2017

Structure-based Prediction of Protein-protein Interaction Networks across Proteomes

Surabhi Maheshwari
Louisiana State University and Agricultural and Mechanical College, surabhimaheshwari89@gmail.com

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

Part of the Life Sciences Commons

Recommended Citation
https://digitalcommons.lsu.edu/gradschool_dissertations/4361

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
STRUCTURE-BASED PREDICTION OF PROTEIN-PROTEIN INTERACTION NETWORKS ACROSS PROTEOMES

A Dissertation

Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

The Department of Biological Sciences

by

Surabhi Maheshwari
BS, SGSITS, 2006
MS, University of Georgia, 2010
August 2012
ACKNOWLEDGEMENT

The process of earning a doctorate and writing a dissertation is long and arduous and it is certainly not done single-handedly. First and foremost, I would like to thank my parents who have given me their unequivocal support throughout and great patience, as always. Without their constant support, encouragement and understanding, it would not have been possible for me to achieve my educational goals. Many thanks to my husband, for being my support system and keeping me sane especially during the last year of grad school.

I would like to express the deepest gratitude to my committee chair, Dr. Michal Brylinski who generously gave his time, expertise and support to all my endeavors. I will always be appreciative of the countless opportunities and experiences I have had as a result of Michal’s support and guidance - he has been a wonderful mentor and never failed to get the best out of me. I would also like to thank the other members of my dissertation committee: Dr. Marcia Newcomer, Dr. Grover Waldrop and Dr. Luigi Marzili. Their insight, feedback and indispensable advice was influential and essential throughout the dissertation writing process.
TABLE OF CONTENTS

ACKNOWLEDGEMENT.........................................................................................................................ii

ABSTRACT..................................................................................................................................................v

CHAPTER 1: INTRODUCTION
  PROTEIN-PROTEIN INTERACTIONS (PPIs) ..................................................................................1
  COMPUTATIONAL METHODS FOR THE PREDICTION OF PPIS ..............................................3
  MODELING OF PROTEIN-PROTEIN INTERACTIONS ...............................................................4
  DISSERTATION SYNOPSIS ..............................................................................................................5
  REFERENCES...........................................................................................................................................8

CHAPTER 2: BINDING SITE PREDICTION
  INTRODUCTION ..............................................................................................................................10
  MATERIALS AND METHODS .........................................................................................................13
  RESULTS AND DISCUSSION ..........................................................................................................23
  CONCLUSION .................................................................................................................................45
  SUPPLEMENTARY INFORMATION .................................................................................................47
  REFERENCES .....................................................................................................................................50

CHAPTER 3: COMPARATIVE ANALYSIS OF BINDING SITE PREDICTION METHODS
  INTRODUCTION ..............................................................................................................................56
  MATERIALS AND METHODS .........................................................................................................71
  RESULTS AND DISCUSSION ..........................................................................................................74
  FUTURE WORK ...............................................................................................................................80
  REFERENCES .....................................................................................................................................80

CHAPTER 4: DOCKING AND RERANKING PROTEIN COMPLEXES
  INTRODUCTION ..............................................................................................................................88
  MATERIALS AND METHODS .........................................................................................................93
  RESULTS ........................................................................................................................................100
  DISCUSSION ...................................................................................................................................112
  CONCLUSION ...................................................................................................................................116
  REFERENCES .....................................................................................................................................117

CHAPTER 5: PPI PREDICTION AND MODELING PIPELINE
  INTRODUCTION ..............................................................................................................................123
  MATERIALS AND METHODS .........................................................................................................127
  RESULTS AND DISCUSSION ..........................................................................................................133
  CONCLUSION ...................................................................................................................................149
  REFERENCES .....................................................................................................................................150
CHAPTER 6: PPI PREDICTION FOR HUMAN PATHWAYS
  INTRODUCTION..................................................................................................................155
  MATERIALS AND METHODS.............................................................................................155
  RESULTS AND DISCUSSION..............................................................................................157
  CONCLUSION....................................................................................................................164
  REFERENCES......................................................................................................................165

CHAPTER 7: CONCLUSIONS.................................................................................................167
  REFERENCES......................................................................................................................169

APPENDIX
  A. LETTER OF PERMISSION TO REPRINT CHAPTER 2 ..................................................171
  B. LETTER OF PERMISSION TO REPRINT CHAPTER 3 ..................................................172
  C. LETTER OF PERMISSION TO REPRINT CHAPTER 1 AND 7 .....................................173
  D. LETTER OF PERMISSION TO REPRINT CHAPTER 4 ..................................................174

VITA........................................................................................................................................175
ABSTRACT

Protein-protein interactions (PPIs) orchestrate virtually all cellular processes, therefore, their exhaustive exploration is essential for the comprehensive understanding of cellular networks. Significant efforts have been devoted to expand the coverage of the proteome-wide interaction space at molecular level. A number of experimental techniques have been developed to discover PPIs, however these approaches have some limitations such as the high costs and long times of experiments, noisy data sets, and often high false positive rate and inter-study discrepancies. Given experimental limitations, computational methods are increasingly becoming important for detection and structural characterization of PPIs. In that regard, we have developed a novel pipeline for high-throughput PPI prediction based on all-to-all rigid body docking of protein structures. We focus on two questions, ‘how do proteins interact?’ and ‘which proteins interact?’. The method combines molecular modeling, structural bioinformatics, machine learning, and functional annotation data to answer these questions and it can be used for genome-wide molecular reconstruction of protein-protein interaction networks. As a proof of concept, 61,913 protein-protein interactions were confidently predicted and modeled for the proteome of E. coli. Further, we validated our method against a few human pathways. The modeling protocol described in this communication can be applied to detect protein-protein interactions in other organisms as well as to construct dimer structures and estimate the confidence of protein interactions experimentally identified with high-throughput techniques.
CHAPTER 1: INTRODUCTION

PROTEIN-PROTEIN INTERACTIONS (PPIs)

Living systems are made up of several molecular entities such as DNA, RNA and proteins, interactions among which leads to complex properties of life that are not attainable by the individual molecules. Understanding biological systems requires detailed knowledge of how these interactions govern cellular events at the molecular level. Many essential cellular processes are mediated by protein-protein interactions, from signal transduction in cellular networks to forming molecular machines like ribosomes, or from catalyzing enzymatic reaction to transcription regulation [1]. Given the unquestionable role of proteins in a cell, significant efforts have been devoted to detect and characterize protein-protein (PPI) sites which is very useful for: (1) elucidating mechanisms that underlie biological function (2) assigning function of unknown proteins based on their interacting partners (3) identification of druggable targets (4) engineering and modification of protein activity (5) interpreting the impact of mutations and allelic variations (among other applications). To achieve these goals, the key steps are to identify which proteins interact with each other and how the interaction takes place. A full understanding of how proteins interact comes only from three-dimensional (3D) structures, as they provide critical atomic details about the binding.

Determination of 3D structures of proteins has been a challenging job from the beginning. The first X-ray structure determined in 1958 took decades to solve [2]. However, the situation has progressed remarkably since then. Upon availability of sufficient data, individual protein structures can now be determined in a matter of days. Current procedures
for overexpression and purification of proteins can supply sufficient material for structural studies of single proteins, but the task is much more complicated for bigger complexes. The reason is quite simple, complex assemblies necessitate precise control and timing in the cell which is very hard to reproduce in a laboratory setting. Also, complex assemblies are not very well understood; therefore, it usually involves years of tinkering to obtain ample quantity of samples and grow crystals which will diffract at a high resolution - a task which is more arduous for complexes as compared to individual proteins. Encouragingly, many advances have begun to address these problems. Improved crystallization techniques require smaller amounts of material to solve large structures. Additionally, newer techniques such as cryo-electron microscopy, can reconstruct low resolution structures for large complexes from much smaller amounts of sample at very low temperatures. However, there is still a large gap between the number of complexes known to exist on the basis of experimental methods and the number for which experimental 3D structures are solved [3].

Despite significant efforts devoted in traditional structure biology and the structural genomics projects that aim at high-throughput complex structure determination [4][5], statistics show that less than ~10% of known protein interactions in the human interactome have an experimental structure associated with them. The number is quite low, given that ~30% of human proteins have been structurally characterized experimentally.

Experimentally, binary interaction data for proteins on a large-scale can be obtained by several methods, such as yeast-two hybrid, affinity purification and protein array techniques. Databases like DIP [6], IntAct [7], MINT [8] and BioGrid [9] have compiled the data obtained by these techniques to assist researchers dealing with protein interactions. However, experimental data is often biased towards complexes with higher stability and
high quantity [10]. Also, there may be inter-study discrepancies between experimental information obtained from different techniques resulting in high false positive rate [11]. Moreover, comparative studies have shown that the repertoires of protein-protein interactions are far from complete. [10]. Therefore, there is a dire need to develop of efficient computational methods not only to discover and model new interactions on a large scale, but also to assess, validate and scrutinize experimentally derived data.

**COMPUTATIONAL METHODS FOR THE PREDICTION OF PPIs**

Wide availability of experimentally determined protein-interaction data has accelerated the development of several computational methods over the last decade. These methods aim to leverage the knowledge derived from experimentally verified known interactions in order to predict new PPIs. Their goal is to predict physical PPIs, where the proteins are actually engaging in physical contact. These approaches can be divided into three general categories: methods based on genomic analysis, protein sequence and three-dimensional protein structure.

Genomic methods analyze patterns such as evolutionary conservation of gene order, gene co-inheritance and co-expression across related organisms in order to identify putative protein interaction pairs [12]. Also, protein-protein interaction annotation can be readily transferred from one organism to another when a pair of proteins had joint sequence identity greater than 80% [13]. On the other hand, sequence-based methods rely on the hypothesis that protein-protein interactions may be mediated through short polypeptide sequences which do not span whole domains but are found repeatedly within the proteins of a cell [14]. Overall, prediction approaches based on genomic methods or sequence
information do not always provide fully reliable answers regarding the putative interacting partners. Looking at the structural details of the putative interaction using an experimentally determined or even a predicted structure can be of great help. This leads to the third class of interaction prediction methods – structure-based approach.

Our knowledge of protein interactions has gotten deeper with the increasing amount of structural data. Structure based methods may look at the global or local structure of the participating proteins in order to predict binary interaction between them. Protein pairs interact with each other using an interface region on their surface. Structure-based methods are often based on geometric description, conservation and electrostatic characterization of these surfaces and critically depend on energy functions used to evaluate proposed conformations. A detailed literature review on structure based prediction of protein interactions is provided in chapter 5.

**MODELING OF PROTEIN-PROTEIN INTERACTIONS**

Computational methods for modeling protein-protein complex structures can be divided into two main classes of algorithms, template-free and template-based methods. Template-free methods, also known as docking, starts with the atomic coordinates of two molecules, typically obtained using x-ray crystallography or nuclear-magnetic resonance, but can also be built using homology modelling. Solving the docking problem involves two main components – pose prediction and pose ranking. Pose prediction corresponds to a search procedure which samples over six different rotational and translational degrees of freedom for favorable binding conformations. Once a pool of candidate conformations is built, the candidates are ranked and filtered using some criteria such as geometric and
physiochemical complementarity in order to identify the best-pose, i.e. the near-native model. On the other hand, template based methods model the structure of a complex by copying and refining the structural framework of “templates”, i.e. related protein complexes whose structures have been experimentally determined. Each approach has its strength and weaknesses. Template-based methods may have more accurate results [15], [16] but they critically depend on template availability, therefore cannot be applied on a proteome-wide scale. Docking methods can be sensitive to large conformational changes upon binding, however they do not require any prior knowledge of structures of related protein complexes to model the association between targets proteins. Therefore, docking approaches provide a higher coverage for large-scale applications that focus on construction and analysis of PPI networks. A more detailed literature review on template-free and template-based docking methods is provided in chapter 4.

In my dissertation, I aimed to develop a novel pipeline for structure-based prediction of protein-protein interactions on a proteome-wide scale. We focused on two major questions: ‘which proteins are interacting with which others’ and ‘how does the interaction take place’. Such an ambitious task requires a new set of tools. In this work, we have developed some highly accurate algorithms for the bottom-up assembly of protein interaction networks. The pipeline is carefully benchmarked against existing experimental data and applied on the entire proteome of E. coli and a few human pathways.
Chapter 2

In this chapter I report an analysis of protein complex structures which demonstrates that binding site locations as well as the interfacial geometry are highly conserved across evolutionarily related proteins. Because the conformational space of protein–protein interactions is highly covered by experimental structures, sensitive protein threading techniques can be used to identify suitable templates for the accurate prediction of interfacial residues. Toward this goal, I developed eFindSitePPI, an algorithm that uses the three-dimensional structure of a target protein, evolutionarily remotely related templates and machine learning techniques to predict binding residues.

Chapter 3

In this chapter I review eFindSitePPI and 9 other methods for protein binding site prediction that are freely available as web servers. In addition, I comparatively evaluate their performance on a common data set comprising different quality target structures. Results show that using experimental structures and high-quality homology models, structure-based methods outperform those using only protein sequences. For moderate-quality models, sequence-based methods often perform better than those structure-based techniques that rely on fine atomic details. We note that post-processing protocols implemented in several methods quantitatively improve the results only for experimental structures, suggesting that these procedures should be tuned up for computer-generated models. However, we observe that eFindSitePPI is fairly tolerant to the structural imperfections in computer-generated models and outperforms other prediction methods for
both high- and moderate-quality models. Finally, we anticipate that advanced meta-prediction protocols are likely to enhance interface residue prediction.

Chapter 4

In this chapter I address the problem of pose-ranking that is frequently observed in docking methods. Despite recent advances in the development of new methods to model macromolecular assemblies, most current methodologies are designed to work with experimentally determined protein structures. However, because only computer-generated models are available for a large number of proteins in a given genome, computational tools should tolerate structural inaccuracies in order to perform the genome-wide modeling of PPIs. To address this problem, we developed eRankPPI, an algorithm for the identification of near-native conformations generated by protein docking using experimental structures as well as protein models. The scoring function implemented in eRankPPI employs multiple features including interface probability estimates calculated by eFindSitePPI and a novel contact-based symmetry score. In comparative benchmarks using representative datasets of homo- and hetero-complexes, we show that eRankPPI consistently outperforms state-of-the-art algorithms improving the success rate by ~10 %.

Chapter 5

In this chapter I describe a pipeline to discover and model protein interactions employing an exhaustive all-to-all docking strategy. This approach integrates molecular modeling, structural bioinformatics, machine learning, and functional annotation filters for the bottom-up assembly of protein interaction networks. In order to demonstrate the utility of this approach on large-scale projects, I modeled dimer structures and predicted PPIs
across the proteome of *E. coli*. The modeling protocol described in this chapter can be applied to detect protein-protein interactions in other organisms as well as to construct dimer structures and estimate the confidence of protein interactions experimentally identified with high-throughput techniques.

Chapter 6

In this chapter I apply the pipeline described in chapter 5 to model structurally characterized protein-protein interaction networks for nine selected human pathways such as diseases of the immune system associated with TLR signaling, myogenesis, synthesis of IP3 and IP4 to name a few.

Chapter 7

I summarize the major findings of my dissertation and discuss their applications in the genome-wide analysis of interactomes.

REFERENCES


CHAPTER 2: BINDING SITE PREDICTION *

INTRODUCTION

Proteins often function in conjugation with other proteins, thus an overwhelming number of biological processes are mediated by protein-protein interactions [1]. For example, interacting proteins are routinely involved in signal transduction, protein transport and folding, DNA replication and repair, and cell division, just to mention a few examples. Consequently, significant efforts have been devoted to study protein-protein interactions because of their importance in elucidating protein function and molecular recognition processes. Also, protein-protein interaction sites are attractive targets for therapeutics as the disruption of crucial interactions may attenuate or even impair the function of pharmacologically relevant proteins [2][3]. In recent years, many experimental and theoretical studies have been conducted to discover and characterize these interactions; however, despite evident progress, salient challenges remain. Experimental methods used to identify interface residues are often low-throughput with associated high costs of instruments and experiments. Therefore, many cost-efficient computational approaches have been developed for the prediction of interaction sites to complement experimental efforts. For instance, computationally predicted protein-protein interaction sites can be used to optimize site directed mutagenesis experiments by reducing the number of mutations needed to be tested in vitro [4][5][6]. Protein-protein docking is another important application of interfacial site prediction.

Taking into account even the approximate location of protein interface can, in principle, reduce the search space, improve the accuracy of modeled complexes, and shorten computing time [7][8][9]. For instance, Li and Kihara showed that docking results obtained by a docking program PI-LZerD are successfully improved even when the accuracy of supplied PPI restraints is significantly low [8]. On the other hand, another study by Shih and Hwang demonstrated that when using bioinformatics-predicted information on interface residues, data-guided protein docking methods perform poorly [10], suggesting that PPI restraints should have a certain accuracy in order to improve protein docking.

Until now, a variety of computational methods have been developed for the prediction of protein-protein interaction sites [11][12][13][14]. Sequence-based methods largely rely on features extracted from sequence profiles constructed by PSI-BLAST [15][16][17]. Other methods extensively utilize remote evolutionary information to detect functionally important sites [18][19][20][21]. For example, the Evolutionary Trace algorithm [20] maps conserved amino acids onto a 3D protein structure and then identifies functional sites by analyzing highly conserved residues in the branches of an evolutionary tree. Identified residues are assumed to be structurally important if they lie in the core of a protein, while those on the surface are relevant for protein function. Finally, as a consequence of the continuously growing structural content in protein databases [22], a number of structure-based approaches have been developed. These algorithms exploit geometrical and physicochemical features derived from the three-dimensional structures of target proteins [23][24][25], e.g. the solvent accessibility, secondary structure states, hydrophobicity, B-factors and the local topology. Furthermore, recent studies demonstrate that the interaction sites tend to be conserved among structural analogs [26], which
stimulates the development of methods for the prediction of protein-protein interaction (PPI) sites based on the global structural similarity between query proteins and those with known dimer structures. For example, a recently developed method called PrePPI derives empirical scores from the interfaces of structural neighbors for the prediction of binary protein-protein interaction [27]. The accuracy and coverage of approaches based on the global structural similarity certainly depend on the availability of experimental structures of target proteins as well as the oligomer complexes of their structural neighbors.

PPI sites can be separated from the rest of the surface by various geometric features, e.g. accessible surface area, planarity and protrusion [28][23], as well as the local structure similarity between query proteins and a repository of known dimers [25]. Consequently, there is an increasing interest in PPI prediction based on the local similarity; for instance, PrISE detects interaction sites using a local surface similarity between query proteins and a collection of structural elements [25]. Notwithstanding the evident progress in the structure-based identification of PPI sites in proteins, these methods have not been widely used in proteome-scale applications, primarily because 1) the number of proteins with known structures is far smaller than the number of known sequences, 2) they may require an additional knowledge of interacting partners, which is often unavailable, and 3) their performance depends on the availability of protein dimers structurally similar to query proteins.

In that regard, continuous efforts are directed towards the development of novel approaches for the prediction of protein-protein interfacial sites. In this study, we describe the development and benchmarking of eFindSitePPI, a new evolution/structure-based
method that can be used to predict PPI sites in proteins with known structures, as well as in gene products whose structures have not yet been solved experimentally. eFindSite\textsuperscript{PPI} effectively integrates sensitive meta-threading techniques with structure alignments and machine learning to accurately detect interfacial residues in query proteins. Its unique feature is the capability to predict positions and types of molecular interactions that target proteins are likely to form with their partners. These include many interactions known to stabilize protein-protein complexes, such as hydrogen bonds, salt bridges, as well as hydrophobic and aromatic contacts. Importantly, eFindSite\textsuperscript{PPI} makes accurate predictions for protein models with diverse quality, which opens up the possibility for structure-based PPI site identification at the proteome scale. Finally, in comprehensive benchmarks, we demonstrate that eFindSite\textsuperscript{PPI} outperforms other methods for the prediction of PPI sites from protein structures.

MATERIALS AND METHODS

Overview of eFindSite\textsuperscript{PPI}

eFindSite\textsuperscript{PPI} is a new evolution/structure-based approach for the prediction of protein binding sites, specific interactions as well as the local interfacial geometry. The flowchart shown in Figure 2.1 illustrates the procedure implemented in eFindSite\textsuperscript{PPI}, which starts with the structure of a target protein (Figure 2.1A). Next, using meta-threading, functionally and structurally related templates are identified in the template library (Figure 2.1B). For each template, eFindSite\textsuperscript{PPI} retrieves its known complexes and maps their interfaces onto the target protein using structure alignments (Figure 2.1C). Then, the algorithm computes five different attributes for each surface residue in the target protein:
the relative accessible area, generic interface propensity, sequence entropy, position specific interface propensity, and the fraction of templates that have an equivalent residue at the protein-protein interface (Figure 2.1D). These attributes are combined into probabilistic scores by machine learning using Support Vector Machines and the Naïve Bayes Classifier (SVM and NBC, respectively; Figure 2.1E). Both classifiers are finally used to distinguish between interface and non-interface residues in the target protein (Figure 2.1F). Below, we describe datasets used in this study, i.e. the template library and various benchmarking sets, provide details on the methods and algorithms implemented in eFindSitePP1, and explain evaluation metrics used to assess its performance in PPI prediction.

Figure 2.1. Flowchart for the PPI site prediction using eFindSitePP1. Details are given in text.

**Dimer template library**

Template library was compiled from all Protein Data Bank (PDB) [29] entries as of September 2012 with biologically relevant arrangements of two protein chains identified using PISA (Protein, Interface, Surfaces, and Assemblies) [30]. The redundancy was removed at 40% pairwise sequence identity by CD-HIT [31], however, two homologous dimers were included in the library if they either had structurally dissimilar receptor proteins with a TM-score (Template Modeling score) of <0.4 [32], non-overlapping interfacial residues with Matthew’s correlation coefficient (MCC) of <0.5, or a different interfacial geometry with an
IS-score (Interfacial Similarity score) of <0.191 [33]. Note that an IS-score of 0.191 indicates a significant interfacial similarity at a *p*-value of 0.05. TM-score is a structure alignment quality measure that ranges from 0 to 1 and has a length independent statistical significance threshold of ≥0.4, which corresponds to a *p*-value of $3.4 \times 10^{-5}$ [32]. Here, TM-score is calculated upon structure alignments constructed by Fr-TM-align [34], whereas the overlap of binding residues and the local structure similarity of binding interfaces (IS-score) are assessed by iAlign [33]. The complete template library comprises 17,792 dimer structures.

**Benchmarking dataset BM4361**

The primary dataset used in eFindSite<sup>PPI</sup> benchmarking, BM4361, consists of complex crystal structures selected from the template library. In each dimer, the longer chain is considered a receptor and shorter chain is a ligand. We selected those dimers, in which the receptor has 50-600 residues. Furthermore, to avoid ambiguity when assessing the accuracy of interfacial residue prediction, we excluded receptors that interact with different ligands through different binding residues or whose close homologues with ≥40% sequence identity form different protein-protein interactions. This procedure resulted in a non-redundant dataset of 4,361 protein dimers with unique and biologically relevant interfaces, referred to as BM4361. In addition to benchmarking simulations, this dataset was used to optimize eFindSite<sup>PPI</sup> parameters and to construct machine learning models.

**Benchmarking dataset BM1905**

This dataset was compiled as a subset of BM4361 to benchmark the accuracy of binding residue prediction against non-native structures. It features three structural forms for each receptor protein: a crystal structure as well as high- and moderate-quality protein
models. Weakly homologous models were generated by template-based modeling using 
\( e\text{Thread} \) [35][36] following a procedure described in Supporting Information. \( e\text{Thread} \) is a 
meta predictor that integrates several single threading algorithms to improve the 
recognition of structurally and functionally related templates [37]. Both models with the 
preferred accuracy were constructed for 1,905 target proteins, thus the corresponding sets 
of crystal structures, high-, and moderate-quality models are referred to as BM1905C, 
BM1905H and BM1905M, respectively.

Other datasets

In addition to the BM4361 and BM1905 datasets, we compare the performance of 
\( e\text{FindSite}^{\text{PPI}} \) to other approaches for interfacial residue prediction on datasets used 
previously in the development and benchmarking of those algorithms. Comparison with 
PrISE is carried out using bound and unbound receptor conformations from the Benchmark 
4.0 dataset [38]. We note that the accuracy of PrISE is assessed only against crystal structures 
in their bound conformational state [25]. We excluded multimeric complexes, in which the 
receptor is either smaller than 50 or larger than 600 residues, forms multiple interfaces, or 
the interface is made up of less than 20 residues. This dataset consists of 170 target proteins, 
95 in bound and 75 in the unbound conformational state. We also assess the performance of 
\( e\text{FindSite}^{\text{PPI}} \) with respect to ET and iJET predictors [20][21] on the Huang dataset [39], 
applying similar criteria as described above. This dataset comprises 52 target proteins 
including 28 homodimers, 17 heterodimers and 7 transient complexes. When applicable, we 
modify \( e\text{FindSite}^{\text{PPI}} \) parameters to match prediction procedures described in the original 
publications of PrISE, ET and iJET.
Selection of dimer templates

*eFindSite^PPI* is a template-based approach, which employs meta-threading using *eThread* [35][36] to identify structurally and functionally related proteins in the template library as described previously [37]. At least one dimer template is required in order to make a prediction. By default, we carry out benchmarking simulations excluding closely related templates, whose sequence identity to the target is >40%. Moreover, we only use templates that structurally align to their targets with a TM-score of ≥0.4 [32] as reported by Fr-TM-align [34]. Note that benchmarking calculations under these conditions are devised to approximate real applications in across-proteome functional annotation, where at most weakly homologous proteins can be identified for the majority of gene products. In addition to the default sequence identity threshold of 40%, we evaluate the performance of *eFindSite^PPI* at 30% and 20% as well.

Interfacial probability score

Each residue in the target protein is assigned an interfacial probability score that estimates the likelihood of this residue position to be at the protein-protein interface. These scores are calculated using machine learning and a set of the following residue-level attributes:

Relative surface accessibility - The relative accessible solvent area (ASA) of each residue is calculated using NACCESS [40]. This program implements a method by Lee and Richards [41], which calculates the atomic accessible surface by rolling a probe of a given size around the van der Waals surface. Residues with a surface accessibility of <5% are considered
buried, thus non-interfacial. Remaining residues are assigned the relative surface accessibility score, RSA.

*Interface propensity* - We use interface residue propensities derived for 20 standard amino acids by Jones and Thornton from a non-redundant set of high-resolution crystal structures of protein-protein complexes [42][43]. Interface propensities, *IP*, describe the statistical likelihood of different amino acids to be found at protein-protein interfaces. These are calculated for each amino acid (*AA_j*) as the relative contribution of *AA_j* to the interfacial ASA compared to the whole surface:

$$IP_j = \frac{\sum_{i=1}^{N_j} ASA_i(j)}{\sum_{i=1}^{N_i} ASA_i}/\frac{\sum_{s=1}^{N_s} ASA_s(j)}{\sum_{s=1}^{N_s} ASA_s}$$ Eq. 2.1

Where, \(\sum ASA_i(j)\) is the sum of ASA of amino acid residues of type *j* at the interface, \(\sum ASA_i\) is the sum of ASA of all amino acids at the interface, \(\sum ASA_i(j)\) is the sum of ASA of amino acid residues of type *j* on the surface, and \(\sum ASA_s\) is the sum of ASA of all amino acids on the surface.

*Sequence entropy* - Functionally important residues tend to be evolutionarily conserved [39][44][45], therefore, we include a conservation score estimating the sequence variability for each target residue. First, multiple sequence alignments generated for the target
sequence by PSI-BLAST [46] are converted to a sequence profile. Then, the conservation score for each residue position, \( SE \), is calculated using the Shannon entropy [47]:

\[
SE = - \sum_{i=1}^{20} p_i \log_2 (p_i)
\]

Eq. 2.2

where \( p_i \) is the fraction of residues of amino acid type \( i \) in a given position in the sequence profile. \( SE \) ranges from 0 (absolute conservation of a particular residue type) to 4.32 bits (maximum entropy for equally distributed amino acids).

