Sequence homology based protein-protein interacting residue predictions and the applications in ranking docked conformations

by

Li Xue

A Dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:

Vasant Honavar, Co-major Professor

Drena Dobbs, Co-major Professor

Robert Jernigan

Guang Song

Peng Liu

Iowa State University

Ames, Iowa

2012

DEDICATION

I would like to dedicate this dissertation

to my hard working parents - my father, for giving me the heart of curiosity about nature; my mother, who sacrificed her years to our family, for her endless love and instilling the importance of diligence;

to my sister for the sunshine of her love, for teaching me the first English letter, sharing my happiness and walking me through the toughness;

to my brother-in-law for being a true brother, for his encouragement, guidance and financial assistance;

to Lu Li for her loving protection, without whom I could not have come to the U.S.A. for this study;

to Ying Zi for being my rock, for every rainy and sunny moment that we shared in the past 13 years.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES x

ACKNOWLEDGEMENTS xxi

ABSTRACT xxiii

CHAPTER 1. Overview 1

1.1 Experimental Methods to Identify Protein-Protein Interface Residues 2

1.2 Protein-Protein Interface Features 4

1.3 Computational Methods to Predict Protein-Protein Interface Residues 5

1.4 Protein-Protein Docking 6

1.5 Interface Residue Conservation 8

1.6 Hypotheses and Overview of This Work 9

1.7 Main Contributions of This Work 11

CHAPTER 2. Conservation Analysis of Protein-Protein Interface Residues . 13

2.1 Introduction 14

2.2 Results 16

2.2.1 Conservation of PPIs in Non-Partner Specific (NPS) Interfaces 16

2.2.2 Conservation of PPIs in Partner-Specific (PS) Interfaces 25

2.3 Discussion 28

2.3.1 Protein Interface Conservation across Structure Space 28

2.3.2 Distance Functions for Identifying Putative Homologs with Conserved Interfaces 29

2.3.3 Conservation of Interfaces in Obligate and Transient Complexes 29

2.3.4 Interface *Residue* Conservation and Interface *Position* Conservation . . 29
2.4 Conclusions . . . 31
2.5 Methods . . . 31
2.5.1 Datasets . . . 31
2.5.2 Interface Definition . . . 33
2.5.3 Mapping Interfaces in Structures to Sequences . . . 33
2.5.4 NCBI BLAST Parameters . . . 34
2.5.5 Interface Conservation (IC) Scores . . . 34
2.6 Acknowledgements . . . 35
CHAPTER 3. HomPPI: A class of Sequence Homology Based Protein-Protein Interface Prediction Methods . . . 39
3.1 Introduction . . . 40
3.2 Results . . . 43
3.2.1 HomPPI - Homologous Sequence-Based Protein-Protein Interface Prediction . . . 43
3.2.2 Performance Evaluation of HomPPI Methods . . . 44
3.3 Discussion . . . 52
3.3.1 Performance of HomPPI Compared with Published Methods . . . 52
3.3.2 Prediction Coverage of HomPPI Methods . . . 54
3.3.3 Parameters for HomPPI Can Be Relaxed for Obligate Interactions . . . 55
3.3.4 Prediction of Binding Partners vs. Prediction of Interface Residues . . . 55
3.3.5 Using Interface Predictions to Steer Docking and to Rank Docked Conformations . . . 56
3.4 Conclusions . . . 57
3.5 Methods . . . 58
3.5.1 Datasets . . . 58
3.5.2 Interface Definition . . . 59
3.5.3 Mapping Interfaces in Structures to Sequences . . . 60
3.5.4 NCBI BLAST Parameters . . . 60

3.5.5 Performance Evaluation . . . 60
3.5.6 NPS-HomPPI . . . 62
3.5.7 PS-HomPPI . . . 65
3.6 Availability . . . 67
3.7 Acknowledgements . . . 67

CHAPTER 4. DockRank: Ranking Docked Models Using Partner-Specific Sequence Homology Based Protein Interface Predictions . . . 68

4.1 Introduction . . . 69
4.2 Results . . . 73
4.2.1 DockRank Outperforms Energy-based Scoring Functions . . . 73
4.2.2 Partner-specific Interface Prediction Improves Ranking . . . 73
4.2.3 DockRank Has Lower I-RMSDs of Top Models . . . 75
4.2.4 DockRank Improves ClusPro Rankings . . . 76
4.3 Discussion . . . 81
4.4 Materials and Methods . . . 86
4.4.1 Decoy sets . . . 86
4.4.2 PS-HomPPI (Partner-Specific Homology based Protein-Protein Interface predictor) . . . 87
4.4.3 Databases Used by PS-HomPPI . . . 89
4.4.4 DockRank: Our Scoring Function for Ranking Docked Conformations . . . 90
4.4.5 Evaluation of Scoring Functions . . . 92

CHAPTER 5. Conclusions and Future Work . . . 106

5.1 Conclusions . . . 107
5.1.1 Protein-Protein Interface Positions Are Highly Conserved in Sequence Alignments . . . 107
5.1.2 A Family of Sequence Homology based Protein-Protein Interface Predictors . . . 108

5.1.3 Partner-Specific Sequence Homology based Interface Prediction Significantly Improves The Ranking of Docked Conformations 109
5.2 Directions for Future Research 110
5.2.1 Improve the Prediction Coverage of HomPPI family 110
5.2.2 Incorporate More Binding Partners Into Predictions 111
5.2.3 Constraining Docking with Partner-Specific Predictions 111
APPENDIX A. THE EXPECTATION AND VARIANCE OF RANDOM SUCCESS RATE AND HIT RATE IN RANKING DOCKED MODELS IN CHAPTER 4 113
BIBLIOGRAPHY 116

LIST OF TABLES

Table 2.1 Variables, Parameter Estimates and Significance Values for the Linear Model for NPS-Interface Conservation. 22

Table 2.2 Variables, Parameter Estimates and Significance Values for the Linear Model for PS-Interface Conservation. 27

Table 2.3 The Proportion of Interface Residues in Datasets Used in Interface Conservation Analysis. 33

Table 2.4 BLAST Substitution Matrices and Gap Costs used for BLASTP searches. 34

Table 3.1 Interface Residue Prediction Performance of NPS-HomPPI on Benchmark180. 46

Table 3.2 Prediction Performance of NPS-HomPPI using Homologs from the Safe, Twilight, Dark Zones. 47

Table 3.3 The Proportion of Interface Residues in Datasets Used in The Evaluation of HomPPI. 59

Table 3.4 Boundaries of Safe, Twilight and Dark Zones used by NPS-HomPPI. . 66

Table 3.5 Boundaries of Safe, Twilight and Dark Zones used by PS-HomPPI. . . 66

Table 4.1 Interface prediction performance of PS-HomPPI on BM3 dataset with three different prediction confidence zones on three levels of conformational change upon binding. Only the interfaces between the receptors and ligands are predicted and used by DockRank in ranking docked models. During the evaluation, we consider each partner-specific predicted receptor-ligand interface as one prediction. For example, for complex A-BC with one receptor A and two ligand chains B and C, we consider four predictions: A|A-B, A|A-C, B|B-A, C|C-A, where A|A-B means the interface of A with its binding partner B. There are totally 379 partner-specific receptor-ligand interfaces, 65.4% of which can be predicted by PS-HomPPI using homo-interologs in Safe, Twilight or Dark Zones. The performance of PS-HomPPI is not affected by the conformational changes upon binding; instead it is clearly correlated with the prediction confidence zones. The most reliable interface predictions are obtained in Safe Zone. Some residues of some proteins may not have interface predictions from PS-HomPPI. These residues are not considered in the evaluation, because these residues are not used by DockRank in ranking docked models. 83

Table 4.2 Average Hit Rates in top 1000 models of different scoring functions on ZDock3-BM3 decoy set in different interface prediction confidence zones. Cases with more than one receptor-ligand chain pairs may have predicted interface from different confidence zones. Cases with solo confidence zones are studied here. 66 cases that have only Safe Zone interface predictions, of which 53 cases have at least one hit. 16 cases have only Twilight interface predictions, of which 12 cases have at least one hit. 3 cases have only Dark interface predictions, of which 2 cases have at least one hit. DockRank has the most reliable performance in terms of Hit Rates for cases with interface prediction confidence in Safe Zone. The average Hit Rate of DockRank Cases with Twilight Zone confidence declined but still outperformed other scoring functions. 85

LIST OF FIGURES

Figure 2.1 Principal Component Analysis of Interface Conservation Scores and Sequence Alignment Statistics. Proteins in the Nr6505 and their homologs were analyzed. The data points in the biplot correspond to the projection of a 6-dimensional vector representing each protein-homolog onto a 2-dimensional space defined by the first and second principal components (PC1 and PC2). Blue lines with red circles at their tips represent the axes of the original 6-dimensional space for the 6 variables used in PCA analysis: -log(EVal), Identity Score, Positive Score, log(LAL), alignment length fractions (LAL/query length) and (LAL/homolog length). Each data point is colored according to its computed interface conservation (IC) score, with higher IC scores (red/orange) indicating higher interface conservation and lower IC scores (blue/green) indicating lower interface conservation (see text for details). The large gray arrow indicates the direction of increasing degree of interface conservation, from Dark to Twilight to Safe Zone. 18

Figure 2.2 EVal is a Good Indicator of Interface Conservation. Each blue dot in the scatter plot corresponds to a query-homolog pair. Red dots are the median values of IC scores for a specific EVal. To avoid log(0), we set log(EVal) = -450 when EVal = 0. 20

Figure 2.3 Interface Conservation (IC) Scores are Linearly Related to the Log of the Local Alignment Score (LAL). Each blue dot in the scatter plot corresponds to a query-homolog pair. Red dots are the median values of IC scores for a specific LAL. Note the trend of increasing median IC score with log(LAL) observed with the transitions from Dark to Twilight to Safe Zone. 21

Figure 2.4 A High BLAST Positive Score Reflects NPS-Interface Conservation. Each blue dot in the scatter plot corresponds to a query-homolog pair. Red dots are the median values of IC scores for a specific Positive Score. Note that medians of IC scores are near zero until Positive Scores become larger than 90 %. 23

Figure 2.5 Principal Component Analysis of Interface Conservation Scores and Sequence Alignment Statistics for Obligate versus Transient Complexes. The PCA biplots shown are for (a) proteins from obligate complexes and (b) proteins from transient complexes. See Figure 2.1 legend for additional details. 36

Figure 2.6 Comparison of Interface Conservation in Proteins from Obligate versus Transient Complexes. Proteins from obligate complexes are analyzed in a, c and e (left panels); proteins from transient complexes are analyzed in b, d, and e (right panels). Scatter plots show IC scores plotted as a function of: (a, b) log (local alignment length); (c, d) log (EVal); and (e, f) Positive Score. Red dots are median values of IC scores for a specific value on the x-axis. 37

Figure 2.7 PS-Interface Conservation in Transient Complexes. Homo-interologs corresponding to complexes in the Trans135 dataset were analyzed (see text for details). (a-c) Scatter plots show IC scores (blue dots) plotted as a function of: (a) log EVal; (b) Positive Score; (c) log LAL. Red dots are median values of IC scores for a specific value on the x-axis. (d) Scatter plot of Positive Score as a function of FracA × FracA'. Each data point (in d only) is colored using according to its IC score. 38

Figure 3.1 Performance of NPS-HomPPI Compared with Web-based PPI Servers. Performance was evaluated on four different protein complex types from Benchmark180: (a) Enzyme-inhibitors, transient. (b) Non-enzyme-inhibitors (NEIT), transient. (c) Hetero-dimers, obligate. (d) Homo-dimers, obligate. Servers compared were: NPS-HomPPI: red circles; Meta-PPISP: green squares; Cons-PPISP: blue triangles; Promate: brown stars; PIER: purple stars; PSIVER: yellow stars. 50

Figure 3.2 Performance of NPS-HomPPI Compared with ANCHOR in Predicting Interface Residues in Disordered Proteins. Two datasets of disordered proteins were used: (a) S1: short disordered proteins. (b) S2: long disordered proteins. NPS-HomPPI: red circles; ANCHOR: green squares. 53

Figure 3.3 Performance Comparison of PS-HomPPI and NPS-HomPPI. Only proteins for which predictions could be generated by both PS-HomPPI and NPS-HomPPI (139 out of 270 chains from Trans135) were used in this evaluation. The lower (Q1), middle (Q2) and upper (Q3) quartiles of each box are 25th, 50th and 75th percentile. Interquartile range IQR is Q3-Q1. Any data value that lies more than 1.5 × IQR lower than the first quartile or 1.5 × IQR higher than the third quartile is considered an outlier, which is labelled with a red cross. The whiskers extend to the largest and smallest value that is not an outlier. Averages are marked by green dots. 53

Figure 3.4 An Example of Interface Residue Prediction using NPS-HomPPI. The sequence of the query protein 1byf chain A is BLASTed against nr_pdbaa_s2c database. In this case, 3 sequences meet the thresholds set by NPS-HomPPI for "close homolog" in Safe Zone or Twilight Zone defined in Table 3.4 . If there are more than K = 10 homologs met the zone thresholds in Table 3.4, regression equation 2.2.1.5 is used to determine the nearest K homologs for final prediction. For each position in the alignment, an amino acid residue in the query sequence is predicted to be an interface residue if the majority of the amino acid residues in the alignment are interface residues. Otherwise, it is predicted to be a non-interface residue. Interface residues are denoted by red 1's; Non-interface residues are denoted by black 0's. Question marks denote residues for which coordinates are missing from PDB files. 63

Figure 4.1 The distribution of the number of actual hits in each case of ZDock3-BM3 decoy set. Docked model with I-RMSD $\leq$ 2.5 angstroms is considered a hit. 54,000 docked models are generated by ZDock 3.0 for each case. The 69 cases that have at least one hit generated by ZDock 3.0 and can be ranked by DockRank using homo-interologs (homologous interacting proteins) in Safe, Twilight or Dark Zone are shown here. 74

Figure 4.2 The Success Rates of DockRank and other scoring functions on ZDock3-BM3 decoy set. The Success Rates of DockRank (red solid line) are compared with those of two energy-based scoring functions: IRAD (black) and ZRank (yellow), and with three other rankings of docked models by combining our scoring fucntion with the predicted NPS-interface from: NPS-HomPPI (purple), PRISE (light green solid), and meta-PPISP (blue). NPS-HomPPI, PRISE and meta-PPISP are NPS-interface predictors, which do not consider the information of the query protein's binding partner when predicting interface residues. DockRank consistently has significantly higher Success Rates than IRAD, ZRank, NPS-HomPPI, PRISE, meta-PPISP. The Success Rates of the ranking of docked models by PS-actual interface residues (dark green dash line) combined with our scoring function and the expectations of the Success Rates of a random pick (dash red line) are plotted to defined the upper and lower bound. Studied here are 69 out of 97 cases that have at least one hit (I-RMSD $\leq$ 2.5 angstroms) and can be ranked by DockRank using homo-interologs in Safe, Twilight and Dark Zone. 93

Figure 4.3 The Hit Rates of DockRank and other scoring functions on ZDock3-BM3 decoy set. The hit rates of top 1000 ranked models selected by DockRank (red) are compared with those of two energy-based scoring functions: IRAD (black) and ZRank (yellow), and with three other rankings of docked models by combining our scoring function with the predicted NPS-interface from: NPS-HomPPI (purple), PRISE (light green solid), and meta-PPISP (brown). NPS-HomPPI, PRISE and meta-PPISP are NPS-interface predictors, which do not consider the information of the query protein's binding partner when predicting interface residues. DockRank consistently has higher Hit Rates than other scoring functions. The Hit Rates of the ranking of docked models by PS-actual interface residues combined with our scoring function (dark green stem dash line) and the expectations of Hit Rates of a random pick (dash red stem dash line) are plotted to define the upper and lower bound. Studied here are 69 out of 97 cases that have at least one hit (I-RMSD $\leq$ 2.5 angstroms) and can be ranked by DockRank using homo-interologs in Safe, Twilight and Dark Zone. 94

Figure 4.4 Pair-wise comparisons of the mean of I-RMSDs of top 1 models selected by different docking scoring methods on ZDock3-BM3 decoy set using the Nemenyi test. Methods that are not significantly different (at significance level) are grouped together (via connecting lines). The average "rank" of each method over docking cases is shown in the table (and also on the x-axis of the plot). The mean of I-RMSDs of top 1 models selected by DockRank is significantly smaller than those selected by IRAD, ZRank, NPS-HomPPI, PRISE, and meta-PPISP. The mean of I-RMSDs of top 1 models selected by DockRank is not significantly different from actual partner-specific interface-based method (PS-Act Int). 95

Figure 4.5 The distribution of the number of actual hits in each case of Cluspro decoy set. Docked model with L-RMSD $\leq$ 10 angstroms is considered a hit. 54,000 docked models are generated by ZDock 3.0 for each case. The 32 cases that have at least one hit generated by ClusPro 2.0 and can be ranked by DockRank using homo-interologs (homologous interacting proteins) in Safe, Twilight or Dark Zone are shown here. 96

Figure 4.6 The Success Rates of DockRank and ClusPro scoring functions on ClusPro2-BM3 decoy set. ClusPro scoring functions (Default Cluser-size based, Center Energy-based, Lowest Energy-based) were applied on the original docked models generated by ClusPro's underlying docking program PIPER. DockRank was applied on the filtered docked models by ClusPro scoring functions. The Success Rates of three ClusPro scoring functions are significantly improved. DockRank (PS-HomPPI interface prediction based) is able to select at least one hit in top 8 ranked models for more than 95 % cases tested here. The Success Rate of PS-actual interface based ranking and the expectation of Success Rate of random rankings are also plotted to show the upper and lower bound of Success Rates. 32 cases that have at least one hit and whose interface can be predicted by PS-HomPPI using Safe, Twilight, or Dark Zone homo-interologs are studied here. Case 1PPE has only 9 models, so the Success Rates of up to top 9 rankings are studied here. 97

Figure 4.7 The Hit Rates in top 1 docked models ranked by DockRank and ClusPro scoring functions on ClusPro2-BM3 decoy set. The average Hit Rates of scoring functions over cases are shown in the table. DockRank improved the average Hit Rates of top 1 docked models from 0.21 of ClusPro, 0.28 of Lowest Energy, and 0.14 of Center Energy, to 0.40. The Hit Rates of PS-Actual interface-based ranking and the expectation of Hit Rates of random rankings (see Appendix for the derivation of the expectation and variance of the random Hit Rate) are calculated to define the upper and lower bound. 32 cases that have at least one hit and whose interacting residues can be predicted by PS-HomPPI using homo-interologs in Safe, Twilight, or Dark Zone are studied here. 98

Figure 4.8 Pair-wise comparisons of different docking scoring methods on ClusPro2-BM3 decoys using the Nemenyi test. Methods that are not significantly different (at significance level alpha = 0.05) are grouped together (via connecting lines). The average "rank" of each method over docking cases is shown in the table (and also on the x-axis of the plot). Pairwise Nemenyi test shows that the average L-RMSDs of top models selected by DockRank are significantly smaller than those selected by ClusPro Center Energies. However, the average L-RMSDs of top models selected by ClusPro, Lowest Energy and DockRank are not significantly different, which indicates that DockRank has limited improvement on top 1 model in term of L-RMSDs when applied to the filtered docked models by ClusPro scoring functions under the definition of a hit as a docked decoy with L-RMSD $\leq$ 10 angstroms. 32 cases that have at least one hit and whose interface can be predicted by PS-HomPPI using Safe, Twilight, or Dark Zone homo-interologs are studied here. 99

Figure 4.9 The difference between the weighted averages L-RMSDs of top models between DockRank and ClusPro Rank on each case of ClusPro2-BM3 decoy set. L-RMSD of each top model is weighted by its ranks. 56 cases with at least one docked model with L-RMSD $\leq$ 20 angstroms and can be ranked by DockRank using homo-interologs in Safe, Twilight or Dark zones are studied here. A positive dot means the top models ranked by DockRank for the specific case have a lower weighted L-RMSD than those ranked by ClusPro. For 40 out of 56 (71.4%) cases, top models ranked by DockRank have lower weighted L-RMSD than ClusPro. . . 100

Figure 4.10 The comparison of top 1 models ranked by ClusPro and DockRank for case 1RLB. The red cartoon is the receptor in the top 1 docked model (the bound and docked receptors are superimposed, and bound receptor is not shown here). The ligand of the top 1 model ranked by ClusPro default cluster-size based method (blue ribbon in the left panel) is near the bound ligand position (pink mesh), however, the ligand of top 1 model selected by DockRank (white ribbon in the right panel) is totally wrong and is on the opposite side of bound ligand position (pink mesh) relative to the receptor (red ribbon). Note that the structure of the receptor is symmetric. So the natural question is that whether it is possible that the ligand might be able to bind on both sides of the symmetric receptor instead of on only one side? 101

Figure 4.11 The top 1 model ranked by DockRank and two bound (native) ligands of case 1RLB. PDB entry 1RLB in fact has two bound ligands - chain E and F (purple ribbons). Chain E is included in BM3 dataset (purple ribbon in the left top corner, also shown as mesh in Figure) but chain F (purple ribbon on the lower right side) is arbitrarily left out of BM3 dataset. The ligand of the top 1 model selected by DockRank (blue ribbon) is right beside the left-out bound ligand. 102

Figure 4.12 The top 5 models ranked by DockRank and ClusPro for case 1RLB. Both bound ligands (pink mesh) in the PDB entry 1RLB are shown here. DockRank (right panel) is able to give top ranks to the models with ligands that are near the native ligand positions (pink mesh) on both binding sides of the receptor (red ribbon). However, ClusPro (left panel) gives top ranks to not only the models with ligands on the two binding sides of the receptor, but also models with irrelevant ligand positions. 103

Figure 4.13 Success Rates in top 1-1000 models of different scoring functions on ZDock3-BM3 decoy set in different interface prediction confidence zones. Cases with more than one receptor-ligand chain pairs may have predicted interface from different confidence zones. Cases with solo confidence zones are studied here. 66 cases have only Safe Zone interface predictions, of which 53 cases have at least one hit. 16 cases have only Twilight interface predictions, of which 12 cases have at least one hit. 3 cases have only Dark interface predictions, of which 2 cases have at least one hit. DockRank (red solid line) achieves the most reliable performance in terms of Success Rates on cases with interface prediction confidence in Safe Zone (right panel). The Success Rates of DockRank on cases with Twilight zone prediction confidence declines, but are still consistently higher other scoring functions from top 1 to 1000 models. For the two cases in Dark Zone, DockRank was able to rank a hit to top 100 for one case, but could not find a hit for another case in top 1000 models. 54,000 models for each case were generated by ZDock 3.0. 104

Figure 4.14 PS-HomPPI: Partner-specific sequence homology based protein-protein interface predictor. PS-HomPPI has two components: PS-interface conservation analysis and PS-interface prediction. PS-interface conservation analysis (shown as the PCA biplot on the left) is based on a dataset of 135 transient dimers with experimentally determined interface residues. For each dimer A-B, sequence homologs with known interfaces are retrieved. Each dot in the PCA biplot represents two sequence alignments: query A - its sequence homolog A', and query's partner B - its sequence homolog B'. Complex A'-B' is called a homo-interolog of A-B. 9 sequence alignment measures (blue lines with a red circle at the end) are calculated. An interface conservation score (IC-score) is calculated based on the similarity of the interfaces of A-B and those of A'-B'. The higher IC-score the more similar the interfaces of A-B and A'-B' are. IC-scores are represented using different colors: red for a high degree of interface conservation, and blue for low conservation. The original interface conservation space with 9 alignment measures and 1 IC-score was mapped to two dimensional PC1-PC2 space, where the relation of IC-score and the sequence alignment measures can be easily observed. Based on the color change (IC-score), three interface conservation zones are identified: Safe Zone for high level of interface conservation, Twilight Zone for medium level, and Dark Zone for low level. A regression model of IC-score with the sequence alignment measures is built. When making an interface prediction of a pair of proteins, a list of homo-interologs with known interfaces is searched. Sequence alignment measures are calculated for each query - homo-interolog. The regression model is used to rank the homo-interologs. Top K homo-interologs are used to make interface transfer, and their conservation zone provides a prediction confidence. 105

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this dissertation.

First I want to thank my major professor Dr. Vasant Honavar for his excellent guidance, support and high standard in research, which spurred me to seek deeper understanding research questions and to try new angles of solutions. It is a great present for me that Dr. Honavar gave me the freedom to explore research topics that I am interested in, and made himself available whenever I need his suggestions and help. His generous financial support has provided me various training opportunities, and his firm guidance and insight has led me through every stage of my research.

I want to express my deepest respect and thanks to my co-major professor Dr. Drena Dobbs for her understanding, patience, inspiring discussions, encouragements and loving guidance. In the early stage of my research she inspired me with research ideas, and walked me through the details of how to conduct research and write research papers. When my steps stumbled she was always there to listen, comfort and help. Her truly scientist intuition and unselfish devotion to students deeply touched me.

I want to thank my committee members: Dr. Robert Jernigan for his insightful questions and comments, which inspired my multiple binding partners discussion in future work; and Dr. Guang Song and Dr. Peng Liu for their encouragement and helpful discussions.

I would like to thank our program coordinator Trish, who is a loving mother to all our BCBers, who tries to protect us from our worries, listens patiently to what we think, and strives to make our program united and strong.

I owe lots of thanks my labmates, without whose collaboration, discussion and suggestions this dissertation could not be finished. I want to thank Rafael for his helpful discussions, his useful database and his processing the large amount of docked conformations; Yasser for his

guidance and careful editing of my paper drafts; Fadi for his patient help with Linux commands and perl; Rasna for her proofreading and helping me with other related projects; and Usha for providing me the benchmarked Cluspro docked decoys. I cherish our friendship developed during these years when we worked hard as a team.

I want to sincerely thank Dr. Balaji Narasimhan and Latrisha Peterson for introducing me to the combinatorial immunology, for their help during our collaboration on the vaccine delivery system, which is not part of this dissertation but has broadened my view of computational biology.

I thank my instructors in the Departments of Biology, Statistics, and Computer Science, who equipped me with, and made me eager for more, valuable knowledge, which is essential to this dissertation and is the foundation of my future research and exploration.

I could not imagine how I could survive the stressful years and challenging moments without my friends' accompany and constant support. The best and worst moments of my studies here have been shared with many people, especially with Monica Richards, who revolutionized our apartment with the first set of furniture, rescued me from many times of dead batteries, enjoys me with all my "mysterious" love for animals, and loves me with all my absent-mindedness. I benefit tremendously from her belief of life and her positive spirit. I thank my roommates and classmates, who are very supportive while fighting on the same road. I thank ISU Horsemen's Association members for the happy and unforgettable memories.

ABSTRACT

Protein-protein interactions play a central role in the formation of protein complexes and the biological pathways that orchestrate virtually all cellular processes. Three dimensional structures of a complex formed by a protein with one or more of its interaction partners provide useful information regarding the specific amino acid residues that make up the interface between proteins. The emergence of high throughput techniques such as Yeast 2 Hybrid (Y2H) assays has made it possible to identify putative interactions between thousands of proteins (but not the interfaces that form the structural basis of interactions or the structures of protein complexes that result from such interactions). Reliable identification of the specific amino acid residues that form the interface of a protein with one or more other proteins is critical for understanding the structural and physico-chemical basis of protein interactions and their role in key cellular processes, for predicting protein complexes, for validating protein interactions predicted by high throughput methods, for ranking conformations of protein complexes generated by docking, and for identifying and prioritizing drug targets in computational drug design.

However, given the high cost of experimental determination of the structures of protein complexes, there is an urgent need for reliable and fast computational methods for identifying interface residues and/or predicting the structure of a complex formed by a protein of interest with its interaction partners. Given the large and growing gap between the number of known protein sequences and the number of experimentally determined structures, sequence-based methods for predicting protein-protein interfaces are of particular interest. Against this background, we develop HomPPI (http://homppi.cs.iastate.edu/), a class of sequence homology based approaches to protein interface prediction. We present two variants of HomPPI: (i) NPS-HomPPI (non-partner-specific HomPPI), which can be used to predict interface residues of a query protein in the absence of knowledge of the interaction partner. NPS-HomPPI is based on the results of a systematic analysis of the conditions under which interface residues of

a query protein are conserved among its sequence homologs (and hence can be inferred from the known interface residues in proteins that are sequence homologs of the query protein). Our experiments suggest that when sequence homologs of the query protein can be reliably identified, NPS-HomPPI is competitive with several state-of-the-art interface prediction servers including those that exploit the structure of the query proteins. (ii) PS-HomPPI (partner-specific HomPPI), which can be used to predict the interface residues of a query protein with a specific target protein. PS-HomPPI is based on a systematic analysis of the conditions under which the interface residues that make up the interface between a query protein and its interaction partner are preserved among their homo-interologs, i.e., complexes formed by their respective sequence homologs. To the best of our knowledge, with the exception of protein-protein docking (which is computationally much more expensive than PS-HomPPI), PS-HomPPI is one of the first partner-specific protein-protein interface predictors. Our experiments with PS-HomPPI show that when homo-interologs of a query protein and its putative interaction partner can be reliably identified, the interface predictions generated by PS-HomPPI are significantly more reliable than those generated by NPS-HomPPI.

Protein-Protein Docking offers a powerful approach to computational determination of the 3-dimensional conformation of protein complexes and protein-protein interfaces. However, the reliability of conformations produced by docking is limited by the efficacy of the scoring functions used to select a few near-native conformations from among tens of thousands of possible conformations, generated by docking programs. Against this background, we introduce DockRank, a novel approach to rank docked conformations based on the degree to which the interface residues inferred from the docked conformation match the interface residues predicted by a partner-specific sequence homology based interface predictor PS-HomPPI. We compare, on a data set of 69 docked cases with 54,000 decoys per case, the ranking of conformations produced using DockRank's interface similarity scoring function applied to predicted interface residues obtained from four protein interface predictors: PS-HomPPI, and three NPS interface predictors NPS-HomPPI, PRISE, and meta-PPISP, with the rankings produced by two state-of-the-art energy-based scoring functions ZRank and IRAD. Our results show that DockRank significantly outperforms these ranking methods. Our results that NPS interface predictors

(homology based and machine learning-based methods) failed to select near-native conformations that are superior to those selected by DockRank (partner-specific interface prediction based), highlight the importance of the knowledge of the binding partners in using predicted interfaces to rank docked models. The application of DockRank, as a third-party scoring function without access to all the original docked models, for improving ClusPro results on two benchmark data sets of 32 and 56 test cases shows the viability of combining our scoring function with existing docking software. An online implementation of DockRank is available at http://einstein.cs.iastate.edu/DockRank/.

CHAPTER 1. Overview

The word *protein* is derived from the Greek word *proteios*, meaning "primary". Proteins are the principal catalytic agents, structural elements, signal transmitters, transporters and molecular machines in cells. Protein-protein interactions play a pivotal role in carrying out virtually all major cellular processes, such as immune responses, cell cycle control, signal transduction, DNA replication, transcription and translation. Proteins do not function alone. Proteins realize their functions by (i) interacting with proteins to serve as molecular messengers, as guards in immune system, or as building blocks; (ii) interacting with DNA to express or replicate the genetic code; (iii) interacting with RNA to regulate the synthesis of proteins or to modify pre-mRNA; and (iv) interacting with small molecules to strengthen inter-cellular communication signals. In this study, we focus on the study and prediction of protein-protein interaction sites (interfaces) - the regions where two proteins interact. The interface of a protein is composed of a set of residues of this protein that form non-covalent contacts with the atoms or residues of other proteins. The distortion of protein-protein interfaces often lead to various diseases. Characterization of protein interfaces is crucial for understanding the molecular, structural, and biophysical bases of protein interactions, for elucidating the mechanisms that underlie signal transduction cascades, and their physiological role in networks and pathways involved in biological processes, and for identifying promising drug targets for the therapeutic interventions [139].

The possible number of protein pairs is huge and even the highest throughput methods are not able to provide meaningful information for such huge numbers – for example, if an organism has 10,000 genes then there would be 100 million pair-wise interactions that need investigating. While there are advances in high throughput methods such as yeast two-hybrid (Y2H) experiments and protein binding microarrays, and increasing numbers of solved high resolution protein structures, nonetheless these types of information are available for only a

small fraction of the interacting proteins [50]. This lack of information about the protein-protein structures places a significant barrier to progress in understanding the functioning of proteins as well as comprehending the topology and complexity of cellular protein interaction networks. The research described in this dissertation aims to overcome this barrier by developing novel, accurate, and efficient computational approaches to predict the likely protein-protein interfaces and using the predicted interfaces to select near-native interaction conformations out of the huge numbers of docked candidate conformations.

Before introducing our work, we briefly review the experimental and computational work in protein-protein interface identification, analysis and predictions.

1.1 Experimental Methods to Identify Protein-Protein Interface Residues

Several different genetic, biochemical, and biophysical methods have been used to identify and characterize protein-protein interacting regions (interfaces). Widely used techniques include

High resolutions techniques:

- X-ray crystallography
- Nuclear Magnetic Resonance (NMR) spectroscopy
- Alanine scanning

Low resolution techniques (Mass spectrometry-based approaches):

- Chemical cross-linking
- Hydrogen/deuterium (H/D) exchange

Both X-ray crystallography and NMR spectroscopy can provide atomic level information of protein structures hence the interacting sites (interfaces). Most resolved structures using X-ray crystallography and NMR are deposited in Protein Data Bank (PDB) [14].

X-ray crystallography [46] determines the special arrangement of atoms in a crystallized protein/protein complex by shedding X-rays beam on the crystal and studying the diffraction pattern caused by the electrons of atoms of the crystal. X-ray crystallography provides high

resolution information of protein structures, which allows direct visualization of the interaction of proteins and their binding sites. However, obtaining crystals of proteins is often the most difficult step, and limits the applications of X-ray crystallography in large scale characterization of protein structures.

NMR Spectroscopy [23, 157] is based on the property of the nuclei of atoms being able to absorb and re-emit electromagnetic radiation when placed in external magnetic field. Different nuclei have different resonance frequency, which provides researchers the information of protein structures. NMR can also study weak protein–protein interaction and protein dynamics during molecular recognition. NMR spectroscopy is highly suited to investigate molecular interactions at a close physiological condition and is particularly suited for the study of low-affinity, transient complexes. NMR is limited by size constraints, and the method is best applied to proteins smaller than 35 kDa.

Alanine scanning [96] determines the interface residues by replacing residues of a protein with alanine and studying the change of binding affinity. The limitation of this method is that without prior knowledge of approximate interface location, doing an exhaustive alanine scanning on all the combination of residues of potential protein pairs can be labor-intensive and slow.

Mass spectrometry (MS) [7] is a key technology in proteomics and was recognized in 2002 Nobel Prize in chemistry (jointly with NMR). MS determines the components in a sample by ionizing the protein/peptide sample, measuring the number of ions at each mass-to-charge ratio, and comparing the sample spectrum with calculated databases of known proteins/peptides. MS has been used to determine the sequences of proteins, to detect protein interaction partners, and to analyze protein post-translational modifications. Furthermore, the combinations of MS with other techniques, such as chemical crosslinking [120], H/D exchange [56], are used to characterize protein interacting regions at low resolutions.

Chemical crosslinking has been used to chemically joining two or more amino acids in the proximity of interacting proteins by a covalent bond. The comparisons of MS spectral profiles of the cross-linked protein-protein complex with those of individual component proteins have allowed the detection of the interacting regions. An advantage of using crosslinking is that

it can be used to stabilize the low binding affinity transient interacting complexes before MS characterization.

H/D exchange is an isotope tagging technique. This method makes use of the facts that (i) exchange of protons between solvent-exposed amide hydrogens and the solvent occurs constantly while the interface regions are less likely to exchange protons; (ii) pairs of isotopes (hydrogen/deuterium) can be differentiated by MS owing to their mass difference; (iii) the ratio of signal intensities for such analyte pairs accurately indicates the abundance ratio for the two analytes. The interacting regions can be revealed by the comparisons between the rates of deuterium exchange of peptides from protein-protein complex and from individual component proteins.

These experiments are extremely valuable and have contributed greatly to our knowledge of protein-protein interfaces. However, the major bottleneck in these techniques remains the efficient purification of protein samples, which makes these experiments labor-intensive, time-consuming, or restricted by various technical difficulties which prevent them from being applied to large scale characterization of protein structures. Therefore, reliable computational approaches to identify interface residues are especially needed.

