the complex with the lowest Situs score is selected as the best model.

3.4.1 Case Studies and the Benchmark Dataset in Modeling Protein Assemblies

We illustrate the construction of human immunodeficiency virus (HIV) protein interactions as case studies. If available, protein structures are taken from the Protein Data Bank (PDB) [259] and EM density maps are from the EMDataBank (EMDB) [216]. In the first case (HIV gp120-CD4-17b), the structure of the assembly is available in the PDB; in the second case, the structure of the antibody is unknown. We used a homology modeling tool, I-Tasser [260, 261], to create structures from the sequence. Structures were modeled with templates not having 100% sequence similarity to consider as if the actual structures did not exist and eliminate any bias in the structure modeling. We used the best homology model in our predictions (**Table 3.7**, PDB files

of the models are available in the Supplementary Data). Chimera [235] version 1.6.2 is used to calculate the correlation between high and low resolution data. RMSD was calculated with respect to backbone heavy atoms (N, Ca, C, O) of all residues. We evaluated interface predictions with IS-score [256]; IS-scores less than 0.12 indicate ‘incorrect’ predictions, IS-scores between 0.12 and 0.17 indicate ‘acceptable’ predictions and IS-scores higher than 0.17 indicate ‘correct’ predictions.

Table 3.8 Scores obtained and templates used in homology modeling

Protein	PDB ID	C-score	Templates
Antibody m36	-	1.47	3ux9B, 3u2sA, 2kh2B, 1igtB, 1yc8A, 1dlfH
Acetylcholine-binding protein type 1	4aodA	1.46	3sq6A, 2qc1B, 4aq5B, 2qc1A, 4aoeA
Heavy chain of antibody VRC-PG04	3se9H	0.96	1dfbA, 3mlrH, 1y0lH, 3se8A, 4hsH, 2r56H
Light chain of antibody VRC-PG04	3se9L	0.94	3se8L
Lidless Mm-cpn	3iyfA	1.86	1q3qA, 4aolH, 1a6dA, 3j1bA
Conjugal transfer protein TrwB	1e9rA	1.53	4ag5A
Circadian clock protein KaiC	3dvlA	-1.79	2dr3D, 2ztsC, 1tf7A

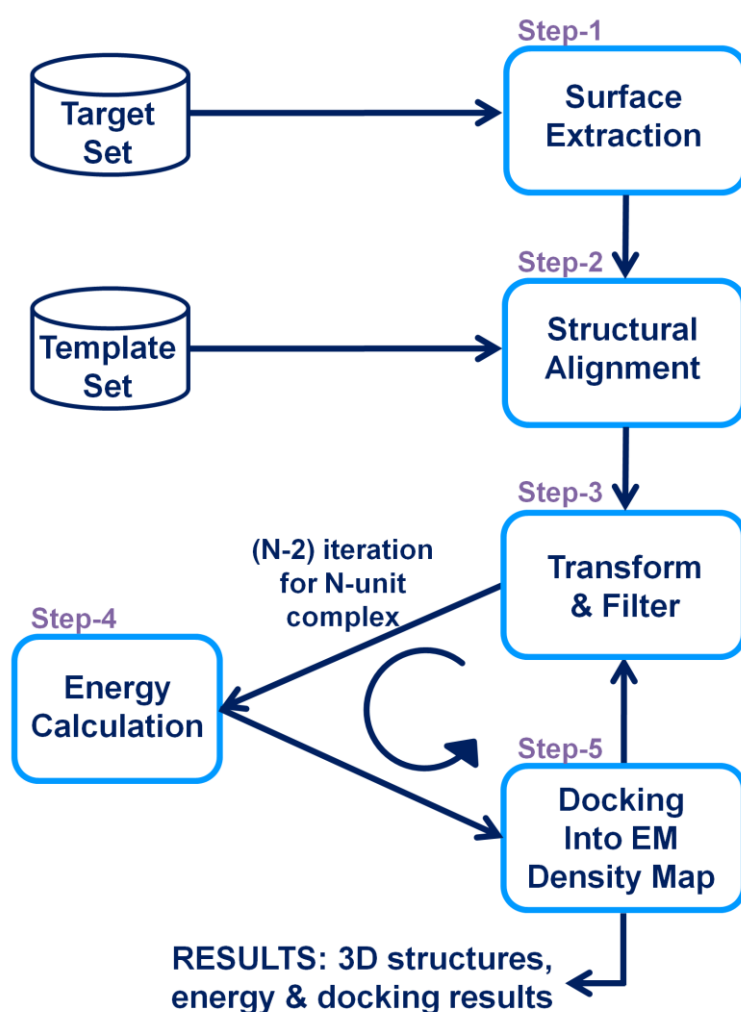


Figure 3.8 Flowchart of assembly construction using EM density map.

Step-1: Surfaces of the target proteins are extracted. Step-2: Surfaces of the target proteins are aligned onto the template interfaces of the known interactions. Step-3: Target proteins are transformed next to each other regarding the template interfaces they match; candidate interactions are filtered regarding hot spot match and clashing. Step-4: Flexible refinement and energy calculation are done. Interactions with favorable energy pass this step. Step-5: Structures are docked into the EM density map and checked if they fit. N-2 iterations are performed through steps 3-5 to construct N-subunit assemblies. At the end, assembly structures together with their energy and docking into EM density map results are given as the result. PRISM predicts binary interaction through steps 1-4. It gives the structures of the interactions and their energy as the result.

In addition, we prepared a benchmark dataset to test our method (**Table 3.8**). The benchmark dataset covers 7 assemblies with 3-8 subunits encompassing 19-521 residues in a single subunit and 286-4168 residues in total in the complex. It includes symmetric/asymmetric and homo-/hetero-complexes. The proteins have 1.4-26.7% similarity and 0.7-15.3% identity with proteins in other complexes. They cover main SCOP (Structural Classification of Proteins [262]) classes: all alpha proteins (all α), all beta proteins (all β), alpha and beta proteins ($\alpha+\beta$), and alpha-beta proteins (α/β), which were found using the server Superfamily 1.75 [263]. For missing structures, the SCOP class was determined through homologs of the proteins. EM density maps have resolution of 6-23 Å. We constructed assemblies starting from unbound (PDB or I-Tasser-modeled) structures of the proteins whose binary interactions can be predicted by PRISM. We also tested those cases using Multifit [154, 188] and Segger [233, 234] through their plug-in versions of Chimera [235]. We fitted structures using Multifit with symmetric or non-symmetric mode and obtained up to 10 models for each case. If the models are not similar to the PDB structure, we selected the first model as the solution. In testing Segger, we first segmented the density map; if we could have desired segments, then we docked the structures using ‘fit to segments’ option of Chimera.

Table 3.9 Benchmark dataset of modeling protein assemblies

Assembly	EMDB ID	EM resolution	PDB ID	Subunit residue #	Assembly residue #	Structure	SCOP class
Saf Pilus- Type A	1494	17 Å	3cre	19, 123, 144	286	Hetero-complex, asymmetric	all β
ParM filament	1980	7.2 Å	4a6j	320	960	Homo-complex, asymmetric	α/β
Biomphalaria glabrata Acetylcholine-binding protein type 1	2055	6 Å	4aod	205	1025	Homo-complex, symmetric	all β
Antibody VRC-PG04 in complex with HIV-1 gp120	2427	23 Å	3se9	208, 228, 353	789	Hetero-complex, asymmetric	all β , $\alpha+\beta$
Lidless Mm-cpn in the Open State	5140	8 Å	3iyf	521	4168	Homo-complex, symmetric	all α , $\alpha+\beta$, α/β
Conjugal transfer protein TrwB	5505	20 Å	1e9r	437	2622	Homo-complex, symmetric	α/β
Circadian Clock Protein KaiC	5672	16 Å	3dvl	519	3114	Homo-complex, symmetric	α/β

Furthermore we tested our method on multimer structures (with more than 2 subunits) in the Multifit benchmark set. We had to eliminate some cases and modify one case in our test. We worked on multimer structures, therefore, did not work on the NADPH binding site on beef liver catalase (7cat), which is a dimer. Nitrite reductase (1nic) is given as a monomer in the PDB; yet, a trimer form can be obtained using crystal symmetry. However, interfaces of the trimer are not large enough regarding our criteria; residues are considered at the interface if their distance is less than summation of their van der Waals distances plus 0.5 Å. Similarly, interfaces of RecA protein (2rec) are also not large enough; dihydropyrimidine dehydrogenase (1gte) is a two dimer structure and there is no interface between dimers (since it is a dimer, we eliminated this entry as well); envelope protein (1urz) is a two trimer structure and we considered it as a trimer structure. We included their case studies methane monooxygenase enzyme (1mty) and GroEL chaperone (1gru) into the benchmark set. EMDB ID 1046 was used for the density map of GroEL chaperone, and the other density maps were created using *Situs*, as described in the *Multifit* paper [188].

Chapter 4

RESULTS OF MODELING PROTEIN ASSEMBLIES

In this chapter, we modeled multi-molecular protein assemblies based on binary interactions predicted by PRISM; the method was covered in **section 3.2**. Protein assembly construction was performed for two scenarios: first, starting from the bound forms of the components; second, starting from the unbound forms of the components. Then, we illustrated our method on a case study of an assembly. After that, we explained the limitations of the method, and finally, presented concluding remarks of the chapter.

4.1 Reconstruction of Assemblies

In the first part, assemblies were decomposed into their components and the components were treated as individual structures. **Figure 3.2** explains how the assembly was constructed based on PRISM predictions. Components (or chains) of the assemblies were submitted as the target set. We reconstructed 3-unit assemblies in the benchmark: PDB IDs: 2e86, 1eer and 1gp2. We calculated the RMSD of predictions compared to the PDB structure and set the energy score of an assembly as the summation of the energy scores calculated at each protein addition. The RMSD vs. energy score is plotted in **Figure 4.1**. The best energy prediction is the best RMSD prediction (e.g. 1eer) or has an RMSD value close to the best RMSD value (e.g. as in 2e86-the best energy and RMSD predictions have 0.53 and 0.51 Å, respectively; and 1gp2 the best energy and RMSD predictions have 0.79 and 0.52 Å, respectively). We considered the energy as the indicator in subsequent steps. We clustered the predictions based on structural similarity. The RMSD values among the predictions were calculated using VMD. We selected the best energy prediction in each cluster as the representative. The best representative predictions (as judged by the similarity to the PDB structures) are given in **Table 4.1** (details in **Table A.3**; the first binary interaction

prediction is given as step 0 and the n^{th} iteration in the assembly construction is step n ; all structurally different predictions are listed in **Table A.4**). The reconstruction of the assemblies suggests that our method works; yet, we need to construct assemblies starting from the unbound forms of proteins to be more realistic.

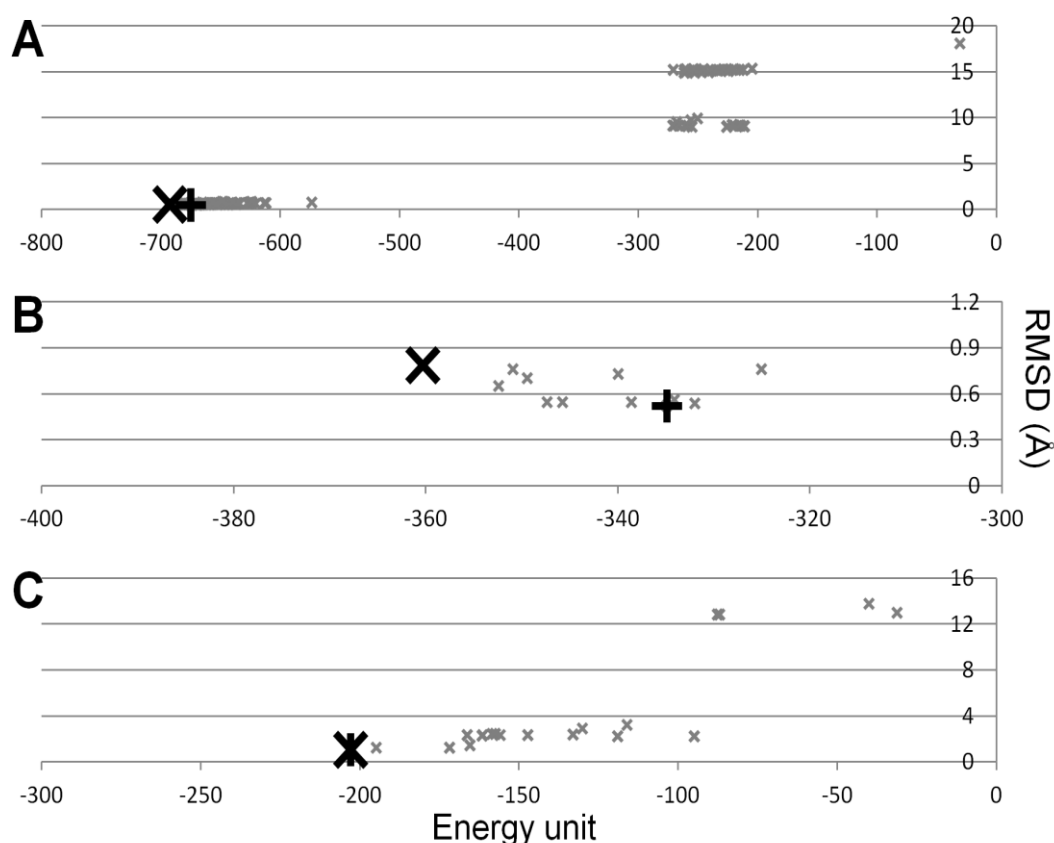


Figure 4.1 Energy score vs. RMSD of predictions [236].

Energy score (energy unit) vs. RMSD (Å) is given for three re-constructed assemblies: A) 2e86 B) 1gp2 C) 1eer. The best energy prediction and the best RMSD prediction are shown with larger black cross (X) and black plus (+) signs, respectively. Others are shown with smaller gray cross. In part A, predictions with energy score lower than -600 have RMSD of 0.5-1.0 Å. The best energy and the best RMSD predictions are the same structure in C.

Table 4.1 Results for construction of protein assemblies starting from their components

PDB structure	No. chain	No. residue	Structurally Different Predictions	Energy unit	RMSD (Å)
2e86	3	1011	9	-692.94	0.53
1eer	3	620	6	-203.02	1.09
1gp2	3	764	1	-360.33	0.79

Predicted PDB structures are given with their PDB IDs, number of chains and number of residues. RMSD is calculated compared to the PDB structures for all backbone atoms and the energy value of an assembly is the summation of energy values calculated for the addition of each protein. These are the best RMSD representatives of the structurally clustered predictions.

4.2 Construction of Assemblies from Unbound Protein Structures

In the second part, PDB structures were assembled starting from the unbound forms of the components. The same procedure was followed as in the first part (**Figure 3.2**). We constructed all assemblies in our benchmark starting from their unbound forms, listed in **Table 3.3**. Homologous chains are given together in the table. Predictions were structurally clustered and the best energy prediction in each cluster was chosen as the representative. The best RMSD representatives, their unbound form sets and the results are listed in **Table 4.2** (details in **Table A.5**; the first binary interaction prediction is given as step 0 and n^{th} iteration in the assembly construction is given as step n ; all structurally different predictions are given in **Table A.6**). If a protein was used twice it is labeled “x2”.

Table 4.2 Results for the construction of protein assembly starting from the unbound forms of the components

PDB structure	No. chain	No. residue	Unbound Forms	Structurally Different Predictions	Energy	RMSD (Å)
2e86	3	1011	1et5A x3	1	-694.70	0.52
1eer	3	620	1buy, 1ernA x2	9	-75.58	5.56
1gp2	3	764	1giaA, 1tbgA, 1tbgE	1	-279.38	3.57
1ado	4	482	1aldA x4	15	-324.79	2.79
1akj	4	615	2clrA, 2clrB, 1cd8A x2	5	-272.70	2.57
1z0k	4	482	2bmeA x2, 1yzmA x2	13	-104.58	3.07
1b0c	5	290	9ptiA x5	23	-109.12	5.15
1wnr	7	658	3nx6A x7	13	-426.80	4.60

Predicted PDB structures are given with their PDB IDs, number of chains and number of residues. If an unbound structure is used twice, it is indicated with “x2”. RMSD is calculated compared to the PDB structures for all backbone atoms and the energy value of an assembly is the summation of energy values calculated for the addition of each protein. These are the best RMSD representatives of the structurally clustered predictions.

Assembly construction of benchmark proteins resulted in up to 23 different structures. Construction of 2e86 and 1gp2 has one representative structure; construction of 1b0c has 23 representative structures. In each case, one of these results matches the PDB structure; the RMSD of the best representative structures ranges between 0.52-5.56 Å, which suggests that our method can construct assemblies starting from their unbound forms. However, we also had results that differ from the PDB structures. These may be different conformations of the assemblies or false positives. For example, one of the results for 1b0c has an RMSD value of 23.29 Å. In the construction of this assembly, the template interface 1aalAB was used four times. 1aalAB is a dimer interface of trypsin inhibitor; yet it is structurally different from interfaces in 1b0c. In 1b0c, trypsin inhibitors interact head-to-head and form a star-shape. However, in 1aal,

the interaction is head-to-tail with a wider angle. Besides structurally different results, we could obtain results structurally close to the PDB structures. However, we need methods to construct assemblies whose structures are unknown. Experimental data of mutations or different techniques like EM, SAXS and FRET can help to select the most appropriate structure as the result, which is covered in **Chapter 6**.

4.3 Modeling Complexes of MEK, KSR, ERK and Raf Proteins

Kinase suppressor of Ras (KSR) is a crucial scaffolding protein which regulates the Raf–MEK–ERK complexes [264-266]. We modeled complexes of MEK, KSR, ERK, and Raf in the mitogen-activated protein kinase (MAPK) pathway. The MAPK cascades constitute a central signaling pathway responsible for the regulation of cellular processes such as growth, proliferation, differentiation, stress response, survival and apoptosis [267-270]. Dysfunction in this pathway can cause a variety of diseases, including cancer [271, 272], diabetes [273], neurodegenerative diseases [274] and developmental diseases [275]. Beyond their roles in putting assemblies together, scaffolding proteins are key components in regulating them via allostery [276].

Barford's group [266] showed that MEK-KSR has a tetramer form (MEK-KSR-KSR-MEK, where MEK-KSR dimers interact through the KSR-KSR interface) which interacts with Raf and they form MEK-KSR-Raf trimer. They crystallized MEK-KSR as a dimer (PDB ID: 2y4i, **Figure 4.2A**) and built the trimer and tetramer structures. We modeled these complexes using MEK1 (PDB ID: 3dy7A), KSR2 (2y4iB, the only PDB ID of KSR including more than one domain) and BRaf (B-Raf, 1uwhA) molecules. We first predicted binary interactions using PRISM and used the whole template set together with the 2y4iBC as a MEK-KSR interaction. Binary interactions with lowest energies were used to construct the complexes. Since the best prediction of KSR-KSR interaction has energy of -9.62 energy unit, we set the cut-off at this value. We selected the energetically most favorable prediction as the model of the assembly. **Figure 4.2B** shows the model of MEK-KSR tetramer. The MEK-KSR interaction in our

model covers 29 out of 35 residues at the interface of the crystal structure MEK-KSR (2y4iBC) and KSR-KSR interaction in our model is centered on residue 718 (Arginine) as indicated [266]. **Figure 4.2C** shows the model of MEK-KSR-Raf trimer. KSR-Raf interaction of our model is centered on Raf residue 509 (Arginine) and KSR residue 718 as indicated [266]. It is also known that ERK interacts with KSR in MEK-KSR-ERK-Raf complex [264, 277, 278]; however, this interaction has not been crystallized yet. PRISM could predict the binary interaction of KSR-ERK, where C-terminus of KSR interacts with ERK, and then we modeled MEK-KSR-ERK-Raf complex using a molecule of each protein: MEK1 (PDB ID: 3dy7A), KSR2 (2y4iB), ERK2 (1wzyA) and BRaf (B-Raf, 1uwhA) (**Figure 4.2D**). Details of the construction of MEK-KSR tetramer, MEK-KSR-Raf trimer and MEK-KSR-ERK-Raf complex are given in **Table A.7** (the first binary interaction prediction is given as step 0 and ‘n’th iteration in the assembly construction is given as step ‘n’; if a structure has more than one chain in the table, they represent the homolog chains).

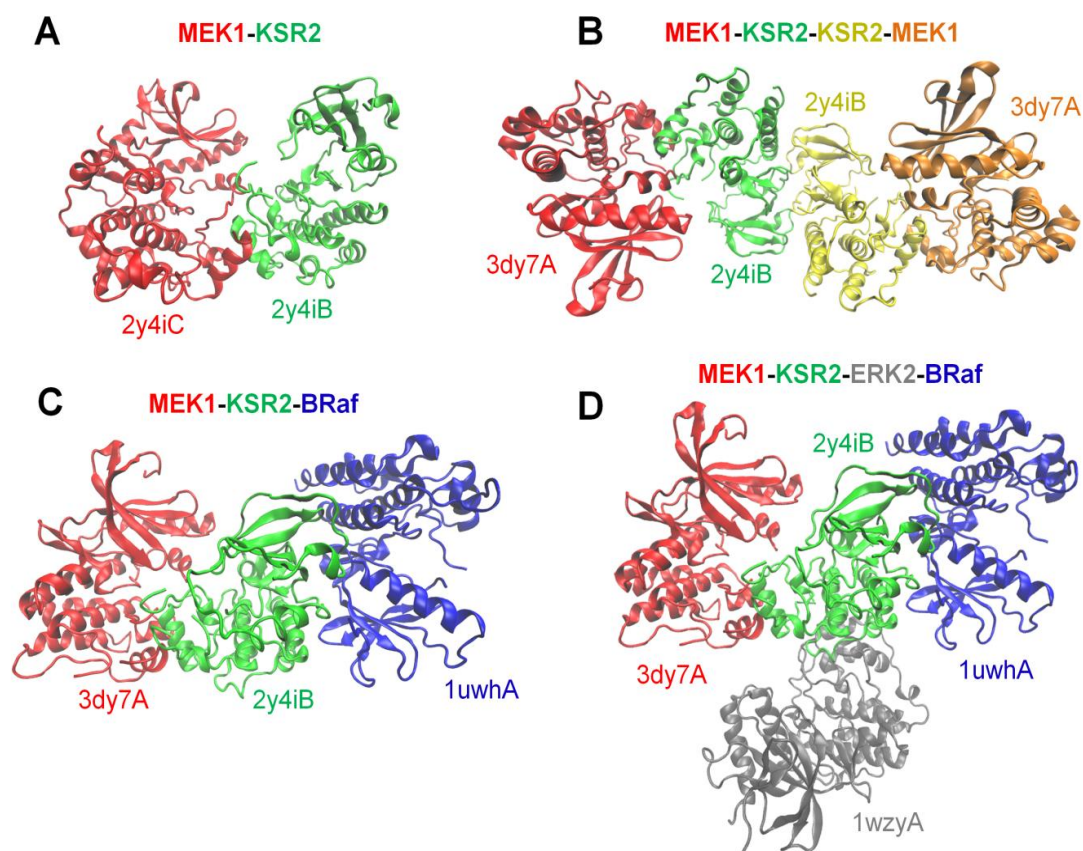


Figure 4.2 Models of complexes of MEK, KSR, ERK and Raf proteins.

A) Crystal structure of MEK-KSR interaction (PDB ID: 2y4i) B) MEK-KSR tetramer. MEK1 proteins (3dy7A) are given in red and orange; KSR2 proteins (2y4iB) are given in green and yellow. C) MEK-KSR-Raf trimer. MEK1 (3dy7A), KSR2 (2y4iB) and BRaf (1uwhA) proteins are given in red, green and blue, respectively. D) MEK-KSR-ERK-Raf complex. MEK1 (3dy7A), KSR2 (2y4iB), ERK2 (1wzyA) and BRaf (1uwhA) proteins are given in red, green, grey and blue, respectively.

4.4 Limitations in Constructing Protein Assemblies

Like all methods, the one followed here has some limitations. First, we constructed protein assemblies based on predictions of binary interactions. There can be cooperativity in the biological complex formation. In addition, there can be many predictions for each binary interaction, and for a protein assembly, considering all combinations can result in a large solution space. Thus, it may not be possible to test each solution. To reduce this space, we select predictions of binary interactions which are energetically more favorable. However, the possibility always exists that we miss biological solutions which while being less favorable for a binary interactions, become more stable upon further monomer additions, as the assembly grows. Such a situation is also encountered hierarchical folding strategies, which are similar to binding [279, 280] and in CombDock [203]. Second, we utilized a knowledge-based method (PRISM) to predict binary interactions. Knowledge-based methods treat structures as rigid bodies, which make the process faster, but disregard protein flexibility. Proteins can allosterically change their conformations upon binding, or following post-translational modifications. Some proteins are also disordered in their unbound states. Protein flexibility increases dramatically the solution space. To partially alleviate this problem, we used alternative conformations in the PDB, which allow us to reach solutions which may be closer to the biological ones. Nonetheless, this is limited by the number of conformations available in the PDB. The fast growth of the PDB is encouraging; however, it can be expected that relevant functional forms are often missing. Third, PRISM has a template set of known interactions and predicts interactions based on these templates. If the appropriate interaction is not assigned as a template, a correct prediction cannot be obtained. Finally, our predictions are based on the availability of PDB structures; however, the structural space is incomplete, and disordered domains are missing, in addition to missing flexible fragments. While it is often possible to model these when handled individually, this is more difficult on a large scale. We are currently taking steps to include high quality modeled structures [248], which may partially alleviate this problem.

4.5 Concluding Remarks

The availability of structures of multimolecular associations, even if the interactions are short-lived, is essential. This is our aim here. Using our method, we first re-constructed 3-unit protein assemblies in the benchmark starting from the assembly components, obtaining low RMSDs. We next tested the modeling of protein assemblies starting from the unbound forms and also obtained good results (0.52 - 5.56 Å). Since we construct assemblies based on binary interactions, the most challenging cases are the symmetric cyclic assemblies. To complete the cyclic structure, the last protein is docked into limited space and interacts with more than one chain, which affects the energy calculation and may cause clashes. We obtained good results also in constructing such symmetric cyclic structures on the cases that we tried. Knowledge-based methods, including PRISM, do not consider the protein flexibility, except in the last refinement step where backbones and side-chains of the structures can be slightly reoriented.

