2014

Prediction of protein-protein interaction types using machine learning approaches

Mina Maleki
University of Windsor

PREDICTION OF PROTEIN-PROTEIN INTERACTION TYPES USING MACHINE LEARNING APPROACHES

by
Mina Maleki

A Dissertation
Submitted to the Faculty of Graduate Studies
through School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy at the
University of Windsor

Windsor, Ontario, Canada
2014

PREDICTION OF PROTEIN-PROTEIN INTERACTION TYPES USING MACHINE LEARNING APPROACHES

by

MINA MALEKI

APPROVED BY:

R. Alhajj, External Examiner
Department of Computer Science, University of Calgary

W. Crosby
Department of Biological Sciences

A. Ngom
School of Computer Science

D. Wu
School of Computer Science

L. Rueda, Advisor
School of Computer Science

13 May, 2014

Declaration of Co-Authorship and Previous Publications

I. Co-Authorship Declaration:

I hereby declare that this thesis incorporates material that is result of joint research undertaken in collaboration with Muhammad Aziz, Gokul Vasudev and Micheal Hall under the supervision of professor Luis Rueda. The collaboration is covered in Chapters 2 to 7 of the thesis. In most cases, the key ideas, primary contributions, experimental designs, applying and optimizing different machine learning methods for prediction, numerical and visual analysis and interpretation and writing the papers, were performed by the author, and the contribution of coauthors was primarily done through the extraction of features for prediction which was also re-implemented by the author.

Also, the paper reported in Chapter 8 of the thesis is written in collaboration with Mohammad Haj Dezfulian under the supervision of professors William Crosby and Luis Rueda. Mohammad prepared the SCF-ligase dataset and helped me in the validation of the biological results and proof-reading the paper while other computational tasks and writing the paper were performed by the author.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have

obtained written permission from each of the co-author(s) to include the above material(s) in my thesis. I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

II. Declaration of Previous Publication:

This thesis includes 7 original papers that have been previously published/submitted for publication in peer reviewed journals and conferences, as follows:

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as graduate student at the University of Windsor.

I declare that, to the best of my knowledge, my thesis does not infringe upon anyones copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Thesis Chapter	Publication title	Publication status
Chapter 2	Md. Aziz, M. Maleki, L. Rueda, M.Raza, S. Banerjee, "Prediction of Biological Protein-protein Interactions using Atom-type and Amino Acid Properties," Wiley-VCH Proteomics, vol. 11, no. 19, pp. 3802-10, Aug. 2011.	published
Chapter 3	M. Maleki, Md. Aziz, L. Rueda, "Analysis of Relevant Physicochemical Properties in Obligate and Non-obligate Protein-protein Interactions," in Workshop on Computational Structural Bioinformatics in conjunction with BIBM 2011, GA, USA, Nov. 2011.	published
Chapter 4	M. Maleki, G. Vasudev, L. Rueda, "The Role of Electrostatic Energies in Prediction of Obligate Protein-Protein Interactions," Journal of BMC Proteome Science, Nov. 2013.	published
Chapter 5	M. Maleki, M. Hall, L. Rueda, "Using Desolvation Energies of Structural Domains to Predict Stability of Protein Complexes," Journal of Network Modeling Analysis in Health Informatics and Bioinformatics (NetMAHIB), vol. 2, no. 4, pp. 267275, Dec. 2013.	published
Chapter 6	M. Maleki, M. Hall, L. Rueda, "Using Structural Domain to Predict Obligate and Non-obligate Protein-protein Interactions," in 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB2012), California, USA, May 2012.	published
Chapter 7	M. Maleki, Md. Aziz, L. Rueda, "Analysis of Obligate and Non-obligate Complexes using Desolvation Energies in Domain-domain Interactions," in 10th International Workshop on Data Mining in Bioinformatics (BIOKDD2011) in conjunction with ACM SIGKDD 2011, San Diego, USA, Aug. 2011.	published
Chapter 8	M. Maleki, M. Dezfulian, W. Crosby, L. Rueda, "Computational Analysis of the Stability of SCF Ligases Employing Domain Information," in 5th ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACMBCB), CA, 2014.	submitted

Abstract

Prediction and analysis of protein-protein interactions (PPI) is an important problem in life science research because of the fundamental roles of PPIs in many biological processes in living cells including regulation of biochemical pathways, signaling cascades, and gene regulation.