*Position-specific interface propensity* - The PSIP score combines generic interface residue propensities, as described above, with evolutionary information included in sequence profiles:

\[
PSIP = \sum_{i=1}^{20} p_i IP_i
\]

Eq. 2.3

where \( p_i \) is the fraction of residues of amino acid type \( i \) at a given position in the profile and \( IP_i \) is the interface propensity for amino acid type \( i \).

*Fraction of templates* - Finally, we include the fraction of templates, \( FT \), that have an interfacial residue in the equivalent position according to template-target structure alignments constructed by Fr-TM-align.

Individual residue-level attributes, \( RSA, IP, SE, PSIP \) and \( FT \), are combined into a single probabilistic score using machine learning. Two different classifiers, Support Vector Machines (SVM) [48] and the Naïve Bayes Classifier (NBC) [49], are trained to predict interfacial residues according to the assignment by iAlign [33]. iAlign assigns interfacial
residues based on interatomic contacts, which occur when any two heavy atoms belonging to residues from different chains are within a distance of 4.5 Å. Both machine learning models are 2-fold cross-validated on the BM4361 dataset. Specifically, dataset proteins are randomly divided into two subsets, A and B; A is used to train a model and then validate it against B, and vice versa, the model trained on B is validated against A. We note that <40% sequence identity between any pair of proteins in the BM4361 dataset ensures that the classifiers are trained and validated using different proteins. Probability thresholds optimized using the BM4361 dataset are 0.202 for the SVM and 0.178 for the NBC predictor. These values were selected to maximize Matthew’s correlation coefficient to 0.428, which corresponds to a true positive rate of 0.464 at the expense of 0.076 false positive rate. A given residue in the target protein is predicted to be at the interface when both probabilities are above their threshold values.

**Calculation of interfacial interactions**

In analyzing interfacial interactions, we consider the following four types of inter-residue contacts: salt bridges, hydrogen bonds, hydrophobic, and aromatic interactions. Salt bridges and hydrogen bonds across protein interfaces are detected by PDB2PQR [50]. Hydrophobic interactions are defined when the distance between any pair of atoms belonging to hydrophobic side chains is ≤5 Å; hydrophobic amino acids include Ala, Ile, Leu, Phe, Pro, Met and Val. Using the same distance threshold, aromatic contacts are identified between the side chains of His, Phe, Trp and Tyr. For each predicted interfacial residue, we calculate the fraction of templates that have a residue in the equivalent position forming a particular type of protein-protein interaction using template-target structure alignments.
constructed by Fr-TM-align. These frequency values calculated for all interaction types correspond to the probabilities of various contacts that target residues may form with protein partners. Thresholds optimized on the BM4361 dataset are 0.001 for salt bridges, 0.021 for hydrogen bonds, 0.041 for hydrophobic contacts, and 0.012 for aromatic interactions. Similar to the interface residue prediction, these threshold values maximize the respective Matthew’s correlation coefficients.

Confidence estimation system

In proteome-level function inference, reliable predictions cannot be obtained for all targeted gene products, therefore, various predictors are required to provide confidence estimates. Every prediction by eFindSite<sup>PPI</sup> is assigned an overall confidence score, CS, defined as:

\[
CS = \frac{1}{N} \sum_{i=1}^{N} SVM_i \times NBC_i
\]

Eq. 2.10

where \( N \) is the total number of predicted binding residues, and \( SVM_i \) and \( NBC_i \) are the binding probability scores assigned to \( i \)-th residue by machine learning using Support Vector Machines and the Naïve Bayes Classifier, respectively. Calibrated ranges are \( CS \geq 0.5 \) for high, \( 0.25 < CS < 0.5 \) for medium, and \( CS \leq 0.25 \) for low confidence predictions.

Performance evaluation metrics

Binding residue prediction by eFindSite<sup>PPI</sup> is assessed using standard evaluation metrics for classification problems:
Sensitivity (true positive rate): \( TPR = \frac{TP}{TP + FN} \)  \hspace{1cm} \text{Eq. 2.4}

Fall-out (false positive rate): \( FPR = \frac{FP}{FP + TN} \)  \hspace{1cm} \text{Eq. 2.5}

Specificity (true negative rate): \( SPC = \frac{TN}{FP + TN} \)  \hspace{1cm} \text{Eq. 2.6}

Precision (positive predictive value): \( PPV = \frac{TP}{TP + FP} \)  \hspace{1cm} \text{Eq. 2.7}

Accuracy: \( ACC = \frac{TP + TN}{TP + FP + TN + FN} \)  \hspace{1cm} \text{Eq. 2.8}

Matthew’s correlation coefficient:

\[
MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + TN)(FP + FN)(TN + FN)}}
\]

\hspace{1cm} \text{Eq. 2.9}

where \( TP \) (True Positives), \( FN \) (False Negatives) and \( FP \) (False Positives) is the number of correctly predicted, under-, and overpredicted binding residues, respectively. \( TN \) (True Negatives) is the number of correctly predicted non-interfacial residues. Binding residues in experimental complex structures (Positives) are defined as those forming protein-protein interfaces according to iAlign [33]. The minimum value is 0 and the maximum value is 1 for all scores, except for MCC that ranges from -1 to 1. MCC quantifies the strength of the correlation between predicted and actual classes; by heavily penalizing both over- and under-predictions, it provides a convenient assessment measure that balances the sensitivity and specificity. In addition to numerical values assessing the classification
accuracy, we analyze the prediction results using Receiver Operating Characteristic (ROC) plots. This technique was developed to evaluate the overall performance of a classifier and shows the tradeoff between sensitivity and specificity. The area under the ROC curve (AUC) quantifies the performance of classifier; larger AUC values indicate a better prediction power of the classification model.

The accuracy of interface residue prediction is compared to that of a random, size-independent classifier. First, for a given target protein, we estimate the size of its interface from the number of exposed residues as described by Martin [51]. Next, we randomly select a patch on the target surface whose size is equivalent to the estimated number of interfacial residues. This patch represents a random interface and includes the correction of a size bias, i.e. smaller proteins have proportionally more residues within the patch, increasing the chances of overlapping with the correct interface.

RESULTS AND DISCUSSION

Accuracy of template selection

eFindSitePPI employs meta-threading and structure alignments to select templates for the prediction of interfacial sites. The prediction accuracy inevitably depends on the quality of the identified set of dimer templates; therefore, using the BM4361 dataset, we first assess the accuracy of template selection. We note that templates used in this study are at most weakly homologous, sharing <40% sequence identity with their targets. Figure 2.2 shows a series of ROC plots cross-validating the accuracy of template selection with respect to several features. Using template confidence as a variable parameter, Figure 2.2A (a solid line) shows the performance of eThread in detecting those templates that are structurally similar to the
target with a TM-score of ≥0.4. Structure similarity is quantified by the TM-score [32] calculated for template-target structure alignments constructed by Fr-TM-align [34]. Detecting structurally similar templates yields the maximum accuracy of 0.746 at a true positive rate of 0.642 and a false positive rate of 0.210, resulting in the area under ROC of 0.754.

Next, in addition to the global structure similarity, we also require a template to have a similar location of the PPI interface in order to be considered a positive. Specifically, we measure the interface overlap between the target and a template by calculating Matthew’s correlation coefficient (MCC) over interfacial residues in both structures with residue equivalences taken from structure alignments. MCC values of ≥0.5 indicate that both the target and a template bind their partners at similar locations. Figure 2.2A (a dashed line) shows that protein templates whose binding interfaces are at similar locations are accurately detected. The corresponding area under ROC is 0.747 with the maximum accuracy of 0.759 obtained at a true positive rate of 0.655 and a false positive rate of 0.215. Finally, we consider the most stringent case, where the interfacial geometry in a template is similar to that in the target with an IS-score of ≥0.191. The IS-score measures interfacial similarity by comparing geometric distances as well as the conservation of contact patterns [33]. Encouragingly, the area under ROC is 0.709, with the maximum accuracy of 0.695 at a true positive rate of 0.778 and a false positive rate of 0.419 (Figure 2.2A, a dotted line). Our results demonstrate that both the interface location and its geometry are conserved across a set of evolutionarily and structurally related proteins, which accords with previous studies [26][33]. Therefore, threading and meta-threading techniques can be effectively utilized to explore remote relationships between proteins using sensitive sequence profile comparisons. This strategy
optimizes the selection of dimer templates for template-based prediction of functional aspects related to protein-protein interactions.

Similarity-based approaches to protein docking use dimer templates, in which both monomers are structurally similar to the target monomers [27][52]. These algorithms employ global structure similarity to construct complex models based on the identified dimer templates. Therefore, we also analyze the capabilities of threading to detect weakly homologous receptor templates that bind globally similar ligands. First we assess the global structure similarity of template ligands, where the interacting partners with a TM-score ≥0.4 to the target ligand are positives. Figure 2.2B (a dashed line) shows that binding ligands are not necessarily structurally similar to the target ligand even when they share the same binding location. The corresponding area under ROC is only 0.538 and the maximum accuracy of 0.483 is obtained at a true positive rate of 0.448 and a false positive rate of 0.373.

Next, we use global sequence similarity to select interacting partners from the identified dimer templates; here, template ligands whose sequence identity to the target ligand is ≥40% are positives. Interestingly, as shown in Figure 2.2B (a solid line), receptor templates with similar binding sites tend to bind homologous proteins with respect to the target ligand. The area under ROC is 0.848 and the maximum accuracy of 0.790 is obtained at a true positive rate of 0.866 and a false positive rate of 0.210. We note that structurally similar ligands with a TM-score of ≥0.4 and homologous ligands with a sequence identity of ≥40% were found for 44% and 0.5% of the cases, respectively. This analysis shows that the interface site can be inferred using the global structure similarity when the sequence
similarity between the target and template ligands is high. Nevertheless, due to the incompleteness of dimer libraries, the coverage of suitable protein targets is rather low.

Figure 2.2. Accuracy of eThread in recognizing templates for PPI site prediction. In (A), correct templates for the receptor (larger subunit) are defined using the global structure similarity with a TM-score of ≥0.4, the overlap of interfacial residues with MCC of ≥0.5, and the local interfacial similarity with an IS-Score of ≥0.191. In (B), we evaluate the recognition of those dimer templates in which the ligand (smaller subunit) is globally similar to the target-bound ligand with a sequence identity of ≥40% and a TM-score of ≥0.4, respectively. Combined curves are calculated using a 2-fold cross-validation against the BM4361 dataset. TPR – true positive rate, FPR – false positive rate. Gray areas correspond to predictions no better than random.

Conservation of interfacial interactions

Since protein complexes are stabilized by a variety of interactions, we analyze the conservation of interaction patterns across weakly related proteins. For each protein in the BM4361 dataset, interfacial interactions in its dimer templates are mapped to the target residues according to the structure alignments of receptor proteins. ROC plots in Figure 2.3
show the structural conservation of interfacial hydrogen bonds, salt bridges, aromatic and hydrophobic contacts at protein-protein interfaces. ROC curves end at certain sensitivity values, because we can only take account of those surface residues having an interacting residue at a structurally aligned position in at least one template. The maximum accuracy obtained for hydrogen bonds, salt bridges, hydrophobic and aromatic interactions is 0.900, 0.945, 0.895 and 0.949, at a true (false) positive rate of 0.684 (0.091), 0.459 (0.049), 0.760 (0.098) and 0.488 (0.044), respectively. Comparison of these ROC plots shows that the conservation of interfacial hydrophobic contacts and hydrogen bonds is higher than aromatic interactions and salt bridges.

The high conservation of hydrophobic contacts is in line with previous studies suggesting that these interactions play a central role in stabilizing protein-protein complexes and the PPIs are dominated by hydrophobic patches [42] [43]. Overall, the results suggest that, in addition to binding residues, the interaction conservation patterns detected across structurally and evolutionarily related proteins can be used to predict various interaction types as well. These features can be used to support protein-protein docking simulations by favoring those assembled dimer conformation, in which highly conserved interactions are formed.

**Prediction of PPI sites using experimental structures**

*eFindSitePPI* extracts PPIs from weakly homologous dimer templates identified by meta-threading for the prediction of protein binding residues, specific interactions as well as the local interfacial geometry. Most of these features are identified by machine learning techniques. Here, we assess the accuracy of binding residue prediction, i.e. the classification
of target residues as either interfacial or non-interfacial, using two machine learning algorithms, Support Vector Machines (SVM) and the Naïve Bayes Classifier (NBC).

Figure 2.3. ROC plot evaluating the conservation of different types of protein-protein interactions across sets of evolutionarily weakly related dimer templates. The following non-covalent interaction types are considered: hydrogen bonds, salt bridges, hydrophobic, and aromatic contacts. A variable parameter is the fraction of templates that form the same interactions as the target in structurally equivalent positions. TPR – true positive rate, FPR – false positive rate. Gray area corresponds to interactions found by a random chance.

As shown in Figure 2.4, the performance of both classifiers on the BM4361 dataset is fairly comparable. The area under ROC for SVM is 0.737, with the maximum MCC of 0.404 at a true (false) positive rate of 0.573 (0.144). For NBC, the area under ROC is 0.773, with the maximum MCC of 0.339 at a true (false) positive rate of 0.628 (0.209). Encouragingly, combining both classifiers using optimized thresholds, labeled as SVM+NBC in Figure 2.4,
further enhances the discriminatory power. Specifically, MCC improves to 0.428, which corresponds to a sensitivity of 0.464 at the expense of only 0.076 false positive rate.

![ROC plot assessing the accuracy of interfacial residue prediction across the BM4361 dataset by eFindSitePPI compared to PINUP. For eFindSitePPI, three prediction protocols are evaluated: SVM only, NBC only and a combination of SVM and NBC. TPR – true positive rate, FPR – false positive rate. Gray area corresponds to predictions no better than random.](image)

We also evaluate the performance of eFindSitePPI in predicting specific interactions that the target protein is likely to form with its partners. The performance of eFindSitePPI in the prediction of interaction types across the BM4361 dataset is shown in Figure 2.5; note that under-predicted interfacial residues count as false negatives in this analysis. Interestingly, despite the fact that closely homologous templates with a sequence identity of >40% were excluded from benchmarking calculations, the prediction of all interaction types is fairly accurate. True positive rates for hydrogen bonds and aromatic interactions are 0.515 and 0.484, with very small false positive rates of 0.048 and 0.037, respectively. For salt bridges and hydrophobic contacts, the true (false) positive rates are 0.330 (0.031) and 0.306
(0.017). These results demonstrate that eFindSitePPI predicts approximately one-half of interfacial hydrogen bonds and aromatic interactions, and one-third of salt bridges and hydrophobic contacts.

**Size and composition of predicted interfaces**

In addition to binding residues and interaction types predicted by eFindSitePPI, in Figure 2.6, we analyze the general properties of interfacial sites, such as their size and amino acid composition. Figure 2.6A shows that the size of interfacial sites predicted by eFindSitePPI for the BM4361 dataset correlates well with the size of experimental interfaces identified by iAlign [33]; the Pearson correlation coefficient is 0.720 with a standard error of 0.118. In Figure 2.6B, we compare the amino acid composition of experimental and predicted protein-protein interfaces. The frequencies of amino acids at the predicted interfaces are in good quantitative agreement with the experimental data; the differences are less than 1% on average. Consequently, interfaces predicted by eFindSitePPI are predominantly hydrophobic, which is consistent with a previous study conducted by Lijnzaad and Argos showing that interfacial sites often contain the largest or second-largest hydrophobic patches on the surface of proteins [53].

Next, we evaluate the composition of amino acids involved in specific interactions at protein-protein interfaces. In general, interfaces are rich in hydrogen bonds, which are the major contributors to electrostatic interactions between proteins [54]. The analysis of the composition of residues involved in the formation of hydrogen bonds at the predicted interfaces reveals that some polar residues are underrepresented, e.g. Arg, Glu, Asp and Ser (by 3.9%, 4.0%, 4.5% and 2.3%, respectively), whilst several hydrophobic residues are over
predicted to form hydrogen bonds, e.g. Leu, Ala, Ile, Phe, Pro and Met (by 4.8%, 2.8%, 2.1%, 2.6%, 2.2% and 1.8%, respectively). The amino acid composition of residues predicted to interact with ligands through salt bridges, hydrophobic and aromatic contacts are comparable to that in the experimental complexes except for Arg and Phe, which are slightly over predicted to form electrostatic and hydrophobic contacts by 5.5% and 5.1%.

**Susceptibility to target-template sequence similarity**

The accuracy of template-based function inference certainly depends on the target-template sequence similarity, therefore, we analyze the performance of eFindSitePPI at different similarity thresholds applied to the selection of evolutionarily related templates. Table 2.1 summarizes the results obtained at 40%, 30% and 20% sequence similarity thresholds. The accuracy of protein interface prediction at 40% and 30% similarity thresholds is comparably high, however, the performance of eFindSitePPI starts deteriorating at lower sequence similarity thresholds. For example, MCC is 0.428, 0.381 and 0.177 at 40%, 30% and 20% sequence similarity, respectively. This corresponds to a true (false) positive rate of 0.464 (0.076), 0.415 (0.077) and 0.151 (0.042). Thus excluding templates with >20% sequence identity to the target leads to an approximately two-fold drop-off in the prediction accuracy compared to higher sequence identity thresholds. We note that this is a common feature of threading-based approaches to protein function inference from evolutionarily related templates and a similar behavior was observed in ligand binding site prediction using eFindSite [55].
Figure 2.5. ROC plot for the prediction of various interaction types by eFindSite$^{PPI}$ for the BM1905C dataset. The following non-covalent interaction types are considered: hydrogen bonds, salt bridges, hydrophobic, and aromatic contacts. TPR – true positive rate, FPR – false positive rate. Gray area corresponds to predictions no better than random.

**Protein models as targets for PPI prediction**

Similar to eFindSite, a recently developed algorithm to ligand-binding site prediction, the design of eFindSite$^{PPI}$ makes it particularly well suited for structure-based PPI prediction using protein models. Therefore, in addition to target crystal structures, we benchmark eFindSite$^{PPI}$ against computer-generated models. The details on model preparation and their structural characteristics are provided as Supporting Information.
Figure 2.6. Size and composition of interfaces predicted by eFindSite<sup>PPI</sup>. (A) The correlation between the size of experimental interfaces identified by iAlign and those predicted by eFindSite<sup>PPI</sup>. (B) Amino acid composition of experimental and predicted interfaces.

Table 2.1. Performance of eFindSite<sup>PPI</sup> in interface residue prediction across the BM1905C dataset at different target-template sequence similarity thresholds.

<table>
<thead>
<tr>
<th>Similarity threshold</th>
<th>Evaluation metric&lt;sup&gt;a&lt;/sup&gt;</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FPR</td>
<td>TPR</td>
<td>ACC</td>
<td>SPC</td>
<td>PPV</td>
<td>MCC</td>
</tr>
<tr>
<td>40%</td>
<td>0.076</td>
<td>0.464</td>
<td>0.835</td>
<td>0.924</td>
<td>0.594</td>
<td>0.428</td>
</tr>
<tr>
<td>30%</td>
<td>0.077</td>
<td>0.415</td>
<td>0.824</td>
<td>0.922</td>
<td>0.563</td>
<td>0.381</td>
</tr>
<tr>
<td>20%</td>
<td>0.042</td>
<td>0.151</td>
<td>0.800</td>
<td>0.957</td>
<td>0.459</td>
<td>0.177</td>
</tr>
</tbody>
</table>

<sup>a</sup> FPR: false positive rate; TPR: sensitivity; ACC: accuracy; SPC: specificity; PPV: precision; MCC: Matthew’s correlation coefficient.
Figure 2.7. Accuracy of interfacial residue identification for predictions assigned different confidence levels. The accuracy is assessed by Matthew's correlation coefficient; boxes end at the quartiles $Q_1$ and $Q_3$ and a horizontal line in each box is the median. Whiskers point at the farthest points that are within $3/2$ times the interquartile range.

Benchmarking results for different quality models from the BM1905 dataset compared to experimental structures are presented in Table 2.2 Since small proteins involve proportionally more residues at interfaces compared to large targets, it is important to eliminate a potential bias caused by this size effect. To address this issue, several techniques for systematic corrections have been recently suggested [51]. Table 2.2 also includes a random background that accounts for the size bias estimated for the BM1905 dataset. Only a fraction of surface residues contribute to PPIs, therefore most residues assigned by a random classifier are true negatives, resulting in a relatively high accuracy (ACC) and
Table 2.2. Comparison of the performance of eFindSite\textsuperscript{PPI} and PINUP using different quality target structures. For eFindSite\textsuperscript{PPI}, three prediction protocols are evaluated: SVM only, NBC only and a combination of SVM and NBC (listed as eFindSite\textsuperscript{PPI}). Values pointing to the best performance are highlighted in bold, except for FPR and TPR that need to be considered jointly.

<table>
<thead>
<tr>
<th>Dataset\textsuperscript{a}</th>
<th>Predictor</th>
<th>Evaluation metric\textsuperscript{b}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FPR</td>
</tr>
<tr>
<td>BM1905C</td>
<td>eFindSite\textsuperscript{PPI} (SVM)</td>
<td>0.150</td>
</tr>
<tr>
<td></td>
<td>eFindSite\textsuperscript{PPI} (NBC)</td>
<td>0.208</td>
</tr>
<tr>
<td></td>
<td>PINUP</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>Random\textsuperscript{c}</td>
<td>0.078</td>
</tr>
<tr>
<td>BM1905H</td>
<td>eFindSite\textsuperscript{PPI} (SVM)</td>
<td>0.161</td>
</tr>
<tr>
<td></td>
<td>eFindSite\textsuperscript{PPI} (NBC)</td>
<td>0.228</td>
</tr>
<tr>
<td></td>
<td>PINUP</td>
<td>0.083</td>
</tr>
<tr>
<td></td>
<td>Random\textsuperscript{c}</td>
<td>0.074</td>
</tr>
<tr>
<td>BM1905M</td>
<td>eFindSite\textsuperscript{PPI} (SVM)</td>
<td>0.169</td>
</tr>
<tr>
<td></td>
<td>eFindSite\textsuperscript{PPI} (NBC)</td>
<td>0.233</td>
</tr>
<tr>
<td></td>
<td>PINUP</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td>Random\textsuperscript{c}</td>
<td>0.076</td>
</tr>
</tbody>
</table>

\textsuperscript{a}BM1905C: crystal structures; BM1905H: high-quality models; BM1905M: moderate-quality models. \textsuperscript{b} FPR: false positive rate; TPR: sensitivity; ACC: accuracy; SPC: specificity; PPV: precision; MCC: Matthew’s correlation coefficient. \textsuperscript{c} Random performance includes the correction of a size bias.
specificity (SPC). However, sensitivity (TPR) and fall-out (FPR) are comparably low and close to the diagonal in a ROC space.

Using the SVM classifier in eFindSitePPI yields slightly better performance than NBC, however, combining predictions from both machine learning algorithms (listed as eFindSitePPI in Table 2.2) gives the highest accuracy. For instance, using target crystal structures, MCC for eFindSitePPI is 0.428. The performance using protein models is only slightly worse with MCC of 0.371 for high- and 0.339 for moderate-quality models. Compared to a random, size-independent classifier, using eFindSitePPI yields MCC values higher by 0.417 for target crystal structures, and 0.352 and 0.309 for high- and moderate-quality models. This analysis demonstrates that eFindSitePPI is capable of tolerating distortions in modeled target structures.

**Prediction confidence**

A reliable confidence index is an essential feature to identify those targets, whose interface is likely to be correctly predicted. eFindSitePPI uses an average probability score assigned by machine learning to target residues to categorize predictions as either high, medium or low confidence. In Figure 2.7, we report the prediction accuracy separately for each confidence group using target crystal structures as well as protein models from the BM1905 dataset. In general, confidence estimates correlate well with the actual prediction accuracy assessed by MCC across all datasets, i.e. the average MCC for high-confidence predictions is significantly higher than those assigned medium- and low-confidence. For high-confidence predictions, using targets from the BM1905C, BM1905H and BM1905M datasets yields the median MCC of 0.623, 0.585 and 0.520, whereas for medium- (low-)
confidence predictions, the median MCC is 0.383 (0.128), 0.246 (0.095) and 0.210 (0.086), respectively.

As expected, the percentage of high-confidence predictions slightly decreases from 32% to 29% (28%) when high- (low-) quality protein models are used instead of the target crystal structures. To that end, eFindSite^{PPI} offers a reliable confidence index, which can be used to select only accurately predicted interfaces for large-scale protein docking simulations and other applications that may require a high precision.

**Comparison with PINUP**

We compare the performance of eFindSite^{PPI} to several structure-based approaches for protein-binding residue prediction. The first one is PINUP [24], a method that employs residue-level energy scores, accessible surface area-dependent interface propensities and conservation scores to derive a set of structural and functional constraints. PINUP effectively combines side chain energy, residue conservation and interface propensity into a single score, which is used to build a consensus region from initial top-ranked patches. The corresponding weight factors were obtained by a linear optimization of the scoring function against a training dataset of 57 protein targets. Figure 2.4 shows that eFindSite^{PPI} is almost twice as sensitive as PINUP on the BM4361 dataset; a true positive rate for eFindSite^{PPI} and PINUP is 0.446 and 0.236, at a comparably low false positive rate of 0.073 and 0.060, respectively. In Table 2.2, we assess the performance of both methods using experimental structures and different quality protein models from the BM1905 dataset. Consistent with benchmarking results against BM4361, eFindSite^{PPI} outperforms PINUP on crystal structures from the BM1905C dataset; for instance, MCC is 0.428 for eFindSite^{PPI} and 0.189.
for PINUP. More importantly, the prediction accuracy for $e$FindSite$^{PPI}$ against protein models from the BM1905H and BM1905M datasets is much higher than for PINUP. When high- (moderate-) quality models are used instead of the experimental structures, MCC for PINUP decreases by 0.109 (0.136), whereas for $e$FindSite$^{PPI}$, MCC decreases only by 0.057 (0.089). Thus, $e$FindSite$^{PPI}$ tolerates structure deformations in protein models more efficiently than PINUP. These unequal performances of $e$FindSite$^{PPI}$ and PINUP can be explained by differences in their prediction techniques. $e$FindSite$^{PPI}$ mainly exploits template-target similarities using global structure alignments, which are fairly insensitive to local distortions in the target proteins, whereas PINUP employs local features, e.g. side chain conformations of individual amino acids as well as solvent accessible surface calculations to predict interface residues. Despite the correct global topology, the local characteristics of computer-generated models may deviate significantly from experimental structures, decreasing the performance of PINUP in binding interface prediction using non-native target conformations.

Next, we compare the performance of $e$FindSite$^{PPI}$ and PINUP separately for 3,896 homo- and 465 heterodimers identified in the BM4361 dataset. Table 2.3 shows that both algorithms perform better on homodimers compared to heterodimers; MCC for $e$FindSite$^{PPI}$ (PINUP) is 0.419 (0.187) for homo- and 0.289 (0.156) for heterodimers. Furthermore, consistent with previous results, $e$FindSite$^{PPI}$ is roughly twice as sensitive as PINUP on both datasets of dimers. We note that the performance of algorithms for PPI site prediction is often different on homo- and heterodimers; for example, Englen et al. [21] reported that the average performance of iJET and ET [20] were better on homodimers compared to
heterodimers. This is due to the fact that homodimers often have a nearly perfect symmetric organization at the interface in contrast to mainly asymmetric interfaces in heterodimers.