Modeling of multi-molecular assemblies from monomeric structures of their components is computationally challenging with a broad solution space. To reduce this space, we select the energetically more favorable predictions of binary interactions. However, the possibility always exists that we miss biological solutions which while being less favorable for binary interactions, become more stable as the assembly grows. Such a situation is also encountered in hierarchical folding strategies [279, 280] as also discussed for CombDock [203], which also suffers from this handicap. One major difference of our approach from CombDock is that we do not consider all possible pairwise interactions but instead use template-based pairwise interactions which is expected to be much less than $N(N-1)/2$. Another problem is choosing the right prediction. Although the energy value can be an indicator, similar to protein folding, this is not always the case. Experimental techniques like Cryo-EM, FRET and SAXS, which provide low resolution data of assemblies, can be used to help select the solutions, which we will cover in **Chapter 6**.

Other caveats relate to PDB structures which do not always represent the entire protein, or the functional state. In addition flexible fragments and disordered domains are missing. While it is often possible to model these when handled individually, it is more difficult on a large scale. We are currently including high quality modeled structures [248], which may partially alleviate this problem. Further, the coverage of the interface architectures by a template set which is based on the PDB, affects PRISM predictions and hence the assembly construction. Nonetheless, despite these limitations, we were able to successfully construct protein assemblies thanks to the current PDB richness.

Chapter 5

RESULTS OF EXPLOITING CONFORMATIONAL ENSEMBLES

In the previous chapter, the method of modeling protein assemblies is presented. More accurate models can be obtained with the improvement in PPI predictions and the assembly modeling approach. In this chapter, we explain how exploiting conformational ensembles improves our predictions in modeling PPIs and protein assemblies.

In **Chapter 2**, we review that computational approaches can be classified into two categories: ‘blind’ docking and knowledge-based methods, and stated that knowledge-based methods are appropriate techniques to construct interactome-scale structural networks. Even though the PDB covers a limited number of protein-protein interactions, a systematic large scale study [281] showed that templates are available to model complexes of structurally characterized proteins. However, on the down side, knowledge-based approaches typically consider the structures as rigid bodies, even though proteins are flexible [282-285] and their preferred conformational states are expected to change with the environment. Neighboring molecules, atoms, or ions can redistribute their ensembles [286, 287]; this redistribution is dynamic [288-292] and is reflected in the observed protein structures. Binding, post-translational modifications, changes in ligand concentration, pH and ionic strength of the medium will affect the conformational distribution. The ensembles confer on the protein the ability to function [290-292]. The bound forms of the proteins can be obtained from the unbound forms in *in silico* experiments [293]; however, protein flexibility and conformational changes upon binding are mostly ignored in large scale knowledge-based methods. This limits

the capability and success in prediction, and therefore, the construction of the structural network.

In principle, different conformations of proteins can be considered by knowledge-based methods, and this can be expected to improve the predictions. The PDB offers many protein structures including different conformations of the same protein. These conformations may include bound, unbound or any alternative forms (e.g. following allosteric post-translational modifications, or bound to different ligands) of the proteins which can be utilized. The collection of all available structures of a protein provides a subset of the repertoire of its conformations under different conditions, and can constitute the input rather than a single structure. These conformations would help to more reliably figure out whether the proteins interact, and how they interact, and identifying binary protein interactions is the first step in the construction of the structural network. We tested a motif-based protein interaction prediction tool, PRISM [88, 173, 174], on a docking benchmark dataset [253], to see if providing different conformations of proteins would indeed increase PRISM's capability to detect interactions. Then, we used this approach in modeling protein assemblies.

5.1 Exploiting Conformational Ensembles in PPIs

5.1.1 PPI Prediction Results

We worked on difficult cases of ZDOCK benchmark dataset [294], where bound and unbound forms of the proteins are given. We obtained two sets of predictions. First, PRISM was run with a 'standard' target set of given unbound structures (**Figure 3.6A**); one conformation for each molecule. Then, the target set was enlarged with the different conformations of the unbound forms of each protein available in the PDB (**Table A.2, Figure 3.6B**). The first target set had two conformations (a size of 2) in all cases; and on average the enlarged target sets had a size of 8.87, and the median values was 6.50, which did not affect considerably the computation process. Using standard target sets, 9 out of 30 'difficult' interactions were predicted as energetically favorable

(-10 energy unit cut-off value; **Table 5.1**). In the second set, when the target sets were enlarged with the different conformations of the given unbound structures, 7 of these 9 interactions were better predicted, with lower global energy values (interactions 2i, 11i, 11ii, 12i, 16, 18ii and 19; **Table 5.1**). Energetically favorable predictions (an increase from 9 to 24) were obtained for additional interactions when enlarged target sets were used (**Table 5.1**). We note that energetically favorable models do not necessarily correspond to the native complex structures of the benchmark. Energetically favorable interactions together with their templates are given in **Tables 5.2 and 5.3**. If the chain is specified, the interacting molecule is the monomer; otherwise, it is the whole structure. If more than one chain is listed, the chains are structurally similar.

Table 5.1 Energy values of PRISM results obtained using standard template set

Case	Energy Results (energy unit)		Size of enlarged target set	Case	Energy Results (energy unit)		Size of enlarged target set
	Standard Target Set	Enlarged Target Set			Standard Target Set	Enlarged Target Set	
1i	-21.85	-21.85	8	12ii	none	-102.61	46
1ii	+4.45	-6.66	9	13	none	-27.80	13
2i	-16.41	-21.69	8	14	+25.97	-144.40	14
2ii	+25.58	-35.96	8	15	none	-63.74	5
3	+22.66	+22.66	2	16	-27.01	-49.32	3
4	+21.64	-68.75	19	17	+5.50	-21.97	6
5	+461.37	-74.40	3	18i	-20.13	-20.13	4
6	none	+8.60	4	18ii	-43.50	-50.84	4
7	+1.35	-8.22	4	19	-13.76	-140.42	7
8	none	-46.33	10	20	-2.90	-58.42	11
9	+31.27	-114.68	11	21	-1.98	-14.77	12
10	0.00	0.00	2	22	+2.29	-40.35	6
11i	-32.28	-63.43	4	23	+4.46	-42.70	10
11ii	-37.70	-64.65	4	24	none	-43.80	3
12i	-31.70	-57.47	21	25	+2.90	+2.90	5

Table 5.2 Favorable PRISM predictions of difficult cases using standard template set and standard target set

Difficult Case	Interacting Molecules of the Standard Target Set		Template Interface	Energy (energy unit)
1i	3aveAB	1fnl	1t83BC	-21.85
2i	1s6pB	2hmiC	1e8oCD	-16.41
11i	1buy	1ernAB	1eerAB	-32.28
11ii	1buy	1ernAB	1eerAC	-37.70
12i	1qfk	1tfhB	1o5dHT	-31.70
16	1qupAB	2jcw	1pu0CD	-27.01
18i	1jzoA	1jpe	1jzdAC	-20.13
18ii	1jpe	1jzoB	1vrsBE	-43.50
19	1hurAB	1r8m	1r8sAE	-13.76

IS-scores and contact fractions of the predictions obtained using alternative conformations are provided in **Table 5.4** (**Table 5.5** shows the results of predictions obtained using the standard target set). As to IS-scores, among the predictions obtained using standard target sets, 7 cases were ‘near native’, 1 case was ‘acceptable’ and 1 case was ‘incorrect’ (**Table S16**), and among predictions obtained using enlarged target sets, 18 cases were ‘near native’ and 6 cases were ‘incorrect’ (**Table 5.1**). When we eliminated incorrect predictions, we had 8 successful predictions with the standard target set and 18 with the enlarged set. In addition, if a different prediction was obtained when the enlarged target set was used, the prediction has a higher IS-score.

Table 5.3 Favorable PRISM predictions of difficult cases using standard template set and enlarged target set

Difficult Case	Interacting Molecules of the Enlarged Target Set		Template Interface	Energy (energy unit)
1i	3aveAB	1fnl	1t83BC	-21.85
2i	1dloA	2hmiC	1qd1AB	-21.69
2ii	2hmiD	3kjbB	1at3AB	-63.70
4	1fpzD	2uueAC	1fq1AB	-68.75
5	1y4hAB	1nyc	1pxvAC	-74.40
8	2y83OPQRT	3dni	2a40DE	-46.33
9	1nvwR	1xd2C	1nvuRS	-114.68
11i	1cn4C	1ebaAB	1eerAB	-63.43
11ii	1cn4C	1ebaAB	1eerAC	-64.65
12i	1jbu	1w0yT	1o5dHT	-57.47
12ii	2zp0L	2zwlT	1wtgLT	-102.61
13	1rgiG	2yjfB	1c0gSA	-27.80
14	1ibrA	1ibrBD	1ibrAB	-144.40
15	1ilr	1iraY	1itbAB	-63.74
16	1jk9BD	2jcw	1pu0CD	-49.32
17	3k65B	1jnjA	2b8nAB	-21.97
18i	1jzoA	1jpe	1jzdAC	-20.13
18ii	3pfu	1jzoB	1vrsBE	-50.84
19	1r8sA	1r8sE	1r8sAE	-140.42
20	2yjfE	1y64B	1y64AB	-58.42
21	1zm3C	1aer	1zm9EF	-14.77
22	1fchB	1c44	1qjcAB	-40.35
23	1ywhG	3bt1A	1pbiAB	-42.70
24	2idoBD	1j54	2b8nAB	-43.80

Table 5.4 Energies, IS-scores and contact fractions of PRISM predictions obtained using standard template set and enlarged target set

Case	Energy (energy unit)	IS- score	Rating	Real Contacts	Model Contacts	Common Contacts	f_{nat}	$f_{\text{non-nat}}$
1i	-21.85	0.2366	Near native	17	27	16	0.94	0.41
2i	-21.69	0.0648	Incorrect	0	39	0	0.00	1.00
2ii	-63.70	0.0751	Incorrect	0	72	0	0.00	1.00
4	-68.75	0.4041	Near native	21	38	18	0.86	0.53
5	-74.40	0.6640	Near native	65	74	55	0.85	0.26
8	-46.33	0.3775	Near native	38	46	21	0.55	0.54
9	-114.68	0.6956	Near native	86	113	80	0.93	0.29
11i	-63.43	0.6691	Near native	47	57	41	0.87	0.28
11ii	-64.65	0.6036	Near native	38	46	34	0.89	0.26
12i	-57.47	0.4789	Near native	15	29	15	1.00	0.48
12ii	-102.61	0.5916	Near native	53	76	47	0.89	0.38
13	-106.01	0.5890	Near native	46	58	41	0.89	0.29
14	-144.40	0.7976	Near native	79	85	68	0.86	0.20
15	-63.74	0.6283	Near native	53	71	48	0.91	0.32
16	-49.32	0.6967	Near native	62	78	56	0.90	0.28
17	-21.97	0.0659	Incorrect	0	32	0	0.00	1.00
18i	-20.13	0.4864	Near native	19	23	18	0.95	0.22
18ii	-50.84	0.4942	Near native	35	38	26	0.74	0.32
19	-140.42	0.9363	Near native	93	100	93	1.00	0.07
20	-58.42	0.6201	Near native	38	44	29	0.76	0.34
21	-14.77	0.1770	Near native	10	21	6	0.60	0.71
22	-40.35	0.0696	Incorrect	0	50	0	0.00	1.00
23	-42.70	0.0679	Incorrect	0	36	0	0.00	1.00
24	-43.80	0.0700	Incorrect	9	43	0	0.00	1.00

Table 5.5 Energies, IS-scores and contact fractions of PRISM predictions obtained using standard template set and standard target set

Difficult Case	Energy (energy unit)	IS-score	Rating	Real Contacts	Model Contacts	Common Contacts	f_{nat}	$f_{\text{non-nat}}$
1i	-21.85	0.2366	Near native	17	27	16	0.94	0.41
2i	-16.41	0.0753	Incorrect	0	70	0	0.00	1.00
11i	-32.28	0.2880	Near native	34	40	22	0.65	0.45
11ii	-37.70	0.3063	Near native	28	36	18	0.64	0.50
12i	-31.70	0.3690	Near native	19	16	15	0.79	0.06
16	-27.01	0.6793	Near native	45	55	42	0.93	0.24
18i	-20.13	0.4864	Near native	19	23	18	0.95	0.22
18ii	-43.50	0.3330	Near native	35	44	25	0.71	0.43
19	-13.76	0.1367	Acceptable	28	33	6	0.21	0.82

PRISM was also run without using the data of the bound forms in the benchmark dataset to eliminate any bias. The results of this scenario (see **section 3.3.2**, “Eliminating bias by discarding information of the bound form”) are given in **Table 5.6**. Enlarging the target set by including alternative conformations increased the number of energetically favorable predictions from 6 to 21 out of 28 interactions (5 of these 6 were predicted better, with lower global energy), with the cut-off energy set at -10 energy unit. Favorable interactions together with their templates and global energy values are listed in **Tables 5.7 and 5.8**. IS-scores and contact fractions are provided in **Tables 5.9 and 5.10**. According to the IS-scores classifications, all 6 predictions obtained using standard target sets were ‘near native’, and among predictions using enlarged target sets, 14 cases were ‘near native’ and 7 cases were ‘incorrect’. When we

eliminated incorrect predictions, there were 6 successful predictions with the standard target set and 14 with the enlarged set. In addition, higher IS-scores were obtained (except for case 16 with better energy but lower IS-score) for predictions with the enlarged target set.

Table 5.6 Energy values of PRISM results obtained using rearranged templates and target sets

Case	Energy Results (energy unit)		Size of enlarged target set	Case	Energy Results (energy unit)		Size of enlarged target set
	Standard Target Set	Enlarged Target Set			Standard Target Set	Enlarged Target Set	
1i	-21.85	-21.85	8	13	none	-27.80	13
1ii	+4.45	-6.66	9	14	+25.97	-36.79	14
3	+22.66	+22.66	2	15	none	-27.67	4
4	+75.32	-63.20	18	16	-27.01	-45.75	3
5	none	-84.17	3	17	+0.50	-21.97	5
6	+8.6	+8.6	4	18i	-26.48	-54.15	4
7	+1.35	+1.35	3	18ii	-43.50	-50.84	4
8	none	-46.33	10	19	none	-13.76	6
9	+31.27	-114.68	10	20	-2.90	-52.20	10
10	0.00	0.00	2	21	-1.98	-14.77	12
11i	-32.28	-49.87	4	22	+2.29	-40.35	5
11ii	-9.53	-57.64	4	23	+4.46	-42.70	11
12i	-31.70	-57.47	21	24	none	none	2
12ii	none	-102.61	46	25	+2.90	+2.90	4

Table 5.7 Favorable predictions using rearranged templates and standard target set

Difficult Case	Interacting Molecules of the Standard Target Set		Template Interface	Energy (energy unit)
1i	3aveAB	1fnl	1t83BC	-21.85
11i	1ernAB	1buy	1cn4BC	-27.44
12i	1qfk	1tfhB	1o5dHT	-31.70
16	1qupAB	2jcw	1pu0CD	-27.01
18i	1jpe	1jzoA	1vrsBE	-26.48
18ii	1jpe	1jzoB	1vrsBE	-43.50

Table 5.8 Favorable predictions using rearranged templates and enlarged target set

Difficult Case	Interacting Molecules of the Enlarged Target Set		Template Interface	Energy (energy unit)
1i	3aveAB	1fnl	1t83BC	-21.85
4	1fpzF	1pye	1nh0AB	-63.20
5	1y4hAB	1nycAB	1y4hAC	-84.17
8	2y83OPQR	3dni	2a40DE	-46.33
9	1nvwR	1xd2C	1nvuRS	-114.68
11i	1ebaAB	1buy	1cn4BC	-49.87
11ii	1ebaAB	1cn4C	1cn4AC	-57.64
12i	1jbu	1w0yT	1o5dHT	-57.47
12ii	2zp0L	2zwlT	1wtgLT	-102.61
13	1rgiG	2yjfB	1c0gSA	-27.80
14	1f59AB	1i2mC	1l9wCD	-36.79
15	1ilr	1iraY	1itbAB	-27.67
16	1ej8	2jcw	1pu0CD	-45.75
17	3k65B	1jnjA	2b8nAB	-21.97
18i	3pfu	1jzoA	1vrsBE	-54.15
18ii	3pfu	1jzo	1vrsBE	-50.84
19	1r8qAB	1r8m	1re0AB	-90.07
20	2yjfE	1ux5	2btfAP	-52.20
21	1zm3C	1aer	1zm9EF	-14.77
22	1fchB	1c44	1qjcAB	-40.35
23	1ywhG	3bt1A	1pbiAB	-42.70

Table 5.9 Energies, IS-scores and contact fractions of PRISM predictions obtained using rearranged templates and standard target set

Difficult Case	Energy (energy unit)	IS-score	Rating	Real Contacts	Model Contacts	Common Contacts	f_{nat}	$f_{\text{non-nat}}$
1i	-21.85	0.2366	Near native	17	27	16	0.94	0.41
11i	-27.44	0.2664	Near native	26	28	17	0.65	0.39
12i	-31.70	0.3690	Near native	19	16	15	0.79	0.06
16	-27.01	0.6793	Near native	45	55	42	0.93	0.24
18i	-26.48	0.4864	Near native	19	23	18	0.95	0.22
18ii	-43.50	0.3330	Near native	35	44	25	0.71	0.43

Table 5.10 Energies, IS-scores and contact fractions of PRISM predictions obtained using rearranged templates and enlarged target set

Difficult Case	Energy (energy unit)	IS-score	Rating	Real Contacts	Model Contacts	Common Contacts	f_{nat}	f_{non-nat}
1i	-21.85	0.2366	Near native	17	27	16	0.94	0.41
4	-63.20	0.0592	Incorrect	21	31	0	0.00	1.00
5	-84.17	0.6186	Near native	51	56	42	0.82	0.25
8	-46.33	0.3775	Near native	38	46	21	0.55	0.54
9	-114.68	0.6956	Near native	86	113	80	0.93	0.29
11i	-49.87	0.3529	Near native	34	48	23	0.68	0.52
11ii	-57.64	0.6495	Near native	38	48	36	0.95	0.25
12i	-57.47	0.4789	Near native	15	29	15	1.00	0.48
12ii	-102.61	0.5916	Near native	53	76	47	0.89	0.38
13	-106.01	0.5890	Near native	46	58	41	0.89	0.29
14	-36.79	0.0696	Incorrect	0	42	0	0.00	1.00
15	-27.67	0.5574	Near native	56	78	48	0.86	0.38
16	-45.75	0.5497	Near native	39	46	32	0.82	0.30
17	-21.97	0.0659	Incorrect	0	32	0	0.00	1.00
18i	-54.15	0.0669	Incorrect	0	49	0	0.00	1.00
18ii	-50.84	0.4942	Near native	35	38	26	0.74	0.32
19	-90.07	0.5517	Near native	56	57	41	0.73	0.28
20	-52.20	0.0574	Incorrect	0	22	0	0.00	1.00
21	-14.77	0.1770	Near native	10	21	6	0.60	0.71
22	-40.35	0.0696	Incorrect	0	50	0	0.00	1.00
23	-42.70	0.0679	Incorrect	0	36	0	0.00	1.00

5.1.2 95% Sequence Homology was Used to Enlarge the Target Set More

Up to here, we considered 100% sequence homology for alternative conformations from the PDB. Analysis of the sequence homology between unbound and bound forms given in the benchmark dataset, illustrates that some are not 100% similar (**Table A.8**). However, reducing sequence homology much below 100% can lead to picking different proteins. We reduced this value to 95% for cases that we could not obtain successful predictions (for 16 interactions: 1ii, 2i, 2ii, 3, 4, 6, 7, 10, 14, 17, 18i, 20, 22, 23, 24 and 25 – for cases 4, 14, 18i and 20, unsuccessful results were obtained when bound data was eliminated from the template and target sets; therefore, we added these to this list) to enlarge the target set. We used the same criteria to identify structurally different conformations as described in Materials and Methods, section “Enlarging the Target Dataset Using Different Conformations”. These additional alternative conformations are given in **Table A.9** (more than one chain name indicates that the molecule is similar to each chain). No additional conformations were detected for the unbound forms of cases 7, 10, 22, 23 and 24. Although additional favorable predictions were obtained for cases 1ii, 2i, 2ii, 3, 4, 17, 20 and 25 by using these conformations (**Table 5.11**), only successful results were obtained for cases 1ii, 3 and 14 (IS-score > 0.12). Energy values, IS-scores and contact fractions are given in **Table 5.12**.

Table 5.11 Favorable PRISM predictions of difficult cases obtained using additional conformations found by 95% sequence homology

Difficult Case	Interacting Molecules		Template Interface	Energy (energy unit)
1ii	1t89A	3ay4C	1t83AC	-19.76
2i	3nbpB	2hmiC	1hjrBD	-30.08
2ii	2hmiD	1rt2B	1at3AB	-31.85
3	1f6mABEF	2tir	1f6mAC	-37.07
4	2iw9C	1fqzE	1n4aAB	-41.16
14	3gjfF	1f59AB	2bkuCD	-19.31
17	1jmO	1mkxK	1h1bAB	-41.90
18i	1vrsB	1jzoA	1jd1AB	-63.00
20	3lueABCD	1ux5	2a40DE	-23.95
25	1yzuAB	2ot3A	1ihoAB	-13.21

Table 5.12 Energies, IS-scores and contact fractions of PRISM predictions obtained using additional conformations found by 95% sequence homology

Difficult case	Energy (energy unit)	IS-score	Rating	Real Contacts	Model Contacts	Common Contacts	f_{nat}	$f_{\text{non-nat}}$
1ii	-19.76	0.4319	Near native	24	38	22	0.92	0.42
2i	-30.08	0.0642	Incorrect	0	35	0	0.00	1.00
2ii	-31.85	0.0692	Incorrect	0	43	0	0.00	1.00
3	-37.07	0.7728	Near native	41	44	39	0.95	0.11
4	-41.16	0.0669	Incorrect	21	38	0	0.00	1.00
14	-19.31	0.2111	Near native	13	33	9	0.69	0.73
17	-41.90	0.0592	Incorrect	0	21	0	0.00	1.00
18i	-63.00	0.0669	Incorrect	0	49	0	0.00	1.00
20	-23.95	0.0674	Incorrect	0	45	0	0.00	1.00
25	-13.21	0.0600	Incorrect	0	27	0	0.00	1.00

To enlarge the target sets, we also tested 90% and 85% sequence homology. Lower sequence homology yields more structures; however, increases the risk of picking a different protein. For target sets enlarged with 90% sequence homology, we tested all templates. Due to computational cost, for target sets enlarged with 85% sequence homology, we tested templates that successfully matched target proteins found with higher sequence homology. At 95% sequence homology, successful results were obtained for 3 cases; the 90% and 85% sequence homology cases failed to give additional successful predictions (data not shown). 95% sequence homology covers sequence difference between unbound and bound forms in the benchmark dataset, where the lowest sequence similarity is 98.8% (**Table A.8**). Therefore, we consider 95% sequence homology as appropriate to further enlarge the target set.

When we could use the structures of bound forms, the success was 18 out of 30 ‘difficult’ interactions at 100% sequence homology, and considering additional successful results of cases 1ii and 3, increased the number to 20. When we did not use the data of the bound forms (see **section 3.3.2**, “Eliminating bias by discarding information of the bound form”), the success was 14 out of 28 ‘difficult’ interactions at 100% sequence homology, and considering additional successful results of cases 1ii and 14, it increased to 16. The final success rates are given in **Table 5.13**. We also present the ratio of successfully predicted results vs. favorable predictions in **Table 5.14**, which indicates the percentage of predictions that successfully match the bound forms in the benchmark dataset. When the standard target set was used together with the standard or modified template set, almost all predictions (except one obtained with the standard target and template sets) matched with the bound forms in the benchmark dataset. When the target set was enlarged with alternative conformations, more than 76% of the favorable predictions matched the bound forms following both two scenarios. **Figure 5.1** presents the distribution of the predictions based on their energies and IS-scores. Predictions with relatively lower energies can have low or high IS-scores; predictions with relatively higher energies have high IS-scores. Predictions with favorable energies

but low IS-scores (red triangles) might correspond to alternative binding modes of the proteins. The benchmark offers favorable interactions of protein pairs; yet, other conformations of the proteins may also interact in a different way than the bound forms given in the benchmark dataset, and we may capture such interactions. We checked if those interactions can be biological using machine learning tools, NOXclass [295], DiMoVo [296] and EPPIC [297] (**Table 5.15**). Default parameters, interface area, interface area ratio and area-based amino acid composition, were used in NOXclass evaluations; cut-off value was chosen as 0.5 in the DiMoVo as indicated there; the EPPIC server directly indicates if the interface is biological or the outcome of crystal packing. NOXclass evaluated 13 interactions, DiMoVo evaluated 2 interactions and EPPIC evaluated 5 out of 18 interactions as biological. 2 interactions were evaluated as biological by all three tools and 13 interactions by at least one tool. Machine learning tools depend on their training datasets; however, evaluation of some interactions as biological indicates that these interactions might be biological although structurally they are not similar to the bound forms given in the benchmark dataset.