Prediction of PPIs has been studied from many different perspectives in solving different problems. One of the important problems surrounding PPIs is the identification and prediction of different types of complexes, which are characterized by properties such as type and numbers of proteins that interact, stability of the proteins, and also duration of the interactions. This thesis focuses on studying the temporal and stability aspects of the PPIs mostly using structural data. We have addressed the problem of predicting obligate and non-obligate protein complexes, as well as those aspects related to transient versus permanent because of the importance of non-obligate and transient complexes as therapeutic targets for drug discovery and development. Generally, non-obligate interactions are more difficult to study and understand due to their instability and short life, while obligate interactions are more stable.

We have presented a computational model to predict-protein interaction types using our proposed physicochemical features of desolvation and electrostatic energies and also structural and sequence domain-based features. To achieve a comprehensive comparison and

demonstrate the strength of our proposed features to predict PPI types, we have also computed a wide range of previously used properties for prediction including physical features of interface area and interface area ratio, chemical features of hydrophobicity and amino acid composition, physicochemical features of solvent-accessible surface area (SASA) and atomic contact vectors (ACV). After extracting the main features of the complexes, a variety of machine learning approaches have been used to predict PPI types mostly based on combinations of classification, clustering and feature selection techniques. The prediction is performed via several state-of-the-art classification techniques, including linear dimensionality reduction (LDR), support vector machine (SVM), naive Bayes (NB) and *k*-nearest neighbor (*k*-NN). Moreover, several feature selection algorithms including gain ratio (GR), information gain (IG), chi-square (Chi2) and minimum redundancy maximum relevance (mRMR) are applied on the available datasets to obtain more discriminative and relevant properties to distinguish between these two types of complexes

Our computational results on different datasets confirm that using our proposed physicochemical features of desolvation and electrostatic energies lead to significant improvements on prediction performance. Moreover, using structural and sequence domains of CATH and Pfam and doing biological analysis help us to achieve a better insight on obligate and non-obligate complexes and their interactions.

Dedication

To my kind parents,

for their endless support and encouragement ...

To my wonderful husband,

for his love and patience ...

To my adorable daughters,

for their better tomorrow ...

Acknowledgements

I would like to express my deepest appreciation to my advisor, Dr. Luis Rueda, a kind, talented and passionate scientist. His steady and invaluable support and advice throughout my research led me to the right way.

I would also like to thank my doctoral committee members, Dr. Reda Alhajj, Dr. William Crosby, Dr. Alioune Ngom and Dr. Dan Wu for their time to review my dissertation, give me insightful comments, and attend my seminars, comprehensive examination, proposal, and defense.

My sincere thanks also goes to Dr. William Crosby and his research fellow, Dr. Mohammad Haj Dezfulian, for their invaluable advice and help during my research.

Special thanks to my friends in the Pattern Recognition and Bioinformatics Lab, especially my collaborators in protein-protein interaction research group, Muhammad Aziz, Michael Hall, Manish Pandit, and Gokul Vasudev. I enjoyed our brainstorming meetings and joint projects that were really useful to share our knowledge and experiences.

Contents

Author's Declaration of Co-Authorship and Previous Publications iii

Abstract vi

Dedication viii

Acknowledgements ix

List of Figures xvii

List of Tables xx

1 Introduction 1

1.1 Protein-protein Interactions 1

1.2 PPI Types 2

1.3 Prediction of PPI Types 4

1.3.1 Feature Extraction 4

1.3.2 Feature Selection 5

1.3.3 Classification 7

1.3.4 Evaluation and Analysis 8

1.4 Motivation and Objective 9

1.4.1 Physicochemical Properties 9
1.4.2 Domain-based Properties 10
1.5 Contributions 13
1.6 Thesis Organization 14

Bibliography **17**

Part 1 Physicochemical Features **22**

2 Prediction of Biological Protein-protein Interactions using Atom-type and Amino Acid Properties (Proteomics 2011) **23**

2.1 Introduction 23
2.2 Materials and Methods 26
2.2.1 Dataset 26
2.2.2 Prediction Properties 27
2.2.3 Prediction Methods 31
2.3 Results and Discussions 34
2.3.1 Experimental Settings 34
2.3.2 Analysis of Prediction 35
2.3.3 Visual Analysis of Desolvation Energy 36
2.3.4 Analysis of Interacting Sub-units 38
2.4 Conclusion 39

Bibliography **43**

3 Analysis of Relevant Physicochemical Properties in Obligate and Non-obligate Protein-protein Interactions (BIBM 2011) 46