Table 2.3. Comparison of the performance of eFindSitePPI and PINUP using homodimers and heterodimers from the BM4361 dataset. Values pointing to the best performance are highlighted in bold, except for FPR and TPR that need to be considered jointly.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Predictor</th>
<th>Evaluation metrica</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FPR</td>
</tr>
<tr>
<td>Homodimer</td>
<td>eFindSitePPI</td>
<td>0.088</td>
</tr>
<tr>
<td></td>
<td>PINUP</td>
<td>0.089</td>
</tr>
<tr>
<td>Heterodimer</td>
<td>eFindSitePPI</td>
<td>0.093</td>
</tr>
<tr>
<td></td>
<td>PINUP</td>
<td>0.090</td>
</tr>
</tbody>
</table>

a FPR: false positive rate; TPR: sensitivity; ACC: accuracy; SPC: specificity; PPV: precision; MCC: Matthew's correlation coefficient.

Comparison with PrISE

In order to eliminate any potential prediction bias using one dataset, we evaluate the performance of eFindSitePPI with respect to other methods on different protein sets. In addition to PINUP, we compare eFindSitePPI with PrISE, a recently developed method that exploits local surface similarities to predict protein interfaces [25]. This method extracts structural elements from a target protein and scans them through two databases of protein quaternary structures and protein-protein interface residues, ProtInDB [56] and PQS [57]. The accuracy of PrISE was previously evaluated using the Protein-Protein Docking Benchmark dataset [58]. We ran eFindSitePPI on the Benchmark 4.0 dataset following the same procedure as used in PrISE benchmarking [25]. In this analysis, we also include results from PINUP reported for the Benchmark 4.0 dataset. Table 2.4 shows that eFindSitePPI
outperforms both PrISE and PINUP; for example, the accuracy (MCC) is 0.909 (0.352), 0.790 (0.279) and 0.783 (0.246), respectively. Moreover, Benchmark 4.0 also provides apo structures for most of the target proteins; we use these conformations to evaluate the performance of eFindSite\textsuperscript{PPI} against unbound experimental structures to complement our previous analysis using protein models from the BM1905 dataset. The accuracy of eFindSite\textsuperscript{PPI} against bound and unbound structures is fairly comparable; using apo conformations only slightly decreases the sensitivity by 0.022 and MCC by 0.014. Thus, eFindSite\textsuperscript{PPI} performs better than other predictors on the Benchmark 4.0 dataset offering a high prediction accuracy using both bound as well as unbound experimental target conformations.

Table 2.4. Comparison of the performance of eFindSite\textsuperscript{PPI}, PINUP and PrISE on the Benchmark 4.0 dataset. Values pointing to the best performance are highlighted in bold, except for FPR and TPR that need to be considered jointly.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Predictor\textsuperscript{a}</th>
<th>Evaluation metric\textsuperscript{b}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FPR</td>
</tr>
<tr>
<td>Bound</td>
<td>eFindSite\textsuperscript{PPI}</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>PINUP</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>PrISE</td>
<td>0.042</td>
</tr>
<tr>
<td>Unbound</td>
<td>eFindSite\textsuperscript{PPI}</td>
<td>0.047</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Results for PINUP and PrISE are taken from ref. [25]. \textsuperscript{b} TPR: sensitivity; ACC: accuracy; PPV: precision; MCC: Matthew’s correlation coefficient.
Comparison with ET and iJET

Finally, we compare eFindSite\textsuperscript{PPI} to evolution-based predictors, ET and iJET [20][21]. Inspired by the Evolutionary Trace approach [20], these methods identify PPI interfaces by detecting and analyzing conserved surface patches on target proteins. Evolutionary conservation is the primary feature for the identification of interface residues by both algorithms, as it reflects the evolutionary selection at interfacial sites to maintain the molecular function across protein families. The comparison with ET and iJET is based on the interface residue prediction for 52 protein chains derived from the Huang dataset [39]. The targets are experimental structures in their bound conformational state and cover three categories of protein-protein interactions: non-transient homodimers, non-transient heterodimers and transient complexes.

Table 2.5 summarizes the performance of eFindSite\textsuperscript{PPI}, ET and iJET in terms of sensitivity, specificity, precision and accuracy. Clearly, eFindSite\textsuperscript{PPI} produces quantitatively better results than ET and iJET across all targets. For instance, the sensitivity of eFindSite\textsuperscript{PPI} is 28.9% (33.8%), 20.8% (14.6%) and 21.2% (7.6%) higher than ET (iJET) on homo-, heterodimers and transient complexes, respectively. However, despite a lower sensitivity for the transient complexes, iJET gives 7.8% higher precision compared to eFindSite\textsuperscript{PPI}. This analysis also shows that similar to ET and iJET, the performance of eFindSite\textsuperscript{PPI} decreases from non-transient homodimers to heterodimers to transient complexes. This is consistent with other studies demonstrating that, in contrast to proteins forming transient complexes, the prediction of non-transient interfaces is less complicated, because they are evolutionarily more conserved, larger and flatter [59][39].
Table 2.5. Comparison of the performance of eFindSite<sup>PPI</sup>, ET and iJET using non-transient homo- and heterodimers as well transient complexes from the ET/iJET dataset. Values pointing to the best performance are highlighted in bold, except for FPR and TPR that need to be considered jointly.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Predictor&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Evaluation metric&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FPR</td>
</tr>
<tr>
<td></td>
<td>eFindSite&lt;sup&gt;PPI&lt;/sup&gt;</td>
<td>0.049</td>
</tr>
<tr>
<td>Homodimer</td>
<td>ET</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>iJET</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>eFindSite&lt;sup&gt;PPI&lt;/sup&gt;</td>
<td>0.071</td>
</tr>
<tr>
<td>Heterodimer</td>
<td>ET</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>iJET</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>eFindSite&lt;sup&gt;PPI&lt;/sup&gt;</td>
<td>0.048</td>
</tr>
<tr>
<td>Transient</td>
<td>ET</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>iJET</td>
<td>0.030</td>
</tr>
</tbody>
</table>

<sup>a</sup> Results for ET and iJET are taken from ref. [21]. <sup>b</sup> TPR: sensitivity; PPV: precision; SPC: specificity; ACC: accuracy.

Case studies

To illustrate the prediction performance of eFindSite<sup>PPI</sup>, we discuss a couple of representative examples. We note that these proteins are not present in the BM4361 dataset, thus have not been used in the construction of machine learning models. The first case study involves a NAD-dependent D-glycerate dehydrogenase (GDH) from *H. methylovorum* (PDB-ID: 1gdh). This enzyme belongs to the family of oxidoreductases and catalyzes the NADH-linked reduction of 3-hydroxypyruvate to D-glycerate in the serine pathway for the assimilation of one-carbon compounds in methylotrophs [60]. The GDH molecule forms a
homodimer composed of two structurally similar subunits related to each other by a 2-fold symmetry [61]. Figure 2.8 presents the PPI interface predicted for a GDH monomer by eFindSitePPI from remotely homologous templates. 59% of interfacial residues are correctly identified, with 0.992 specificity, 0.951 precision, and 0.909 accuracy (Figure 2.8A). Moreover, eFindSitePPI correctly predicted 7 out of 16 hydrogen bonds as well as 2 out of 5 salt bridges present at the GDH interface. Figure 2.8B illustrates selected correctly identified interactions, including a salt bridge between the side chains of R129-chain A and D277-chain B, and hydrogen bonds between the side chain of R127-chain A and T281-chain B.

Figure 2.8. Example of PPI prediction by eFindSitePPI for a homodimer (PDB-ID: 1gdh). (A) The surface representation of a monomer chain; true positives, true negatives, false positives, and false negatives are colored in green, gray, red, and cyan, respectively. (B) Interface residues correctly predicted to form specific interactions; dashed blue lines represent salt bridges and red lines represent hydrogen bonds.
The second example is a mouse T cell receptor protein (TCR) (PDB-ID: 1tcr), which is localized on the surface of T cells and is responsible for their activation [62]. These molecules participate in the recognition of antigens bound to major histocompatibility complexes [63][64]. TCR is a membrane-anchored heterodimer composed of alpha and beta chains [65]; we use eFindSite\textsuperscript{PPI} to predict interfacial residues separately for both chains. Figure 2.9 shows that eFindSite\textsuperscript{PPI} correctly identified 65\% of interfacial residues in chain alpha, with 0.946 specificity, 0.420 precision, and 0.929 accuracy (Figure 2.9A). For chain beta, 46\% of interfacial residues are correctly predicted, with 0.815 specificity, 0.959 precision, and 0.817 accuracy (Figure 2.9B). Importantly, most false positives and false negatives in both chains are located at the rim of interface patches, thus the prediction of the core interfacial residues is highly accurate. This is evident in Figure 2.9C, which shows the heterodimer structure composed of alpha and beta chains interacting via two interfaces. Residues overpredicted and missed by eFindSite\textsuperscript{PPI} are mainly positioned either within the interfacial cavity or at the interface edge, whereas those predicted correctly make up the core of the TCR alpha-beta interface. Furthermore, eFindSite\textsuperscript{PPI} accurately identified 3 out of 6 interfacial hydrogen bonds and 1 out of 2 salt bridges stabilizing the dimer complex according to the experimental structure. Figure 2.9D illustrates two correctly predicted interactions: a salt bridge between the side chains of D137-alpha and R187-beta, and a hydrogen bond between the main-chain of D157-alpha and the side chain of Y173-beta. These examples demonstrate the capability of eFindSite\textsuperscript{PPI} to predict PPI sites, residues, and interaction types for homo- as well as heterodimers using weakly homologous templates.
Figure 2.9. Example of PPI prediction by eFindSite\textsuperscript{PPI} for a heterodimer (PDB-ID: 1tcr). The surface representations of alpha and beta chains are shown in (A) and (B); the dimer complex is displayed in (C). True positives, true negatives, false positives, and false negatives are colored in green, gray/tan, red, and cyan, respectively. Interfacial residues in both chains correctly predicted to form specific interactions are shown in (D). Dashed blue lines represent salt bridges and red lines represent hydrogen bonds.

**CONCLUSION**

The analysis of evolutionarily weakly related dimer proteins reported in this study strongly suggests that the locations of their binding sites are highly conserved, irrespectively
of the global structure similarity of protein-protein complexes. Furthermore, the interfacial geometry is preserved as well, thus can be predicted with a high accuracy. This is consistent with previous studies demonstrating that surface regions responsible for protein binding are conserved among structural neighbors [26]. Exploiting these insights, we developed eFindSite\textsuperscript{PPI}, a new approach for the prediction of protein binding sites using information derived from evolutionarily and structurally related templates.

\textit{eFindSite}\textsuperscript{PPI} employs sensitive meta-threading by eThread [35] to identify evolutionarily related templates and extensively uses various machine learning techniques to detect interfacial residues on a query protein surface. A higher degree of conservation of local interface compared to the global structure of protein complexes forms the basis for an accurate prediction of interfacial binding sites. In addition to these conservation patterns, eFindSite\textsuperscript{PPI} also employs other residue-level descriptors to effectively discriminate between interfacial and non-interfacial residues. For instance, it incorporates the relative solvent accessible area and the interfacial propensities of amino acids, which have been already successfully used by several other interfacial site prediction algorithms [66][24]. A high accuracy in extracting structural information from the “twilight zone” templates motivated us to further extend the capabilities of eFindSite\textsuperscript{PPI} to predict specific interactions as well. That is, eFindSite\textsuperscript{PPI} also detects the types of molecular interactions that target proteins are likely to form with their interacting partners; this is demonstrated for hydrogen bonds, salt bridges as well as hydrophobic and aromatic contacts. Comparative benchmarking calculations on several datasets of protein dimers show that eFindSite\textsuperscript{PPI} outperforms other methods for protein binding residue prediction. Equally important, it is designed to work with protein models, so that the interfacial site can be efficiently predicted even when the
experimental structure of a query protein is unavailable. Finally, a carefully tuned confidence estimation system identifies those predictions that are likely to be correct. eFindSitePPI is freely available to the academic community as a user-friendly web-server and a well-documented stand-alone software distribution on our website which also provides all benchmarking datasets and results reported in this paper.

Website: http://www.brylinski.org/efindsiteppi

SUPPLEMENTARY INFORMATION

![Graphs](image)

Figure 2.10. Quality of structure models generated by eThread for proteins in the BM4361 dataset. (A) The distribution of estimated TM-score values, (B) the correlation between the estimated and real TM-score values calculated against crystal structures. Dotted lines delineate a TM-score of 0.4.

**Computer-generated protein models**

Weakly homologous structure models were constructed for the BM4361 dataset using eThread excluding closely related templates whose sequence similarity to the target is >40%. We assembled up to 20 models for each receptor target, 10 using eThread/Modeller
and 10 using eThread/TASSER-Lite. Figure 2.10 A shows that for the majority of BM4361 proteins, the top-ranked models are confidently predicted; the estimated TM-score is >0.7 and 0.4-0.7 for 45% and 28% of the models, respectively. In addition to the template-based assembly of full-length structures, eThread provides reliable confidence estimates for the quality assessment. As demonstrated in Figure 2.10 B, these estimates correlate well with the actual TM-score values calculated versus experimental structures. The Pearson correlation coefficient between real and estimated TM-score values is 0.84, which is consistent with previous benchmarking results.

From the pool of conformations generated for BM4361 proteins, we compiled two sets of non-native target structures. For each receptor protein, a high-quality model with a TM-score to native of >0.7 was randomly selected; similarly, a structure with a TM-score to native within the range of 0.4-0.7 was selected as the moderate-quality model. Both models with the preferred accuracy were constructed for 1,905 target proteins, thus the resulting datasets are referred to as BM1905H and BM1905M. The characteristics of the BM1905H and BM1905M datasets are summarized in Table 2.6. The former comprises high-quality structures, most of which were generated by eThread/TASSER-Lite with an average TM-score to native of 0.76. The latter contains moderate-quality models with an average TM-score of 0.53; roughly three-quarters of these structures were constructed by eThread/Modeller. In addition, we also compiled a corresponding set of experimental structures, BM1905C. We note that BM1905 contains fewer proteins than BM4361 because of two reasons. First, no models with the preferred quality were generated for a subset of targets. Second, for some models, particularly those with moderate-quality structures, structurally similar dimer templates at a TM-score of 0.4 are unavailable. Nevertheless, three
non-redundant at 40% sequence identity BM1905 datasets provide a sufficient number of targets to perform a thorough assessment of the structure-based prediction of PPI sites using protein models.

Figure 2.11. Percentage of high, medium and low confidence predictions for (A) BM1905C, (B) BM1905H and (C) BM1905M datasets. Low, medium and high confidence predictions are colored in white, light gray and dark gray, respectively.

Table 2.6. Percentage of models constructed by a particular protocol and the structure quality of two datasets of protein models used in addition to crystal structure as targets for PPI site prediction.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Modeller (%)</th>
<th>TASSER-Lite (%)</th>
<th>TM-scorea</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM1905H</td>
<td>35.6</td>
<td>64.4</td>
<td>0.76 ±0.05</td>
</tr>
<tr>
<td>BM1905M</td>
<td>73.2</td>
<td>26.8</td>
<td>0.53 ±0.08</td>
</tr>
</tbody>
</table>

a Mean ± standard deviation.
Figure 2.12. Amino acid composition of residue positions involved in the formation of (A) hydrogen bonds, (B) salt bridges, (C) hydrophobic, and (D) aromatic contacts at the experimental and predicted interfaces.

REFERENCES


CHAPTER 3: COMPARATIVE ANALYSIS OF BINDING SITE PREDICTION METHODS *

INTRODUCTION

Proteins do not operate in isolation, rather they interact with each other either directly or indirectly to carry out their functions [1]. In fact, protein-protein interactions (PPIs) play a pivotal role in cellular functions mediating virtually all biological processes. Therefore, significant efforts have been devoted to characterize and catalog PPIs to improve our understanding of molecular recognition and reveal the mechanisms by which proteins work. Mapping these interactions facilitates the modeling of the entire functional proteome and its constituent pathways. Moreover, linking PPIs to diseased states and other phenotypes helps develop drugs that directly target protein-protein interfaces [2][3]. Computationally inferred information about interfacial residues also aids the design of mutants for the experimental verification of interactions [4][5] as well as it enhances the prediction of complex structures through homology modeling and protein docking [6–8].

Given that numerous biological applications require information about surface regions involved in PPIs, a wide range of experimental techniques have been designed to identify interfacial residues, with much efforts devoted to the development of high-throughput methods [9–11]. Nonetheless, these techniques are often tedious, labor intensive and are associated with high costs of experiments. In addition, many experimental techniques have been shown to suffer from high false positive and false negative rates, as well as inter-study discrepancies [12–14].

*This chapter previously appeared as: Surabhi Maheshwari and Michal Brylinski, ‘Predicting Protein Interface Residues Using Easily Accessible on-Line Resources’, Briefings in Bioinformatics, 16.6 (2015), 1025–34. It is reprinted by permission of the publisher, see the permission letter for proper acknowledgment phrase.
On the other hand, the ongoing proteomics and structural genomics research routinely generates massive amounts of data, which need to be interpreted at a fast pace. Hence, there is a dire need for computational methods to effectively identify PPIs, and to assess, validate and scrutinize experimentally collected data. One of the first attempts to predict residues located at the interface was made by Jones and Thornton [15]. Since then a number of methods for predicting protein-protein interface residues have been reported. These approaches use diverse techniques for the identification of PPI sites and may vary in terms of the attributes used to distinguish interacting sites and the implemented learning/prediction algorithms [16–19]. In general, computational methods can be broadly divided into sequence- and structure-based approaches. Sequence-based methods often use sliding window frames in order to calculate the specific features associated with residues based on their neighbors [20–22]. These methods employ various residue-level properties, such as the degree of evolutionary conservation, physicochemical features, energetics, etc., to construct scoring functions. Furthermore, the availability of protein tertiary structures allows for the integration of a variety of structural information, e.g. solvent accessibility, B-factors, and secondary structure, to improve the prediction accuracy [23].

Many recently published reviews provide insights into the fundamentals of protein binding and docking and discuss the mechanics of PPI prediction. Zhou and Qin give a comprehensive overview of the underlying principles used by different methods and discuss the challenges faced by the community [24]. Vries and colleagues provide a critical assessment of the state-of-the-art in PPI prediction, compare different approaches, and explain difficulties in assessing the absolute and relative performance of various predictors due to differences in the choice of data and evaluation criteria [25]. A review by Ezkurdia et
al. examines the weak points of current PPI prediction methods arising from the incomplete structural information on transient complexes, which remain largely under-represented in the Protein Data Bank (PDB) [26]. Finally, Wang and colleagues focus on machine learning-based techniques and outline the key components of an effective prediction pipeline to infer protein interaction sites [19]. Since the majority of research studies concentrate on the experimental structures of target proteins in their bound and unbound conformations, significantly fewer reviews touch on issues related to using protein models in the structure-based prediction of PPI sites. Certainly, the unavailability of structural data may impose constraints on research projects involving PPIs. Using protein models mitigates this issue, however, assuming that PPI prediction methods tolerate imperfections in the target structures. Therefore, in this communication, we describe ten freely accessible web servers for PPI site prediction and comparatively evaluate their performance on a common dataset assessing the effect of using computer-generated models on the prediction accuracy.

Types of protein complexes

Protein-protein complexes can be divided into obligatory and transient assemblies based on their overall interaction strength and stability. Obligatory complexes are functional only in their coupled state, and the monomers do not exist as stable structures in vivo. The interaction partners also have a high shape complementarity and their interface residues resemble the hydrophobic core of globular proteins. In contrast, transient complexes are formed by proteins that may be functional even in their unbound monomeric state. The interface of such complexes is stabilized by weak interactions, the partners have a lower geometrical complementarity, and the interface area between them is relatively small compared to obligatory complexes. Also, the hydrophobicity of residues that make up the
interface of transient associations is indistinguishable from the remaining protein surface. With respect to the sequence identity between monomers, protein assemblies can be divided into homo- and hetero-complexes. The former consist of two or more identical chains, while the latter are composed of protein chains with different amino acid sequences. Obligatory associations can be homo- and hetero-complexes, whereas, the majority of transient assemblies are hetero-complexes that comprise different chains. In general, interfacial sites in obligatory complexes are easier to detect as they are generally larger, flatter, more hydrophobic and more conserved than transient interfaces [27–29].

**Interfacial regions of protein surface**

Proteins interact with one another via interfacial sites predominantly composed of surface residues. Interface residues tend to be more conserved than other positions, however, this signal is weakened for residues below a certain solvent accessibility. Therefore, the definition of surface residues plays a pivotal role in the creation of databases for methods exploiting evolutionary conservation. The prediction accuracy also strongly depends on how surface residues are defined; as a common practice, residues are classified as surface residues if their relative solvent accessibility (RSA) is above some threshold. Different studies use different cutoffs, which typically range from 5% to 16% [26,28] with higher thresholds leading to a lower number of solvent-exposed residues. Based on the three-dimensional structure of a protein complex, PPI sites are identified from the subset of surface residues either using interatomic distances between non-hydrogen atoms in different protein chains, or by calculating the change in the solvent accessible area upon complex formation. In both cases, empirically optimized thresholds are often used; for instance, distance-based methods typically use cutoff values of 4Å, 4.5Å or 5Å [30][31][32],
whereas surface-based approaches define interfacial residues as those whose accessible surface area changes by more than 20Å² [28].

**Characteristic features of interface residues**

Comparison of interfacial and non-interfacial regions on protein surfaces reveals a number of intrinsic characteristics of residues involved in the formation of quaternary structures. These features are commonly used by PPI prediction algorithms, and can be broadly classified into the following three categories:

- **Sequence-based features** are derived from the amino acid sequence alone and use various physicochemical properties of residues to identify the interface regions. Examples of these features are interface propensity [33][34], hydrophobicity and electrostatic desolvation [35], as well as structural attributes predicted from sequence, such as secondary structure and solvent accessibility [23][36].

- **Structure-based features** are derived from the tertiary structures of target proteins. These attributes include, but are not limited to, solvent accessible surface area [37,38], secondary structure [39], crystallographic B-factors [40], local geometries [41], as well as the spatial distribution of hydrophobic and polar surface patches [42].

- **Evolutionary features** are calculated by comparing the sequence of a query protein to the sequences of its homologs. Interface residues tend to be highly conserved, in contrast to non-interfacial surface residues that are subjected to a notably lower selection pressure [43][44]. Thus, the sequence conservation reflects the evolutionary selection at interfacial sites to maintain protein function. These attributes have a high discriminatory power towards interfacial residues; for
example, protein sequence entropy is a conservation score that estimates sequence variability, thus it is often used in PPI site prediction [45].

**Feature integration and the prediction of PPI sites**

While a number of discriminatory features have been explored, individual attributes provide only a weak signal, thus no single feature can be used to unambiguously identify the interaction regions in proteins [24]. Since these attributes may provide a complementary discriminatory power with respect to each other, many PPI residue predictors combine different features in order to more effectively identify interfacial regions. Individual features are often integrated using scoring functions and machine learning techniques. The optimization of a relatively small number of attributes can be done by constructing a discriminant function that either linearly or non-linearly combines individual features [15,28,38,46]. More recently, machine learning strategies have become popular, especially for the optimal combination of a large number of attributes. The most commonly used machine learning algorithms include Neural Networks (NNs) [17,20,32,47], Support Vector Machines (SVMs) [30,48][31], Random Forests (RFs) [22], and Naïve Bayesian Classifiers (NBCs) [39].

Most PPI site predictors fall into two major categories, residue- and patch-based methods. Residue-based techniques assign each residue in the target protein with a score corresponding to the probability to be a part of the interface [39,49,50] [31]. These residues need not necessarily be adjacent on the protein surface, however, clustering algorithms are sometimes used to impose a spatial proximity. The output from such methods often contains raw interface/non-interface scores calculated for all residues in the target protein as well as a separate list of predicted interface residues that have their score above some pre-defined
threshold. Methods that employ machine learning usually adopt the residue-based approach as the input data can be conveniently mapped to the feature space. On the other hand, patch-based methods partition a target protein surface into a set of discrete patches/clusters [15]. These surface patches are then analyzed and ranked based on a combined score calculated using individual features with the top-ranked group taken as the predicted interface. In addition to interface/non-interface scores assigned to individual residues, the output from patch-based methods often contains a confidence score derived for the entire cluster of residues. A weak point of many patch-based strategies is that the predicted patches are generally circular, whereas biological interfaces tend to be rather irregular in shape. Furthermore, these methods also require estimating the size of a putative interfacial site, nevertheless, this information can be reliably obtained from a correlation between the number of interfacial residues and the target protein length [15][51].

**Intrinsic disorder in protein interactions**

While the main focus of this review is on the structure-based prediction of interface residues, other methods for the identification of PPIs involving intrinsically disordered proteins attract significant attention owing to the fact that the flexibility and disorder play an important role in molecular recognition. Briefly, the term “intrinsic disorder” refers to those proteins and protein segments that fail to self-fold into fixed tertiary structures [52]. Attributed to unique characteristics of interactions mediated by intrinsically disordered proteins, the involvement of disordered regions in complex PPI networks has become increasingly apparent in recent years. For instance, these molecules can recognize multiple partners upon the adoption of different conformations contributing to binding diversity [53]. Moreover, due to a relatively lower binding affinity compared to classical binding,
interactions involving intrinsically disordered segments are fully reversible while maintaining the high specificity [54]. Interestingly, binding motifs located in longer intrinsically disordered protein regions, called Molecular Recognition Features (MoRFs), undergo disorder-to-order transitions upon binding [55]. Several prediction methods have been developed to identify MoRFs from protein primary sequence, e.g. SLiMPred [56], MoRFpred [57], and ANCHOR [58]. The implications of the protein intrinsic disorder in molecular recognition and binding functions are comprehensively discussed in a recent review [59].

Web server for PPI site prediction

A number of algorithms for PPI site prediction are freely available to the scientific community as user-friendly web servers. Here, we selected ten resources (listed in Table 3.1) that represent a variety of methods and were up and running at the time of this study. Moreover, these web servers offer a possibility to process datasets of moderate sizes in the order of a couple of hundreds of proteins using either web-based interfaces or command line tools that can query remote services. The selected web servers are arranged in four groups: (I) primarily sequence profile-based techniques that additionally use the accessible solvent area (ASA), (II) those approaches using residue-level characteristics, (III) algorithms employing sub-residue physicochemical and structural features, and (IV) template-based methods that incorporate global structure alignments. Below, we review the design of individual web servers according to this classification.

Group I. We assigned two algorithms to this group, cons-PPISP and PSIVER. The original PPISP (Protein-Protein Interaction Site Predictor) algorithm [47] was developed to effectively exploit evolutionary information from sequence profiles constructed by PSI-
BLAST [60] and the residue solvent exposure calculated by DSSP [61]. It uses an NN classifier, in which the nodes are fed with a series of scores including those calculated for spatial neighbors on the protein surface. It is noteworthy that PPISP was demonstrated to maintain its accuracy when unbound structures are used as the targets for interfacial residue prediction. The problem of over- and under-predictions was subsequently addressed by using consensus classification by multiple NN models. This improved method, called cons-PPISP, employs a series of models ranging from a high accuracy with low coverage to a low accuracy with high coverage, and a new procedure for the spatial clustering of predicted interface residues [32]. Cons-PPISP not only offers a higher accuracy at an increased coverage compared to PPISP, but also shows a good agreement with experimental data as demonstrated for several proteins whose protein-protein complexes were characterized by NMR chemical shift perturbation.

The second method in this group is a sequence-based approach, PSIVER (Protein-protein interaction SItes prediction seVER) [36]. It employs an NBC and a set of sequence features to predict protein interaction sites, focusing on transient and heterodimer complexes. Two separate classification models are implemented in PSIVER for sequence profiles obtained from PSI-BLAST [60] and ASA. Since PSIVER is a sequence-based method, rather than calculating ASA directly from structure, these values are predicted for target sequences using SABLE [62]. Both NBCs calculate conditional probabilities using the kernel density estimation method. Leave-one-out cross-validation demonstrated that combining individual sequence profile- and ASA-based classifiers significantly improves the overall performance of PSIVER. Evaluated on an independent dataset of proteins selected from the
Protein Docking Benchmark Set 3.0 [63], PSIVER outperformed the ISIS server [29] and the sequence-based version of SPPIDER [17].