Table 5.13 The success of PRISM predictions

Cases	Total Interaction	Predicted Interactions Using Standard Target Set		Predicted Interactions Using Enlarged Target Set	
	Number	Number	Percentage	Number	Percentage
Difficult cases (also using data of bound forms)	30	8	26.7%	20	66.7%
Difficult cases (not using data of bound forms)	28	6	21.4%	16	57.1%

Table 5.14 Successful PRISM predictions in energetically favorable PRISM predictions

Predictions	Favorable PRISM Predictions (Energy < -10 energy unit)	Successful PRISM Predictions (IS-score > 0.12)	Percentage of favorable PRISM results matching bound forms
Standard Target Set, Standard Template Set	9	8	88.9%
Enlarged Target Set, Standard Template Set	26	20	76.9%
Standard Target Set, Unbiased Template Set	6	6	100%
Enlarged Target Set, Unbiased Template Set	21	16	76.2%

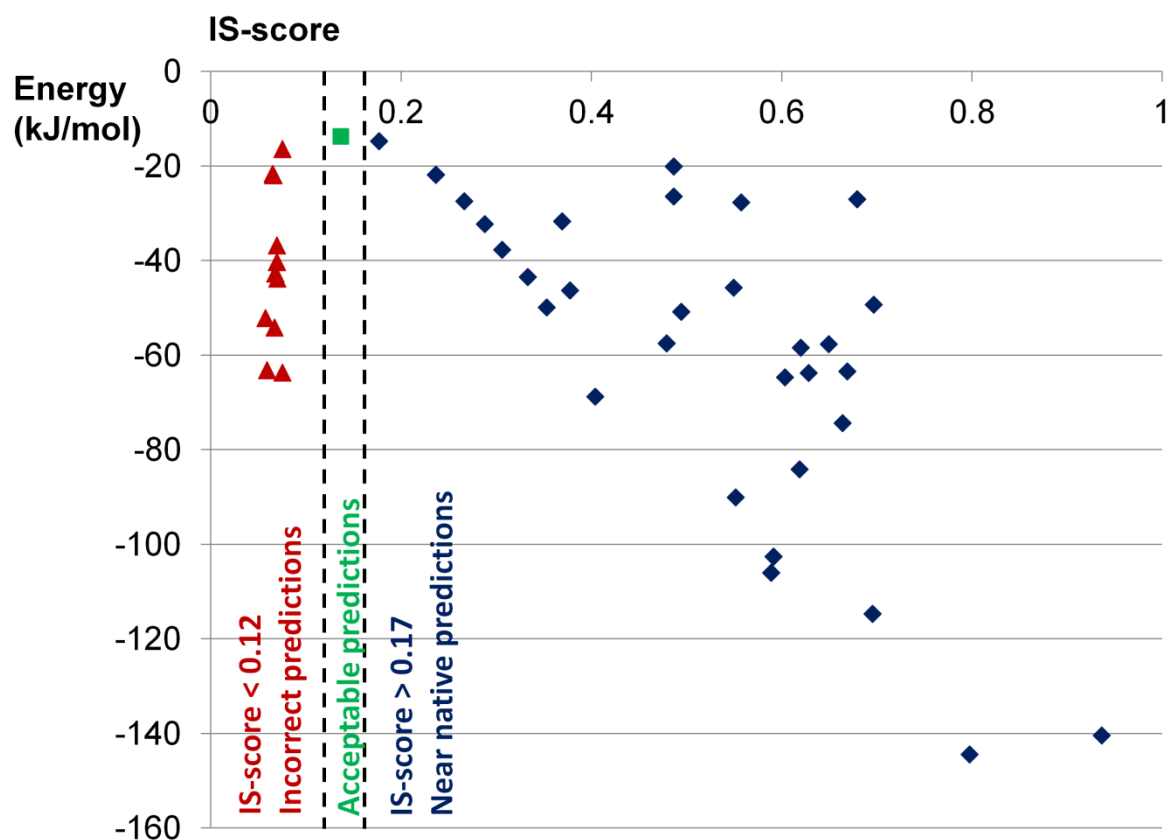


Figure 5.1 Distribution of predictions based on energy and IS-score [236].

Red triangles show incorrect predictions (IS-score < 0.12). Green square represents acceptable prediction ($0.12 < \text{IS-score} < 0.17$). Blue diamonds show near native predictions (IS-score > 0.17). They are distributed in a triangle profile where predictions with higher energies do not have low IS-scores. We should note that the predictions with favorable energies but low IS-scores (red triangles) might correspond to alternative binding modes of the proteins.

Table 5.15 Biological/non-biological interaction test results of predictions with favorable energies but lower IS-scores

Case	Pair	NOXclass score (%)	NOXclass evaluation	DiMoVo score	DiMoVo evaluation	EPPIC evaluation
2i	1dloA - 2hmiC	2.00	non-biological	-0.15848	non-biological	non-biological
2i	3nbpB - 2hmiC	91.82	biological	0.080515	non-biological	biological
2i	1s6pB - 2hmiC	78.26	biological	0.299148	non-biological	non-biological
2ii	2hmiD - 3kjbB	97.69	biological	0.167662	non-biological	non-biological
2ii	2hmiD - 1rt2B	97.69	biological	0.167662	non-biological	non-biological
4	1fpzF - 1pye	84.81	biological	0.150784	non-biological	non-biological
4	2iw9C - 1fqzE	37.42	non-biological	0.020412	non-biological	non-biological
14	1f59AB - 1i2mC	96.09	biological	0.277509	non-biological	non-biological
17	3k65B - 1jmiA	45.14	non-biological	-0.18798	non-biological	non-biological
17	1jmo - 1mkxK	99.98	biological	0.352058	non-biological	biological
18i	3pfu - 1jzoA	57.26	biological	0.877925	biological	biological
18i	1vrsB - 1jzoA	88.07	biological	0.234299	non-biological	biological
20	2yjfE - 1ux5	27.56	non-biological	0.12399	non-biological	non-biological
20	1ux5 - 3lueABCD	44.02	non-biological	0.002495	non-biological	non-biological
22	1fchB - 1c44	69.45	biological	0.143824	non-biological	non-biological
23	1ywhG - 3bt1A	88.67	biological	-0.14501	non-biological	non-biological
24	2idoBD - 1j54	99.97	biological	0.86499	biological	biological
25	1yzuAB - 2ot3A	60.38	biological	0.24899	non-biological	non-biological

5.1.3 A Case Study: ERK Interactions

We tested this method on the extracellular signal-regulated kinases (ERK) interactions. In the KEGG database [298], interactions among the proteins ERK, MEK1, MEK2, MP1, RSK2 and Mnk1/2 are given as in **Figure 5.2**. In this ‘classical’ edge-and-node representation, nodes are the proteins and edges are the interactions (**Figures 5.2 and 5.3A**). This protein-protein interaction network provides information related to which proteins interact; but not how they interact. **Figure 5.3A** presents as example: MEK1 and MEK2 interact with ERK; however, it is unclear whether these two proteins can (or cannot) interact with ERK simultaneously to form a trimer. In contrast, structural data provide the contacting residues and the interfaces of the interacting proteins, which allow distinguishing between distinct and overlapping protein-protein interactions [248]. Panchenko's group has also shown that the structurally inferred interaction network is more functionally coherent [299]. We constructed the structural network of the ERK interactions, first using one conformation of each protein (**Figure 5.3B**), and then including alternative conformations (**Figure 5.3C**). Interacting conformations are represented by an edge and binding sites are shown by small rectangles on the nodes [300]. As we show below, considering alternative conformations can improve the network. Alternative conformations were found in the PDB using 100% sequence homology. PDB IDs of the conformations are given in **Table 5.16**. The best results with respect to their energy values for the query proteins and their conformations are shown in **Tables 5.9 and 5.10** (if the chain name is not indicated, it is the whole structure). In the first case, 5 predictions were obtained and 4 of them were energetically favorable when the cut-off energy value was -10 energy unit (**Figure 5.3B, Table 5.17**). One of these was between MEK1 and MEK2, shown in **Figure 4.9A**. However, in the second case, 7 interactions were predicted, all favorable and better results (with lower energy values) were obtained for the previous predictions (**Figure 5.3C, Table 5.18**). For example, the energy of MEK1-MEK2 interaction prediction decreased from to -23.84 to -78.99 energy unit, when a different conformation (PDB ID 3e8n instead of 3eqc) of MEK1 was utilized. Compared to the

previous conformation of MEK1, this conformation is a better complementary structure to MEK2 (**Figure 5.4B**), which indicates that utilizing alternative conformations can improve the predictions and thus the structural pathways.

Table 5.16 Different conformations of ERK interacting proteins

Protein name	Query protein	Other conformations
ERK	2ojgA	3i5zA, 3i60A, 2e14A, 2ojjA
MEK1	3eqcA	3dv3A, 3eqiA, 3eqgA, 3eqdA, 3eqfA, 3eqhA, 2p55A, 3eqbA, 3e8nA, 1s9jA, 3dy7A, 3mb1A, 3pp1A
MEK2	1s9iA	1s9iB
MP1	3cptA	1skoA, 1tvmA
RSK2	2z7qA	2z7rA, 2z7sA
Mnk1/2	2hw7A	2ac5A

Table 5.17 Energetically most favorable PRISM predictions of the query ERK interacting proteins

Protein-1	PDB ID-1	Protein-2	PDB ID-2	Template Interface	Energy (energy unit)
MEK2	1s9iA	MEK1	3eqc	1s9iAB	-23.84
ERK	2ojg	MEK1	3eqc	1uaxAB	-10.78
MEK1	3eqc	RSK2	2z7q	1zxeAE	-12.27
MEK2	1s9iA	RSK2	2z7q	1zxeAE	-4.64
Mnk1/2	2hw7	ERK	2ojg	1fvrAB	-10.13

Table 5.18 Energetically most favorable PRISM predictions obtained using alternative conformations of the ERK interacting proteins

Protein-1	PDB ID-1	Protein-2	PDB ID-2	Template Interface	Energy (energy unit)
MEK2	1s9iA	MEK1	3e8n	1s9iAB	-78.99
MP1	1sko	MEK1	3dv3	1unhAB	-26.50
ERK	2ojg	MEK1	3mbl	1uaxAB	-16.09
MEK1	3eqc	RSK2	2z7s	1zxeAE	-31.40
MEK2	1s9iB	RSK2	2z7r	1zxeAE	-32.18
Mnk1/2	2hw7	ERK	2ojg	1fvrAB	-10.13
RSK2	2z7q	ERK	2ojj	1fvrAB	-30.78

Some of the predicted interactions have been shown experimentally. MEK1 activates ERK1 via phosphorylation. MEK1 residues 33-393, which cover all interacting residues in our model, interact with ERK [301]. Deletion of ERK residues 241-272, which in our model contribute 10 out of the 13 interacting residues, abolishes the interaction. However, this deletion does not affect the ERK-Mnk association, and in our model the interaction is not through this portion of the surface. MP1 is a MEK1 scaffolding protein which regulates cell spreading. In the MEK1-MP1 interaction, binding is through MEK1 residues 220-393, which in our model include all interacting residues (without contribution from MEK1 residues 1-219) [302]. No MEK2-MP1 interaction was detected in the two-hybrid system in agreement with our predictions. Kinase suppressor of Ras (KSR) is a scaffolding protein for the Raf-MEK-ERK complex. KSR2 promotes phosphorylation of MEK1 and the crystal structure of this heterodimer was obtained (PDB ID: 2y4i) [266]. Since the KEGG database does not include KSR2 in the MAPK pathway, we also did not include this protein in our predictions. However, PRISM could predict the KSR2-MEK1 interaction (data not shown). MEK1-MEK2 interaction determines the strength and duration of the ERK signal [303] and the ERK-RSK interaction is observed in response to various stimuli, like growth factors,

polypeptide hormones, neurotransmitters, and chemokines [304]. MEK1 and MEK2 signaling stimulate ribosomal S6 kinase (RSK) [305].

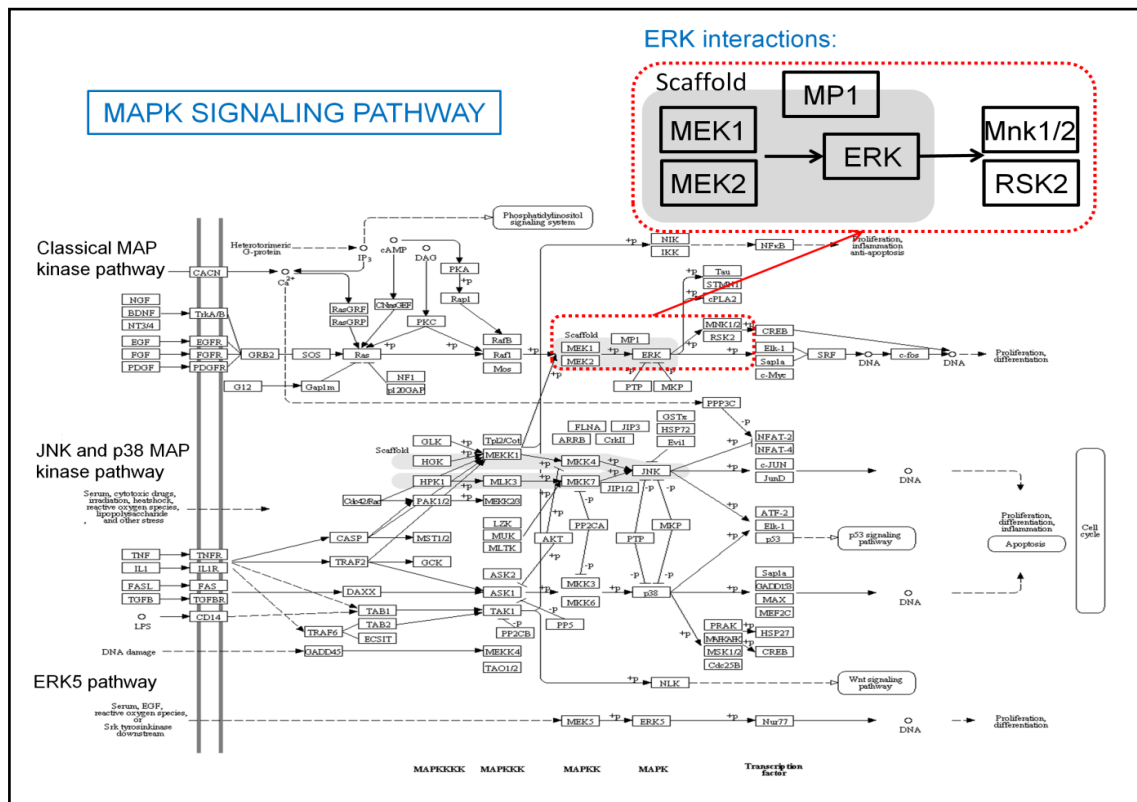
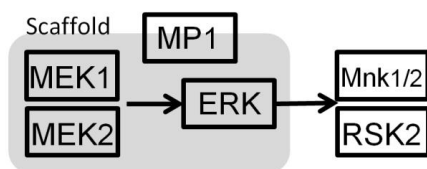


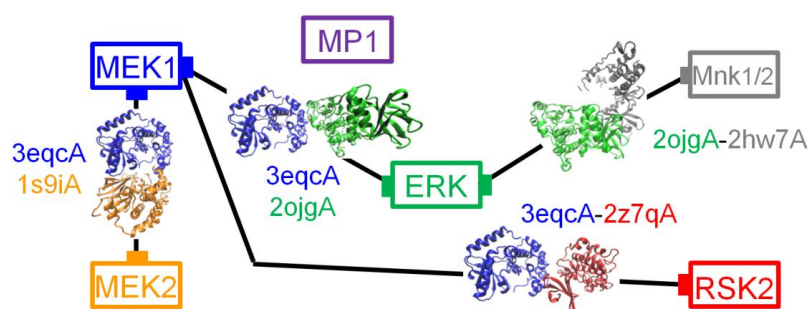
Figure 5.2 MAPK signaling pathway given in the KEGG database.

Nodes are the proteins and edges are the interactions. Gary lines indicate the membrane. The data is taken from the KEGG database. The focus is on ERK interactions with MEK1, MEK2, MP1, Mnk1/2 and RSK2.

A) ERK interacting proteins given in the KEGG database



B) PRISM predictions using one conformation of each protein



C) PRISM predictions using alternative conformations in the PDB

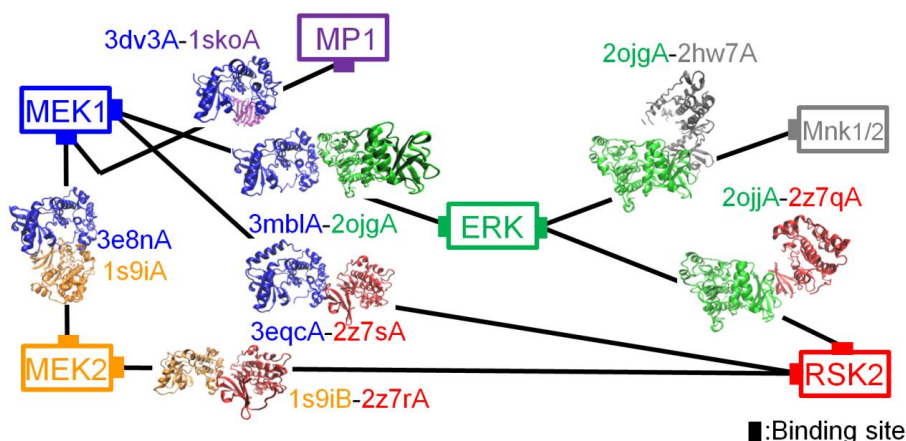


Figure 5.3 Interactions of ERK protein as given in the KEGG database and predicted by PRISM [252].

A) ERK interactions with MEK1, MEK2, MP1, Mnk1/2 and RSK2. The data is taken from the KEGG database. B) PRISM predictions obtained by using one conformation of the ERK interacting proteins. Different colors are for different proteins. Binding sites are shown as little rectangles. If two interactions are through the same binding site, edges are connected to the same little box. C) PRISM predictions obtained by using alternative conformations of the ERK interacting proteins. Alternative conformations are found from the PDB using 100% sequence homology.

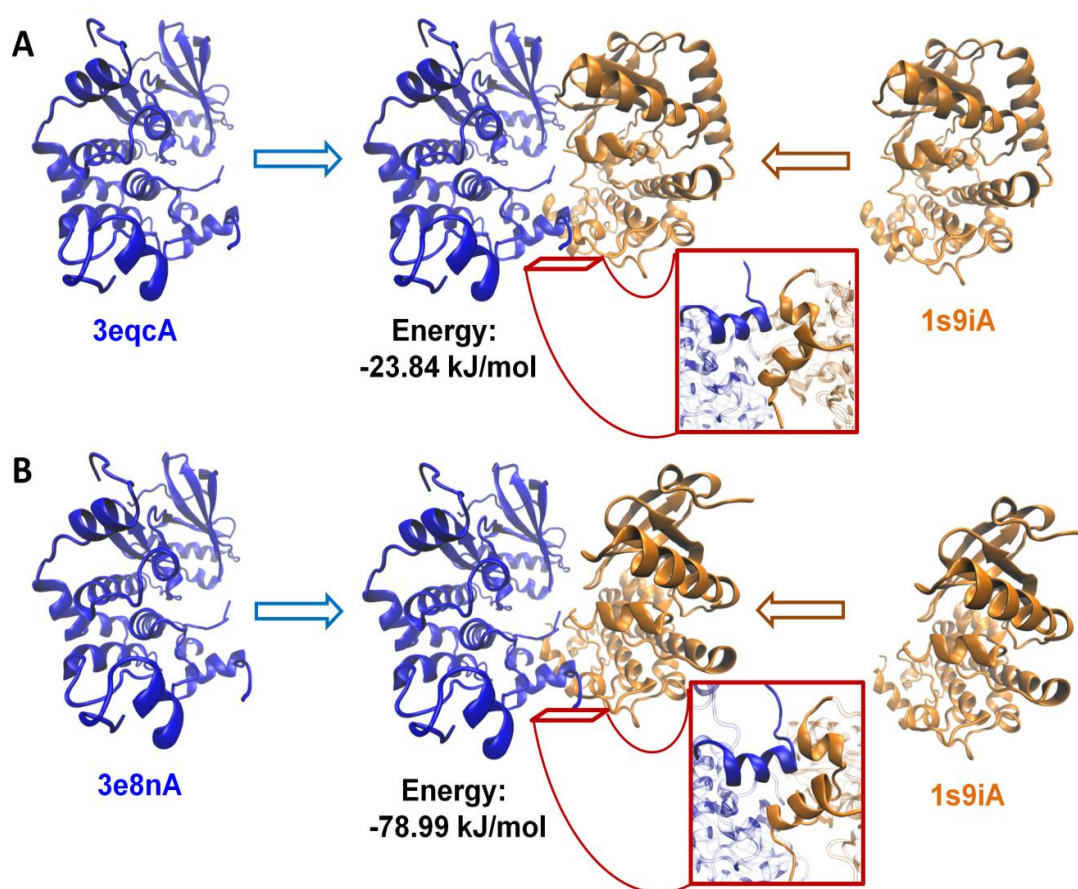


Figure 5.4 Predicted interactions of MEK1-MEK2 [252].

A) Predicted interaction of MEK1-MEK2, PDB IDs are 3eqcA (MEK1) and 1s9iA (MEK2). PRISM calculated the energy as -23.84 energy unit. Red box focuses on the helices where two structures are most close to each other. B) Predicted interaction of MEK1-MEK2, PDB IDs are 3e8nA (MEK1) and 1s9iA (MEK2). PRISM calculated the energy as -78.99 energy unit, lower than the energy calculated in part A. Red box focuses on the helices where two structures are most close to each other. It shows helices are more complementary to each other compared to the interaction in part A.

Using a structural protein-protein interaction network, we are able to observe which interactions can (or cannot) occur simultaneously. If two proteins interact through different sites of a third protein and there is no spatial clash, the two interactions can occur simultaneously, forming a trimer. For example, a trimer of ERK-MEK1-MEK2

can form, as shown in **Figure 5.5**. The energetically most favorable predictions were 3mblA-2ojgA (PDB codes) for MEK1-ERK and 3e8nA-1s9iA for MEK1-MEK2. These formed via two different conformations of MEK1, 3mblA and 3e8nA. In the construction of the trimer, we considered one of these MEK1 conformations, 3mblA. The most favorable interaction of 3mblA with a MEK2 was 3mblA-1s9iB (energy: -62.75 kJ/mol, template: 1s9iAB). Since ERK and MEK2 bind through different sites of MEK1 and their residues did not clash, a trimer could be obtained based on MEK1-ERK and MEK1-MEK2 interactions. The trimer structure was modeled by superimposing MEK1 (3mblA) of MEK1-ERK (3mblA-2ojgA) and MEK1-MEK2 (3mblA-1s9iB). **Figure 5.6** presents the most favorable interacting conformations of ERK, obtained by using alternative conformations. Different conformations of the same protein are given by different structures in the same color. Collectively, these show which conformation of a protein is more favorable to interact with another protein, and indicate that the protein switches its conformation while switching its interacting partner, information that is valuable for figuring out pathway regulation.

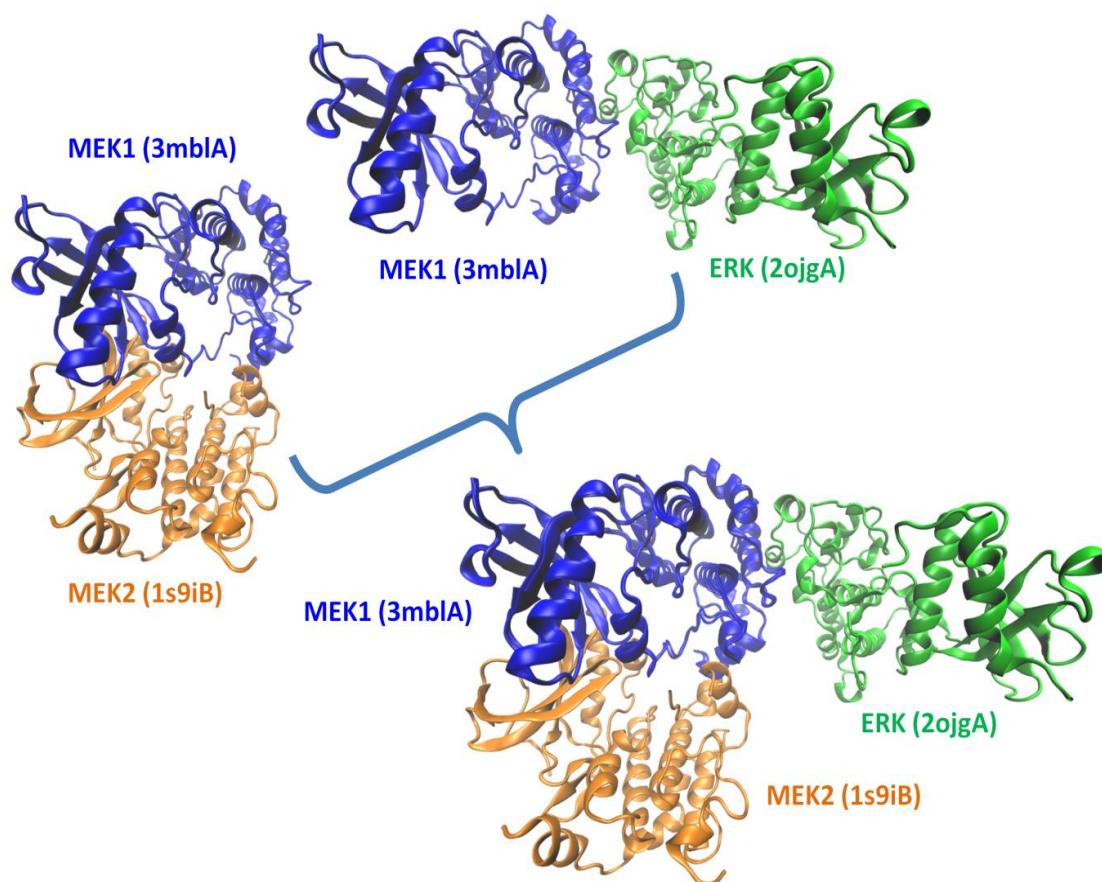


Figure 5.5 Trimer structure of ERK, MEK1 and MEK2.

Trimer structure was obtained via superimposing MEK1 protein (3mblA) of MEK1-ERK (3mblA-2ojgA) and MEK1-MEK2 (3mblA-1s9iB) interactions predicted by PRISM. Blue: MEK1 protein, green: ERK protein and orange: MEK2 protein.