1.2 Protein-Protein Interface Features

Several research groups have explored the utility of various protein sequence and structural features [71, 73, 72, 79, 11, 17, 24, 36, 84, 118, 123, 58] in predicting protein-protein interface residues. Such features include, but not limited to, amino acid propensities of interfaces [58], secondary structure of interfaces [61], accessible surface area, hydrophobicity, and protrusion [73]. Ofran and Rost [103] showed that interface residues tend to cluster together. Jones and Thornton [73] studied six parameters of interfaces using surface patch analysis; however none of the parameters were definitive. Yan et al. [148] found that interfaces favor hydrophobic residues (particularly aromatic residues), the opposite charge pairs, hydrophobic pairs and Pro-Trp pair. Tuncbag et al. [124] found that residue occlusion from solvent in the complexes and pairwise potentials were important discriminative features in protein interface hot spot [1]

[1] Hot spots are interfial residues that contribute more to the binding free energy.

prediction. Chakrabarti and Janin [69, 24] found that the core of an interface patch has an amino acid composition that differs from that of the rim. Mintseris and Weng [93] studied the relative frequencies of the different atom–atom contacts, and used it to distinguish homodimers and crystal contacts and to separate transient complexes from permanent oligomeric ones.

Discriminative features can be classified into two types: structure-based features and sequence-based features. Structure-based features are extracted from the 3D conformations of proteins, such as solvent accessible area, residue/atom propensity of surface patches and geometric attributes of protein surfaces. Commonly used sequence-based features are hydrophobicity, residue propensity in sequence windows, PSSM generated from multiple sequence alignment (MSA), predicted solvent accessibility, and predicted structural features.

These studies show that protein-protein interacting residues indeed share some features that are different from non-interface residues. However, no single feature can reliably discriminate interface residues from non-interface residues. Hence, it is of interest to explore computational interface predictors that combine these features as needed.

1.3 Computational Methods to Predict Protein-Protein Interface Residues

A large number of in silico approaches to protein-protein interface prediction have been explored in the literature in the past decade (reviewed in [37, 52, 153]). Based on whether or not they require the knowledge of 3D conformation of input proteins, the protein-protein interface predictors can be classified into sequence-based classifiers, structure-based classifiers and hybrid classifiers. The majority of predictors are structure-based or hybrid methods. However, structure-based methods have several critical disadvantages: 1) limited applications. Structure-based predictors require the knowledge of protein structures. Due to the difficulty of experimental characterization of protein structures, the vast majority of proteins, especially for transient binding proteins and intrinsically disordered proteins, do not have experimentally determined 3D structures; 2) limited robustness for interactions subject to substantial conformational changes. Structural features may not hold unchanged before and after the formation of protein-protein complexes due to the conformational changes induced by interactions. Therefore, the development of sequence-based methods, which can reliably differentiate interface residues from non-

interacting ones without requiring the knowledge of 3D protein structures, is of great interest.

Despite the considerable efforts dedicated to the development of sophisticated and advanced protein-protein interface predictors, most of them ignored the fact that many proteins use different interacting regions to interact with different binding partners (partner-specificity). High degree of partner-specificity is especially true in the case of transient interactions [142]. Transient interactions provide a mechanism for the cell to quickly respond to extracellular stimuli, and are essential in the regulation of many disease-related pathways [101, 5]. The high degree of partner-specificity of transient interactions is appealing for the discovery and development of target-specific therapeutic inhibitors. Therefore, developing reliable partner-specific interface predictors is urgently needed.

1.4 Protein-Protein Docking

The 3D structures of complexes formed by interacting proteins are valuable sources of information needed to understand the structural basis of interactions and their role in complexes and pathways that orchestrate key cellular processes, to validate interactions determined using high throughput methods such as yeast-2-hybrid assays, and to identify and prioritize drug targets in computational drug design. Because of the expenses and efforts associated with X-ray crystallography or NMR experiments to determine 3D structures of protein complexes, protein-protein docking methods are often used to predict the 3D conformation of complexes formed by two or more interacting proteins and hence interfaces of the component proteins. Docking programs search through the conformation space to generate large numbers of candidate conformations, and rank the resulting conformations based on a criteria such as the energy of the conformation and structural or physico-chemical complementarity of the interface between the proteins that make up the complex.

Despite the promise of protein-protein docking in predicting 3D structures of interacting proteins, the following problems need to be solved to make its use feasible for large-scale applications: 1) Docking processes tend to be computationally expensive; 2) The existing docking programs require experimentally solved or computationally predicted structures of the component proteins; 3) It is challenging for existing docking programs to generate meaningful con-

formations for proteins subject to large conformational changes upon binding; 4) The goal of singling out near-native conformations from the vast number of candidate conformations within a reasonable computational time is far from being satisfactorily solved.

To lower the computational cost of docking programs and to generate more near-native conformations in the conformational sampling stage of a docking process, there is an increasing effort to constrain the docking process with experimentally determined or computationally predicted interaction sites [41, 38, 44, 132, 39, 81]. The success of these methods highlights the importance of the knowledge of interacting sites. As we discussed earlier, experimental identification of the interface residues is not a trivial task, therefore, reliable in silico identifications of interface residues are urgently needed.

Another direction for improving docking programs focuses on improving the reliability of scoring functions. Scoring functions reported in the literature can be broadly classified into four types: (1) geometric complementarity-based scoring functions, such as FFT-based methods [75] and geometric hashing [141, 47]; (2) energy-based scoring functions designed to approximate the binding free energy of protein-protein assemblies [133, 110, 59]; (3) Knowledge-based scoring functions (i.e. knowledge-based pairwise potentials [95, 85, 82], knowledge-based weighted correlations [63, 109], machine learning classifiers of native/non-native protein-protein assemblies [18, 87], and predicted interface-based scoring functions [99, 64]); (4) Hybrid functions that combine the scoring functions of the previous three types [88, 35, 76, 77]. Despite the large number of advanced and sophisticated scoring approaches that are currently used by docking programs, the goal of selecting near-native conformations from the large number of candidates is far from solved [62, 128].

We are particularly insterested in the scoring functions using predicted interface residues to rank docked conformations. This approach is based on the hypothesis that docked conformations with interacting sites that are highly similar to predicted interfaces are more likely to be near-native conformations. However, this type of approaches relies heavily on the reliability of interface predictions. Li and Kihara [81] ranked conformations using interfaces predicted by a start-of-the-art *non-partner-specific* interface prediction method - meta-PPISP [112]. They concluded that "Blind PPI site predictions cannot be used for improving docking prediction with

the post-filtering procedure. On average it will only deteriorate prediction accuracy." However, a natural question is whether or not the predicted partner-specific interfaces are reliable enough to rank docked conformations?

Against this background, we focus our efforts on the design of a fast and reliable *partner-specific* interface predictor, and on the incorporation of the predicted interfaces into a reliable and computational efficient scoring function for ranking docked conformations.

1.5 Interface Residue Conservation

Although homology based approaches have been successfully applied in many areas (such as protein structure predictions using homology modeling [57], the prediction of protein interaction partners [134, 150], and function annotation [86]), published studies disagree on whether or not protein-protein interfaces are more conserved than other surface residues or the rest of the protein sequences. Grishin and Phillips [60], after examining five enzyme families, concluded that the degree of conservation of interface residues is *same* as that of protein sequences as a whole. Caffrey et al. [22], based on their study of 64 protein-protein interacting chains found that interface residues are *slightly* more conserved than the rest of the protein surface residues. Valdar and Thornton [129] concluded that interface conservation of homodimers is *higher* than other surface residues after studying six homodimers families. Choi et al. [32] based on sequence conservation analysis of 2,646 protein interfaces concluded that protein interface residues are *more conserved* than other surface residues.

With the exception of the study conducted by Grishin and Phillips in 1994 [60] these studies focused on the comparison of interface residue conservation relative to other surface residues, and excluded the residues that are buried inside the protein surface. However, the determination of surface residues is not a trivial task if protein structures are not available. In light that the number of known protein sequences is much larger than the number of protein structures, it is of interest to study the conservation of interface residues compared with non-interface residues. Such analysis may benefit the development of reliable *sequence* homology based predictors of interface residues.

Another important fact is that although the partner-specific property of protein-protein

interactions has been long recognized, it was ignored by previous conservation studies. Many transient protein-protein bindings serve as molecular messengers, carrying out cell signaling among different parts of the cell. In order to carry out the complicated signaling, the catalytic activity must be highly regulated and the interaction sites can be highly partner-specific. Besides the factor of small datasets used in the previous studies, the controversy of previous interface conservation studies might be due to this property that transient binding proteins tend to use different interfaces when interacting different partners. Hence, a *partner-specific* analysis of interface conservation of transient binding proteins is of interest.

1.6 Hypotheses and Overview of This Work

Against this background, we formulate the following hypotheses:

1) The interface residues of interfactions are conserved among sequence homologs or sequence homo-interologs (homologous interacting proteins). Specifically, (i) the interfaces of transient interactions are highly conserved among sequence homo-interologs and are highly Partner-Specific (PS); and (ii) the interfaces of IDPs (Intrinsically Disordered Proteins) are highly conserved among sequence homologs and are non-partner-specific (NPS).

2) Given that transient protein interactions are highly partner-specific, it should be possible to improve the reliability of predictors of interface residues by taking information about binding partners into account.

3) We hypothesize that among the large number of docked conformations of protein-protein complexes, conformations with interacting sites that agree well with the reliably predicted interfaces are more likely to be near-native conformations. Therefore, predicted interface residues can be used to rank docked conformations.

These hypotheses are explored in greater details, which form the following four chapters of this dissertation.

Chapter 2: Conservation analysis of protein-protein interface residues. We introduce a novel measure of interface conservation that captures the degree to which interface residues in each protein are conserved among its sequence homologs. First, we describe the results

of our analysis of the interface conservation among homologous sequences using several large non-redundant datasets of protein-protein interfaces extracted from the Protein Data Bank (PDB), including datasets that allow us to compare "obligate" versus "transient" interfaces. To explore the extent to which interface conservation can be exploited in the prediction of interface residues, we systematically examined the relationship between interface conservation and multiple protein sequence similarity variables. In one set of experiments, we examined binding interfaces in homologous proteins *without* specifying a specific interaction partner (i.e., non-partner-specific, NPS-interfaces). The results of this analysis indicated that interfaces in obligate complexes are, in general, more highly conserved than those in transient complexes. In a complementary set of experiments, we examined interfaces in complexes between specific pairs of proteins (i.e., partner-specific, PS-interfaces). In contrast to the results for NPS-interfaces, by focusing on the interface of each query protein with a *specific* binding partner, we discovered a high degree of interface conservation in transient PS-interfaces. This analysis revealed that transient interfaces tend to be highly partner-specific.

Chapter 3: The design and evaluation of our Sequence Homology based Protein-Protein Interface Predictor - HomPPI. Based on the results of protein interface conservation analysis in Chapter 2 we propose HomPPI, a class of sequence homology based approaches to protein interface prediction. We present two variants of HomPPI: (i) NPS-HomPPI (non-partner-specific HomPPI), which can be used to predict interface residues of a query protein in the absence of knowledge of the interaction partner; and (ii) PS-HomPPI (partner-specific HomPPI), which can be used to predict the interface residues of a query protein with a specific target protein. The performance of both HomPPI methods was evaluated on several benchmark datasets, including a large non-redundant set of transient complexes. Due to the increasing importance of intrinsically disordered proteins in understanding molecular recognition mechanics and in rational drug design and discovery [53, 92, 121, 127], we also tested NPS-HomPPI on two datasets of intrinsically disordered proteins. The HomPPI web server is available at http://homppi.cs.iastate.edu/

Chapter 4: DockRank- ranking docked models using partner-specific sequence homology based protein interface prediction. We introduce a novel method DockRank for ranking docked conformations based on the degree of similarity between the interface residues of a docked

conformation formed by a receptor and a ligand with the set of interface residues predicted by our partner-specific interface predictor. DockRank utilizes PS-HomPPI (described in Chapter 3), a sequence homology based method that, given a query protein and its putative interacting partner (target protein), predicts the residues of the query protein that are likely to interact with the target protein. Our experiments with several benchmark decoy sets show that the quality of the ranking of docked conformations using DockRank is consistently superior to several state-of-the-art scoring functions. Our results also show that NPS (Non-Partner-Specific) interface predictors (homology-based and machine learning-based methods), cannot reliably select near-native conformations for transient interactions. Besides, our results on a set of ClusPro decoys show that DockRank, as a third party scoring function without accessing to all the docked conformations, significantly improved the rankings of pre-filtered top conformations ranked by ClusPro energy scoring functions in terms of Success Rate and Hit Rate. DockRank webserver is available at: http://einstein.cs.iastate.edu/DockRank/.

Chapter 5: Conclusions and future work. We review and discuss the major results and conclusions derived in the previous chapters. We discuss some directions for future work. Also we discuss the *limitations* of homology-based interface prediction methods, and propose several possible ways to improve the prediction coverage.

1.7 Main Contributions of This Work

The main contributions of this work are:

1. A novel partner-specific measure of conservation of residues at the interface between a pair of interacting proteins among their homo-interologs [142]. Analysis of conservation of residues in transient interfaces using this *partner-specific* measure shows that transient interfaces are in fact highly conserved. On the contrary, another set of conservation analysis on the same set of transient complexes using a *non-partner-specific* measure can only detect *weak* interface conservation. The contrast of the results obtained from these two sets of experiments highlight the importance of taking into account of the information of binding partners for the conservation analysis of transient interfaces, and it might explain the previous controversy on the degree of interface conservations.

2. Application of the preceding observation to develop PS-HomPPI, the first *sequence* based partner-specific interface predictor, which in our preliminary studies has been shown to provide among the most reliable predictors of interface residues of a hypothetical transient complex formed by a protein A with its putative interaction partner B whenever the homo-interologs of A-B can be reliably identified. PS-HomPPI, unlike most existing computational approaches to prediction of protein-protein interface residues (with the exception of protein-protein docking which is computationally far more expensive and hence infeasible to use on thousands of proteins and their interaction partners) can differentiate between the interfaces formed by a protein with different interacting partners.

3. A novel use of predicted interfaces to rank docked conformations based on the agreement between the interfaces of a docked conformation and the PS-HomPPI predicted *PS*-interfaces. Our results show that the PS-HomPPI predicted interfaces significantly and consistently improve the likelihood of singling out near-native conformations out of the large number of candidate conformations.

Collectively, our results suggest the possibility of developing purely sequence-based methods for reliably predicting protein-protein interfaces. Currently the coverage of interface prediction methods is limited to those cases where the interface of query proteins can be inferred from their sequence homologs with experimentally determined interfaces. However, this limitation may be partially alleviated with the expected increase in the number of structures in PDB. In addition, development of hybrid approaches that combine homology based predictors with their machine learning based trained counterparts can be expected to further improve the coverage of the resulting predictors. In the case of transient interactions, which tend to be partner-specific, the resulting improvement in the reliability and coverage of interface predictions can be expected to yield substantial improvements in ranking docked conformations. Additional advances can be expected in validating protein-protein interaction networks, in guiding the mutagenesis experiments with multi-faced hub proteins, and in developing target-specific therapeutic interventions.

CHAPTER 2. Conservation Analysis of Protein-Protein Interface Residues

A paper titled "HomPPI: a class of sequence homology based protein-protein interface prediction methods", BMC Bioinformatics 2011, 12:244

Li C. Xue, Drena Dobbs and Vasant Honavar

Abstract Although evolutionary conservation of protein-protein interfaces have been investigated on different datasets in the past decades, no agreed-upon conclusions were drawn on whether and to what degree interface residues are more conserved than the rest of the residues. The previous studies were conducted in a non-partner-specific (NPS) way (the conservation of interfaces were examined without specifying a specific interaction partner), and only on surface residues, which requires the knowledge of protein structures. However, on one hand, many biological pivotal protein-protein interactions, such as transient interactions, are highly regulated and partner-specific (PS). Detecting the conservation of interfaces that are highly partner-specific requires a PS interface conservation analysis that takes the knowledge of binding partners into account. On the other hand, to make use of the interface conservation results in inferring interface residues from protein primary sequences, a rule of interface conservation relative to the rest of residues (both surface and interior residues) is needed. To this end, we studied more than 300,000 pair-wise alignments of protein sequences from structurally characterized protein complexes, including both obligate and transient complexes, in both NPS and PS way. We identified sequence similarity criteria required for accurate sequence homology based inference of interface residues in a query protein sequence.

2.1 Introduction

The relation between sequence conservation and various aspects of protein structure, interaction, expression, and function has been the focus of many studies over the past decades. Because proteins with similar sequences often have similar structures and similar functions sequence homology based methods have been used, among other things, for structure prediction, homology modelling, and function prediction. Sander and Schneider [117] defined the HSSP curve to describe the relationship between sequence identity and alignment length. Rost [115]demonstrated the relationship between alignment length and the extent of protein structure similarity. Several authors have used methods that use the amino acid sequence of the target protein and the 3D structure of a homologous protein to model a 3D structure of the target protein (homology modelling) [57]. Homologs are also widely used for protein function annotation [4, 10, 86, 130]. Nair and Rost [98] have identified conserved sequence signatures that can be used to predict subcellular localization of proteins. Several authors have used conserved structure and sequence features to predict protein-protein interacting partners [89, 134, 150].

Thus, it is natural to ask whether protein-protein interface residues can be reliably identified using sequence homology based methods. Published studies disagree on whether protein-protein interfaces are more conserved than the rest of the protein sequences. Grishin and Phillips [60], after examining five enzyme families, concluded that the degree of conservation of interfaces is same as that of protein sequences as a whole. The studies by Caffrey et al. [22] as well as Reddy and Kaznessis [113], found that the interacting surface-patches are not significantly more conserved than other surface-patches. Caffrey et al. [22], based on their study of 64 protein-protein interacting chains, found that interface residues are slightly more conserved than the rest of the protein surface residues. Reddy and Kaznessis [113], based on their study of 28 hetero transient and non-transient complexes, found that the fraction of highly conserved interface residues is greater than that of highly conserved non-interface surface residues. They suggested that the number of conserved residue positions is more predictive of protein-protein binding sites than the average conservation index of residues in the target patch. Choi et al. [32] analyzed 2,646 protein interfaces based on a conservation score that measures the position-specific evolutionary

rate estimated using a phylogenetic tree, and concluded that protein interface residues are more conserved than non-interface surface residues.

All these protein-protein interface conservation studies share several features, first two of which might explain their conflicted results. First, these studies were conducted in a non-partner-specific way, i.e., the interface conservation of a protein was studied without taking into account a specific interaction partner. However, many proteins use different interfaces to interact different binding partners. Ignoring the partner-specific property of interfaces may result in accurate estimation of interface conservations. Second, small datasets were used. All studies but Choi 2009 [32] have used small datasets, which may not be large enough to draw a general applicable conclusion of interface conservation. Third, these studies except Grishin and Phillips [60] have been focused on the comparison of conservation degrees of interface residues relative to other surface residues, which requires the knowledge of protein structures. To make the interface conservation applicable for sequence-based protein-protein interface predictions (which take the whole protein sequence as input and without the knowledge of protein structures, hence surface residues), we need to identify the relation of interface conservation and the sequence similarity (without differentiating the surface and the interior residues).

Against this background, we introduce a novel measure of interface conservation that captures the degree to which interface residues in each protein are conserved among its sequence homologs. First, we describe the results of our analysis of the interface conservation among homologous sequences using several large non-redundant datasets of protein-protein interfaces extracted from the Protein Data Bank (PDB) [14], including datasets that allow us to compare "obligate" versus "transient" interfaces. To explore the extent to which interface conservation can be exploited in the prediction of interface residues, we systematically examined the relationship between interface conservation and six sequence-based variables. In one set of experiments, we examined binding interfaces in homologous proteins without specifying a specific interaction partner (i.e., non-partner specific, *NPS-interfaces*). The results of this analysis indicated that interfaces in obligate complexes are, in general, more highly conserved than those in transient complexes. In a complementary set of experiments, we examined interfaces in complexes between specific pairs of proteins (i.e., partner-specific, *PS-interfaces*). In contrast to the results

for NPS-interfaces, by focusing on the interface of each query protein with a specific binding partner, we discovered a high degree of sequence conservation in transient PS-interfaces. This analysis revealed that transient interfaces tend to be highly partner-specific.

2.2 Results

To define conditions under which it should be possible to infer protein-protein interface (PPI) residues using conservation of interfaces in homologous proteins and/or complexes, we systematically examined the relationship between interface residue conservation and sequence similarity (based on BLAST alignments). Our analyses are based on the following datasets: Nr6505 (a large non-redundant dataset of protein chains extracted from PDB), Oblig94 and Trans135 (a non-redundant obligate/transient binding dataset taken from [94]), and nr_pdbaa_s2c (BLAST database) (see Methods for additional details).

2.2.1 Conservation of PPIs in Non-Partner Specific (NPS) Interfaces

First, we examined the conservation of PPI residues in the absence of knowledge of interaction partners. For this study, we analyzed interfaces in putative homologs (hereafter, we refer to putative homologs as "homologs" for simplicity) of each protein in a large non-redundant dataset, Nr6505. After removing chains with interfaces containing fewer than 3 amino acids, we were left with 5853 chains. For each of the 5853 remaining proteins, we extracted homologs from the nr_pdbaa_s2c database using BLASTP [9] with expectation value (EVal) ≤ 10 from the resulting set of homologs, we eliminated those that were nearly identical to the query sequence (to ensure an accurate estimate of conservation). To ensure that the interface residues of the homologs could be reliably determined, we retained only those homologs that were part of complexes with resolution 3.5 Å or better. For each query-homolog pair in sequence alignments generated by BLASTP, we used the interface residues of the homolog(s) to predict the interface residues of the query protein. We calculated the correlation coefficient (CC) between the predicted and actual interface residues of the query protein, and refer to this value as the interface conservation (IC) score, i.e., the degree of conservation of interface residues between the query protein and its homologs (see Methods for details).

We examined the dependence of the interface conservation score on six NCBI BLAST alignment statistics: Expectation value (EVal), Identity Score, Positive Score, Local Alignment Length (LAL) and two Alignment Length Fractions (LAL/Query Length) and (LAL/Homolog Length). The EVal is a statistic that estimates the number of hits expected by chance when searching database of a particular size; the lower the EVal value, the more significant the score. The Identity Score is a measure of the degree of sequence identity between two amino acid sequences. The Positive Score returned by BLASTP is the number of positive-scoring matches in an alignment. It takes into account observed substitutions that preserve the physicochemical properties of the original residue. The LAL is the length of the local alignment; Alignment Length Fractions are LAL normalized by the length of the query or the length of the identified homologous sequence. We represent each query-homolog pair as a six dimensional vector defined by these six variables.

2.2.1.1 Principal Components Analysis of NPS-interface conservation space

As a first exploratory step, PCA (Principal Component Analysis) was applied to visualize the relationships between the interface conservation (IC) scores and the six BLAST alignment statistics. PCA, which is a dimensionality reduction technique, is typically used to represent dimensions that explain maximum variability and provide a simple and parsimonious description of the covariance structure [70].

Figure 2.1 shows a PCA biplot in which each data point, representing a query-homolog pair, is projected from the original 6-dimensional space to a 2-dimensional space defined by the first and second principal components (PC1 and PC2). A large fraction (88.58%) of the variance is explained by the first two principal components (48.75% + 39.83%). Based on IC scores, the PCA biplot can be subdivided into three regions that correspond to: (i) Dark Zone: containing query-homolog pairs with poorly conserved interface residues (blue and green data points), corresponding to low values of the CC between predicted and actual interfaces and thus low IC scores; (ii) Twilight Zone: containing pairs with moderately conserved interfaces (yellow and orange data points); and (iii) Safe Zone: containing pairs with highly conserved interfaces (red data points).

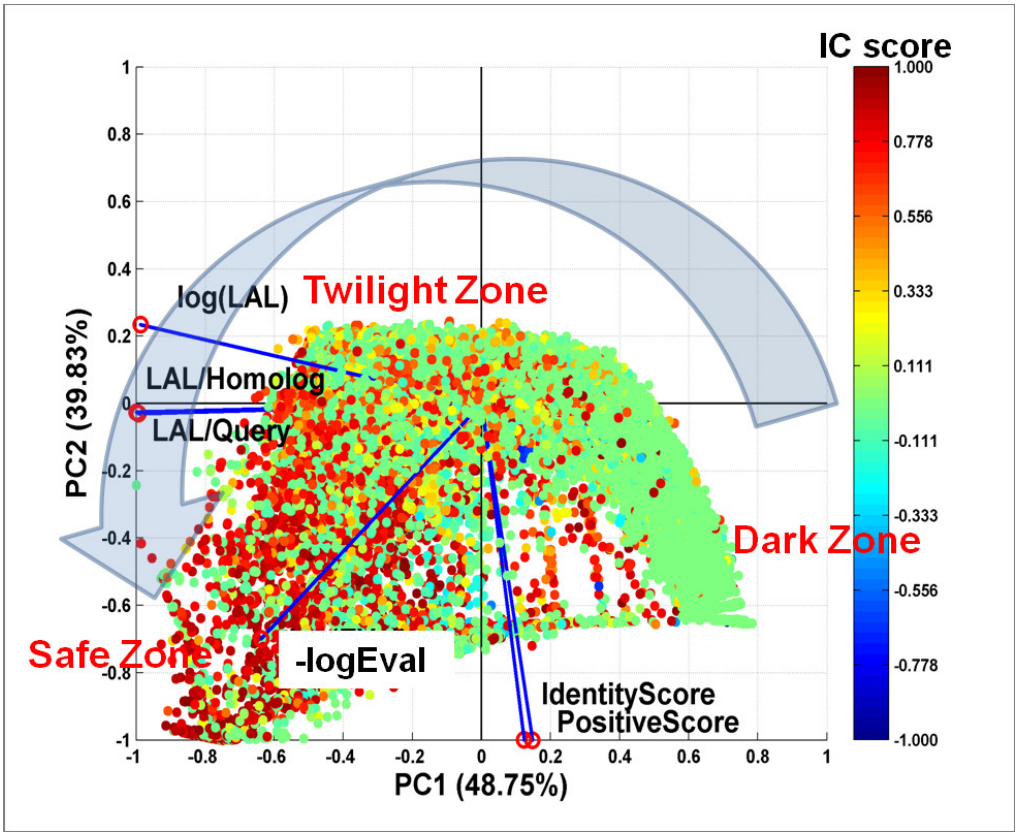


Figure 2.1 Principal Component Analysis of Interface Conservation Scores and Sequence Alignment Statistics. Proteins in the Nr6505 and their homologs were analyzed. The data points in the biplot correspond to the projection of a 6-dimensional vector representing each protein-homolog onto a 2-dimensional space defined by the first and second principal components (PC1 and PC2). Blue lines with red circles at their tips represent the axes of the original 6-dimensional space for the 6 variables used in PCA analysis: -log(EVal), Identity Score, Positive Score, log(LAL), alignment length fractions (LAL/query length) and (LAL/homolog length). Each data point is colored according to its computed interface conservation (IC) score, with higher IC scores (red/orange) indicating higher interface conservation and lower IC scores (blue/green) indicating lower interface conservation (see text for details). The large gray arrow indicates the direction of increasing degree of interface conservation, from Dark to Twilight to Safe Zone.

The PCA analysis allows us to identify highly correlated explanatory variables. In Figure 2.1, the axes of the original 6 dimensional space are represented as blue vectors with red circles at their tips in the 2-dimensional space defined by PC1 and PC2. Highly correlated vectors (variables) have small angles between them. This type of analysis reveals, for example, that the two Alignment Length Fractions are highly correlated with each other, as are the Positive Score and Identity Score. Explanatory variables that are highly correlated with each other make similar contributions to the IC score.

2.2.1.2 BLAST EVal is a strong indicator of NPS-interface conservation

We studied the relationship of each individual variable with interface conservation. A scatter plot in which the IC score for each query-homolog pair is plotted against log(EVal) is shown in Figure 2.2. One can see that log(EVal) is a good indicator of protein interface conservation. When log(EVal) $>$ -50 the median values of IC scores cluster around 0 (low conservation). In the region of log(EVal) $\leq$-50 (that is, EVal $\leq$ 1.9287E-022) the medians of IC scores increase as the log (EVal) decreases. When log(EVal) $<$ -100 the medians of IC scores tend to be greater than 0.5 (strong conservation).

2.2.1.3 NPS-interface conservation in Twilight/Safe Zone is strongly positively correlated with log(LAL)

Figure 2.3 is a scatter plot showing the IC score for each query-homolog pair plotted against the log of its LAL value. We can clearly see that when log(LAL) is larger than 4, the medians of IC score show a strong positive correlation with log(LAL). When the LAL is shorter than 55 residues (log(LAL) $<$4), the probability that interface is conserved in these homologs is low (the medians of the IC scores are ˜0). We define this region as the Dark Zone. When the LAL is longer than 700 residues (log(LAL)$>$6.55), interface conservation is high (the medians of IC scores are usually $>$ 0.7). We define this region as the Safe Zone.

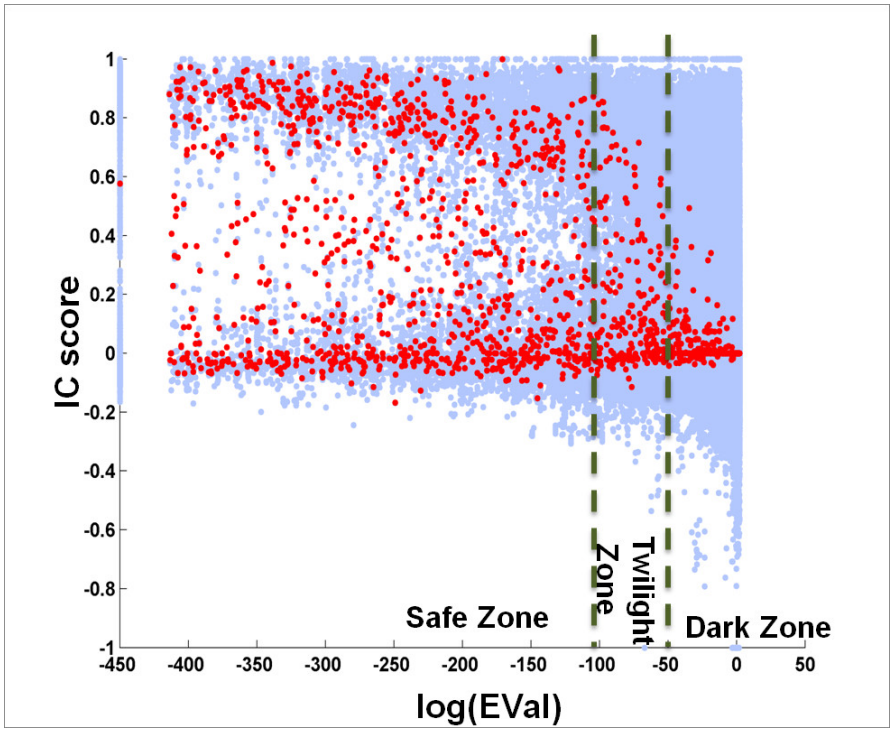


Figure 2.2 EVal is a Good Indicator of Interface Conservation. Each blue dot in the scatter plot corresponds to a query-homolog pair. Red dots are the median values of IC scores for a specific EVal. To avoid log(0), we set log(EVal) = -450 when EVal = 0.

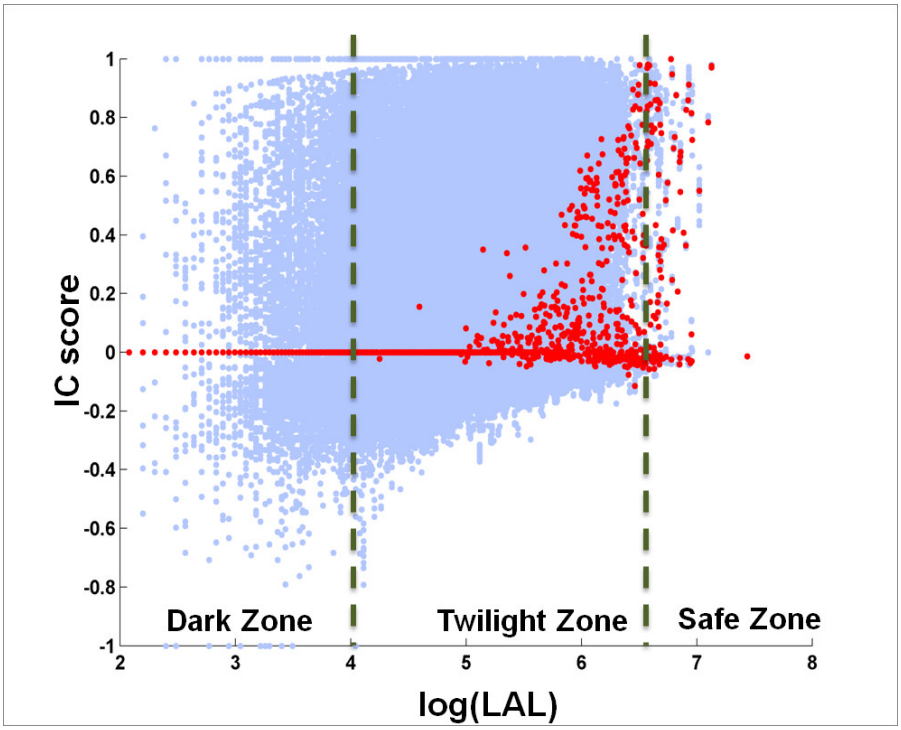


Figure 2.3 Interface Conservation (IC) Scores are Linearly Related to the Log of the Local Alignment Score (LAL). Each blue dot in the scatter plot corresponds to a query-homolog pair. Red dots are the median values of IC scores for a specific LAL. Note the trend of increasing median IC score with log(LAL) observed with the transitions from Dark to Twilight to Safe Zone.

2.2.1.4 A high BLAST Positive Score reflects NPS-interface conservation

The relationship between IC scores and the Positive Scores of query-homolog alignments is shown in Figure 2.4. The median values of the IC scores begin to increase at a BLAST Positive Score of ~ 90%.

We also studied the relationship of IC score with the Identity Score, and the Local Alignment Length Fractions (LAL/Query Length) and (LAL/Homolog Length). As expected, the Identity Score results were similar to those for the Positive Score. The IC score was not as strongly linearly related to LAL fraction as it was to the log(LAL) (data not shown). Taken together, these results provide guidelines for choosing sequence similarity thresholds that reflect the degree of conservation in NPS interfaces.

2.2.1.5 NPS-Interface conservation as a function of sequence alignment

We built a linear model for NPS-interface conservation based on the most important sequence alignment statistics identified in the PCA analysis: logEVal, Positive Score, logLAL.

The model is

$$IC\ Score = \beta_0 + \beta_1 log(EVal) + \beta_2 PositiveS + \beta_3 log(LAL) \quad (2.2.1.5)$$

Variables, parameter estimates and coefficients are shown in Table 2.1. All the coefficients are significant.

Table 2.1 Variables, Parameter Estimates and Significance Values for the Linear Model for NPS-Interface Conservation.

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
β_0	-0.5655	0.0080	-70.66	<.0001
β_1	-0.0004	0.0000	-23.3	<.0001
β_2	0.0037	0.0000	54.44	<.0001
β_3	0.1057	0.0011	94.62	<.0001

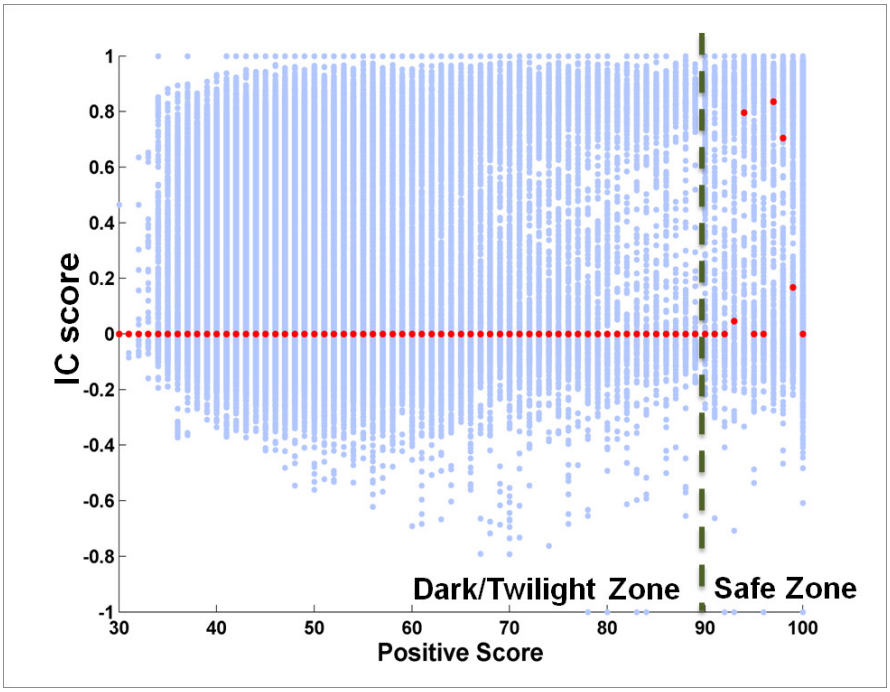


Figure 2.4 A High BLAST Positive Score Reflects NPS-Interface Conservation. Each blue dot in the scatter plot corresponds to a query-homolog pair. Red dots are the median values of IC scores for a specific Positive Score. Note that medians of IC scores are near zero until Positive Scores become larger than 90 %.