**Group II.** Among many residue-level attributes, interface propensities derived for individual amino acids are frequently used in interfacial residue prediction, as exemplified by several methods in this group. For instance, *InterProSurf* [37] employs interfacial propensities for amino acids calculated from a dataset of 72 dimer structures [64]. Different from other approaches, InterProSurf first partitions the target protein surface defined by the GetArea program [65] using either a cluster or a patch analysis, and then applies a scoring function to find surface regions with high interface propensities. The number of high-ranking clusters in the clustering method and a radius in the patch analysis were optimized empirically to balance the sensitivity and precision of interface residue prediction. In addition to benchmarking simulations, InterProSurf successfully predicted interaction sites for the Anthrax toxin and measles virus hemagglutinin protein as validated by sequence analysis and mutagenesis experiments [37].

*SPPIDER* (Solvent accessibility based Protein-Protein Interface iDEntification and Recognition) [17] is an NN method that uses a set of 19 attributes derived from the sequence and structure of a query protein, and its evolutionary profiles. Predicted solvent accessibility fingerprints are a novel feature implemented in SPPIDER. Interestingly, the difference between the observed and predicted ASA is highly informative and can be used to increase the predictive power of solvent accessibility-based features. The integration of the enhanced RSA predictions by SABLE [66] with high-resolution structural data led to the development of RSA-based fingerprints of protein interactions, which were found to significantly improve the discrimination between interacting and non-interacting sites. Similar to cons-PPISP,
SPPIDER is a consensus-based classifier that combines ten cross-validated NN models with a $k$-nearest neighbor selection procedure to filter out misclassified residues.

A recent study indicated that Voronoi diagrams provide more accurate description of the exposed residue environment than techniques based on Euclidian distances and sequence sliding windows [40]. This observation led to the development of VORFFIP (Voronoi Random Forest Feedback Interface Predictor), a novel method for protein binding site prediction. It integrates heterogeneous data including various residue-level structural and energetic characteristics, the evolutionary sequence conservation calculated by AL2CO [66], and crystallographic B-factors. VORFFIP employs a two-step RF classifier and a set of residue- and environment-based features to assign surface residues with interfacial scores. Cross-validation benchmarks performed on a dataset derived from the Protein Docking Benchmark Set 3.0 [63] demonstrated that combining different features with Voronoi diagrams used as the environment descriptor yields the best performance. VORFFIP was also found to outperform other methods for binding interface prediction, SPPIDER [17] and WHISCY [46].

The last method in this group, WHISCY (What Information does Sequence Conservation Yield?) [46] employs a linear regression (LR) method to combine residue conservation and structural information to effectively discriminate between interfacial and non-interfacial residues. The conservation is computed from multiple sequence alignments obtained from the HSSP database [67]. WHISCY takes into account structural information such as interface propensities and considers the properties of surface neighbors in order to remove isolated high-scored residues. The implemented simple LR model offers a high flexibility by allowing users to choose which characteristic should be included in the
prediction procedure. In a validation study, WHISCY and ProMate [39] were used to generate input for a data-driven protein docking program, HADDOCK [68]. Near-native structures constructed by docking simulations using unbound receptor conformations from the Protein Docking Benchmark Sets 1.0 and 2.0 [51][69] demonstrate that incorporating the predicted PPI sites in data-driven docking yields an improved accuracy of the protein quaternary structure modeling.

**Group III.** Statistical properties are usually derived for individual amino acids, however, these can be also calculated at the sub-residue level of atomic groups. For example, PIER (Protein IntErface Recognition) [28] applies a partial least square regression (PLS-R) algorithm to optimize desolvation parameters [70] for 12 significant atomic groups whose ASA is calculated by ICM [71]. PIER initially divides the surface of a target protein into a set of individual patches. In the alignment-independent mode, a decision score indicating the likelihood of being at the protein interaction site is computed as a linear combination of the physical descriptors. Furthermore, sequence alignment information was incorporated in order to evaluate the strength of evolutionary signal. Specifically, in the alignment-dependent mode, surface patches are additionally assigned several features calculated from sequence alignments constructed by the Zero End-gap Global Alignment (ZEGA) method [72]. Interestingly, adding evolutionary information only marginally influenced the prediction performance of PIER and for certain classes of proteins, the evolutionary signal even deteriorated the prediction accuracy [28].

Atomic level descriptors are implemented in ProMate [39], an NBC method that identifies interface regions using composite probabilities derived from protein sequences and structures. ProMate employs Connolly's MS program [73] to identify surface atoms,
which are subsequently extended to so-called circles. In order to classify these regions as interfacial, non-interfacial, or boundary, an optimal combination of scoring terms was identified from a set 13 different properties comprising the chemical composition of binding interfaces, geometric properties, and specific information obtained from crystallographic data. Based on this classification, the neighboring circles are merged and clustered to predict interface patches. The algorithm was demonstrated to successfully predict the interface location for the majority of benchmarking transient hetero-complexes. Importantly, the identified biophysical properties were found to be largely independent of a particular receptor conformation, therefore, the success rate of ProMate was almost equal for target proteins experimentally solved in their bound and unbound states.

**Group IV.** The last group of methods for protein interface residue prediction comprises template-based predictors, eFindSitePPI and PredUs. eFindSitePPI capitalizes on the tendency of the location of binding sites to be highly conserved across evolutionarily related protein dimers [31]. It employs a collection of effective algorithms, including meta-threading by eThread [74], structural alignments by Fr-TM-align [75], and machine learning using SVMs and NBCs [76]. Each residue in a query protein is assigned a probability to be at the interface using residue-level attributes as well as structure and sequence conservation scores derived from evolutionarily related complexes. In addition, eFindSitePPI effectively detects specific molecular interactions at the interface, such as hydrogen bonds, aromatic interactions, salt bridges and hydrophobic contacts. Previous comparative benchmarks demonstrated that it outperforms PINUP (Protein INterface residUe Prediction) [38] using experimental protein structures as well as computer-generated models. The performance of eFindSitePPI was also better than several other PPI site prediction programs, including PrlSE (Prediction of
protein-protein Interface residues using Structural Elements) [41], ET (Evolutionary Trace) [21] and JET (Joint Evolutionary Trees) [77].

Interface conservation is most significant among proteins that have a clear evolutionary relationship, however, it has been shown that a notable level of conservation exists among remote structural neighbors as well [49]. These structural insights are exploited by PredUs, a structure-based method that predicts surface residues likely to participate in the binding of other proteins [78]. For a given protein of interest, PredUs employs a structure alignment program Ska [79] to identify those structural neighbors forming complexes according to the Protein Quaternary Structures database [80] and the PDB [81]. Interfaces from neighbors are used to calculate contact frequencies, which along with ASAs computed by SURFace [82] make a feature vector for SVMs [83]. PredUs offers several unique interactive features so that a prediction can be tailored to a particular hypothesis. For example, users can upload the structure of a binding partner to include structural neighbors of the partner in PPI residue prediction. Moreover, since proteins may interact with different partners at distinct regions to perform various molecular functions [84], the list of structural neighbors can be filtered based on functional information according to Gene Ontology [85], Structural Classification of Proteins [86], Pfam [87], and InterPro [88]. Comparative benchmarks demonstrated that PredUs outperforms several other algorithms, including PINUP [38], cons-PPIISP [32] and ProMate [39].
Table 3.1. Summary of the design and implementation of ten web servers for the prediction of protein interface residues.

<table>
<thead>
<tr>
<th>Group</th>
<th>Web server</th>
<th>Local features&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Global features&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Classifier&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Clustering</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Propensity level</td>
<td>ASA&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Sequence profiles</td>
<td>Structure alignments</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>Cons-PPISP</td>
<td>-</td>
<td>DSSP</td>
<td>PSI-BLAST</td>
<td>-</td>
<td>NN</td>
</tr>
<tr>
<td></td>
<td>PSIVER</td>
<td>-</td>
<td>SABLE&lt;sup&gt;e&lt;/sup&gt;</td>
<td>PSI-BLAST</td>
<td>-</td>
<td>NBC</td>
</tr>
<tr>
<td>II</td>
<td>InterProSurf</td>
<td>Residue</td>
<td>GetArea</td>
<td>-</td>
<td>-</td>
<td>Product</td>
</tr>
<tr>
<td></td>
<td>SPPIDER</td>
<td>Residue</td>
<td>DSSP</td>
<td>PSI-BLAST</td>
<td>-</td>
<td>NN</td>
</tr>
<tr>
<td></td>
<td>VORFFIP</td>
<td>Residue</td>
<td>DSSP</td>
<td>AL2CO</td>
<td>-</td>
<td>RF</td>
</tr>
<tr>
<td></td>
<td>WHISCY</td>
<td>Residue</td>
<td>NACCESS</td>
<td>HSSP</td>
<td>-</td>
<td>LR</td>
</tr>
<tr>
<td>III</td>
<td>PIER</td>
<td>Sub-residue</td>
<td>ICM</td>
<td>BLAST, ZEGA</td>
<td>-</td>
<td>PLS-R</td>
</tr>
<tr>
<td></td>
<td>ProMate</td>
<td>Atom</td>
<td>Connolly's MS</td>
<td>PSI-BLAST</td>
<td>-</td>
<td>NBC</td>
</tr>
<tr>
<td>IV</td>
<td>eFindSite&lt;sup&gt;PPi&lt;/sup&gt;</td>
<td>Residue</td>
<td>NACCESS</td>
<td>PSI-BLAST</td>
<td>Fr-TM-align</td>
<td>SVM, NBC</td>
</tr>
<tr>
<td></td>
<td>PredUs</td>
<td>-</td>
<td>SURface</td>
<td>-</td>
<td>Ska</td>
<td>SVM</td>
</tr>
</tbody>
</table>

<sup>a</sup> Derived for amino acids, groups of atoms, or individual atoms.  
<sup>b</sup> Derived from sequence or structure alignments of the target protein and its homologs or structural neighbors.  
<sup>c</sup> NN – Neural Network, NBC – Naïve Bayesian Classifier, Product – average interface propensity weighted by ASA, RF – Random Forest, LR – Linear Regression, PLS-R – Partial Least Square Regression, SVM – Support Vector Machines.  
<sup>d</sup> Accessible Solvent Area.  
<sup>e</sup> Predicted from sequence.
MATERIALS AND METHODS

Head-to-head comparison of web servers

Comparing the performance of various algorithms for PPI site prediction reported in literature may not be straightforward as their accuracy was often assessed using different datasets and evaluation metrics. Moreover, most benchmarking studies focus on experimental structures in their bound and/or unbound conformations with significantly fewer assessments carried out for close and remote homology models. Yet, using protein models as the targets in PPI interface prediction is particularly relevant for across-proteome studies, where only sequence information is available for the vast majority of proteins. Therefore, in this review, we include a direct comparison of ten web servers using a common testing dataset composed of experimental and computer-generated structures.

Target proteins were selected from the Protein Docking Benchmark Set 4.0 [89]. We followed similar criteria to those used in our previous study [31], i.e. we excluded multimeric complexes, in which the receptor is either smaller than 50 or larger than 600 amino acids, the interface is made up of less than 20 residues, or multiple interfaces are present. This procedure resulted in a set of 90 target proteins forming heterodimers (42 enzyme/inhibitor or enzyme/substrate, one antibody/antigen, and 47 other complexes). In addition to the experimental structures, we constructed high- and moderate-quality protein models for each target. Specifically, weakly homologous models were generated by eThread [74][90] excluding closely related templates whose sequence similarity to the target is >40%. High-quality models have a TM-score [91] to native of >0.7, whereas the TM-score of moderate-quality models is within a range of 0.4-0.7. These sets of crystal structures, high-, and moderate-quality models are referred to as BM90C, BM90H and BM90M, respectively. We
queried the web servers with all BM90 structures using either web interfaces that allow for multiple target submissions or command-line tools and scripts. Because PSIVER is a sequence-based method, we queried it using BM90 sequences. The predictions were collected and assessed using several commonly accepted evaluation metrics that are derived from a confusion matrix as described below.

**Accuracy measures for PPI residue prediction**

Predicting interfacial residues can be formulated as a binary classification problem, where each protein residue can be either interfacial (positive, $P$) or non-interfacial (negative, $N$). Evaluation of the classification performance generally considers those cases that are correctly and incorrectly predicted for each class, which is quantified by the number of true positives ($TP$), false positives ($FP$), true negatives ($TN$), and false negatives ($FN$). Several metrics are commonly used to represent these four figures as a single measure of the binary classification performance:

- *Accuracy* ($ACC$) evaluates the effectiveness of a predictor by the fraction of correct predictions:

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \quad \text{Eq. 3.1}$$

- *Precision* (also *Positive Predictive Value*, $PPV$) evaluates the fraction of predicted interface residues forming an interface in the experimental complex structure:

$$PPV = \frac{TP}{TP+FP} \quad \text{Eq. 3.2}$$
• **Sensitivity** (also **True Positive Rate**, **TPR**) and **Specificity** (**SPC**) evaluate the effectiveness of the predictor for each class. **TPR** measures the fraction of correctly predicted interface residues, while **SPC** evaluates the fraction of correctly predicted non-interface residues:

\[
\text{TPR} = \frac{TP}{TP+FN} \quad \text{Eq. 3.3}
\]

\[
\text{SPC} = \frac{TN}{FP+TN} \quad \text{Eq. 3.4}
\]

• **Fall-out** (also **False Positive Rate**, **FPR**) evaluates the fraction of predicted interface residues, which are not at the interface:

\[
\text{FPR} = \frac{FP}{FP+TN} \quad \text{Eq. 3.5}
\]

• **Matthew's Correlation Coefficient** (**MCC**) is a measure that balances the sensitivity and specificity, evaluating the strength of the correlation between predicted and the actual classes. Its values range from -1 to 1, where 1 corresponds to a perfect prediction, 0 to a random prediction, and -1 to a perfectly inverse prediction:

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+TN)(FP+FN)(FN+TN)}} \quad \text{Eq. 3.6}
\]
• *Receiver Operating Characteristic (ROC)* plots, representing the relation between *FPR* and *TPR* on a single graph, is another widely used performance assessment method for binary classification problems.

**RESULTS AND DISCUSSION**

**Performance of web servers using experimental structures**

We carried out a comparative assessment of the performance of ten freely available PPI prediction servers using experimental target structures (BM90C) as well as their high- and moderate-quality models (BM90H and BM90M, respectively). Full *ROC* plots were constructed for those servers that provide continuous residue scores; here, we also found the optimal threshold values that maximize *MCC*. Additionally, some servers use post-processing procedures, e.g. clustering and re-ranking, to compile a list of predicted residues, therefore, the performance was also assessed using the default list of predicted interfacial residues when this information was available. For these servers, the better performance (either optimized or default) was used in the comparative analysis.

Table 3.2 shows that using BM90C, the ranking of web servers based on *MCC* is PredUs, *eFindSite*<sup>PPI</sup>, cons-PPISP, SPIDER, ProMate, WHISCY, PIER, VORFFIP, PSIVER, and InterProSurf. PredUs with *MCC* of 0.384 is the best performing server on this dataset, *eFindSite*<sup>PPI</sup> is second with *MCC* of 0.376, and cons-PPISP is third with *MCC* of 0.247. While *MCC* for PredUs is slightly better than that for *eFindSite*<sup>PPI</sup>, *SPC*, *PPV* and *ACC* for *eFindSite*<sup>PPI</sup> are higher than those for PredUs by 0.111, 0.156 and 0.075, respectively. Moreover, we point out that post-processing procedures implemented in several web servers often considerably improve their performance for crystal structures; note that diamonds representing the default predictions in Figure 3.1A are above the corresponding continuous lines calculated
from raw residue scores. For example, the improvement in MCC for SPIDDER (cons-PPISP) on the BM90C dataset is 0.093 (0.078).

**Performance of web servers using computer-generated models**

Nine out of ten web servers described in this review are structure-based methods, i.e. they require the structure of a target protein. The performance of these predictors certainly depends on the quality of input structures. Despite a continuous growth of protein structure databases, there is still a huge gap between the number of known sequences and the number of solved structures. When the experimental structures of query proteins are unavailable, computer-generated models can be used in structure-based PPI residue prediction, however, assuming that the predictor tolerates distortions in modeled structures. In order to assess the impact of the quality of input structures on the prediction accuracy, we submitted high- (BM90H) and moderate-quality (BM90M) models of the target proteins to nine structure-based web servers.

Table 3.2 shows that all predictors give the best performance when experimental structures are used. The prediction accuracy of most algorithms significantly decreases from crystal structures to protein models. Interestingly, the ranking of web servers based on MCC is quite similar for all three BM90 datasets, except for eFindSitePPI, which outperforms PredUs for BM90H and BM90M. For BM90H, the ranking is eFindSitePPI, PredUs, con-PPISP, SPPIDER, PIER, ProMate, WHISCY, PSIVER, VORFFIP, and InterProSurf. Using high-quality models, eFindSitePPI yields the best results with ACC of 0.898 and MCC of 0.340, thus its performance only slightly deteriorates with respect to the BM90C dataset. PredUs is also fairly insensitive to small distortions in the input structures and still gives relatively high ACC of 0.827 and MCC of 0.309, in contrast to the remaining web servers; see Figure 3.1B.
Figure 3.1. ROC plots assessing the accuracy of interface residue prediction by ten web servers across three BM90 datasets. (A) Crystal structures, BM90C; (B) high-quality models, BM90H; and (C) moderate-quality models, BM90M. Continuous ROC lines are calculated using raw residue scores with triangles corresponding to the best performance of raw scores. Default predictions by web servers, including post-processing, are shown as diamonds and circles; circles are used for those web servers that do not provide continuous residue scores. Asterisks mark the accuracy of a pseudo-meta approach that combines the best predictions produced by individual algorithms.
Table 3.2. Comparison of the performance of ten web servers for the prediction of protein interface residues using different quality target structures. For each dataset, web servers are sorted by MCC values. A pseudo-meta approach combines the best predictions produced by individual methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Web server</th>
<th>MCC</th>
<th>TPR</th>
<th>FPR</th>
<th>SPC</th>
<th>PPV</th>
<th>ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM90C</td>
<td>Pseudo-meta</td>
<td>0.481</td>
<td>0.692</td>
<td>0.094</td>
<td>0.905</td>
<td>0.417</td>
<td>0.887</td>
</tr>
<tr>
<td>BM90C</td>
<td>PredUs</td>
<td>0.383</td>
<td>0.701</td>
<td>0.156</td>
<td>0.843</td>
<td>0.302</td>
<td>0.831</td>
</tr>
<tr>
<td>BM90C</td>
<td>eFindSitePPi</td>
<td>0.375</td>
<td>0.396</td>
<td>0.045</td>
<td>0.954</td>
<td>0.459</td>
<td>0.905</td>
</tr>
<tr>
<td>BM90C</td>
<td>cons-PPISP</td>
<td>0.247</td>
<td>0.279</td>
<td>0.052</td>
<td>0.947</td>
<td>0.338</td>
<td>0.888</td>
</tr>
<tr>
<td>BM90C</td>
<td>SPPIDER</td>
<td>0.173</td>
<td>0.340</td>
<td>0.125</td>
<td>0.875</td>
<td>0.208</td>
<td>0.827</td>
</tr>
<tr>
<td>BM90C</td>
<td>WHISCY</td>
<td>0.164</td>
<td>0.130</td>
<td>0.025</td>
<td>0.975</td>
<td>0.334</td>
<td>0.900</td>
</tr>
<tr>
<td>BM90C</td>
<td>PIER</td>
<td>0.118</td>
<td>0.066</td>
<td>0.012</td>
<td>0.987</td>
<td>0.342</td>
<td>0.906</td>
</tr>
<tr>
<td>BM90C</td>
<td>VORFFIP</td>
<td>0.117</td>
<td>0.531</td>
<td>0.401</td>
<td>0.598</td>
<td>0.337</td>
<td>0.579</td>
</tr>
<tr>
<td>BM90C</td>
<td>PSIVER</td>
<td>0.103</td>
<td>0.645</td>
<td>0.463</td>
<td>0.536</td>
<td>0.118</td>
<td>0.546</td>
</tr>
<tr>
<td>BM90C</td>
<td>InterProSurf</td>
<td>0.100</td>
<td>0.435</td>
<td>0.291</td>
<td>0.709</td>
<td>0.163</td>
<td>0.677</td>
</tr>
<tr>
<td>BM90H</td>
<td>Pseudo-meta</td>
<td>0.443</td>
<td>0.680</td>
<td>0.108</td>
<td>0.891</td>
<td>0.380</td>
<td>0.872</td>
</tr>
<tr>
<td>BM90H</td>
<td>PredUs</td>
<td>0.309</td>
<td>0.571</td>
<td>0.147</td>
<td>0.852</td>
<td>0.272</td>
<td>0.827</td>
</tr>
<tr>
<td>BM90H</td>
<td>eFindSitePPi</td>
<td>0.340</td>
<td>0.377</td>
<td>0.051</td>
<td>0.948</td>
<td>0.414</td>
<td>0.898</td>
</tr>
<tr>
<td>BM90H</td>
<td>cons-PPISP</td>
<td>0.207</td>
<td>0.251</td>
<td>0.058</td>
<td>0.941</td>
<td>0.294</td>
<td>0.881</td>
</tr>
<tr>
<td>BM90H</td>
<td>SPPIDER</td>
<td>0.164</td>
<td>0.464</td>
<td>0.216</td>
<td>0.783</td>
<td>0.171</td>
<td>0.755</td>
</tr>
<tr>
<td>BM90H</td>
<td>WHISCY</td>
<td>0.127</td>
<td>0.101</td>
<td>0.023</td>
<td>0.976</td>
<td>0.291</td>
<td>0.899</td>
</tr>
<tr>
<td>BM90H</td>
<td>PIER</td>
<td>0.137</td>
<td>0.234</td>
<td>0.088</td>
<td>0.911</td>
<td>0.204</td>
<td>0.852</td>
</tr>
<tr>
<td>BM90H</td>
<td>ProMate</td>
<td>0.132</td>
<td>0.463</td>
<td>0.278</td>
<td>0.721</td>
<td>0.189</td>
<td>0.689</td>
</tr>
<tr>
<td>BM90H</td>
<td>PSIVER</td>
<td>0.103</td>
<td>0.645</td>
<td>0.463</td>
<td>0.536</td>
<td>0.118</td>
<td>0.546</td>
</tr>
<tr>
<td>BM90H</td>
<td>VORFFIP</td>
<td>0.092</td>
<td>0.681</td>
<td>0.576</td>
<td>0.423</td>
<td>0.284</td>
<td>0.488</td>
</tr>
<tr>
<td>BM90H</td>
<td>InterProSurf</td>
<td>0.075</td>
<td>0.405</td>
<td>0.293</td>
<td>0.706</td>
<td>0.145</td>
<td>0.673</td>
</tr>
<tr>
<td>BM90M</td>
<td>Pseudo-meta</td>
<td>0.290</td>
<td>0.563</td>
<td>0.158</td>
<td>0.841</td>
<td>0.225</td>
<td>0.816</td>
</tr>
<tr>
<td>BM90M</td>
<td>PredUs</td>
<td>0.135</td>
<td>0.366</td>
<td>0.177</td>
<td>0.822</td>
<td>0.165</td>
<td>0.782</td>
</tr>
<tr>
<td>BM90M</td>
<td>eFindSitePPi</td>
<td>0.242</td>
<td>0.303</td>
<td>0.064</td>
<td>0.935</td>
<td>0.312</td>
<td>0.880</td>
</tr>
<tr>
<td>BM90M</td>
<td>cons-PPISP</td>
<td>0.077</td>
<td>0.152</td>
<td>0.076</td>
<td>0.923</td>
<td>0.160</td>
<td>0.855</td>
</tr>
<tr>
<td>BM90M</td>
<td>SPPIDER</td>
<td>0.103</td>
<td>0.645</td>
<td>0.463</td>
<td>0.536</td>
<td>0.118</td>
<td>0.546</td>
</tr>
<tr>
<td>BM90M</td>
<td>WHISCY</td>
<td>0.101</td>
<td>0.571</td>
<td>0.417</td>
<td>0.582</td>
<td>0.162</td>
<td>0.580</td>
</tr>
<tr>
<td>BM90M</td>
<td>PIER</td>
<td>0.096</td>
<td>0.537</td>
<td>0.371</td>
<td>0.628</td>
<td>0.122</td>
<td>0.620</td>
</tr>
<tr>
<td>BM90M</td>
<td>VORFFIP</td>
<td>0.078</td>
<td>0.072</td>
<td>0.025</td>
<td>0.974</td>
<td>0.215</td>
<td>0.895</td>
</tr>
<tr>
<td>BM90M</td>
<td>InterProSurf</td>
<td>0.058</td>
<td>0.625</td>
<td>0.555</td>
<td>0.445</td>
<td>0.245</td>
<td>0.485</td>
</tr>
<tr>
<td>BM90M</td>
<td>InterProSurf</td>
<td>0.034</td>
<td>0.354</td>
<td>0.302</td>
<td>0.697</td>
<td>0.115</td>
<td>0.663</td>
</tr>
</tbody>
</table>

BM90C – crystal structures; BM90H – high-quality models; BM90M – moderate-quality models; FPR – false positive rate; TPR – sensitivity; ACC – accuracy; SPC – specificity; PPV – precision; MCC – Matthew’s correlation coefficient.
For moderate-quality structures from the BM90M dataset, the \( MCC \)-based ranking of web servers is \( e\text{FindSite}^{\text{PPI}} \), PredUs, PSIVER, ProMate, SPPIDER, con-PPISP, WHISCY, PIER, VORFFIP, and InterProSurf. Notably, the performance of most web servers for the BM90M dataset is significantly lower than for BM90C and BM90H, suggesting that these algorithms are sensitive to moderate distortions in the input structures. Also, while post-processing enhances the performance across all target structures, the improvement for protein models is not as good as that obtained for crystal structures. Figure 3.1C demonstrates that \( e\text{FindSite}^{\text{PPI}} \) has the highest tolerance to structural deformations with \( ACC \) and \( MCC \) for the BM90M dataset of 0.889 and 0.242, respectively. Similar to BM90H, PredUs is ranked second with \( ACC \) of 0.782 and \( MCC \) of 0.135. Note that the performance of sequence-based PSIVER is independent on the quality of input structures, thus remains constant across all BM90 datasets. For the BM90C and BM90H datasets, PSIVER is ranked 9th and 8th, respectively. Nonetheless, it is ranked as high as third on the BM90M dataset, suggesting that the performance of most structure-based methods using moderate-quality structures is lower than that of sequence-based approaches. Amongst the algorithms tested here, \( e\text{FindSite}^{\text{PPI}} \) and PredUs are the only exceptions to this limitation.

We believe that the main reason for the high sensitivity to distortions in target structures of many structure-based approaches to PPI residue prediction is their strong dependence on fine atomic details. For instance, PIER employs local statistical properties of protein surface derived at the level of atomic groups, therefore, its high \( ACC \) of 0.906 for BM90C drops to 0.852 (0.714) for BM90H (BM90M). Similarly, \( ACC \) for SPPIDER, which utilizes atomic-level RSA-based fingerprints, drops by over 7% (20%) when high- (moderate-) quality models are used instead of experimental structures. In contrast,
eFindSite\textsuperscript{PPI} and PredUs use global structure alignments by Fr-TM-align and Ska, respectively, which make these predictors fairly insensitive to even moderate structural distortions in computer-generated models. Therefore, except for eFindSite\textsuperscript{PPI} and PredUs, most web servers require high-quality structural data in order to provide accurate PPI residue predictions.