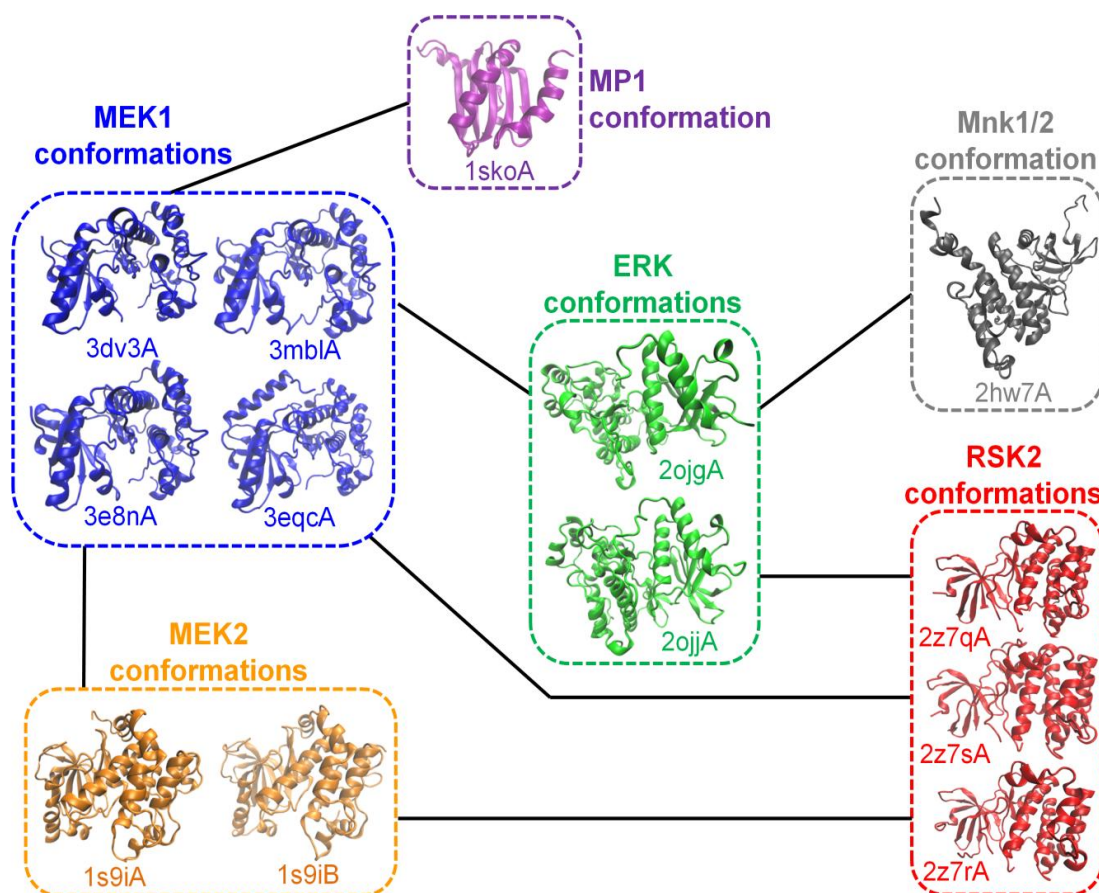


Figure 5.6 Interacting conformations of ERK interacting proteins.

Interactions are energetically most favorable PRISM predictions obtained by using alternative conformations. Different conformations of the same protein are given by different structures in the same color.

5.2 Exploiting Conformational Ensembles in Modeling Protein Assemblies

We applied the method also in modeling protein assemblies. Proteins are flexible and can change their conformations upon binding. Therefore, constructing a protein assembly starting from unbound forms of the components, illustrated in **section 4.2**, may lead to unsuccessful predictions. An assembly can be constructed more successfully with the help of different conformations of the query proteins [306]. The PDB offers different conformations of proteins, including their unbound, bound or any

alternative forms such as mutants, those obtained following post-translational modifications or of different crystal forms. We identified PDB structures with 100% sequence homology to the query proteins in the benchmark set used in **Chapter 4** (**Table 3.2**) and determined the structurally different ones using structural alignment. We obtained better results in the construction of 1eer and 1gp2 using alternative conformations: 3.23 Å RMSD rather than 5.56 Å for 1eer, and 2.50 Å RMSD rather than 3.57 Å for 1gp2. In this part, predictions with the lowest energy value were selected for each binary interaction and the assemblies were constructed based on only these binary interactions. Alternative conformations of these proteins can be found in **Table 5.19** and the results of the assembly construction using these alternative conformations are given in **Table 5.20** (details in **Tables 3.4 and A.10**; the first binary interaction prediction is given as step 0 and n^{th} iteration).

Table 5.19 Alternative conformations found from the PDB

Assembly	Unbound form (query protein)	Alternative conformations found from the PDB
1eer	1buyA	1cn4C
1eer	1ernA	1ebaA
1gp2	1giaA	1agrD, 1cipA, 1gfiA, 1gg2A, 3ffaA, 1gitA, 2pz3A, 1y3aA, 1gddA
1gp2	1tbgABCD	-
1gp2	1tbgEFGH	1gotG

Table 5.20 Results for the construction of protein assemblies starting from alternative conformations of the unbound forms

PDB structure	No. chain	No. residue	Target set	Energy unit	RMSD (Å)
1eer	3	620	1gia, 1tbgA, 1tbgE	-75.58	5.56
1eer	3	620	1gia, 1gg2A, 1tbgA, 1tbgE	-122.19	3.23
1gp2	3	764	1buyA, 1ernA x2	-279.38	3.57
1gp2	3	764	1buyA, 1ernA x2, 1cn4C, 1ebaA x2	-376.18	2.50

Predicted PDB structures are given with their PDB IDs, number of chains and number of residues. Target set includes unbound forms and their alternative conformations found in the PDB. If an unbound structure is used twice, it is indicated with “x2”. Different target sets are used: i) target set of unbound forms ii) target set of unbound forms and their alternative structures. Their best energy predictions are compared with respect to energy scores and RMSD values. RMSD is calculated for all backbone atoms and the energy value of an assembly is the summation of energy values calculated for the addition of each protein.

5.3 Concluding Remarks

Identification of PPIs is a first step in the elucidation of the interactome-scale structural network. Toward this aim, a tool which can help in the modeling of the structural pathway can be very useful. Here, we investigated whether different conformations of the query proteins found from the PDB could increase success in PPI identification by a motif, knowledge-based protein-protein interaction prediction tool. Knowledge-based methods can predict successfully an interaction between proteins and its three dimensional structure, if a physically and biologically meaningful interaction exists in the database. They neither consider the conformational space nor model conformational changes. Flexibility is considered only in the refinement process, which is the last step in the prediction, where side chain and slight backbone re-orientation are carried out. In principle, the capability of knowledge-based methods, like PRISM, to consider protein flexibility is limited by the supply of different conformations to the target set. Adding different conformations to the target set and searching interactions

among these can be expected to improve the predictions. In this study, we tested such a strategy. Possible conformations of the query proteins were extracted from the PDB and treated as individual structures in the predictions. These structures were guaranteed to belong to the same protein by requiring high percentage (first 100%, and then 95%) sequence homology. Different conformations were identified by structural similarity to eliminate redundancy.

Adding sufficiently different conformations of proteins to the target set may allow predicting the interaction of two proteins in any form. However, two factors limit the capability of template-based methods to predict correctly the native interaction: (i) the template set organization, and (ii) availability of the ‘right’ conformations of the query proteins. For the first, it is important that the template structures represent the interactions in their corresponding clusters, otherwise, appropriate structural similarity cannot be found between the target and template structures. Second, in particular, a major limitation of the prediction success is the coverage of the PDB. If the appropriate conformation exists in the PDB, it can be found and the interaction can be predicted successfully; in its absence the prediction will fail. The exponentially increasing numbers of structures in the PDB promises a high likelihood of finding the ‘right’ conformations of the query proteins. More successful predictions will be obtained as the PDB covers more structures. Even with these limitations, knowledge-based approaches offer successful predictions. A large scale, systematic study [281] indicates that knowledge-based approaches are more reliable than ‘blind’ docking strategies. Appropriate templates can be extracted to model nearly all complexes of structurally characterized proteins, although there are currently a limited number of protein–protein complexes in the PDB.

In the specific example here with PRISM, successful prediction was achieved for almost all the ‘easy’ cases (87 out of 88) of the docking benchmark dataset [307], and considering other conformations in the PDB, ‘difficult’ cases of this dataset were predicted with 66.7% success rate. A success rate 57.1% was obtained in unbiased

predictions. Previous studies predicted interactions in signaling pathways (76% and 78% accuracies in the ubiquitination [245] and apoptosis [246] pathways, respectively) and constructed structural networks [248]. Here, we modeled the structural network of ERK, first based on one conformation of each protein; then considering alternative conformations in the PDB an improved structural network was obtained. Given the limited number of architectural motifs in single chain proteins and as we have shown earlier in protein-protein interfaces [237, 238, 284, 308], and the difficulties facing large-scale applications of ‘blind’ docking methods, knowledge-based approaches are a reasonable venue for proteome-scale cellular network constructions. Fast incorporation of experiment-based flexibility into these is advantageous. While vastly incomplete, and with limited sampling, it is a reasonable and reliable strategy to adopt. We also showed that exploiting alternative conformations leads better predictions in modeling protein assemblies.

Chapter 6

RESULTS OF EXPLOITING ELECTRON MICROSCOPY DATA

Choosing the right prediction is an important problem in the assembly construction approaches. As we mentioned in **Chapter 4**, the energy value can be an indicator, which is not always the case. However, experimental techniques providing low resolution data of assemblies, like Cryo-EM, FRET and SAXS, can be used to help select the solutions. With the recent advances in the technology, EM has come forward among others and become a preferred method to identify structures of protein assemblies. It can be combined with computational approaches in two main ways: proteins can be computationally fitted into EM density maps or docking solutions obtained computationally can be filtered through EM data. As an example for the latter one, we constructed protein assemblies using the assembly construction method as introduced in **Chapter 4** and selected the solution model based on the EM density map. One of the cases in the benchmark used in **Chapter 4** (introduced in **section 3.3.2**) is 1wnr. 1wnr is a heptamer of 10kDa chaperonin. In the construction of 1wnr starting from its unbound form, we obtained thirteen different structures. We exploited EM density map to select the solution. EM density map of co-chaperonin protein 10 complexed with GroEL and ADP, where 10 kDa chaperonin is at the top of the structure, is available in the EMDDataBank (EMD 1531). We docked thirteen solutions into the EM density map using Chimera. Only one result matched with the top of the density map (**Figure 6.1**), Chimera calculated the correlation as 0.85, and that result has the lowest RMSD (4.60 Å) among those thirteen structures. It did not fit perfectly into the density map, since it does not have a perfectly symmetrical cyclic shape. However, RMSD of 4.60 Å is still acceptable. The density map helped us to choose the structure

with the best RMSD among those thirteen, which suggests that experimental data like EM density maps can help in choosing the solution.

Based on the success in the case of chaperonin, PDB ID: 1wnr, we introduced a modified version of the method to construct multi-protein assemblies; it exploits EM density maps to filter improper predictions. Filtering structures through EM data at each protein addition step prevents assembly construction based on wrong structures and saves computational time. We illustrate our method on HIV protein interactions and tested it on benchmark datasets. Then, we discuss advantages and shortcomings of our method.

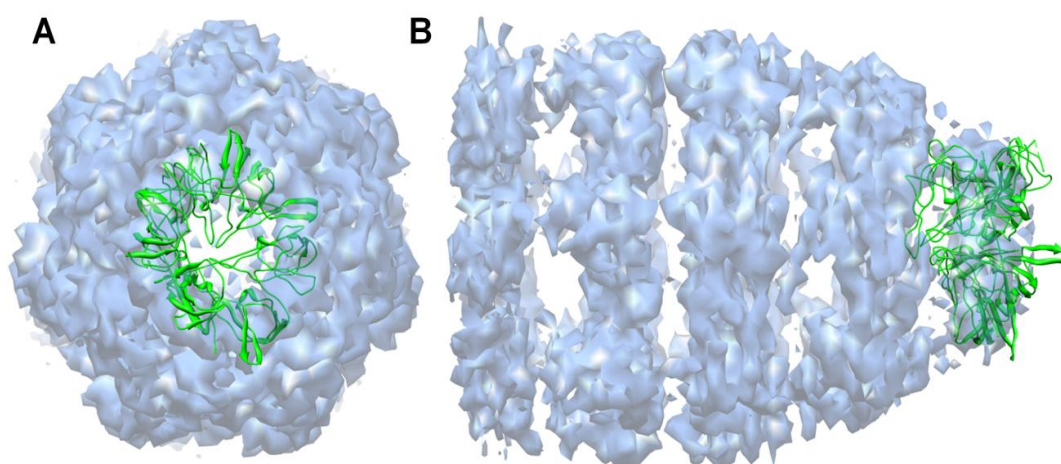


Figure 6.1 Predicted structure docked in the EM density map [236].

The structure (green) is docked into the EM density map (blue, EMD ID: 1531) using Chimera. 30 solutions are created using “fit” command. Top (A) and side (B) views are given.

6.1 Results

We first illustrated the construction of assemblies using our method on HIV protein interactions: (i) HIV protein complexed with the human membrane protein CD4 and the antibody 17b, whose complex structure is available in the PDB (PDB ID 1gc1), (ii) HIV protein complexed with the human membrane protein CD4 and the antibody m36 whose structure is not available in the PDB. Then, we tested our method on a benchmark dataset of various assemblies. We constructed assemblies in the benchmark set and evaluated our predictions with respect to PDB structures and EM density maps. We also tested Multifit [154, 188] and Segger [233, 234], and compared our results with theirs.

6.1.1 Case Study 1. HIV Protein gp120 – Membrane Protein CD4 – Antibody 17b Complex

The PDB structure 1gc1 has four chains and shows the interaction of HIV gp120 with the human membrane protein CD4 and heavy and light chains of the antibody 17b. The electron microscopy density map of that assembly is available as EMDB ID 5020. We started to construct the assembly using the PDB structures 3jwoA (HIV gp120), 3cd4A (CD4) and 1rz8AB (17b). 3jwo shows the interaction of HIV gp120 with another antibody 48d; 3cd4 is the unbound form of human membrane protein CD4; 1rz8 is the dimer form of heavy and light chains of the antibody 17b. We modeled quadromer structures. The model having the best fit has a Situs score of 1.07 Å. The correlation calculated by Chimera is 0.91. IS-score of interfaces (gp120-CD4, gp120-17b heavy chain and 17b heavy chain-17b light chain) are 0.71, 0.19 and 0.53, which are greater than 0.17 and indicate interfaces as ‘correct’. **Figures 6.2A and 6.2B** present the best model fitted into the density map.

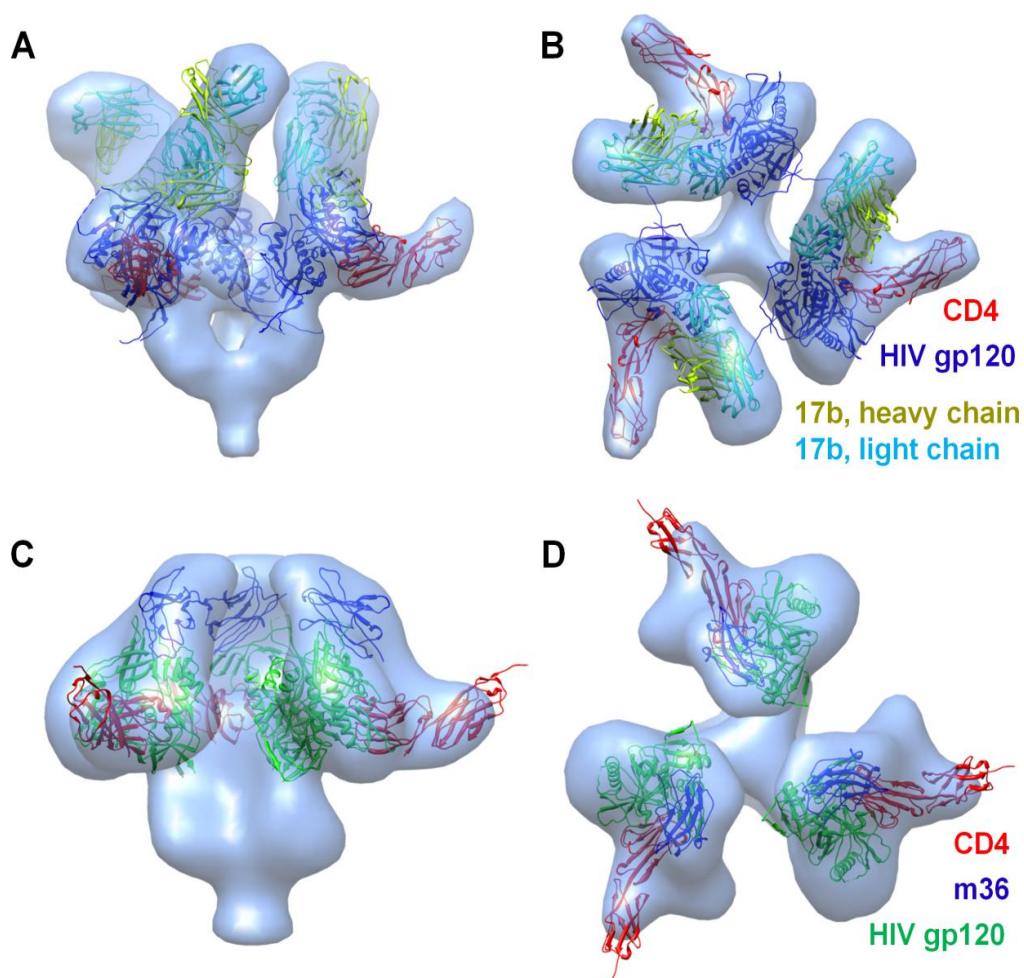


Figure 6.2 HIV interactions.

The best models of HIV gp120-CD4-17b (perspective (A) and top (B) views) and HIV gp120-CD4-m36 (perspective (C) and top (D) views) are shown as fitted into their EM density maps (EMDB IDs: 5020 and 5554, respectively). In “A” and “B”, HIV gp120 is shown as blue, CD4 as red, heavy chain of 17b as yellow and light chain of 17b as cyan. In “C” and “D”, HIV gp120 is shown as green, CD4 as red and m36 as blue. Three symmetric complexes are created using Chimera based on the symmetry of the density maps.

6.1.2 Case Study 2. HIV Protein gp120 – Membrane Protein CD4 – Antibody m36 Complex

We also tested our method to construct an assembly which includes a protein whose structure is not available in the PDB. EMDB ID 5554 presents the assembly of HIV gp120 protein complexed with the human membrane protein CD4 and the antibody m36. Unlike HIV gp120 and CD4, the structure of the antibody m36 is unknown. We first modeled the structure of m36 using the homology modeling tool I-Tasser. We constructed assemblies using the PDB structures 4dkuA (HIV gp120), 1gc1C (CD4) and the m36 model I-Tasser created. Here we used different PDB structures of HIV gp120 and CD4 rather than the ones we used in the construction of HIV gp120-CD4-17b; 4dku is the dimer HIV gp120 and 1gc1 shows the interaction of HIV gp120-CD4-17b. We modeled trimer structures using our method. The best assembly model has a Situs score of 3.01 Å and a Chimera correlation of 0.89. **Figures 6.2C and 6.2D** present the best model fitted into the density map. HIV gp120-CD4 interface has an IS-score of 0.77, which indicates the interface as ‘correct’. The antibody m36 in our model placed in the same direction and in the same space in the density map as indicated by the previous study [309]. In that study, the density map could not be fitted by a complex structure due to the absence of a PDB structure or a model. Here we presented not just the locations but also the full structure of the complex.

6.1.3 Benchmark

We tested our method on the benchmark dataset including various proteins (**Table 3.2**) in a similar way as we illustrated in the construction of HIV protein interactions. As we noted above, we constructed assemblies starting from unbound structures of the proteins if available or from the homology models. We used our default template set and included specific interactions for some cases (**Table 6.1**); 2cnzAB is the interaction of Saf Pulin complexed to Saf Pulin peptide, 3ikuHJ is the dimer interface of ParM filament, 3ngbGH is the interface of HIV gp120 and the antibody VRC01, and 3kfkCD is the dimer interface of a chaperonin. We selected the model with the lowest Situs

score as the best model. **Table 6.1** provides the results. We compared our predictions with the PDB structures and presented docking results with Situs score and correlation (**Figure 6.3**).

Table 6.1 Assembly construction results of the benchmark dataset

EMDB ID	Predicted structure	Target set	Template Set	PDB structure, Situs result	Model, RMSD	Model, IS-score	Model, Situs result	Model, correlation
1494	3mer	3crfB, 2co1AB	Default set +2cnzAB	1.16 Å	4.92 Å	0.80 – 0.83	1.36 Å	0.79
1980	3mer	2zhcA	Default set +3ikuHJ	6.16 Å	3.83 Å	0.14 – 0.21	4.17 Å	0.59
1980	4mer	2zhcA	Default set +3ikuHJ	6.16 Å	4.16 Å	0.12 – 0.20	7.38 Å	0.51
2055	5mer	4aodA-model	Default set	2.22 Å	2.16 Å	0.33 – 0.38	1.95 Å	0.89
2427	3mer	4lspG, 3se9H-model, 3se9L-model	Default set +3ngbGH	1.78 Å	4.64 Å	0.14 – 0.33	2.01 Å	0.84
5140	8mer	3iyfA-model	Default set +3kfkCD	2.35 Å	7.35 Å	0.24 – 0.37	2.95 Å	0.78
5505	6mer	1e9rA-model	Default set	3.04 Å	2.44 Å	0.67 – 0.71	1.07 Å	0.84
5672	6mer	3dvlA-model	Default set	1.53 Å	1.86 Å	0.41 – 0.51	4.50 Å	0.96

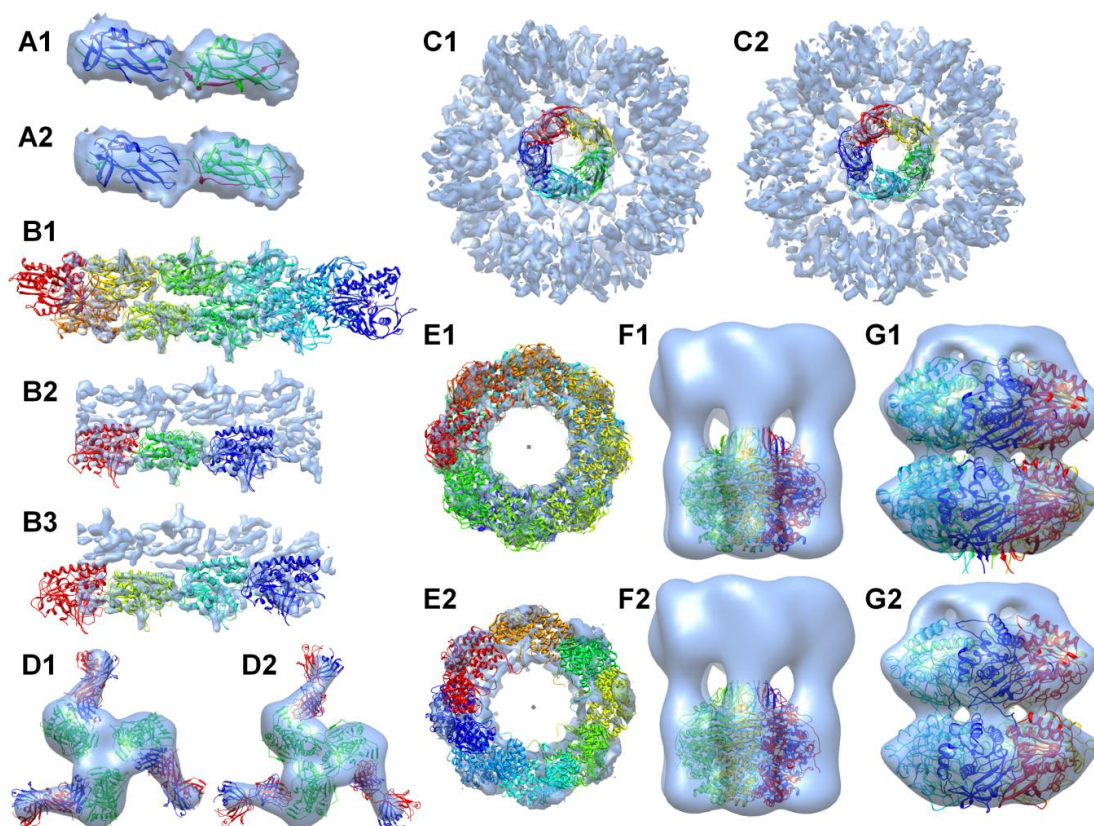


Figure 6.3 Benchmark assemblies docked into the EM maps.

The PDB structures and the best models fitted into EM density maps are given. Each chain is shown in different colors. Case 1: Saf Pilus-Type A, PDB ID 3cre (A1) and the model (A2) are fitted into EMDB ID 1494. Case 2: ParM filament, PDB ID 4a6j (B1), trimer model (B2) and quadromer model (B3) are fitted into EMDB ID 1980. Case 3: Acetylcholine-binding protein type 1, PDB ID 4aod (C1) and the model (C2) are fitted into EMDB ID 2055; EM density map level was decreased to 1.80 to obtain an image with the Chimera fit. Case 4: VRC-PG04 in complex with HIV-1 gp120, PDB ID 3se9 (D1) and the model (D2) are fitted into EMDB ID 2427; three complexes are created using Chimera based on the symmetry of the density map. Case 5: Lidless Mm-cpn, PDB ID 3iyf (E1) and the model (E2) are fitted into EMDB ID 5140. Case 6: Conjugal transfer protein TrwB, PDB ID 1e9r (F1) and the model (F2) are fitted into EMDB ID 5505; EM density map level was decreased to 0.6 to have distinctive image of the density map. Case 7: Circadian Clock Protein KaiC, PDB ID 3dvl (G1) and the model (G2) are fitted into EMDB ID 5672; EM density map level was decreased to 0.1 to have distinctive image of the density map.