2.2.1.6 NPS-Interface conservation in transient versus obligate binding proteins

In light of reports that protein interfaces in transient complexes are not as conserved as those in obligate (permanent) complexes [32], it is interesting to ask whether the query-homolog pairs with near-zero IC scores (Figures 2.2 and 2.3) tend to involve proteins that participate in transient interactions. To address this question, we further studied the differences in protein interface conservation among proteins that participate in transient versus obligate interactions.

To compare protein interfaces in transient and obligate complexes, we used the Trans135 and Oblig94 dataset obtained from [94], which includes a total of 270 chains from transient and 188 chains from obligate complexes. We extracted the homologs of each chain from nr_pdbaa_s2c using BLASTP with EVal ≤ 10. Query and homolog proteins with interfaces containing fewer than 3 amino acids were removed, as were homologs that were nearly identical to the query proteins. We extracted 43,115 query-homolog pairs containing chains that participate in transient interactions and 24,212 pairs containing chains that participate in obligate interactions.

In agreement with previous studies [32], our analyses showed that PPIs are conserved in both obligate and transient binding proteins. As before, we performed PCA to examine the conservation of interfaces as a function of log(EVal), Identity Score, Positive Score, log(LAL), and alignment length fractions (LAL/Query Length) and (LAL/Homolog Length). The PCA biplots in Figure 2.5 show that data points corresponding to different IC scores (different colors) are partially segregated, indicating that the six alignment statistics can distinguish query-homolog pairs with highly conserved interface residues (red) from those in which interface residues are not conserved (blue or green).

The results in Figure 2.5 also reveal that interface residues in proteins from obligate complexes (left panel) are more conserved among their sequence homologs than those from transient complexes (right panel). Figure 2.6 further illustrates differences in interface conservation in obligate (left) versus transient complexes (right). The median values of IC scores plotted as a function of log (LAL) are more frequently above 0 for pairs that involve obligate binding proteins (Figure 2.6a) than for those that involve transient binding proteins (Figure 2.6b). Regression

analysis of these data confirms that log(LAL) for the obligate dataset has a larger coefficient (0.095) than that for the transient dataset (0.052), which confirms that protein interfaces are more conserved in the obligate complexes than in transient complexes analyzed in this study.

Figure 2.6 c reveals an obvious pattern of interface conservation in obligate binding proteins: a strong trend of increasing median IC score with decreasing log(EVal). In contrast, Figure 2.6 d shows that for transient binding proteins, more of the median values of IC scores cluster around 0, indicating that log(EVal) has little relation to interface conservation in transient complexes.

Also, comparison of Figure 2.6 e and f reveals that the Positive Score is a good indicator of interface conservation in the case of proteins from obligate complexes; however, this is not the case for proteins from transient complexes. For obligate binding proteins, when the Positive Score exceeds 45%, the medians of IC scores begin to show an increasing trend (Figure 2.6 e). In contrast, in the case of transient binding proteins, medians of IC scores do not begin to increase until the Positive Score approaches 70% (Figure 2.6 f).

It is important to emphasize that all of the interfaces analyzed above are what we refer to as "non partner-specific" (NPS). That is, the interface residues of a query protein represent the complete set of its interface residues with all of its partners. However, a given query protein can interact with different binding partners through different interfaces. A possible explanation for the low IC scores for NPS-transient interfaces is that the union of all interface residues of a transient binding protein are not highly conserved across its homologs. This does not preclude the possibility that such interfaces are conserved in the context of partner-specific interactions. We investigate this possibility in the following section.

2.2.2 Conservation of PPIs in Partner-Specific (PS) Interfaces

To examine the conservation of partner-specific (PS) interfaces in transient protein complexes, we again used the Trans135 dataset of protein pairs that participate in transient interactions. For each of the proteins in an interacting pair, we separately extracted the corresponding homologs, using BLASTP with expectation value EVal$\leq$10 against the nr_pdbaa_s2c database. We removed homologs that are part of complexes with resolution worse than 3.5 Å. If query proteins A and B form a complex A-B, and have homologs A' and B' that interact in a

complex A'-B', we consider A'-B' as a homo-interolog of A-B. To ensure an accurate estimate of conservation, from the resulting set of homo-interologs, we eliminated those that were within the same PDB complex as the query proteins, and those that were nearly identical to the query pairs (see Methods for additional details). For each protein chain in a query pair, we use the interface residues of its homolog in a homo-interolog to infer the PS interface residues of the query protein chain. Thus, we use the interface residues of A' in the homo-interolog (A'-B') of query pair A-B to infer the interfaces of A with B, based on the sequence alignment between A and A' obtained using BLASTP. We measure the similarity between a pair of interacting proteins A-B and its homo-interolog A'-B', in terms of the metrics for the quality of sequence alignment between A and A' and between B and B', using the six BLAST alignment statistics described above.

We used PCA of 3, 456 candidate homo-interologs to explore the relationship between interface conservation (IC score) and the six alignment statistics computed from the predicted PS interfaces, e.g., of chain A when it interacts with B, using known interfaces of A' with B'. This analysis revealed that much of the observed variance in IC scores is explained by three factors: (i) the average log (EVal); (ii) the average Positive Score of the homo-interolog and (iii) the alignment fractions FracA, FracA', FracB, and FracB' computed from the alignments of constituent chains (A with A' and B with B') (see Methods for additional details).

The results in Figure 2.7 show that transient interfaces are highly conserved in homo-interologs. The trend of increasing median IC scores, as a function of decreasing logEval (Figure 2.7 a) or increasing Positive Score (Figure 2.7 b) or the combination of Positive Score and FracA $\times$ FracA' is clear (Figure 2.7 d). The trend of increasing IC scores as a function of FracB $\times$ FracB' is similar to that as a function of FracA $\times$ FracA' (data not shown). In contrast, the, logLAL, which is the average of alignment length between A and A', and between B and B', is not strongly correlated with interface conservation for PS-interfaces (Figure 2.7 c).

A comparison of the results for PS-interface conservation in transient complexes here (Figure 2.7 a and b) with those obtained for NPS-interface conservation in transient complexes above (Figure 2.6 d and f), reveals that the conservation of transient interfaces can be detected easily when the binding partner sequence information is utilized. The seemingly weak conservation

of interfaces in transient complexes shown in Figure 2.6 is thus a consequence of the specificity of transient interfaces for different partners. Therefore, we conclude that interfaces in transient complexes are both highly partner-specific and highly conserved, when their partner-specificity is taken into account.

2.2.2.1 PS-Interface conservation as a function of sequence alignment

We built a linear model for PS-interface conservation based on the important sequence alignment statistics identified in the PCA analysis: logEVal, Positive Score, FracAA' and FracBB', where

$$logEVal = \frac{log(EVal_{AA'})+log(EVal_{BB'})}{2}$$

$$PositiveS = \frac{PositiveS_{AA'}+PositiveS_{BB'}}{2}$$

$$Frac_{AA'} = Frac_A \times Frac_{A'}$$

$$Frac_{BB'} = Frac_B \times Frac_{B'}$$

$$Frac_A = \frac{LAL_{AA'}}{length_A},\ Frac_{A'} = \frac{LAL_{AA'}}{length_{A'}},\ Frac_B = \frac{LAL_{BB'}}{length_B},\ Frac_{B'} = \frac{LAL_{BB'}}{length_{B'}}.$$

A-B is query protein pair and A'-B' is the homo-interolog of A-B. $EVal_{AA'}$ and $EVal_{BB'}$are the $EVal$ between A and A', and between B and B'. $positiveS_{AA'}$and $positiveS_{BB'}$ are the BLAST Positive Score between A and A', between B and B'. The model is

$$IC\ Score = \beta_0 + \beta_1 logEVal + \beta_2 PositiveS + \beta_3 Frac_{AA'} + \beta_4 Frac_{BB'} \quad (2.2.2.1)$$

Variables, parameter estimates and coefficients are shown in Table 2.2. All the coefficients are significant.

Table 2.2 Variables, Parameter Estimates and Significance Values for the Linear Model for PS-Interface Conservation.

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
β_0	-0.505	0.040	-12.62	<.0001
β_1	0.001	0.000	6.16	<.0001
β_2	0.009	0.001	14.6	<.0001
β_3	0.341	0.027	12.54	<.0001
β_4	0.205	0.028	7.4	<.0001

2.3 Discussion

2.3.1 Protein Interface Conservation across Structure Space

The study of protein interface conservation among proteins with similar structures has received considerable attention in recent years. By analyzing the structural similarity of representative protein-protein interfaces in dimeric proteins, Gao and Skolnick [55] showed that the vast majority of native interfaces have a close structural neighbor with similar backbone $C\alpha$ geometry and interface contact pattern.

In a related study, Zhang et al. [152] explored the conservation of interface residues among structural neighbors of a query protein (i.e., proteins that share the same SCOP family, superfamily or fold, or a high degree of structural similarity regardless of their SCOP classification). They showed that: (i) interfaces are indeed conserved among structural neighbors; (ii) the degree of interface conservation is most significant among proteins that have a clear evolutionary relationship. They further showed that conservation of interface residues among structural neighbors can be successfully exploited to predict protein-protein interfaces based on protein structure information.

To investigate the extent to which conservation of interface residues can be used to improve the prediction of protein-protein interfaces based on protein sequence information, we systematically studied interface conservation across sequence space. Our results demonstrate that protein interfaces from different binding types are conserved among proteins with homologous sequences. We further showed that the degree of conservation of interfaces is even greater when putative interaction partners are taken into account. The IC score, our measure of interface conservation, unlike those used in previous studies [32] (e.g., residue conservation in sequence alignments), makes direct use of experimentally determined interface residues to measure the degree of interface conservation. Specifically, the IC score directly measures the extent to which the interface residues of sequence homologs of a query protein are predictive of the interface residues of a query protein. Hence, the IC score provides the basis for setting the parameters of our sequence homology based interface prediction methods.

2.3.2 Distance Functions for Identifying Putative Homologs with Conserved Interfaces

Because we do not know the IC score for a query sequence with unknown interface residues, we identified several statistics associated with the BLASTP alignment of a query sequence with its homologs that are correlated with the IC score. We found that interface residues of a query protein can be reliably predicted from the known interfaces of its homologs (and in the case of partner-specific predictions, the homologs of its interaction partner as well) when the homologs are selected taking into account measures of quality of sequence alignment, specifically NCBI BLAST sequence alignment statistics. The HomPPI methods presented here use simple linear combinations of BLAST sequence alignment statistics, determined using PCA analysis of the relationship between the statistics and the IC score. It would be interesting to explore optimal, perhaps non-linear, combinations of parameters to maximize the desired performance criteria (e.g., sensitivity, specificity, or some combination thereof).

2.3.3 Conservation of Interfaces in Obligate and Transient Complexes

We found interface residues to be more highly conserved than non-interface residues, in both obligate and transient complexes. We also found that when information regarding the specific binding partner of a query protein is not taken into account in estimating the conservation score, interfaces in transient complexes appear to be less highly conserved than those in obligate complexes. Our results further show that transient interfaces are highly partner-specific, and that the partner-specific interfaces in transient complexes are, in fact, highly conserved.

2.3.4 Interface *Residue* Conservation and Interface *Position* Conservation

Two different but related methods can be used to measure protein interface conservations: interface *residue* conservation and interface *position* conservation. Each one has its advantages and disadvantages.

Most, if not all, previous interface conservation analyses were conducted in the former way -interface *residue* conservation. Specifically, for each protein sequence, a multiple se-

quence/structure alignment (MSA) is constructed. The conservation score of a residue in a specific position in the alignment can be calculated by counting the frequency of the same or similar residues in the aligned sequences/structures appear in the same position [1]. The advantage of this method is that a query protein sequence is aligned with a large database of protein sequences, which can be relatively easily determined. However, this method has two potential challenges when applying to protein-protein interface predictions. 1) For each query protein sequence, it requires a number of quality homologs aligned in order to reasonably estimate residue conservation scores of residues of a protein. 2) A residue that has a high conservation score defined by this method is not necessarily a protein-protein interface residue, since other functional residues and residues in the protein core can also be highly conserved. Therefore, the conservation score derived from this method may not be used alone to infer protein interface residues out of the whole protein sequence.

Different from the above interface *residue* conservation analysis method, the latter method – interface *position* conservation method - is used in this study. For each query-homolog alignment pair, we calculate a Interface Conservation score (i.e. the similarity of their interface vectors , which essentially estimate whether a specific position in the alignment is conserved as an interface position in putative sequence homologs [2]). Tens of thousands of IC scores are calculated for each protein with experimentally determined interfaces in a large non-redundant dataset against each of the putative sequence homologs with experimentally determined interfaces. A regression model built on these IC scores and the sequence similarities of query-homologs can be later used to predict IC score (the confidence using the experimentally determined interface residues of a putative sequence homolog to infer interfaces of a query protein sequence). One advantage is that when using the regression model to predict IC score, it only requires one protein with experimentally determined interfaces aligned with the query protein sequence, in that the conservation score of the set of interfaces of a protein is calculated for each query-homolog alignment pair. Another advantage is that as the experimentally determined interface residues

[1] Usually the conservation scores are not explicitly calculated in the form of frequency, but using other finer measurements in the same spirit, for example, the Shannon Entropy.

[2] The interface residues of both query protein and the homolog protein are determined using their solved 3D structures. Interface vector is a binary vector with one for interface residues and zero for non-interface residues. Residues aligned with gaps and residues that are not aligned with the other protein are ignored.

of putative sequence homologs are used to calculate the conservation score, the differentiation of surface residues and interior residues, a problem that the first method faces, is naturally solved. Besides, the conservation conclusion of this method is more closely related to the design of a homology-based interface predictor in that the conservation score itself is essentially the evaluation of a prediction using the interface of one homolog as predictions.

However, interface *position* conservation is limited by the availability of experimentally solved structures when applying to protein interface predictions. Some query proteins may not even find one putative sequence homologs with experimentally determined interfaces and with satisfactory predicted IC score. We estimate this limitation of interface *position* conservation analysis in applying to predictions and discuss possible ways for improvement in Chapter 3.

2.4 Conclusions

We studied a large number of sequence alignments between protein pairs with known interfaces to explore the conditions under which conservation of protein interface residues, as determined by the alignment of a query sequence against its homologs/homo-interologs, can be used to reliably predict protein-protein interfaces. We showed that the PCA biplot is a convenient tool to visualize the multivariate relation between the interface conservation score and the sequence similarity measures. We proposed a novel method to study partner-specific protein interface conservations, and detected that transient interfaces are highly partner-specifically conserved.

2.5 Methods

2.5.1 Datasets

Three datasets were used in our conservation analysis in this chapter:

• Nr6505 - For analyzing the protein interface conservation.

• Oblig94 and Trans135 - For comparing the degree of conservation of protein interfaces in transient/obligate binding proteins.

• nr_pdbaa_s2c - For BLASTP searching for close sequence homologs

Nr6505

We extracted a maximal non-redundant set of known protein-protein interacting chains from the Protein Data Bank (PDB) [14] available on 2/4/2010. We used the following steps to build Nr6505 to eliminate the influence of over-represented protein families in PDB:

1. Extract all the X-ray derived protein structures with resolution 3.5 Å or better in PDB. Remove proteins with less than 40 residues. We obtained 102,853 protein chains.

2. Remove redundancy of the resulting dataset in step 1 using PISCES [137]. All the remaining sequences have less than or equal to 30% sequence similarity. We obtained 6505 chains.

Oblig94 and Trans135

This dataset of 94 obligate protein-protein dimer complexes and the dataset of 135 transient dimer complexes was obtained from a large non-redundant dataset of 115 obligate complexes and 212 transient complexes (3.25 Å or better resolution, determined using X-ray crystallography) previously generated by Mintseris and Weng [94] to study the conservation of protein-protein interfaces. In ordered to exclude the influence of other types of interfaces, we extracted 94 obligate dimers and 135 transient dimers from the original dataset and get Oblig94 and Trans135. In Oblig94, 1QLA has been superseded by 2BS2. In Trans135, 1DN1 and 1IIS have been superseded by 3C98 and 1T83, respectively, and 1F83, 1DF9, 4CPA and 1JCH have since been deemed as obsolete and hence discarded from PDB.

BLAST nr_pdbaa_s2c

This dataset is used for BLASTP searches. We used the fasta files from S2C database [136] to generate our BLAST database nr_pdbaa_s2c. We removed proteins with resolution worse than 3.5 Å from S2C fasta formatted database. We built a non-redundant database for BLAST queries from the S2C fasta formatted database. To generate the non-redundant BLAST database, we grouped proteins with identical sequences into one entry. We used the resulting database to search for homologs of a query sequence using BLASTP 2.2.22+ [9]. There are

36,352 sequences and 9,549,671 total residues in nb_pdbaa_s2c.

2.5.2 Interface Definition

This paper adopts a stringent definition of protein-protein interfaces. Surface residues are defined as residues that have the relative solvent accessible area (RASA) at least 5% [111]. Interface residues are defined as surface residues with at least one atom that is within a distance of 4 Å from any of the atoms of residues in the chain. The ratios of interface residues versus the total number of residues for the datasets used in this work are summarized in Table 2.3. Interface information was extracted from the ProtInDB server http://protInDB.cs.iastate.edu .

Table 2.3 The Proportion of Interface Residues in Datasets Used in Interface Conservation Analysis.

Dataset	**Number of Interface Residues**[a]	**Total Number of Residues**[b]	**% Interface Residues**
Nr 6505	145,498	1,377,630	10.6%
Oblig94	10,273	55,400	18.5%
Trans135	6,460	55,217	11.7%

[a]When a chain interacts with more than one other chain, the interfaces are counted separately. For example, for protein complex 2phe C:AB, the interface of C with A and the interface of C with B are regarded as two interfaces.

[b]Residues that missing from PDB structures are not counted.

2.5.3 Mapping Interfaces in Structures to Sequences

We label the protein sequences as interface or non-interface residues (according to the definition of interface residues given above) as follows: We first calculate the relevant distances between atoms using the atom coordinates in ATOM section in PDB files. Then, by associating the ATOM section to residues in the SEQRES section, we can map the corresponding residues to protein sequences. However, various errors in PDB files make this a non-trivial task. Hence, we used the mapping files from S2C database, which offers corrected mapping information from ATOM section to residues in the SEQRES section of PDB files, to map interfaces determined in structures to full sequences.

2.5.4 NCBI BLAST Parameters

The amino acid substitution matrix and gap cost are essential parameters that need to be specified in BLAST searches. In this study, we used the substitution matrices and gap costs recommended for the different query lengths [1] (See Table 2.4).

Table 2.4 BLAST Substitution Matrices and Gap Costs used for BLASTP searches.

Query Length	**Substitution Matrix**	**Gap Costs**
<35	PAM-30	(9,1)
35-50	PAM-70	(10,1)
50-85	BLOSUM-80	(10,1)
85	BLOSUM-62	(10,1)

2.5.5 Interface Conservation (IC) Scores

In protein interface conservation analysis, we used the Matthews correlation coefficient (CC) as a measure of the extent to which the interface residues in query protein are similar to those in a putative homolog. We treat a query protein as a test protein, and the aligned interface residues of its putative homolog is the prediction. For clarity, we refer this measure as the Interface Conservation (IC) score.

$$IC\,score = CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FP) \times (TN+FN)}}$$

where TP, FP, TN and FN are respectively the number of interface residues of a test protein that are correctly predicted to be interface residues, the number of residues of the test protein that are incorrectly predicted to be interface residues, the number of residues of the test protein that are correctly predicted to be non-interface residues, and the number of residues of the test protein that are incorrectly predicted to be non-interface residues.

Note that the part of the query protein that is not aligned with the putative homolog is not used in calculating $IC\,score$.

2.6 Acknowledgements

This work was funded in part by the National Institutes of Health grant GM066387 to Vasant Honavar and Drena Dobbs and in part by a research assistantship funded by the Center for Computational Intelligence, Learning, and Discovery.

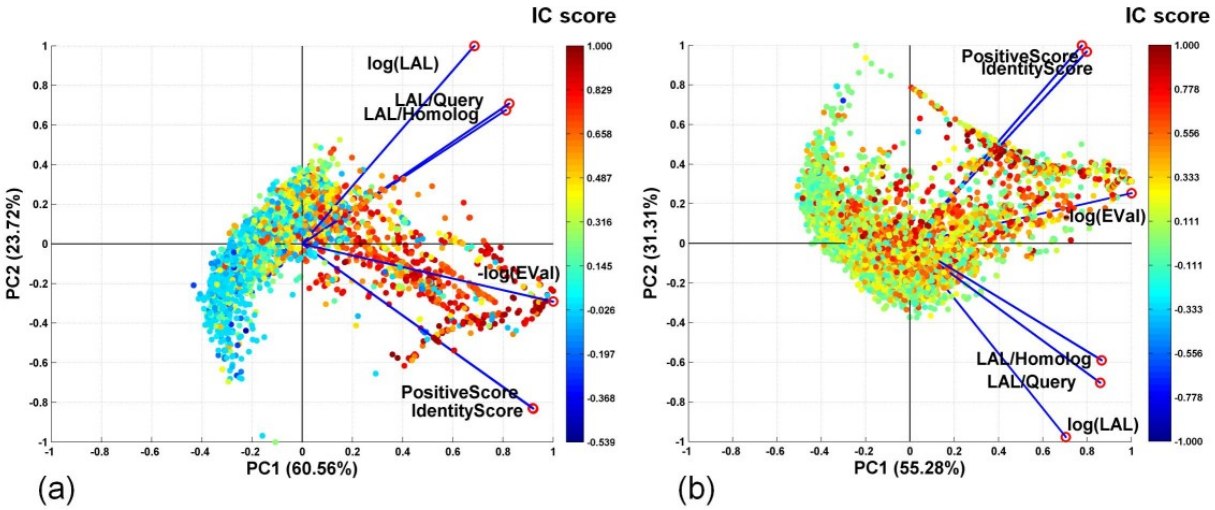


Figure 2.5 Principal Component Analysis of Interface Conservation Scores and Sequence Alignment Statistics for Obligate versus Transient Complexes. The PCA biplots shown are for (a) proteins from obligate complexes and (b) proteins from transient complexes. See Figure 2.1 legend for additional details.

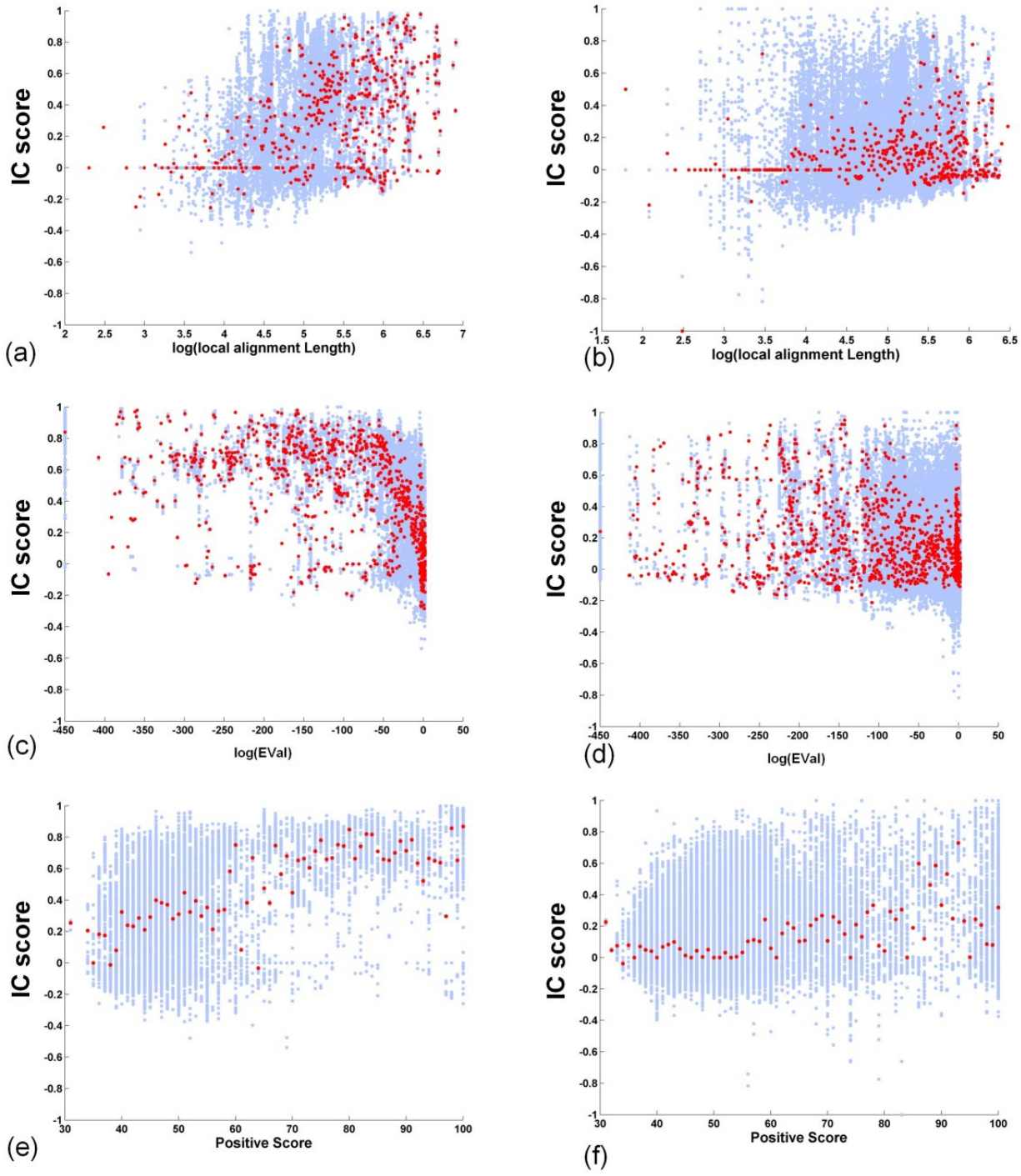


Figure 2.6 Comparison of Interface Conservation in Proteins from Obligate versus Transient Complexes. Proteins from obligate complexes are analyzed in a, c and e (left panels); proteins from transient complexes are analyzed in b, d, and e (right panels). Scatter plots show IC scores plotted as a function of: (a, b) log (local alignment length); (c, d) log (EVal); and (e, f) Positive Score. Red dots are median values of IC scores for a specific value on the x-axis.

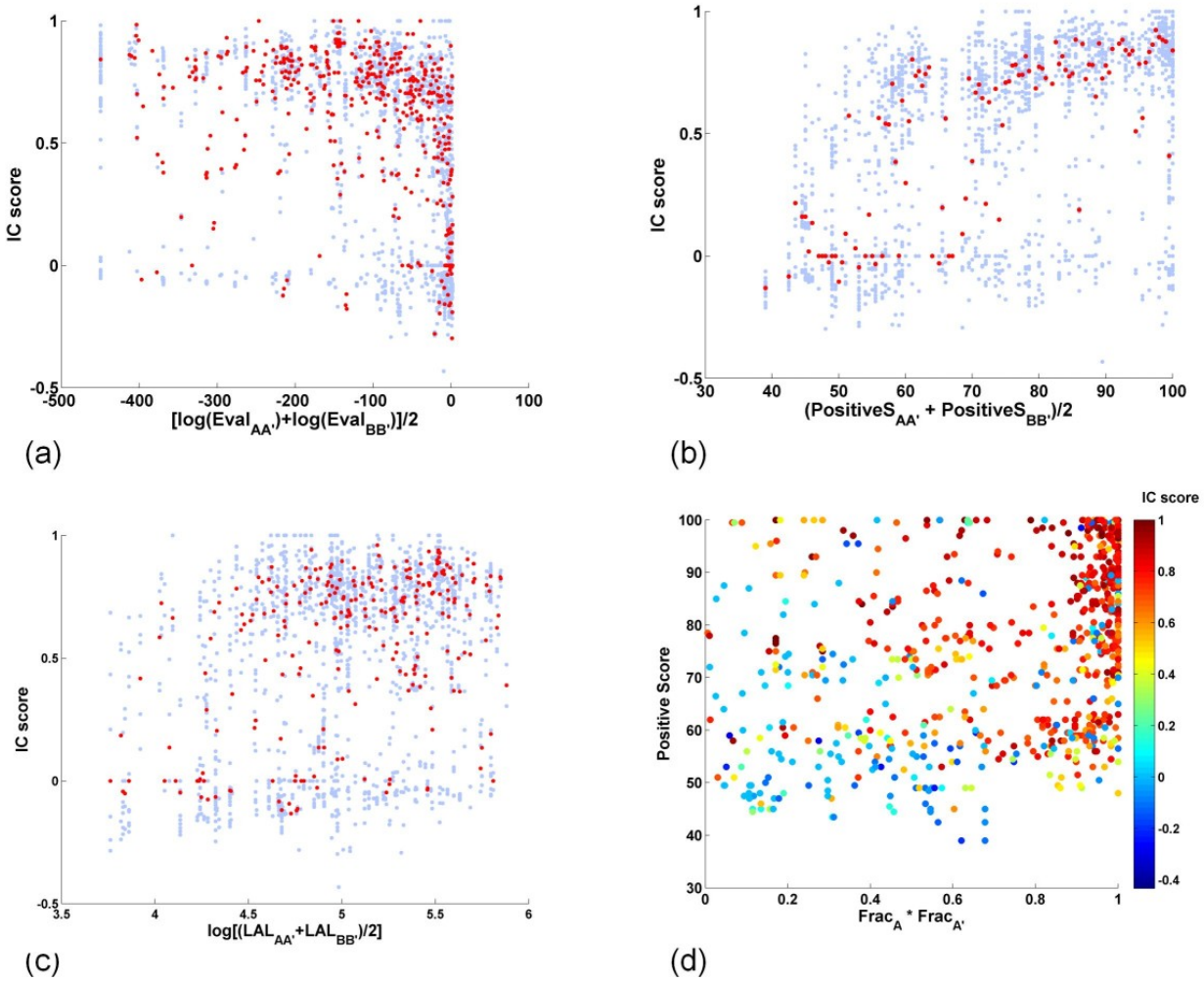


Figure 2.7 PS-Interface Conservation in Transient Complexes. Homo-interologs corresponding to complexes in the Trans135 dataset were analyzed (see text for details). (a-c) Scatter plots show IC scores (blue dots) plotted as a function of: (a) log EVal; (b) Positive Score; (c) log LAL. Red dots are median values of IC scores for a specific value on the x-axis. (d) Scatter plot of Positive Score as a function of FracA × FracA'. Each data point (in d only) is colored using according to its IC score.

CHAPTER 3. HomPPI: A class of Sequence Homology Based Protein-Protein Interface Prediction Methods

A paper titled "HomPPI: a class of sequence homology based protein-protein interface prediction methods", BMC Bioinformatics 2011, 12:244

Li C. Xue, Drena Dobbs and Vasant Honavar

Abstract Although homology-based methods are among the most widely used methods for predicting the structure and function of proteins, the question as to whether interface sequence conservation can be effectively exploited in predicting protein-protein interfaces has been a subject of debate. In Chapter 2, we systematically studied the interface conservation in sequence homologs, and we identified sequence similarity criteria required for accurate homology-based inference of interface residues in a query protein sequence. Based on these analyses, we developed HomPPI, a class of sequence homology based methods for predicting protein-protein interface residues. We present two variants of HomPPI: (i) NPS-HomPPI (Non partner-specific HomPPI), which can be used to predict interface residues of a query protein in the absence of knowledge of the interaction partner; and (ii) PS-HomPPI (Partner-specific HomPPI), which can be used to predict the interface residues of a query protein with a specific target protein.

Our experiments on a benchmark dataset of obligate homodimeric complexes show that NPS-HomPPI can reliably predict protein-protein interface residues in a given protein, with an average correlation coefficient (CC) of 0.76, sensitivity of 0.83, and specificity of 0.78, when sequence homologs of the query protein can be reliably identified. NPS-HomPPI also reliably predicts the interface residues of intrinsically disordered proteins. Our experiments suggest that NPS-HomPPI is competitive with several state-of-the-art interface prediction servers including those that exploit the structure of the query proteins. The partner-specific classifier,

PS-HomPPI can, on a large dataset of transient complexes, predict the interface residues of a query protein with a specific target, with a CC of 0.65, sensitivity of 0.69, and specificity of 0.70, when homologs of both the query and the target can be reliably identified. The HomPPI web server is available at http://homppi.cs.iastate.edu/.

Our results show that sequence homology based methods offer a class of computationally efficient and reliable approaches for predicting the protein-protein interface residues that participate in either obligate or transient interactions. For query proteins involved in transient interactions, the reliability of interface residue prediction can be improved by exploiting knowledge of putative interaction partners.

3.1 Introduction

Protein-protein interactions are central to protein function; they constitute the physical basis for formation of complexes and pathways that carry out virtually all major cellular processes. These interactions can be relatively permanent or "obligate" (e.g., in subunits of an RNA polymerase complex) or "transient" (e.g., kinase-substrate interactions in a signalling network). Both the distortion of protein interfaces in obligate complexes and aberrant recognition in transient complexes can lead to disease.

With the increasing availability of high throughput experimental data, two related problems have come to the forefront of research on protein interactions: i) prediction of protein-protein interaction partners; and ii) prediction of protein binding sites or protein-protein interfaces (PPIs). Although most effort to date has focused on one or the other of these problems, it is possible to use information from predicted protein-protein interaction networks as input for interface prediction methods, and predicted interface residues can be used as input for interaction partner predictions, a concept explored in a recent study of Yip et al. [149]. In the current study, we focus on the prediction of protein-protein interfaces, specifically, the use of sequence homology based methods to predict which residues of a query protein participate in its physical interaction with a partner protein or proteins.

Computational Prediction of Protein-Protein Interfaces

Several different genetic, biochemical, and biophysical methods have been used to identify and characterize protein interfaces [16, 42, 46, 23, 157, 7, 56, 120, 96]. These experiments are very valuable and have contributed greatly to our knowledge of protein-protein interfaces. However, the high cost in time and resources required for these experiments call for reliable computational approaches to identify interface residues. In addition to providing important clues to biological function of novel proteins, computational predictions can reduce the searching space required for docking two polypeptides [41, 38, 81, 122].

To distinguish interface residues from non-interface surface residues, a wide range of sequence, physicochemical and structural features have been investigated [71, 73, 72, 79, 11, 17, 24, 36, 84, 118, 123, 58], and many in silico approaches to protein-protein interface prediction have been explored in the literature (reviewed in [37, 52, 153]). Protein-protein interface prediction algorithms can be classified into three categories: (i) sequence-based methods, which use only the primary amino acid sequence of the query protein as input [114, 119, 108, 97, 146, 147, 29, 28]; (ii) structure-based methods, which make use of information derived from the structure of the query protein [99, 68]; and (iii) methods that use both sequence and structure derived information in making predictions [111, 26, 83].

Several sequence-based protein-protein interface prediction methods have been explored in the literature [114, 119, 108, 97, 146, 147, 29]. Most, if not all, of these methods, extract for each residue in the query protein, a fixed length window that includes the target residue and a fixed number of its sequence neighbours. Each residue is classified as an interface residue or a non-interface residue based on features of the amino acids in the corresponding window. Various methods differ both in the specific machine learning algorithms or statistical methods employed and in terms of the specific features of the amino acids used. Commonly used features include the identity of the amino acids in the window [147], the amino acid composition of interfaces [103] , the physicochemical properties of the amino acids [29], and the degree of conservation of the amino acids (obtained by aligning the query sequence with homologous sequences) [104]. Some studies report substantial improvements in interface residue prediction when predicted

structural properties, e.g., solvent surface accessibility and secondary structure of the residues are utilized [37].

A number of structure-based methods [99, 68] or hybrid methods that combine both sequence and structure-derived information [111, 26, 83] have been proposed for predicting protein interfaces. The performance of the best-performing sequence-based methods is generally lower than that of structure-based methods (see [37] for a comparison). A possible explanation for the difference in the performance of sequence-based and structure-based protein interface residue predictors is that the latter can trivially eliminate non-surface residues from the set of candidate interface residues and potentially exploit a rich set of features derived from the 3D structures.

The use of structure-based methods, however, is limited to proteins for which the structure of the query protein is available, and the number of solved structures significantly lags behind the number of protein sequences [29]. Even when the structure of a query protein is available, the application of structure-based prediction methods is complicated by conformational changes that take place when some proteins bind to their partners. Structure-based methods rely on structural features extracted from the structure in the unbound state or from a bound complex that has been separated into constituent proteins. It is unclear whether such structural features are indeed reliable predictors of interfaces for proteins that undergo significant conformational changes upon binding [153, 140]. Moreover, higher organisms have a large number of intrinsically disordered proteins/regions (IDPs/IDRs) that undergo induced folding only after binding to their partners [49]. Such disordered regions - for which experimental structure information is, by definition, lacking - participate in many important cellular recognition events, and are believed to contribute to the ability of some hub proteins to interact with multiple partners in protein-protein interaction networks [48]. Hence, there is an urgent need for sequence-based methods for reliable prediction of protein-protein interfaces.