**Rationale for a meta-predictor**

It has been reported that combining predictions by WHISCY and ProMate into an integrated approach called WHISCYMATE yields an improved accuracy of the identification of protein interface residues [46]. Another study demonstrated that meta-PPISP, a meta-predictor built upon PINUP, cons-PPISP and ProMate, outperforms its component methods [92]. In the present study, we perform a similar analysis to determine whether or not combining ten web servers improves the prediction accuracy over individual algorithms using experimental and computer-generated structures. To address this issue, we first applied the Friedman test, a non-parametric alternative to the repeated measures ANOVA [93], to $MCC$ values calculated for web server predictions. $P$-values obtained for the BM90M, BM90M and BM90M datasets are $2.19 \times 10^{-12}$, $1.36 \times 10^{-10}$ and $5.07 \times 10^{-09}$, respectively, indicating that individual algorithms produce statistically different results. Next, we selected the most accurate prediction for each target protein, referred to as a pseudo-meta approach. Note that this protocol is not a true meta-predictor; rather, it helps estimate the upper bound for the prediction accuracy given an optimal combination of individual algorithms. As presented in Figure 3.1 (black asterisks) and Table 3.2, the pseudo-meta approach systematically outperforms individual web servers with $MCC$ for the BM90C, BM90H and BM90M datasets of 0.481, 0.443 and 0.290, respectively. The top three contributors to the
best predictions are PredUs (38% for BM90C, 30% for BM90H and 13% for BM90M), eFindSite\textsuperscript{PPI} (29% for BM90C, 31% for BM90H and 36% for BM90M), and cons-PPISP (18% for BM90C, 8% for BM90H and 13% for BM90M). Lastly, we tested the differences between individual web servers and the pseudo-meta approach using the Wilcoxon signed-rank test, a non-parametric alternative to the paired Student's \( t \)-test \cite{94}. In all cases, the pseudo-meta protocol outperforms web servers with statistically highly significant \( p \)-values of \( \ll 0.01 \).

**FUTURE WORK**

Currently available web servers represent a diverse collection of algorithms for PPI residue prediction. Despite their relatively high accuracy obtained for experimentally solved target structures, using computer-generated models clearly yields less accurate predictions. Based on the results of our analysis, we suggest that post-processing protocols, which seem to quantitatively improve the results only for experimental structures, should be revisited and perhaps tuned up for the homology models of target proteins. Furthermore, meta-predictors should be systematically explored, for example using techniques already extensively studied in protein threading \cite{95,74} and ligand binding site prediction \cite{96,97}. Here, we show that even a simple combination of outputs from various web servers gives a chance to outperform the best single method. More advanced meta-prediction techniques using non-linear machine learning models are likely to further improve the accuracy of PPI residue prediction.

**REFERENCES**


CHAPTER 4: DOCKING AND RERANKING PROTEIN COMPLEXES *

INTRODUCTION

Most proteins work by interacting with other proteins to fulfill their molecular functions, therefore, quaternary assemblies are the key components of the vast majority of biological processes. Consequently, the structural characterization of protein-protein complexes provides valuable insights into protein function and association mechanisms, immensely contributing to the understanding of cellular interaction networks. The knowledge of atomic-level details of protein-protein interactions (PPIs) is required for a number of practical applications, for instance, it is critical for the design of therapeutics targeting protein interfaces [1][2]. X-ray crystallography and NMR spectroscopy are the most widely used experimental techniques to determine protein complex structures. Nonetheless, these methods cannot keep pace with the rapidly growing number of protein interactions identified by high-throughput approaches such as yeast two-hybrid [3] and affinity purification techniques (co-immunoprecipitation [4], tandem affinity purification [5]) followed by mass spectrometry. The low stability of many complexes as well as significant efforts and high costs associated with experiments certainly impede the systems-level exploration of the molecular structures of protein assemblies. On that account, computational tools for PPI structure modeling bridge the gap between the volume of sequence data, the evidence of binary interactions, and the atomic details of pharmacologically relevant protein complexes.

88

*This chapter previously appeared as: Surabhi Maheshwari and Michal Brylinski, ‘Predicted Binding Site Information Improves Model Ranking in Protein Docking Using Experimental and Computer-Generated Target Structures’, BMC Structural Biology, 15.1 (2015), 23. It is reprinted by permission of the publisher, see the permission letter for proper acknowledgment phrase.
Quaternary structure modeling to find the best relative orientation of monomers forming a stable complex can be performed using template-based or template-free techniques. Template-based methods use the similarity to known complex structures to model the interaction between a given pair target proteins. This strategy involves superposing target proteins onto the identified templates using either global or interfacial structure alignments [6]. For instance, PRISM models quaternary structures by matching target proteins to a template interface selected from a representative database of the experimental structures of PPI complexes [7][8]. In contrast, template-free approaches do not use any quaternary information from similar protein complexes; instead, these methods perform docking of the tertiary structures of receptor and ligand proteins. A typical docking calculation comprises two successive steps. First, a rigid-body sampling of six translational and rotational degrees of freedom generates a large set of candidate dimer conformations, in which the constituent monomers are in contact avoiding steric clashes. In the second step, a scoring function is used to rank the disparate collection of docked poses in order to identify near-native models. Current docking algorithms employ a variety of conformational search techniques including a fast Fourier transform [9]–[11], Monte Carlo methods [12], and the geometric hashing [13], [14]; for recent reviews see [15][16][17]. Significant efforts have also been devoted to develop reliable scoring functions, many of which assess the stability of the assembled dimers by combining multiple scoring terms such as the geometric shape [18][19][20][21], chemical and electrostatic complementarity [22][23][24][25][26]. Nevertheless, despite the advances in pose prediction and scoring, docking programs still face significant difficulties in identifying the best solution from a pool of candidates generated through conformational sampling [22], [27]. Therefore, the development of new
approaches to more reliably distinguish between near-native and decoy conformations represents a practical strategy to improve the accuracy of protein docking.

To address the problem of model scoring, the prediction of protein quaternary structures is often supported by a variety of experimental and computational data [28]–[30]. Several strategies to incorporate experimental data in protein docking have been developed. For instance, upper bounds for distances between residues in interacting protein chains can be identified by NMR spectroscopy [31] and chemical crosslinking [32]. Moreover, simultaneous screening for mutations that disrupt yeast two-hybrid interactions was proposed to identify critical interface residues for multiple interacting partners [33]. Experimental data can be subsequently transformed into distance constrains to narrow the search space and to guide the selection of docking poses [34],[35]. Indeed, data-driven docking has been demonstrated to considerably improve the accuracy of dimer structure modeling [36], nonetheless, a limited availability of experimental data remains the major drawback of large-scale investigations of PPI networks. Although computational methods for interface residue prediction [37],[38] can support the complex assembly through PPI prediction-driven docking strategies, [38],[39] the predicted PPI site information is not always accurate leading to spurious results generated by a misguided conformational sampling.

Interaction symmetry is another commonly used form of constraints to model homo-oligomeric complexes. Symmetry is a prevalent feature of the global arrangement between subunits in homo-oligomer complexes formed by two or more identical protein chains. Homo-dimers are important parts of biochemical pathways that are found to occur more
frequently than by chance [40]. Approximately 50-70% of the available datasets comprise homo-oligomers whose structural symmetry is remarkably well conserved [40]–[43]. The symmetric organization of proteins is known to confer structural and functional advantages providing stability, control over accessibility and specificity of active sites [44]. It also provides the ability to avoid unwanted aggregation, which is responsible for a number of pathological conditions, such as Alzheimer's and prion diseases [45], [46]. Furthermore, the symmetric self-association provides an opportunity for cooperative interactions and multivalent binding [47]. Since the cyclic symmetry containing a single rotational axis is the most common type of regularity observed in protein quaternary structures, symmetrical docking a priori restricts the conformational search space only to symmetric transformations [48][10].

In recent years, a two-stage ranking strategy has gained significant attention. Here, a standard protocol is first employed to rapidly scan for putative dimer conformations and to identify a subset of plausible candidates. Subsequently, an additional scoring system is used to re-rank the docked conformations in order to improve the ranking of near-native poses. These methods integrate a variety of features including sophisticated energy calculations, experimental and predicted binding site locations, statistical potentials derived from databases of complex structures, and evolutionary information [28][49]. For instance, ZRANK [50] combines van der Waals, electrostatic and desolvation energy terms to re-rank the initial docking predictions generated by ZDOCK [9], whereas DECK [51] employs a distance and environment dependent knowledge-based potential to refine predictions from GRAMMX [52]. Furthermore, the accuracy of HADDOCK [29] was improved by applying a scoring function based on a Voronoi tessellation of protein structures and machine learning
[53]. Other examples include T-PioDock [54], which uses interface prediction to assist the ranking of docked poses, and ClusPro [55] that re-ranks the top 2,000 solutions generated either by ZDOCK or DOT [56] using a greedy clustering technique. Most of the available re-ranking protocols were designed and subsequently benchmarked using the experimentally determined structures in their bound and unbound conformational state. Since the structure-based reconstruction of across-proteome interaction networks involves docking of various quality homology models, re-ranking strategies should ideally tolerate inaccuracies in the atomic coordinates of interacting monomers.

In that regard, we developed eRank\textsuperscript{PPI}, an algorithm for the selection of correct docking conformations constructed by protein docking using not only experimental monomer structures but also protein models. A scoring function implemented in eRank\textsuperscript{PPI} combines in a novel way certain features such as residue-level interface probabilities estimated by eFindSite\textsuperscript{PPI} [57], protein docking potentials [58], and a new contact-based symmetry score. Although, the predicted interface location was already successfully employed to improve the ranking accuracy for docked conformations [54], most previously reported benchmarking calculations were carried out against relatively small datasets of experimental structures [59]–[61]. In contrast, in this study, we perform a comprehensive analysis using non-redundant and representative sets of crystal structures as well as various quality protein models. In large-scale benchmarks using homo- and hetero-complexes, the accuracy of eFindSite\textsuperscript{PPI} is compared to state-of-the-art scoring methods.
MATERIALS AND METHODS

Datasets and tools

The algorithm for the re-ranking of docking models is trained and tested on the BM1905 dataset of 1,905 proteins, which was compiled previously to evaluate the accuracy of interface residue prediction [57]. This dataset contains experiment target structures (BM1905C) as well as high- and moderate- quality models (BM1905H and BM1905M, respectively). The quality of computer-generated models was assessed by TM-score [62], which ranges from 0 to 1 with values ≥0.4 indicating a significant structure similarity to the native protein. BM1905M and BM1905H datasets comprise models whose TM-score is in the range of 0.4-0.7 and 0.7-0.9 respectively. Furthermore, the BM1905 dataset contains 1,755 homo-dimers (BM1755) and 150 hetero-dimers (BM150).

ZDOCK [9] version 3.0.2 is used to generate rigid-body docking conformations with the default search parameters. It has consistently been among the best performing algorithms in the Critical Assessment of Prediction of Interactions (CAPRI) [27], [63]-[66], a community-wide project assessing the accuracy of protein-protein docking algorithms. ZDOCK employs a fast Fourier transform (FFT) correlation-based method, which performs a systematic search in the six-dimensional space created by 3 rotational and 3 translational degrees of freedom. Docking conformations are predicted based on the desolvation and electrostatics contributions to the complex formation as well as the pairwise shape complementarity. Prior to docking, both the receptor and ligand structures are randomly translated and rotated to avoid any bias towards initial orientations. We collect 2,000 highest scoring conformations reported by ZDOCK for each protein.
In this study, putative interfacial sites are predicted for the benchmarking receptors by eFindSite\textsuperscript{PPI} [57], a recently developed structure/evolution-based approach to detect interface residues. eFindSite\textsuperscript{PPI} exploits a general tendency of the location and geometry of binding sites to be highly conserved in evolutionarily weakly related dimer proteins. It employs a collection of effective algorithms, including meta-threading by eThread [67], structure alignments by Fr-TM-align [62], and machine learning using Support Vector Machines (SVMs) and a Naïve Bayes Classifier (NBC) [68]. Each residue in the query protein is assigned a probability to be at the interface using residue-level attributes in combination with sequence and structure conservation scores derived from evolutionarily related templates.

**Training attributes**

eRank\textsuperscript{PPI} developed in this study employs a series of attributes to re-rank docking conformations, including residue-level interface probabilities, protein docking contact potentials, and energy-based scores. The training and evaluation is performed separately for homo- and hetero-dimers as the modeling of homo-complex structures additionally takes account of symmetry constraints. Individual features are described below.

*Interface scores* - eRank\textsuperscript{PPI} incorporates interface probability estimates for the receptor protein. We use probability scores assigned to each residue in the target protein by eFindSite\textsuperscript{PPI} to estimate the likelihood to be at the protein-protein interface. Interfacial residues in docking models constructed by ZDOCK are identified by iAlign [69], which uses a distance-based criterion to identify the interface in a given multimer structure. The interface score is the sum of probabilities calculated over interface residues; two scores are
computed using SVC and NBC. In general, these scores favor docking conformations with a substantial coverage of surface regions assigned a high interfacial probability by eFindSitePPI.

Protein docking potential - In addition to the interface scores, we employ a protein docking potential previously developed using a linear programming technique [58]. In this model, the side chain center of mass, the backbone carbonyl oxygen, and the amide group are considered interaction sites for each residue. Inter-residue contacts are defined using distance thresholds of 6.8 Å, 4.0 Å and 5.6 Å for side chain, backbone and backbone/side chain sites, respectively. 253 independent pairwise parameters were optimized in order to efficiently discriminate between hits and non-hits across protein-protein ensembles constructed by rigid-body docking.

![Figure 4.1. Calculation of the contact-based symmetry score. The schematics illustrate pairwise residue contacts in (A) a completely symmetric dimer and (B) a partially symmetric dimer. Ax → By denotes that the residue number x in chain A is in contact with the residue number y in chain B.](image)

ZDOCK energy score - Conformational ensembles of putative dimers are constructed by ZDOCK, as described above. The scoring function implemented in ZDOCK is a linear weighted sum of van der Waals attractive and repulsive energies, short- and long-range attractive and
repulsive electrostatic energies, and desolvation. The optimal set of weight factors that maximizes the discriminatory capabilities of ZDOCK was obtained by training the scoring function on the Benchmark 1.0 set [70], followed by a cross-validation against non-homologous cases selected from the Benchmark 2.0 set [71]. We use the total energy score reported by ZDOCK as one of the components of the scoring function in eRankPPI.

Symmetry score - The vast majority of homo-dimers form symmetric interfaces, therefore, we include the deviation from an ideal point group cyclic symmetry in the scoring function to re-rank the homo-complex models. Specifically, we developed a new metric to measure the degree of symmetry at the protein-protein interface, called the contact-based symmetry score (CBS). Figure 4.1 shows two complexes of identical protein chains A (dark gray) and B (light gray) with residues numbered as A1, A2 ... A5 and B1, B2 ... B5, respectively. A complex shown in Figure 4.1A is perfectly symmetrical at the interface, whereas that presented in Figure 4.1B deviates from the ideal symmetry. To quantify this deviation, we first find all inter-residue contacts, defined as those residue pairs, for which any two non-hydrogen atoms are within a distance of 10 Å. For example, in the complex shown in Figure 4.1B, interacting residue pairs are A3: B4, A4: B3, A5: B2, and A5: B1; the notation Ax: By means that the residue number x in chain A is in contact with the residue number y in chain B where x ≠ y. Next, we divide residue pairs into two sets, S1 and S2, so that S1 contains pairs with x < y and S2 contains pairs with x > y. For the complex shown in Figure 4.1B, this gives us S1 = {A3: B4} and S2 = {A4: B3, A5: B2, A5: B1}. Finally, the CBS score is calculated as the Jaccard index to measure the similarity between S1 and S2:

$$CBS = \frac{|S1 \cap S2|}{|S1 \cup S2|} \quad \text{Eq. 4.1}$$
Essentially, the Jaccard index is a ratio of the intersection and the union between the two sets of interacting residue pairs, where $Ax: By$ is considered a match for $Ay: Bx$. CBS ranges from 1 for perfectly symmetrical interfaces to 0 for completely asymmetrical complexes. For example, CBS scores calculated for homo-dimers shown in Figures 4.1A and 4.1B, are 1 (perfect symmetry) and $\frac{1}{3}$ (one-third of a perfect symmetry), respectively. The CSB scores are used only for homo-dimers, therefore five features are computed by $e$Rank$^{pP}$ for homo-complexes, whereas four features are used for hetero-dimers.

**Supervised learning**

The scoring function implemented in $e$Rank$^{pP}$ is trained and cross-validated on docking ensembles generated by ZDOCK separately for the BM1755C and BM150C datasets. Specifically, we calculate the set of either five (homo-dimers) or four (hetero-dimers) attributes for statistical learning in order to rank individual conformations so that near-native structures are assigned lower ranks compared to decoy complexes. The learning procedure is supervised by an iRMSD-based ranking, where the iRMSD is a root-mean-square deviation from the experimental complex structure calculated over the Cα atoms of interface residues. Consequently, the ranking problem can be formulated as the prediction of iRMSD values from individual attributes using a regression analysis. We note that all benchmarking calculations are carried out using a two-fold cross validation protocol by randomly splitting dataset proteins to avoid memorization effects in machine learning. We tested several linear and non-linear models and found that for homo-dimers, Support Vector Regression, epsilon-SVR, with a radial basis function kernel from the LIBSVM version 3.14 [72] yields the best performance. Because of a much smaller dataset size, we use a linear
regression (LR) model [73] for hetero-dimers. Furthermore, individual attributes are standardized independently for each target complex in order to account for proteins of different lengths forming interfaces of different sizes. Specifically, a raw attribute value $x$ is converted to the standard score ($Z$-score) as follows:

$$Z\text{-score} = \frac{x - \bar{x}}{\sigma_x} \quad \text{Eq. 4.2}$$

where $\bar{x}$ is the mean attribute value calculated across the dimer ensembles generated for a given pair of target proteins by ZDOCK, and $\sigma_x$ is the corresponding standard deviation.

**Evaluation of docking predictions**

The quality of model dimer structures is assessed using two metrics, iRMSD and a contact-based score. The iRMSD is a standard evaluation measure in CAPRI corresponding to the interface Cα-RMSD between a ligand in the predicted complex and the ligand in the experimental structure upon the superposition of the receptor structures. In iRMSD calculations, interface residues are defined as those having at least one atom within 10 Å from any atom in the other protein chain. In addition to the iRMSD, the accuracy of complex structures can be evaluated at the level of pairwise residue contacts. Previously, $f_{\text{nat}}$ and $f_{\text{non-nat}}$ have been used to assess the quality of predicted interface interactions [74]. The former is defined as the number of correct (native) residue-residue contacts in the predicted complex divided by the total number of contacts in the experimental structure, whereas the latter is the fraction of non-native contacts in the predicted complex divided by the total number of contacts in that model. Note that $f_{\text{nat}}$ alone may be insufficient to reliably assess the model accuracy because of possible over-predicted interface contacts, which are
revealed by $f_{\text{non-nat}}$. Because, a single metric is more convenient to evaluate the accuracy of protein docking predictions, we formulated a Pairwise Contact Score (PCS). Similar to the iRMSD, pairs of residues on different chains are in contact if any of their atoms are within 10 Å from each other. PCS employs Matthew’s correlation coefficient (MCC) to evaluate the strength of a correlation between the predicted and actual classes:

$$ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+TN)(FP+FN)(TN+FN)}} \quad \text{Eq. 4.3} $$

where $TP$ (True Positives), $FN$ (False Negatives) and $FP$ (False Positives) is the number of correctly predicted, under-, and over-predicted pairwise contacts, respectively. $TN$ (True Negatives) is the number of correctly predicted non-contacting residue pairs. Importantly, PCS considers both the accuracy and error rates, and it is less affected by the imbalanced numbers of positives (pairwise interface contacts) and negatives (non-contacting pairs). Theoretically, MCC ranges from -1 to 1, where 1 corresponds to a perfect prediction and -1 is a perfectly inverse prediction; in practice, PCS scores vary from about 0 to 1.

**Assessment of model ranking**

Protein docking algorithms typically construct multiple dimer models for a given pair of protein structures. Therefore, a reliable scoring function is critical to rank the predicted models so that near-native structures can be selected from a large set of decoys. In that regard, we evaluate the ranking capability using the following measures:

*Percentage of successful cases* - This metric reports the percentage of docking cases for which at least one hit is ranked within the top 10 models. Hits are defined as those conformations having iRMSD below a given cutoff varying from 0 to 15 Å. In addition to the iRMSD, we also
calculate the percentage of successful cases using PCS as the hit criterion where the respective cutoff changes from 1 to 0.

*Hit count* - Hit count gives the average number of hits within the top 10 docking models across the benchmarking dataset. Hits are predictions whose iRMSD is below a given cutoff ranging from 0 to 15 Å. Thus the hit count measures the overall enrichment of the top ranked models with near-native conformations.

*Success rate* - The docking success rate is defined as the percentage of targets for which at least one correct model is ranked within the top \( n \) conformations, where \( n \) changes from 1 to 1,000. The acceptance criteria for correct predictions are an iRMSD of \( \leq 2.5 \) Å, \( \leq 8.5 \) Å and \( \leq 9.5 \) Å for experimental structures, high- and moderate-quality models, respectively.

**RESULTS**

**Symmetry in homo-dimers**

*eRank* employs a new measure, called CBS, which quantifies the deviation from an ideal cyclic symmetry using inter-residue contacts rather than purely geometrical features. First, we calculated the distribution of CBS scores across the experimental homo-dimer structures from the BM1755C dataset. Figure 4.2 demonstrates that the fraction of proteins self-interacting through symmetrical interfaces is notably higher than those having an asymmetric arrangement of their quaternary structures. For instance, 86.6% of the protein complexes have a CBS of \( \geq 0.7 \), compared to only 8.7% with a CBS below 0.5. These results concur with previous studies presenting the symmetry as a rule in the global arrangement of homo-dimers [41][47]. Next, we calculated CBS scores for dimers assembled by ZDOCK.
Here, we separately analyze two subsets of models, 2,000 randomly selected near-native structures whose iRMSD from the corresponding experimental complexes is ≤5 Å, and 2,000 random decoys with an iRMSD of >20 Å. As shown in Figure 4.2, the near-native models tend to deviate from an ideal symmetry to a lesser degree compared to decoys; for example, 50% of near-native structures have a CBS of at least 0.33, whereas only 3.6% of decoys are found at this CBS threshold.

Figure 4.2. Distribution of contact-based symmetry scores across the BM1755 dataset. The results are presented as cumulative fraction of homo-dimers with a contact-based symmetry (CBS) score larger than or equal to the value displayed on the x-axis. CBS quantifies the deviation of a homo-dimer from an ideal cyclic symmetry. Near-native structures and random decoys are those dimer models whose iRMSD from the corresponding experimental complexes is ≤5 Å and >20 Å, respectively.

These findings encouraged us to use the CBS as one of the features to improve the ranking of homo-dimers. As a matter of fact, the concept of symmetry is widely used to
construct homo-dimer complexes. Several protein docking programs were developed to model homo-oligomer structures by performing a systematic space search exclusively for symmetric conformations, e.g. M-ZDOCK [10], SymmRef [75] and SymmDock [48], [76]. These programs commonly use the symmetry to narrow the search space, however, eRank\textsuperscript{PPI} employs a different approach. First, it incorporates the deviation from an ideal symmetry as a feature to improve the ranking of near-native models within docking ensembles generated through an unrestricted conformational search. Second, eRank\textsuperscript{PPI} exploits a contacts-based symmetry rather than geometric regularities, which is more suitable for complex assembly using computer-generated monomers whose tertiary structures are somewhat distorted compared to experimental structures. To our knowledge, the pairwise contact-based symmetry is a novel feature used by eRank\textsuperscript{PPI} in the modeling of homo-dimers.

**Quality of predicted binding interfaces**

The knowledge of PPI sites can be used to improve the success rate in protein docking [28], [36], [77]. Several groups integrated experimentally determined PPI information into their docking algorithms either to restrict the docking space during pose prediction or to filter the constructed conformations as a post-processing step. Moreover, due to the limited availability of experimental data, predicted PPI sites can be used instead. Nonetheless, the predicted PPI information is not always highly accurate and using erroneous data may lead to failed predictions. Ideally, docking strategies utilizing predicted PPI sites should tolerate to some extent only partially accurate constraints. In eRank\textsuperscript{PPI}, we use interface residue prediction by eFindSite\textsuperscript{PPI} that produces a continuous range of probability estimates over surface residues in target proteins rather than just a binary classification of interacting and
non-interacting residues. These probability estimates are used to calculate the cumulative interface score for a given docking model, which is advantageous over the binary classification as it better tolerates a weaker signal from PPI prediction with moderate and low accuracy.

Figure 4.3. Accuracy of PPI site prediction for the BM1905 dataset. The results are presented as the cumulative fraction of proteins with Matthew's correlation coefficient (MCC) between predicted and experimental interface residues larger than or equal to the value displayed on the x-axis. A dotted vertical line marks an MCC of 0.3.

Since the quality of predicted binding interfaces is important for the subsequent modeling of dimer structures, we first inspect the distribution of the PPI prediction accuracy across benchmarking datasets. For each protein target, we calculate Matthew’s correlation coefficient between interface residues in the experimental complex and those predicted by eFindSite\textsuperscript{PPI}. The results for BM1755C (homo-dimers) and BM149C (hetero-dimers) are presented in Figure 4.3. For example, PPI interfaces are predicted with an MCC of ≥0.3 for
58% and 39% of BM1755C and BM149C targets, respectively. We note that PPI residues are identified using evolutionarily weakly homologous templates at the 40% sequence identity threshold. Similar to other template-based PPI residue predictors [78][79], the overall performance of eFindSitePPI for homo-complexes is notably better than that for hetero-complexes, which are underrepresented in the PDB.

![Graph](image)

Figure 4.4. Effect of the PPI prediction accuracy on dimer ranking by eRankPPI. The BM1755C and BM58C datasets are divided into two subsets with respect to the accuracy of interface residue prediction (MCC ≥0.3 and MCC < 0.3). The average hit count ± standard deviation is then calculated separately for each subset. An asterisk indicates that the ranking capability of eRankPPI for hetero-dimers is significantly affected by the accuracy of PPI residue prediction with a p-value of <0.05.

Next, we investigate the effect of the PPI prediction accuracy on the quality of dimer models selected by eRankPPI from docking ensembles constructed by ZDOCK. Specifically, we divide each dataset based on the MCC of PPI site prediction using a cutoff of 0.3 and compare the ranking capability of eRankPPI. Figure 4.4 shows the average hit count and the standard deviation calculated at an iRMSD of 2.5 Å for homo-dimers (BM1755C) and hetero-dimers (BM150C). The average hit count for the BM1755C dataset is 1.35 and 0.94 considering those
target proteins whose PPI residues are predicted with an MCC of \( \geq 0.3 \) and \(< 0.3 \), respectively. For the BM150C dataset, the average hit count is 1.79 at an MCC of \( \geq 0.3 \) and 0.67 at an MCC of \(< 0.3 \). To assess the statistical significance of these differences, we calculated the corresponding \( p \)-values using the Wilcoxon signed-rank test, a non-parametric alternative to the paired Student’s \( t \)-test [13]. At the 5% significance level, the accuracy of PPI residue prediction for hetero-dimers affects the ranking capability of eRank\(^{PPI} \) with a \( p \)-value of 0.027. In contrast, a \( p \)-value of 0.121 indicates that the selection of near-native models for homo-dimers is less affected by the quality of the PPI interfaces predicted by eFindSite\(^{PPI} \).

The main reason for the higher tolerance of inaccurately annotated interface residues for homo-dimers is the additional score, CBS, which helps eliminate the majority of asymmetric decoys.