Comparing with the PDB structures, our models have RMSD of 1.86-7.35 Å and docking into density maps resulted in Situs scores of 1.07-7.38 Å and correlations of 0.51-0.96. Excluding the result of Lidless Mm-cpn, our results have RMSD less than 5 Å and excluding the results of ParM filament, we obtained correlation close to or higher than 0.8. We evaluated each interface with IS-score; 28 interfaces were evaluated as ‘correct’, 4 interfaces as ‘acceptable’ and none as ‘incorrect’ (all IS-scores are given in **Table S29**). Two interfaces could not be obtained due to imperfect cyclic models. Here, we investigated the cases in which we could not obtain very good results. ParM filament density map, EMDB ID 1980, shows two stripes and each covers three complete and one partial subunits. First we modeled the trimer structure starting from the unbound form 2zhc. Correlation of our trimer model (**Figure 6.3B2**) is 0.59, which is quite low; but the Situs score is 4.17 Å and RMSD compared to the PDB structure, 4a6j, is 3.83 Å, which are good results. Then, based on the trimer model we obtained the quadromer model, which is larger than the density map (**Figure 6.3B3**). Docking into the density map indicated not very successful results, correlation of 0.59 and Situs score of 7.38 Å. However, when we compared our quadromer model with the PDB structure, it has an RMSD of 4.16 Å, and when we docked the PDB structure into the density map, the Situs score is 6.16 Å, which is close to the Situs score of our quadromer model, 7.38 Å. Here, our models have similar correlation and fitting results to the PDB structure. In another case, we modeled open state of lidless Mm-cpn, PDB ID 3iyf and EMDB ID 5140, and obtained an octamer model (**Figure 6.3E2**) which does not have a perfect cyclic structure (IS-scores are given in **Table S29**). It has an RMSD of 7.35 Å. However, when we docked our model and the PDB structure into the density map, we obtained very similar Situs scores, 2.95 Å and 2.35 Å, respectively. This indicates that our model can fit into the density map in a similar way that the PDB structure does. Another case that we could not have a perfect cyclic model is conjugal transfer protein TrwB (IS-scores are given in **Table S29**). However, we obtained a hexamer model very similar to the PDB structure, 1e9r, and the RMSD is 2.44 Å.

Docking into the density map also resulted in good results, a Situs score of 1.07 Å and a correlation of 0.84.

We also tested Multifit and Segger on our benchmark dataset. Multifit could not model pentamer of acetylcholine-binding protein type 1 and octamer of lidless Mm-cpn starting from the homology models and quadromer of ParM filament starting from the PDB structure, 2zhc. It could successfully predict hexamer of circadian clock protein KaiC and only one interface in the trimer of ParM filament regarding the IS-score (IS-scores are given in **Table S30**). When we used Segger, we could obtain the desired segments only for pentamer of acetylcholine-binding protein type 1 and octamer of lidless Mm-cpn in the open state. We could also obtain desired number of segments for ParM filament, conjugal transfer protein TrwB and circadian clock protein KaiC; but segmentation was not properly (segmentation results are given in **Figure 6.4**). Only two interfaces of lidless Mm-cpn were evaluated as ‘correct’ regarding the IS-score (IS-scores are given in **Table S31**). Models obtained by using Multifit and Segger are shown in **Figure 6.5**. We also found the correlation and calculated RMSD values compared to the PDB structures for the models we could have (**Table 6.2**). Comparing the performance of our method with theirs, we could obtain 28 ‘correct’ and 4 ‘acceptable’ interface predictions, and could not model 2 interfaces (there is no interface or not enough contacting residues of two proteins). For the 5 assembly models that Multifit could provide, 7 ‘correct’ and 10 ‘incorrect’ interfaces were obtained, and 1 interface could not be modeled. Segger could predict models of 2 assemblies, and provided 2 ‘correct’ and 5 ‘incorrect’ interface predictions; 6 interfaces could not be modeled. If we consider that a successful prediction of an assembly is obtained when all subunits are connected by successfully predicted interfaces, which are labeled as ‘correct’ or ‘acceptable’, we could successfully construct all 8 assemblies; however, Multifit could successfully predict only one of them and Segger could construct none. Correlations results are about the same on average but comparing to the PDB structures, our RMSD results are lower than the results of other two methods.

Finally we tested our method on the Multifit benchmark set and evaluated our performance, given in **Table 6.3**. We reconstructed those assemblies also using Multifit by Chimera to calculate their RMSD and correlation values. RMSD values we calculated are very similar to the values given in the Multifit paper [188]. In the assembly construction, we could obtain better results compared to Multifit's performance (correlation: 0.81-0.98, RMSD: 0.90-4.62 Å); correlation values of our models are between 0.97-0.99, equal or higher than correlation of Multifit results and RMSD values are between 0.50-1.68 Å, lower than Multifit results in each case. We also evaluated interfaces of the models via IS-scores (**Table S32**). Based on IS-scores, we could obtain more correctly predicted interfaces than Multifit. We predicted all 51 interfaces as 'correct' and Multifit had 33 'correct', 1 'acceptable' and 5 'incorrect' interface predictions (12 interfaces could not be predicted). Regarding successfully predicted interfaces, which are labeled as 'correct' or 'acceptable', we could successfully construct all 8 assemblies, however, Multifit could successfully predict 5 of them (1qu9, 1urz, 1mtu, 1oel and 1gru).

Table 6.2 Assembly construction results of three methods in our benchmark set

EMDB ID	PDB ID	Our method			Multifit			Segger		
		correlation	IS-score results	RMSD	correlation	IS-score results	RMSD	correlation	IS-score results	RMSD
1494	3cre	0.79	2 correct	4.92 Å	0.76	2 incorrect	22.5 Å	-		
1980	4a6j	0.59	1 correct 1 acceptable	3.83 Å	0.68	1 correct 1 incorrect	3.20 Å	-		
1980	4a6j	0.51	1 correct 2 acceptable	4.16 Å	-			-		
2055	4aod	0.89	5 correct	2.16 Å	-			0.85	5 incorrect	5.07 Å
2427	3se9	0.84	1 correct 1 acceptable	4.64 Å	0.84	1 incorrect 1 no interface	16.14 Å	-		
5140	3iyf	0.78	7 correct 1 no interface	7.35 Å	-			0.86	2 correct 6 no interface	30.3 Å
5505	1e9r	0.84	5 correct 1 no interface	2.44 Å	0.91	6 incorrect	27.4 Å	-		
5672	3dvl	0.96	6 correct	1.86 Å	0.79	6 correct	2.34 Å	-		

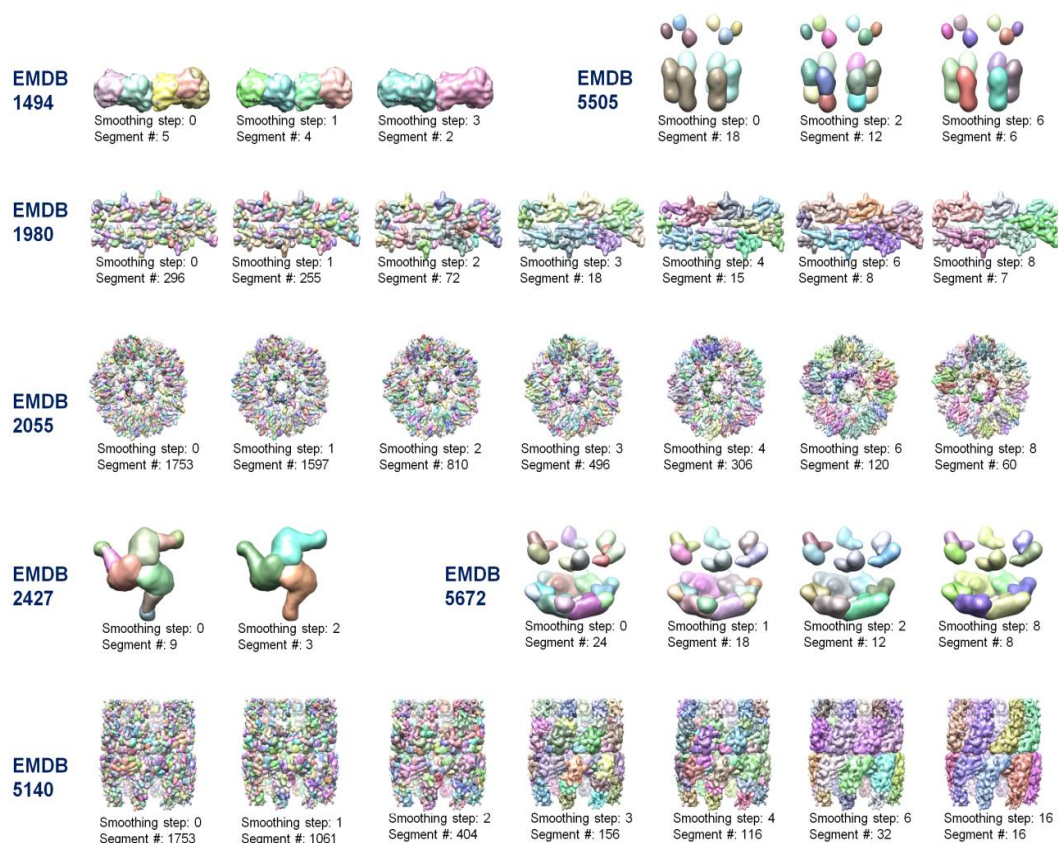


Figure 6.4 Results of segmentation by Segger

Segger gives different segmentation for different smoothing step number. Smoothing step and segment numbers are given below each figure. Figures obtained with different smoothing step number but giving the same result are not shown here; for example, smoothing step 2 gave the same result as smoothing step 1. We could obtain proper segmentation for EMDB IDs 2055 and 5140. We could obtain desired number of segments for EMDB IDs 1980, 5505 and 5672, but segmentation was not properly. Using different contour level of density maps did not change the segmentation.

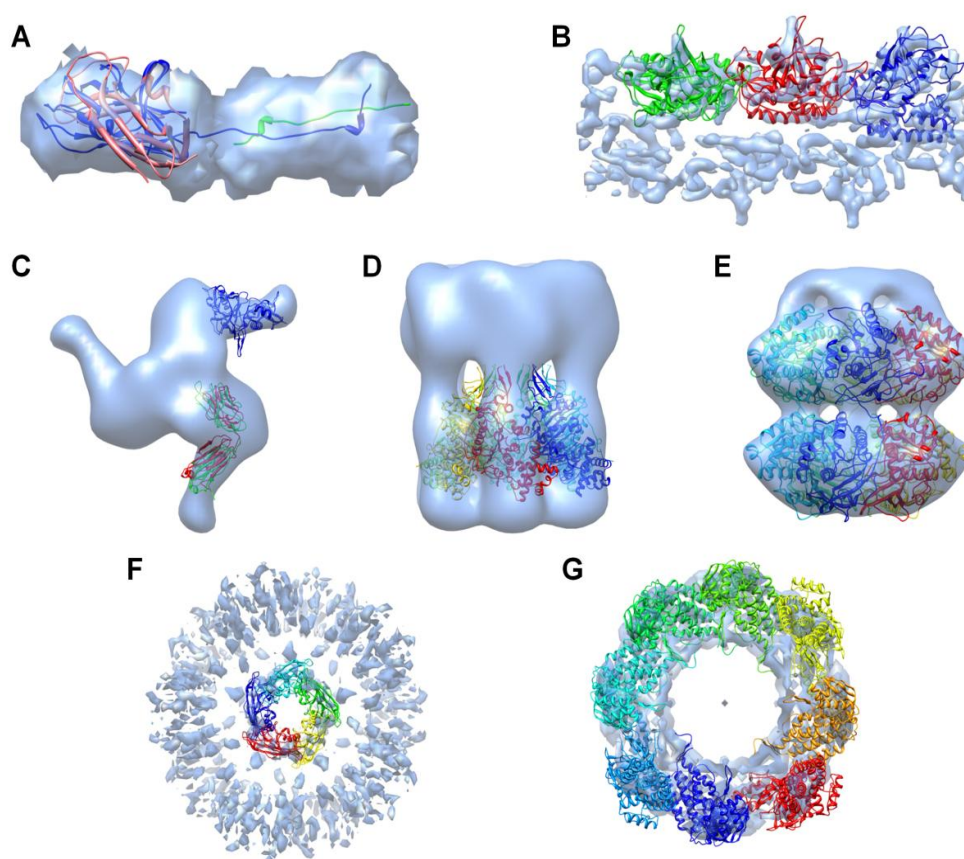


Figure 6.5 Models obtained by using Multifit and Segger.

Models are fitted into EM density maps. Chains are given in different colors. (A) Multifit model of Saf Pilus fitted into EMDB ID 1494. (B) Multifit model of ParM filament fitted into EMDB ID 1980. (C) Multifit model of HIV-1 gp120/VRC-PG04 complex fitted into EMDB ID 2427. (D) Multifit model of conjugal transfer protein TrwB fitted into EMDB ID 5505. (E) Multifit model of circadian clock protein KaiC fitted into EMDB ID 5672. (F) Segger model of acetylcholine-binding protein type 1 fitted into EMDB ID 2055. (G) Segger model of lidless Mm-cpn fitted into EMDB ID 5140.

Table 6.3 Assembly construction results in the Multifit benchmark set

PDB ID	Structure	Multifit		Our method	
		RMSD	Correlation	RMSD	Correlation
1qu9	3mer	0.90 Å	0.98	0.50 Å	0.98
1urz	3mer	0.98 Å	0.98	0.74 Å	0.98
1z5s	4mer	4.62 Å	0.86	1.68 Å	0.97
1e6v	6mer	2.88 Å	0.93	0.70 Å	0.99
1mtv	6mer	2.08 Å	0.98	0.81 Å	0.99
1tyq	7mer	3.31 Å	0.85	1.31 Å	0.98
1oel	7mer	2.40 Å	0.97	1.03 Å	0.98
1gru	7mer	3.36 Å	0.81	1.15 Å	0.98

6.2 Concluding Remarks

Experimental techniques differ in the structural data that they provide; some produce high resolution data of smaller proteins and some give low resolution of larger proteins and protein complexes. Here we present a method to bridge the gap between these two. Our method constructs protein assemblies based on high resolution data and filters them with low resolution data. A protein is added to the complex regarding an energy score threshold and the structure constructed is tested as to whether it fits into the density map after each addition of a protein to the assembly, forestalling wrong assignments. While this does not reorder the ranking of the solutions, in principle, if none is observed, the threshold can be lowered to gather lower stability interfaces with the complex further stabilized as the assembly grows. Even though our method can produce a number of solutions, EM data directs us to a certain solution model.

We exploited PDB structures in the construction of the assembly if their unbound forms are available. We also modeled protein structures if they are not available in the PDB. Constructing assemblies also from homology models illustrates that our method can work even if the structures of proteins are unknown. However, we have limitations due to protein structures we used. The PDB is not complete, some structures are missing a domain or a part of the protein, and modeling structures is error-prone. Constructing

assemblies of incomplete proteins is harder than predicting their binary interactions. Binary interaction of incomplete structures can be predicted successfully if the missing parts do not cover the binding site; yet, a protein in a complex can interact with more than one protein and it is more likely that the missing part is a segment of one of its binding sites. It is also possible that the segment obstructs a binding site or creates steric hindrance with the protein partners.

We tested our method on a benchmark dataset that covers proteins differing in the number of chains, numbers of residues and SCOP classes, representing a subset of all proteins. We especially selected symmetrical cyclic structures for the benchmark dataset. Symmetrical cyclic proteins are challenging structures for our method, which constructs assemblies based on binary interactions and is not specialized on symmetrical structures. At the addition of the final protein, more than one interface should be considered and a small error in binary predictions would cost to complete the cyclic structure. In two cases (EMDB IDs 5140 and 5505) we could not obtain perfect cyclic structures. However, comparing with the PDB structures or docking into the density maps indicated them as successful predictions. Some methods, like *Multifit*, were designed particularly for symmetrical structures. They can provide models with a proper cyclic shape. However, they may sacrifice in interface predictions while coordinating subunits with a certain angle, like in the case EMD ID 5505. Its model has a better cyclic shape compared to our model yet the interfaces are ‘incorrect’ regarding the IS-score and our model has ‘correct’ interfaces. On the other hand, *Segger* could not do segmentation properly for the complexes that include buried (EMDB ID 1494) or entwined (EMDB ID 2427) subunits. It could predict correct segments in two cases, EMD IDs 2055 and 5140. However, the interfaces obtained after docking proteins into the segments were ‘incorrect’ with respect to IS-score. Methods whose priority is to dock proteins into the density map may create models with high correlation with the density map; however, they may give false positive predictions of protein interfaces, end up with many clashes of the residues and are more likely to

obtain unfavorable assembly structures. Our method depends on interface predictions and the docking the complex into the EM density map, which is likely to be more robust. We have shown that the accuracy of our interface predictions is high and combining it with an exhaustive search docking method led us to obtain accurate prediction of protein assembly structures.

We had better predictions also in the *Multifit* benchmark set. The density maps were very specific and biased for the PDB structures in that set, since they were created from the PDB structures using *Situs* (except for one case, PDB ID 1gru, EMDB ID: 1046). However, in reality, density maps can be very huge and represent larger structures, like in our benchmark set. Docking structures into such larger EM density maps is more challenging compared to docking into the density maps very specific to the protein structures. While using *Multifit* by *Chimera*, we got errors in some cases, yet, could obtain models when we reoriented the structures. Errors may be due to the starting positions of protein structures with respect to the density map; if so, we could obtain better models for different orientation of the structures and density maps. On the other hand, our method does not work for the small interface complexes (less than 10 residues at the interface, as in the example of 2rec in the *Multifit* benchmark set); however, *Multifit* does not depend on the size of the interface and could construct such complexes.

Our methodology can be considered as three prediction processes: prediction of protein structures (if necessary), prediction of protein interactions and docking structures into the EM density maps. Each step is error-prone: the success of the homology modeling tool determines the prediction of protein structures, the accuracy of the protein-protein interaction prediction tool determines the prediction of protein interactions and the success of the docking tool determines the docking of structures into the EM density maps. If the proper template does not present in the homology modeling, we cannot obtain a proper structure; if a similar interface does not present in the PDB and the

template set organization of PRISM was not done correctly, we cannot predict protein interactions accurately. Each of the tools that we used (*I-Tasser*, *PRISM* and *Situs*) has its shortcomings. Even though here we were successful, using them together increases the likelihood of an error. Another limitation comes from the rigid-body treatment of the structures. Exploiting alternative conformations available in the PDB is an alternative to consider protein flexibility [236, 252] but limited with the richness of the PDB. In our method, we add proteins one by one and consider them as rigid bodies for simplicity. However, conformational change upon binding and allostery are known phenomena and affect the whole assembly structure.

Chapter 7

CONCLUSION

In this dissertation we aimed to model multi-molecular protein assemblies. We developed a novel method that can construct symmetric/asymmetric, cyclic/acyclic and hetero-/homo-complexes without a limitation in the number of the proteins. It is based on a knowledge-based approach, which brings applicability in large scale studies. We modified the method to improve our predictions; we exploited different conformations of the protein to consider protein flexibility and EM data to determine the solution structure. Consequently, we introduced two approaches: one is to construct assemblies of the given proteins and the other is to construct assemblies for the EM data by using the given proteins; in addition, assemblies can be modeled using different conformations of the proteins to consider flexibility.

Our methods model structures of protein assemblies, which is important to understand how proteins come together and interact to each other physicochemically. Most, if not all, proteins function when associated in assemblies. Therefore, identifying the protein assembly structures is necessary to understand how biological processes are performed and the mechanism beneath the cellular reactions. Protein structures are also important for drug design. Cavities or binding sites for the drugs, interfaces that a drug binds to and interactions that a drug can disrupt are known through the structural data. Moreover, EM is a foremost experimental method to obtain low resolution data of protein complexes and experimentalists need to verify their density maps by fitting high resolution data into them. EM density maps together with high resolution data help not only to understand the structure more accurately, but also to identify the regions where conformational changes occur and to determine regions of proteins whose structures are uncharacterized or can be characterized under certain conditions. If the PDB does not

include the protein complex, models can be used to dock into EM density maps. As mentioned in **Chapter 6**, Meyerson et al. [309] did not have the structure of the whole protein complex, but we were able to provide a model based on their EM data. EM maps show the outer contour of the complex; however, our model provides also the interfaces in the complex. That was a striking example showing that our method complementarily assists EM data in identifying the structure of multi-molecular protein assemblies.

Limitations of our methods are related to template organization, richness of the PDB, protein flexibility and docking structures into EM density maps. Since our assembly construction method is based on binary protein interactions, improvement in binary interactions would lead to more accurate modeling in protein assemblies. Better organization of the template set would give better binary interactions and then better assembly models. A possible future study might be re-organization of the template set. Our method constructs assemblies by adding a protein at each step. It would be modified in such a way that the method adds any sub-complex to any other one instead of adding a protein to sub-complexes at each step, which would decrease the computation time. For example, a quadromer model could be obtained by adding two dimers, which would end in two iterations of the method. This alternative method would be more effective if the template set includes multimer interfaces instead of binary interfaces. However, that does not guarantee that a better model could not be obtained by adding proteins one by one. Those possibilities would also be checked to be sure and then the computation time would be longer than the current version. That is why; we set up the method in a way that it adds a protein at each step. Moreover, we utilize PDB structures in modeling protein assemblies. In the lack of the structure in the PDB, we first need to model the protein structure, which brings the error possibility in structure modeling. The better structures of individual proteins obtained in homology modeling, the better models of the protein assemblies. Therefore, the materials we use, i.e. protein structures, directly affect our models. Another limiting factor is about the protein flexibility. Since knowledge-based methods treat structures as rigid bodies, we

exploit alternative conformations available in the PDB to consider flexibility. However, we are limited with the richness of the PDB again. An alternative solution may be creating an ensemble of the proteins using molecular dynamics or other techniques. However, that would bring additional computational cost and wrong conformations would result in improper interactions. Furthermore, it is known that protein interaction is very similar to protein folding [17-19, 35]. Therefore, the method can be modified to be applicable in creating different conformations in a similar process to the assembly construction, which would improve predictions of protein interactions. The query protein could be split into sub-structures; this separation could be done through hinge points which are the most stable points of the protein. The same process would be followed as in modeling protein assemblies; however, the inputs would be sub-structures instead of proteins. The template set might be the same but parameters should be arranged with respect to the size of the proteins and sub-structures. A possible future study could be done to model protein interactions and protein assemblies through simultaneous folding and interaction prediction processes. Another point is our method includes external programs to do surface extraction, structural alignment, energy calculation and flexible refinement, and docking structures into EM density maps. Those codes could be re-written and inserted into the main code, which would provide computational advantage.

This dissertation presents methods to construct multi-molecular protein assemblies by overcoming limitations of other approaches, and we have shown that our methods could successfully model protein complexes. Protein interaction networks enriched by structural data shows how proteins interact and our methods will help to construct such structural networks. They will assist to understand biological mechanism and cellular processes and serve to pathway analyses, drug design and functional genomics through a structural perspective.

BIBLIOGRAPHY

1. Nussinov, R., *How do dynamic cellular signals travel long distances?* Molecular Biosystems, 2012. **8**(1): p. 22-26.
2. Grunberg, R., J. Leckner, and M. Nilges, *Complementarity of structure ensembles in protein-protein binding.* Structure, 2004. **12**(12): p. 2125-36.
3. Lawrence, M.C. and P.M. Colman, *Shape complementarity at protein/protein interfaces.* J Mol Biol, 1993. **234**(4): p. 946-50.
4. Lo Conte, L., C. Chothia, and J. Janin, *The atomic structure of protein-protein recognition sites.* J Mol Biol, 1999. **285**(5): p. 2177-98.
5. Sheinerman, F.B., R. Norel, and B. Honig, *Electrostatic aspects of protein-protein interactions.* Curr Opin Struct Biol, 2000. **10**(2): p. 153-9.
6. Zhang, Q., M. Sanner, and A.J. Olson, *Shape complementarity of protein-protein complexes at multiple resolutions.* Proteins, 2009. **75**(2): p. 453-67.
7. Keskin, O., et al., *A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications.* Protein Sci, 2004. **13**(4): p. 1043-55.
8. Gong, S., et al., *PSIbase: a database of Protein Structural Interactome map (PSIMAP).* Bioinformatics, 2005. **21**(10): p. 2541-3.
9. Gadkari, R.A., D. Varughese, and N. Srinivasan, *Recognition of interaction interface residues in low-resolution structures of protein assemblies solely from the positions of C(alpha) atoms.* PLoS One, 2009. **4**(2): p. e4476.
10. Jones, S. and J.M. Thornton, *Principles of protein-protein interactions.* Proc Natl Acad Sci U S A, 1996. **93**(1): p. 13-20.
11. Jones, S., A. Marin, and J.M. Thornton, *Protein domain interfaces: characterization and comparison with oligomeric protein interfaces.* Protein Eng, 2000. **13**(2): p. 77-82.
12. Jones, S. and J.M. Thornton, *Analysis of protein-protein interaction sites using surface patches.* J Mol Biol, 1997. **272**(1): p. 121-32.

13. Keskin, O., et al., *Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach*. Phys Biol, 2005. **2**(2): p. S24-35.
14. Chakrabarti, P. and J. Janin, *Dissecting protein-protein recognition sites*. Proteins, 2002. **47**(3): p. 334-43.
15. Chothia, C. and J. Janin, *Principles of protein-protein recognition*. Nature, 1975. **256**(5520): p. 705-8.
16. Keskin, O. and R. Nussinov, *Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways*. Protein Eng Des Sel, 2005. **18**(1): p. 11-24.
17. Tsai, C.J., et al., *Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect*. Protein Sci, 1997. **6**(1): p. 53-64.
18. Tsai, C.J. and R. Nussinov, *Hydrophobic folding units at protein-protein interfaces: implications to protein folding and to protein-protein association*. Protein Sci, 1997. **6**(7): p. 1426-37.
19. Tsai, C.J., D. Xu, and R. Nussinov, *Structural motifs at protein-protein interfaces: protein cores versus two-state and three-state model complexes*. Protein Sci, 1997. **6**(9): p. 1793-805.
20. Negi, S.S. and W. Braun, *Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces*. J Mol Model, 2007. **13**(11): p. 1157-67.
21. Gruber, J., et al., *Computational analyses of the surface properties of protein-protein interfaces*. Acta Crystallogr D Biol Crystallogr, 2007. **63**(Pt 1): p. 50-7.
22. Headd, J.J., et al., *Protein-protein interfaces: properties, preferences, and projections*. J Proteome Res, 2007. **6**(7): p. 2576-86.
23. Xu, D., C.J. Tsai, and R. Nussinov, *Hydrogen bonds and salt bridges across protein-protein interfaces*. Protein Eng, 1997. **10**(9): p. 999-1012.