Overview of the Chapter

Based on the results of protein interface conservation analysis in chapter 2 we propose HomPPI, a class of sequence homology based approaches to protein interface prediction. We present two variants of HomPPI: (i) NPS-HomPPI (non partner-specific HomPPI), which can

be used to predict interface residues of a query protein in the absence of knowledge of the interaction partner; and (ii) PS-HomPPI (partner-specific HomPPI), which can be used to predict the interface residues of a query protein with a specific target protein. The performance of both HomPPI methods was evaluated on several benchmark datasets, including a large non-redundant set of transient complexes. Due to the increasing importance of intrinsically disordered proteins in understanding molecular recognition mechanics and in rational drug design and discovery [121, 92, 53, 127], we also tested NPS-HomPPI on two datasets of intrinsically disordered proteins.

We compare the performance of HomPPI with that of other web-based servers for interface residue prediction, using several performance measures that assess the reliability of correctly predicting, on average, interface and non-interface residues in a given protein. We discuss the relative advantages and limitations of homology-based methods for interface residue prediction.

3.2 Results

3.2.1 HomPPI - Homologous Sequence-Based Protein-Protein Interface Prediction

Based on the results of our analysis of protein interface conservation described in Chapter 2, we developed HomPPI, a family of sequence homology based algorithms for protein interface prediction. We implemented two variants of HomPPI:

1. NPS-HomPPI - Given a query protein sequence, NPS-HomPPI searches the nr_pdbaa_s2c database (for details see Methods in Chapter 2) to identify homologous proteins that are components of experimentally determined complexes with one or more other proteins. NPS-HomPPI labels a residue of the query sequence as an "interface" residue if a majority of residues in a selected subset of homologs in alignment of the query sequence with its homologs are interface residues, and as "non-interface" residue otherwise. Specifically, given a query protein, we first use NPS-HomPPI to search for sequence homologs within the Safe Zone. If at least one homolog in the Safe Zone is found, NPS-HomPPI uses the Safe homolog(s) to infer the interfaces of the query protein. Otherwise, the process is repeated to search for homologs in the Twilight

Zone or the Dark Zone. If no homologs of the query protein can be identified in any of the three zones, NPS-HomPPI does not provide any predictions. The Safe, Twilight, and Dark Zone homologs of the query protein sequence to be used for interface prediction are identified by searching the nr_pdb_s2c database using BLASTP with thresholds based on the interface conservation analysis (see Methods Section for details) (after removing the query sequence and any highly similar sequences from the same species as the query sequence, in order to allow unbiased evaluation of the performance of NPS-HomPPI).

2. PS-HomPPI - Given the sequences of a query protein A and its putative binding partner B, PS-HomPPI searches the nr_pdbaa_s2c database to identify homologous complexes i.e., the homo-interologs of A-B. PS-HomPPI labels a residue of the query sequence as an "interface" residue (with respect to its putative binding partner) if a majority of the residues in the corresponding position in homologous complexes are interface residues, and as "non-interface" residues otherwise. PS-HomPPI uses homo-interologs in Safe and Twilight Zones to make predictions. The PS-HomPPI prediction process is thus analogous to that for NPS-HomPPI, using thresholds for "close homo-interologs" based on the results of interface conservation analysis of PS-interface conservation (see Methods Section for additional details).

3.2.2 Performance Evaluation of HomPPI Methods

We report several performance measures that provide estimates of the reliability of interface (and non-interface) residue predictions obtained using the HomPPI family of predictors. We compare the performance of HomPPI predictors with several state-of-the-art interface prediction methods on a benchmark dataset. We evaluate the effectiveness of HomPPI in predicting the interface residues of disordered proteins. Finally, we compare the partner-specific and non-partner-specific versions of HomPPI.

We focus our discussion on results using several performance measures that assess the effectiveness of the methods in reliably predicting, on average, the interface and non-interface residues of any given protein (See Methods for details). However, because several of the published studies report performance measures that assess the effectiveness of the methods in reliably assigning interface versus non-interface labels, on average, to any given protein residue, we

also include results using "residue-based" performance measures in Supplementary Materials (See http://homppi.cs.iastate.edu/supplementaryData.html).

(i) NPS-HomPPI Performance on the Benchmark180 Dataset

Among the 180 protein sequences in the Benchmark180 dataset (taken from [20]), 125 sequences had at least one homolog that met the thresholds for the Safe or Twilight Zones, based on zone boundaries determined using Trans135 (Table 3.4). We examined the performance of NPS-HomPPI in predicting interface residues on each of the four different protein complex types in Benchmark180. As shown in Table 3.1, NPS-HomPPI performed best on obligate homodimers, in terms of CC (0.76), sensitivity (0.83), specificity (0.78) and accuracy (0.94). Performance on obligate heterodimers was comparable, although slightly lower. NPS-HomPPI performance on transient interfaces was substantially lower than on obligate interfaces. For transient enzyme inhibitor complexes, the accuracy was 0.86, with a CC of 0.53; for transient non enzyme-inhibitor complexes, the accuracy was 0.83, with a CC of 0.45. These results are consistent with the finding from our statistical analyses in Chapter 2 that NPS-obligate interfaces are more conserved than NPS-transient interfaces in their homologs.

We also evaluated the prediction performance of NPS-HomPPI using homologs with different degrees of sequence homology. In Table 3.2, the prediction performance is shown separately for sets of test proteins for which HomPPI can identify at least one homolog in Safe, Twilight, or Dark Zones. As expected, Safe Zone homologs consistently gave the most reliable prediction performance for all four types of complexes (CC values ranged from 0.55 to 0.84). Both obligate and transient interfaces were predicted with moderate to high reliability (CC values ranged from 0.12 to 0.67) even using only distant homologs from the Twilight or Dark Zones.

(ii) Comparison of NPS-HomPPI with other PPI Prediction Servers

Direct comparison of NPS-HomPPI with other methods described in the literature is complicated by the limited availability of implementations of the underlying methods (many of which are available only in the form of servers), and differences in the choice of training and evaluation datasets, evaluation procedures and evaluation measures [21]. Hence, we limit our comparisons

Table 3.1 Interface Residue Prediction Performance of NPS-HomPPI on Benchmark180.

Binding Type	**Homology Zone**	**Prediction coverage**[a]	CC^P	$Sensitivity^P$	$Specificity^P$	$Accuracy^P$
Enzyme-inhibitor,-Transient	Safe/Twilight	67% (24/36)	0.53	0.67	0.58	0.86
Non-enzyme-inhibitor -Transient (NEIT)	Safe/Twilight	60% (18/30)	0.45	0.54	0.58	0.83
Hetero-dimer - Obligate	Safe/Twilight	85% (23/27)	0.63	0.72	0.69	0.88
Homo-dimer - Obligate	Safe/Twiight	69% (60/87)	0.76	0.83	0.78	0.94

[a]Prediction coverage reflects that predictions are made only on proteins for which HomPPI can identify Safe/Twilight Zone homologs.

Table 3.2 Prediction Performance of NPS-HomPPI using Homologs from the Safe, Twilight, Dark Zones.

Binding Type	**Homolog Zone**	**Prediction coverage**	CC^P	$Sensitivity^P$	$Specificity^P$	$Accuracy^P$
Enzyme-inhibitor, -Transient	Safe	14% (5/36)	0.55	0.62	0.57	0.94
	Twilight	50% (18/36)	0.52	0.69	0.58	0.83
	Dark	19% (7/36)	0.12	0.23	0.20	0.83
	Total	**83% (30/36)**	**0.44**	**0.58**	**0.50**	**0.85**
Non-enzyme-inhibitor, -Transient (NEIT)	Safe	23% (7/30)	0.56	0.64	0.60	0.91
	Twilight	37% (11/30)	0.37	0.48	0.57	0.78
	Dark	33% (10/30)	0.36	0.37	0.50	0.86
	Total	**93% (28/30)**	**0.42**	**0.48**	**0.55**	**0.84**
Hetero-dimer, - Obligate	Safe	52% (14/27)	0.70	0.81	0.72	0.91
	Twilight	30% (8/27)	0.52	0.58	0.64	0.82
	Dark	15% (4/27)	0.44	0.66	0.47	0.80
	Total	**96% (26/27)**	**0.60**	**0.71**	**0.66**	**0.86**
Homo-dimer, - Obligate	Safe	38% (33/87)	0.84	0.90	0.84	0.96
	Twilight	31% (27/87)	0.67	0.74	0.71	0.91
	Dark	28% (24/87)	0.36	0.47	0.44	0.84
	Total	**97% (84/87)**	**0.65**	**0.73**	**0.68**	**0.91**

of HomPPI with five state-of-the-art methods available as web-based servers: Promate [99], Cons-PPISP [26, 155], meta-PPISP [112], PIER [78] and PSIVER [97]. All of these methods except PSIVER take advantage of both sequence and experimentally determined protein structure of the query proteins. They have been reported to be among the best performing methods currently available for predicting PPIs (see [153][37] for reviews). PSIVER is one of the most recently published methods for interface residue prediction that only uses protein sequence-derived information. Although direct comparisons of the data representation and the algorithms used by PSIVER with those used by other sequence-based interface residue predictors are currently not available, PSIVER has been reported to outperform two other sequence-based servers: ISIS [104] and the sequence-based variant (made available as an experimental version in 2008) of SPPIDER [111].

Promate samples the protein surface using circular patches around a set of anchoring dots and estimates the probability that each surface dot belongs to an interface, based on the distribution of various physicochemical properties within interface and non-interface patches. Cons-PPISP is a consensus method that combines six neural networks trained on six datasets. Meta-PPISP is a consensus method that combines the output from cons-PPISP, Promate, and PINUP [83]. PIER relies on partial least squares (PLS) regression of surface patch properties of the query protein. PSIVER uses PSSM profiles and predicted solvent accessibility as input features, and uses a Naïve Bayes classifier with parameters obtained using kernel density estimation. Because NPS-HomPPI does not take structural information into account, to compare its performance with the structure-based servers, we mapped the interfaces predicted by each server onto the full sequence of each query protein in order to evaluate prediction performance on the entire protein sequence.

We compared the performance of NPS-HomPPI with all five PPI servers on a subset of the Benchmark180 dataset, specifically, 125 out of 180 proteins for which NPS-HomPPI was able to identify homologs in the Safe or Twilight zones. The sensitivity-specificity plots (also called precision-recall plots) are shown in Figure 3.1. Each data point corresponds to a different classification threshold value. The prediction score of NPS-HomPPI is simply the normalized vote (for each residue total votes for interfaces from homologs are normalized by the number

of homologs) from 10 (or fewer available) homologs. Thus, NPS-HomPPI produces a limited number of distinct prediction scores.

For the two transient complex types, enzyme-inhibitors (Figure 3.1a) and transient non-enzyme-inhibitors, transient (Figure 3.1b), NPS-HomPPI consistently outperforms Promate, PIER, meta-PPISP, cons-PPISP, and PSIVER except for sensitivity values lower than 0.2 (which is very low to be useful in practice). On both obligate heterodimers (Figure 3.1c) and homodimers (Figure 3.1d), NPS-HomPPI outperforms all five servers across the full range of sensitivity and specificity values for which it can generate homology-based predictions. It should be noted that structure-based methods predict which surface residues are interface residues. In contrast, sequence-based methods have the more challenging task of identifying interface residues from the set of all residues. In other words, structure-based methods can trivially eliminate all non-surface residues from the set of candidate interface residues. Viewed in this light, the observed predictive performance of NPS-HomPPI, a purely sequence-based method, suggests that it is possible to make reliable non-partner-specific interface residue predictions using only the sequences of a protein by taking advantage of the conservation of interfaces in the context of non-partner-specific interactions.

(iii) Performance of NPS-HomPPI on Intrinsically Disordered Proteins

Intrinsically disordered proteins (IDPs) and proteins containing intrinsically disordered regions (IDRs) are attractive targets for drug discovery [92]. The lack of defined tertiary structure in IDPs/IDRs poses a major challenge to structure-based interface prediction methods. Hence, we compared the performance of NPS-HomPPI with ANCHOR [45], a recently published method for the prediction of binding regions in disordered proteins. For this comparison, we used two non-redundant disordered protein datasets, S1 and S2, recently collected by Meszaros et al. [91]. Some of the test proteins are based on data from NMR structures. In order to compare NPS-HomPPI with ANCHOR on the largest possible number of cases available to us, we extracted interface residues from these NMR cases; however, we used only sequence homologs with interface residues determined from X-ray structures to make predictions.

Figure 3.2 shows the performance comparison of NPS-HomPPI with ANCHOR on the pre-

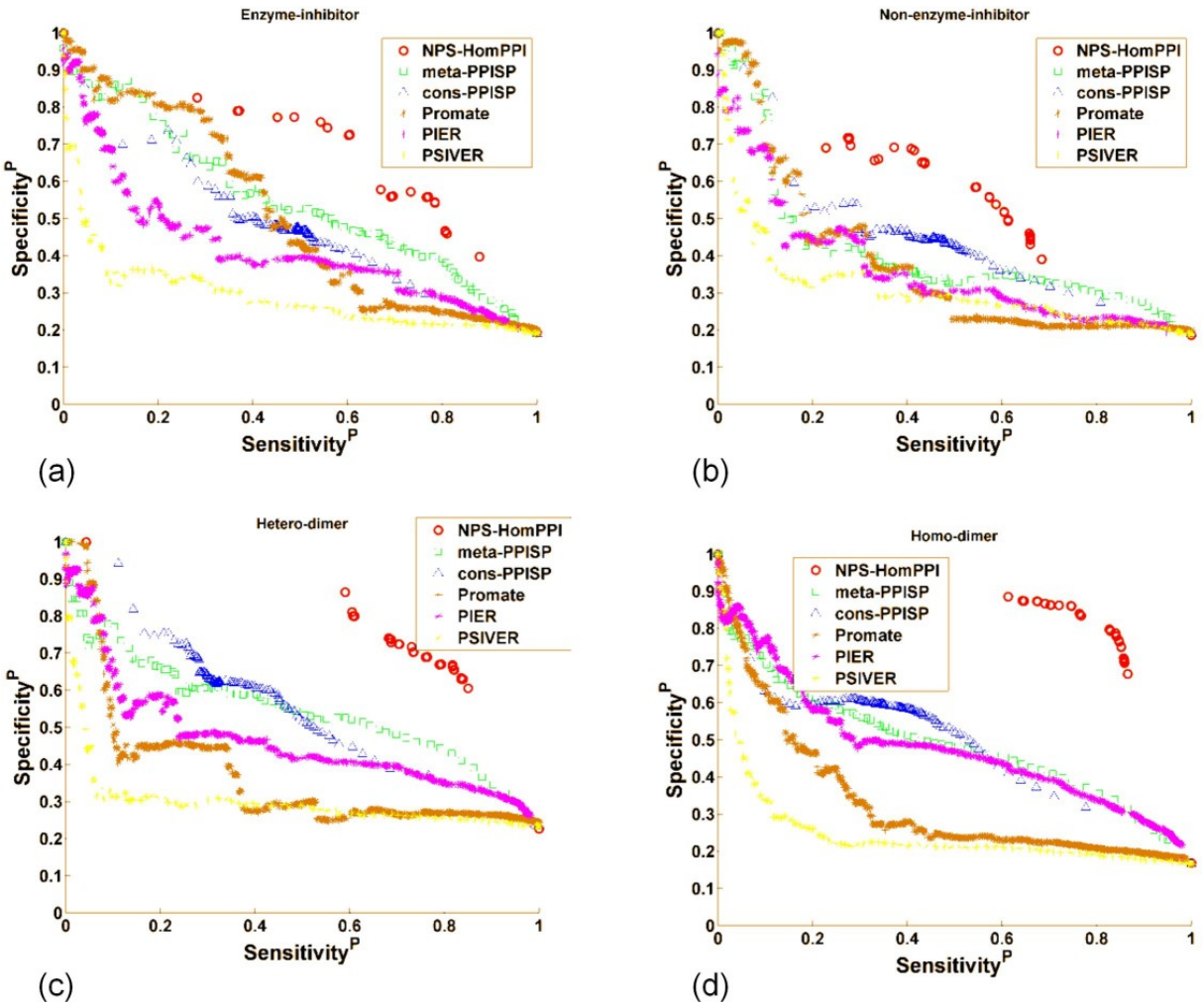


Figure 3.1 Performance of NPS-HomPPI Compared with Web-based PPI Servers. Performance was evaluated on four different protein complex types from Benchmark180: (a) Enzyme-inhibitors, transient. (b) Non-enzyme-inhibitors (NEIT), transient. (c) Hetero-dimers, obligate. (d) Homo-dimers, obligate. Servers compared were: NPS-HomPPI: red circles; Meta-PPISP: green squares; Cons-PPISP: blue triangles; Promate: brown stars; PIER: purple stars; PSIVER: yellow stars.

diction of interface residues in disordered proteins. NPS-HomPPI significantly outperforms ANCHOR over a broad range of sensitivity and specificity for both short as well as long disordered proteins for which sequence homologs are available in Safe, Twilight or Dark Zones (Figure 3.2 a and b respectively). For example, as shown in Figure 3.2 b, on the S2 dataset, at a prediction sensitivity value of 0.70, ANCHOR achieves a specificity of ~0.40, whereas NPS-HomPPI achieves a specificity of ~0.64.

At present, NPS-HomPPI has relatively high prediction coverage for long disordered proteins (78%; 31 out of 40 interfaces of disordered proteins), but lower coverage for short disordered proteins (50%; 28 out of 56 interfaces of disordered proteins). This is in part due to that fact that many disordered proteins available in the PDB (Protein Data Bank) [14] have only NMR structures, which were excluded from the current study. Incorporation of data from NMR structures in the future can be expected to increase the coverage of NPS-HomPPI for disordered proteins.

(iv) Performance of NPS-HomPPI versus PS-HomPPI

Our analysis of the conservation of PS-transient interfaces described in Chapter 2 suggests that many interfaces in transient protein complexes are highly partner-specific. Thus, we implemented a variant of HomPPI, designated PS-HomPPI, to evaluate the possibility that prediction of interface residues, especially in transient complexes, can be improved by using sequence information about specific binding partners, when available.

We first evaluated the performance of PS-HomPPI on a transient complex dataset, Trans135 (dimers from the dataset in [94]). PS-HomPPI found at least one homo-interolog that meets the Safe or Twilight similarity thresholds for 60% (162/270) proteins in the Trans135 dataset. Overall, PS-HomPPI had an average CC of 0.65, sensitivity of 0.69, specificity of 0.70 and accuracy of 0.92.

To investigate whether the partner information is, in fact, helpful in predicting interfaces we directly compared the performance of PS-HomPPI with NPS-HomPPI on the Trans135 dataset. In Trans135, there were 139 out of 270 chains that for which predictions could be generated by both NPS-HomPPI (using homologs) and PS-HomPPI (using homo-interologs) from the Safe

or Twilight zones (see Methods for details).

The results shown in Figure 3.3 indicate that, at least for transient interfaces in the Trans135 dataset, PS-HomPPI outperforms NPS-HomPPI. Although the average values over proteins (green dots) for CC, sensitivity and specificity are similar, the median values (the red bar in the box) for PS predictions (left panel) are much higher than that for NPS predictions (right panel). Also, the observed variance (length of the box) of PS predictions (left panel) is much smaller than that of NPS predictions (right panel). These results suggest that the reliability of interface residue predictions can be improved by exploiting the knowledge of the binding partner of a query protein.

3.3 Discussion

3.3.1 Performance of HomPPI Compared with Published Methods

Our results show that whenever the interfaces of the close sequence homologs of a query protein are available, NPS-HomPPI outperforms several state-of-the-art protein interface prediction servers (many of which take advantage of the structure of the query protein), over a broad range of sensitivity and specificity values. In the case of transient complexes (Figure 3.1 a and b), NPS-HomPPI consistently outperforms Promate, PIER, meta-PPISP, cons-PPISP, and PSIVER except for sensitivity values lower than 0.2. On obligate dimers (Figure 3.1 c and d), NPS-HomPPI significantly outperforms all five servers across the full range of sensitivity and specificity values for which it can generate homology-based predictions. These results strongly suggest that it is possible to reliably predict protein interface residues using only sequence information whenever the interface residues of sequence homologs of the query protein are known. Each of the webbased PPI servers with which we compared our NPS-HomPPI server, except PSIVER, take advantage of the structure of the query proteins to determine surface residues, and restrict the predicted interface residues to a subset of the surface residues. This trivially reduces the number of false positive interface residue predictions (relative to the total number of residues in the query protein) which, in turn, yields a substantial increase in the specificity of interface predictions produced by structure-based servers. Consequently, purely sequence-

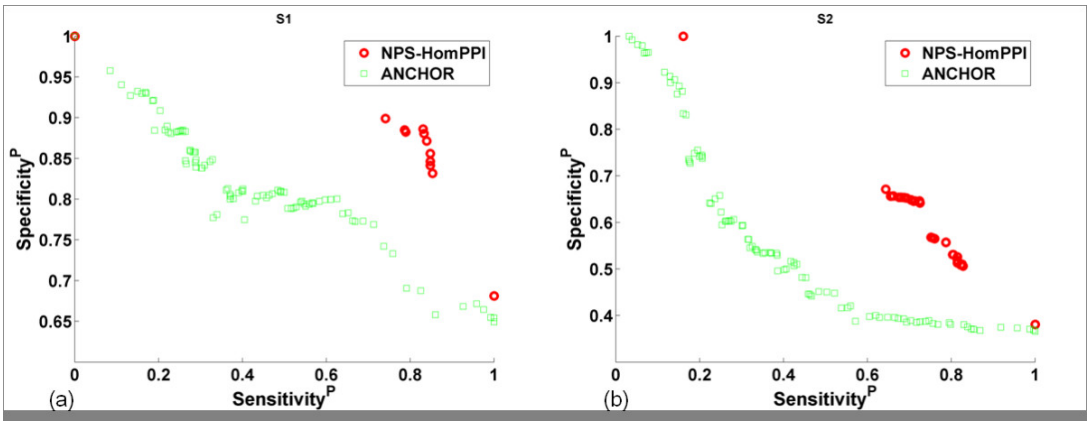


Figure 3.2 Performance of NPS-HomPPI Compared with ANCHOR in Predicting Interface Residues in Disordered Proteins. Two datasets of disordered proteins were used: (a) S1: short disordered proteins. (b) S2: long disordered proteins. NPS-HomPPI: red circles; ANCHOR: green squares.

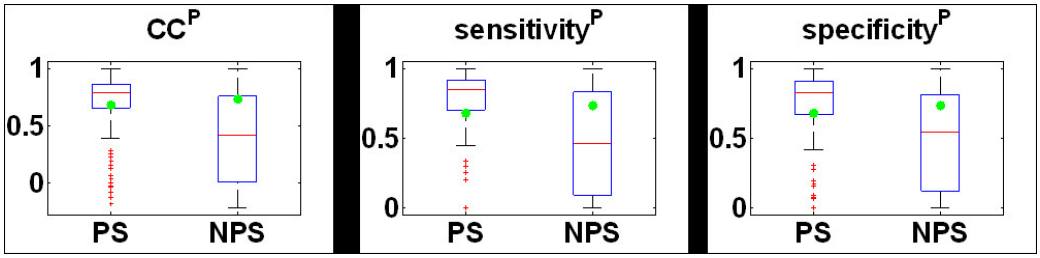


Figure 3.3 Performance Comparison of PS-HomPPI and NPS-HomPPI. Only proteins for which predictions could be generated by both PS-HomPPI and NPS-HomPPI (139 out of 270 chains from Trans135) were used in this evaluation. The lower (Q1), middle (Q2) and upper (Q3) quartiles of each box are 25th, 50th and 75th percentile. Interquartile range IQR is Q3-Q1. Any data value that lies more than 1.5 $\times$ IQR lower than the first quartile or 1.5 $\times$ IQR higher than the third quartile is considered an outlier, which is labelled with a red cross. The whiskers extend to the largest and smallest value that is not an outlier. Averages are marked by green dots.

based protein interface prediction servers have a handicap relative to structure-based prediction servers. When viewed in this light, performance of NPS-HomPPI relative to the state-of-the-art protein interface prediction methods is especially impressive.

The HomPPI methods for interface residue prediction do have an important limitation, however, in that they rely on the availability putative homologs for which experimentally-determined structures of bound complexes are available in the PDB. One may ask whether the coverage of the HomPPI family of protein-protein interface prediction methods is broad enough to be sufficiently useful in practice. We address this question below.

3.3.2 Prediction Coverage of HomPPI Methods

The current coverage of HomPPI protein interface prediction methods can be assessed from our results as follows:

NPS-HomPPI

• Benchmark180 dataset: NPS-HomPPI found at least one homolog that meets the similarity thresholds for Safe or Twilight Zones for 73% (83/114) of the obligate binding chains (homo and hetero-dimers). Among these, 82% (68/83) were predicted with both sensitivity and specificity $\geq$0.50, simultaneously. Similarly, at least one homolog was found for 62% (42/66) of transient binding chains (enzyme-inhibitors and non-enzyme inhibitors) in this dataset. Among these 55% (23/42) were predicted with both sensitivity and specificity $\geq$0.5.

• Trans135 dataset: In the case of transient query proteins in the Trans135 dataset, NPS-HomPPI found at least one homolog that meets the similarity thresholds for Safe or Twilight Zones for 75% (202/270) of chains. Among these, 37% (74/202) were predicted with both sensitivity and specificity $\geq$0.5.

• Disordered protein datasets S1 and S2: In the case of disordered proteins, NPS-HomPPI found at least one homolog that meets the similarity thresholds for Safe or Twilight or Dark Zones for 50% (26/52) of interfaces of disordered proteins in S1, the short disordered protein set, and 75% (30/40) of interfaces of disordered proteins in S2.

PS-HomPPI

• Trans135 dataset: PS-HomPPI found at least one homo-interolog that meets the Safe

or Twilight similarity thresholds for 60% (162/270) proteins in the Trans135 dataset. Among these, 80% (130/162) where predicted with sensitivity and specificity ≥0.5, simultaneously.

Based on these results, we estimate that, at present, the coverage of the HomPPI protein interface prediction methods is in the range of 60-70% of all query proteins. As the structural genomics projects currently underway generate increasing numbers of structures of protein-protein complexes [25], we can expect corresponding increases in the coverage of HomPPI family of protein interface prediction methods. In the meantime, one can envision hybrid methods that combine HomPPI with one or more machine learning based methods that do not require the availability of putative homologs for which experimentally determined structures of bound complexes are available in the PDB.

3.3.3 Parameters for HomPPI Can Be Relaxed for Obligate Interactions

The current default parameters for HomPPI are intentionally rather stringently set based on the results of our statistical analysis of interface conservation using Trans135, which is a dataset of transient binding proteins. Our analyses suggest that NPS-HomPPI has wider Safe and Twilight Zones for obligate binding proteins than for transient binding proteins. Furthermore, even Dark Zone homologs yield interface predictions that are accurate enough to be useful in practice, with average specificity of 0.47 and sensitivity of 0.66 for hetero-obligate dimers, average specificity of 0.44 and sensitivity of 0.47 for homo-obligate dimers (see Table 3.2). Therefore, for obligate interactions, if a query protein has little sequence similarity with proteins in the PDB, the thresholds of NPS-HomPPI can be relaxed to allow identification of more distant homologs with potentially conserved interfaces that still provide reliable interface predictions.

3.3.4 Prediction of Binding Partners vs. Prediction of Interface Residues

Protein interface (binding site) predictions and protein interaction (partner) predictions answer closely related, but different questions. Non partner-specific protein interface predictors are designed to identify the residues in a query protein that are likely to make contact with the residues of one or more unspecified interaction partner proteins. Partner-specific protein interface predictors are designed to identify the residues in a query protein that are likely

to make contact with residues of a putative interaction partner protein. In contrast, protein interaction predictors are designed to predict whether or not a given pair of proteins is likely to interact [106, 65, 31, 131]. Although our study does not directly address the latter question, it is possible to use PS-HomPPI predictions to determine whether or not two query proteins interact: Given a pair of protein sequences, say A and B, we can first use PS-HomPPI to predict the interface residues of A with its putative partner B; and the interface residues of B with its putative partner A. If, in both cases, some number of interface residues are predicted, we can infer that proteins A and B are likely to interact with each other. Conversely, it is possible to use information from predicted protein-protein interactions to refine interface predictions. Yip et al. [149] have proposed an approach to utilize residue level information to improve the accuracy of protein level predictions, and vice versa. They have shown that a two-level machine learning framework that allows information flow between the two levels through shared features yields predictions that are more accurate than those obtained independently at each of the levels.

3.3.5 Using Interface Predictions to Steer Docking and to Rank Docked Conformations

Reliable partner-specific interface predictions can be used to restrict the search space for protein-protein docking by specifying the contacts that need to be preserved in the docked conformation. It is also possible to rank the conformations produced by docking, based on the degree of overlap between the interface of a query protein and its binding partner in the docked conformation with the interface generated by a partner-specific interface prediction method, e.g. PS-HomPPI. In related work [143], we have shown that PS-HomPPI provides reliable interface predictions on a large subset of a Docking Benchmark Dataset, and is both fast and robust in the face of conformational changes induced by complex formation. The quality of the ranking of docked conformations by PS-HomPPI interface prediction is consistently superior to that produced using ClusPro cluster-size-based and energy-based criteria for 61 out of 64 docking complexes for which PS-HomPPI produces interface predictions [143].

3.4 Conclusions

Based on the results of our interface conservation analyses, we developed HomPPI, a simple sequence-based method for predicting interface residues based on the known interface residues in homologous sequences. HomPPI has two variants: NPS-HomPPI (for predicting interface residues of a query protein with unspecified interaction partners) and PS-HomPPI (for predicting interface residues of query proteins with a specified putative interaction partner).

Our systematic evaluation of NPS-HomPPI showed that, when close homologs can be identified, NPS-HomPPI can reliably predict interface residues in both obligate and transient complexes, with a performance that rivals several state-of-the-art structure-based interface prediction servers. NPS-HomPPI can also be used as a reliable tool for identifying disordered binding regions. In this regard, NPS-HomPPI has an advantage over structure-based interface predictors, which cannot be used to predict binding sites in disordered regions of proteins because they do not form stable structures in their unbound state. In addition, the HomPPI family of interface prediction methods are fast enough for proteome-wide analyses.

Many studies on in silico identification of protein interfaces have been published in the past decade. However, despite the fact that many proteins are very specific in their choice of binding partners, the majority of studies focus on only one side of the bound complex. In this study, we implemented a novel partner-specific protein interface prediction method, PS-HomPPI, which infers interface residues based on known interfaces in the homo-interologs, i.e., complexes formed by homologs of the query protein and its putative interaction partner. When homo-interologs can be identified, PS-HomPPI can reliably predict highly partner-specific transient interfaces.

Although our focus in this study was on prediction of protein-protein interfaces, these methods could be useful in other settings, such as sequence-based prediction of protein-DNA, protein-RNA, and protein-ligand interfaces, and the prediction of B and T cell epitopes.

Both NPS-HomPPI and PS-HomPPI have been implemented in a server available at: *http* : *//homppi.cs.iastate.edu/*.

3.5 Methods

3.5.1 Datasets

Three datasets were used in this chapter:

• Benchmark180 - For evaluating the prediction performance of HomPPI.

• S1 and S2 - For evaluating the performance of NPS-HomPPI on interfaces of disordered proteins.

• Trans135 - For comparing the perfomrance of NPS-HomPPI and PS-HomPPI on the highly partner-specific interfaces of transient interactions.

Benchmark180

We tested NPS-HomPPI on a benchmark dataset manually collected and used as evaluation dataset by Bradford and Westhead [20]. This dataset consists of 180 protein chains taken from 149 complexes; 36 of these are involved in enzyme-inhibitor interactions, 27 in hetero-obligate interactions, 87 in homo-obligate interactions, and 30 in non-enzyme-inhibitor transient (NEIT) interactions.

Disordered protein datasets S1 and S2

We evaluated the performance of NPS-HomPPI on a non-redundant disordered dataset that has been recently collected by Meszaros et al. [91]. S1 consists of 46 complexes of short disordered and long globular proteins. S2 consists of 28 complexes of long disordered and long globular proteins. Note that a protein complex e.g., 1fv1 C:AB formed by a disordered protein C with two ordered proteins A and B, yields two sets of interface residues for C (corresponding to interfaces between C with A and C with B). As a result, 46 complexes in S1 and 28 complexes in S2 (respectively) correspond to 56 and 40 interfaces of disordered proteins. We focused on cases in which NPS-HomPPI is able to identify Safe/Twilight/Dark zone homologs for the query proteins resulting in NPS-HomPPI interface predictions for 28 out of 56 and 31 out of 40 interfaces of disordered proteins in S1 and S2 respectively.

Trans135

This dataset of 135 transient dimer complexes was obtained from a large non-redundant dataset of 212 transient complexes (3.25 Å or better resolution, determined using X-ray crystallography) previously generated by Mintseris and Weng [94] to study the conservation of protein-protein interfaces. In ordered to exclude the influence of other types of interfaces, we extracted 135 transient dimers from the original dataset and get Trans135. In Trans135, 1DN1 and 1IIS have been superseded by 3C98 and 1T83, respectively, and 1F83, 1DF9, 4CPA and 1JCH have since been deemed as obsolete and hence discarded from the PDB.

3.5.2 Interface Definition

This paper adopts a stringent definition of protein-protein interfaces. Surface residues are defined as residues that have the relative solvent accessible area (RASA) at least 5% [111]. Interface residues are defined as surface residues with at least one atom that is within a distance of 4 Å from any of the atoms of residues in the chain. The ratios of interface residues versus the total number of residues for the datasets used in this work are summarized in Table 3.3. Interface information was extracted from the ProtInDB server $http://protInDB.cs.iastate.edu$.

Table 3.3 The Proportion of Interface Residues in Datasets Used in The Evaluation of HomPPI.

Dataset	Number of Interface Residues[a]	Total Number of Residues[b]	% Interface Residues
Benchmark180	6,401	43,013	14.90%
Trans135	6,460	55,217	11.70%
Disordered S1[c]	585	1,171	50.00%
Disordered S2	1,797	11,400	15.80%

[a]When a chain interacts with more than one other chain, the interfaces are counted separately. For example, for protein complex 2phe C:AB, the interface of C with A and the interface of C with B are regarded as two disordered interfaces.

[b]Residues that missing from PDB structures are not counted.

[c]For disordered interface datasets S1 and S2, only interfaces of IDPs are counted. Interfaces of IDPs' binding parnters are not counted.

3.5.3 Mapping Interfaces in Structures to Sequences

We label the protein sequences as interface or non-interface residues (according to the definition of interface residues given above) as follows: We first calculate the relevant distances between atoms using the atom coordinates in ATOM section in PDB files. Then, by associating the ATOM section to residues in the SEQRES section, we can map the corresponding residues to protein sequences. However, various errors in PDB files make this a non-trivial task. Hence, we used the mapping files from S2C database, which offers corrected mapping information from ATOM section to residues in the SEQRES section of PDB files, to map interfaces determined in structures to full sequences.

3.5.4 NCBI BLAST Parameters

In this study, we used the substitution matrices and gap costs recommended for the different query lengths (See Table 2.4).

3.5.5 Performance Evaluation

To evaluate the extent to which protein interfaces are conserved in query-homolog pairs and to estimate the performance of HomPPI and other predictors that we compare with in predicting the interface residues of a novel protein (i.e., one not used to train the predictor), we consider several standard performance measures including sensitivity (recall), specificity (precision), accuracy and Matthews correlation coefficient (CC) [12]. Specifically, for each test protein i, we calculate the corresponding performance measures for each protein i as follows:

$$sensitivity_i = \frac{TP_i}{TP_i+FN_i}$$

$$specificity_i = \frac{TP_i}{TP_i+FP_i}$$

$$accuracy_i = \frac{TP_i+TN_i}{TP_i+FP_i+FN_i+TN_i}$$

$$CC_i = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i+FN_i)(TP_i+FP_i)}}$$

where TP_i, FP_i, TN_i, and FN_i are respectively the number of interface residues of protein i that are correctly predicted to be interface residues, the number of residues of protein i that

are incorrectly predicted to be interface residues, the number of residues of protein i that are correctly predicted to be non-interface residues, and the number of residues of protein i that are incorrectly predicted to be non-interface residues.

We calculate the protein-based overall performance measures as follows:

$$sensitivity^P = \frac{\sum_{i=1}^{N} sensitivity_i}{N}$$

$$specificity^P = \frac{\sum_{i=1}^{N} specificity_i}{N}$$

$$accuracy^P = \frac{\sum_{i=1}^{N} accuracy_i}{N}$$

$$CC^P = \frac{\sum_{i=1}^{N} CC_i}{N}$$

where N is the total number of test proteins.