3.1 Introduction 46

3.2 Materials and Methods 49

3.2.1 Datasets and Properties 49

3.2.2 The Prediction Methods 51

3.2.3 The Feature Selection Methods 52

3.3 Results and Discussions 54

3.3.1 Experimental Settings 54

3.3.2 Analysis of MRMR-based Feature Selection 55

3.3.3 Analysis of Biologically-guided Feature Selection Methods 57

3.3.4 Visual Analysis of Relevant Features 58

3.4 Conclusion 60

Bibliography 63

4 The role of electrostatic energy in prediction of obligate protein-protein interactions (BMC 2013) 66

4.1 Background 66

4.2 Methods 69

4.2.1 Datasets 69

4.2.2 Prediction properties 70

4.2.3 Prediction methods 73

4.2.4 Feature selection methods 76

4.3 Results and discussion 78

4.3.1 Analysis of prediction properties 79
4.3.2 Analysis of distance cutoffs 81
4.3.3 Analysis of feature selection 85
4.3.4 Visual analysis 87
4.4 Conclusions 91

Bibliography **93**

Part 2 Domain-based Features-CATH **98**

5 Using Desolvation Energies of Structural Domains to Predict Stability of Protein Complexes (NetMAHIB 2013) **99**
5.1 Introduction 99
5.2 Datasets and Prediction Properties 102
5.2.1 Desolvation Energy 102
5.2.2 Domain-based Properties 103
5.3 Prediction Methods 107
5.3.1 Linear Dimensionality Reduction 107
5.3.2 Support Vector Machines Based on SMO 109
5.3.3 *k*-Nearest Neighbor 109
5.3.4 Naive Bayes 110
5.4 Results and Discussion 110
5.4.1 Analysis of the Prediction Properties 110
5.4.2 Analysis of Structural Properties 113
5.5 Conclusion 116

5.6 Supplementary Materials 116

Bibliography **129**

6 Using Structural Domains to Predict Obligate and Non-obligate Protein-protein Interactions (CIBCB 2012) **133**

6.1 Introduction 133

6.2 Prediction Methods 136

6.2.1 Linear Dimensionality Reduction 136

6.2.2 Support Vector Machines 137

6.3 Datasets and Prediction Properties 139

6.3.1 Desolvation Energy 139

6.3.2 Domain-based Properties 140

6.4 Results and Discussions 143

6.4.1 Experimental Settings 143

6.4.2 Analysis of Prediction 144

6.4.3 Analysis of DDIs 146

6.5 Conclusion 148

Bibliography **151**

Part 3 Domain-based Features-Pfam **154**

7 Analysis of Obligate and Non-obligate Complexes using Desolvation Energies in Domain-domain Interactions (BIOKDD 2011) **155**

7.1 Introduction 155

7.2 Materials and Methods . . . 158
7.2.1 Dataset . . . 158
7.2.2 Features . . . 161
7.2.3 Prediction Methods . . . 162
7.3 Results and Discussions . . . 164
7.3.1 Experimental Settings . . . 164
7.3.2 Analysis of Prediction . . . 166
7.3.3 Analysis of DDIs . . . 167
7.4 Conclusion . . . 168

Bibliography **171**

8 Computational Analysis of the Stability of SCF Ligases Employing Domain Information (ACMBCB 2014) **173**

8.1 Introduction . . . 173
8.2 Materials and Methods . . . 175
8.3 SCF-Ligases Dataset . . . 175
8.3.1 Prediction Properties . . . 176
8.3.2 Prediction Method . . . 178
8.3.3 Feature Selection . . . 178
8.4 Results and Discussions . . . 179
8.4.1 Analysis of Interaction Types . . . 179
8.4.2 Analysis of the Prediction Properties . . . 181
8.4.3 Analysis of the Feature Selection . . . 182
8.5 Conclusion . . . 182

Bibliography **184**

9 Conclusions & Future Works **187**

9.1 Conclusion 187

9.2 Future Work 190

Vita Auctoris **192**

*

List of Figures

1.1 A general framework used to predict PPI types. 5

1.2 Quaternary structure of an obligate complex, PDB-ID *1h8e*, along with its interacting chains A and D and containing Pfam domains of each chain. Chains A and D are shown in light green and light blue respectively. Chain A has three domains of *PF02874* (orange), *PF00006* (red), and *PF00306* (green). Similarly, chain D has the same number and types of Pfam domains represented in purple, blue and yellow. The figure was generated using ICM browser [34]. 12

2.1 Heatmaps of desolvation energies of (a) interacting atom pairs and (b) interacting amino acid pairs. 41

2.2 3D structure of (a) obligate complexes and (b) non-obligate complexes visualized using ICM Browser [32]. 42