**Ranking using experimental structures**

In order to evaluate the performance of eRank\(^{PPI} \), we first re-ranked the top 2,000 models assembled by ZDOCK from monomers in their bound conformational state. We use iRMSD and PCS to assess the native-likeness of modeled dimer structures and analyze the results in terms of the percentage of successful cases, the hit count and the success rate. First, we evaluate the ranking capability of eRank\(^{PPI} \) compared to ZDOCK and ZRANK against homo-dimers from the BM1755C dataset. Table 4.1 shows that using eRank\(^{PPI} \), at least one model with an iRMSD below 2.5 Å is found within the top 10 ranked conformations for 58.1% of the benchmarking cases. This performance represents an improvement over ZDOCK and ZRANK, which give the percentage of successful cases of 51.1% and 55.2% respectively. We also assessed the contribution of the symmetry score to the overall success; removing the
symmetry score from the scoring function yields the percentage of successful cases of 56.1%. Moreover, using PCS with a cutoff of 0.65 as the success criterion, eRankPPI improves model ranking by 17.2% (8.6%) with respect to ZDOCK (ZRANK).

Table 4.1. Comparison of the success rates for different scoring functions against experimental target structures.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Scoring function</th>
<th>Success rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM1755C</td>
<td>iRMSD = 2.5 Å</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCS = 0.65</td>
<td></td>
</tr>
<tr>
<td>eRankPPI</td>
<td>58.08</td>
<td>58.86</td>
</tr>
<tr>
<td>ZDOCK</td>
<td>51.13</td>
<td>51.68</td>
</tr>
<tr>
<td>ZRANK</td>
<td>55.18</td>
<td>55.49</td>
</tr>
<tr>
<td>BM58C</td>
<td>iRMSD = 2.5 Å</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCS = 0.65</td>
<td></td>
</tr>
<tr>
<td>eRankPPI</td>
<td>84.42</td>
<td>84.48</td>
</tr>
<tr>
<td>ZDOCK</td>
<td>67.75</td>
<td>67.24</td>
</tr>
<tr>
<td>ZRANK</td>
<td>75.86</td>
<td>75.86</td>
</tr>
</tbody>
</table>

Further comparison of the overall performance of eRankPPI, ZDOCK and ZRANK is shown in Figure 4.5. Figures 4.5A and 4.5B demonstrate that the percentage of successful cases within the top 10 conformations for eRankPPI is higher than that for ZDOCK and ZRANK over a range of iRMSD and PCS threshold values used to define correct predictions. The same holds true for the hit count and the success rate; for instance, Figure 4.5C shows that using eRankPPI yields an average number of 2.21 hits per target within the top 10 ranked predictions at an iRMSD cutoff of 5 Å, whereas the hit count for ZDOCK and ZRANK is 1.60 and 1.68, respectively. Model ranking by eRankPPI is consistently better than that by ZDOCK.
and ZRANK not only for the top 10 but also considering lower ranks, which can be evaluated using the success rate shown in Figure 4.5D. These results suggest that compared to other algorithms, the scoring function implemented in eRankPPI more reliably identifies near-native models of homo-dimer complexes across docking ensembles.

Figure 4.5. Performance of eRankPPI, ZDOCK and ZRANK on the BM1755 dataset. Ranking accuracy is assessed by the percentage of successful cases based on (A, E, I) iRMSD and (B, F, J) PCS, (C, G, K) the hit count, and (D, H, L) the success rate. Each algorithm is evaluated against (A-D) experimental structures, as well as (E-H) high-quality and (I-L) moderate-quality protein models. Black dashed lines shown for the percentage of successful cases correspond to the upper bound estimated by taking the best of all 2,000 models constructed for each target.

Next, we turn over to hetero-dimers and compare the performance of eRankPPI, ZDOCK and ZRANK for the BM155 dataset. The success rate of ZDOCK, ZRANK and eRankPPI against BM155C targets is 53.7, 67.1 and 58.4, respectively. The analysis of the quality of predicted binding interfaces on the docking accuracy presented above indicates that
eRank\textsuperscript{PPI} is sensitive to inaccuracies in PPI annotation for hetero-complexes. Therefore, we use a subset of 58 targets selected from BM155 whose interface residues are predicted with an MCC of \( \geq 0.3 \); we refer to this dataset as BM58. Figure 4.6A shows that the ranking capability of eRank\textsuperscript{PPI} for the BM58C dataset is better than that of ZDOCK and ZRANK. For example, Table 4.1 shows that at an iRMSD threshold of 2.5 Å, the percentage of successful cases for eRank\textsuperscript{PPI}, ZDOCK and ZRANK is 84.4%, 67.8% and 75.9% respectively. Similar improvements are observed for the PCS used as the success criterion in Figure 4.6D; using eRank\textsuperscript{PPI} improves the ranking by ZDOCK (ZRANK) by 13.8% (5.2%). We note that in contrast to homo-dimers, eRank\textsuperscript{PPI} does not improve model ranking for those targets whose binding interfaces are poorly annotated, therefore, a sufficiently high accuracy of PPI residue prediction is critical for the construction of hetero-dimer structures.

**Ranking using computer-generated models**

Genome-wide determination of protein interaction networks is an important step in the elucidation of cellular regulatory mechanisms [80], [81]. Although constituent interactions can be modeled through a structure-based dimer assembly, the performance of scoring functions for model selection certainly depends on the quality on input structures. So far, we discussed the ranking of dimer models constructed from experimental monomer structures. Nonetheless, despite the exponential growth of the PDB, experimentally determined structures of a vast majority of gene products are not yet available. This necessitates using computer-generated models in protein docking, however, assuming that a docking program is capable to reliably construct complexes using theoretical monomer structures. Previously, a low-resolution docking method was applied to protein models [82]
as a starting point for the subsequent high-resolution refinement to address the challenges of PPI modeling at a proteome-wide scale.

Here, we investigate how different docking scoring strategies cope with inaccuracies in the computer-generated models of query proteins. Undoubtedly, docking using protein models represents a difficult task and the quality of the resulting dimers cannot be higher than the quality of monomer structures. An iRMSD cutoff of 2.5 Å is widely accepted as a criterion for near-native models using experimental structures. However, different threshold values need to be used to evaluate dimer structures assembled from computer-generated models in order to account for distortions in individual monomers. Therefore, we first calculated the distribution of hits with an iRMSD of 2.5 Å across the top 2,000 docking models constructed by ZDOCK using experimental monomer structures. A black dashed line in Figure 4.5A shows that at least one assembled dimer has an iRMSD of 2.5 Å for about 70% of the target proteins. We found that an iRMSD cutoff of 8.5 Å (9.5 Å) gives a similar coverage when high- (moderate-) quality models are used in protein docking. Furthermore, we established PCS cutoffs in a similar fashion so that ~70% of the cases have at least one hit within docking ensembles; the corresponding threshold values are 0.65, 0.30 and 0.25 for crystal structures, high- and moderate-quality models, respectively.

Using these iRMSD and PCS cutoffs to define accurate predictions, we evaluate the ranking capability of eRankPPI, ZDOCK and ZRANK on the BM1755H and BM1755M datasets of homo-dimers. Table 4.2 shows that eRankPPI places at least one model with an iRMSD of ≤8.5 Å (≤9.5 Å) within the top 10 conformations for 42.7% (42.3%) of the high- (moderate) quality models. This performance represents a significant improvement over both ZDOCK
and ZRANK, which give the percentage of successful cases of 27.6% (26.9%) and 22.5% (24.6%), respectively. Furthermore, the overall performance of eRank\textsuperscript{PPI}, ZDOCK and ZRANK for homo-dimer targets is compared in Figure 4.5. Figures 4.5E, 4.5F, 4.5I and 4.5J demonstrate that the percentage of successful cases within the top 10 conformations for eRank\textsuperscript{PPI} is closer to the estimated upper limit than for ZDOCK and ZRANK over a range of iRMSD and PCS threshold values defining correct predictions. We note that the black dashed lines in Figures 4.5 and 4.6 represent upper bounds for the docking accuracy calculated by selecting the best dimer from the entire ensemble of 2,000 structures constructed by ZDOCK for a given target protein.

Figure 4.6. Performance of eRank\textsuperscript{PPI}, ZDOCK and ZRANK on the BM58 dataset. Ranking accuracy is assessed by the percentage of successful cases based on (A-C) iRMSD and (D-F) PCS. Each algorithm is evaluated against (A, D) experimental structures, as well as (B, E) high-quality and (C, F) moderate-quality protein models. Black dashed lines correspond to the upper bound estimated by taking the best of all 2,000 models constructed for each target.
Similar performance improvements are observed for the hit count and the success rate. For instance, Figure 4.5G and 5K show that using eRank$^{\text{PPI}}$ yields an average number of 1.36 and 1.35 hits per target for the BM1755H and BM1755M datasets at the iRMSD cutoffs of 8.5 Å and 9.5 Å, respectively. For comparison, the corresponding hit counts for ZDOCK (ZRANK) are only 0.66 (0.69) and 0.46 (0.47). Furthermore, in Figure 4.6, we examine the performance of eRank$^{\text{PPI}}$, ZDOCK and ZRANK on the BM58H and BM58M datasets of heterodimers. For instance, Figures 4.6B and 4.6C show that the percentage of successful cases at an iRMSD of 8.5 Å (9.5Å) obtained by eRank$^{\text{PPI}}$, ZDOCK and ZRANK for BM58H (BM58M) is 50.0% (34.4%), 29.31% (27.5%) and 34.5% (17.2%) respectively. This comprehensive analysis using various evaluation measures demonstrates that dimer ranking by eRank$^{\text{PPI}}$ is consistently better than that by ZDOCK and ZRANK not only using experimental monomer structures, but also computer-generated models.

Table 4.2. Comparison of the success rates for different scoring functions against high- and moderate-quality protein models.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Scoring function</th>
<th>Success rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM1755H</td>
<td></td>
<td>iRMSD = 8.5 Å</td>
</tr>
<tr>
<td></td>
<td>eRank$^{\text{PPI}}$</td>
<td>42.71</td>
</tr>
<tr>
<td></td>
<td>ZDOCK</td>
<td>27.61</td>
</tr>
<tr>
<td></td>
<td>ZRANK</td>
<td>22.55</td>
</tr>
<tr>
<td>BM1755M</td>
<td></td>
<td>iRMSD = 9.5 Å</td>
</tr>
<tr>
<td></td>
<td>eRank$^{\text{PPI}}$</td>
<td>42.31</td>
</tr>
<tr>
<td></td>
<td>ZDOCK</td>
<td>26.99</td>
</tr>
<tr>
<td></td>
<td>ZRANK</td>
<td>24.60</td>
</tr>
</tbody>
</table>
DISCUSSION

The identification of near-native conformations across docking ensembles remains a challenging problem in the structure-based modeling of protein-protein interactions. Docking strategies need accurate scoring functions to rank the predicted conformations. Many current approaches employ the geometric, chemical and electrostatic complementarity as well as knowledge-based interaction potentials as components of their scoring functions. In this communication, we describe eRank\textsuperscript{PPI}, a new scoring method for protein-protein docking that integrates predicted binding site information, protein docking potentials, energy-based scoring and a contact-based symmetry constraints (for homodimers). Although these attributes have been used previously in protein docking, we combined them in eRank\textsuperscript{PPI} as a single, machine learning-based scoring function. The results demonstrate that eRank\textsuperscript{PPI} reliably selects near-native conformations from a large number of decoys generated by ZDOCK [9]. Moreover, comparative benchmarks show that eRank\textsuperscript{PPI} consistently outperforms the state-of-the-art algorithms, ZDOCK and ZRANK, for both homo- and hetero-complexes yielding notably higher hit counts and success rates.

In addition to experimental target structures, we performed a series of benchmarking simulations using computer-generated models. Interestingly, ZRANK performs better than ZDOCK only against experimental target structures. The main reason for this high sensitivity to distortions in target structures is likely a strong dependence on atomic potentials, therefore, ZRANK requires high-quality structural data in order to provide accurate ranking. In contrast, eRank\textsuperscript{PPI} outperforms both ZDOCK and ZRANK not only using experimental structures, but also computer-generated models. This is an important feature of eRank\textsuperscript{PPI}
owing to the fact that protein models represent the most challenging targets for molecular docking.

The analysis of the linear regression model used by eRankPPI to rank hetero-dimers shows that the optimized weights for the SVC and NBC interface scores assigned by eFindSitePPI, the protein-docking potential and the ZDOCK score are 171.9, 891.8, 122.7 and 2.2, respectively. Therefore, the predicted binding site information is a major contributor to the improvement of model ranking in protein docking. Since the success of eRankPPI depends on the accuracy of protein interface prediction, using a robust PPI prediction program is essential. Here, we use eFindSitePPI, a recently developed template-based approach that effectively exploits the tendency of the location of binding sites to be highly conserved across evolutionarily related protein dimers [57]. eFindSitePPI uses the three-dimensional structure of a query protein, evolutionarily remotely related templates and machine learning to predict interfacial sites. It was also shown to outperform several PPI site prediction programs [83]. Also, different from other prediction techniques, eFindSitePPI tolerates structural imperfections in computer-generated models. These characteristics make eFindSitePPI a preferred PPI predictor to support dimer ranking in across-proteome docking studies using eRankPPI.

We conclude this study discussing several examples that illustrate the key features of eRankPPI. Figure 4.7 shows how predicted PPI site information helps improve the ranking of near-native models. The experimental structure of aromatic amino acid aminotransferase homo-dimer (ARAT, PDB-ID: 1ay4, chains A and B) [84] is presented in Figure 4.7A. Figures 4.7B and 4.7C show selected docked conformations with residues in the receptor protein are
colored according to the predicted probability to be at the interface (green and blue correspond to the high and low interfacial probability, respectively). Only a partial overlap between the predicted and docked interface is apparent in Figure 4.7B as a large chunk of the predicted interface area is exposed to the solvent. This conformation has an iRMSD of 23.64 Å and was ranked 1st by ZDOCK, whereas eRankPPI placed it at the rank 413. In contrast, the docked interface shown in Figure 4.7C has a substantial overlap with that predicted by eFindSitePPI; the iRMSD of this model is 6.11 Å and it is ranked 1st and 14th by eRankPPI and ZDOCK, respectively.

Figure 4.7. Model ranking for ARAT homo-dimer. The experimental complex structure is shown in (A) with the chain A colored in blue and the chain B colored in yellow. The top ranked models by ZDOCK and eRankPPI are shown in (B) and (C), respectively. In (B, C), the surface of the chain A is colored according to interface probability estimated by eFindSitePPI with the scale given in the bottom right corner (blue/white/green for the high/intermediate/low probability).
Figure 4.8. Model ranking for repressor protein cI homo-dimer. The experimental complex structure is shown in (A) with chain A colored in blue and chain B colored in red. The top ranked models by ZDOCK (chain B is yellow) and eRank^PPI (chain B is green) are shown in (B) and (C), respectively. A cartoon representation is used for both chains with interface residues presented as a solid surface.

Next, we present a case study that illustrates how contact-based symmetry improves the ranking of near-native models for homo-dimers. Figure 4.8A shows the crystal structure of λ repressor C-terminal domain (repressor protein cI, PDB-ID: 1f39, chains A and B) [85], whereas Figures 4.8B and 4.8C present the top ranked conformations by eRank^PPI and ZDOCK, respectively. The symmetry score implemented in eRank^PPI ranges from 0 (no symmetry) to 1 (perfect symmetry); the native complex has a perfect symmetry as indicated by a CBS of 1.00. The top ranked model by ZDOCK has an iRMSD of 14.89 Å and a symmetry score of 0.00. The lack of symmetry is evident in Figure 4.8B; eRank^PPI placed this model at rank 806 because of the low CBS score. On the other hand, the top ranked model by eRank^PPI shown in Figure 4.8C has a high symmetry score of 0.85 and it is indeed the best model constructed for this target with an iRMSD of 1.27 Å. ZDOCK placed this model at rank 286, therefore, the symmetry score was critical to improve the ranking of this near-native conformation. We note that the contact-based symmetry score is not only intuitive as it ranges from 0 to 1, but also it can be calculated for any protein complex, including those constructed using computer-generated monomer structures.
Finally, we discuss an example of the hetero-dimer complex between the human cyclin-dependent kinase 2 and cell cycle-regulatory protein CksHs1; the crystal complex structure is shown in Figure 4.9A (CDK2, PDB-ID: 1buh, chains A and B) [86]. Figure 4.9B shows the structure of the top ranked conformation by ZDOCK, which has an iRMSD of 18.53 Å and was ranked 6th by eRankPPi. Figure 4.9C presents the structure of the nearest-native complex found within the set of 2,000 conformations generated by ZDOCK that has an iRMSD of 0.98 Å. This model is ranked 28th by ZDOCK, whereas eRankPPi placed it at rank 2. MCC of PPI site prediction for this target is only 0.39, nonetheless, despite the moderate accuracy of interface residue prediction, eRankPPi ranked this nearest-native conformation much higher than ZDOCK.

**CONCLUSION**

In this study, we developed eRankPPi, an algorithm for the selection of correct docking conformations constructed by rigid-body protein docking. eRankPPi features a new scoring function that integrates the predicted interface location with protein docking potentials and
a contact-based symmetry score. Comprehensive benchmarking calculations show that eRankPPI has a high tolerance to structural imperfections in computer-generated protein models, therefore, it opens up a possibility to conduct the exhaustive structure-based reconstruction of PPI networks across proteomes.

Availability of Supporting Data: The methods and datasets used in this study are available at www.brylinski.org/erankppi.

REFERENCES


INTRODUCTION

Protein-protein interactions (PPIs) are ubiquitous and play crucial roles in all biological processes within and between cells by mediating signaling pathways in cellular networks and controlling intracellular communication [1]. Since complex biological systems are governed by sophisticated networks of PPIs, associations between proteins ultimately determine the behavior of the cell. Genome-sequencing projects provide comprehensive datasets of biological sequences and numerous post-genomic projects are largely focused on the exploration and analysis of PPIs across proteomes [2], [3]. The number of possible PPIs in an organism can be scaled as the square of the total number of monomeric proteins, yielding an estimated number of disparate protein complexes in the order of millions. High-throughput approaches allow the large-scale detection of protein-interaction partners in many organisms. Although the PPI data is being produced at a swift pace, the major issues in using the current genome-wide PPI data are a low coverage and high false positive rates [4], [5]. Moreover, inter-study discrepancies between different experimental approaches applied to the same biological system are not uncommon [6]. Last but not least, while these high-throughput methods identify proteins interacting with one another, they do not provide structural information on biologically relevant protein complexes.

On the other hand, interaction details, which can only be obtained from three-dimensional structures, are crucial to fully comprehend interaction mechanisms at the atomic level. Unfortunately, despite ongoing efforts in structural genomics projects to determine complex structures, structural biology is lagging behind in the current trends of
high-throughput methods. While the repertoire of monomeric protein structures solved by X-ray crystallography and NMR spectroscopy is increasing exponentially, the structural space of interacting proteins is still far from complete. In fact, there is an increasing gap between the number of identified interactions and the number of 3D structures of these associations. Thus, it is imperative to develop and continuously improve computational techniques to accurately identify interacting proteins and the corresponding complex structures.

A number of computational approaches have been developed to discover and model new interactions at a system level. Modeling complex structures can be accomplished using two distinct types of techniques, template-free and template-based. The former methods, also known as protein docking, construct a complex model by assembling the monomeric structures of target proteins through a conformational search followed by the selection of high scoring binding orientations. In contrast, template-based approaches build complex structures by mapping monomeric targets to experimentally solved template complexes often followed by the refinement of the initial structural framework. Both methods have advantages and disadvantages. Template-based approaches can construct dimeric models directly from target sequences, therefore, monomer structures may not be required. Further, these techniques select templates based on sequence [7], [8], sequence-to-structure [9] and structure alignments [10][11] often yielding more accurate results than template-free docking [12], [13]. Although dimer templates are available in the Protein Data Bank (PDB) [14] to model all complexes in which the monomer structures are either known or can independently be modeled [15], the success rate of template-based docking is only about 23% when no closely homologous templates with a sequence identity to the target of >40%
can be found for at least one monomer chain. Analogous interaction templates cannot be identified in the current PDB to effectively guide template-based docking in those failed cases [16]. The fact that suitable templates are available only for a limited number of interactions significantly lowers the coverage of proteome-scale datasets.

In contrast, template-free methods are, in principle, applicable to those protein targets whose monomer structures are either solved experimentally or can be generated with homology modeling. These techniques do not require the structures of related complexes to model the association between targets proteins. Consequently, template-free approaches provide a higher coverage in large-scale applications focusing on the construction and analysis of PPI networks. Although template-free modeling is often applied to a pair of proteins known to interact with one another, several studies have successfully employed the exhaustive rigid-body protein docking and post-docking analysis to predict PPIs and PPI networks [17]–[19]. For instance, a docking experiment comparing the distribution of docking scores collected for proteins known to interact to those between putatively non-interacting proteins was reported [20].

Another study attempted to predict the protein-protein interaction network of the bacterial chemotaxis signaling pathway using an all-to-all docking approach [21]. Here, two docking tools, MEGADOCK [17] and ZDOCK [22], were employed to conduct rigid-body docking of all possible combinations of 101 proteins belonging to 13 families, which are known to be part of the chemotaxis signaling pathway. Based on a previous observation that the decoys of interacting proteins form dense clusters as opposed to the lack of dense clusters formed by non-interacting proteins [17], [18], clustering high-scoring decoys was
used to evaluate protein binding affinity and to predict the PPI network. Encouragingly, combining positive predictions from both docking tools correctly identified almost all core-signaling interactions in bacterial chemotaxis. Although the aforementioned methods were shown to discriminate true protein interactions from likely non-interacting pairs, the native complexes of interacting proteins have not been recovered mainly due to an insufficient ranking accuracy of docking algorithms. Further, the reported benchmarking calculations conducted using relatively small datasets of experimental structures may not be indicative of the performance of the proteome-scale identification of molecular interactions.

In that regard, we developed a new approach to discover and model PPIs across proteomes employing an exhaustive all-to-all docking strategy. This pipeline comprises six major steps including protein threading and homology modelling, the prediction of binding interfaces, a rigid body docking, the flexible refinement and scoring of the modeled interfaces, and a series of function annotation filters. Our approach was carefully benchmarked on a large and representative dataset of experimental structures and computer-generated models of target proteins. In order to demonstrate its utility in large-scale projects, we modeled dimer structures and predicted PPIs across the proteome of *E. coli*. Interaction data generated for *E. coli* is primed for experimental validation and further computational analyses. Encouragingly, our results demonstrate that protein docking can be used not only to identify near-native complexes but also to predict interaction partners. Overall, this study shows that combining computational modeling, structural bioinformatics, machine learning, and function annotation provides a powerful methodology for the bottom-up assembly of protein-protein interaction networks.
MATERIALS AND METHODS

Datasets

The pipeline to model PPIs is benchmarked on the BM1905 dataset (available at http://brylinski.cct.lsu.edu/content/efindsiteppi-datasets), which was previously compiled to evaluate the accuracy of interface residue prediction and the re-ranking of docked models [23], [24]. This dataset contains experimental target structures (BM1905C) as well as high-quality computer-generated models (BM1905H). The quality of monomer models was assessed by the root-mean-square deviation (RMSD) and the Template Modeling score (TM-score) [25]. The latter ranges from 0 to 1 with values >0.4 indicating a significant structural similarity to the native conformation. BM1905H comprises models whose mean Co-RMSD is 6.94 Å ±4.61 and mean TM-score is 0.72 ±0.15.

The algorithm to predict binary interactions is trained and validated against a non-redundant and representative dataset of 18,162 protein dimers selected from the PDB. First, all dimers having at least 20 interface residues were categorized as either homo-dimers whose individual chains share at least 85% sequence identity or hetero-dimers when the sequence identity was below 85%. Next, each subset was clustered with CD-HIT [26] at 80% sequence identity. Finally, redundant dimers that have similar interfaces with the Matthews correlation coefficient (MCC) calculated over interface residues of >0.5 were removed from each cluster. This procedure resulted in a set of 14,944 homodimers (HOM14944) and a set of 3,519 heterodimers (HET3519). In addition, the algorithm to predict binary interactions is tested on 1,688 non-interacting protein pairs derived from the Negatome 2.0 database [27]. Computer models of individual proteins in Negatome 2.0 were built with Modeller [28].
using templates identified by eThread [29], followed by a high-resolution structure refinement with ModRefiner [30].

The developed pipeline to predict PPI networks is validated using *Escherichia coli* as a model organism. Protein interaction data for *E. coli* consisting of 13,374 known interactions formed by 2,994 bacterial proteins were downloaded from the Database of Interacting Proteins (DIP) [31] in March 2016. We removed from the original dataset redundant proteins as well as those targets longer than 600 residues, which may be difficult to model with threading, and shorter than 50 residues because these molecules are likely peptides. The final *E. coli* dataset consists of 2,300 proteins forming 6,341 interactions. DIP provides the sequences of interacting proteins, therefore, we constructed monomer structures with Modeller [28] using templates identified by eThread [29], followed by a high-resolution structure refinement with ModRefiner [30].

**Protein docking, ranking and refinement**

For a given pair of protein targets, a collection of docking solutions is generated with the FFT-based rigid body docking program ZDOCK version 3.02 [32]. We use the default parameters to exhaustively search the 3D grid space around the receptor by rotating and translating the ligand. Subsequently, the top 2,000 conformations reported by ZDOCK are re-ranked with eRankPPI [24], a recently developed algorithm to identify near-native conformations from the high-scoring hits. The scoring function implemented in eRankPPI employs multiple features including residue-level interface probability estimates, protein docking potentials, and energy-based scores. Surface residues in target receptors are annotated with interface probability estimates by eFindSitePPI [23], a structure/evolution-
based approach to detect interface residues. eFindSite\textsuperscript{PPI} builds on a strong conservation of the location and geometry of binding sites in evolutionarily related dimers and employs meta-threading, structural alignments, and machine learning to predict interfacial residues for a target protein. The top 10 models selected by eRank\textsuperscript{PPI} are finally subjected to a flexible refinement with FiberDock [33]. FiberDock mimics the induced fit by accounting for both side-chain and backbone flexibility. The side-chain flexibility is modeled using a rotamer library, whereas a normal mode procedure is used to model the backbone flexibility.

**Assessing the quality of protein complex models**

The accuracy of dimer models is primarily assessed with iAlign [34] against experimental complex structures retrieved from the PDB. iAlign evaluates the quality of structural models with the Interface Similarity score (IS-score) combining Cartesian distances with the overlap of interfacial contact patterns. IS-score ranges from 0 to 1 with values greater than 0.210, 0.311 and 0.473 indicating a statistically significant interface similarity at \( p \)-values of \( 10^{-2} \), \( 10^{-5} \) and \( 10^{-10} \), respectively. In addition, the quality of dimer models is assessed with iRMSD, a standard evaluation measure in CAPRI. iRMSD is the interfacial C\( \alpha \)-RMSD between ligands in the predicted and experimental complexes upon the superposition of receptor structures. In iRMSD calculations, interface residues are defined as those having at least one atom within 10 Å from any atom in the other protein chain. The docking success rate is defined as the percentage of targets for which at least one correct model is ranked within the top 10 conformations. The acceptance criteria for correct predictions are an iRMSD of \( \leq 2.5 \) Å for experimental structures and \( \leq 8.5 \) Å for computer-generated models [24].
Protein-protein interaction prediction with supervised learning

The scoring function to identify biologically relevant assemblies was trained and cross-validated against the HET3519 dataset of experimental hetero-dimers used as positives and a simulated dataset of 14,944 likely non-interacting pairs used as negatives. The negative dataset was constructed by randomly swapping ligands within the HOM14944 dataset. Since HOM14944 proteins share less than 80% sequence identity, this procedure resulted in a random set of hetero-dimers referred to as RND14944. Uniformly choosing random protein pairs excluding experimental interactions produces an unbiased estimate of the distribution of negatives in the prediction of protein-protein interactions [35]. Hence, this procedure is a common practice to generate negative datasets containing at most a negligible fraction of interacting proteins [36][37][38]. FiberDock calculates several binding energy scores, including attractive and repulsive van de Waals forces, the atomic contact energy, partial electrostatics, hydrogen and disulfide bonds, π-stacking, and aliphatic interactions. These scores were used as a feature vector to train a Random Forest Classifier (RFC) returning a single probabilistic score to assess whether two interacting proteins are biologically relevant. The machine learning model was 10-fold cross-validated against the positive set HET3519 and the negative set RND14944.