These measures describe different aspects of predictor performance. The overall sensitivity is the probability, on average, of correctly predicting the interface residues of a given protein. The overall specificity is the probability, on average, that a predicted interface residue in any given protein is in fact an interface residue. The overall accuracy corresponds to the fraction of residues in any given protein, on average, that are correctly predicted. The overall Matthews correlation coefficient measures of how predictions correlate, on average, with true interfaces and non-interfaces.

Often it is possible to trade off one performance measure (e.g., specificity) against another (e.g., sensitivity) by varying the threshold that is applied to the prediction score to generate the binary (interface versus non-interface) predictions. Hence, we include of the overall sensitivity against overall specificity for different choices of the threshold. The resulting specificity-sensitivity plots or precision-recall plots show the trade-off between sensitivity and specificity and hence provide a much more complete picture of predictive performance.

The performance measures described above provide an estimate of the reliability of the predictor in predicting interface residues of a novel protein. It is worth noting that most of the papers in the literature on interface residue prediction report performance measures by averaging over residues (as opposed to proteins). The residue-based overall performance measures are calculated as follows:

$$sensitivity^R = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FN_i)}$$

$$specificity^R = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N}(TP_i+FP_i)}$$
$$accuracy^R = \frac{\sum_{i=1}^{N}(TP_i+TN_i)}{\sum_{i=1}^{N}(TP_i+FP_i+FN_i+TN_i)}$$
$$CC^R = \frac{\sum_{i=1}^{N} TP_i \times \sum_{i=1}^{N} TN_i - \sum_{i=1}^{N} FP_i \times \sum_{i=1}^{N} FN_i}{\sqrt{\sum_{i=1}^{N}(TP_i+FN_i) \times \sum_{i=1}^{N}(TP_i+FP_i) \times \sum_{i=1}^{N}(TN_i+FP_i) \times \sum_{i=1}^{N}(TN_i+FN_i)}}$$

Residue-based specificity-sensitivity plots in this case show how the trade-off between specificityR and specificityR is obtained by varying the threshold applied to the prediction score. The residue-based performance measures provide an estimate of the reliability of the predictor in correctly labelling a given residue. However, in practice, it is useful to know how well a predictor can be expected to perform on a given protein sequence as opposed to a residue. $sensitivity^P$, $specificity^P$, $accuracy^P$, and CC^P are more informative than their residue-based counterparts. Hence, in this paper, we report results based on the protein-based measures although, for the purpose of comparison with other published methods, we include the results based on the residue-based measures in Supplementary Materials in HomPPI website.

3.5.6 NPS-HomPPI

NPS-HomPPI is a Non-Partner-Specific Homologous Sequence-Based Protein-Protein Interface Prediction algorithm. NPS-HomPPI is based on the conclusion from statistical analysis of protein interface conservation on several non-redundant large datesets - Nr6505, Trans135 and Oblig94 (for details of these datasets see Methods in Chapter 2), i.e., that protein interfaces are conserved across close sequence homologs.

As illustrated in Figure 3.4, NPS-HomPPI predicts interface residues in a query protein based on the known interface residues of a selected subset of homologs in a sequence alignment. Homologs of the query protein sequence are identified by searching the nr_pdb_s2c database using BLASTP. Note that, in our experiments, in order to allow unbiased evaluation of the performance of NPS-HomPPI, the query sequence itself and sequences that share a high degree ($\geq$95%) of amino acid sequence identity with, and are from the same species as the query sequence are deleted from the set of putative homologs.

If at least one homolog in the Safe Zone is found by the BLASTP search, NPS-HomPPI uses the Safe Zone homolog(s) to infer the interfaces of the query protein. Otherwise, the search is

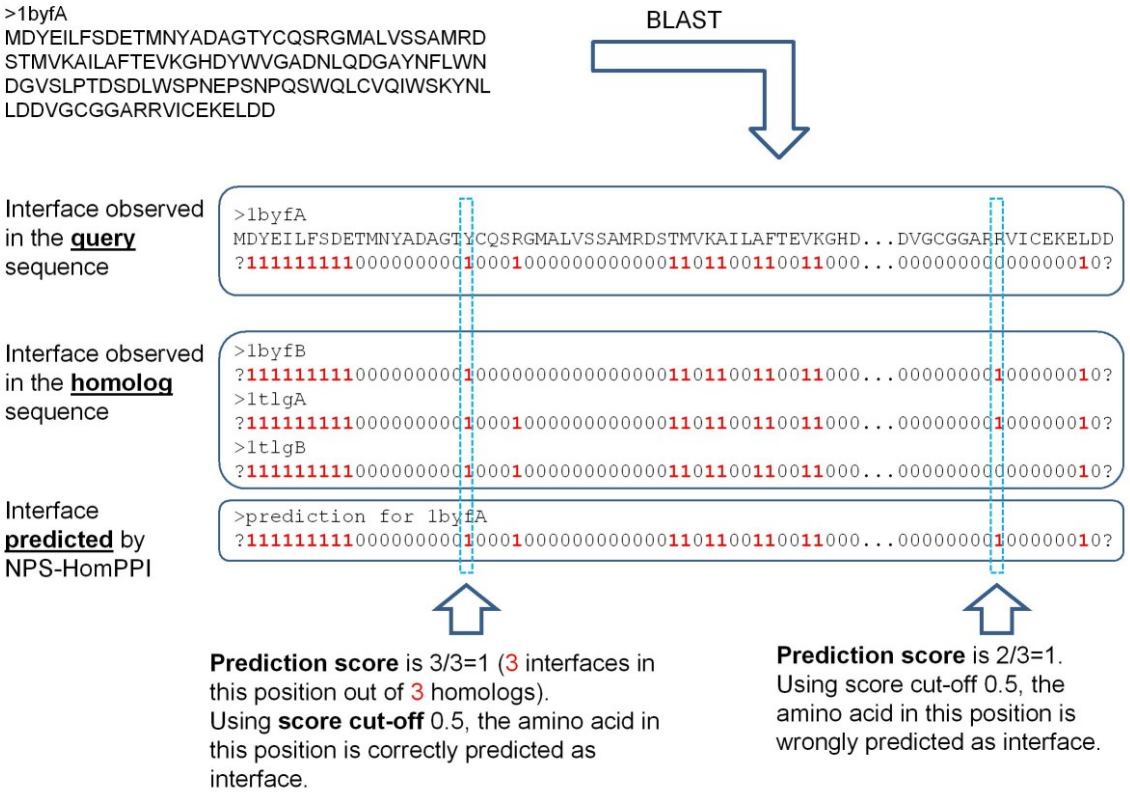


Figure 3.4 An Example of Interface Residue Prediction using NPS-HomPPI. The sequence of the query protein 1byf chain A is BLASTed against *nr_pdbaa_s2c* database. In this case, 3 sequences meet the thresholds set by NPS-HomPPI for "close homolog" in Safe Zone or Twilight Zone defined in Table 3.4 . If there are more than K = 10 homologs met the zone thresholds in Table 3.4, regression equation 2.2.1.5 is used to determine the nearest K homologs for final prediction. For each position in the alignment, an amino acid residue in the query sequence is predicted to be an interface residue if the majority of the amino acid residues in the alignment are interface residues. Otherwise, it is predicted to be a non-interface residue. Interface residues are denoted by red 1's; Non-interface residues are denoted by black 0's. Question marks denote residues for which coordinates are missing from PDB files.

repeated for homologs in the Twilight and Dark Zones. If NPS-HomPPI cannot find homologs in any of the three zones, it does not provide any predictions. The default zone boundaries used by NPS-HomPPI (and hence the parameters used in NPS-HomPPI search for homologs of a query sequence) is based on our interface conservation analysis on the dataset of transient dimers Trans135 (Table 3.4). The choice of these default parameter thresholds for NPS-HomPPI is intentionally rather conservative; the thresholds can be relaxed if additional information is available (e.g., if we know that the query protein is an obligate binding protein). The IC score of each of the homologs of a query sequence in the alignment returned by BLASTP is predicted using the regression model for the IC score (see eq. 2.2.1.5) from the BLASTP statistics for the alignment of each homolog with the query sequence. For a given query sequence, at most K closest (Safe, Twilight, or Dark Zone homologs, as the case may be, in that order) are selected from the alignment of the query sequence with its homologs to be used to infer the interface residues of the query sequence. In our experiments, K, the maximum number of homologs used in the prediction was set equal to 10. At most K homologs of the query sequence are determined by ranking the homologs in the alignment in decreasing order of their predicted IC scores and choosing (at most) K Safe zone homologs (or Twilight zone homologs if no Safe zone homologs exist or Dark zone homologs if neither Safe nor Twilight zone homologs exist). Once the (at most) K closest homologs to be used for predicting the interface residues of the query sequence are chosen, each residue in the query sequence is labelled as an interface or non-interface residue based on the majority (over the set of at most K closest homologs of the query sequence) of the labels associated with the corresponding position in the alignment. More specifically, each of the at most K homologs provides a positive vote for a given position in the query sequence if the corresponding residue of the homolog is an interface residue; and a negative vote if it is a non-interface residue. The prediction score of NPS-HomPPI for that position in the query sequence is simply the number of positive votes divided by the total number of votes. A query sequence residue with a HomPPI score ≥ 0.5 is predicted to be an interface residue (See Figure 3.4 for an example); otherwise, it is predicted to be a non-interface residue. This procedure can be seen as an application of the (at most) K nearest neighbor classifier at each residue of the query sequence.

3.5.7 PS-HomPPI

PS-HomPPI predicts the interface residues in a protein chain based on the known interface residues of its closest homo-interologs. Given a query protein A and its interaction partner B, PS-HomPPI first identifies the set homo-interologs of A-B using BLASTP to identify the homologs of A and homologs of B. From the BLASTP results, we identify a set of homo-interologs that meet sequence similarity thresholds (determined based on the results of our partner-specific interface conservation analysis, as described in the Results Section). We discard the whole PDB complex that contains A-B, to ensure an objective assessment of the reliability of our prediction procedure. For query A-B and its homologous interacting pair A'-B', we also discard the interacting protein pair A'-B' if A and A' or B and B' share $\geq$95% sequence identity and belong to the same species.

PS-HomPPI uses homo-interologs in the Safe and Twilight Zones to make predictions. The zone boundaries were determined using Trans135 and are shown in Table 3.5. The PS-HomPPI prediction process is similar to that of NPS-HomPPI in that it progressively searches for homointerologs from higher, then lower, homology zones: i.e., if PS-HomPPI cannot find at least one homo-interolog in the Safe Zone, it next looks for homo-interologs in the Twilight Zone.

PS-HomPPI predicts whether an amino acid in query sequence A is an interface residue or not based on the corresponding position in its alignment with (at most) K of the closest homo-interologs of A-B (based on their predicted IC scores). In our experiments, K was set equal to 10. Given a query-partner pair A-B, we label each position in the amino acid sequence of protein A as an interface or non-interface based on whether or not a majority of the corresponding positions of the homologs of A within the homo-interologs of A-B are interface residues. More specifically, each of the at most K homo-interologs provides a positive vote for a given position in the query protein sequence A if the corresponding residue of its homolog A' in its homo-interolog is an interface residue; and a negative vote if it is a non-interface residue. The prediction score of PS-HomPPI for that position in the query sequence is simply the number of positive votes divided by the total number of votes. A residue in the query protein A with a prediction score $\geq$0.5, is predicted as interface, otherwise, it is predicted as non-interface.

Table 3.4 Boundaries of Safe, Twilight and Dark Zones used by NPS-HomPPI.

Zones	**Sequence Similarity**	**Thresholds**[a]
Safe Zone	$log(EVal)$	≤ -100
	$Positive\,Score$	$\geq 80\%$
	$log(LAL)$	≥ 5.2
Twilight Zone 1[b]	$log(EVal)$	≤ -50
	$Positive\,Score$	$\geq 65\%$
	$log(LAL)$	≥ 4
Twilight Zone 2	$log(EVal)$	≤ 1
	$Positive\,Score$	$\geq 60\%$
	$log(LAL)$	≥ 4
Dark Zone	$log(EVal)$	≤ 1
	$Positive\,Score$	≥ 0
	$log(LAL)$	≥ 0

[a]Thresholds were chosen based on interface conservation analysis of proteins in Trans135.

[b]Twilight Zone is divided into two sub-zones: Twilight Zone 1 with stricter thresholds, and Twilight Zone 2 with looser thresholds.

Table 3.5 Boundaries of Safe, Twilight and Dark Zones used by PS-HomPPI.

Zones	**Sequence Similarity**	**Thresholds**
Safe Zone	$logEVal$	≤ -100
	$PositiveS$	$\geq 70\%$
	$Frac_{AA'}$	$\geq 80\%$
	$Frac_{BB'}$	$\geq 80\%$
Twilight Zone 1	$logEVal$	≤ -50
	$PositiveS$	$\geq 60\%$
	$Frac_{AA'}$	$\geq 60\%$
	$Frac_{BB'}$	$\geq 60\%$
Twilight Zone 2	$logEVal$	≤ 1
	$PositiveS$	$\geq 55\%$
	$Frac_{AA'}$	$\geq 40\%$
	$Frac_{BB'}$	$\geq 40\%$
Dark Zone	$logEVal$	≤ 1
	$PositiveS$	≥ 0
	$Frac_{AA'}$	≥ 0
	$Frac_{BB'}$	≥ 0

3.6 Availability

HomPPI family of protein-protein interface predictors: NPS-HomPPI and PS-HomPPI are available as webservers at $http://homppi.cs.iastate.edu/$

3.7 Acknowledgements

This work was funded in part by the National Institutes of Health grant GM066387 to Vasant Honavar and Drena Dobbs and in part by a research assistantship funded by the Center for Computational Intelligence, Learning, and Discovery. The authors sincerely thank Irina Kufareva in Abagyan Lab at the University of California for providing PIER prediction results. The authors also thank Rafael Jordan and Fadi Towfic for helpful discussions and assistance with the web server implementation. The work of Vasant Honavar while working at the National Science Foundation was supported by the National Science Foundation. Any opinion, finding, and conclusions contained in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation.

CHAPTER 4. DockRank: Ranking Docked Models Using Partner-Specific Sequence Homology Based Protein Interface Predictions

A paper titled "DockRank: Ranking Docked Models Using Partner-Specific Sequence Homology Based Protein Interface Predictions", to be submitted to PloS Computational Biology

Li C. Xue, Rafael Jordan, Yasser El-Manzalawy, Drena Dobbs and Vasant Honavar

Abstract Docking programs offer a valuable approach to computational determination of the 3-dimensional conformation of protein complexes and protein-protein interfaces. However, selecting near-native conformations from the immense number of possible conformations generated by docking programs within reasonable time presents a significant challenge in practice. We introduce DockRank, a novel approach to rank docked conformations based on the degree to which the interface residues inferred from the docked conformation match the interface residues predicted by a *partner-specific* sequence homology based interface predictor PS-HomPPI.

We compare, on a data set of 69 docked cases with 54,000 decoys per case, the ranking of conformations produced using DockRank's interface similarity scoring function applied to predicted interface residues obtained from four protein interface predictors: PS-HomPPI, and Non-Partner-Specific (NPS) interface predictors NPS-HomPPI, PRISE, and meta-PPISP, with the rankings produced by two state-of-the-art energy-based scoring functions (ZRank and IRAD). Our results show that DockRank significantly outperforms these ranking methods. Our results that NPS interface predictors (homology based and machine learning-based methods) failed to select near-native conformations that are superior to those selected by DockRank (partner-specific interface prediction based), highlight the importance of the knowledge of the binding

partners in using predicted interfaces to rank docked conformations. The application of DockRank, as a third-party scoring function without access to all the original docked models, for improving ClusPro results on two benchmark data sets of 32 and 56 test cases shows the viability of combining our scoring function with existing docking software. An online implementation of DockRank is available at http://einstein.cs.iastate.edu/DockRank/.

4.1 Introduction

The 3-D structures of complexes formed by interacting proteins are valuable sources of information needed to understand the structural basis of interactions and their role in complexes and pathways that orchestrate key cellular processes. High-throughput methods such as Yeast-2-Hybrid (Y2H) assays provide a source of information about possible pairwise interactions between proteins, but not the structures of the corresponding complexes. Because of the expense and effort associated with X-ray crystallography or NMR experiments to determine 3D structures of protein complexes, the gap between the number of possible interactions (e.g., determined using Y2H assays) and the number of experimentally determined structures is rapidly expanding. Hence, there is considerable interest in computational methods for determining the structures of complexes formed by proteins. When the structures of individual proteins are known or can be predicted with sufficiently high accuracy, docking methods can be used to predict the 3D conformation of complexes formed by two or more interacting proteins, to identify and prioritize drug targets in computational drug design, and to potentially validate interactions determined using high throughput methods such as Yeast-2-Hybrid (Y2H) assays.

In general, solving the docking problem computationally includes [62, 80, 128, 90]: 1) The initial sampling of the conformational space stage. At this stage, the proteins to be docked are rotated and translated with intervals. Usually a vast number of (thousands to tens of thousands) putative conformations are generated. Less computationally intensive ways of ranking these models are usually used at this stage, for example, the FFT (Fast Fourier Transformation) for geometric complementarity; 2) Refinement stage. The structures of the filtered decoys in step 1) are further refined and re-ranked using a higher resolution and more computational demanding scoring functions; 3) Cluster stage. Usually at the final stage of a docking procedure, top ranked

conformations (hundreds) are clustered, and often ranked by the cluster sizes: the cluster with the most members is ranked highest.

Substantial efforts have been dedicated to the design of scoring (ranking) functions for docking programs. Scoring functions in the literature can be broadly classified into four types: 1) geometric complementarity-based scoring functions; 2) energy-based scoring functions; 3) Knowledge-based scoring functions. 4) Hybrid functions that combine the scoring functions of the first three types [88, 35, 76, 77].

Geometric complementarity based scoring functions represent an early generation of scoring functions used in docking programs. Vakser and coworkers [75] introduced FFT to calculate the geometric fit between the receptor and ligand. The fast processing speed of FFT made the full conformational space search possible. This type of scoring functions were successfully applied on bound protein-protein docking but could not perform well for unbound protein-protein docking because of the conformational changes upon binding.

Energy-based scoring functions [133, 110, 59] are designed to approximate the binding free energy of protein-protein assemblies. They usually consist of the weighted energy terms of the van der Waals interaction, electrostatic interactions and solvation energies.

Knowledge-based functions can be grouped into three sub-types. a) Knowledge-based weighted correlations [63, 109]. This type of scoring functions takes into consideration of the complementarity of chemico-physical properties to overcome the limitations of scoring functions that rely on geometric complementarity alone. b) Knowledge-based pairwise potentials [95, 85, 82]. Pairwise potentials are derived from observed statistical frequency of amino acid/atom contacts in databases of solved protein structures. c) Machine-learning based methods: c1) Classifiers for directly predicting whether a query docked model as near-native or non-native. The classifiers are trained using protein complexes that are labeled as native or near-native and non-native conformations [18, 87]. c2) Classifiers that first predict protein interacting residues and then use predicted interacting residues to rank docked models [64]. First, protein-protein interface predictors are trained using protein sequences and/or structures with experimentally determined interface residues. The resulting classifiers are used to predict interface residues of the complex formed by the interacting proteins. Then the conformations are ranked using the predicted

interface residues [99, 64]. c3) Consensus scoring methods [19, 15] that combine the output of several scoring functions to give a final score.

Despite the large number of advanced and sophisticated scoring approaches that are currently used by docking programs, the goal of selecting near-native conformations from the large number of candidates is far from being solved [62]. Existing docking scoring functions often fail to rank near-native conformations higher than the rest of the conformations; and the prohibitive computational cost of existing high-resolution *atom-based* and *structure-based* scoring functions imposes practical limitations on their use in practice. Thus, selecting near-native conformations from the large number of poses generated by a docking program within a reasonable computation time presents a major obstacle for the applications of docking programs. Hence, there is a need for computationally efficient scoring functions for reliable ranking of docked conformations.

Against this background, we propose DockRank, a novel approach to rank docked conformations based on the degree to which the interface residues inferred from the docked conformation match the interface residues predicted by a partner-specific interface predictor, PS-HomPPI. Given a docking case, i.e., a pair of proteins A and B that are to be docked against each other, we use PS-HomPPI to predict the interface residues between A and B. We compare the binary vectors of interface residues between A and B predicted by PS-HomPPI with the interface residues between A and B in each of the conformations of the complex A-B produced by the docking program (See Methods for details). The greater the similarity of the interface of docked conformation with the predicted interface between A and B, the higher the rank of the corresponding conformation among the docked conformations of A with B. A novel aspect of DockRank is the partner-specific nature of the interface residue predictions. While a broad range of computational methods for prediction of protein-protein interfaces have been proposed in the literature (reviewed in [153, 51, 125]), barring a few exceptions [138, 33, 8], the vast majority of such methods focus on predicting the protein-protein interface residues of a query protein, without taking into account its specific interacting partner(s). To take into consideration the partner-specific nature of protein-protein interactions [100], DockRank makes use of PS-HomPPI [142], a sequence homology based predictor of interface residues between a given pair of potentially interacting proteins. PS-HomPPI has been shown to be able to reliably

predict the interface residues between a pair of interacting proteins whenever a homo-interolog, i.e., complex structure formed by the respective sequence homologs of the given pair of proteins. PS-HomPPI has been shown to be effective at predicting interface residues in transient complexes associated with relatively weak and reversible, often highly specific, interactions. This is especially interesting in light of the widely held belief that although transient interactions play important roles in cellular function [6, 102], transient interfaces are more difficult to predict than permanent interfaces [107, 102]. Hence, PS-HomPPI offers an especially attractive protein-protein interface prediction method for ranking docked conformations, including those that represent transient interactions.

We compare, on a data set of 69 docking cases (protein complexes) with 54,000 decoys (docked conformations) per case generated by ZDock 3.0 [67], the ranking of conformations produced using DockRank with the rankings produced by two state-of-the-art energy-based scoring functions (ZRank [67, 110] and IRAD [133]). Our results show that DockRank applied to protein interface residues predicted by PS-HomPPI, a sequence homology based partner-specific (PS) interface prediction method, significantly outperforms ZRank and IRAD. Also, to assess the importance of the knowledge of binding partners in using predicted interfaces to rank docked models, we compare and discuss the ranking results of our scoring function using PS interfaces predicted by PS-HomPPI and NPS (non-partner-specific) interfaces predicted by three state-of-the-art NPS interface predictors. Besides, we show that DockRank, when used as a third-party scoring function to rank the docked models returned by ClusPro, improves upon the the results of ClusPro obtained using three different ClusPro scoring functions on two benchmark data sets of 32 and 56 test cases generated by Cluspro 2.0 webserver [34, 35, 77, 76]. We examine the performance of DockRank on protein complexes with different degrees of conformational change upon binding; and complexes on which interfaces are predicted with different degrees of prediction confidence by PS-HomPPI. An online implementation of DockRank is available at http://einstein.cs.iastate.edu/DockRank/.

4.2 Results

4.2.1 DockRank Outperforms Energy-based Scoring Functions

We report the comparison of DockRank with two energy-based docked model scoring functions ZRank and IRAD on 69 cases obtained from ZDock3-BM3 decoy set (see Methods for the description of this decoy set). In this data set, each case has 54,000 decoys generated by ZDOCK3 program and at least one near-native structure (i.e., a hit). Figure 4.1 summarizes the distribution of the number of hits for each case. In this experiment, we used the hit definition by the IRAD 2011 paper [133]: A hit is a docked model with interface $C\alpha$ atom Root Mean Square Deviation $I-RMSD \leq 2.5$ angstroms. ZRank and IRAD are two energy-based scoring functions developed by ZDOCK group. ZRANK is a linear combination of atom-based potentials, and it has been proven to be one of top scoring functions in several studies [67, 38, 85]. IRAD is a recent improved version of ZRANK by adding residue-based potentials into ZRANK scoring functions [133].

Figure 4.2 shows the Success Rate plot for DockRank and other scoring functions on ZDock3-BM3 decoy set. DockRank consistently has a higher Success Rate than IRAD and ZRank over top ranks from 1 to 1000. If we limit our comparison to the top 1 ranked models, the Success Rate for DockRank is 40% while the Success Rate for ZRank and IRAD scoring functions is 9% and 12%, respectively. Considering the top 10 models of each scoring approach, DockRank has a 56% Success Rate against 19% and 30% Success Rates for ZRank and IRAD, respectively. Considering this decoy set is not clustered and has highly redundant conformations, this performance of DockRank in top 10 models is very encouraging. And for most cases DockRank (red) is able to rank a large proportion of total hits generated by ZDock to top 1000 ranks[1] (Figure 4.3) .

4.2.2 Partner-specific Interface Prediction Improves Ranking

Given the reliance of PS-HomPPI on the availability of homo-interologs for interface prediction, a natural question that arises is whether the general approach to ranking docking models

[1] We used such a high number of top ranks (1000) because this decoy set is not clustered and highly redundant.

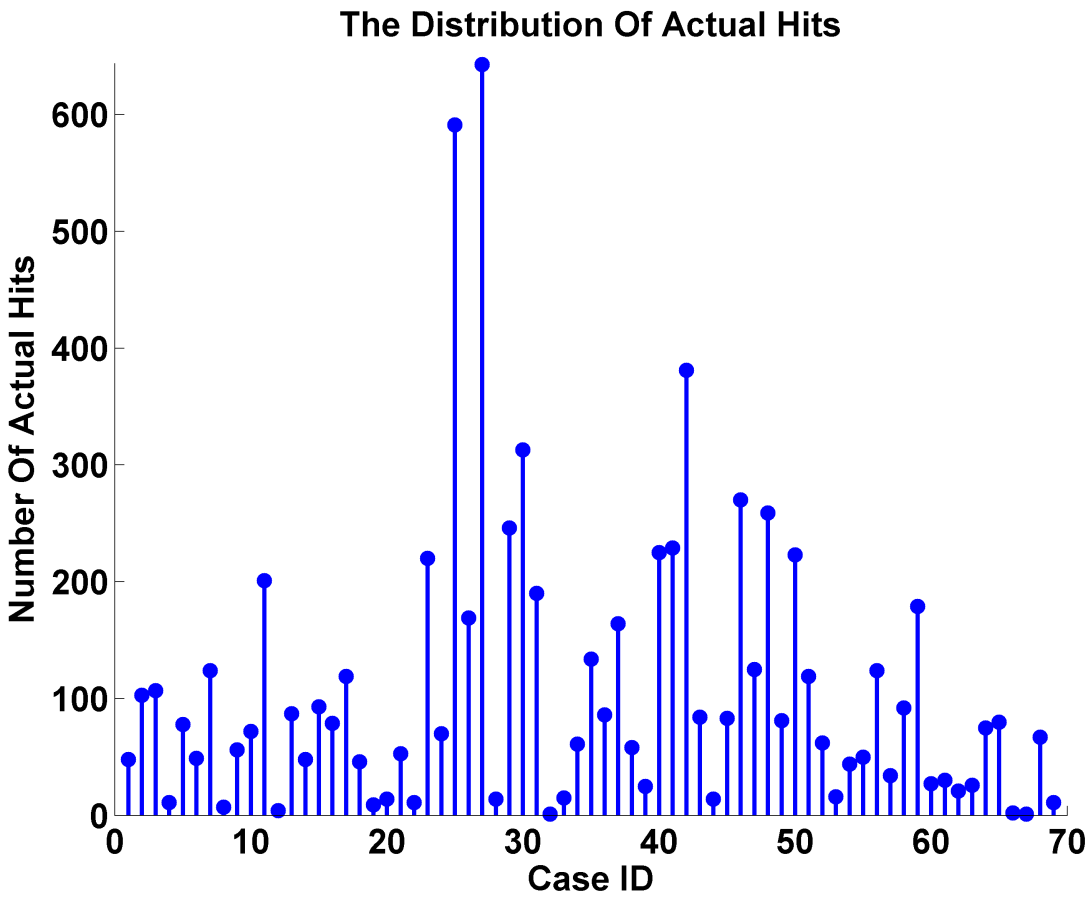


Figure 4.1 The distribution of the number of actual hits in each case of ZDock3-BM3 decoy set. Docked model with I-RMSD $\leq$ 2.5 angstroms is considered a hit. 54,000 docked models are generated by ZDock 3.0 for each case. The 69 cases that have at least one hit generated by ZDock 3.0 and can be ranked by DockRank using homo-interologs (homologous interacting proteins) in Safe, Twilight or Dark Zone are shown here.

based on the similarity of predicted interface residues with the interface residues of the corresponding docked models is applicable in cases where homo-interologs are not available. To address this question, we further investigated how well the interface predictions from the state-of-the-art machine learning based protein-protein interface predictors rank the docked models in combination with our scoring function. We compared our proposed scoring function using PS-HomPPI predictions, DockRank, with variants of our proposed scoring function using three state-of-the-art protein-protein interface residue predictors: i) NPS-HomPPI [142], a sequence homology based method for predicting protein-protein interfaces ; ii) PRISE [74], a structure

homology based method for predicting protein-protein interfaces; iii) meta-PPISP [112], consensus method that takes the scores of three other structure-based predictors cons-PPISP [156, 27], PINUP [83], and Promate [99] as input. The three predictors represent different approaches for predicting protein-protein interface residues. However, a common feature among the three predictors is that information of the interacting partner protein is not used for predicting the interfaces. Our results show that, with the same scoring function, the ranking of docked models using PS-HomPPI predicted PS-interfaces is significantly more reliable than scores based on the predicted NPS-interfaces by NPS-HomPPI, meta-PPISP and PRISE in terms of Success Rate and Hit Rate (Figure 4.2 and Figure 4.3). We conclude that incorporating knowledge of binding partners into interface predictors dramatically improves the rankings of docked models.

Clearly the performance of our proposed scoring function depends on the reliability of the predicted partner-specific protein-protein interfaces. In order to estimate an upper-limit for the performance of our scoring function, we tested the ranking performance of our scoring function by using the actual partner-specific interfaces calculated from the bound complexes (referred as PS-Actual interfaces) to rule out the effect of the uncertainty of interface predictions. Figure 4.2 shows that PS-Actual interface based ranking can identify at least one hit in top 10 models for about 90% cases while DockRank identifies at least one hit in top 10 models for almost 60% cases. Figure 4.3 shows that PS-Actual interface based ranking (green stem plot) is able to rank $>$50% total hits generated by ZDock to top 1000 for most cases. These results show that our scoring function indeed can reliably rank near native conformations to top when it is provided with reliable PS interface information, which can come from various sources, such as predictions from a PS interface predictor or experimental data. Furthermore, the gap between PS-Actual interface based ranking (green dash line) and PS-HomPPI based ranking (DockRank, red line) suggest that there is still more room for improving our ranking performance using more reliable partner-specific protein-protein interface prediction methods.

4.2.3 DockRank Has Lower I-RMSDs of Top Models

Besides using the Success Rate and Hit Rate to evaluate and compare different scoring schemes on ZDock3-BM3 decoy set, we also studied the I-RMSDs of top ranked models selected

by these scoring schemes [43]. Specifically, we treated each of the 69 docked cases as a data set; The performance of each scoring function over each case is reported as the I-RMSD of the top scored model. For each case, the different scoring functions are ranked based on their observed I-RMSD (i.e., lower ranks corresponds to lower I-RMSDs). Our analysis shows that the null hypothesis that the top 1 models selected by different scoring functions have the same means of I-RMSDs can be rejected with high confidence (p-value < 0.0001). We further applied the Nemenyi test to determine whether the means of I-RMSDs of top 1 models selected by any given pair of ranking schemes are significantly different. The critical difference determined by Nemenyi test at significance level 0.05 is 1.08. Hence, the difference between any pair of docking scoring methods is statistically significant provided the difference between their corresponding average ranks is more than 1.08. Figure 4.4 summarizes the results of our ad hoc test which indicates that there is no significant difference between our scoring function using real PS-interfaces and using PS-HomPPI predicted interfaces. And the performance of these two scoring functions is significantly better than the performance of the other scoring functions. Furthermore, no statistically significant difference is observed among the variants of our scoring function using three NPS predictors NPS-HomPPI, PRISE, meta-PPISP and the two energy-based scoring functions, IRAD and ZRank.

4.2.4 DockRank Improves ClusPro Rankings

Existing docking programs have their own embedded scoring functions. Here, we evaluated whether and to what degree DockRank, as a third party scoring function, can improve the original rankings of pre-filtered docked conformations output by a docking program, ClusPro 2.0 [34, 35, 77, 76]. Our choice of ClusPro is motivated by its superior performance reported in CAPRI competitions [77]. Briefly, ClusPro is built on top of a FFT-based rigid docking program PIPER. PIPER rotates and translates the ligand with about 10^9positions relative to the receptor. PIPER's scoring function, which contains terms of shape complementarity, electrostatic and pairwise potentials, is applied to these candidate conformations, and returns top 1000 conformations to Cluspro's clustering algorithm. Cluspro ranks the conformations (models) by cluster size.

DockRank improves the Success Rate and the average Hit Rate of rankings of ClusPro ranking schemes. We applied ClusPro to the 119 cases in Docking Benchmark 3.0. For 47 out of the 119 cases, ClusPro 2.0 generated at least one hit (e.g., a model with $L - RMSD \leq 10$ angstroms). Among these 47 cases, PS-HomPPI returns predictions for 32 cases. Thus, our experiment is limited to only these 32 cases, referred to as ClusPro2-BM3 decoy set. Figure 4.5 summarizes the distribution of the number of hits for each case.

For each case, ClusPro returned about 30 decoys ranked. We re-ranked ClusPro output using DockRank and compared our new ranking with the original ClusPro rankings provided using three different scoring schemes: ClusPro, Lowest Energy, and Center Energy. Figure 4.6 compares the Success Rates of our scoring function based on actual partner-specific interfaces (PS-Actual Interface curve, green dash with circles) and PS-HomPPI predictions (DockRank, red curve) with three ClusPro supported scoring schemes: ClusPro, Lowest Energy, and Center Energy. For each scoring scheme, we plot the Success Rate against the number of top models considered. We had to limit the number of top models to 9 models, since ClusPro returned only 9 models for one of the docked cases, 1PPE. Figure 4.6 shows that the DockRank curve dominates the three curves representing ClusPro scoring schemes. In addition, DockRank improves the average Hit Rate in top 1 models selected by ClusPro scoring functions from 0.21(ClusPro Ranking), 0.28 (ClusPro Lowest Energy Ranking), and 0.14 (ClusPro Center Energy Ranking) to 0.40 (Figure 4.7). It worth noting that this result should not interpreted as a direct comparison between DockRank and ClusPro because ClusPro scoring functions have access to 10^9decoys, generated by PIPER [76], while DockRank has access only to a very small representative set (~ 30 decoys) of these 10^9 decoys.

DockRank improves the L-RMSDs of top models filtered by ClusPro scoring schemes. We applied the Friedman test to determine whether top 1 models selected by DockRank and other ranking schemes have the same mean L-RMSDs. Our analysis shows that the null hypothesis that the mean L-RMSDs of top 1 models selected by different scoring functions are the same is rejected with high confidence (p-value $<$ 1.8667e-005). We also applied the Nemenyi test to determine whether the differences of L-RMSDs of top 1 models selected by any given pair of ranking schemes are statistically significant. The critical difference determined by Nemenyi test

at significance level of 0.05 is 1.13, which determines whether the difference between the mean L-RMSDs of top 1 models selected by two scoring schemes are statistically significant or not. The results (Figure 4.8) suggest at significance level of 0.05, the L-RMSDs of top models selected by DockRank are significantly smaller than those selected by ClusPro Center Energy. Although DockRank has a lower average L-RMSD of top 1 models, it is not significantly different from those of ClusPro Rank, ClusPro Lowest Energy Rank. This result indicates that to achieve a significant lower L-RMSD DockRank should be directly applied on the original large number of docked models instead of on the filtered docked models. On the other hand, this "not significant" result may be due to the limited power of the non-parametric test that is undermined by the small number of cases (32) and the small number of docked models (~ 9-25) that DockRank has access to.

Improving the ranking of a set of small number (~ 9-25) of clustered and pre-filtered docked models by a docking program imposes a big challenge on the third party scoring function, in that the power of the third party ranking program is limited by the ability of the embedded scoring function of the docking program to select near-native conformations in the first place. ClusPro generated at least one near-native model with L-RMSD $\leq$ 10 angstroms for only a small proportion of testing cases (47 out of 119). A natural question is when these is no hit returned by ClusPro, how well can a third party scoring function still rank reasonably good models to the top ranks? Also, ClusPro decoy set are the representative models from clustered docked models with similar 3D conformations. When a case has no hit models (L-RMSD $\leq$ 10 angstroms) returned by ClusPro, if L-RMSD of a representative model is small enough, there might be some hits in the cluster where it is chosen from. Considering that ligands in docked models with L-RMSD $\leq$ 20 angstroms usually spatially overlap with the native ligands, we therefore extend our study to any cases that have at least one docked model with the L-RMSD $\leq$ 20 angstroms in order to study DockRank's ability to select potential good clusters which might contain actual hits but are not selected out by ClusPro in the first place.