3.1 Prediction accuracy of the ZH-AA and MW-AA datasets using MRMR feature selection method. 56

3.2 Heatmaps of desolvation energies of interacting amino acid pairs in (a) the MW-AA dataset and (b) the ZH-AA dataset. Amino acids are grouped based on their polarity. 62

4.1 Quaternary structure of an obligate complex, PDB-ID *1b8j*, visualized with ICM Browser, along with its interacting chains A and B. Positive and negative charges are represented in red and blue respectively. Interface atoms of the interacting chains are represented in yellow and purple, respectively. . 73

4.2 ROC curves for the (a) MW-AT and (b) ZH-AT datasets using desolvation energy (blue line) and electrostatic energy (red line) as properties for prediction by using LDR. 82

4.3 Prediction accuracy for NB on MW-AT (red line) and ZH-AT (blue line) using desolvation energy as the prediction property and different distance cutoffs ranging from 5Å to 13Å. 83

4.4 Prediction accuracy for NB on MW-AT (red line) and ZH-AT (blue line) using electrostatic energy as the prediction property and different distance cutoffs ranging from 7Å to 13Å. 84

4.5 Prediction accuracy for LDR on MW-AT (red line) and ZH-AT (blue line) using electrostatic energy plotted against the number of features selected by GR. 86

4.6 Plot of solvent accessible surface by electrostatic potential of an obligate complex, PDB-ID *2min*, before and after the interaction takes place; (a) Electrostatic potential of chain A of *2min*, (b) Electrostatic potential of chain B of *2min*, (c) Electrostatic potential of chains A and B of *2min*. . . . 89

4.7 Plot of solvent accessible surface area by electrostatic potential of a non-obligate complex, PDB-ID *1a2k*, before and after the interaction takes place. (a) Electrostatic potential of chains AB of *1a2k*, (b) Electrostatic potential of chain C of *1a2k*, (c) Electrostatic potential of chains AB and C of *1a2k*. . 90

5.1 Four levels of the CATH hierarchy (Class, Architecture, Topology and Homologous superfamily). 104

5.2 ROC curves and AUC values for all subsets of features of (a) MW and (b) ZH datasets. 114

6.1 Schematic view of levels 2 and 3 of CATH DDIs present in the MW and ZH datasets. 149

7.1 Schematic view of the DDI pairs in obligate and non-obligate interactions. . 169

8.1 A schematic view of a SCF-ligase. 176

8.2 Number and type of interactions for two groups of non-obligate (left) and obligate (right) SCF-ligase complexes. 180

List of Tables

1.1 Properties employed in different studies for prediction of obligate and non-obligate interactions (or interfaces). 6

1.2 Pfam domains of chains A and D of complex *1h8e*. 11

2.1 Datasets used in this study 27

2.2 BPPI dataset containing 213 obligate and 303 non-obligate binary complexes. 28

2.3 Description of the subsets of features used in this study. 31

2.4 Prediction results for SVM and LDR on the BPPI dataset. 36

2.5 Prediction results for LDR classifiers on the BPPI dataset after using visual pair selection. 38

2.6 Analysis of interacting sub-units in obligate and non-obligate complexes. . 39

3.1 Description of datasets used in this study. 50

3.2 Prediction results for LDR classifier by using different MRMR-based feature selection methods. 57

3.3 Prediction results for LDR classifier by using biologically guided feature selection methods for the MW and ZH datasets. 58

3.4 Post-processing analysis of features selected by MRMR^{pro} for the MW and ZH datasets. 60

4.1 Description of datasets used in this study 74
4.2 Comparison of accuracies for electrostatic and desolvation energies as properties 80
4.3 Prediction accuracies using desolvation energy and different distance cutoffs 81
4.4 Prediction accuracies for electrostatic energy and different distance cutoffs . 84
4.5 Prediction accuracies for electrostatic energy and different feature selection methods 87

5.1 Prediction accuracies of SVM-SMO, NB, *k*-NN and LDR for all domain-based subsets of features of the ZH and MW datasets. 112
5.2 A summary of the number of CATH DDIs from level 2 present in the ZH and MW datasets, categorized by their class types. 115
5.3 List of feature vectors for the ZH-L2 dataset 117
5.4 List of feature vectors for the ZH-L3 dataset 117
5.5 List of feature vectors for the ZH-L2+L3 dataset 120
5.6 List of feature vectors for the MW-L2 dataset 121
5.7 List of feature vectors for the MW-L3 dataset 123
5.8 List of feature vectors for the MW-L2+L3 dataset 127