Annotation filters

Positive predictions are further subjected to filtering with Gene Ontology (GO) terms. GO is a hierarchically organized database providing a controlled vocabulary to characterize gene products, divided into three sub-ontologies: cellular component (CC), biological process (BP) and molecular function (MF) [39]. Here, we use GO slims, which are cut-down
versions of the GO ontologies without the detail of the specific fine grained terms. GO slims were extracted from the PANTHER classification system [40], whereas annotations for *E. coli* proteins were obtained from the EcoCyc database [41] in May 2016. We tested whether CC, BP and MF slims can be used to refine prediction results by considering proteins localized in the same cellular component, assigned to the same biological process, and having different molecular functions.

**Performance evaluation metrics**

PPI prediction is assessed using standard evaluation metrics for classification problems:

- **True positive rate:** 
  \[ TPR = \frac{TP}{TP + FN} \quad \text{Eq. 1} \]

- **False positive rate:** 
  \[ FPR = \frac{FP}{FP + TN} \quad \text{Eq. 2} \]

- **Accuracy:** 
  \[ ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{Eq. 3} \]

- **Matthews correlation coefficient:** 
  \[ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + TN)(FP + FN)(TN + FN)}} \quad \text{Eq. 4} \]

where *TP* (True Positives), *FN* (False Negatives) and *FP* (False Positives) are the number of correctly predicted, under-, and over-predicted PPIs, respectively. *TN* (True Negatives) is the number of correctly predicted non-interacting partners. The MCC quantifies the strength of
the correlation between predicted and actual classes; by heavily penalizing both over- and under-predictions, it provides a convenient assessment measure that balances the sensitivity and specificity.

Figure 5.1. Flowchart of the across-proteome modeling of dimer structures and the prediction of protein-protein interactions. (A) Query protein structures are first built with homology modeling. (B) Subsequently, a binding site is identified in the receptor and initial dimer models are generated through rigid body docking. (C) Initial models are then re-ranked by eRank$^{PPI}$ taking into account the binding site information and (D) subjected to a flexible refinement. (E) Machine learning followed by (F) annotation filters are finally employed to identify biologically relevant protein assemblies (G).
RESULTS AND DISCUSSION

The goal of this study was to develop and test a new protocol to model putative protein complex structures across proteomes that can subsequently be used to assemble protein-protein interaction networks. The modeling procedure for a pair of proteins is presented in Figure 5.1. The construction of a hetero-dimer starts with the prediction of 3D structures of individual monomer chains using eThread and Modeller (Figure 5.1A). Here, the larger monomer is the receptor and the smaller monomer is the ligand; the size is proportional to the number of amino acid residues. Subsequently, eFindSite\textsuperscript{PPI} is employed to predict a protein binding site in the receptor structure and, simultaneously, a rigid-body docking of the ligand to the receptor is performed with ZDOCK (Figure 5.1B). In the next step, docking conformations are filtered and re-ranked with eRank\textsuperscript{PPI} utilizing the binding interface predicted by eFindSite\textsuperscript{PPI} (Figure 5.1C). The identified putative dimers are then subjected to a flexible refinement with FiberDock (Figure 5.1D) followed by the evaluation of binding energies with the RFC in order to select the final model (Figure 5.1E). A probability score reported by the RFC is used together with annotation filters according to Gene Ontology terms (Figure 5.1F) to make the final decision whether or not the constructed dimer is biologically relevant (Figure 5.1G).

Although the comprehensive benchmarks of eFindSite\textsuperscript{PPI} and eRank\textsuperscript{PPI} have been already reported [23], [24], we found that a flexible refinement improves the accuracy of dimers assembled from experimental as well as computer-generated monomer structures. In addition, using machine learning to evaluate the refined interfaces is shown to reliably detect biologically relevant protein complexes. Finally, we demonstrate that annotation
filters can successfully be employed in genome-wide projects to further refine the classification results and more accurately identify putative pairs of interacting proteins.

Figure 5.2. Analysis of success and failure rates in docking of crystal structures and protein models. Successful docking cases shown in green correspond to those predictions for which at least one native-like configuration with an IS-score greater than a value display on the x-axis is ranked within the top 10 poses. The remaining cases represent two types of docking failures. Scoring failures shown in red correspond to those predictions in which at least one native-like configuration is present in a set of 2,000 dimer models, however, it was not ranked within the top 10 poses. Sampling failures shown in yellow correspond to the remaining cases for which no native-like configurations have been generated.

**Sampling and scoring in template-free docking**

In this work, the structures of protein complexes are modeled via a protocol utilizing template-free docking with ZDOCK. Template-free docking consists of two successive tasks, sampling and scoring. Sampling employs a rigid-body search over different rotational-translational degrees of freedom, whereas the purpose of scoring is to rank the sampled poses in order to identify near-native configurations. Consequently, sampling and scoring failures are two major reasons for the lack of success in protein docking. The former are caused by an insufficient sampling, viz. near-native conformations are not generated by a sampling algorithm, therefore, reliable dimer models cannot be constructed. These errors can frequently be corrected simply by increasing the sampling exhaustiveness. Scoring
failures are unsuccessful docking calculations, in which at least one near-native conformation is generated, however, it is not selected by a scoring function as a feasible solution; correcting these errors is more challenging compared to sampling failures. eRank\textsuperscript{PPI} was developed specifically to address scoring failures by improving the accuracy of dimer ranking in protein docking [24].

Here, we assess docking success rates, sampling and scoring failures for crystal structures as well as computer-generated models for the BM1905 dataset. The results are shown as IS-score spectrum plots in Figure 5.2. For instance, at an IS-score of 0.210 corresponding to a \( p \)-value of \( 10^{-2} \), the success rate of ZDOCK against crystal structures is 73.4\%, with the remaining 26.6\% cases classified as scoring failures (Figure 5.2A). Re-ranking of the docked poses with eRank\textsuperscript{PPI} increases the success rate to 88.1\%, decreasing the rate of scoring failures to only 11.9\% (Figure 5.2B). For computer-generated models, the success rates (scoring failures) are 64.4\% (35.6\%) for ZDOCK and 71.9\% (28.1\%) for eRank\textsuperscript{PPI} (Figures 2C and 2D, respectively). Note that the lack of sampling failures at an IS-score of 0.210 suggests that rigid-body docking successfully samples the conformational space of dimers assembled with experimental as well as computer-generated models of monomer proteins. Sampling failures come into sight only at higher IS-score values, for example, conformations with an IS-score of at least 0.473 corresponding to a \( p \)-value of \( 10^{-10} \) are not constructed by ZDOCK for 19.1\% and 61.1\% of the cases when experimental monomer structures and computer-generated models are used, respectively. However, one should keep in mind that the models of individual monomers may already contain significant inaccuracies, thus interfaces highly similar to those in experimental structures simply cannot be constructed by rigid-body docking. Overall, this analysis shows that scoring failures are
responsible for the majority of unsuccessful docking calculations and that eRank\textsuperscript{PPI} improves the success rate by reducing the number of scoring failures by 14.7% for crystal structures and 7.5% for protein models.

![Graphs showing performance of ZDOCK, eRank\textsuperscript{PPI} and FiberDock on the BM1905 dataset.](image)

Figure 5.3. Performance of ZDOCK, eRank\textsuperscript{PPI} and FiberDock on the BM1905 dataset. Dimer complexes are constructed using (A) experimentally solved monomer structures (BM1905C) and (B) computer generated monomer models (BM1905H). The results are presented as the cumulative fraction of proteins with the IS-score between predicted and experimental complex structures larger than or equal to the value displayed on the x-axis.

**Dimers constructed from experimental monomer structures**

Interface quality in the modeled dimer structures is assessed in Figure 5.3 by the distribution of IS-scores [34] across the BM1905 dataset. Figure 5.3A shows the accuracy of complex models constructed from experimental monomeric structures with ZDOCK alone, ZDOCK followed by FiberDock, eRank\textsuperscript{PPI}, and eRank\textsuperscript{PPI} followed by FiberDock. For each receptor-ligand pair, we first selected the top 10 highest scoring ZDOCK models and picked the model with the best IS-score. At least one model with a statistically highly significant IS-
score of 0.473 is found in 34.9% of the cases. This percentage increases to 42.4% when the initial dimers are refined by FiberDock. Next, we re-ranked the top 2,000 models from ZDOCK with eRank\textsuperscript{PPI} in order to more reliably identify near-native structures. Encouragingly, in 50.5% of the cases, at least one model having an IS-Score higher than 0.473 is now found within the top 10 dimers re-ranked by eRank\textsuperscript{PPI}. Further refinement with FiberDock increases this fraction to as high as 57.5%. In addition to the IS-score, Table 5.1 shows that success rates measured with iRMSD increase when eRank\textsuperscript{PPI} and FiberDock are included in the modeling protocol.

Table 5.1. Comparison of the success rates for protein dimers assembled from the crystal structures and computer-generated models of monomers. The acceptance criteria for correct predictions are an iRMSD of ≤2.5 Å for crystal structures and ≤8.5 Å for protein models.

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Crystal structures</th>
<th>Protein models</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZDOCK</td>
<td>51.5%</td>
<td>28.1%</td>
</tr>
<tr>
<td>ZDOCK + eRank\textsuperscript{PPI}</td>
<td>58.3%</td>
<td>43.7%</td>
</tr>
<tr>
<td>ZDOCK + eRank\textsuperscript{PPI} + FiberDock</td>
<td>72.8%</td>
<td>52.4%</td>
</tr>
</tbody>
</table>

Altogether, eRank\textsuperscript{PPI} and FiberDock generate the most accurate dimers in these benchmarking calculations. Figure 5.3A and Table 5.1 show that re-ranking with eRank\textsuperscript{PPI} places more near-native structures within the top-ranked models compared to ZDOCK, which is in accordance with our previous studies [24] reporting ~10% improvement in the success rate. In general, the refinement by FiberDock considering both backbone and sidechain flexibility consistently improves the model accuracy, however, the improvement
clearly depends on the quality of the top-ranked dimers. Most significant improvement for models selected by eRankPPI is achieved when the IS-Score of the initial dimers is in the range of 0.4-0.8.

**Dimers constructed from computer-generated monomer structures**

The unavailability of experimentally determined structures for a vast majority of gene products necessitates using computer-generated models for genome-wide determination of PPIs. On that account, we investigate how protein docking, and dimer re-ranking and refinement are affected when computer-generated models are used instead of experimental structures. Figure 5.3B shows the accuracy of dimer models constructed using four protocols described above. Since monomers are weakly homologous models containing structural inaccuracies, the modeling results are evaluated with a lower, yet still statistically significant IS-score threshold of 0.311. We find that in 22.3% and 31.0% of the cases, at least one model with an IS-score of ≥0.311 is found within the top 10 conformations ranked by ZDOCK and eRankPPI, respectively. Furthermore, a flexible refinement with FiberDock increases the percentage of successful cases to 32.2% for ZDOCK and to 48.7% for eRankPPI. Table 5.1 shows that similar results are obtained with the iRMSD used to measure the success rate. Therefore, not only dimer models re-ranked by eRankPPI and additionally refined by FiberDock are the most accurate, but also the refinement procedure yields better improvements for eRankPPI compared to ZDOCK. Despite the fact that protein docking using weakly homologous monomer structures is a difficult task and the dimer accuracy cannot be expected to be higher than the accuracy of the monomers, our analysis demonstrates that, in
many cases, using a protocol combining eRankPPI and FiberDock constructs reliable complexes as assessed by the IS-score and the iRMSD.

**Predicting biologically relevant interactions**

Macromolecular complexes are stabilized by a variety of interactions including solvation effects, changes in the internal energy upon binding, electrostatics, van der Waals interactions, hydrogen bonds, π-stacking, and hydrophobic contacts across the interface. These interactions are prevalently found in the crystal structures of protein assemblies deposited in the PDB. Given that protein crystals mimic the actual interactions in an aqueous solution, biologically relevant complex structures can be predicted based on these contributions to the binding energy. Figure 5.4 shows the distribution of various energy terms calculated by FiberDock for the positive dataset HET3519 and the negative dataset RND14944. Note a clear distinction in the distribution of most energies between interacting and non-interacting protein pairs suggesting that these scores can be utilized to identify biologically relevant interactions. For example, the median attractive (repulsive) van der Waals energy is -0.230 (-0.187) and 0.214 (-0.195) for interacting and non-interacting pairs, respectively. Another highly discriminatory term is the hydrogen bond energy with the median value of -0.068 for interacting and 0.418 for non-interacting pairs, which is consistent with other studies reporting that the hydrogen bond potential greatly improves the recognition of correctly docked protein-protein complexes from large sets of alternative structures [42].

Next, we combine various interactions at the interface for the top 3 refined models in order to evaluate the complex stability and to predict whether the interaction is biologically
relevant or not. Specifically, the RFC is employed to estimate a probability that a given complex model represents a true interaction. Figure 5.5 shows a receiver operating characteristic plot evaluating the performance of a classifier separating true interactions within the HET3519 dataset from negative pairs present in the RND14944 dataset. Using the top-ranked model, the area under the curve for the prediction of biologically relevant interactions is 0.72. The probability threshold of 0.13 maximizes the MCC to a value of 0.43 at a true positive rate of 0.51 and a false positive rate of 0.14. Essentially, this threshold corresponds to a point in the ROC space farthest from the diagonal representing the performance of a random classifier.

![Boxplots](image)

Figure 5.4. Distribution of various components to the binding energy calculated with FiberDock. Negative pairs from the RND14944 dataset and positive pairs from the HET3519 dataset are shown as white and gray boxes, respectively. The following normalized (Z-score) energy terms are shown: (A) global energy, (B) attractive van der Waals potential, (C) repulsive van der Waals potential, (D) atomic contact energy, (E) internal energy, and (F) hydrogen bond potential. Boxes end at quartiles Q₁ and Q₃ and a horizontal line in each box is the median. Whiskers point at the farthest points that are within $3/2$ of the interquartile range.
Next, we improved the classification procedure by employing up to top 5 ranked models constructed for a given pair of receptor and ligand proteins. A pair is predicted to represent a true interaction if a positive predictive score is greater than the optimized probability threshold of 0.13 for at least one out of top \( n \) models. Table 5.2 shows that this strategy indeed enhances the discriminatory power. Considering the top 3 models maximizes the MCC to a value of 0.61 with a true positive rate of 0.81 and a false positive rate of 0.19 (a solid circle in Figure 5.5). Finally, we independently test our classification protocol against the Negatome 2.0 database, which provides a collection of protein pairs unlikely to physically interact with each other [27]. We obtained a false positive rate of 0.23, i.e. 23% of non-interacting pairs included in Negatome 2.0 are predicted as interacting proteins. This false positive rate is similar to that calculated for the HET3519 and RND14944 datasets suggesting that the RFC classifier is robust and its performance is independent on the validation dataset. Overall, the classifier performance is sufficiently high to be applicable at a proteome scale.

**Modeling protein-protein complex structures for \( E. coli \)**

All-against-all docking of 2,300 proteins in \( E. coli \) produced 2,643,850 possible binary PPIs with 3 putative dimer models generated for each unique receptor-ligand pair, totaling 7,931,550 3D complex structures of bacterial proteins. Applying the RFC trained on the HET3519 and RND14944 datasets predicted 425,412 biologically relevant interactions corresponding to 18.2% of all possible PPIs. Note that although the experimentally covered PPI space provided by DIP [31] is very limited with only 6,341 validated interactions, our
structure-based pipeline correctly identified 3,930 (62%) of these true PPIs. According to the BioGRID Database Statistics, an estimated number of 164,717 non-redundant interactions are present in *E. coli*, suggesting that that additional filters are required to further refine the set of predicted interactions. On that account, we added annotation filters from Gene Ontology to support the identification of biologically relevant dimers constructed for the *E. coli* proteome.

Figure 5.5. Receiver operating characteristic (ROC) plot evaluating the accuracy of the prediction of biologically relevant PPIs for the HET3519 and RND14944 datasets. The solid line corresponds to the performance of a Random Forest Classifier employing the top-ranked models with the black triangle pointing out the highest accuracy. Circles represent the performance achieved by considering the top 2, 3, 4 and 5 ranked models for each target complex. The gray area shows the performance of a random classifier.
Table 5.2. Accuracy of the prediction of biologically relevant PPIs for the HET3519 and RND14944 datasets. Here, we consider up to top 5 ranked models constructed for a given pair of receptor and ligand proteins.

<table>
<thead>
<tr>
<th>Number of models</th>
<th>MCC</th>
<th>TPR</th>
<th>FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.43</td>
<td>0.53</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.58</td>
<td>0.74</td>
<td>0.14</td>
</tr>
<tr>
<td>3</td>
<td>0.61</td>
<td>0.81</td>
<td>0.19</td>
</tr>
<tr>
<td>4</td>
<td>0.58</td>
<td>0.85</td>
<td>0.20</td>
</tr>
<tr>
<td>5</td>
<td>0.58</td>
<td>0.88</td>
<td>0.22</td>
</tr>
</tbody>
</table>

MCC – Matthews correlation coefficient; TPR – true positive rate; FPR – false positive rate.

**Integrating structure-based prediction with Gene Ontology**

First, we tested whether CC, BP and MF slims can be used as filters to identify interacting proteins by comparing GO annotations in positive and negative protein pairs. Here, the positive set contains known protein interactions according to the DIP database, whereas the negative set is compiled by randomly pairing *E. coli* proteins included in the DIP database. Those protein pairs having at least one common GO slim pass the annotation filter. About 82% of positives pass the CC filter that requires two proteins to co-localize in order to form a physical interaction. In contrast, only 58% of negatives are located in the same cellular component. Further, as many as 93% of positives are part of the same biological process, whereas 66% of negatives pass the BP filter. These results are in line with previous studies demonstrating that proteins localized in the same cellular compartment are more likely to interact than those residing in spatially distant compartments [43], [44]. Similarly, proteins involved in the same biological process have on average a higher chance to interact compared to molecules functioning in different biological processes. Thus, both CC and BP filters retain the majority of true interactions and reject a number of non-interacting protein pairs leading to a better classification performance. In contrast, molecular function cannot
be used to improve the identification of biologically relevant interactions because a similar percentage of positives (48%) and negatives (52%) pass the MF filter. To further corroborate these results, we applied both CC and BP filters to the HET3519 and RND14944 datasets. Encouragingly, as many as 91% and 93% of HET3519 complexes passed CC and BP filters, respectively. In contrast, significantly fewer pairs from the random dataset RND14944 passed CC (63%) and BP (44%) filters. The discriminatory performance of GO filters applied to HET3519 and RND14944 is consistent with that obtained for the *E. coli* dataset.

**Assembly and analysis of PPI network in *E. coli***

In order to assemble the network of protein-protein interactions in *E. coli*, we first applied the CC filter to 425,412 putative hetero-dimers identified by the RFC bringing this number down to 253,230 interactions between proteins localized in the same cellular compartment. Next, we selected only those protein pairs involved in the same biological process further reducing the number of putative hetero-dimers to 81,280. Although the BP filter is highly sensitive correctly identifying 93% of true interactions, this significant reduction of the number of positive predictions is mainly attributed to the fact that BP annotations are available for only 1,294 out of 2,300 proteins. Combining structure-based prediction of PPIs with both annotation filters results in 61,913 biologically relevant interactions. Note that GO filters are frequently employed to automatically refine large sets of protein interactions. For instance, the F-measure assessing the accuracy of PPI prediction for the bacterial chemotaxis signaling pathway increased from 0.52 to 0.69 when the protein localization was taken into consideration [21]. Our final set of protein interactions with
confidently modeled dimer conformations provide a tremendous source of structural data relating to the network of protein-protein interactions in *E. coli*.

Figure 5.6. Hive plots of PPI networks for the proteome of *E. coli*. Turquoise circles (nodes) represent individual proteins connected by interactions (edges). Three types of interactions are denoted by edges in different colors, positive predictions are gray, true positives (predicted interactions also present in the DIP database) are green, and false negatives (DIP interactions that are not predicted) are red. (A) Network constructed by modeling the structures of hetero-dimer complexes followed by the classification of interfaces with machine learning. (B) Random network comprising the same number of nodes and edges as the structure-based network, however, with interactions randomly assigned to pairs of nodes. *E. coli* proteins are assigned to three axes based on their degree $d$, low-degree ($d < 50$) on the $x$-axis, medium-degree ($50 \leq d \leq 80$) on the $y$-axis, and high-degree ($d > 80$) on the $z$-axis. Each axis is then split into two identical axes in order to show interactions within each group. Further, nodes on the axes are sorted by the increasing clustering coefficient $c$ with the maximum value of $c$ shown next to each axis (note the significant scale difference between A and B).

Subsequently, we investigated several properties of the PPI network constructed for *E. coli* in comparison with a random network comprising the same number of nodes and edges. The only difference between the predicted and random networks is that the latter is built on interactions randomly assigned to pairs of proteins. For the PPI network predicted for *E. coli* by the structure-based approach, the degree, diameter, and clustering coefficient
are 110.5, 6, and 0.30, respectively. Although the random network has a similar degree of 111.4, its diameter is 3 and the clustering coefficient is only 0.11. This analysis reveals that the global topology of the constructed network significantly differs from that of a random network. Specifically, the predicted PPIs tend to cluster together forming functional units around highly connected hubs, whereas PPIs are distributed more uniformly in a random network. In order to further corroborate these findings, we constructed a PPI network from experimental interactions included in the DIP database and the corresponding random network having the same number of nodes and edges. Here the degree, diameter and clustering coefficient calculated for the DIP (random) network are 6.9 (6.8), 12 (7), and 0.08 (0.004), respectively. The differences between the network predicted by a structure-based approach and that built on interaction data from DIP result from the incompleteness of the latter, i.e. the DIP network is sparse, having about 17 times less connections per node than the predicted network. Nonetheless, the deviations of both networks from their random counterparts are qualitatively similar showing a notable tendency to form clusters and sub-networks.

Figure 5.6 shows hive plots [46] generated for the predicted (Figure 5.6A) and random (Figure 5.6B) networks of PPIs in *E. coli*. In both plots, true positives and false positives with respect to experimentally validated interactions from the DIP database are colored in green and red, respectively. First, the structure-based approach including GO filters correctly identifies the majority of experimental interactions (green lines), whereas these connections are largely missed in the random network (red lines). Second, the axes in both hive plots are sorted by the clustering coefficient of individual nodes and the axis scales in Figures 6A and 6B are significantly different. Third, considering the global network
topology, the majority of nodes in the random network are assigned to a medium-degree group (y-axis) forming extensive connections to themselves as well as to low- (x-axis) and high-degree (z-axis) groups. In contrast, extensive connections between all groups are present in the network predicted by the modeling of quaternary structures. These hive plots effectively visualize differences between the predicted and random networks described above.

**Examples of dimer models selected from the *E. coli* network**

Since the PPI network for the *E. coli* proteome is assembled by the modeling of interactions between proteins, we discuss a couple of representative examples of the modeled dimer structures. Note that experimentally solved structures are unavailable for these proteins, therefore, the presented molecular assemblies have been constructed solely from the primary sequences of individual monomers. Although monomer models are built on templates whose sequence identity to the target protein is less than 40%, the estimated Global Distance Test (GDT) [47] is greater than 0.7 indicating that these computer-generated structures are highly confident. The first example is a hetero-dimer assembled from fadJ and fadI proteins involved in the fatty acid beta oxidation pathway, which is part of lipid metabolism. This interaction was proposed to increase the efficiency of anaerobic beta-oxidation by favoring substrates of different chain length [48]. Even though there is experimental evidence that these two proteins interact with one another [49], no structural data is available for the individual proteins nor the complex. The modeling procedure developed in this study correctly identified these proteins to be interaction partners with the putative fadJ/fadI hetero-dimer shown in Figure 5.7. A protein binding site confidently
predicted by eFindSitePPI on fadJ comprises 11 residues, out of which 9 are also found at the interface in the modeled fadJ/fadI complex. Moreover, fadJ has a NAD binding domain according to the Pfam database [50]. Interestingly, we were able to not only identify a binding pocket for NAD in the fadJ structure model with eFindSite [51], but also to dock a NAD molecule to this pocket using our in-house ligand docking software eSimDock [52].

Figure 5.7. Example of PPI prediction for a hetero-dimer. Cartoon representation of the dimer complex of fadI (yellow) and fadJ (purple). Interface residues predicted for the receptor are shown as a solid surface. A small molecule ligand (NAD) docked to fadJ is shown as sticks colored by atom type.

The second example is glutaminase 2 (glsA2), an amidohydrolase enzyme responsible for generating glutamate from glutamine, demonstrated to be a self-assembling protein [53]. The GDT of the glsA2 monomer estimated by eThread is 0.78 indicating a confident structure model. Next, we predicted the structure of glsA2 homo-dimer as a symmetric complex shown in Figure 5.8. A unique feature of eFindSitePPI is that it not only detects interaction sites, but
also points out specific molecular interactions that stabilize a putative complex. Molecular interactions predicted by eFindSitePPI for glsA2 include a salt bridge between the side chains of R232 (chain A) and E82 (chain B) as well as aromatic contacts between W252 (chain A) and W252 (chain B), which are found in the top-ranked complex model selected by eRankPPI.

![Image of PPI prediction for a homo-dimer]

Figure 5.8. Example of PPI prediction for a homo-dimer. Cartoon representation of the dimer complex of YneH with chains A and B colored in green and blue, respectively. Protein interfaces predicted for the monomers are shown as a solid surface. Residues predicted to be involved in a salt bridge R32(A)-E28(B) and aromatic contact W525(A)-W525(B) are shown as balls and sticks.

CONCLUSION

In this work, we developed a new method combining molecular modeling, structural bioinformatics, machine learning, and functional annotation data to predict PPIs across proteomes. We first comprehensively tested this protocol on representative datasets of experimental structures and computer-generated models of protein dimers and then we applied this methodology to predict PPIs across the proteome of *E. coli* and within the human immune disease pathway. Our results indicate that protein docking supported by
evolutionary restraints and machine learning can be used to reliably identify and model biologically relevant protein assemblies. Furthermore, the accuracy of the identification of interaction partners can greatly be improved by including only those protein pairs co-localized in the same cellular compartment and involved in the same biological process. The proposed method can be applied to detect PPIs in other organisms and pathways as well as to construct structure models and estimate the confidence of interactions experimentally identified with high-throughput techniques. Finally, with the growing volume of structural data, experimentally confirmed protein interactions, and functional annotation, we expect the coverage and accuracy of our approach to increase over time.

REFERENCES


CHAPTER 6: PPI PREDICTION FOR HUMAN PATHWAYS

INTRODUCTION

Interaction between proteins are critical to numerous biological processes, thus they are considered as a core of the cellular interactome [1]. Exhaustive exploration of protein-protein interactions is one of the central objectives of systems biology, with the goal to elucidate the role of human genes in health and in disease. While the estimated size of the human interactome ranges from 130,000 [2] to 650,000 [3] PPI, databases report only 41,000 binary interaction between human proteins and quite a few of them may be in error because of the limitations of high-throughput experimental interaction discovery methods. In that regard we have developed a high-throughput PPI-network prediction pipeline as described in chapter 5. The prediction method adopts an all-to-all docking strategy to predict structurally characterized PPI networks. Identification of PPI networks using exhausting docking calculation requires massive computational resources, however, availability of supercomputers makes these large-scale calculations feasible. In this study, we applied our prediction method to reconstruct protein-protein interaction networks for nine human pathways. The pathways selected for the analysis are: cellular response to hypoxia, diseases of immune system, diseases of glycosylation, metabolism of porphyrins, myogenesis, nitric oxide metabolism, oncogene induced senescence, regulation of beta cell development and synthesis of IP3 and IP4.