24. De, S., et al., *Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different*. BMC Struct Biol, 2005. **5**: p. 15.
25. Teyra, J. and M.T. Pisabarro, *Characterization of interfacial solvent in protein complexes and contribution of wet spots to the interface description*. Proteins, 2007. **67**(4): p. 1087-95.
26. Rodier, F., et al., *Hydration of protein-protein interfaces*. Proteins, 2005. **60**(1): p. 36-45.
27. Kar, G., et al., *Allostery and population shift in drug discovery*. Curr Opin Pharmacol, 2010. **10**(6): p. 715-22.
28. Ozbabacan, S.E., et al., *Transient protein-protein interactions*. Protein Eng Des Sel, 2011. **24**(9): p. 635-48.
29. Guharoy, M. and P. Chakrabarti, *Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions*. Bioinformatics, 2007. **23**(15): p. 1909-18.
30. Mintseris, J. and Z. Weng, *Structure, function, and evolution of transient and obligate protein-protein interactions*. Proc Natl Acad Sci U S A, 2005. **102**(31): p. 10930-5.
31. Ofra, Y. and B. Rost, *Analysing six types of protein-protein interfaces*. J Mol Biol, 2003. **325**(2): p. 377-87.
32. Vacic, V., et al., *Composition Profiler: a tool for discovery and visualization of amino acid composition differences*. BMC Bioinformatics, 2007. **8**: p. 211.
33. Zhanhua, C., et al., *Protein subunit interfaces: heterodimers versus homodimers*. Bioinformation, 2005. **1**(2): p. 28-39.
34. Tuncbag, N., et al., *Architectures and functional coverage of protein-protein interfaces*. J Mol Biol, 2008. **381**(3): p. 785-802.
35. Haliloglu, T., et al., *How similar are protein folding and protein binding nuclei? Examination of vibrational motions of energy hot spots and conserved residues*. Biophys J, 2005. **88**(3): p. 1552-9.

36. Keskin, O. and R. Nussinov, *Similar binding sites and different partners: implications to shared proteins in cellular pathways*. Structure, 2007. **15**(3): p. 341-54.
37. Kuzu, G., et al., *Expanding the conformational selection paradigm in protein-ligand docking*. Methods Mol Biol, 2012. **819**: p. 59-74.
38. Veselovsky, A.V., et al., *Protein-protein interactions: mechanisms and modification by drugs*. J Mol Recognit, 2002. **15**(6): p. 405-22.
39. Tsuchiya, Y., H. Nakamura, and K. Kinoshita, *Discrimination between biological interfaces and crystal-packing contacts*. Advances and Applications in Bioinformatics and Chemistry, 2008. **2008**(1): p. 99 - 113.
40. Kuhn, D., et al., *From the similarity analysis of protein cavities to the functional classification of protein families using cavbase*. J Mol Biol, 2006. **359**(4): p. 1023-44.
41. Binkowski, T.A., S. Naghibzadeh, and J. Liang, *CASTp: Computed Atlas of Surface Topography of proteins*. Nucleic Acids Res, 2003. **31**(13): p. 3352-5.
42. Dundas, J., et al., *CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues*. Nucleic Acids Res, 2006. **34**(Web Server issue): p. W116-8.
43. An, J., M. Totrov, and R. Abagyan, *Pocketome via comprehensive identification and classification of ligand binding envelopes*. Mol Cell Proteomics, 2005. **4**(6): p. 752-61.
44. Li, X., et al., *Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking*. J Mol Biol, 2004. **344**(3): p. 781-95.
45. Veselovsky, A.V. and A.I. Archakov, *Inhibitors of protein-protein interactions as potential drugs*. Curr Comput Aided Drug Des, 2007. **3**(1): p. 51-8.
46. Larsen, T.A., A.J. Olson, and D.S. Goodsell, *Morphology of protein-protein interfaces*. Structure, 1998. **6**(4): p. 421-7.

47. Nguyen, M.N., J.C. Rajapakse, and K. Duan, *Amino acid features for prediction of protein-protein interface residues with support vector machines*. Proceedings of EvoBIO, 2007. **2007**: p. 187 - 196.
48. Zhanhua, C., et al., *Identification of critical heterodimer protein interface parameters by multi-dimensional scaling in euclidian space*. Front Biosci, 2005. **10**: p. 844-52.
49. Von Eichborn, J., S. Gunther, and R. Preissner, *Structural features and evolution of protein-protein interactions*. Genome Inform, 2010. **22**: p. 1-10.
50. Baussand, J. and A.C. Camproux, *Deciphering the shape and deformation of secondary structures through local conformation analysis*. BMC Struct Biol, 2011. **11**: p. 9.
51. Neuvirth, H., R. Raz, and G. Schreiber, *ProMate: a structure based prediction program to identify the location of protein-protein binding sites*. J Mol Biol, 2004. **338**(1): p. 181-99.
52. Horton, N. and M. Lewis, *Calculation of the free energy of association for protein complexes*. Protein Sci, 1992. **1**(1): p. 169-81.
53. Tuncbag, N., et al., *A survey of available tools and web servers for analysis of protein-protein interactions and interfaces*. Brief Bioinform, 2009. **10**(3): p. 217-32.
54. Henrick, K. and J.M. Thornton, *PQS: a protein quaternary structure file server*. Trends Biochem Sci, 1998. **23**(9): p. 358-61.
55. Shakhnovich, B.E., et al., *Functional fingerprints of folds: evidence for correlated structure-function evolution*. J Mol Biol, 2003. **326**(1): p. 1-9.
56. Torrance, J.W., et al., *Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families*. J Mol Biol, 2005. **347**(3): p. 565-81.
57. Caffrey, D.R., et al., *Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?* Protein Sci, 2004. **13**(1): p. 190-202.

58. Valdar, W.S. and J.M. Thornton, *Conservation helps to identify biologically relevant crystal contacts*. J Mol Biol, 2001. **313**(2): p. 399-416.
59. Thorn, K.S. and A.A. Bogan, *ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions*. Bioinformatics, 2001. **17**(3): p. 284-5.
60. Pupko, T., et al., *Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues*. Bioinformatics, 2002. **18 Suppl 1**: p. S71-7.
61. Grishin, N.V. and M.A. Phillips, *The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences*. Protein Sci, 1994. **3**(12): p. 2455-8.
62. Ouzounis, C., et al., *Are binding residues conserved?* Pac Symp Biocomput, 1998: p. 401-12.
63. Bartlett, G.J., et al., *Analysis of catalytic residues in enzyme active sites*. J Mol Biol, 2002. **324**(1): p. 105-21.
64. Ma, B., et al., *Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces*. Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5772-7.
65. Konc, J. and D. Janezic, *Protein-protein binding-sites prediction by protein surface structure conservation*. J Chem Inf Model, 2007. **47**(3): p. 940-4.
66. Watson, J.D., R.A. Laskowski, and J.M. Thornton, *Predicting protein function from sequence and structural data*. Curr Opin Struct Biol, 2005. **15**(3): p. 275-84.
67. Gonzalez-Ruiz, D. and H. Gohlke, *Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding*. Curr Med Chem, 2006. **13**(22): p. 2607-25.
68. Wang, B., et al., *Prediction of protein-protein interacting sites by combining SVM algorithm with Bayesian method*. ICNC 2007, 2007. **2**: p. 329 - 33.

69. Laskowski, R.A., J.D. Watson, and J.M. Thornton, *Protein function prediction using local 3D templates*. J Mol Biol, 2005. **351**(3): p. 614-26.
70. Laskowski, R.A., J.D. Watson, and J.M. Thornton, *ProFunc: a server for predicting protein function from 3D structure*. Nucleic Acids Res, 2005. **33**(Web Server issue): p. W89-93.
71. Kortemme, T. and D. Baker, *A simple physical model for binding energy hot spots in protein-protein complexes*. Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14116-21.
72. Chung, J.L., W. Wang, and P.E. Bourne, *Exploiting sequence and structure homologs to identify protein-protein binding sites*. Proteins, 2006. **62**(3): p. 630-40.
73. Keskin, O., B. Ma, and R. Nussinov, *Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues*. J Mol Biol, 2005. **345**(5): p. 1281-94.
74. Illergard, K., D.H. Ardell, and A. Elofsson, *Structure is three to ten times more conserved than sequence--a study of structural response in protein cores*. Proteins, 2009. **77**(3): p. 499-508.
75. Valas, R.E., S. Yang, and P.E. Bourne, *Nothing about protein structure classification makes sense except in the light of evolution*. Curr Opin Struct Biol, 2009. **19**(3): p. 329-34.
76. Zhang, Q.C., et al., *Protein interface conservation across structure space*. Proc Natl Acad Sci U S A, 2010. **107**(24): p. 10896-901.
77. Tsai, C.J., et al., *A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique*. J Mol Biol, 1996. **260**(4): p. 604-20.
78. Marchler-Bauer, A., et al., *CDD: a curated Entrez database of conserved domain alignments*. Nucleic Acids Research, 2003. **31**(1): p. 383-387.

79. Marchler-Bauer, A., et al., *CDD: a database of conserved domain alignments with links to domain three-dimensional structure*. Nucleic Acids Research, 2002. **30**(1): p. 281-283.
80. Kim, W.K., et al., *The many faces of protein-protein interactions: A compendium of interface geometry*. PLoS Comput Biol, 2006. **2**(9): p. e124.
81. Korkin, D., F.P. Davis, and A. Sali, *Localization of protein-binding sites within families of proteins*. Protein Sci, 2005. **14**(9): p. 2350-60.
82. Stein, A., A. Ceol, and P. Aloy, *3did: identification and classification of domain-based interactions of known three-dimensional structure*. Nucleic Acids Res, 2011. **39**(Database issue): p. D718-23.
83. Aung, Z., et al., *PPiClust: efficient clustering of 3D protein-protein interaction interfaces*. J Bioinform Comput Biol, 2008. **6**(3): p. 415-33.
84. Gao, M. and J. Skolnick, *Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected*. Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(52): p. 22517-22522.
85. Clackson, T. and J.A. Wells, *A hot spot of binding energy in a hormone-receptor interface*. Science, 1995. **267**(5196): p. 383-6.
86. Bogan, A.A. and K.S. Thorn, *Anatomy of hot spots in protein interfaces*. J Mol Biol, 1998. **280**(1): p. 1-9.
87. Shulman-Peleg, A., et al., *Spatial chemical conservation of hot spot interactions in protein-protein complexes*. BMC Biol, 2007. **5**: p. 43.
88. Aytuna, A.S., A. Gursoy, and O. Keskin, *Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces*. Bioinformatics, 2005. **21**(12): p. 2850-2855.
89. Moreira, I.S., P.A. Fernandes, and M.J. Ramos, *Hot spots--a review of the protein-protein interface determinant amino-acid residues*. Proteins, 2007. **68**(4): p. 803-12.

90. Arkin, M.R. and J.A. Wells, *Small-molecule inhibitors of protein-protein interactions: progressing towards the dream*. Nat Rev Drug Discov, 2004. **3**(4): p. 301-17.
91. Hu, Z., et al., *Conservation of polar residues as hot spots at protein interfaces*. Proteins, 2000. **39**(4): p. 331-42.
92. Ahmad, S., et al., *CCRXP: exploring clusters of conserved residues in protein structures*. Nucleic Acids Res, 2010. **38**(Web Server issue): p. W398-401.
93. Halperin, I., H. Wolfson, and R. Nussinov, *Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking*. Structure, 2004. **12**(6): p. 1027-38.
94. Gao, Y., R. Wang, and L. Lai, *Structure-based method for analyzing protein-protein interfaces*. J Mol Model, 2004. **10**(1): p. 44-54.
95. Guerois, R., J.E. Nielsen, and L. Serrano, *Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations*. J Mol Biol, 2002. **320**(2): p. 369-87.
96. Huo, S., I. Massova, and P.A. Kollman, *Computational alanine scanning of the 1:1 human growth hormone-receptor complex*. J Comput Chem, 2002. **23**(1): p. 15-27.
97. Rajamani, D., et al., *Anchor residues in protein-protein interactions*. Proc Natl Acad Sci U S A, 2004. **101**(31): p. 11287-92.
98. Moreira, I.S., P.A. Fernandes, and M.J. Ramos, *Hot spot occlusion from bulk water: a comprehensive study of the complex between the lysozyme HEL and the antibody FVD1.3*. J Phys Chem B, 2007. **111**(10): p. 2697-706.
99. Guney, E., et al., *HotSprint: database of computational hot spots in protein interfaces*. Nucleic Acids Res, 2008. **36**(Database issue): p. D662-6.
100. Ofra, Y. and B. Rost, *Protein-protein interaction hotspots carved into sequences*. PLoS Comput Biol, 2007. **3**(7): p. e119.

101. Acuner Ozbabacan, S.E., et al., *Conformational ensembles, signal transduction and residue hot spots: application to drug discovery*. Curr Opin Drug Discov Devel, 2010. **13**(5): p. 527-37.
102. Brenke, R., et al., *Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques*. Bioinformatics, 2009. **25**(5): p. 621-7.
103. Ma, B. and R. Nussinov, *Trp/Met/Phe hot spots in protein-protein interactions: potential targets in drug design*. Curr Top Med Chem, 2007. **7**(10): p. 999-1005.
104. Ma, B., et al., *The contribution of the Trp/Met/Phe residues to physical interactions of p53 with cellular proteins*. Phys Biol, 2005. **2**(2): p. S56-66.
105. Wells, J.A. and C.L. McClendon, *Reaching for high-hanging fruit in drug discovery at protein-protein interfaces*. Nature, 2007. **450**(7172): p. 1001-9.
106. Bruckner, A., et al., *Yeast two-hybrid, a powerful tool for systems biology*. Int J Mol Sci, 2009. **10**(6): p. 2763-88.
107. Pande, J., M.M. Szewczyk, and A.K. Grover, *Phage display: concept, innovations, applications and future*. Biotechnol Adv, 2010. **28**(6): p. 849-58.
108. Katz, C., et al., *Studying protein-protein interactions using peptide arrays*. Chem Soc Rev, 2011. **40**(5): p. 2131-45.
109. Kim, E.D., et al., *Predicting direct protein interactions from affinity purification mass spectrometry data*. Algorithms Mol Biol, 2010. **5**: p. 34.
110. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update*. Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
111. Ceol, A., et al., *MINT, the molecular interaction database: 2009 update*. Nucleic Acids Res, 2010. **38**(Database issue): p. D532-9.
112. Willis, R.C. and C.W. Hogue, *Searching, viewing, and visualizing data in the Biomolecular Interaction Network Database (BIND)*. Curr Protoc Bioinformatics, 2006. **Chapter 8**: p. Unit 8 9.
113. Stark, C., et al., *The BioGRID Interaction Database: 2011 update*. Nucleic Acids Res, 2011. **39**(Database issue): p. D698-704.

114. Aranda, B., et al., *The IntAct molecular interaction database in 2010*. Nucleic Acids Res, 2010. **38**(Database issue): p. D525-31.
115. Raman, K., *Construction and analysis of protein-protein interaction networks*. Autom Exp, 2010. **2**(1): p. 2.
116. Lievens, S., et al., *Large-scale protein interactome mapping: strategies and opportunities*. Expert Rev Proteomics, 2010. **7**(5): p. 679-90.
117. Bandyopadhyay, S., et al., *A human MAP kinase interactome*. Nature Methods, 2010. **7**(10): p. 801-U50.
118. Shahriari, M., et al., *The Arabidopsis ESCRT protein-protein interaction network*. Plant Molecular Biology, 2011. **76**(1-2): p. 85-96.
119. Acuner Ozbabacan, S.E., et al., *Enriching the human apoptosis pathway by predicting the structures of protein-protein complexes*. J Struct Biol, 2012.
120. Zhang, Y.X., P. Gao, and J.S. Yuan, *Plant Protein-Protein Interaction Network and Interactome*. Current Genomics, 2010. **11**(1): p. 40-46.
121. Sakai, Y., et al., *Protein Interactome Reveals Converging Molecular Pathways Among Autism Disorders*. Science Translational Medicine, 2011. **3**(86).
122. Kar, G., et al., *Human proteome-scale structural modeling of E2-E3 interactions exploiting interface motifs*. J Proteome Res, 2012. **11**(2): p. 1196-207.
123. Nguyen, T.N. and J.A. Goodrich, *Protein-protein interaction assays: eliminating false positive interactions*. Nat Methods, 2006. **3**(2): p. 135-9.
124. You, Z.H., et al., *Increasing Reliability of Protein Interactome by Combining Heterogeneous Data Sources with Weighted Network Topological Metrics*, in *Advanced Intelligent Computing Theories And Applications*. 2010, Springer-Verlag Berlin: Berlin. p. 657-663.
125. Chou, K.C. and Y.D. Cai, *Predicting protein-protein interactions from sequences in a hybridization space*. Journal of Proteome Research, 2006. **5**(2): p. 316-322.
126. Shen, J., et al., *Predicting protein-protein interactions based only on sequences information*. Proc Natl Acad Sci U S A, 2007. **104**(11): p. 4337-41.

127. Dong, Q., S. Zhou, and X. Liu, *Prediction of protein-protein interactions from primary sequences*. International Journal of Data Mining and Bioinformatics, 2010. **4**(2): p. 211-227.
128. Guo, Y., et al., *PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment*. BMC Res Notes, 2010. **3**: p. 145.
129. Shi, M.-G., et al., *Predicting protein-protein interactions from sequence using correlation coefficient and high-quality interaction dataset*. Amino Acids, 2010. **38**(3): p. 891-899.
130. Xia, J.-F., X.-M. Zhao, and D.-S. Huang, *Predicting protein-protein interactions from protein sequences using meta predictor*. Amino Acids, 2010. **39**(5): p. 1595-1599.
131. Venkatesan, K., et al., *An empirical framework for binary interactome mapping*. Nature Methods, 2009. **6**(1): p. 83-90.
132. Fourme, R., et al., *High-resolution structures and properties of biomolecules under high pressures probed by X-ray crystallography*. High Pressure Research, 2010. **30**(1): p. 100-103.
133. Blundell, T.L. and L. Johnson, eds. *Protein Crystallography*. 1976, Academic Press: New York.
134. Bonvin, A., R. Boelens, and R. Kaptein, *NMR analysis of protein interactions*. Current Opinion in Chemical Biology, 2005. **9**(5): p. 501-508.
135. Stahlberg, H. and T. Walz, *Molecular electron microscopy: State of the art and current challenges*. Acs Chemical Biology, 2008. **3**(5): p. 268-281.
136. Joo, C., et al., *Advances in single-molecule fluorescence methods for molecular biology*, in *Annual Review of Biochemistry*. 2008. p. 51-76.
137. Mertens, H.D.T. and D.I. Svergun, *Structural characterization of proteins and complexes using small-angle X-ray solution scattering*. Journal of Structural Biology, 2010. **172**(1): p. 128-141.

138. Berman, H.M., et al., *The Protein Data Bank*. Nucleic Acids Res, 2000. **28**(1): p. 235-42.
139. Dunitz, J.D. and A. Gavezzotti, *Molecular recognition in organic crystals: directed intermolecular bonds or nonlocalized bonding?* Angew Chem Int Ed Engl, 2005. **44**(12): p. 1766-87.
140. Grant, T.D., et al., *Small Angle X-ray Scattering as a Complementary Tool for High-Throughput Structural Studies*. Biopolymers, 2011. **95**(8): p. 517-530.
141. Robinson, C.V., A. Sali, and W. Baumeister, *The molecular sociology of the cell*. Nature, 2007. **450**(7172): p. 973-982.
142. Janin, J., et al., *CAPRI: A Critical Assessment of PRedicted Interactions*. Proteins-Structure Function And Bioinformatics, 2003. **52**(1): p. 2-9.
143. Chen, R., L. Li, and Z.P. Weng, *ZDOCK: An initial-stage protein-docking algorithm*. Proteins-Structure Function and Genetics, 2003. **52**(1): p. 80-87.
144. Kozakov, D., et al., *PIPER: An FFT-based protein docking program with pairwise potentials*. Proteins-Structure Function And Bioinformatics, 2006. **65**(2): p. 392-406.
145. Tovchigrechko, A. and I.A. Vakser, *GRAMM-X public web server for protein-protein docking*. Nucleic Acids Research, 2006. **34**: p. W310-W314.
146. Lyskov, S. and J.J. Gray, *The RosettaDock server for local proteinprotein docking*. Nucleic Acids Research, 2008. **36**: p. W233-W238.
147. Fiorucci, S. and M. Zacharias, *Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT*. Proteins-Structure Function And Bioinformatics, 2010. **78**(15): p. 3131-3139.
148. Mashich, E., R. Nussinov, and H.J. Wolfson, *FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking*. Nucleic Acids Res, 2010. **38**(Web Server issue): p. W457-61.
149. Schneidman-Duhovny, D., et al., *PatchDock and SymmDock: servers for rigid and symmetric docking*. Nucleic Acids Research, 2005. **33**: p. W363-W367.

150. Ben-Zeev, E. and M. Eisenstein, *Weighted geometric docking: Incorporating external information in the rotation-translation scan*. Proteins-Structure Function and Genetics, 2003. **52**(1): p. 24-27.
151. Cheng, T.M.K., T.L. Blundell, and J. Fernandez-Recio, *pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein-protein docking*. Proteins-Structure Function And Bioinformatics, 2007. **68**(2): p. 503-515.
152. De Vries, S.J., M. van Dijk, and A. Bonvin, *The HADDOCK web server for data-driven biomolecular docking*. Nature Protocols, 2010. **5**(5): p. 883-897.
153. Gerega, S.K. and K.M. Downard, *PROXIMO - a new docking algorithm to model protein complexes using data from radical probe mass spectrometry (RP-MS)*. Bioinformatics, 2006. **22**(14): p. 1702-1709.
154. Tjioe, E., et al., *MultiFit: a web server for fitting multiple protein structures into their electron microscopy density map*. Nucleic Acids Research, 2011. **39**: p. W167-W170.
155. Pons, C., et al., *Present and future challenges and limitations in protein-protein docking*. Proteins-Structure Function And Bioinformatics, 2010. **78**(1): p. 95-108.
156. Kastritis, P.L. and A. Bonvin, *Are Scoring Functions in Protein-Protein Docking Ready To Predict Interactomes? Clues from a Novel Binding Affinity Benchmark*. Journal Of Proteome Research, 2010. **9**(5): p. 2216-2225.
157. Gursoy, A., O. Keskin, and R. Nussinov, *Topological properties of protein interaction networks from a structural perspective*. Biochem Soc Trans, 2008. **36**(Pt 6): p. 1398-403.
158. Tuncbag, N., A. Gursoy, and O. Keskin, *Prediction of protein-protein interactions: unifying evolution and structure at protein interfaces*. Physical Biology, 2011. **8**(3).
159. Hue, M., et al., *Large-scale prediction of protein-protein interactions from structures*. BMC Bioinformatics, 2010. **11**.

160. Aloy, P. and R.B. Russell, *Interrogating protein interaction networks through structural biology*. Proceedings Of The National Academy Of Sciences Of The United States Of America, 2002. **99**(9): p. 5896-5901.
161. Aloy, P. and R.B. Russell, *InterPreTS: protein Interaction Prediction through Tertiary Structure*. Bioinformatics, 2003. **19**(1): p. 161-162.
162. Kundrotas, P.J., M.F. Lensink, and E. Alexov, *Homology-based modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles*. International Journal of Biological Macromolecules, 2008. **43**(2): p. 198-208.
163. Kundrotas, P.J., Z.W. Zhu, and I.A. Vakser, *GWIDD: Genome-wide protein docking database*. Nucleic Acids Research, 2010. **38**: p. D513-D517.
164. Talavera, D., R.A. Laskowski, and J.M. Thornton, *WSsas: a web service for the annotation of functional residues through structural homologues*. Bioinformatics, 2009. **25**(9): p. 1192-1194.
165. Shoemaker, B.A., et al., *IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins*. Nucleic Acids Research, 2012. **40**(D1): p. D834-D840.
166. Tyagi, M., et al., *Large-scale mapping of human protein interactome using structural complexes*. EMBO Rep, 2012. **13**(3): p. 266-71.
167. Lu, L., H. Lu, and J. Skolnick, *MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading*. Proteins-Structure Function and Genetics, 2002. **49**(3): p. 350-364.
168. Chen, H.L. and J. Skolnick, *M-TASSER: An algorithm for protein quaternary structure prediction*. Biophysical Journal, 2008. **94**(3): p. 918-928.
169. Shoemaker, B.A., A.R. Panchenko, and S.H. Bryant, *Finding biologically relevant protein domain interactions: Conserved binding mode analysis*. Protein Science, 2006. **15**(2): p. 352-361.
170. Davis, F.P., et al., *Protein complex compositions predicted by structural similarity*. Nucleic Acids Research, 2006. **34**(10): p. 2943-2952.