Out of the 119 cases in Docking Benchmark 3.0, there are 76 cases that have at least one docked model with L-RMSD $\leq$ 20 angstroms in the decoy set generated using ClusPro programs. Out of these 76 cases, PS-HomPPI returned interface predictions for only 56 cases. We focus on

these 56 cases in the following experiment. For each case, we calculate the average of weighted L-RMSD of top models selected by a scoring function using:

$$ave\_L-RMSD_j = \frac{1}{N_j}\sum_{i=1}^{N_j} weighted\_L-RMSD_i$$

$$weighted\_L-RMSD_i = L-RMSD_i \times Rank_i = L-RMSD_i \times i$$

where ave_L-RMSD_j is the average of weighted L-RMSD of top models for case j, N_j is the total number of models with L-RMSD $\leq$ 20 angstroms for case j, $L-RMSD_i$ is the L-RMSD of the ith ranked model by a scoring function, $Rank_i$ is the rank of ith model and it is equal to i. In this way, L-RMSD of a model is weighted by its rank, because we are more interested in top ranked models than the models ranked at the tail.

Figure 4.9 shows the difference between the average of weighted L-RMSDs of top models between DockRank and ClusPro Rank on each case. Positive numbers are cases where DockRank has a lower average L-RMSD of top models than ClusPro Rank. For 40 out of 56 (71.4%) cases, top models selected by DockRank have lower weighted L-RMSD than ClusPro. A pair-wise Wilcoxon signed rank test shows that top models selected by DockRank have significantly lower weighted L-RMSD than those selected by ClusPro (p-value = 7.5475e-004).

4.2.4.1 Error analysis on case 1RLB

From Figure 4.9, we observed that for case number 37 (PDB ID 1RLB), the top models selected by ClusPro have a much lower average L-RMSD than those by DockRank. By examining the top 1 models ranked by ClusPro (left panel) and DockRank (right panel) scoring functions (See Figure 4.10), we can see that the docked ligand position of the top 1 model selected by DockRank (white ribbon in the right panel) is totally "wrong", and it is on the opposite side of bound (correct) ligand position (pink mesh) relative to the receptor (red cartoon). However, the interface predictions of PS-HomPPI used by DockRank are *Safe Zone* interface predictions, which means high interface prediction confidence, for all receptor-ligand protein chain pairs of 1RLB case (for Zone boundaries of PS-HomPPI see [142]). And we noticed that the structure

of the receptor (red cartoon in Figure 4.10) is symmetric. So the natural question is whether it is possible that the ligand might be able to bind on both sides of the symmetric receptor instead of on only one side?

We looked into the PDB file of 1RLB downloaded from PDB(the Protein Data Bank) [13]. We realized that in fact 1RLB bound complex has two identical ligands (chain E and chain F), which bind on both sides of the receptor. It turns out that the authors of BM3 (Docking Benchmark 3.0) only included chain E into their benchmark dataset as bound ligand and excluded chain F, and PS-HomPPI is able to reliably predict the interface sites on both sides of the receptor as shown in Figure 4.11. Figure 4.11 shows the two bound ligands and the top 1 selected model by DockRank, from which we can see that the ligand of the top 1 model selected by DockRank is right beside the other bound ligand (chain F).

We recalculated the L-RMSD for all the docked models of case 1RLB by regarding either of the two identical bound ligands as native ligand positions. The smaller value of L-RMSDs between a docked ligand and two bound ligands is used as final L-RMSD for this model. After the recalculation, we found that two models have L-RMSD $\leq$ 10 angstroms (with DockRank's rank of 3 and 4 compared with ClusPro's rank of 3 and 13), and four models have L-RMSD $\leq$ 20 angstroms (with DockRank's rank of 3 , 4 , 9, 11, compared with ClusPro's rank of 3, 7, 11, 13).

Figure 4.12 shows the top 5 models selected by ClusPro and DockRank for case 1RLB. DockRank improves the rankings of ClusPro by selecting models with ligands binding on either side of the receptor.

To facilitate comparisons with future developed scoring functions, we made the ClusPro decoy set available to the community at http://einstein.cs.iastate.edu/DockRank/supplementaryData.html, including Docked models generated using four different ClusPro energy functions, L-RMSD for each docked model, DockRank scores, ClusPro scores and the recalculated L-RMSDs of models of 1RLB after including both identical bound ligand chains.

4.3 Discussion

Selecting near-native conformations from thousands of decoys generated by a docking program remains a challenging problem in computational molecular docking [62]. In this study, we presented DockRank - a novel predicted interface based scoring method for protein-protein docking. The proposed scoring function relies on a measure of similarity between interfaces of docked models and predicted interfaces by a PS interface predictor PS-HomPPI.

Comparisons of DockRank with two state-of-the-art energy based docking scoring functions, ZRank and IRAD, demonstrated an impressive superior performance of DockRank over ZRank and IRAD on a decoy set of 69 cases with 54,000 decoys per case. These results suggest the viability of predicted interface based scoring functions as an alternative to complicated and computationally expensive energy based scoring functions. The observation that our scoring function using PS-HomPPI significantly outperforms three variants of our scoring functions using three state-of-the-art non-partner specific protein-protein interface predictors underscores the importance of partner-specific protein-protein interface prediction methods for providing reliable interface predictions in general and for providing reliable protein-protein docking scoring functions.

The reason why NPS interface predictors do not perform as well as PS interface predictors in selecting near-native conformations to top ranks for a large portion of the test complexes in our study might be as follows. The complexes in our study that are used to generate the decoys are transient binding proteins, the interfaces of which are mostly highly partner-specific. Suppose that we have a perfect NPS-interface predictor that can correctly predict the union of all the actual interface residues of one query protein with all its possible interacting partners. Many proteins tend to use different residues to interact with different partners, which is especially true for transient bindings (see Xue et al. [142] for the comparison of obligate and transient interface conservations, and the comparison of PS-interface and NPS-interface conservations of transient interactions). Therefore, the prediction of the perfect NPS-predictor not only includes the actual interface between receptor A and ligand B that we are interested in, but also other interface residues that may be far away from the interface between A and B. And

these "false positive interface residues" may falsely give top ranks to the docked models that have docked interfaces near these "false positive interacting areas". Therefore, to reliably select near native conformations for transient interactions using predicted interfaces, a reliable PS-interface predictor is needed.

DockRank is insensitive to conformational changes upon binding. Large conformational changes upon binding impose big challenges onto the predicted interface based scoring schemes, in that it is challenging for docking programs to generate detectable numbers of near native conformations, and it is challenging to make reliable interface predictions for the interface predictor used by the scoring function. Therefore, it is of great interest to evaluate the performance of scoring functions on complexes with different conformational change levels. Although BM3 dataset classified each complex into three conformational change upon binding groups, to systematically study the differences of ranking performances with respect to conformational changes upon binding is difficult in practice because of the limited capability of docking programs to generated at least one hit for complexes with large conformational changes. Therefore, we studied the interface prediction performance of our underlying interface predictor PS-HomPPI with respect to different conformational change levels to *indirectly* study the performance of DockRank in ranking docked models of cases with different conformational changes. For each interface prediction, PS-HomPPI also provides an interface prediction confidence zone (Safe/Twilight/Dark Zone) where the homo-interologs used for interface inferences lie. The effects of conformational changes and the prediction confidence zones on the performance of PS-HomPPI may be confounded. So we summarized the interface prediction performance of PS-HomPPI into 9 subgroups of three levels of conformational changes and three prediction confidence zones (Table 4.1). From Table 4.1, we can see that the performance of PS-HomPPI is insensitive to conformational changes upon binding and it is clearly correlated with the prediction confidence zones: the higher confidence the more reliable the interface predictions are. Hence, one can expect the performance of DockRank in ranking docked models is insensitive to the conformational changes, because both the interface prediction of PS-HomPPI and the calculation of ranking scores are insensitive to the conformational changes.

Since DockRank is insensitive to conformational changes (discussed above), we are able to

Table 4.1 Interface prediction performance of PS-HomPPI on BM3 dataset with three different prediction confidence zones on three levels of conformational change upon binding. Only the interfaces between the receptors and ligands are predicted and used by DockRank in ranking docked models. During the evaluation, we consider each partner-specific predicted receptor-ligand interface as one prediction. For example, for complex A-BC with one receptor A and two ligand chains B and C, we consider four predictions: A|A-B, A|A-C, B|B-A, C|C-A, where A|A-B means the interface of A with its binding partner B. There are totally 379 partner-specific receptor-ligand interfaces, 65.4% of which can be predicted by PS-HomPPI using homo-interologs in Safe, Twilight or Dark Zones. The performance of PS-HomPPI is not affected by the conformational changes upon binding; instead it is clearly correlated with the prediction confidence zones. The most reliable interface predictions are obtained in Safe Zone. Some residues of some proteins may not have interface predictions from PS-HomPPI. These residues are not considered in the evaluation, because these residues are not used by DockRank in ranking docked models.

Zones	**Conformational Change Upon Binding**	**Num of Predictions (out of 379)**	**CC**	**F1**	**Specificity**	**Sensitivity**	**Accuracy**
Safe	Rigid	148	0.72	0.75	0.72	0.77	0.96
	Medium	24	0.77	0.8	0.75	0.85	0.95
	Difficulty	30	0.73	0.75	0.69	0.81	0.95
Twilight	Rigid	32	0.66	0.71	0.74	0.68	0.92
	Medium	4	0.46	0.52	0.62	0.45	0.88
	Difficulty	4	0.66	0.72	0.75	0.69	0.9
Dark	Rigid	4	0.24	0.33	0.34	0.33	0.84
	Medium	0	-	-	-	-	-
	Difficulty	2	0.51	0.6	0.54	0.68	0.85
Total		248/379 = 65.4%	0.71	0.74	0.71	0.76	0.95

investigate the performance of DockRank on different confidence zones independent of the conformational change factor. There are 66 cases that have only Safe Zone interface predictions, of which 53 cases have at least one hit. There are 16 cases that have only Twilight interface predictions, of which 12 cases have at least one hit. There are 3 cases that have only Dark interface predictions, of which 2 cases have at least one hit. When a case has more than two protein chains (hence more than one pair of query proteins), it may have the predicted interfaces by PS-HomPPI from different confidence zones. When a case has different confidence zones, it is not further analyzed. We studied the Success Rates and Hit Rates of DockRank on cases with only Safe, Twilight, and Dark Zone interface predictions, respectively. As we expected, DockRank performs best with Safe Zone interface predictions relative to other confidence zones. In Twilight Zone, the Success Rate and average Hit Rate of DockRank declines, but still outperforms other scoring functions that we compared with. In Dark Zone (2 cases), DockRank is able to rank a hit at the rank of 100 out of 54,000 models for one case, but fails to find any hits for the other case in top 1000 models. The average Hit Rate of top 1000 models of DockRank (0.01) in Dark Zone is lower than IRAD (0.09) and ZRank (0.07). See Figure 4.13 for the Success Rate plots and Table 4.2 for the average Hit Rates of top 1000 models of DockRank and other scoring functions in different confidence zones.

Different weights can be assigned to predicted interfaces with different prediction confidences when calculating the scores for each docked model. For a docked model with more than two protein chains, it is possible that the interface predictions for these proteins are from different interface prediction confidence zones. It is reasonable to set different weights for interface predictions from different prediction confidence zones. In our study here, we used weights 1, 1, and 0.001 for the Safe, Twilight, and Dark Zone interface predictions, respectively. Our web server allows the users to set different weights for interface predictions from PS-HomPPI when ranking docked models: higher weight for Safe Zone predicted interfaces and lower weight for Dark Zone predicted interfaces.

Like any homology based method, an important limitation of our scoring method DockRank is that the current implementation of DockRank using PS-HomPPI might fail to score some cases when homologs for inferring partner-specific interfaces are not available. For example,

Table 4.2 Average Hit Rates in top 1000 models of different scoring functions on ZDock3-BM3 decoy set in different interface prediction confidence zones. Cases with more than one receptor-ligand chain pairs may have predicted interface from different confidence zones. Cases with solo confidence zones are studied here. 66 cases that have only Safe Zone interface predictions, of which 53 cases have at least one hit. 16 cases have only Twilight interface predictions, of which 12 cases have at least one hit. 3 cases have only Dark interface predictions, of which 2 cases have at least one hit. DockRank has the most reliable performance in terms of Hit Rates for cases with interface prediction confidence in Safe Zone. The average Hit Rate of DockRank Cases with Twilight Zone confidence declined but still outperformed other scoring functions.

Zones	Num of Cases	PS-Act Int	**DockRank**	NPS-HomPPI	PRISE	Meta-PPISP	IRAD	ZRank
Safe	53	0.85	**0.72**	0.32	0.15	0.08	0.15	0.12
Twilight	12	0.84	**0.56**	0.25	0.12	0.15	0.15	0.1
Dark	2	0.76	**0.01**	0.49	0.04	0.00	0.09	0.07

PS-HomPPI returns interface predictions for 87 out 119 cases comprising Docking Benchmark 3.0 (coverage is 73.1%). Hence, current implementation of DockRank using PS-HomPPI might fail to score some cases when homologs for inferring partner-specific interfaces are not available. We expect that the coverage of DockRank will increase as more complexes are deposited in PDB. Also, to improve the coverage of DockRank, one might use a hybrid interface prediction method that combines the interface predictor PS-HomPPI with machine learning based *PS-interface* prediction methods that do not require the availability of putative homo-interologs with experimentally determined interfaces. Our previous and current study shows that taking into account the information of binding partner is very important for reliably predicting the interfaces of transient binding cases [142] and in ranking docked models of such cases (Figures 4.2 and 4.3). The development of machine learning based protein-protein interface predictors that exploit the information of both the query protein and its binding partner is urgently needed.

Lastly, PS-HomPPI interface prediction can be used as constraints for docking. Because of the high computational cost of exploring the large conformational space of complexes formed by several protein chains, there has been increasing interest in utilizing knowledge of the actual or predicted interface residues between a pair of proteins to constrain the exploration of docked

configurations to those that are consistent with the predicted interfaces (thus improving the computational efficiency of docking and the accuracy of docking [39, 40, 44, 38, 80]). Our analysis shows that our interface predictor PS-HomPPI can reliably identify the interfaces between receptors and ligands when the homo-interologs of the receptor-ligand protein pair can be reliably identified (Table 4.1). Also, another advantage of the interface predictions of PS-HomPPI is that it is not affected by the conformational changes upon binding (as shown in Table 4.1), because the input of PS-HomPPI is protein sequences. Therefore, one may expect that using the predicted interfaces from PS-HomPPI as constraints to the docking process might help docking procedures to generate hits for complexes even with large conformational changes upon binding.

4.4 Materials and Methods

4.4.1 Decoy sets

In this study, for different purposes we used two decoys sets : ZDock3-BM3 [66] and ClusPro2-BM3. ZDock3-BM3 decoys are used to compare DockRank with other scoring functions, and ClusPro2-BM3 decoys are used to evaluate whether and how well DockRank can improve the pre-filtered docked models. These two decoy sets represent different aspects of two different state-of-the-art docking programs. ZDock3-BM3 decoy set faithfully reflects the initial population of decoys generated by ZDock 3.0 [67] before clustering. ClusPro2-BM3 decoys are representative decoys of top ~30 clustered decoys generated by ClusPro 2.0 [34, 35, 77], which represent the common clustered outputs of a docking program.

1) ZDock3-BM3 decoy set

Docking Benchmark 3.0 (BM3) consists of a set of non-redundant transient complexes (3.25 Å or better resolution, determined using X-ray crystallography) from three biochemical categories: enzyme-inhibitor, antibody-antigen, and “others”. This data set includes complexes that are categorized into three difficulty groups for benchmarking docking algorithms: Rigid-body (88 complexes), Medium (19), and Difficult (17), based on the conformational change

upon binding. Obligate complexes are filtered out manually. BM3 originally had 124 cases. 2VIS (rigid-body), 1K4C (rigid-body), 1FC2 (rigid-body), 1N8o (rigid-body) were deleted because the bound complexes and the corresponding unbound complexes have different number of chains. 1K74 (rigid-body) was deleted because the sequence of chain D in the bound complex is different from the corresponding unbound chain 1ZGY_B. There are finally 119 docking complexes: Rigid-body (83 complexes), Medium (19), and Difficult (17). A set of 54,000 decoys for each case was generated using ZDock 3.0 . Despite the large number of generated decoys, there are only 97 cases that have at least one near-native structure (e.g., a decoy with interface $C\alpha$ atom Root Mean Square Deviation $I - RMSD \leq 2.5$ angstroms). Out of these 97 cases, our homology based protein-protein interface predictor, PS-HomPPI, returned interface predictions for only 69 cases. Therefore, our final decoy set consists of decoys generated for these 69 cases.

2) ClusPro-BM3 decoy sets

This decoy set was also generated from the 119 cases in BM3 using ClusPro 2.0 program. For each case ClusPro returned 9-25 decoys. ClusPro-BM3_32 decoy set of 32 cases were generated using the following selection criteria: i) each case should have at least one hit (i.e., a decoy with L-RMSD $\leq$ 10 angstroms); ii) PS-HomPPI interface predictions are available for the proteins in that complex. Another decoy set of 56 cases, ClusPro-BM3_56 was generated by relaxing the definition of a hit to include decoys $\leq$ 20 angstroms .

4.4.2 PS-HomPPI (Partner-Specific Homology based Protein-Protein Interface predictor)

DockRank uses the predicted interfaces by PS-HomPPI to rank docked models. PS-HomPPI is a sequence homology based method for partner-specific (PS) protein-protein interface residue prediction [142, 144]. PS-HomPPI uses the experimentally determined interfaces of homo-interologs (homologous interacting proteins) to infer those of a query protein pair. PS-HomPPI is described in details in [142], and we briefly summarize it here.

PS-HomPPI consists of two major components: PS-interface conservation and PS-interface prediction.

PS-interface conservation

The prediction of PS-HomPPI is based on our PS-interface conservation analysis on a large non-redundant transient interacting proteins dataset Trans135 (a non-redundant transient binding dataset taken from [94]), the interfaces of which are experimentally determined. For a query, i.e., a protein A and its interaction partner B in Trans135, we identify their homo-interologs (homologous interacting proteins) using BLASTP [9] to identify the homologs of A and homologs of B in the PDB with $E-Value \leq 10$ and $PositiveS \leq 100$. Because the interfaces of both the query protein pair and of their homo-interologs are known, we can calculate $IC-score$ (Interface Conservation Score) for each query - homo-interolog pair. Each query - homo-interolog pair can be represented as a 10×1 vector with 9 sequence alignment measures and one $IC-score$. We explored the functional relation between $IC-score$ and 9 sequence alignment measures in the first two PCs (Principal Components) space. As we can see from the PCA biplot in Figure 4.14, the degree of PS-interface conservation is correlated with the 9 sequence alignment measures (blue lines with red circles on the tip). Based on the color trend of $IC-score$, we can divide the conservation space into Safe, Twilight and Dark Zones, hence we established the alignment quality criteria that must be met for a certain protein-protein interface prediction confidence.

PS-interface prediction

The prediction of PS-HomPPI is based on the alignment quality criteria established in PS-interface conservation analysis. Figure 4.14 illustrates the prediction process of PS-HomPPI. Given a query protein A and its interaction partner B, PS-HomPPI first identifies the set of homo-interologs (homologous interacting protein pairs) of A-B using BLASTP to identify the homologs A' of A and homologs B' of B in the PDB. Based on the alignment quality between the query A-B and a homo-interolog A'-B', we can know which zone this homo-interolog lies. PS-HomPPI ranks the homo-interologs based on their sequence similarity to the query protein pair. PS-HomPPI uses $K = 10$ (or fewer if 10 homo-interologs meeting the similarity thresholds are not available) nearest homo-interologs to infer the interfaces of the query protein pair. PS-

HomPPI predicts interface residues in a protein chain based on the known interface residues of its closest homo-interologs. Specifically, a residue in query sequence A is predicted to be an interface residue with respect to an interaction partner B, if a majority of the residues in the corresponding position in its alignment with K of the closest homo-interologs of A-B are interface residues. If at least one homo-interolog in the Safe Zone is found by the BLASTP search, PS-HomPPI uses the Safe Zone homo-interolog(s) to infer the interfaces of the query protein. Otherwise, the search is repeated for homo-interologs in the Twilight and Dark Zones. The zone area of homo-interologs used in predictions provides the prediction confidence level.

To objectively evaluate the performance of PS-HomPPI in ranking docked models, highly similar homo-interologs were removed. Specifically, for query A-B and its homologous interacting pair A'-B', we also discard the interacting protein pair A'-B' if A and A' or B and B' share $\geq$ 95% sequence identity and belong to the same species. Each case in the decoy sets have bound and unbound proteins. Unbound proteins were used by docking programs to generated docked models and their sequences were used by PS-HomPPI to predict interfaces. The bound complexes were used to evaluate the ranking schemes of docked models. The bound complex of each case (although most bound complexes are probably removed in the first filter of highly similar homologs) was also explicitly deleted from the homo-interolog list, and was not used in later prediction.

The default parameters of PS-HomPPI were used in this study. For detailed parameter settings please refer to Xue et al. 2011 [142].

4.4.3 Databases Used by PS-HomPPI

Four databases are used by PS-HomPPI to make interface predictions.

ProtInDB [2] (version Aug 2011) and S2C DB [3](version Sep 23rd, 2011): Used by PS-HomPPI to calculate the interface residues of homo-interologs. ProtInDB is a protein-protein interface residues database. It only contains the protein complexes with at least one pair of interacting chains in PDB. ProtInDB web server is at http://einstein.cs.iastate.edu/protInDb/. S2C DB is used to map the calculated interface residues based on ProtInDB to the whole protein sequences.

BLAST nr_pdbaa_s2c: For BLASTP 2.2.25+ searching for close sequence homologs. It is built based on ProtInDB in Aug 2011 and S2C DB on Sep 23rd 2011. Only protein chains existing in ProtInDB are included into nr_pdbaa_s2c. We built a non-redundant database for BLAST queries from the S2C fasta formatted database. To generate the non-redundant BLAST database, we grouped proteins with identical sequences into one entry. Now nr_pdbaa_s2c contains 31,527 sequences and 7,854,391 total letters.

SIFTS taxonomy file (version Sep 23rd, 2011): Used to determine the species of a protein chain in a PDB entry when we remove highly similar homologs and homo-interologs. Some protein chains are missing from SIFTS taxonomy file. If a homolog is missing from SIFTS taxonomy file, we assume it is from the same species as the query protein to keep a stringent criteria of highly similar homologs.

Interface Definition

Interface residues are defined as residues with at least one atom that is within a distance of 5 angstroms from any of the atoms of residues in the interaction partner chain.

4.4.4 DockRank: Our Scoring Function for Ranking Docked Conformations

Given a pair of proteins A and B that are to be docked against each other by a docking program, we use PS-HomPPI to predict the interface residues between A and B. We represent predicted/docked interfaces as binary vectors with 1s meaning interface residues and 0s non-interface residues. We then compare the binary vectors of interface residues between A and B predicted by PS-HomPPI with the interface residues between A and B in each of the conformations of the complex A-B produced by the docking program. The docked conformation with the greatest similarity of interface vectors with the predicted interface residues is assigned the top rank.

Many similarity measures for binary vectors have been proposed (See [151] for a review). Among these, only Russell-Rao, SoKal-Michener and Rogers-Tanmoto(-a) measures are defined in the case when both sequences consist of all 0 elements (which is the case when there are no interface residues between the corresponding protein chains and both PS-HomPPI and the

docking model correctly predict no interface residues). Because the numbers of interface and non-interface residues are highly unbalanced, we used weighted SoKal-Michener metric to measure the similarity between the interface and non-interface residues in a protein chain A (with chain B) encoded in the form of binary sequences $\vec{v_A}$and $\vec{v_B}$ based on PS-HomPPI predictions and the docked conformation, respectively,

$$S(\vec{v_A}, \vec{v_B}) = \frac{S_{11} + \beta S_{00}}{N}$$

where S_{11} and S_{00} are the numbers of positions where the two sequences match with respect to interface residues and non-interface residues, respectively, and β is a weighting factor, $0 < \beta < 1$, that is used to balance the number of matching interface residues against the number of matching non-interface residues.

The weighting factor β is defined as a PS-interface residue ratio. For example, for docking a protein consisting of a single ligand chain A with a receptor protein consisting of chains B and C,

$$\beta = \frac{\#int\ A|A:B}{length(A)} + \frac{\#int\ A|A:C}{length(A)} + \frac{\#int\ B|B:A}{length(B)} + \frac{\#int\ C|C:A}{length(C)}$$

where "$\#int\ A|A:B$" denotes the number of interface residues of protein chain A with respect to its binding partner B.

In this study, we set $\beta = 0.08$, which is calculated using a set of transient interaction proteins with experimentally determined interfaces.

Only the interface residues between the receptor and the ligand are used to rank docked models. When predicted interface vector is a zero vector, it is NOT used in ranking docked models. When actual interface vector is a zero vector, it is USED in ranking docked models.

For each docked conformation we calculate one score using our scoring function. When a protein complex consists of multiple chains, multiple interface similarities were calculated, and they were weighted based on the prediction confidence zones of PS-HomPPI (The weight of the interface similarity is 1 if the predicted interface is from Safe Zone of PS-HomPPI, 1 for Twilight Zone, and 0.001 for Dark Zone), and averaged by pairing each chain of the receptor

with each chain of the ligand.

4.4.5 Evaluation of Scoring Functions

We used Root Mean Square Deviations (RMSDs) to measure the difference of structures between each decoy and the corresponding bound complex (target complex): L-RMSD (Ligand-RMSD) is the backbone RMSD between the ligand in the docked decoy and the bound ligand after superimposing the receptor of the decoy and that of bound complex; I-RMSD (Interface-RMSD) is the backbone RMSD calculated by superimposing the backbone of the docked interface and the bound interface.

We also used Success Rate and Hit Rate to evaluate different ranking schemes. Success Rate is defined as the percentage of cases that have at least one hit (near-native conformation) in top n ranks. Hit Rate is defined as the percentage of hits that are detected in the top n ranks. Hit Rate measures the enrichment of hits in top ranked models.

Upper bounds of Success Rate and Hit Rate: We used actual PS-interfaces to rank the decoys to obtain the upper bounds of Success Rate and Hit Rate.

Lower bounds of Success Rate and Hit Rate: The number of hits X in the top K random picks from total N generated decoys with total M hits follows Hypergeometric distributions: $X \sim HG(N, M, K)$. We calculated the expectations and variances of the Success Rate and Hit Rate of a random pick as the lower bound of scoring functions (see Appendix 1 for details).

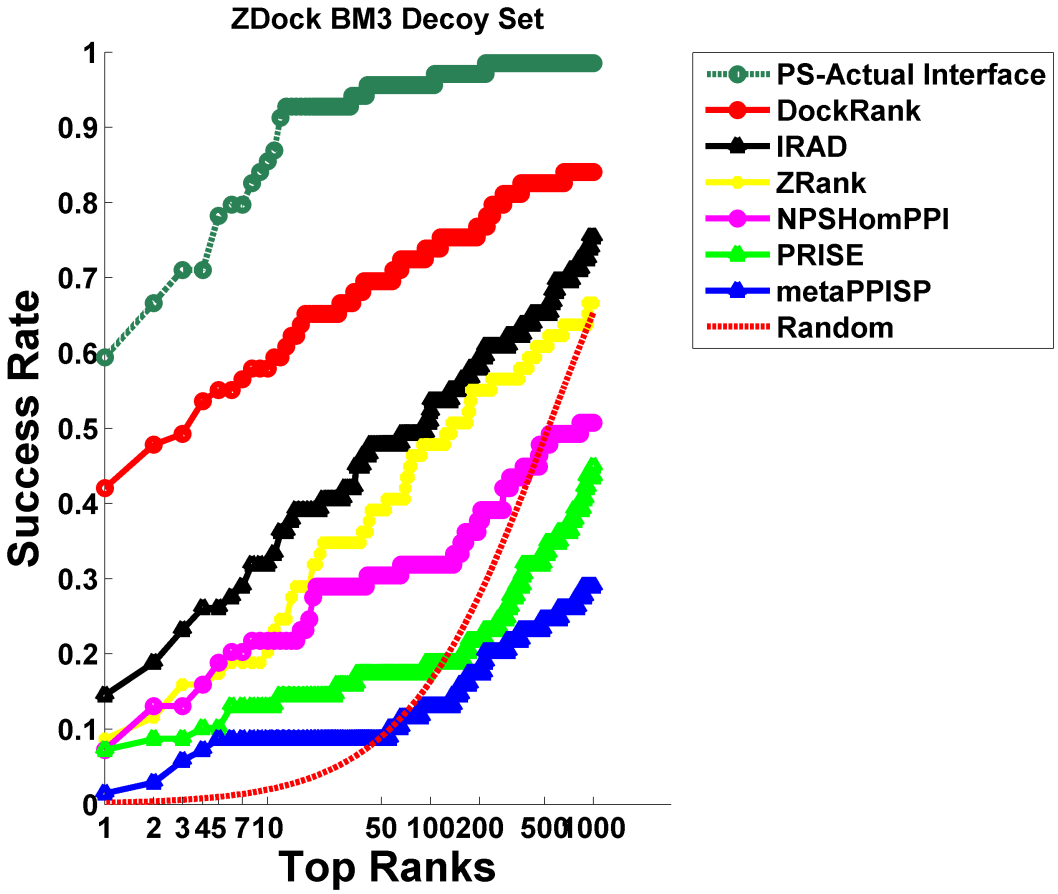


Figure 4.2 The Success Rates of DockRank and other scoring functions on ZDock3-BM3 decoy set. The Success Rates of DockRank (red solid line) are compared with those of two energy-based scoring functions: IRAD (black) and ZRank (yellow), and with three other rankings of docked models by combining our scoring fucntion with the predicted NPS-interface from: NPS-HomPPI (purple), PRISE (light green solid), and meta-PPISP (blue). NPS-HomPPI, PRISE and meta-PPISP are NPS-interface predictors, which do not consider the information of the query protein's binding partner when predicting interface residues. DockRank consistently has significantly higher Success Rates than IRAD, ZRank, NPS-HomPPI, PRISE, meta-PPISP. The Success Rates of the ranking of docked models by PS-actual interface residues (dark green dash line) combined with our scoring function and the expectations of the Success Rates of a random pick (dash red line) are plotted to defined the upper and lower bound. Studied here are 69 out of 97 cases that have at least one hit (I-RMSD $\leq$ 2.5 angstroms) and can be ranked by DockRank using homo-interologs in Safe, Twilight and Dark Zone.

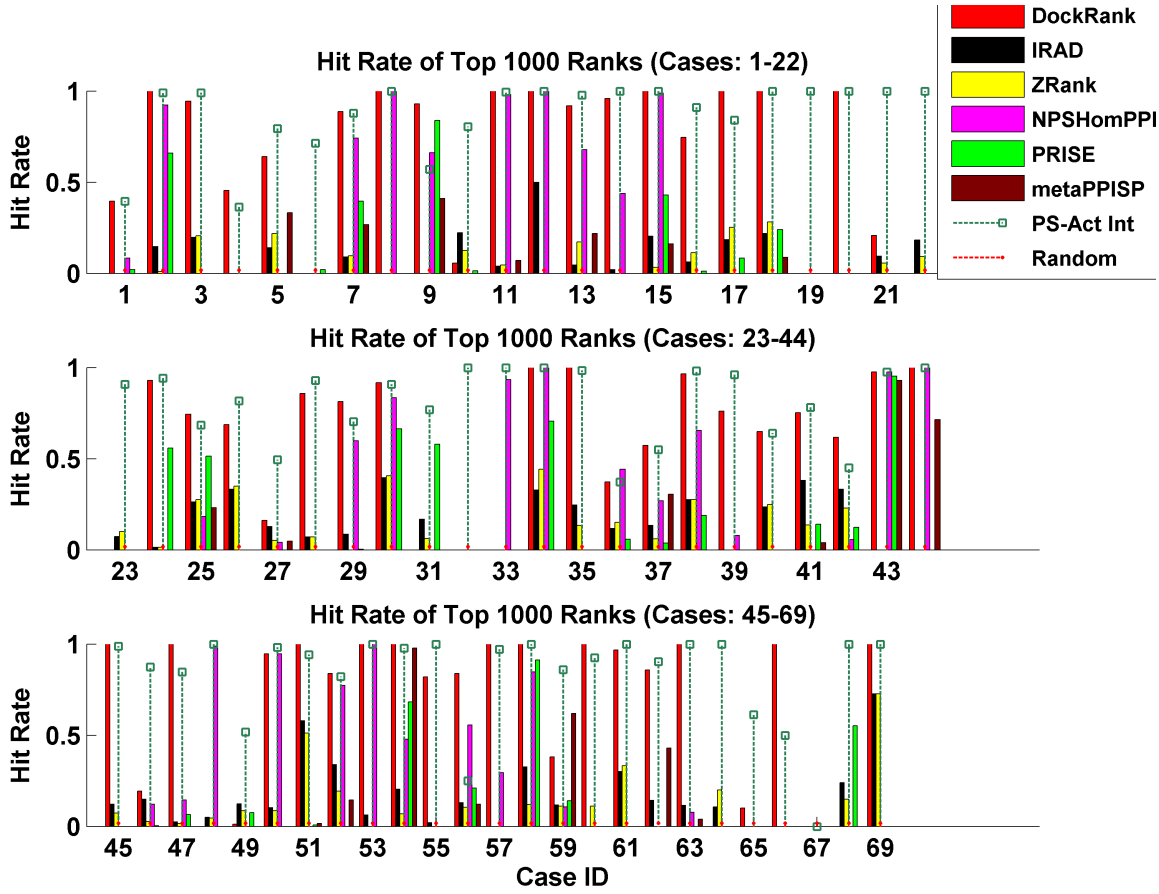
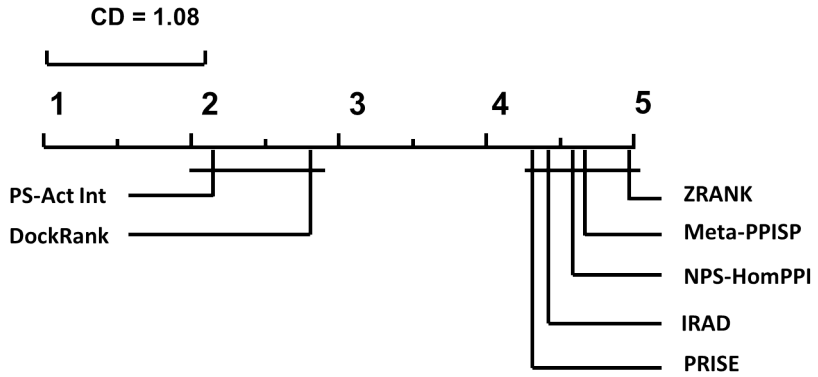


Figure 4.3 The Hit Rates of DockRank and other scoring functions on ZDock3-BM3 decoy set. The hit rates of top 1000 ranked models selected by DockRank (red) are compared with those of two energy-based scoring functions: IRAD (black) and ZRank (yellow), and with three other rankings of docked models by combining our scoring function with the predicted NPS-interface from: NPS-HomPPI (purple), PRISE (light green solid), and meta-PPISP (brown). NPS-HomPPI, PRISE and meta-PPISP are NPS-interface predictors, which do not consider the information of the query protein's binding partner when predicting interface residues. DockRank consistently has higher Hit Rates than other scoring functions. The Hit Rates of the ranking of docked models by PS-actual interface residues combined with our scoring function (dark green stem dash line) and the expectations of Hit Rates of a random pick (dash red stem dash line) are plotted to define the upper and lower bound. Studied here are 69 out of 97 cases that have at least one hit (I-RMSD $\leq$ 2.5 angstroms) and can be ranked by DockRank using homo-interologs in Safe, Twilight and Dark Zone.



	PS-Act Int	DockRank	PRISE	IRAD	NPS-HomPPI	Meta-PPISP	ZRANK
Average Rank	2.14	2.80	4.35	4.46	4.59	4.65	4.99

Figure 4.4 Pair-wise comparisons of the mean of I-RMSDs of top 1 models selected by different docking scoring methods on ZDock3-BM3 decoy set using the Nemenyi test. Methods that are not significantly different (at significance level) are grouped together (via connecting lines). The average "rank" of each method over docking cases is shown in the table (and also on the x-axis of the plot). The mean of I-RMSDs of top 1 models selected by DockRank is significantly smaller than those selected by IRAD, ZRank, NPS-HomPPI, PRISE, and meta-PPISP. The mean of I-RMSDs of top 1 models selected by DockRank is not significantly different from actual partner-specific interface-based method (PS-Act Int).