6.1 Datasets and their number of complexes used in this study. 141
6.2 Subsets of features used in this study. 142
6.3 Prediction results for LDR and SVM classifiers for the MW and ZH datasets. 145
6.4 A summary of the number of CATH DDIs of level 1 present in the ZH and MW datasets. 146

7.1 Binary-PPID dataset (146 obligate and 169 non-obligate binary complexes). 160

7.2 Description of the subsets of features used in this study. 162

7.3 Prediction results for SVM and LDR classifiers on binary-PPID dataset. . . 167

8.1 Dataset of SCF-ligase complexes. 177

8.2 A summary of the average number of interactions for obligate and non-obligate complexes of the SCF-ligase dataset categorized by their interaction types. 181

8.3 Prediction accuracies of SVM-SMO for all domain-based subsets of features of the SCF-ligase datase. 182

9.1 Experimental settings employed in our different studies for prediction of obligate and non-obligate complexes. 189

Chapter 1

Introduction

1.1 Protein-protein Interactions

Proteins are large molecules that constitute the bulk of the cellular machinery of any living organism or biological system. They play important roles in fundamental and essential biological processes such as DNA synthesis, transcription, translation, and splicing. Proteins perform their functions by interacting with molecules such as DNA, RNA, and other proteins. Regulation of biochemical pathways, signaling cascades and transduction, cellular motion, gene regulation, forming a protein complex, modifying or carrying another protein are some of the essential biological processes in living cells performed by protein-protein interactions (PPIs). As a consequence, to understand the complex cellular mechanisms involved in a biological system, it is necessary to study the nature of these interactions at the molecular level, in which prediction of PPIs plays a significant role.

Although prediction of PPIs has been studied from many different perspectives, the main aspects that are studied include [1]: sites of interfaces (where), arrangement of proteins in a complex (how), type of protein complex (what), molecular interaction events (if),

and temporal and spatial trends (dynamics). These problems have been investigated in various ways, involving both experimental (*in vivo* or *in vitro*) and computational (*in silico*) approaches. Experimental approaches such as yeast two-hybrid and affinity purification followed by mass spectrometry tend to be costly, labor intensive and suffer from noise. Nonetheless, these techniques have been successfully used to produce high-throughput protein interaction data for many organisms [2, 3]. Typically, structural information in the main databases such as the Protein Data Bank (PDB) [4] is derived through costly techniques such as X-ray crystallography or NMR (for smaller proteins). Therefore, using computational approaches for prediction of PPIs is a good choice for many reasons [5]. To date, a variety of machine learning approaches have been used to predict PPIs, mostly based on combinations of classification, clustering and feature selection techniques. These systems, in general, represent objects (complexes, sites, patches, protein chains, domains or motifs) as features or properties.

Among these research problems in the field of prediction of PPIs, this thesis focuses on computational prediction of PPI types.

1.2 PPI Types

There are different types of protein-protein interactions that provide different levels of information on different biological processes. Based on the type and numbers of proteins that interact, stability of the proteins, duration of the interaction, the following PPI types can be defined [6]:

- Based on the similarities of sub-units (chains):
  - Homo-oligomeric complexes (homomer): Interaction between identical chains.

  - Hetero-oligomeric complexes (heteromers): Interaction between non-identical chains.

- Based on the number of interacting sub-units:
  - dimers (two sub-units), trimers (three sub-units), tetramer (four sub-units), pentamer (five sub-units) and so on.

- Based on the affinity and stability of PPIs:
  - Non-obligate complexes: binding components (proteins) can form stable structures and cannot exist in *vivo* independently.
  - Obligate complexes: components do not form stable functional structures on their own and can be stable in vivo independently.

- Based on the duration and life time of the interactions [6]:
  - Transient complexes: the interactions associate/dissociate temporarily in *vivo*.
  - Permanent complexes: the interactions are stable and irreversible.

Stability of complexes can be quantified in terms of their disassociation rates. Disassociation rates of obligate complexes are in the range of nM (10^{-9}Mol) while for non-obligate complexes this rate is in the range of μM (10^{-6}Mol) [7]. In general, all obligate complexes are permanent. Similarly, except from some examples of permanent non-obligate interactions such as enzyme-inhibitor interactions, all non-obligate interactions can be considered as transient complexes [7].

In this thesis, we focus on the prediction of obligate (permanent) and non-obligate (transient) complexes. It is important to be able to distinguish between obligate and non-obligate

complexes, since non-obligate interactions are more difficult to study and understand due to their instability and short life, while obligate interactions are more stable [8].