171. Aloy, P., et al., *Structure-based assembly of protein complexes in yeast*. Science, 2004. **303**(5666): p. 2026-2029.
172. Stein, A., A. Ceol, and P. Aloy, *3did: identification and classification of domain-based interactions of known three-dimensional structure*. Nucleic Acids Research, 2011. **39**: p. D718-D723.
173. Tuncbag, N., et al., *Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM*. Nature Protocols, 2011. **6**(9): p. 1341-1354.
174. Ogmen, U., et al., *PRISM: protein interactions by structural matching*. Nucleic Acids Research, 2005. **33**: p. W331-W336.
175. Gunther, S., et al., *Docking without docking: ISEARCH-Prediction of interactions using known interfaces*. Proteins-Structure Function And Bioinformatics, 2007. **69**(4): p. 839-844.
176. Sinha, R., P.J. Kundrotas, and I.A. Vakser, *Docking by structural similarity at protein-protein interfaces*. Proteins-Structure Function and Bioinformatics, 2010. **78**(15): p. 3235-3241.
177. Tuncbag, N., et al., *Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement*. Proteins, 2012. **80**(4): p. 1239-49.
178. Abbott, A., *Proteomics: The society of proteins*. Nature, 2002. **417**(6892): p. 894-896.
179. Cramer, P., et al., *Structure of eukaryotic RNA polymerases*, in *Annual Review of Biophysics*. 2008. p. 337-352.
180. Schmeing, T.M. and V. Ramakrishnan, *What recent ribosome structures have revealed about the mechanism of translation*. Nature, 2009. **461**(7268): p. 1234-1242.
181. Horwich, A.L. and W.A. Fenton, *Chaperonin-mediated protein folding: using a central cavity to kinetically assist polypeptide chain folding*. Quarterly Reviews of Biophysics, 2009. **42**(2): p. 83-116.

182. Murata, S., H. Yashiroda, and K. Tanaka, *Molecular mechanisms of proteasome assembly*. Nat Rev Mol Cell Biol, 2009. **10**(2): p. 104-15.
183. Aloy, P., et al., *Structure-based assembly of protein complexes in yeast*. Science, 2004. **303**(5666): p. 2026-9.
184. Alber, F., et al., *The molecular architecture of the nuclear pore complex*. Nature, 2007. **450**(7170): p. 695-701.
185. Topf, M., et al., *Refinement of protein structures by iterative comparative modeling and CryoEM density fitting*. J Mol Biol, 2006. **357**(5): p. 1655-68.
186. Topf, M., et al., *Protein structure fitting and refinement guided by cryo-EM density*. Structure, 2008. **16**(2): p. 295-307.
187. Lasker, K., et al., *Inferential Optimization for Simultaneous Fitting of Multiple Components into a CryoEM Map of Their Assembly*. Journal of Molecular Biology, 2009. **388**(1): p. 180-194.
188. Lasker, K., A. Sali, and H.J. Wolfson, *Determining macromolecular assembly structures by molecular docking and fitting into an electron density map*. Proteins-Structure Function and Bioinformatics, 2010. **78**(15): p. 3205-3211.
189. Lindert, S., P.L. Stewart, and J. Meiler, *Hybrid approaches: applying computational methods in cryo-electron microscopy*. Current Opinion in Structural Biology, 2009. **19**(2): p. 218-225.
190. Qian, B., et al., *High-resolution structure prediction and the crystallographic phase problem*. Nature, 2007. **450**(7167): p. 259-64.
191. Taverner, T., et al., *Subunit architecture of intact protein complexes from mass spectrometry and homology modeling*. Acc Chem Res, 2008. **41**(5): p. 617-27.
192. Bowers, P.M., C.E.M. Strauss, and D. Baker, *De novo protein structure determination using sparse NMR data*. Journal of Biomolecular Nmr, 2000. **18**(4): p. 311-318.
193. Raman, S., et al., *NMR Structure Determination for Larger Proteins Using Backbone-Only Data*. Science, 2010. **327**(5968): p. 1014-1018.

194. Taylor, D.J., et al., *Comprehensive molecular structure of the eukaryotic ribosome*. Structure, 2009. **17**(12): p. 1591-604.
195. Lasker, K., et al., *Integrative structure modeling of macromolecular assemblies from proteomics data*. Mol Cell Proteomics, 2010. **9**(8): p. 1689-702.
196. Forster, F., et al., *An atomic model AAA-ATPase/20S core particle sub-complex of the 26S proteasome*. Biochem Biophys Res Commun, 2009. **388**(2): p. 228-33.
197. Tuukkanen, A., et al., *Structural modeling of histone methyltransferase complex Set1C from Saccharomyces cerevisiae using constraint-based docking*. Proteomics, 2010. **10**(23): p. 4186-4195.
198. Pierce, B., W.W. Tong, and Z.P. Weng, *M-ZDOCK: a grid-based approach for C-n symmetric multimer docking*. Bioinformatics, 2005. **21**(8): p. 1472-1478.
199. Pierce, B. and Z.P. Weng, *ZRANK: Reranking protein docking predictions with an optimized energy function*. Proteins-Structure Function and Bioinformatics, 2007. **67**(4): p. 1078-1086.
200. Schneidman-Duhovny, D., et al., *Taking geometry to its edge: Fast unbound rigid (and hinge-bent) docking*. Proteins-Structure Function and Genetics, 2003. **52**(1): p. 107-112.
201. Schneidman-Duhovny, D., et al., *Geometry-based flexible and symmetric protein docking*. Proteins-Structure Function and Bioinformatics, 2005. **60**(2): p. 224-231.
202. Karaca, E., et al., *Building Macromolecular Assemblies by Information-driven Docking*. Molecular & Cellular Proteomics, 2010. **9**(8): p. 1784-1794.
203. Inbar, Y., et al., *Prediction of multimolecular assemblies by multiple docking*. Journal of Molecular Biology, 2005. **349**(2): p. 435-447.
204. Eisenstein, M., et al., *Modeling supra-molecular helices: Extension of the molecular surface recognition algorithm and application to the protein coat of the tobacco mosaic virus*. Journal of Molecular Biology, 1997. **266**(1): p. 135-143.

205. Berchanski, A., D. Segal, and M. Eisenstein, *Modeling oligomers with C-n or D-n symmetry: Application to CAPRI Target 10*. Proteins-Structure Function and Bioinformatics, 2005. **60**(2): p. 202-206.
206. Berchanski, A. and M. Eisenstein, *Construction of molecular assemblies via docking: Modeling of tetramers with D-2 symmetry*. Proteins-Structure Function and Genetics, 2003. **53**(4): p. 817-829.
207. Comeau, S.R. and C.J. Camacho, *Predicting oligomeric assemblies: N-mers a primer*. Journal of Structural Biology, 2005. **150**(3): p. 233-244.
208. Huang, P.S., J.J. Love, and S.L. Mayo, *Adaptation of a fast Fourier transform-based docking algorithm for protein design*. Journal of Computational Chemistry, 2005. **26**(12): p. 1222-1232.
209. Andre, I., et al., *Prediction of the structure of symmetrical protein assemblies*. Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(45): p. 17656-17661.
210. Carragher, B., et al., *Rapid, routine structure determination of macromolecular assemblies using electron microscopy: current progress and further challenges*. Journal of Synchrotron Radiation, 2004. **11**: p. 83-85.
211. Orlova, E.V. and H.R. Saibil, *Structural Analysis of Macromolecular Assemblies by Electron Microscopy*. Chemical Reviews, 2011. **111**(12): p. 7710-7748.
212. Rachel, R., U. Jakubowski, and W. Baumeister, *Electron-microscopy of unstained, freeze-dried macromolecular assemblies*. Journal of Microscopy-Oxford, 1986. **141**: p. 179-191.
213. Hashem, Y., et al., *High-resolution cryo-electron microscopy structure of the Trypanosoma brucei ribosome*. Nature, 2013. **494**(7437): p. 385-389.
214. Zhang, J.J., et al., *Mechanism of folding chamber closure in a group II chaperonin*. Nature, 2010. **463**(7279): p. 379-U130.

215. Subramaniam, S., *Structure of trimeric HIV-1 envelope glycoproteins*. Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(45): p. E4172-E4174.
216. Lawson, C.L., et al., *EMDataBank.org: unified data resource for CryoEM*. Nucleic Acids Research, 2011. **39**: p. D456-D464.
217. Garzon, J.I., et al., *ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage*. Bioinformatics, 2007. **23**(4): p. 427-433.
218. Jiang, W., et al., *Bridging the information gap: Computational tools for intermediate resolution structure interpretation*. Journal of Molecular Biology, 2001. **308**(5): p. 1033-1044.
219. Navaza, J., et al., *On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation*. Acta Crystallographica Section D-Biological Crystallography, 2002. **58**: p. 1820-1825.
220. Siebert, X. and J. Navaza, *UROX 2.0: an interactive tool for fitting atomic models into electron-microscopy reconstructions*. Acta Crystallographica Section D-Biological Crystallography, 2009. **65**: p. 651-658.
221. Roseman, A.M., *Docking structures of domains into maps from cryo-electron microscopy using local correlation*. Acta Crystallographica Section D-Biological Crystallography, 2000. **56**: p. 1332-1340.
222. Rossmann, M.G., *Fitting atomic models into electron-microscopy maps*. Acta Crystallographica Section D-Biological Crystallography, 2000. **56**: p. 1341-1349.
223. Rossmann, M.G., R. Bernal, and S.V. Pletnev, *Combining electron microscopic with X-ray crystallographic structures*. Journal of Structural Biology, 2001. **136**(3): p. 190-200.
224. Wriggers, W., R.A. Milligan, and J.A. McCammon, *Situs: A package for docking crystal structures into low-resolution maps from electron microscopy*. Journal of Structural Biology, 1999. **125**(2-3): p. 185-195.

225. Rusu, M., S. Birmanns, and W. Wriggers, *Biomolecular pleiomorphism probed by spatial interpolation of coarse models*. Bioinformatics, 2008. **24**(21): p. 2460-2466.
226. Kovacs, J.A., et al., *Fast rotational matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom*. Acta Crystallographica Section D-Biological Crystallography, 2003. **59**: p. 1371-1376.
227. Hoang, T.V., X. Cavin, and D.W. Ritchie, *gEMfitter: A highly parallel FFT-based 3D density fitting tool with GPU texture memory acceleration*. Journal of Structural Biology, 2013. **184**(2): p. 348-354.
228. Baker, M.L., et al., *Modeling protein structure at near atomic resolutions with Gorgon*. Journal of Structural Biology, 2011. **174**(2): p. 360-373.
229. Ceulemans, H. and R.B. Russell, *Fast fitting of atomic structures to low-resolution electron density maps by surface overlap maximization*. Journal of Molecular Biology, 2004. **338**(4): p. 783-793.
230. Baker, M.L., et al., *Automated segmentation of molecular subunits in electron cryomicroscopy density maps*. Journal of Structural Biology, 2006. **156**(3): p. 432-441.
231. Burger, V., I. Bahar, and C. Chennubhotla, *Hierarchical Elastic Network Modeling of cryo-EM Data*. Biophysical Journal, 2011. **100**(3): p. 534-534.
232. Volkman, N., *A novel three-dimensional variant of the watershed transform for segmentation of electron density maps*. Journal of Structural Biology, 2002. **138**(1-2): p. 123-129.
233. Pintilie, G.D., et al., *Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions*. Journal of Structural Biology, 2010. **170**(3): p. 427-438.
234. Pintilie, G. and W. Chiu, *Comparison of Segger and other methods for segmentation and rigid-body docking of molecular components in Cryo-EM density maps*. Biopolymers, 2012. **97**(9): p. 742-760.

235. Pettersen, E.F., et al., *UCSF chimera - A visualization system for exploratory research and analysis*. Journal of Computational Chemistry, 2004. **25**(13): p. 1605-1612.
236. Kuzu, G., et al., *Modeling Protein Assemblies in the Proteome*. Molecular & Cellular Proteomics, 2014. **13**(3): p. 887-896.
237. Tsai, C.J., et al., *Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences*. Crit Rev Biochem Mol Biol, 1996. **31**(2): p. 127-52.
238. Keskin, O. and R. Nussinov, *Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways*. Protein Engineering Design & Selection, 2005. **18**(1): p. 11-24.
239. Tuncbag, N., et al., *Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example*. Mol Biosyst, 2009. **5**(12): p. 1770-8.
240. Tuncbag, N., O. Keskin, and A. Gursoy, *HotPoint: hot spot prediction server for protein interfaces*. Nucleic Acids Research, 2010. **38**: p. W402-W406.
241. Hubbard SJ, T.J., 'NACCESS', *Computer Program*, in *Department of Biochemistry and Molecular Biology*. 1993, University College: London.
242. Nussinov, R. and H.J. Wolfson, *Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques*. Proc Natl Acad Sci U S A, 1991. **88**(23): p. 10495-9.
243. Shatsky, M., R. Nussinov, and H.J. Wolfson, *A method for simultaneous alignment of multiple protein structures*. Proteins, 2004. **56**(1): p. 143-56.
244. Mashich, E., R. Nussinov, and H.J. Wolfson, *FiberDock: Flexible induced-fit backbone refinement in molecular docking*. Proteins, 2010. **78**(6): p. 1503-19.
245. Kar, G., et al., *Human Proteome-scale Structural Modeling of E2-E3 Interactions Exploiting Interface Motifs*. Journal of Proteome Research, 2012. **11**(2): p. 1196-1207.

246. Acuner Ozbabacan, S.E., et al., *Enriching the human apoptosis pathway by predicting the structures of protein-protein complexes*. Journal of Structural Biology, 2012(0).
247. Tuncbag, N., et al., *Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example*. Molecular Biosystems, 2009. **5**(12): p. 1770-1778.
248. Kuzu, G., et al., *Constructing structural networks of signaling pathways on the proteome scale*. Current Opinion in Structural Biology, 2012. **22**(3): p. 367-377.
249. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics*. Journal of Molecular Graphics & Modelling, 1996. **14**(1): p. 33-38.
250. Barton, G.J., *SCOP - Structural classification of proteins*. Trends in Biochemical Sciences, 1994. **19**(12): p. 554-555.
251. Shatsky, M., R. Nussinov, and H.J. Wolfson, *A method for simultaneous alignment of multiple protein structures*. Proteins-Structure Function and Bioinformatics, 2004. **56**(1): p. 143-156.
252. Kuzu, G., et al., *Exploiting Conformational Ensembles in Modeling Protein-Protein Interactions on the Proteome Scale*. Journal of Proteome Research, 2013. **12**(6): p. 2641-2653.
253. Hwang, H., et al., *Protein-protein docking benchmark version 4.0*. Proteins-Structure Function and Bioinformatics, 2010. **78**(15): p. 3111-3114.
254. Baglin, T.P., et al., *Crystal structures of native and thrombin-complexed heparin cofactor II reveal a multistep allosteric mechanism*. Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(17): p. 11079-11084.
255. Kar, G., A. Gursoy, and O. Keskin, *Human Cancer Protein-Protein Interaction Network: A Structural Perspective*. Plos Computational Biology, 2009. **5**(12).
256. Gao, M. and J. Skolnick, *New benchmark metrics for protein-protein docking methods*. Proteins-Structure Function and Bioinformatics, 2011. **79**(5): p. 1623-1634.

257. Lensink, M.F., R. Mendez, and S.J. Wodak, *Docking and scoring protein complexes: CAPRI 3rd edition*. Proteins-Structure Function and Bioinformatics, 2007. **69**(4): p. 704-718.
258. Gao, M. and J. Skolnick, *iAlign: a method for the structural comparison of protein-protein interfaces*. Bioinformatics, 2010. **26**(18): p. 2259-2265.
259. Berman, H.M., et al., *The Protein Data Bank*. Acta Crystallographica Section D-Biological Crystallography, 2002. **58**: p. 899-907.
260. Roy, A., A. Kucukural, and Y. Zhang, *I-TASSER: a unified platform for automated protein structure and function prediction*. Nature Protocols, 2010. **5**(4): p. 725-738.
261. Zhang, Y., *I-TASSER server for protein 3D structure prediction*. BMC Bioinformatics, 2008. **9**.
262. Hubbard, T.J.P., et al., *SCOP: A structural classification of proteins database*. Nucleic Acids Research, 1997. **25**(1): p. 236-239.
263. Morais, D.A.d.L., et al., *SUPERFAMILY 1.75 including a domain-centric gene ontology method*. Nucleic Acids Research, 2011. **39**: p. D427-D434.
264. Kolch, W., *Coordinating ERK/MAPK signalling through scaffolds and inhibitors*. Nature Reviews Molecular Cell Biology, 2005. **6**(11): p. 827-837.
265. Claperon, A. and M. Therrien, *KSR and CNK: two scaffolds regulating RAS-mediated RAF activation*. Oncogene, 2007. **26**(22): p. 3143-3158.
266. Brennan, D.F., et al., *A Raf-induced allosteric transition of KSR stimulates phosphorylation of MEK*. Nature, 2011. **472**(7343): p. 366-U134.
267. Shaul, Y.D. and R. Seger, *The MEK/ERK cascade: From signaling specificity to diverse functions*. Biochimica Et Biophysica Acta-Molecular Cell Research, 2007. **1773**(8): p. 1213-1226.
268. Raman, M., W. Chen, and M.H. Cobb, *Differential regulation and properties of MAPKs*. Oncogene, 2007. **26**(22): p. 3100-3112.
269. Pimienta, G. and J. Pascual, *Canonical and alternative MAPK signaling*. Cell Cycle, 2007. **6**(21): p. 2628-2632.

270. Plotnikov, A., et al., *The MAPK cascades: Signaling components, nuclear roles and mechanisms of nuclear translocation*. Biochimica Et Biophysica Acta-Molecular Cell Research, 2011. **1813**(9): p. 1619-1633.
271. Dhillon, A.S., et al., *MAP kinase signalling pathways in cancer*. Oncogene, 2007. **26**(22): p. 3279-3290.
272. Montagut, C. and J. Settleman, *Targeting the RAF-MEK-ERK pathway in cancer therapy*. Cancer Letters, 2009. **283**(2): p. 125-134.
273. Tanti, J.F. and J. Jager, *Cellular mechanisms of insulin resistance: role of stress-regulated serine kinases and insulin receptor substrates (IRS) serine phosphorylation*. Current Opinion in Pharmacology, 2009. **9**(6): p. 753-762.
274. Kim, E.K. and E.J. Choi, *Pathological roles of MAPK signaling pathways in human diseases*. Biochimica Et Biophysica Acta-Molecular Basis of Disease, 2010. **1802**(4): p. 396-405.
275. Tidyman, W.E. and K.A. Rauen, *The RASopathies: developmental syndromes of Ras/MAPK pathway dysregulation*. Current Opinion in Genetics & Development, 2009. **19**(3): p. 230-236.
276. Nussinov, R., B. Ma, and C.J. Tsai, *A broad view of scaffolding suggests that scaffolding proteins can actively control regulation and signaling of multienzyme complexes through allostery*. Biochim Biophys Acta, 2013. **1834**(5): p. 820-9.
277. Morrison, D.K. and R.J. Davis, *Regulation of map kinase signaling modules by scaffold proteins in mammals*. Annual Review of Cell and Developmental Biology, 2003. **19**: p. 91-118.
278. Dhanasekaran, D.N., et al., *Scaffold proteins of MAP-kinase modules*. Oncogene, 2007. **26**(22): p. 3185-3202.
279. Tsai, C.J. and R. Nussinov, *Hydrophobic folding units at protein-protein interfaces: Implications to protein folding and to protein-protein association*. Protein Science, 1997. **6**(7): p. 1426-1437.

280. Inbar, Y., et al., *Protein structure prediction via combinatorial assembly of sub-structural units*. Bioinformatics, 2003. **19**: p. i158-i168.
281. Kundrotas, P.J., et al., *Templates are available to model nearly all complexes of structurally characterized proteins*. Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(24): p. 9438-9441.
282. Bennett, W.S. and R. Huber, *Structural and functional-aspects of domain motions in proteins*. Crc Critical Reviews in Biochemistry, 1984. **15**(4): p. 291-384.
283. Jacobs, D.J., et al., *Protein flexibility predictions using graph theory*. Proteins-Structure Function And Genetics, 2001. **44**(2): p. 150-165.
284. Keskin, O., et al., *Principles of protein-protein interactions: What are the preferred ways for proteins to interact?* Chemical Reviews, 2008. **108**(4): p. 1225-1244.
285. Ma, B.Y., et al., *Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations*. Protein Science, 2002. **11**(2): p. 184-197.
286. Dill, K.A., *Polymer principles and protein folding*. Protein Science, 1999. **8**(6): p. 1166-1180.
287. Frauenfelder, H., S.G. Sligar, and P.G. Wolynes, *THE ENERGY LANDSCAPES AND MOTIONS OF PROTEINS*. Science, 1991. **254**(5038): p. 1598-1603.
288. Ozbabacan, S.E.A., et al., *Conformational ensembles, signal transduction and residue hot spots: Application to drug discovery*. Current Opinion in Drug Discovery & Development, 2010. **13**(5): p. 527-537.
289. Tsai, C.J., A. Del Sol, and R. Nussinov, *Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms*. Molecular Biosystems, 2009. **5**(3): p. 207-216.
290. Ma, B.Y., et al., *Folding funnels and binding mechanisms*. Protein Engineering, 1999. **12**(9): p. 713-720.
291. Tsai, C.J., et al., *Folding funnels, binding funnels, and protein function*. Protein Science, 1999. **8**(6): p. 1181-1190.

292. Tsai, C.J., B.Y. Ma, and R. Nussinov, *Folding and binding cascades: Shifts in energy landscapes*. Proceedings Of The National Academy Of Sciences Of The United States Of America, 1999. **96**(18): p. 9970-9972.
293. Bakan, A. and I. Bahar, *Computational generation inhibitor-bound conformers of p38 map kinase and comparison with experiments*. Pac Symp Biocomput, 2011: p. 181-92.
294. Hwang, H., et al., *Protein-protein docking benchmark version 4.0*. Proteins-Structure Function and Bioinformatics, 2010. **78**(15): p. 3111-3114.
295. Zhu, H.B., et al., *NOXclass: prediction of protein-protein interaction types*. BMC Bioinformatics, 2006. **7**.
296. Bernauer, J., et al., *DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological proteinprotein interactions*. Bioinformatics, 2008. **24**(5): p. 652-658.
297. Duarte, J.M., et al., *Protein interface classification by evolutionary analysis*. BMC Bioinformatics, 2012. **13**: p. 334.
298. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes*. Nucleic Acids Res, 2000. **28**(1): p. 27-30.
299. Tyagi, M., et al., *Large-scale mapping of human protein interactome using structural complexes*. Embo Reports, 2012. **13**(3): p. 266-271.
300. Gursoy, A., O. Keskin, and R. Nussinov, *Topological properties of protein interaction networks from a structural perspective*. Biochemical Society Transactions, 2008. **36**: p. 1398-1403.
301. Robinson, F.L., et al., *Identification of novel point mutations in ERK2 that selectively disrupt binding to MEK1*. Journal of Biological Chemistry, 2002. **277**(17): p. 14844-14852.
302. Schaeffer, H.J., et al., *MP1: A MEK binding partner that enhances enzymatic activation of the MAP kinase cascade*. Science, 1998. **281**(5383): p. 1668-1671.

303. Catalanotti, F., et al., *A Mek1-Mek2 heterodimer determines the strength and duration of the Erk signal*. Nature Structural & Molecular Biology, 2009. **16**(3): p. 294-303.
304. Hauge, C. and M. Frodin, *RSK and MSK in MAP kinase signalling*. Journal of Cell Science, 2006. **119**(15): p. 3021-3023.
305. Carriere, A., et al., *Oncogenic MAPK signaling stimulates mTORC1 activity by promoting RSK-mediated Raptor phosphorylation*. Current Biology, 2008. **18**(17): p. 1269-1277.
306. Kuzu, G., et al., *Exploiting Conformational Ensembles in Modeling Protein-Protein Interactions on the Proteome Scale*. J Proteome Res, 2013.
307. Tuncbag, N., et al., *Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement*. Proteins, 2011.
308. Keskin, O., et al., *Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach*. Physical Biology, 2005. **2**(2): p. S24-S35.
309. Meyerson, J.R., et al., *Molecular structures of trimeric HIV-1 Env in complex with small antibody derivatives*. Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(2): p. 513-518.

VITA

Güray Kuzu was born in Kırklareli, Turkey, on January 27th, 1984. He graduated from Çanakkale Science High School, Çanakkale, in 2002 and received his B.Sc. Degree from Department of Chemical Engineering in Boğaziçi University, Istanbul, in 2007. He received his M.Sc. Degree from the same department in 2009. He worked as a teaching and research assistant at Boğaziçi University, and studied on “Effects of rapamycin and caffeine on TOR signaling pathway in *Saccharomyces cerevisiae*” during his M.Sc. He visited the Department of Biochemistry in the University of Cambridge as a guest researcher in 2008 and 2009. He enrolled to the Ph.D. program at the Department of Computational Sciences and Engineering in Koç University, Istanbul, in 2010, and has worked a teaching and research assistant there. He visited the National Cancer Institute, Frederick, MD, USA, in 2011, the National Institute of Health, Bethesda, MD, USA, in 2012, and the University of California - San Francisco, CA, USA, in 2013, as a guest researcher. He will do his post-doctoral study at Harvard Medical School, Massachusetts General Hospital and Brown University, USA.