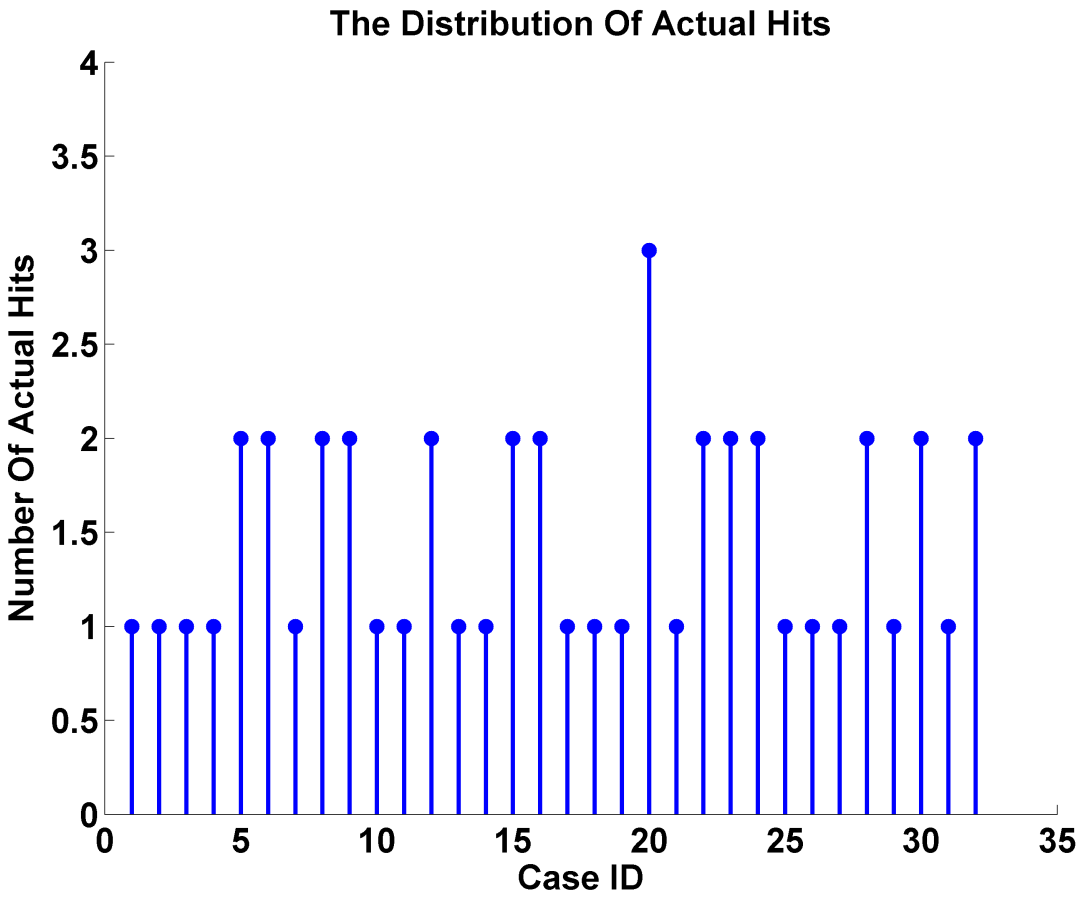


Figure 4.5 The distribution of the number of actual hits in each case of Cluspro decoy set. Docked model with L-RMSD $\leq$ 10 angstroms is considered a hit. 54,000 docked models are generated by ZDock 3.0 for each case. The 32 cases that have at least one hit generated by ClusPro 2.0 and can be ranked by DockRank using homo-interologs (homologous interacting proteins) in Safe, Twilight or Dark Zone are shown here.

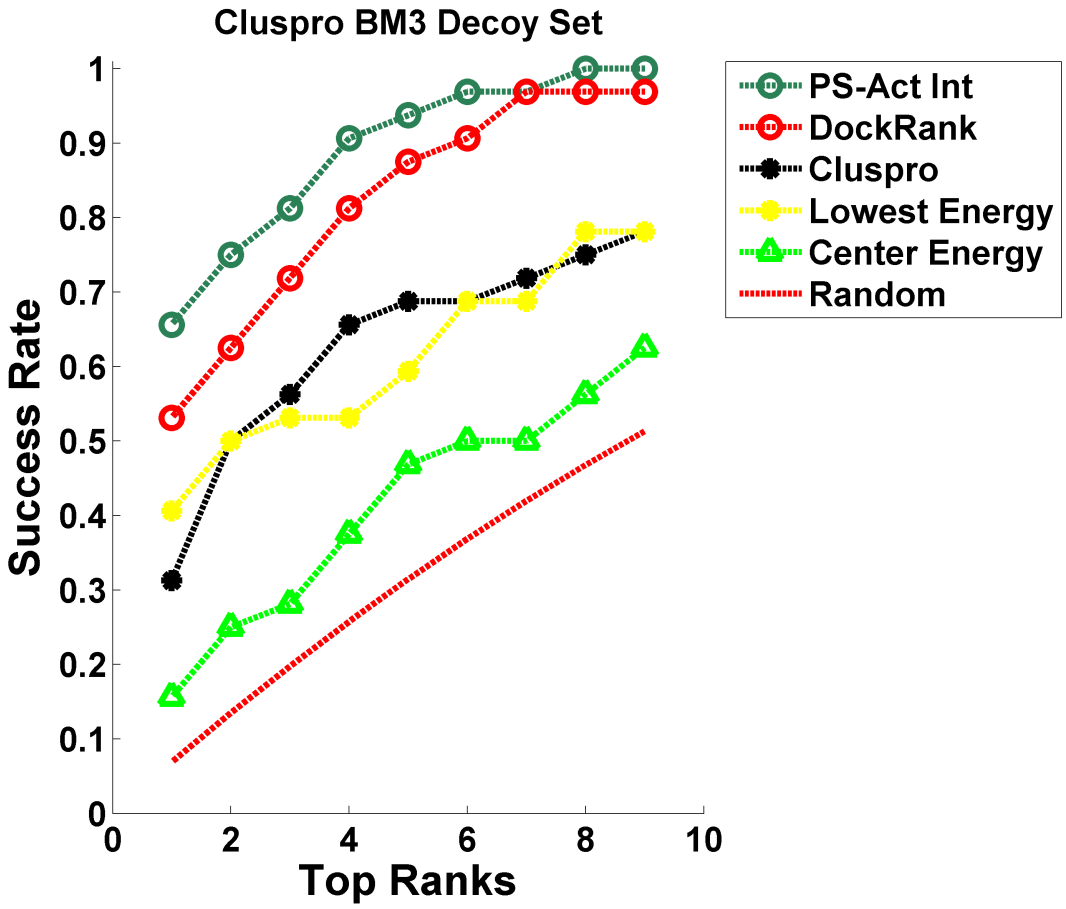
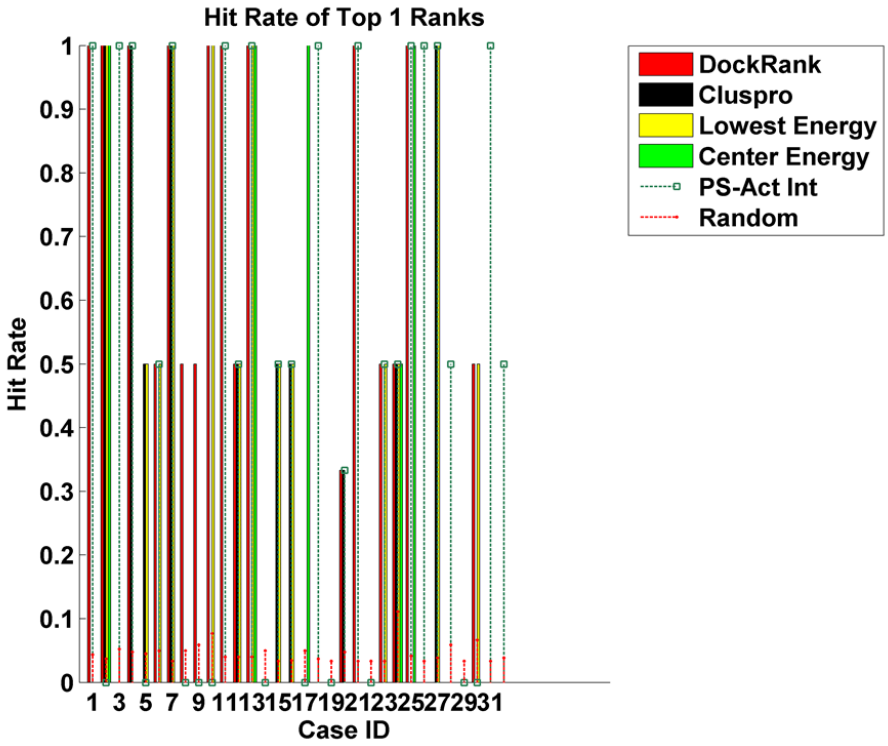
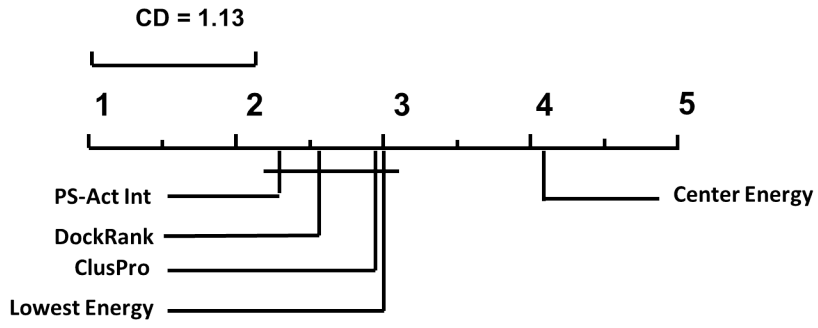


Figure 4.6 The Success Rates of DockRank and ClusPro scoring functions on ClusPro2-BM3 decoy set. ClusPro scoring functions (Default Cluser-size based, Center Energy-based, Lowest Energy-based) were applied on the original docked models generated by ClusPro's underlying docking program PIPER. DockRank was applied on the filtered docked models by ClusPro scoring functions. The Success Rates of three ClusPro scoring functions are significantly improved. DockRank (PS-HomPPI interface prediction based) is able to select at least one hit in top 8 ranked models for more than 95 % cases tested here. The Success Rate of PS-actual interface based ranking and the expectation of Success Rate of random rankings are also plotted to show the upper and lower bound of Success Rates. 32 cases that have at least one hit and whose interface can be predicted by PS-HomPPI using Safe, Twilight, or Dark Zone homo-interologs are studied here. Case 1PPE has only 9 models, so the Success Rates of up to top 9 rankings are studied here.



	DockRank	Cluspro	Lowest Energy	Center Energy
Average Hit Rate	0.40	0.21	0.28	0.14

Figure 4.7 The Hit Rates in top 1 docked models ranked by DockRank and ClusPro scoring functions on ClusPro2-BM3 decoy set. The average Hit Rates of scoring functions over cases are shown in the table. DockRank improved the average Hit Rates of top 1 docked models from 0.21 of ClusPro, 0.28 of Lowest Energy, and 0.14 of Center Energy, to 0.40. The Hit Rates of PS-Actual interface-based ranking and the expectation of Hit Rates of random rankings (see Appendix for the derivation of the expectation and variance of the random Hit Rate) are calculated to define the upper and lower bound. 32 cases that have at least one hit and whose interacting residues can be predicted by PS-HomPPI using homo-interologs in Safe, Twilight, or Dark Zone are studied here.



	PS-Act Int	DockRank	ClusPro	Lowest Energy	Center Energy
Average Rank	2.34	2.53	2.98	3.00	4.14

Figure 4.8 Pair-wise comparisons of different docking scoring methods on ClusPro2-BM3 decoys using the Nemenyi test. Methods that are not significantly different (at significance level alpha = 0.05) are grouped together (via connecting lines). The average "rank" of each method over docking cases is shown in the table (and also on the x-axis of the plot). Pairwise Nemenyi test shows that the average L-RMSDs of top models selected by DockRank are significantly smaller than those selected by ClusPro Center Energies. However, the average L-RMSDs of top models selected by ClusPro, Lowest Energy and DockRank are not significantly different, which indicates that DockRank has limited improvement on top 1 model in term of L-RMSDs when applied to the filtered docked models by ClusPro scoring functions under the definition of a hit as a docked decoy with L-RMSD $\leq$ 10 angstroms. 32 cases that have at least one hit and whose interface can be predicted by PS-HomPPI using Safe, Twilight, or Dark Zone homo-interologs are studied here.

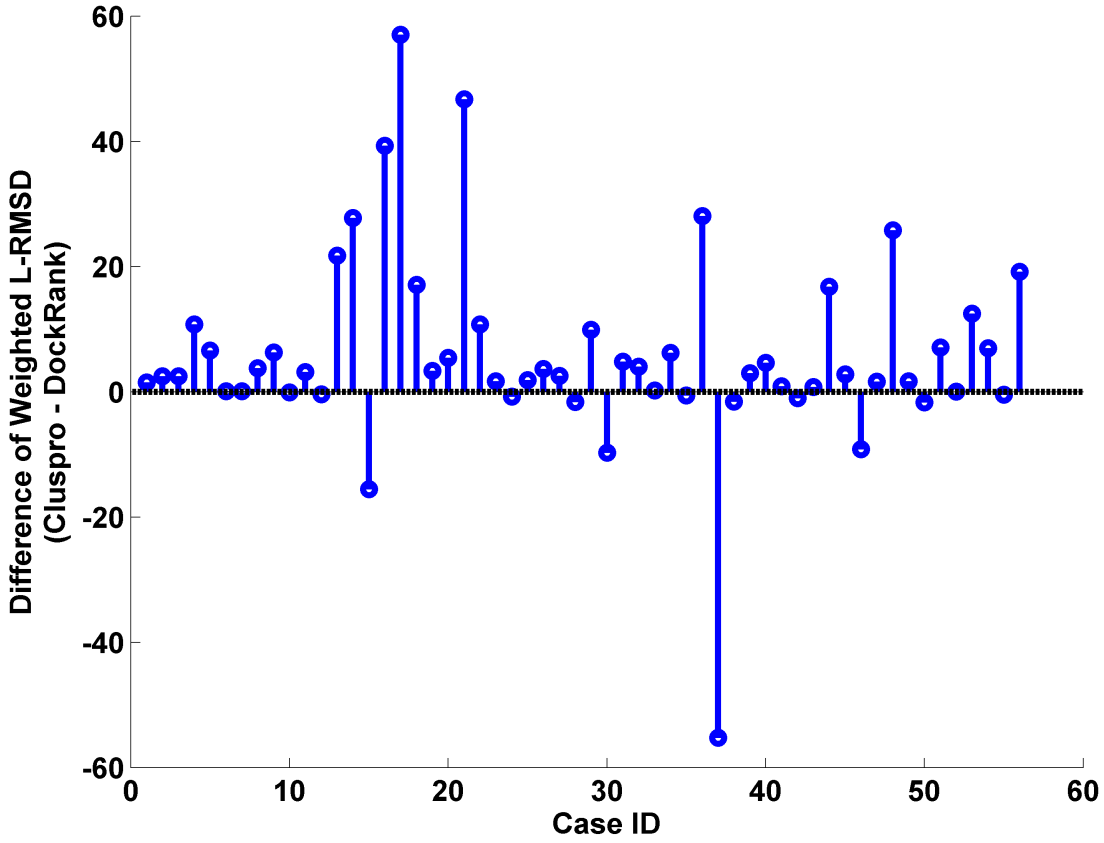


Figure 4.9 The difference between the weighted averages L-RMSDs of top models between DockRank and ClusPro Rank on each case of ClusPro2-BM3 decoy set. L-RMSD of each top model is weighted by its ranks. 56 cases with at least one docked model with L-RMSD $\leq$ 20 angstroms and can be ranked by DockRank using homo-interologs in Safe, Twilight or Dark zones are studied here. A positive dot means the top models ranked by DockRank for the specific case have a lower weighted L-RMSD than those ranked by ClusPro. For 40 out of 56 (71.4%) cases, top models ranked by DockRank have lower weighted L-RMSD than ClusPro.

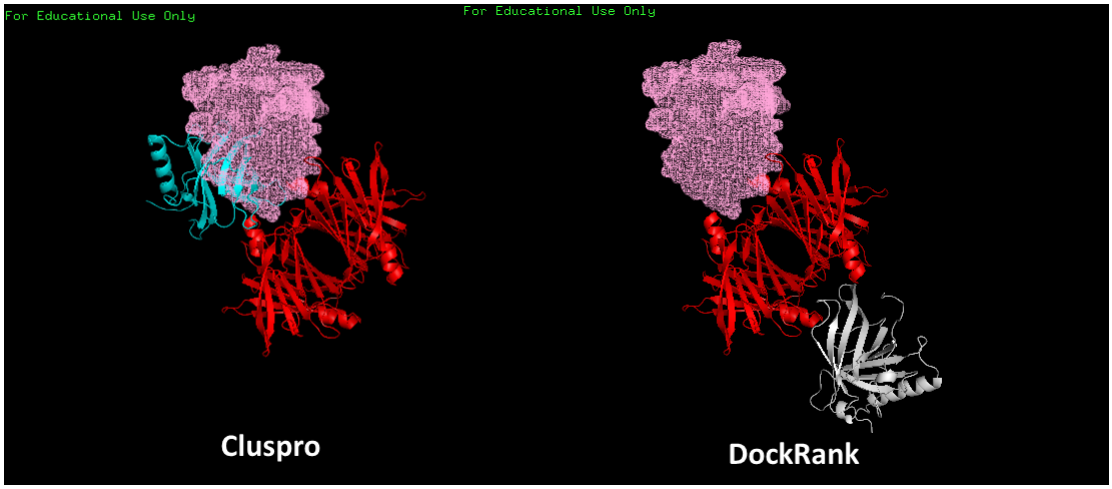


Figure 4.10 The comparison of top 1 models ranked by ClusPro and DockRank for case 1RLB. The red cartoon is the receptor in the top 1 docked model (the bound and docked receptors are superimposed, and bound receptor is not shown here). The ligand of the top 1 model ranked by ClusPro default cluster-size based method (blue ribbon in the left panel) is near the bound ligand position (pink mesh), however, the ligand of top 1 model selected by DockRank (white ribbon in the right panel) is totally wrong and is on the opposite side of bound ligand position (pink mesh) relative to the receptor (red ribbon). Note that the structure of the receptor is symmetric. So the natural question is that whether it is possible that the ligand might be able to bind on both sides of the symmetric receptor instead of on only one side?

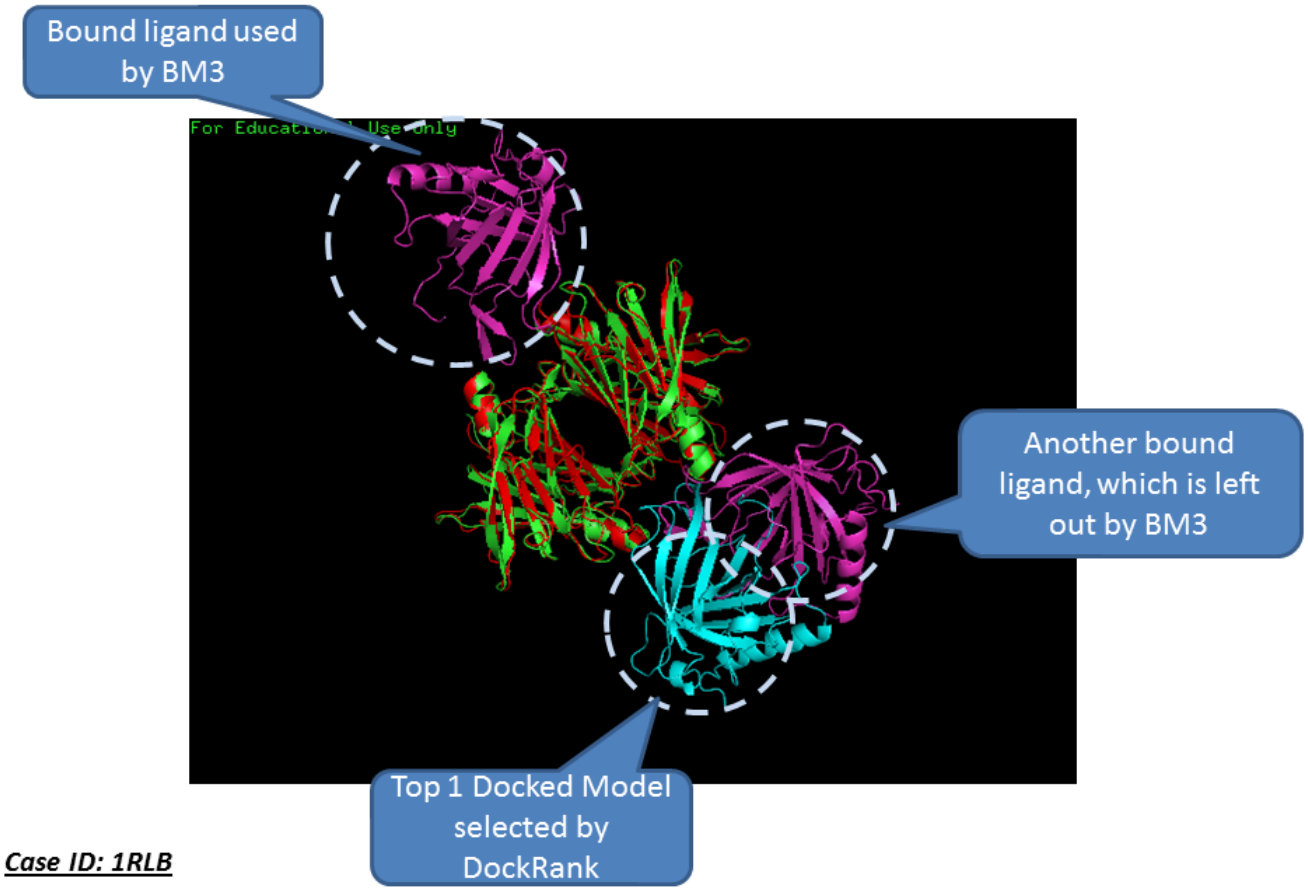


Figure 4.11 The top 1 model ranked by DockRank and two bound (native) ligands of case 1RLB. PDB entry 1RLB in fact has two bound ligands - chain E and F (purple ribbons). Chain E is included in BM3 dataset (purple ribbon in the left top corner, also shown as mesh in Figure) but chain F (purple ribbon on the lower right side) is arbitrarily left out of BM3 dataset. The ligand of the top 1 model selected by DockRank (blue ribbon) is right beside the left-out bound ligand.

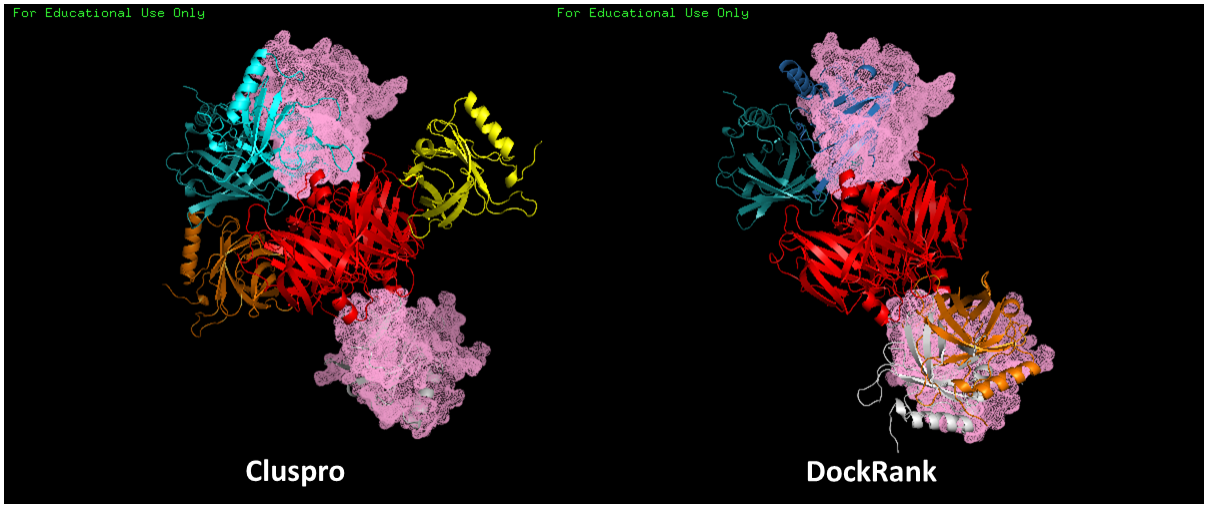


Figure 4.12 The top 5 models ranked by DockRank and ClusPro for case 1RLB. Both bound ligands (pink mesh) in the PDB entry 1RLB are shown here. DockRank (right panel) is able to give top ranks to the models with ligands that are near the native ligand positions (pink mesh) on both binding sides of the receptor (red ribbon). However, ClusPro (left panel) gives top ranks to not only the models with ligands on the two binding sides of the receptor, but also models with irrelevant ligand positions.

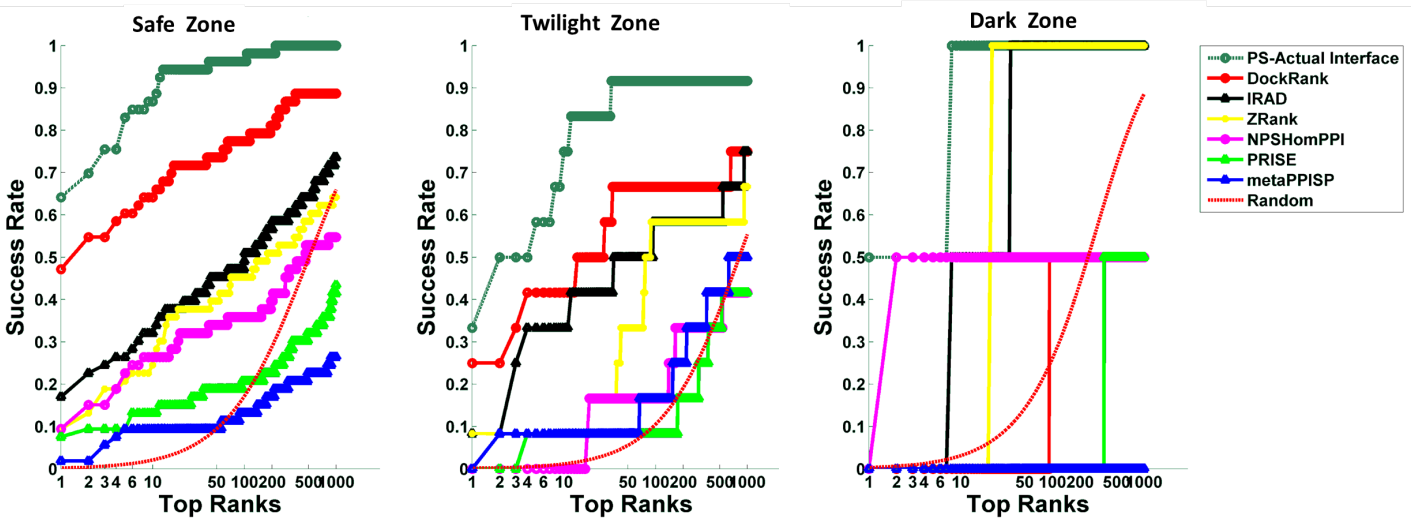


Figure 4.13 Success Rates in top 1-1000 models of different scoring functions on ZDock3-BM3 decoy set in different interface prediction confidence zones. Cases with more than one receptor-ligand chain pairs may have predicted interface from different confidence zones. Cases with solo confidence zones are studied here. 66 cases have only Safe Zone interface predictions, of which 53 cases have at least one hit. 16 cases have only Twilight interface predictions, of which 12 cases have at least one hit. 3 cases have only Dark interface predictions, of which 2 cases have at least one hit. DockRank (red solid line) achieves the most reliable performance in terms of Success Rates on cases with interface prediction confidence in Safe Zone (right panel). The Success Rates of DockRank on cases with Twilight zone prediction confidence declines, but are still consistently higher other scoring functions from top 1 to 1000 models. For the two cases in Dark Zone, DockRank was able to rank a hit to top 100 for one case, but could not find a hit for another case in top 1000 models. 54,000 models for each case were generated by ZDock 3.0.

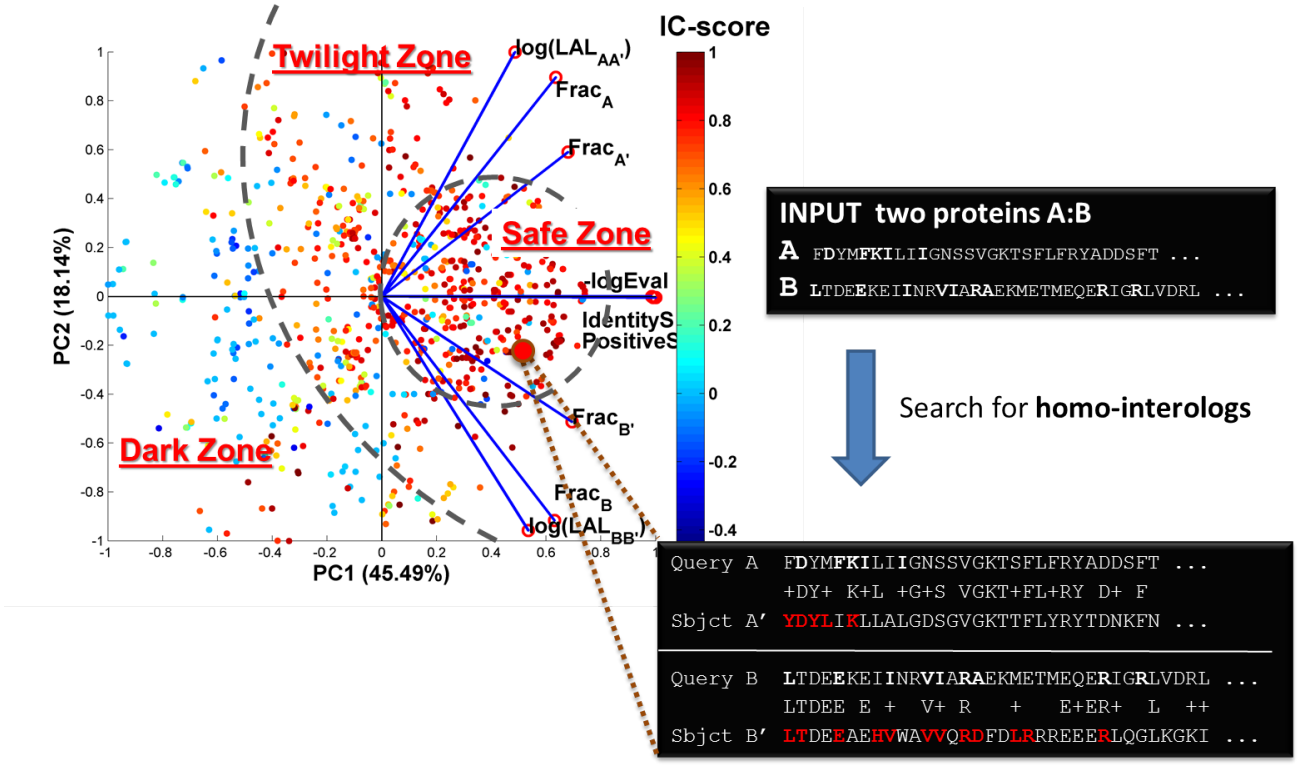


Figure 4.14 PS-HomPPI: Partner-specific sequence homology based protein-protein interface predictor. PS-HomPPI has two components: PS-interface conservation analysis and PS-interface prediction. PS-interface conservation analysis (shown as the PCA biplot on the left) is based on a dataset of 135 transient dimers with experimentally determined interface residues. For each dimer A-B, sequence homologs with known interfaces are retrieved. Each dot in the PCA biplot represents two sequence alignments: query A - its sequence homolog A', and query's partner B - its sequence homolog B'. Complex A'-B' is called a homo-interolog of A-B. 9 sequence alignment measures (blue lines with a red circle at the end) are calculated. An interface conservation score (IC-score) is calculated based on the similarity of the interfaces of A-B and those of A'-B'. The higher IC-score the more similar the interfaces of A-B and A'-B' are. IC-scores are represented using different colors: red for a high degree of interface conservation, and blue for low conservation. The original interface conservation space with 9 alignment measures and 1 IC-score was mapped to two dimensional PC1-PC2 space, where the relation of IC-score and the sequence alignment measures can be easily observed. Based on the color change (IC-score), three interface conservation zones are identified: Safe Zone for high level of interface conservation, Twilight Zone for medium level, and Dark Zone for low level. A regression model of IC-score with the sequence alignment measures is built. When making an interface prediction of a pair of proteins, a list of homo-interologs with known interfaces is searched. Sequence alignment measures are calculated for each query - homo-interolog. The regression model is used to rank the homo-interologs. Top K homo-interologs are used to make interface transfer, and their conservation zone provides a prediction confidence.

CHAPTER 5. Conclusions and Future Work

Our dissertation is guided by three hypotheses: 1) The interface residues are conserved among sequence homologs; 2) The conservation of interfaces among sequence homologs can be utilized to infer interfaces of interacting proteins; and 3) The predicted partner-specific interface residues can be used to reliably rank docked conformations.

To investigate these hypotheses, initially we conduct a systematic analysis of interface conservations of different types of protein-protein interactions, which indicates that interface residues are highly conserved among sequence homologs. Based on this result, we design and implement a family of reliable and computationally efficient predictors of protein-protein interface residues - HomPPI. One variant of this family predicts interface residues of a protein without specifying the binding partners, and the other variant predicts interface residues considering the information of potential binding partners. Our results show that the performance of the HomPPI family of predictors is superior to or compete with that of several state-of-the-art methods. Based on the success of our interface predictors, we design a scoring scheme to rank conformations generated by docking programs. This approach is based on the similarity between the predicted partner-specific interface residues and the interfaces formed in the docked conformations. Several experiments show that this ranking scheme selects a greater number of near-native conformations than several state-of-the-art energy-based scoring functions.

The results of this work will contribute to a better understanding of the physical and structural basis of protein interactions, and will facilitate the discovery and design of target-specific inhibitive drugs.

5.1 Conclusions

5.1.1 Protein-Protein Interface Positions Are Highly Conserved in Sequence Alignments

We conducted a large scale conservation analysis of experimentally determined interface residues. We studied more than 300,000 pair-wise alignments of protein sequences. These sequences include proteins that form obligate as well as transient interactions. We identified sequence similarity criteria for three interface conservation zones: Safe zone (interfaces are highly conserved among homologs), Twilight zone (medium level conservation) and Dark zone (low level conservation), which correspond to three interface prediction confidence zones.

Our main conclusions regarding protein-protein interface conservations are as follows:

1) *Interfaces of transient interacting proteins are highly conserved and partner-specific.* We studied the conservation of both non-partner-specific (NPS) and partner-specific (PS) interface of transient complexes. NPS-interface residues of a protein correspond to the union of the sets of its residues that make up its interfaces with all its known binding partners. PS-interface residues of a protein correspond to its residues that interact with a specific binding partner. Our results show that the PS-interface of transient complexes are clearly more conserved than the NPS-interface, indicating that transient interfaces are highly partner-specific, and that the partner-specific interfaces in transient complexes are, in fact, highly conserved.

2) *Interfaces of intrinsically disordered proteins (IDPs) are highly conserved and non-partner-specific.* The fact that intrinsically disordered interfaces can be reliably inferred by NPS-HomPPI (Figure 3.2) indicates that disordered interfaces are highly conserved and are non-partner-specific [1]. This conclusion about the interface of IDPs is consistent with findings that IDPs are able to bind a broad range of ligands through common binding regions [105, 116]. The high degree of conservation of interface (binding) regions in IDPs is consistent with their many important biological functions. Our finding is also consistent with the hypothesis that the flexibility of disordered binding regions may facilitate the binding of IDPs using the same

[1] Note that this result only applies to the interface residues of IDPs not their binding partners. The interfaces of IDPs' binding partners might/might not be partner-specific. We did not test NPS-HomPPI on the interface residues of IDP's binding partner, so no such conclusions can be extended to them.

set of binding residues to different binding partners (at different times) [54].

5.1.2 A Family of Sequence Homology based Protein-Protein Interface Predictors

We developed two variants of HomPPI, a family of sequence-based methods for predicting interface residues based on the experimentally determined interface residues in homologous sequences: NPS-HomPPI (Non-Partner-Specific HomPPI) and PS-HomPPI (Partner-Specific).

• NPS-HomPPI is a non-partner-specific sequence homology based protein-protein interface predictor. It predicts interface residues of a query protein without specified binding partner. Based on our evaluation results on two IDPs datasets in Chapter 3, NPS-HomPPI can be used to reliably predict the interfaces of IDPs (Intrinsically Disordered Proteins, which do not form structures in their unbound state, and render the structure-based interface predictors inoperable). IDPs have been implicated in cancer, cardiovascular disease, neurodegenerative disease, and diabetes [92, 126], and have become important drug targets [30, 92].