1.3 Prediction of PPI Types

A general model to predict PPI types is shown in Figure 1.1. The dataset is a list of PPIs with their pre-defined types (classes). To predict PPI types, first of all, the prediction properties (features) of each complex in the dataset are extracted employing different PPI databases. Then, the extracted features are passed through a feature selection module used to remove noisy, irrelevant, and redundant features and select the most powerful and discriminative ones for prediction. After that, the selected features are used for classification and the outputs of the classification module are the predicted PPI types. Finally, the performance of the prediction model can be evaluated using different numerical performance metrics and visual analysis tools. More details about the four main parts, feature extraction, feature selection, classification, and evaluation and analysis of the presented prediction model are discussed below.

1.3.1 Feature Extraction

Features are the observed properties of each sample (complex) which are used for prediction. Using the most relevant features is very important for successful prediction. Some studies in PPIs consider the analysis of a wide range of properties for predicting types of complexes or types of protein-protein interfaces (binding sites), including physical [9, 10], chemical [9–12], physicochemical [13], geometric [9, 10, 12, 14], sequence-based [10, 15, 16], and domain-based features [9, 17, 18]. A summary of the feature types employed for

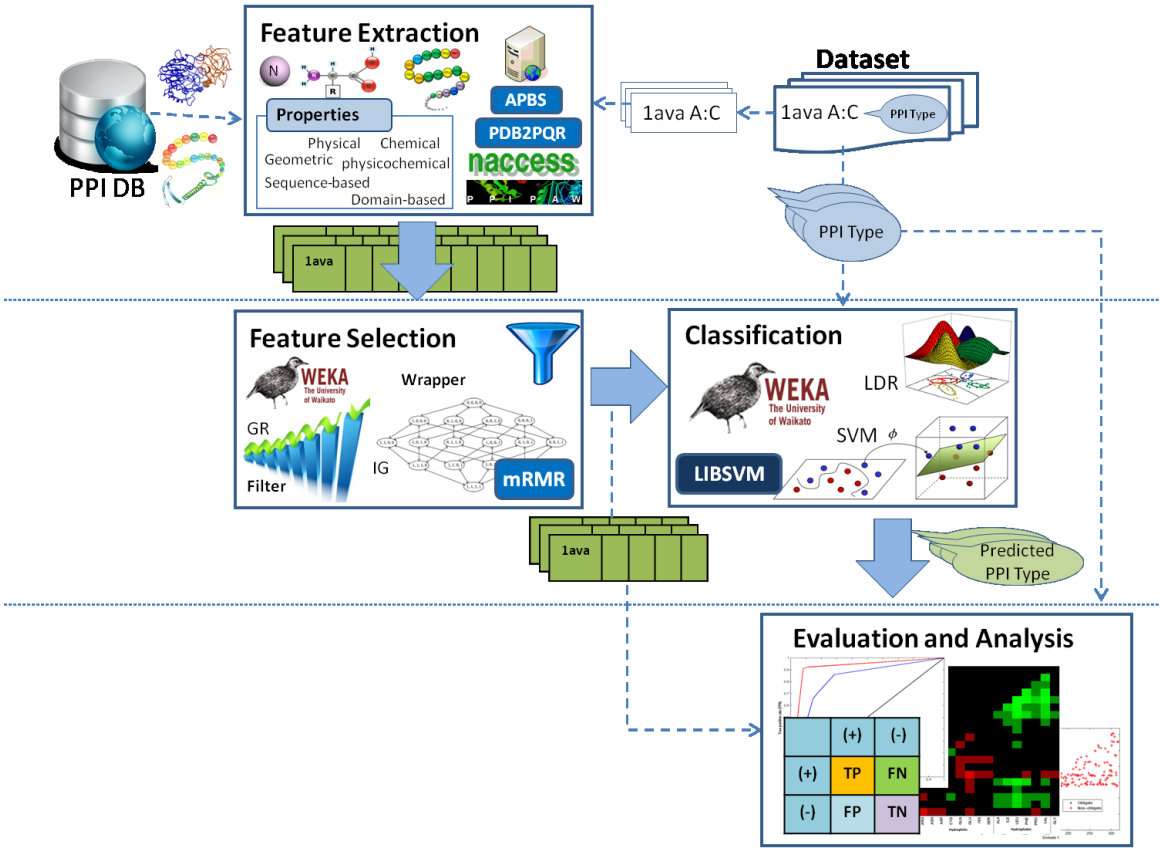


Figure 1.1: A general framework used to predict PPI types.

prediction in different studies along with the characteristics of obligate and non-obligate interactions (or interfaces) based on using those features is shown in Table 1.1.