Appendix A

Supplementary Data

Table A.1 Difficult cases of the docking benchmark dataset

Case	Molecule 1		Molecule 2		Bound State
1i	3aveA	Ig gamma-1 chain C region	1fnlA	Low affinity immunoglobulin gamma FC region receptor III-B	1e4kA:C
1ii	3aveA	Ig gamma-1 chain C region	1fnlA	Low affinity immunoglobulin gamma FC region receptor III-B	1e4kB:C
2i	2hmiC	FAB fragment of monoclonal antibody 28	1s6pB	POL polyprotein	2hmiC:B
2ii	2hmiD	FAB fragment of monoclonal antibody 28	1s6pB	POL polyprotein	2hmiD:B
3	1cl0A	Thioredoxin reductase	2tirA	Thioredoxin	1f6mA:C
4	1b39A	Cell division protein kinase 2	1fpzF	Cyclin-dependent kinase inhibitor 3	1fq1A:B
5	1x9yA	Cysteine proteinase	1nycA	Cysteine protease inhibitor	1pxvA:C
6	1kwmA	Procarboxypeptidase B	2jtoA	Carboxypeptidase inhibitor	1zliA:B
7	1zm8A	Nuclease	1j57A	Nuclease A inhibitor	2o3bA:B
8	1ijjB	Actin, alpha skeletal muscle	3dniA	Deoxyribonuclease I	1atnA:D
9	1ctqA	Transforming protein P21/H-RAS-1	2ii0A	Son of sevenless homolog 1	1bkdR:S
10	1a6zA	HFE	1cx8A	Transferrin receptor	1de4A:C
11i	1buyA	Erythropoietin	1ernA	Erythropoietin receptor	1eerA:B
11ii	1buyA	Erythropoietin	1ernA	Erythropoietin receptor	1eerA:C
12i	1qfkH	Coagulation factor VIIA heavy chain	1tfhB	Tissue factor	1fakH:T

12ii	1qfkL	Coagulation factor VIIA light chain	1tfhB	Tissue factor	1fakL:T
13	1ijjB	Actin, alpha skeletal muscle	1d0nB	Horse plasma gelsolin	1h1vA:G
14	1qg4A	RAN protein	1f59A	Importin beta-1	1ibrA:B
15	1g0yR	Interleukin-1 receptor, type 1	1ilr1	Interleukin-1 receptor antagonist protein	1iraY:X
16	1qupA	Superoxide dismutase 1 copper chaperone	2jcwA	Cu/Zn superoxide dismutase	1jk9A:B
17	1jmjA	Heparin cofactor II	2cn0H	Prothrombin precursor	1jmoA:H
18i	1jzoA	Thiol:disulfide interchange protein DsbC	1jpeA	DsbD-alpha	1jzdA:C
18ii	1jzoA	Thiol:disulfide interchange protein DsbC	1jpeA	DsbD-alpha	1jzdB:C
19	1hurA	ADP-ribosylation factor 1	1r8mE	Arno	1r8sA:E
20	2fxuA	Actin, alpha skeletal muscle	1ux5A	BNI1 protein	1y64A:B
21	1n0vC	Elongation factor 2	1xk9A	Exotoxin A	1zm4A:B
22	1fchA	Peroxisomal targeting signal 1 receptor	1c44A	Sterol carrier protein 2	2c0lA:B
23	1ywhA	Urokinase plasminogen activator surface receptor	2i9aA	Urokinase-type plasminogen activator	2i9bA:E
24	1j54A	DNA polymerase III, epsilon chain	1se7A	Homologue of the theta subunit of DNA polymerase III	2idoA:B
25	1yzuA	Ras-related protein Rab-21	1txuA	Rab5 GDP/GTP exchange factor	2ot3B:A

Table A.2 Other conformations of given molecules in difficult cases of the benchmark dataset, found using 100% sequence homology

Case	Query Protein	Other Conformations
1i, 1ii	3aveA	2j6eB, 1h3yB, 1h3tB, 2ig2H, 3agvA, 3ry6B, 1mcoH
1i, 1ii	1fnlA	
2i	2hmiC	
2ii	2hmiD	
2i, 2ii	1s6pB	2i5jA, 1hysA, 1dloA, 3kk3B, 3kk1B, 3kqvB
3	1cl0A	
3	2tirA	
4	1b39A	2uueC, 1jsuA, 1pyeA, 1p2aA, 1jvpP, 2g9xC, 1oiqA, 3ddqC, 3dogC, 3bhvC, 3my5C, 1oiuC, 1h25C, 2wmaC, 1h26C, 2wmbC, 3pxfA
4	1fpzF	1fpzABCDE, 1fq1A
5	1x9yA	1y4hA
5	1nycA	
6	1kwmA	
6	2jtoA	2k2xA, 1zliB
7	1zm8A	
7	1j57A	1ktuA, 2o3bB
8	1ijjB	2yjfBE, 2yjeA, 2vypB, 2y83R, 2q31AB, 3tpqE, 3mfpA, 3g37R
8	3dniA	
9	1ctqA	1nvwR, 1iaqA, 3lo5AE, 2x1vA
9	2ii0A	1xd4AB, 3ksyA, 1xd2C, 1xdvAB, 1bkdS
10	1a6zA	
10	1cx8A	
11i, 11ii	1buyA	1cn4C
11i, 11ii	1ernA	1ebaA
12i	1qfkH	1jbuH
12ii	1qfkL	1danL, 1dvaL, 1f7mA, 1fakL, 1ff7A, 1ffmA, 1jbuL, 1kliL, 1kljL, 1o5dL, 1wqvL, 1wssL, 1wtgL, 1wunL, 1wv7L, 1ygcL, 2a2qL, 2aeiL, 2b7dL, 2f9bL, 2flbL, 2flrL, 2zp0L, 2zwlL, 2zzuL, 3elaL
12i, 12ii	1tfhB	1ahwF, 1fakT, 1jpsT, 1o5dT, 1tfhAB, 1w0yT, 1w2kT, 1wqvT, 1wssT, 1wtgT, 1wunT, 1wv7T, 2b7dT, 2f9bT, 2flbT, 2flrT, 2zp0T, 2zwlT, 2zzuT, 3elaT
13	1ijjB	2yjfBE, 2yjeA, 2vypB, 2y83R, 2q31AB, 3tpqE, 3mfpA, 3g37R
13	1d0nB	1rgiG, 2fh3A, 2fh4AB
14	1qg4A	1k5gJ, 1i2mAC, 3ch5A, 3a6pCH, 1ibrAC, 1a2kC, 1qbkC, 3gj0B, 1rrpC
14	1f59A	1ibrB, 2q5dB, 2qnaA

15	1g0yR	1itbB, 1iraY
15	1ilr1	1irpA
16	1qupA	1jk9B
16	2jcwA	
17	1jnjA	1jnjB, 1jmoA
17	2cn0H	3nnpA, 3k65B, 2hntE
18i, 18ii	1jzoA	1jzoB
18i, 18ii	1jpeA	3pfuA, 2hyxC
19	1hurA	1r8qA, 3o47AB, 3o47A, 1r8sA, 1u81A
19	1r8mE	1r8sE
20	2fxuA	2q31AB, 2zwhA, 2y83S, 3tpqE, 3g37P, 2yjfBCE, 2vypB, 2yjeA
20	1ux5A	1y64B
21	1n0vC	2p8wT, 1u2rA, 3b8hE, 1zm3C, 3dnyT, 1s1hT, 2npfB
21	1xk9A	1dmaB, 1ikqA, 1aerB
22	1fchA	1fchB, 2j9qAB, 2c0mC
22	1c44A	1qndA, 2c0lB
23	1ywhA	1ywhCEGIKMO, 3bt1U, 2fd6U, 2i9bEGH
23	2i9aA	2i9aBCD, 2i9bABCD, 3bt2A, 1urkA
24	1j54A	
24	1se7A	2idoB
25	1yzuA	1z0iA, 2ot3B, 1yztAB
25	1txuA	

Table A.3 Construction of assemblies step by step starting from their components

Assembly	Structure 1	Structure 2	Template	Energy unit	Prediction
2e86	2e86A	2e86A	2avfEF	-234.68	2e86_step0
	2e86_step0	2e86A	2avfEF	-458.26	2e86_step1
1eer	1eerA	1eerB	1eerAB	-129.35	1eer_step0
	1eer_step0	1eerC	1eerAC	-73.67	1eer_step1
1gp2	1gp2B	1gp2G	1xhmAB	-248.31	1gp2_step0
	1gp2_step0	1gp2A	1gotAB	-112.02	1gp2_step1

Table A.4 Structurally different predictions constructed from bound forms

Clusters	PDB structure	Structurally different predictions	Energy unit	RMSD (Å)
A1	2e86	9	-692.94	0.531
A2	2e86	9	-238.73	15.216
A3	2e86	9	-226.1	8.969
A4	2e86	9	-238.34	15.132
A5	2e86	9	-261.47	15.164
A6	2e86	9	-271.05	9.124
A7	2e86	9	-270.94	15.213
A8	2e86	9	-258.36	9.128
A9	2e86	9	-30.45	18.043
B1	1eer	6	-203.02	1.094
B2	1eer	6	-171.78	1.217
B3	1eer	6	-166.33	2.311
B4	1eer	6	-87.66	12.824
B5	1eer	6	-130.03	2.895
B6	1eer	6	-40.12	13.749
C1	1gp2	1	-360.33	0.788

Table A.5 Construction of assemblies step by step starting from the unbound forms

Assembly	Structure 1	Structure 2	Template	Energy unit	Prediction
2e86	1et5	1et5	2avfEF	-247.69	2e86_step0
	2e86_step0	1et5	2avfEF	-447.01	2e86_step1
1eer	1buy	1ernA	1eerAB	-27.67	1eer_step0
	1eer_step0	1ernA	1eerAC	-18.40	1eer_step1
1gp2	1gia	1tbgA	1gotAB	-25.66	1gp2_step0
	1gp2_step0	1tbgE	1xhmAB	-253.72	1gp2_step1
1ado	1ald	1ald	1zahBC	-88.97	1ado_step0
	1ald	1ado_step0	1xdlBC	-43.67	1ado_step1
	1ald	1ado_step1	1zahBC	-192.15	1ado_step2
1akj	2clrA	2clrB	1tmcAB	-128.70	1akj_step0
	1akj_step0	1cd8	1nezAH	-37.26	1akj_step1
	1cd8	1akj_step1	1nezGH	-106.74	1akj_step2
1z0k	2bmeA	2bmeA	2bmeAC	-33.94	1z0k_step0
	1z0k_step0	1yzm	1z0jAB	-39.33	1z0k_step1
	1z0k_step1	1yzm	1z0jAB	-31.31	1z0k_step2
1b0c	9pti	9pti	1p2mBD	-22.25	1b0c_step0
	1b0c_step0	9pti	2hexBC	-25.42	1b0c_step1
	1b0c_step1	9pti	2hexBC	-23.81	1b0c_step2
	1b0c_step2	9pti	1bhfAF	-37.64	1b0c_step3
1wnr	3nx6	3nx6	1wf4st	-49.01	1wnr_step0
	3nx6	1wnr_step0	1wf4st	-87.85	1wnr_step1
	3nx6	1wnr_step1	1wf4st	-82.99	1wnr_step2
	3nx6	1wnr_step2	1wf4st	-58.26	1wnr_step3
	1wnr_step3	3nx6	1wf4st	-76.51	1wnr_step4
	1wnr_step4	3nx6	1wf4st	-72.18	1wnr_step5

Table A.6 Structurally different predictions constructed from unbound forms

Cluster	PDB ID	Unbound forms	Structurally different predictions	Energy unit	RMSD (Å)
A1	2e86	1et5A x3	1	-694.70	0.521
B1	1eer	1buy, 1ernA x2	9	-75.58	5.561
B2	1eer	1buy, 1ernA x2	9	-85.43	7.740
B3	1eer	1buy, 1ernA x2	9	-36.03	13.824
B4	1eer	1buy, 1ernA x2	9	-35.62	13.351
B5	1eer	1buy, 1ernA x2	9	-34.79	13.321
B6	1eer	1buy, 1ernA x2	9	-41.36	12.917
B7	1eer	1buy, 1ernA x2	9	-43.32	13.615
B8	1eer	1buy, 1ernA x2	9	-27.35	13.851
B9	1eer	1buy, 1ernA x2	9	-32.80	12.956
C1	1gp2	1gia, 1tbgA, 1tbgE	1	-279.38	3.566
D1	1ado	1aldA x4	15	-324.79	2.790
D2	1ado	1aldA x4	15	-211.63	3.724
D3	1ado	1aldA x4	15	-190.31	5.138
D4	1ado	1aldA x4	15	-165.13	4.317
D5	1ado	1aldA x4	15	-159.27	6.488
D6	1ado	1aldA x4	15	-179.60	7.454
D7	1ado	1aldA x4	15	-120.90	8.779
D8	1ado	1aldA x4	15	-127.20	9.891
D9	1ado	1aldA x4	15	-144.36	11.462
D10	1ado	1aldA x4	15	-142.43	13.424
D11	1ado	1aldA x4	15	-169.77	13.281
D12	1ado	1aldA x4	15	-143.63	14.173
D13	1ado	1aldA x4	15	-128.81	14.258
D14	1ado	1aldA x4	15	-122.13	19.225
D15	1ado	1aldA x4	15	-85.23	22.583
E1	1akj	2clrA, 2clrB, 1cd8A x2	5	-272.70	2.572
E2	1akj	2clrA, 2clrB, 1cd8A x2	5	-186.69	4.040
E3	1akj	2clrA, 2clrB, 1cd8A x2	5	-86.49	3.612
E4	1akj	2clrA, 2clrB, 1cd8A x2	5	-168.45	3.676
E5	1akj	2clrA, 2clrB, 1cd8A x2	5	-141.21	3.272
F1	1z0k	2bmeA x2, 1yzmA x2	13	-104.58	3.072
F2	1z0k	2bmeA x2, 1yzmA x3	13	-122.78	4.149
F3	1z0k	2bmeA x2, 1yzmA x4	13	-98.80	5.128

F4	1z0k	2bmeA x2, 1yzmA x5	13	-106.29	6.777
F5	1z0k	2bmeA x2, 1yzmA x6	13	-117.62	3.886
F6	1z0k	2bmeA x2, 1yzmA x7	13	-90.91	4.491
F7	1z0k	2bmeA x2, 1yzmA x8	13	-89.13	9.947
F8	1z0k	2bmeA x2, 1yzmA x9	13	-103.52	4.503
F9	1z0k	2bmeA x2, 1yzmA x10	13	-115.62	12.204
F10	1z0k	2bmeA x2, 1yzmA x11	13	-97.30	13.054
F11	1z0k	2bmeA x2, 1yzmA x12	13	-99.47	15.390
F12	1z0k	2bmeA x2, 1yzmA x13	13	-65.24	15.830
F13	1z0k	2bmeA x2, 1yzmA x14	13	-81.14	19.942
G1	1b0c	9ptiA x5	23	-109.12	5.148
G2	1b0c	9ptiA x5	23	-111.35	5.945
G3	1b0c	9ptiA x5	23	-122.12	6.894
G4	1b0c	9ptiA x5	23	-143.59	7.701
G5	1b0c	9ptiA x5	23	-135.93	8.448
G6	1b0c	9ptiA x5	23	-131.79	17.419
G7	1b0c	9ptiA x5	23	-120.54	10.837
G8	1b0c	9ptiA x5	23	-112.96	11.824
G9	1b0c	9ptiA x5	23	-112.54	8.534
G10	1b0c	9ptiA x5	23	-106.85	13.505
G11	1b0c	9ptiA x5	23	-118.58	10.51
G12	1b0c	9ptiA x5	23	-112.29	15.617
G13	1b0c	9ptiA x5	23	-118.80	17.26
G14	1b0c	9ptiA x5	23	-113.11	17.596
G15	1b0c	9ptiA x5	23	-114.71	19.16
G16	1b0c	9ptiA x5	23	-97.63	19.76
G17	1b0c	9ptiA x5	23	-98.84	21.143
G18	1b0c	9ptiA x5	23	-117.45	22.571
G19	1b0c	9ptiA x5	23	-99.72	23.287
G20	1b0c	9ptiA x5	23	-98.97	24.27
G21	1b0c	9ptiA x5	23	-59.04	25.457
G22	1b0c	9ptiA x5	23	-76.73	28.604
G23	1b0c	9ptiA x5	23	-91.99	30.266
H1	1wnr	3nx6A x7	13	-426.80	4.599
H2	1wnr	3nx6A x7	13	-422.69	4.988
H3	1wnr	3nx6A x7	13	-434.98	6.198
H4	1wnr	3nx6A x7	13	-431.48	7.365
H5	1wnr	3nx6A x7	13	-484.95	7.747

H6	1wnr	3nx6A x7	13	-392.36	9.338
H7	1wnr	3nx6A x7	13	-392.74	9.971
H8	1wnr	3nx6A x7	13	-383.32	8.638
H9	1wnr	3nx6A x7	13	-350.15	7.089
H10	1wnr	3nx6A x7	13	-304.07	14.261
H11	1wnr	3nx6A x7	13	-351.21	5.458
H12	1wnr	3nx6A x7	13	-258.57	17.671
H13	1wnr	3nx6A x7	13	-337.93	18.905

Table A.7 Construction of complexes of MEK, KSR, Raf, ERK proteins step by step

Assembly	Structure 1	Structure 2	Template	Energy (energy unit)	Prediction
MEK-KSR-KSR-MEK (MKKM)	2y4iB	2y4iB	1uwhAB	-9.62	MKKM_step0
	MKKM_step0	3dy7	2y4iBC	-41.23	MKKM_step1
	MKKM_step1	3dy7	2y4iBC	-53.71	MKKM_step2
MEK-KSR-Raf (MKR)	1uwhAB	2y4iB	1uwhAB	-25.37	MKR_step0
	MKR_step0	3dy7	2y4iBC	-62.61	MKR_step1
MEK-KSR-ERK-Raf (MKER)	1uwhAB	2y4iB	1uwhAB	-25.37	MKER_step0
	1wzy	MKER_step0	1qjcAB	-47.91	MKER_step1
	MKER_step1	3dy7	2y4iBC	-65.78	MKER_step2

Table A.8 Sequence similarity of bound and unbound forms given in the benchmark dataset

Case	Unbound form	Bound from	Sequence Identity (%)	Case	Unbound form	Bound from	Sequence Identity (%)
1i	3aveA	1e4kA	100	12ii	1qfkL	1fakL	100
	1fnlA	1e4kC	100		1tfhB	1fakT	100
1iii	3aveA	1e4kB	100	13	1ijjB	1h1vA	100
	1fnlA	1e4kC	100		1d0nB	1h1vG	100
2i	2hmiC	2hmiC	100	14	1qg4A	1ibrA	100
	1s6pB	2hmiB	100		1f59A	1ibrB	100
2ii	2hmiD	2hmiD	100	15	1g0yR	1iraY	100
	1s6pB	2hmiB	100		1ilr1	1iraX	100

3	1cl0A	1f6mA	99.7	16	1qupA	1jk9B	99.3
	2tirA	1f6mC	99.1		2jcwA	1jk9A	100
4	1b39A	1fq1B	100	17	1jmjA	1jmoA	100
	1fpzF	1fq1A	100		2cn0H	1jmoH	100
5	1x9yA	1pxvA	99.4	18i	1jzoA	1jzdA	100
	1nycA	1pxvC	100		1jpeA	1jzdC	99.2
6	1kwmA	1zliA	100	18ii	1jzoA	1jzdB	100
	2jtoA	1zliB	100		1jpeA	1jzdC	99.2
7	1zm8A	2o3bA	100	19	1hurA	1r8sA	100
	1j57A	2o3bB	100		1r8mE	1r8sE	100
8	1ijjB	1atnA	100	20	2fxuA	1y64A	100
	3dniA	1atnD	99.6		1ux5A	1y64B	100
9	1ctqA	1bkdR	100	21	1n0vC	1zm4A	100
	2ii0A	1bkdS	100		1xk9A	1zm4B	100
10	1a6zA	1de4A	100	22	1fchA	2c0lA	100
	1cx8A	1de4C	100		1c44A	2c0lB	100
11i	1buyA	1eerA	98.8	23	1ywhA	2i9bE	100
	1ernA	1eerB	99.5		2i9aA	2i9bA	100
11ii	1buyA	1eerA	98.8	24	1j54A	2idoA	100
	1ernA	1eerC	99.5		1se7A	2idoB	100
12i	1qfkH	1fakH	100	25	1yzuA	2ot3B	100
	1tfhB	1fakT	100		1txuA	2ot3A	99.3

Table A.9 Additional conformations of the query molecules found when the sequence homology is lowered from 100% to 95%

Case	Query Protein	Additional Conformations
1ii	3aveA	2rcjLK
1ii	1fnlA	3sgjC, 3sgkC, 3ay4C
2i	2hmiC	-
2ii	2hmiD	-
2i, 2ii	1s6pB	1hmvBDFH, 1eetA, 3nbpB, 3dlgB, 1c0tAB, 1ep4AB, 1rt2B, 1tl1A, 1rtiA, 3ig1A, 3qo9A, 3lp0B, 3hvtB, 1lweA, 1lwfB
3	1cl0A	1f6mA
3	2tirA	-
4	1b39A	2iw9C, 2iw8C
4	1fpzF	-
6	1kwmA	-

6	2jtoA	2k2zA
7	1zm8A	-
7	1j57A	-
10	1a6zA	-
10	1cx8A	-
14	1qg4A	3ea5A, 3gjxFC
14	1f59A	-
17	1jmjA	-
17	2cn0H	1mkxK
18i	1jzoA	-
18i	1jpeA	1jzdCD, 1vrsABC
20	2fxuA	3lueA, 3byhA
20	1ux5A	1ux4B
22	1fchA	-
22	1c44A	-
23	1ywhA	-
23	2i9aA	-
24	1j54A	-
24	1se7A	-
25	1yzuA	-
25	1txuA	2ot3A

Table A.10 Construction of assemblies step by step considering alternative forms of the unbound forms

Assembly	Structure 1	Structure 2	Template	Energy (energy unit)	Prediction
1eer	1cn4C	1ebaAB	1eerAB	-61.11	1eer_step0
	1eer_step1	1ebaAB	1eerAC	-61.08	1eer_step1
1gp2	1gg2A	1tbgABCD	1gotAB	-95.90	1gp2_step0
	1gp2_step1	1tbgEFGH	1tbgBF	-280.28	1gp2_step1

Table A.11 Evaluation of interfaces predicted by our method

PDB ID	PDB, interface	Model	Model, interface	IS-score	Result
3cre	AB	3cre-model	AB	0.83	correct
	BC		BC	0.80	correct
4a6j	AC	4a6j-3mer-model	AB	0.21	correct
			BC	0.14	acceptable
		4a6j-4mer-model	AB	0.14	acceptable
			BC	0.20	correct
			CD	0.12	acceptable
4aod	AB	4aod-model	AB	0.38	correct
			BC	0.34	correct
			CD	0.38	correct
			DE	0.33	correct
			EA	0.35	correct
3se9	GH	3se9-model	GH	0.14	acceptable
	HL		HL	0.33	correct
3iyf	AB	3iyf-model	AH	0.24	correct
			HG	-	-
			GD	0.37	correct
			DF	0.27	correct
			FE	0.33	correct
			EC	0.26	correct
			CB	0.37	correct
			BA	0.28	correct
1e9r	AB	1e9r-model	AB	0.70	correct
			BC	0.69	correct
			CD	0.71	correct

			DE	0.71	correct
			EF	0.67	correct
			FA	-	-
3dvl	AB	3dvl-model	AB	0.41	correct
			BC	0.42	correct
			CD	0.39	correct
			DE	0.46	correct
			EF	0.51	correct
			FA	0.41	correct
1gc1	GH	1gc1-model	GB	0.1925	correct
	HL		BA	0.5326	correct
	GC		GC	0.7067	correct
	GC	m36-model	AC	0.7701	correct

Table A.12 Evaluation of interfaces predicted by Multifit

PDB ID	PDB, interface	Model	Model, interface	IS-score	Result
3cre	AB	3cre-model	AB	0.07	incorrect
	BC		BC	0.07	incorrect
4a6j	AC	4a6j-3mer-model	AB	0.10	incorrect
			BC	0.23	correct
3se9	GH	3se9-model	GH	-	-
	HL		HL	0.12	acceptable
1e9r	AB	1e9r-model	AB	0.08	incorrect
			BC	0.08	incorrect
			CD	0.08	incorrect
			DE	0.08	incorrect
			EF	0.08	incorrect
			FA	0.08	incorrect
3dvl	AB	3dvl-model	AB	0.36	correct
			BC	0.36	correct
			CD	0.36	correct
			DE	0.36	correct
			EF	0.36	correct
			FA	0.36	correct

Table A.13 Evaluation of interfaces predicted by Segger

PDB ID	PDB, interface	Model	Model, interface	IS-score	Result
4aod	AB	4aod-model	AB	0.08	incorrect
			BC	0.08	incorrect
			CD	0.08	incorrect
			DE	0.08	incorrect
			EA	0.08	incorrect
3iyf	AB	3iyf-model	AB	-	-
			BC	-	-
			CD	0.33	correct
			DE	0.34	correct
			EF	-	-
			FG	-	-
			GH	-	-
			HA	-	-

Table A.14 Evaluation of interfaces in the Multifit benchmark set

PDB ID	Interface	Our predictions		Multifit predictions	
		IS-score	Result	IS-score	Result
1qu9	AB	0.84	correct	0.73	correct
	AC	0.84	correct	0.73	correct
	BC	0.79	correct	0.73	correct
1urz	AB	0.76	correct	0.85	correct
	AC	0.69	correct	0.80	correct
	BC	0.70	correct	0.83	correct
1z5s	AB	0.27	correct	-	
	AC	0.60	correct	0.11	incorrect
	AD	0.88	correct	0.04	incorrect
	BD	0.34	correct	-	
1e6v	AB	0.84	correct	0.15	acceptable
	AC	0.84	correct	0.23	correct
	AD	0.89	correct	0.34	correct
	AE	0.85	correct	0.09	incorrect
	BC	0.90	correct	0.34	correct
	BD	0.86	correct	0.55	correct

	BE	0.82	correct	-	
	DE	0.87	correct	0.09	incorrect
	DF	0.85	correct	-	
	EF	0.91	correct	-	
1mty	BC	0.84	correct	0.74	correct
	BD	0.92	correct	0.66	correct
	BG	0.82	correct	0.49	correct
	CE	0.89	correct	0.49	correct
	CH	0.82	correct	0.46	correct
	DE	0.85	correct	0.28	correct
	DG	0.86	correct	0.39	correct
	EH	0.89	correct	0.43	correct
1tyq	AB	0.48	correct	-	
	AD	0.84	correct	0.21	correct
	AE	0.76	correct	-	
	BF	0.86	correct	-	
	BG	0.55	correct	-	
	CF	0.77	correct	-	
	CG	0.64	correct	-	
	DF	0.87	correct	0.10	incorrect
1tyq	FG	0.72	correct	-	
1oel	AB	0.78	correct	0.51	correct
	BC	0.79	correct	0.56	correct
	CD	0.66	correct	0.52	correct
	DE	0.88	correct	0.51	correct
	EF	0.80	correct	0.54	correct
	FG	0.75	correct	0.52	correct
	GA	0.80	correct	0.51	correct
1gru	AB	0.80	correct	0.35	correct
	BC	0.81	correct	0.34	correct
	CD	0.78	correct	0.33	correct
	DE	0.79	correct	0.33	correct
	EF	0.71	correct	0.33	correct
	FG	0.75	correct	0.33	correct
	GA	0.49	correct	0.35	correct