• PS-HomPPI predicts protein interface residues of a protein with respect to a specific putative binding partner. This is especially useful in the case of transient protein-protein interfaces, which tend to be highly partner-specific. Reliable partner-specific interface residue predictions have important implications in guiding site-directed mutagenesis of multi-faced hub proteins, in scoring docked conformations and guiding protein-protein docking process, and in designing inhibitors of interactions of specific proteins involved in disease pathways.

The results of HomPPI predictors show that:

1) *HomPPI outperforms several state-of-the-art machine learning-based interface predictors.* As shown by our comparison of NPS-HomPPI with five state-of-the-art protein interface prediction servers, including four servers that take advantage of the structures of the query proteins, NPS-HomPPI outperforms other methods, when the sequence homologs with experimentally determined interfaces of a query protein can be reliably identified. And PS-HomPPI further improves the prediction reliability on a transient dataset.

2) Unlike structure-based methods, *HomPPI is insensitive to conformational changes upon binding.* Medium to large conformational changes upon binding commonly exist in protein-protein interactions, and impose serious challenge upon in-silico prediction of interface residues

[154] and protein-protein dockings [140]. Whenever sequence homologs are available, HomPPI is able to provide reliable interface information for proteins regardless their conformational changes (Table 4.1), and we expect such information will efficiently reduce docking sampling space, and dramatically save computational time for further flexible refinement of docked conformations.

However, the HomPPI methods for interface residue prediction do have an important *limitation*, in that they rely on the availability putative homologs for which experimentally-determined structures of bound complexes are available in PDB. We report and discuss the prediction coverage of the HomPPI family of protein-protein interface prediction methods in "Directions for Future Research" section.

HomPPI is available as a set of freely accessible webservers at http://homppi.cs.iastate.edu/.

5.1.3 Partner-Specific Sequence Homology based Interface Prediction Significantly Improves The Ranking of Docked Conformations

Selecting near-native conformations from thousands of decoys generated by a docking program remains a challenging problem in computational molecular docking [62]. In this study, we presented DockRank - a novel predicted interface based scoring method for protein-protein docking. The proposed scoring function relies on a measure of similarity between interfaces of docked models and predicted interfaces by a PS interface predictor PS-HomPPI.

We evaluate DockRank on several representative docking decoys generated by different state-of-the-art docking programs. Our results show that:

1) *DockRank significantly and consistently outperforms several energy-based scoring functions in selecting near-native conformations when putative sequence homo-interologs can be reliably identified.* Comparisons of DockRank with two state-of-the-art energy based docking scoring functions, ZRank and IRAD, show that DockRank consistently outperforms ZRank and IRAD on a decoy set of 69 docking cases (with 54,000 decoys per case) in which PS-HomPPI can return predictions for the interfaces between the receptor and the ligand. These results suggest the viability of DockRank as an alternative to complex and computationally expensive energy based scoring functions in cases where it is possible to obtain reliable partner-specific interface predictions.

2) *Partner-specific interface prediction is significantly better than non-partner-specific interface predictions in ranking docked conformations.* We compared the performance of DockRank using PS-HomPPI predicted interface residues, with variants of DockRank using interface residues predicted by three state-of-the-art non-partner-specific protein interface residue predictors, including our sequence homology based method NPS-HomPPI [142], a local structural homology based method PRISE [74], a machine learning based consensus method meta-PPISP [112]. Our results show that, DockRank using interface residues predicted by the non-partner-specific interface predictors cannot compete with DockRank using interface residues predicted by PS-HomPPI and energy-based scoring functions IRAD and ZRank. This result is consistent with the observation made by Li and Kihara [81] that NPS interface predictors cannot efficiently rank docked conformations. However, the performance of DockRank using partner-specific interface predictions suggests that predicted interfaces can indeed be used to rank docked conformations more reliably than other state-of-the-art scoring functions.

DockRank is available as a freely accessible webserver at: http://einstein.cs.iastate.edu/DockRank/

.

5.2 Directions for Future Research

5.2.1 Improve the Prediction Coverage of HomPPI family

As any homology-based methods, HomPPI family of protein-protein interface predictors has an important limitation in that they rely on the availability of the sequence homologs with experimentally determined interfaces. The current prediction coverage of HomPPI accessed from our results on several non-redundant datasets are in the range of 60-70% of all query proteins (Different binding types have slightly different coverage, for example, obligate interactions and IDPs have better coverage than transient interactions (See "Prediction Coverage of HomPPI Methods" in the Discussion section of Chapter 3 for more details.).

One might improve the prediction coverage of HomPPI by combining HomPPI with sequence or structure based machine learning methods. Some preliminary evidence in support of this possibility is offered by our recent study of hybrid methods that combines a NPS sequence

homology based RNA-binding interface predictor and another machine learning classifier (Naive Bayes in [145] and SVMs in [135]) with extracted sequence features. Our hybrid methods of RNA-binding interface predictors not only solved the prediction coverage limitation of our homology-based method, but also are more reliable than some state-of-the-art structure-based predictors.

A particular interest is the design of partner-specific machine learning based interface predictors. Such methods that can efficiently make use of the information of binding partners could significantly advance the state-of-the-art interface predictions and their applications in protein-protein docking.

5.2.2 Incorporate More Binding Partners Into Predictions

The assumption of PS-HomPPI that the interfaces of protein interactions are independent and additive may not always hold. Currently our predictor PS-HomPPI predicts protein-protein interface residues for a protein with a specific putative binding partner. For an interaction that involves at least three proteins, suppose protein A, B and C are known to form one interaction complex, the user may input three query pairs A:B, A:C and B:C into PS-HomPPI. PS-HomPPI provides predictions for each of these three pairs assuming these queries pairs are not related. However, in nature, the interface of two proteins (or domains) may be affected by the presence of another nearby proteins (or domains), and some domains may compete with another domain on the same interfaces. Hence, methods that can take into account the information of multiple binding partners are of interest.

5.2.3 Constraining Docking with Partner-Specific Predictions

In protein-protein docking, there is an increasing interest in utilizing the information of interface residues to limit the search space of docked conformations around the known binding site [39, 40, 41, 38, 132, 81]. Constrained docking seeks to improve the quality of docked conformations and to lower the expensive computational cost of docking programs spent on globally sampling protein surfaces and on the scoring/ranking of the formidable large number of docked models.

Information about interface residues for constrained docking can be obtained from experimental data (e.g., alanine scanning and Mass Spectrometry) or from interface residue prediction methods. The quality of the resultant docked conformations relies on the reliability of the provided interface residues [DataDriveDocking05]. For interactions that are highly partner-specific (especially transient interactions), taking the interacting partner into account provides more accurate predictions than non-partner-specific predictions [142, 8], and hence one may expect predicted interfaces by such methods are more reliable for constraining data-driven docking programs.

APPENDIX A. THE EXPECTATION AND VARIANCE OF RANDOM SUCCESS RATE AND HIT RATE IN RANKING DOCKED MODELS IN CHAPTER 4

X_i: The number of hits in top K_i random ranked models.

N_i: Total docked models for case i.

M_i: The number of total hits in all N_i docked models of case i.

We know that $X_i \sim HG(N_i, M_i, K_i)$.

$X_1, X_2, ..., X_I$ are mutually independent, where I is the total number of cases.

$$Y_i = \begin{cases} 1 & X_i \geq 1 \\ 0 & o.w. \end{cases}$$

Y_i is the indicator of a success.

Because X_i $(i = 1, 2, ...I)$ are independent, $Y_i (i = 1, 2, ...I)$ are also mutually independent.

$Y = \sum_{i=1}^{I} Y_i$ is the number of cases that have at least one hit in top $K_i$ranked models, where I is the total number of cases.

Random Success Rate $Z = Y/I$.

The expectation of Random Success Rate Z:

$$
\begin{aligned}
E(Z) &= E(Y/I) = \frac{1}{I}E(Y) \\
&= \frac{1}{I}E(\sum_{i=1}^{I} Y_i) \\
&= \frac{1}{I}\sum_{i=1}^{I} E(Y_i) \\
&= \frac{1}{I}\sum_{i=1}^{I} Y_i P(Y_i = 1) \\
&= \frac{1}{I}\sum_{i=1}^{I} P(Y_i = 1) \\
&= \frac{1}{I}\sum_{i=1}^{I} P(X_i \geq 1) \\
&= \frac{1}{I}\sum_{i=1}^{I} [1 - P(X_i = 0)] \\
&= \frac{1}{I}\sum_{i=1}^{I} \left[1 - \frac{\begin{pmatrix} M_i \\ x_i = 0 \end{pmatrix} \begin{pmatrix} N_i - M_i \\ K_i - x_i \end{pmatrix}}{\begin{pmatrix} N_i \\ K_i \end{pmatrix}} \right] \\
&= \frac{1}{I}\sum_{i=1}^{I} \left[1 - \frac{\begin{pmatrix} N_i - M_i \\ K_i \end{pmatrix}}{\begin{pmatrix} N_i \\ K_i \end{pmatrix}} \right]
\end{aligned}
$$

The variance of Random Success Rate Z:

$$
\begin{aligned}
Var(Z) \quad &= \quad Var(Y/I) = \frac{1}{I^2}Var(Y) = \frac{1}{I^2}Var(\sum_{i=1}^{I} Y_i) \\
(independent) \quad &= \frac{1}{I^2}\sum_{i=1}^{I} Var(Y_i) \\
&= \frac{1}{I^2}\sum_{i=1}^{I}[E(Y_i^2) - E^2(Y_i)] \\
&= \frac{1}{I^2}\sum_{i=1}^{I}\{P(Y_i = 1) - [P(Y_i = 1)]^2\} \\
&= \frac{1}{I^2}\sum_{i=1}^{I}\{P(X_i \geq 1) - [P(X_i \geq 1)]^2\} \\
&= \frac{1}{I^2}\sum_{i=1}^{I}\left\{\left[1 - \frac{\binom{N_i - M_i}{K_i}}{\binom{N_i}{K_i}}\right] - \left[1 - \frac{\binom{N_i - M_i}{K_i}}{\binom{N_i}{K_i}}\right]^2\right\}
\end{aligned}
$$

Remark: Log and Exp were used to to overcome the overflow.

The Random Hit Rate for case i : $H_i = \frac{X_i}{M_i}$.

The expectation of Random Hit Rate H_i:

$$E(H_i) \quad = \quad E(\frac{X_i}{M_i}) = \frac{E(X_i)}{M_i} = \frac{\frac{K_i M_i}{N_i}}{M_i} = \frac{K_i}{N_i}$$

The variance of Random Hit Rate H_i:

$$Var(H_i) \quad = \quad Var(\frac{X_i}{M_i}) = \frac{1}{M_i^2}Var(X_i) = \frac{1}{M_i^2}K_i\frac{M_i(N_i - M_i)(N_i - K_i)}{N_i^2(N_i - 1)}$$

BIBLIOGRAPHY

[1] Blast substitution matrix [http://www.ncbi.nlm.nih.gov/blast/html/sub_matrix.html]. 34

[2] ProtinDB-PROTein-protein INterface residues Data Base [http://protindb.cs.iastate.edu/index.py]. 89

[3] S2C-A database correlating sequence and atomic coordinate residue numbering in the protein data bank [http://dunbrack.fccc.edu/guoli/s2c/index.php]. 89

[4] Abascal, F. and Valencia, A. (2003). Automatic annotation of protein function based on family identification. *Proteins: Structure, Function, and Bioinformatics*, 53(3):683–692. 14

[5] Acuner Ozbabacan, S., Engin, H., Gursoy, A., and Keskin, O. (2011a). Transient protein–protein interactions. *Protein Engineering Design and Selection*, 24(9):635–648. 6

[6] Acuner Ozbabacan, S., Engin, H., Gursoy, A., and Keskin, O. (2011b). Transient protein–protein interactions. *Protein Engineering Design and Selection*, 24(9):635–648. 72

[7] Aebersold, R., Mann, M., et al. (2003). Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207. 3, 41

[8] Ahmad, S. and Mizuguchi, K. (2011). Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. *PLoS ONE*, 6(12):e29104. 71, 112

[9] Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402. 16, 32, 88

[10] Andrade, M. (1999). Position-specific annotation of protein function based on multiple homologs. In *ISMB*, volume 7, pages 28–33. 14

[11] Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Engineering*, 2(2):101–113. 4, 41

[12] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412. 60

[13] Berman, H. (2007). The protein data bank: a historical perspective. *Acta Crystallographica Section A: Foundations of Crystallography*, 64(1):88–95. 80

[14] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242. 2, 15, 32, 51

[15] Bernauer, J., Aze, J., Janin, J., and Poupon, A. (2007). A new protein protein docking scoring function based on interface residue properties. *Bioinformatics*, 23(5):555. 71

[16] Blundell, T., Jhoti, H., and Abell, C. (2002). High-throughput crystallography for lead discovery in drug design. *Nature Reviews Drug Discovery*, 1(1):45–54. 41

[17] Bogan, A. and Thorn, K. (1998). Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1):1–9. 4, 41

[18] Bordner, A. and Gorin, A. (2007). Protein docking using surface matching and supervised machine learning. *Proteins: Structure, Function, and Bioinformatics*, 68(2):488–502. 7, 70

[19] Bourquard, T., Bernauer, J., Azé, J., and Poupon, A. (2011). A collaborative filtering approach for protein-protein docking scoring functions. *PloS one*, 6(4):e18541. 71

[20] Bradford, J. and Westhead, D. (2005). Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487. 45, 58

[21] Bradford, J. and Westhead, D. (2008). Machine learning in computational biology. 45

[22] Caffrey, D., Somaroo, S., Hughes, J., Mintseris, J., and Huang, E. (2004). Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science*, 13(1):190–202. 8, 14

[23] Cavanagh, J. (2007). *Protein NMR spectroscopy: Principles and practice*. Academic Pr. 3, 41

[24] Chakrabarti, P. and Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins: Structure, Function, and Bioinformatics*, 47(3):334–343. 4, 5, 41

[25] Chandonia, J. and Brenner, S. (2006). The impact of structural genomics: expectations and outcomes. *Science*, 311(5759):347. 55

[26] Chen, H. and Zhou, H. (2005a). Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against nmr data. *Proteins: Structure, Function, and Bioinformatics*, 61(1):21–35. 41, 42, 48

[27] Chen, H. and Zhou, H. (2005b). Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against nmr data. *Proteins: Structure, Function, and Bioinformatics*, 61(1):21–35. 75

[28] Chen, P. and Li, J. (2010). Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information. *BMC bioinformatics*, 11(1):402. 41

[29] Chen, X. and Jeong, J. (2009). Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics*, 25(5):585–591. 41, 42

[30] Cheng, Y., LeGall, T., Oldfield, C., Mueller, J., Van, Y., Romero, P., Cortese, M., Uversky, V., and Dunker, A. (2006). Rational drug design via intrinsically disordered protein. *Trends in biotechnology*, 24(10):435–442. 108

[31] Cho, Y. and Zhang, A. (2010). Identification of functional hubs and modules by converting interactome networks into hierarchical ordering of proteins. *BMC bioinformatics*, 11(Suppl 3):S3. 56

[32] Choi, Y., Yang, J., Choi, Y., Ryu, S., and Kim, S. (2009). Evolutionary conservation in multiple faces of protein interaction. *Proteins: Structure, Function, and Bioinformatics*, 77(1):14–25. 8, 14, 15, 24, 28

[33] Chung, J., Wang, W., and Bourne, P. (2007). High-throughput identification of interacting protein-protein binding sites. *BMC bioinformatics*, 8(1):223. 71

[34] Comeau, S., Gatchell, D., Vajda, S., and Camacho, C. (2004a). Cluspro: a fully automated algorithm for protein–protein docking. *Nucleic acids research*, 32(suppl 2):W96–W99. 72, 76, 86

[35] Comeau, S., Gatchell, D., Vajda, S., and Camacho, C. (2004b). Cluspro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, 20(1):45. 7, 70, 72, 76, 86

[36] Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *Journal of molecular biology*, 285(5):2177–2198. 4, 41

[37] de Vries, S. and Bonvin, A. (2008). How proteins get in touch: interface prediction in the study of biomolecular complexes. *Current protein and peptide science*, 9(4):394–406. 5, 41, 42, 48

[38] de Vries, S. and Bonvin, A. (2011). Cport: a consensus interface predictor and its performance in prediction-driven docking with haddock. *PLoS One*, 6(3):e17695. 7, 41, 73, 86, 111

[39] de Vries, S., van Dijk, A., and Bonvin, A. (2006). Whiscy: What information does surface conservation yield? application to data-driven docking. *Proteins: Structure, Function, and Bioinformatics*, 63(3):479–489. 7, 86, 111

[40] de Vries, S., van Dijk, A., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T., and Bonvin, A. (2007). Haddock versus haddock: new features and performance of haddock2. 0 on the capri targets. *Proteins: structure, function, and bioinformatics*, 69(4):726–733. 86, 111

[41] de Vries, S., van Dijk, M., and Bonvin, A. (2010). The haddock web server for data-driven biomolecular docking. *nature protocols*, 5(5):883–897. 7, 41, 111

[42] Deisenhofer, J., Epp, O., Miki, K., Huber, R., and Michel, H. (1984). X-ray structure analysis of a membrane protein complex:: Electron density map at 3 å resolution and a model of the chromophores of the photosynthetic reaction center from rhodopseudomonas viridis. *Journal of molecular biology*, 180(2):385–398. 41

[43] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30. 76

[44] Dominguez, C., Boelens, R., and Bonvin, A. (2003). Haddock: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7):1731–1737. 7, 86

[45] Dosztányi, Z., Mészáros, B., and Simon, I. (2009). Anchor: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, 25(20):2745–2746. 49

[46] Drenth, J. (1999). *Principles of protein X-ray crystallography*. Springer Verlag. 2, 41

[47] Duhovny, D., Nussinov, R., and Wolfson, H. (2002). Efficient unbound docking of rigid molecules. *Algorithms in Bioinformatics*, pages 185–200. 7

[48] Dunker, A., Obradovic, Z., et al. (2001). The protein trinity-linking function and disorder. *Nature biotechnology*, 19(9):805–806. 42

[49] Dunker, A., Obradovic, Z., Romero, P., Garner, E., Brown, C., et al. (2000). Intrinsic protein disorder in complete genomes. *GENOME INFORMATICS SERIES*, pages 161–171. 42

[50] Ezkurdia, I., Bartoli, L., Fariselli, P., Casadio, R., Valencia, A., and Tress, M. (2009a). Progress and challenges in predicting protein–protein interaction sites. *Briefings in bioinformatics*, 10(3):233–246. 2

[51] Ezkurdia, I., Bartoli, L., Fariselli, P., Casadio, R., Valencia, A., and Tress, M. (2009b). Progress and challenges in predicting protein–protein interaction sites. *Briefings in bioinformatics*, 10(3):233–246. 71

[52] Ezkurdia, I., Bartoli, L., Fariselli, P., Casadio, R., Valencia, A., and Tress, M. (2009c). Progress and challenges in predicting protein-protein interaction sites. *Briefings in bioinformatics*, 10(3):233–246. 5, 41

[53] Fong, J. and Panchenko, A. (2010). Intrinsic disorder and protein multibinding in domain, terminal, and linker regions. *Mol. BioSyst.*, 6(10):1821–1828. 10, 43

[54] Fong, J., Shoemaker, B., Garbuzynskiy, S., Lobanov, M., Galzitskaya, O., and Panchenko, A. (2009). Intrinsic disorder in protein interactions: insights from a comprehensive structural analysis. *PLoS computational biology*, 5(3):e1000316. 108

[55] Gao, M. and Skolnick, J. (2010). Structural space of protein–protein interfaces is degenerate, close to complete, and highly connected. *Proceedings of the National Academy of Sciences*, 107(52):22517. 28

[56] Garcia, R., Pantazatos, D., and Villarreal, F. (2004). Hydrogen/deuterium exchange mass spectrometry for investigating protein-ligand interactions. *Assay and drug development technologies*, 2(1):81–91. 3, 41

[57] Ginalski, K. (2006). Comparative modeling for protein structure prediction. *Current opinion in structural biology*, 16(2):172–177. 8, 14

[58] Glaser, F., Steinberg, D., Vakser, I., and Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 43(2):89–102. 4, 41

[59] Gray, J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C., and Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology*, 331(1):281–299. 7, 70

[60] Grishin, N. and Phillips, M. (1994). The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *protein Science*, 3(12):2455–2458. 8, 14, 15

[61] Guharoy, M. and Chakrabarti, P. (2007). Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein–protein interactions. *Bioinformatics*, 23(15):1909–1918. 4

[62] Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. (2002). Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics*, 47(4):409–443. 7, 69, 71, 81, 109

[63] Heuser, P. and Schomburg, D. (2006). Optimised amino acid specific weighting factors for unbound protein docking. *BMC bioinformatics*, 7(1):344. 7, 70

[64] Huang, B. and Schroeder, M. (2008). Using protein binding site prediction to improve protein docking. *Gene*, 422(1-2):14–21. 7, 70, 71

[65] Hue, M., Riffle, M., Vert, J., and Noble, W. (2010). Large-scale prediction of protein-protein interactions from structures. *BMC bioinformatics*, 11(1):144. 56

[66] Hwang, H., Pierce, B., Mintseris, J., Janin, J., and Weng, Z. (2008). Protein–protein docking benchmark version 3.0. *Proteins: Structure, Function, and Bioinformatics*, 73(3):705–709. 86

[67] Hwang, H., Vreven, T., Pierce, B., Hung, J., and Weng, Z. (2010). Performance of zdock and zrank in capri rounds 13–19. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3104–3110. 72, 73, 86

[68] Janin, J. (2010). Protein-protein docking tested in blind predictions: the capri experiment. *Molecular bioSystems*, 6(12):2351. 41, 42

[69] Janin, J., Bahadur, R., and Chakrabarti, P. (2008). Protein-protein interaction and quaternary structure. *Quarterly reviews of biophysics*, 41(02):133–180. 5

[70] Johnson, R. and Wichern, D. (2002). *Applied multivariate statistical analysis*, volume 4. Prentice Hall Upper Saddle River, NJ. 17

[71] Jones, S., Marin, A., and M Thornton, J. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Engineering*, 13(2):77. 4, 41

[72] Jones, S. and Thornton, J. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13. 4, 41

[73] Jones, S. and Thornton, J. (1997). Analysis of protein-protein interaction sites using surface patches. *Journal of molecular biology*, 272(1):121. 4, 41

[74] Jordan, R., Yasser, E., Dobbs, D., and Honavar, V. (2012). Predicting protein-protein interface residues using local surface structural similarity. *BMC bioinformatics*, 13(1):41. 74, 110

[75] Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A., Aflalo, C., and Vakser, I. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences*, 89(6):2195. 7, 70

[76] Kozakov, D., Brenke, R., Comeau, S., and Vajda, S. (2006). Piper: An fft-based protein docking program with pairwise potentials. *Proteins: Structure, Function, and Bioinformatics*, 65(2):392–406. 7, 70, 72, 76, 77

[77] Kozakov, D., Hall, D., Beglov, D., Brenke, R., Comeau, S., Shen, Y., Li, K., Zheng, J., Vakili, P., Paschalidis, I., et al. (2010). Achieving reliability and high accuracy in automated protein docking: Cluspro, piper, sdu, and stability analysis in capri rounds 13–19. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3124–3130. 7, 70, 72, 76, 86

[78] Kufareva, I., Budagyan, L., Raush, E., Totrov, M., and Abagyan, R. (2007). Pier: protein interface recognition for structural proteomics. *Proteins: Structure, Function, and Bioinformatics*, 67(2):400–417. 48

[79] Larsen, T., Olson, A., and Goodsell, D. (1998). Morphology of protein-protein interfaces. *Structure*, 6(4):421–427. 4, 41

[80] Lensink, M., Méndez, R., and Wodak, S. (2007). Docking and scoring protein complexes: Capri 3rd edition. *Proteins: Structure, Function, and Bioinformatics*, 69(4):704–718. 69, 86

[81] Li, B. and Kihara, D. (2012). Protein docking prediction using predicted protein-protein interface. *BMC bioinformatics*, 13(1):7. 7, 41, 110, 111

[82] Li, L., Chen, R., and Weng, Z. (2003). Rdock: Refinement of rigid-body protein docking predictions. *Proteins: Structure, Function, and Bioinformatics*, 53(3):693–707. 7, 70

[83] Liang, S., Zhang, C., Liu, S., and Zhou, Y. (2006). Protein binding site prediction using an empirical scoring function. *Nucleic acids research*, 34(13):3698–3707. 41, 42, 48, 75

[84] Lijnzaad, P., Berendsen, H., and Argos, P. (1996). Hydrophobic patches on the surfaces of protein structures. *Proteins: Structure, Function, and Bioinformatics*, 25(3):389–397. 4, 41

[85] Liu, S. and Vakser, I. (2011). Deck: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. *BMC bioinformatics*, 12(1):280. 7, 70, 73

[86] Loewenstein, Y., Raimondo, D., Redfern, O., Watson, J., Frishman, D., Linial, M., Orengo, C., Thornton, J., Tramontano, A., et al. (2009). Protein function annotation by homology-based inference. *Genome Biol*, 10(2):207. 8, 14

[87] London, N. and Schueler-Furman, O. (2008). Funnel hunting in a rough terrain: learning and discriminating native energy funnels. *Structure*, 16(2):269–279. 7, 70

[88] Martin, O. and Schomburg, D. (2008). Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 70(4):1367–1378. 7, 70

[89] Matthews, L., Vaglio, P., Reboul, J., Ge, H., Davis, B., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for

conserved protein-protein interactions or "interologs". *Genome research*, 11(12):2120–2126. 14

[90] Méndez, R., Leplae, R., De Maria, L., and Wodak, S. (2003). Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins: Structure, Function, and Bioinformatics*, 52(1):51–67. 69

[91] Mészáros, B., Simon, I., and Dosztányi, Z. (2009). Prediction of protein binding regions in disordered proteins. *PLoS computational biology*, 5(5):e1000376. 49, 58

[92] Metallo, S. (2010). Intrinsically disordered proteins are potential drug targets. *Current opinion in chemical biology*, 14(4):481–488. 10, 43, 49, 108

[93] Mintseris, J. and Weng, Z. (2003). Atomic contact vectors in protein-protein recognition. *Proteins: Structure, Function, and Bioinformatics*, 53(3):629–639. 5

[94] Mintseris, J. and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):10930. 16, 24, 32, 51, 59, 88

[95] Moont, G., Gabb, H., and Sternberg, M. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins: Structure, Function, and Bioinformatics*, 35(3):364–373. 7, 70

[96] Morrison, K. and Weiss, G. (2001). Combinatorial alanine-scanning. *Current opinion in chemical biology*, 5(3):302–307. 3, 41

[97] Murakami, Y. and Mizuguchi, K. (2010). Applying the naïve bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics*, 26(15):1841. 41, 48

[98] Nair, R. and Rost, B. (2002). Sequence conserved for subcellular localization. *Protein Science*, 11(12):2836–2847. 14

[99] Neuvirth, H., Raz, R., and Schreiber, G. (2004). Promate: a structure based prediction program to identify the location of protein-protein binding sites. *Journal of molecular biology*, 338(1):181–199. 7, 41, 42, 48, 71, 75

[100] Nooren, I. and Thornton, J. (2003a). Diversity of protein–protein interactions. *The EMBO journal*, 22(14):3486–3492. 71

[101] Nooren, I. and Thornton, J. (2003b). Structural characterisation and functional significance of transient protein-protein interactions. *Journal of molecular biology*, 325(5):991–1018. 6

[102] Nooren, I. and Thornton, J. (2003c). Structural characterisation and functional significance of transient protein-protein interactions. *Journal of molecular biology*, 325(5):991–1018. 72

[103] Ofran, Y. and Rost, B. (2003). Predicted protein-protein interaction sites from local sequence information. *FEBS letters*, 544(1-3):236–239. 4, 41

[104] Ofran, Y. and Rost, B. (2007). Isis: interaction sites identified from sequence. *Bioinformatics*, 23(2):e13–e16. 41, 48

[105] Oldfield, C., Meng, J., Yang, J., Yang, M., Uversky, V., and Dunker, A. (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *Bmc Genomics*, 9(Suppl 1):S1. 107

[106] Pan, X., Zhang, Y., and Shen, H. (2010). Large scale prediction of human protein protein interactions from amino acid sequence based on latent topic features. *Journal of Proteome Research*. 56

[107] Panchenko, A., Kondrashov, F., and Bryant, S. (2004). Prediction of functional sites by analysis of sequence and structure conservation. *Protein science*, 13(4):884–892. 72

[108] Pazos, F., Helmer-Citterich, M., Ausiello, G., Valencia, A., et al. (1997). Correlated mutations contain information about protein-protein interaction. *Journal of molecular biology*, 271(4):511. 41

[109] Philipp, H. and Dietmar, S. (2007). Combination of scoring schemes for protein docking. *BMC Bioinformatics*, 8:279. 7, 70

[110] Pierce, B. and Weng, Z. (2007). Zrank: reranking protein docking predictions with an optimized energy function. *PROTEINS: Structure, Function, and Bioinformatics*, 67(4):1078–1086. 7, 70, 72

[111] Porollo, A. and Meller, J. (2007). Prediction-based fingerprints of protein–protein interactions. *PROTEINS: Structure, Function, and Bioinformatics*, 66(3):630–645. 33, 41, 42, 48, 59

[112] Qin, S. and Zhou, H. (2007). meta-ppisp: a meta web server for protein-protein interaction site prediction. *Bioinformatics*, 23(24):3386. 7, 48, 75, 110

[113] Reddy, B. and Kaznessis, Y. (2005). A quantitative analysis of interfacial amino acid conservation in protein-protein hetero complexes. *Journal of Bioinformatics and Computational Biology*, 3(5):1137–1150. 14

[114] Reš, I., Mihalek, I., and Lichtarge, O. (2005). An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, 21(10):2496. 41

[115] Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein engineering*, 12(2):85–94. 14

[116] Russell, R. and Gibson, T. (2008). A careful disorderliness in the proteome: sites for interaction and targets for future therapies. *FEBS letters*, 582(8):1271–1275. 107

[117] Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, 9(1):56–68. 14

[118] Sheinerman, F., Norel, R., and Honig, B. (2000). Electrostatic aspects of protein-protein interactions. *Current Opinion in Structural Biology*, 10(2):153–159. 4, 41

[119] Šikić, M., Tomić, S., and Vlahoviček, K. (2009). Prediction of protein–protein interaction sites in sequences and 3d structures by random forests. *PLoS computational biology*, 5(1):e1000278. 41

[120] Sinz, A. (2003). Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *Journal of mass spectrometry*, 38(12):1225–1237. 3, 41

[121] Tompa, P., Fuxreiter, M., Oldfield, C., Simon, I., Dunker, A., and Uversky, V. (2009). Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*, 31(3):328–335. 10, 43

[122] Tovchigrechko, A. and Vakser, I. (2006). Gramm-x public web server for protein–protein docking. *Nucleic acids research*, 34(suppl 2):W310–W314. 41

[123] Tsai, C., Lin, S., Wolfson, H., and Nussinov, R. (1997). Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect. *Protein Science*, 6(1):53–64. 4, 41

[124] Tuncbag, N., Gursoy, A., and Keskin, O. (2009a). Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, 25(12):1513–1520. 4

[125] Tuncbag, N., Kar, G., Keskin, O., Gursoy, A., and Nussinov, R. (2009b). A survey of available tools and web servers for analysis of protein–protein interactions and interfaces. *Briefings in Bioinformatics*, 10(3):217–232. 71

[126] Uversky, V., Oldfield, C., and Dunker, A. (2008). Intrinsically disordered proteins in human diseases: introducing the d2 concept. *Annu. Rev. Biophys.*, 37:215–246. 108

[127] Vacic, V., Oldfield, C., Mohan, A., Radivojac, P., Cortese, M., Uversky, V., and Dunker, A. (2007). Characterization of molecular recognition features, morfs, and their binding partners. *Journal of proteome research*, 6(6):2351–2366. 10, 43

[128] Vajda, S. and Kozakov, D. (2009). Convergence and combination of methods in protein-protein docking. *Current opinion in structural biology*, 19(2):164–170. 7, 69

[129] Valdar, W. and Thornton, J. (2001). Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins: Structure, Function, and Bioinformatics*, 42(1):108–124. 8

[130] Valencia, A. (2005). Automatic annotation of protein function. *Current Opinion in Structural Biology*, 15(3):267–274. 14

[131] Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, 12(3):368–373. 56

[132] Van Dijk, A., De Vries, S., Dominguez, C., Chen, H., Zhou, H., and Bonvin, A. (2005). Data-driven docking: Haddock's adventures in capri. *Proteins: Structure, Function, and Bioinformatics*, 60(2):232–238. 7, 111

[133] Vreven, T., Hwang, H., and Weng, Z. (2011). Integrating atom-based and residue-based scoring functions for protein–protein docking. *Protein Science*. 7, 70, 72, 73

[134] Walhout, A., Sordella, R., Lu, X., Hartley, J., Temple, G., Brasch, M., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science*, 287(5450):116. 8, 14

[135] Walia, R., Xue, L. C., Wilkins, K., Yasser, E., Dobbs, D., and Honavar, V. (2012). Robust prediction of rna-binding sites in proteins using a combination of sequence homology and machine learning methods. *To be submitted*. 111

[136] Wang, G., Arthur, J., and Dunbrack, R. (2002). S2c: a database correlating sequence and atomic coordinate numbering in the protein data bank. 32

[137] Wang, G. and Dunbrack, R. (2003). Pisces: a protein sequence culling server. *Bioinformatics*, 19(12):1589. 32

[138] Wang, H., Segal, E., Ben-Hur, A., Li, Q., Vidal, M., and Koller, D. (2007). Insite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome biology*, 8(9):R192. 71

[139] Wells, J. and McClendon, C. (2007). Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature*, 450(7172):1001–1009. 1

[140] Wodak, S. et al. (2004). Prediction of protein-protein interactions: the capri experiment, its evaluation and implications. *Current opinion in structural biology*, 14(2):242–249. 42, 109

[141] Wolfson, H. and Rigoutsos, I. (1997). Geometric hashing: An overview. *Computational Science & Engineering, IEEE*, 4(4):10–21. 7

[142] Xue, L., Dobbs, D., and Honavar, V. (2011a). Homppi: a class of sequence homology based protein-protein interface prediction methods. *BMC bioinformatics*, 12(1):244. 6, 11, 71, 74, 79, 81, 85, 87, 89, 110, 112

[143] Xue, L., Jordan, R., El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2011b). Ranking docked models of protein-protein complexes using predicted partner-specific protein-protein interfaces: A preliminary study. In *BCB '11 Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 441–445. ACM. 56

[144] Xue, L., Jordan, R., Yasser, E., Dobbs, D., and Honavar, V. (2010). Ranking docked models of protein-protein complexes using predicted partner-specific protein-protein interfaces: A preliminary study. 87

[145] Xue, L., Walia, R., El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2011c). Improving protein-rna interface prediction by combining sequence homology based method with a naive bayes classifier: preliminary results. In *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 556–558. ACM. 111

[146] Yan, C., Dobbs, D., and Honavar, V. (2003). Identification of surface residues involved in protein-protein interaction–a support vector machine approach. *Intelligent Systems Design and Applications (ISDA-03)*, pages 53–62. 41

[147] Yan, C., Dobbs, D., and Honavar, V. (2004). A two-stage classifier for identification of protein–protein interface residues. *Bioinformatics*, 20(suppl 1):i371–i378. 41

[148] Yan, C., Wu, F., Jernigan, R., Dobbs, D., and Honavar, V. (2008). Characterization of protein-protein interfaces. *The protein journal*, 27(1):59–70. 4

[149] Yip, K., Kim, P., McDermott, D., and Gerstein, M. (2009). Multi-level learning: improving the prediction of protein, domain and residue interactions by allowing information flow between levels. *BMC bioinformatics*, 10(1):241. 40, 56

[150] Yu, H., Luscombe, N., Lu, H., Zhu, X., Xia, Y., Han, J., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: protein–protein interologs and protein–dna regulogs. *Genome research*, 14(6):1107–1118. 8, 14

[151] Zhang, B. and Srihari, S. (2003). Binary vector dissimilarity measures for handwriting identification. In *Proc. of SPIE Vol*, volume 5010, page 29. 90

[152] Zhang, Q., Petrey, D., Norel, R., and Honig, B. (2010). Protein interface conservation across structure space. *Proceedings of the National Academy of Sciences*, 107(24):10896. 28

[153] Zhou, H. and Qin, S. (2007a). Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, 23(17):2203–2209. 5, 41, 42, 48, 71

[154] Zhou, H. and Qin, S. (2007b). Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, 23(17):2203–2209. 109

[155] Zhou, H. and Shan, Y. (2001a). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Structure, Function, and Bioinformatics*, 44(3):336–343. 48

[156] Zhou, H. and Shan, Y. (2001b). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Structure, Function, and Bioinformatics*, 44(3):336–343. 75

[157] Zuiderweg, E. (2002). Mapping protein-protein interactions in solution by nmr spectroscopy. *Biochemistry*, 41(1):1–7. 3, 41