1.3.2 Feature Selection

Feature selection is the process of choosing the best subset of relevant and discriminative features that represents the whole set of features efficiently after removing redundant and/or irrelevant ones. Applying feature selection before running a classifier is useful in reducing the dimensionality of the data and, thus, reducing the prediction time while improving the prediction performance.

Table 1.1: Properties employed in different studies for prediction of obligate and non-obligate interactions (or interfaces).

Type	Property	Non-obligate	Obligate	Ref.
Chemical	Hydrophobicity	Less hydrophobic residues	More hydrophobic residues	[9–12]
	Polarity	High	Low	
Physical	Interface area (IA) Interface area ratio	Small interfaces $< 1500\ \AA^2$	Large and twisted interfaces from 1500 to 10000 $\AA^2$	[9, 10]
Physicochemical	Atomic contacts	Smaller number of contacts	Larger number of contacts	[11]
	Association bonds	Salt bridges	hydrogen bonds and Covalent disulphide bridges	[9]
	Binding affinity	Weak PPIs ($K_d > 10^{-6}$M)	Strong PPIs ($K_d < 10^{-6}$M)	[7, 12]
Geometric	Low-ASA pairs	Small (45 ± 20.6)	Large (83 ± 53.2)	[13]
	Gap volume	Larger gap volume	Larger complementary interfaces	[10]
	Secondary structure	Turns, more Helix	B-sheet, less Helix	[9, 12]
Sequence-based	B-factor	More flexible (Large B-factor)	Rigid (Small B-factor)	[16]
Evolutionary	Conservation score Sequence profile	Less conserved	Evolve slowly and at similar rates	[10, 11, 16, 19]
Domain-based	Interacting folds	Domain-polypeptide interactions	Domain-domain interactions	[7]

There are two different ways of doing feature selection: using wrapper methods and filter methods [20].

In filter-based feature selection methods, the quality of the selected features are scored and ranked independently of the classification algorithm by using some statistical criteria based on their relevance. Although feature selection based on filter methods is fast, it does not consider dependency of features from each other; a feature that is not useful by itself can be very useful when combined with others. Some of the most well-known filter methods are the Minimum Redundancy Maximum Relevance (mRMR), Information Gain (IG), Gain Ratio (GR) and Chi Square (χ^2) [20].

However, the aim of wrapper methods is to find the best subset of features using a particular predictive model (classifier) to score feature subsets. Since doing an exhaustive search to find the best subset of features is computationally intensive, some heuristic search methods can be employed to find an optimal feature subset for a specific dataset such as forward selection and backward elimination [21].

1.3.3 Classification

After extracting and selecting the most discriminating features, a classifier can be applied in order to assign the class labels (PPI types). For this, the samples are first divided into train and test sets using different methods such as *m*-fold cross-validation or leave-one-out methods. Classification method design follows two phases of processing for training and testing. In the training phase, the training samples are used to build a model that is a description of each training class. Then, in the testing phase, that model is used to predict the classes of the test samples. There are a variety of classification methods, of which some of the commonly used methods in the thesis are Linear Dimensionality Reduction

(LDR), Support Vector Machines (SVMs), k-Nearest Neighbor (k-NN) and Naive Bayes. The reader is referred to [22] for more details.

1.3.4 Evaluation and Analysis

Finally, the performance of the prediction model can be evaluated using numerical performance metrics and visual analysis tools. One of the well-known numerical performance metrics is accuracy, which can be computed as follows:

$$Accuracy = \frac{TP + TN}{N + P} \tag{1.1}$$

where TP and TN are the total numbers of true positive (true obligate) and true negative (true non-obligate) predictions, respectively. P and N are the total number of complexes in the positive and negative classes, respectively. For unbalanced class problems, the performance can be analyzed in terms of specificity ($SP = TN/N$), sensitivity ($SN = TP/P$), or geometric mean ($G_m = \sqrt{SN \times SP}$). Moreover, the receiver operating characteristic (ROC) curve is a visual tool that can be plotted based on the true positive rate (TPR), aka "sensitivity", vs. the false positive rate (FPR), or "1 − specificity", at various threshold settings. To generate the ROC curves, the sensitivity and specificity of each subset of features are determined for different parameter values of the employed classifier. Then, by applying a simple algorithm, the FPR and TPR points are filtered as follows: (a) for the same FPR values, the largest TPR value (top point) is chosen, and (b) for the same TPR values, the smallest FPR value (left point) is chosen. A polynomial of degree 2 is then fitted to the selected points. ROC analysis is suitable for unbalanced class problems and yields a better insight than simple performance metrics.