MATERIALS AND METHODS

Datasets

Pathway information for the human genome was obtained from the Reactome database [4] in June 2016. A total of nine Reactome pathways were utilized for PPI prediction
and structural characterization. Computer generated models of individual proteins in each pathway were built by threading and homology modeling using eThread2.0 [5] and Modeller [6]. The quality of the models was measured by its estimated global distance test score (GDT) [7], a widely-used measure to estimate the deviation between structure model and the unknown native structure. GDT ranges from 0 to 1, values close to 1 suggest near perfect structural alignments. Table 6.1 shows the number of proteins modeled for each pathway, the average size of the proteins involved (proportional to the number of residues), standard deviation and the quality of the protein models.

Table 6.1. Description of the modeled proteins in each human pathway.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Number of proteins</th>
<th>Average protein size</th>
<th>Protein size, std. dev.</th>
<th>% cases, GDT_TS &gt; 0.4</th>
<th>% cases, GDT_TS &gt; 0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response to hypoxia</td>
<td>19</td>
<td>370.696</td>
<td>243.688</td>
<td>86.9565</td>
<td>47.8261</td>
</tr>
<tr>
<td>Disease of immune system</td>
<td>17</td>
<td>607.25</td>
<td>225.148</td>
<td>87.5</td>
<td>20.8333</td>
</tr>
<tr>
<td>Diseases of glycosylation</td>
<td>18</td>
<td>456.676</td>
<td>222.018</td>
<td>72.973</td>
<td>18.9189</td>
</tr>
<tr>
<td>Metabolism of porphyrins</td>
<td>17</td>
<td>386.529</td>
<td>133.051</td>
<td>94.1176</td>
<td>47.0588</td>
</tr>
<tr>
<td>Myogenesis</td>
<td>15</td>
<td>519.682</td>
<td>257.705</td>
<td>68.1818</td>
<td>18.1818</td>
</tr>
<tr>
<td>Nitricoxide metabolism</td>
<td>13</td>
<td>411.647</td>
<td>287.764</td>
<td>88.2353</td>
<td>64.7059</td>
</tr>
<tr>
<td>Oncogene senescence</td>
<td>23</td>
<td>426.034</td>
<td>246.47</td>
<td>62.069</td>
<td>41.3793</td>
</tr>
<tr>
<td>Regulation of beta cell</td>
<td>22</td>
<td>380.357</td>
<td>139.992</td>
<td>57.1429</td>
<td>32.1429</td>
</tr>
<tr>
<td>development</td>
<td>Synthesis of IP3 and IP4</td>
<td>10</td>
<td>605.211</td>
<td>293.465</td>
<td>84.2105</td>
</tr>
</tbody>
</table>

**PPI prediction**

interaction using random forest classification. The PPI prediction pipeline is described in detail in chapter 5.

**Evaluation of prediction performance**

The prediction results are evaluated using true positives (TP), false negative (FN) and true positive rate (TPR). TP is the number of true interactions correctly predicted. False negative is the number of true interaction not predicted by the pipeline.

True positive rate: \[ TPR = \frac{TP}{TP + FN} \]

**RESULTS AND DISCUSSION**

Here we have conducted PPI network prediction for nine different human pathways using an exhaustive all-to-all docking strategy. All the proteins found within a pathway were docked against all others via a rigid body docking, using ZDOCK. The top 500 docked models obtained from ZDOCK were then re-ranked by a recently developed re-ranking function called eRankPPI, which helps in the selection of top 10 best rigid-body docked solutions. Since flexibility and dynamics play an important role in PPIs, such as in the induced fit models, we used a soft docking software, FiberDock on the top 10 models to allow for backbone and side chain flexibility. Finally, we analyzed various interactions at the interface for the top 3 flexibly refined models in order to evaluate the complex stability and to predict whether the interaction is physically feasibly or not. A random forest classifier is employed to evaluate the stability of the complex and estimate a probability that a given complex model represents a true interaction.
PPI detection

Any two proteins A and B in a given pathway were docked against each other twice. In the first case, rigid body exhaustive search of orientation was conducted between fixed receptor A with respect to a mobile ligand B, let’s call the resulting complex AB. In the second case A serves as the mobile ligand while B is the fixed receptor and the resulting complex is called BA. Note that AB and BA are reciprocal interactions. The stability of the interfaces of these reciprocal interactions are evaluated using the random forest classifier and if either one of them has a positive predictive probability greater than the predefined threshold, then protein A and B are predicted to be a true interacting pair. In our previous study, we obtained an optimum threshold of 0.13 to maximize the Matthew’s correlation coefficient on our benchmark dataset, so we used the same threshold for the current human pathway study. Table 6.2 shows the assessment of our PPI prediction using the aforementioned scheme. The total number of known PPIs found in the nine selected human pathways is 274, out of which we could correctly predict 184 PPIs, which yields a TPR of 67.15%. While the true positive rate is high, we observed that 68.06% of the total unique PPI combinations built by our pipeline are predicted as positives. Clearly, the positive prediction rate obtained using this strategy is too high, therefore we decided to test other strategies in order to reduce the positive prediction rate without having to compromise much on the true positive rate.

Towards this goal, we imposed additional restraints to the PPI classification scheme. Instead of having at least one docking result of the reciprocal interaction meet the threshold criterion, we imposed the threshold criterion on both the reciprocals, i.e. two protein A and B were predicted to have a true interaction if the docked models AB and BA both had a positive predictive probability greater than a predefined threshold. Using a threshold of 0.13
in the new scheme yielded a TPR of 40% at a positive prediction rate of 32.4%. Although the new strategy reduced the positive prediction rate by half, the TPR also decreased by almost 27%. Thus, the stringent strategy reduced the predictive power of the method. One potential way to overcome this problem was to relax the restrain by slightly reducing the threshold.

Table 6.2. Prediction results using original prediction scheme (AB or BA > 0.13).

<table>
<thead>
<tr>
<th>Pathway name</th>
<th>Unique combinations</th>
<th>Positive predictions</th>
<th>Known Positives</th>
<th>True Positives</th>
<th>False Negatives</th>
<th>TPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cellular response to hypoxia</td>
<td>182</td>
<td>121</td>
<td>93</td>
<td>56</td>
<td>37</td>
<td>0.60</td>
</tr>
<tr>
<td>Diseases of immune system</td>
<td>153</td>
<td>122</td>
<td>58</td>
<td>48</td>
<td>10</td>
<td>0.82</td>
</tr>
<tr>
<td>Diseases of glycosylation</td>
<td>308</td>
<td>196</td>
<td>11</td>
<td>8</td>
<td>3</td>
<td>0.72</td>
</tr>
<tr>
<td>Metabolism of porphyrins</td>
<td>149</td>
<td>85</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>0.71</td>
</tr>
<tr>
<td>Myogenensis</td>
<td>94</td>
<td>73</td>
<td>43</td>
<td>31</td>
<td>12</td>
<td>0.72</td>
</tr>
<tr>
<td>Nitricoxide metabolism</td>
<td>90</td>
<td>50</td>
<td>14</td>
<td>6</td>
<td>8</td>
<td>0.42</td>
</tr>
<tr>
<td>Oncogene induced senescence</td>
<td>276</td>
<td>174</td>
<td>41</td>
<td>25</td>
<td>16</td>
<td>0.60</td>
</tr>
<tr>
<td>Regulation of beta cell development</td>
<td>246</td>
<td>194</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>0.75</td>
</tr>
<tr>
<td>Synthesis of IP3 and IP4</td>
<td>55</td>
<td>42</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0.66</td>
</tr>
<tr>
<td>All Total</td>
<td>1553</td>
<td>1057</td>
<td>274</td>
<td>184</td>
<td>90</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Table 6.3 shows the effect of decreasing thresholds on the overall TPR and Positive Prediction rate of the method. As expected, decreasing the threshold from 0.13 to 0.10 in the new strategy improved the true positive rate. Although the positive prediction rate increased as well, it was not as high as that observed in the original scheme. At a threshold of 0.10,
Table 6.3. Alternative prediction strategies and their overall result.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Threshold</th>
<th>True positive rate</th>
<th>Positive prediction rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB or BA</td>
<td>0.13</td>
<td>0.67</td>
<td>0.68</td>
</tr>
<tr>
<td>AB and BA</td>
<td>0.13</td>
<td>0.40</td>
<td>0.32</td>
</tr>
<tr>
<td>AB and BA</td>
<td>0.12</td>
<td>0.43</td>
<td>0.38</td>
</tr>
<tr>
<td>AB and BA</td>
<td>0.11</td>
<td>0.48</td>
<td>0.46</td>
</tr>
<tr>
<td>AB and BA</td>
<td>0.1</td>
<td>0.57</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Table 6.4. Prediction results using the modified prediction scheme (AB and BA > 0.10).

<table>
<thead>
<tr>
<th>Pathway name</th>
<th>Unique PPIs</th>
<th>Positive predictions</th>
<th>Known Positives</th>
<th>True Positives</th>
<th>False Negatives</th>
<th>TPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response to hypoxia</td>
<td>182</td>
<td>80</td>
<td>93</td>
<td>41</td>
<td>52</td>
<td>0.44</td>
</tr>
<tr>
<td>Diseases of immune system</td>
<td>153</td>
<td>91</td>
<td>58</td>
<td>38</td>
<td>20</td>
<td>0.65</td>
</tr>
<tr>
<td>Diseases of glycosylation</td>
<td>308</td>
<td>111</td>
<td>11</td>
<td>8</td>
<td>3</td>
<td>0.72</td>
</tr>
<tr>
<td>Metabolism of porphyrins</td>
<td>149</td>
<td>66</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>0.71</td>
</tr>
<tr>
<td>Myogenesis</td>
<td>94</td>
<td>57</td>
<td>43</td>
<td>25</td>
<td>18</td>
<td>0.58</td>
</tr>
<tr>
<td>Nitricoxide metabolism</td>
<td>90</td>
<td>38</td>
<td>14</td>
<td>7</td>
<td>7</td>
<td>0.5</td>
</tr>
<tr>
<td>Oncogene senescence</td>
<td>276</td>
<td>146</td>
<td>41</td>
<td>27</td>
<td>14</td>
<td>0.65</td>
</tr>
<tr>
<td>Regulation of beta cell</td>
<td>246</td>
<td>163</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>0.75</td>
</tr>
<tr>
<td>Synthesis of IP3 and IP4</td>
<td>55</td>
<td>33</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>All Total</strong></td>
<td><strong>1553</strong></td>
<td><strong>785</strong></td>
<td><strong>274</strong></td>
<td><strong>117</strong></td>
<td><strong>150</strong></td>
<td><strong>0.57</strong></td>
</tr>
</tbody>
</table>

A TPR of 57.29 is obtained at a positive prediction rate of 50.05%. The TPR obtained is 10% less than the one obtained in the original scheme, however the positive prediction rate decreases by 18%. Thus, we decided to use a threshold of 0.10 in the improvised classification method, which requires the docking complexes of both the reciprocal interactions (AB and BA) to have a positive predictive probability greater than the threshold. The detailed results obtained for individual pathways using this scheme are shown in Table 6.4.
**Analysis of PPIs in the human immune disease pathway**

We modeled protein complex structures for the human immune disease pathway associated with the TLR signaling cascade. TLRs are sensors of the innate immune system recognizing pathogen-associated molecular patterns [12], [13]. These molecular sensors participate in the first line of defense against invading pathogens by promoting the activation and nuclear translocation of certain transcription factors to induce the secretion of inflammatory cytokines. Out of 26 gene products involved in this pathway, we included the following 17 proteins whose 3D structures have been modeled (estimated GDT values are given in parentheses): P58753 (0.64), Q15399 (0.45), Q9Y2C9 (0.46), P08571 (0.48), P16671 (0.59), O15111 (0.56), O14920 (0.54), Q99836 (0.48), Q9NWZ3 (0.65), O60602 (0.49), Q15653 (0.71), Q00653 (0.32), Q04206 (0.52), P25963 (0.70), P19838 (0.33), Q9BXR5 (0.41), and Q9Y6Y9 (0.77). The remaining 9 structures have not been modeled due to either their large size, the unavailability of reliable templates, a significant content of transmembrane regions, or because these are membrane proteins. Although the total number of possible interactions for this dataset is 153, only 58 are confirmed experimentally according to the Reactome database. Figure 6.1 shows the network structure and a binary interaction matrix for PPIs predicted for this pathway. The structure-based approach predicted a total of 90 unique interactions (dashed blue connections in Figure 6.1A) including 38 known interactions (solid green connections in Figure 6.1A). Only 20 known interactions have not been predicted by the quaternary structure modeling (dotted red connections in Figure 6.1A). Therefore, about two-thirds of true PPIs were correctly recovered by the modeling of the complex structures of proteins involved in the human immune disease pathway.
Figure 6.1. Structure-based prediction of PPIs for the human immune disease pathway. (A) Network diagram of the human immune disease pathway. Yellow circles (nodes) represent individual proteins connected by interactions (edges). Three types of interactions are denoted by edges in different colors, positive predictions are blue, true positives (predicted interactions also present in the Reactome database) are green, and false negatives (interactions from Reactome that are not predicted) are red. (B) Matrix of binary interactions including positive predictions (blue), true positives (green), and false negatives (red). Circles marked with a star and a dot show those protein pair that pass and fail the CC filter, respectively. UniProt IDs of proteins involved in this pathway according to the Reactome database are shown in both A and B.

In addition, positive predictions, true positives and false negatives are shown as a binary interaction matrix in Figure 6.1B. Here, we also mapped GO Slims for the cellular component to individual proteins in order to improve the PPI prediction accuracy by including function annotation filters. Since GO annotations were available only for 8 proteins, the CC filter was applied to 17 hetero-dimer models constructed by our structure-based approach. Encouragingly, 12 of the predicted complexes passed the CC filter (black stars in Figure 6.1B), while only 5 failed (black dots in Figure 6.1B). Although, the GO annotation filter can be applied only to a small fraction of structure-based predictions for this pathway,
it turned out to be quite accurate. Therefore, we expect that new function annotations available in the future will selectively reduce the number of positive predictions leading to more accurate PPI prediction results.

**Analysis of PPIs in the synthesis of IP3 and IP4 pathway**

Figure 6.2 shows the network structure for PPIs predicted for the metabolic pathway for the synthesis of inositol trisphosphate (IP3) and tetrakisphosphate (IP4) in the cytosol. IP3 and IP4 molecules are involved in calcium signaling and are synthesized by the action of various kinases and phosphatases in the cytosol. Out of 19 gene products involved in this pathway, we included the following 10 proteins whose 3D structures have been modeled (estimated GDT values are given in parentheses): O43314 (0.49), O95989 (0.75), Q13572 (0.70), Q6PFW1 (0.34), Q8NFP7 (0.93), Q92551 (0.63), Q96G61 (0.93), Q96PC2 (0.54), Q9H8X2 (0.58), Q9NZJ9 (0.73). Although the total number of possible interactions for this dataset is 55, only 3 homo-protein complexes are confirmed experimentally according to the Reactome database. The structure-based approach predicted a total of 33 unique interactions (dashed blue connections in Figure 6.2) including the 3 known homo-protein interactions (solid green loops in Figure 6.2). The large number of positive predictions can be attributed to the fact that all these proteins are either kinases and phosphatases, interactions amongst which is known to underpin cellular regulation. Prediction results for this pathway can be particularly useful to restrict the search space before utilizing expensive PPI analysis methods especially because 6 out of 10 proteins do not have an experimentally determined structure available in the PDB [14] and very limited experimental data is available on their protein-protein interactions.
Figure 6.2. Network diagram of the synthesis of IP3 and IP4 pathway. Yellow circles (nodes labeled with Uniprot ID of the protein) represent individual proteins connected by interactions (edges). Two types of interactions are denoted by edges in different colors, positive predictions are blue, true positives (predicted interactions also present in the Reactome database) are green.

CONCLUSION

We conducted a reconstruction of protein-protein interaction network using our PPI prediction pipeline for nine human pathways. The results show that the proposed PPI prediction pipeline can be used for large-scale characterization of PPIs in any organism. The major novel aspect of the proposed pipeline is that it approaches PPI networks from a structure-oriented perspective, which provides a comprehensive picture of complex biological pathway at the fundamental level of molecular interactions. Finally, with the growing volume of structural data and increasing coverage of functional annotation we expect the accuracy of our method to increase over time.
REFERENCES


CHAPTER 7: CONCLUSIONS

It has been more than a decade since the completion of the Human Genome Project that provided us with a complete list of human proteins. The next obvious task is to figure out how various parts interact with each other. Protein interaction networks are the cornerstone to our understanding of the complex genome-to-phenome relationship, both in health and disease. On that account, I have designed and developed a novel pipeline for structure-based prediction of protein-protein interaction networks with the goal to expand the coverage of the interaction space and unveil the structural details at atomic resolution. It is noteworthy that the proposed method can be easily adapted for modeling and estimating the reliability of experimentally identified interactions. Combining prediction data with high-throughput experimental data would lower false positive rates and yield more accurate results.

In chapter 2, I reported an analysis which showed that the location of binding sites as well as the interfacial geometry is highly conserved in evolutionarily weakly related dimer proteins, irrespective of the global structure similarity. Exploiting these insights, I developed a novel approach called eFindSite$^{PPI}$ to predict protein-binding sites. eFindSite$^{PPI}$ integrates sensitive meta-threading techniques with structure alignments and machine learning to locate putative interfacial sites in target proteins. A novel feature of eFindSite$^{PPI}$ is that it also detects the types of molecular interactions that target proteins are likely to form with their interacting partners; this is demonstrated for hydrogen bonds, salt bridges as well as hydrophobic and aromatic contacts. Knowledge of binding regions of a protein can inform both experiments and other types of predictions. For instance, mutagenesis experiments can be guided to pinpoint functionally important residues of binding proteins and receptors.
Also, knowledge of binding region location can reduce the size of the conformation space to search during the structure prediction of a protein complex. eFindSite$^{\text{PPI}}$ is available as a web server and a stand-alone software package. The web application provides the scientific community with a user-friendly interface for job submission as well as the interpretation of results and data download. The stand-alone package can be installed locally for high throughput computations. (http://brylinski.cct.lsu.edu/efindsiteppi)

In chapter 3, I performed a comprehensive comparative analysis of the performance of eFindSite$^{\text{PPI}}$ with other interface prediction methods such as ProMate [1], PredUS [2], cons-PPISP [3], WHISCY [4], PriSE[5] and PINUP [6] to name a few. I found that although structure-based prediction algorithms perform better than sequence-based methods, their accuracy strongly depends on the quality of query protein structures. However, in contrast to other structure-based algorithms, eFindSite$^{\text{PPI}}$ tolerates small and moderate distortions in the structure of the query protein. Furthermore, I also showed that combining the outputs from various prediction methods typically outperforms the best single algorithm, therefore, consensus predictions by meta-predictors are likely to significantly improve the accuracy of interface residue prediction.

In chapter 4, I developed eRank$^{\text{PPI}}$, an algorithm for the selection of correct docking conformations constructed by rigid-body protein docking. eRank$^{\text{PPI}}$ features a new scoring function that integrates the predicted interface location with protein docking potentials and a contact-based symmetry score. Comprehensive benchmarking calculations show that eRank$^{\text{PPI}}$ has a high tolerance to structural imperfections in computer-generated protein models. eRank$^{\text{PPI}}$ is designed to bridge the gap between the volume of sequence data, the evidence of binary interactions, and the atomic details of pharmacologically relevant protein
complexes. Tolerating structure imperfections in computer-generated models opens up a possibility to conduct the exhaustive structure-based reconstruction of PPI networks across proteomes.

In chapter 5, I designed a pipeline that combines molecular modeling, structural bioinformatics, machine learning, and functional annotation data for the prediction and modeling of protein-protein interactions on a proteome-wide level. I tested our method on a benchmark dataset and then applied it on the *Escherichia coli* proteome. I validated three necessary assumptions taken in our approach namely, (1) Docking can be used for identifying interaction partners, (2) interacting proteins co-localize and (3) interacting proteins function in the same biological process.

In chapter 6, I applied the pipeline describe in chapter 5 on nine human pathways. Results indicate that protein docking supported by evolutionary restraints and machine learning can be used to reliably identify and model biologically relevant protein assemblies. Finally, with the growing availability of experimentally determined structural data and its improving coverage, the accuracy of this approach is expected to increase over time.

REFERENCES


APPENDIX A: LETTER OF PERMISSION TO REPRINT CHAPTER 2

Copyright permission for including chapter 2 in dissertation.

<table>
<thead>
<tr>
<th>JOHNS WILEY AND SONS LICENSE TERMS AND CONDITIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar 20, 2017</td>
</tr>
</tbody>
</table>

This Agreement between Surabhi Maheshwari ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

<table>
<thead>
<tr>
<th>License Number</th>
<th>4073160185713</th>
</tr>
</thead>
<tbody>
<tr>
<td>License date</td>
<td></td>
</tr>
<tr>
<td>Licensed Content Publisher</td>
<td>John Wiley and Sons</td>
</tr>
<tr>
<td>Licensed Content Publication</td>
<td>Journal of Molecular Recognition</td>
</tr>
<tr>
<td>Licensed Content Title</td>
<td>Prediction of protein-protein interaction sites from weakly homologous template structures using meta-threading and machine learning</td>
</tr>
<tr>
<td>Licensed Content Author</td>
<td>Surabhi Maheshwari,Michal Brylinski</td>
</tr>
<tr>
<td>Licensed Content Date</td>
<td>Jan 29, 2015</td>
</tr>
<tr>
<td>Licensed Content Pages</td>
<td>14</td>
</tr>
<tr>
<td>Type of use</td>
<td>Dissertation/Thesis</td>
</tr>
<tr>
<td>Requestor type</td>
<td>Author of this Wiley article</td>
</tr>
<tr>
<td>Format</td>
<td>Print and electronic</td>
</tr>
<tr>
<td>Portion</td>
<td>Full article</td>
</tr>
<tr>
<td>Will you be translating?</td>
<td>No</td>
</tr>
<tr>
<td>Expected completion date</td>
<td>May 2017</td>
</tr>
<tr>
<td>Expected size (number of pages)</td>
<td>250</td>
</tr>
</tbody>
</table>
APPENDIX B: LETTER OF PERMISSION TO REPRINT CHAPTER 3

Copyright permission for including chapter 3 in dissertation.

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Mar 20, 2017

This Agreement between Surabhi Maheshwari ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

<table>
<thead>
<tr>
<th>License Number</th>
<th>4073160385092</th>
</tr>
</thead>
<tbody>
<tr>
<td>License date</td>
<td></td>
</tr>
<tr>
<td>Licensed content publisher</td>
<td>Oxford University Press</td>
</tr>
<tr>
<td>Licensed content publication</td>
<td>Briefings in Bioinformatics</td>
</tr>
<tr>
<td>Licensed content title</td>
<td>Predicting protein interface residues using easily accessible on-line resources</td>
</tr>
<tr>
<td>Licensed content author</td>
<td>Maheshwari, Surabhi; Brylinski, Michal</td>
</tr>
<tr>
<td>Licensed content date</td>
<td>2015-03-21</td>
</tr>
<tr>
<td>Type of Use</td>
<td>Thesis/Dissertation</td>
</tr>
<tr>
<td>Institution name</td>
<td></td>
</tr>
<tr>
<td>Title of your work</td>
<td>Structure-based prediction of protein-protein interaction networks across proteomes.</td>
</tr>
<tr>
<td>Publisher of your work</td>
<td>n/a</td>
</tr>
<tr>
<td>Expected publication date</td>
<td>May 2017</td>
</tr>
<tr>
<td>Permissions cost</td>
<td>0.00 USD</td>
</tr>
<tr>
<td>Value added tax</td>
<td>0.00 USD</td>
</tr>
<tr>
<td>Total</td>
<td>0.00 USD</td>
</tr>
<tr>
<td>Requestor Location</td>
<td>Surabhi Maheshwari</td>
</tr>
<tr>
<td></td>
<td>1315 Bob Petit Blvd</td>
</tr>
<tr>
<td></td>
<td>Apt 39</td>
</tr>
<tr>
<td></td>
<td>BATON ROUGE, LA 70820</td>
</tr>
<tr>
<td></td>
<td>United States</td>
</tr>
<tr>
<td></td>
<td>Attn: Surabhi Maheshwari</td>
</tr>
<tr>
<td>Publisher Tax ID</td>
<td>GB125506730</td>
</tr>
</tbody>
</table>
**APPENDIX C: LETTER OF PERMISSION TO REPRINT CHAPTER 1 AND 7**

Copyright permission for including excerpts in chapter 1 and 7 in dissertation.

---

**ELSEVIER LICENSE TERMS AND CONDITIONS**

Mar 20, 2017

This Agreement between Surabhi Maheshwari ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

<table>
<thead>
<tr>
<th>License Number</th>
<th>4073160529780</th>
</tr>
</thead>
<tbody>
<tr>
<td>License date</td>
<td></td>
</tr>
<tr>
<td>Licensed Content Publisher</td>
<td>Elsevier</td>
</tr>
<tr>
<td>Licensed Content Publication</td>
<td>Methods</td>
</tr>
<tr>
<td>Licensed Content Title</td>
<td>Template-based identification of protein–protein interfaces using eFindSitePPI</td>
</tr>
<tr>
<td>Licensed Content Author</td>
<td>Surabhi Maheshwari, Michal Brylinski</td>
</tr>
<tr>
<td>Licensed Content Date</td>
<td>15 January 2016</td>
</tr>
<tr>
<td>Licensed Content Volume</td>
<td>93</td>
</tr>
<tr>
<td>Licensed Content Issue</td>
<td>n/a</td>
</tr>
<tr>
<td>Licensed Content Pages</td>
<td>8</td>
</tr>
<tr>
<td>Start Page</td>
<td>64</td>
</tr>
<tr>
<td>End Page</td>
<td>71</td>
</tr>
<tr>
<td>Type of Use</td>
<td>reuse in a thesis/dissertation</td>
</tr>
<tr>
<td>Intended publisher of new work</td>
<td>other</td>
</tr>
<tr>
<td>Portion</td>
<td>full article</td>
</tr>
<tr>
<td>Format</td>
<td>both print and electronic</td>
</tr>
<tr>
<td>Are you the author of this Elsevier article?</td>
<td>Yes</td>
</tr>
<tr>
<td>Will you be translating?</td>
<td>No</td>
</tr>
<tr>
<td>Order reference number</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX D: LETTER OF PERMISSION TO REPRINT CHAPTER 4

Copyright permission for including chapter 4 in dissertation.

Chapter 4 is a reprint of materials that appear in:


Copyright of this open access article published by BioMed Central is retained by the authors. It appears as chapter 4 in this dissertation with permission from the authors. The dissertation author is the primary author of this paper.
VITA

Surabhi Maheshwari was born and raised in Indore, India. She graduated with a Bachelor of Science degree in Biomedical Engineering from Shri Govindram Sekseria Institute of technology and science Indore, India in July 2010. She received a Master of Science degree in Bionformatics from University of Georgia in 2012. She began her doctoral research in Department of Biological Sciences at Louisiana State University under the guidance of Dr. Michal Brylinski in Fall 2012. Surabhi is expected to graduate with the Degree of Philosophy in Biochemistry in May 2017.