1.4 Motivation and Objective

Prediction and analysis of protein-protein interactions and specifically types of PPIs is an important problem in life science research because of the fundamental roles of PPIs in many biological processes in living cells. In addition, because of the importance role of non-obligate interactions as drug targets, understanding the mechanism of binding two (or more) proteins and especially attempt to achieve accurate prediction of PPI types are worth further investigation.

However, most of the properties employed to predict obligate and non-obligate PPIs listed in Table 1.1 are not accurate enough. For example, in Figure 2 of [10], it has been shown that although most of the non-obligate complexes have a small interface area (less than 1500 $\AA^2$), there are still some non-obligate complexes with interface area greater than 3500 $\AA^2$. As a consequence, we have proposed new features for predicting of obligate and non-obligate PPIs as follows.

1.4.1 Physicochemical Properties

As mentioned ealier, obligate complexes are more stable than non-obligate ones; the dissociation rate of obligate complexes are in the range of nM (10^{-9}Mol) while for non-obligate complexes this rate is in the range of μM (10^{-6}Mol) [7]. On the other hand, the stability of each protein can be quantified in terms of the energy associated with the forces that form the different interactions. Thus, as in [23], the binding free energy ΔG_{bind} is defined as follows:

$$\Delta G_{bind} = \Delta E_{elec} + \Delta G_{des}, \tag{1.2}$$

where ΔE_{elec} is the total electrostatic energy and ΔG_{des} is the total desolvation energy. Desolvation energy is defined as the knowledge-based contact potential (accounting for hydrophobic interactions), self-energy change upon desolvation of charged and polar atom groups, and side-chain entropy loss. In addition, electrostatic interactions are important in understanding inter-molecular interactions, since they are long-range and because of their influence in charged molecules. This is the main motivation for using electrostatic energy for prediction of PPI types.

In this thesis, both sub-types of binding free energy, desolvation and electrostatic energies, are employed as the properties to predict obligate and non-obligate complexes. These two features are included in the group of physicochemical features, which consider both chemical and physical characteristics of the interacting residues as the prediction properties.

1.4.2 Domain-based Properties

Domains are the minimal and fundamental units of proteins. These functional units often have a biological role and serve some specific purpose, such as signal binding or manipulation of a substrate within cells [24, 25].

Recent studies focus on employing domain knowledge to predict protein-protein interactions [26–30]. It it has been claimed that only a few highly conserved residues are crucial for PPIs [9, 17], and also most domains and domain-domain interactions (DDIs) are evolutionarily conserved [31]. Thus, it can be concluded that physical interactions between proteins are mostly controlled by their domains. As a consequence, we have also proposed a domain-based model to predict obligate and non-obligate complexes to achieve a better insight of the PPIs.

Table 1.2: Pfam domains of chains A and D of complex *1h8e*.

Chain A			Chain D		
Pfam ID	Start	End	Pfam ID	Start	End
PF02874	24	92	PF02874	13	79
PF00006	148	372	PF00006	135	355
PF00306	384	488	PF00306	368	475

There are few domain family resources that can be applied for this purpose including databases such as Pfam [32] and CATH (Class, Architecture, Topology and Homologous superfamily) [33].

The Pfam database contains domains that are derived from sequence homology with other known structures, whereas CATH domains are based on structural homology. The sequence domains in Pfam are identified using multiple sequence alignments and hidden Markov models (HMMs). Other classifications of Pfam entries are (a) families of related protein regions, (b) short unstable units that can make a stable structure when repeated multiple times, and (c) short motifs present in disordered parts of a protein.

In Figure 1.2, the quaternary structure of an obligate complex, PDB-ID *1h8e*, along with its interacting chains A and D and containing Pfam domains of each chain is shown in different colors. Also, the Pfam ID, start and end residue numbers of all domains of each chain are listed in Table 1.2. From the table, it is clear that each chain has three similar Pfam domains of *PF02874*, *PF00006* and *PF00306*.

In contrast, the structural domains in the CATH database are organized in a hierarchical fashion, which can be visualized as a tree with levels numbered from 1 to 8. Domains at upper levels of the tree represent more general classes of structures than those at lower levels. For example, “roll”, “beta barrels”, and “2-layer sandwich” are three different sample architectures of domains (level 2) in class of “mixed alpha-beta” domains (level 1) in the

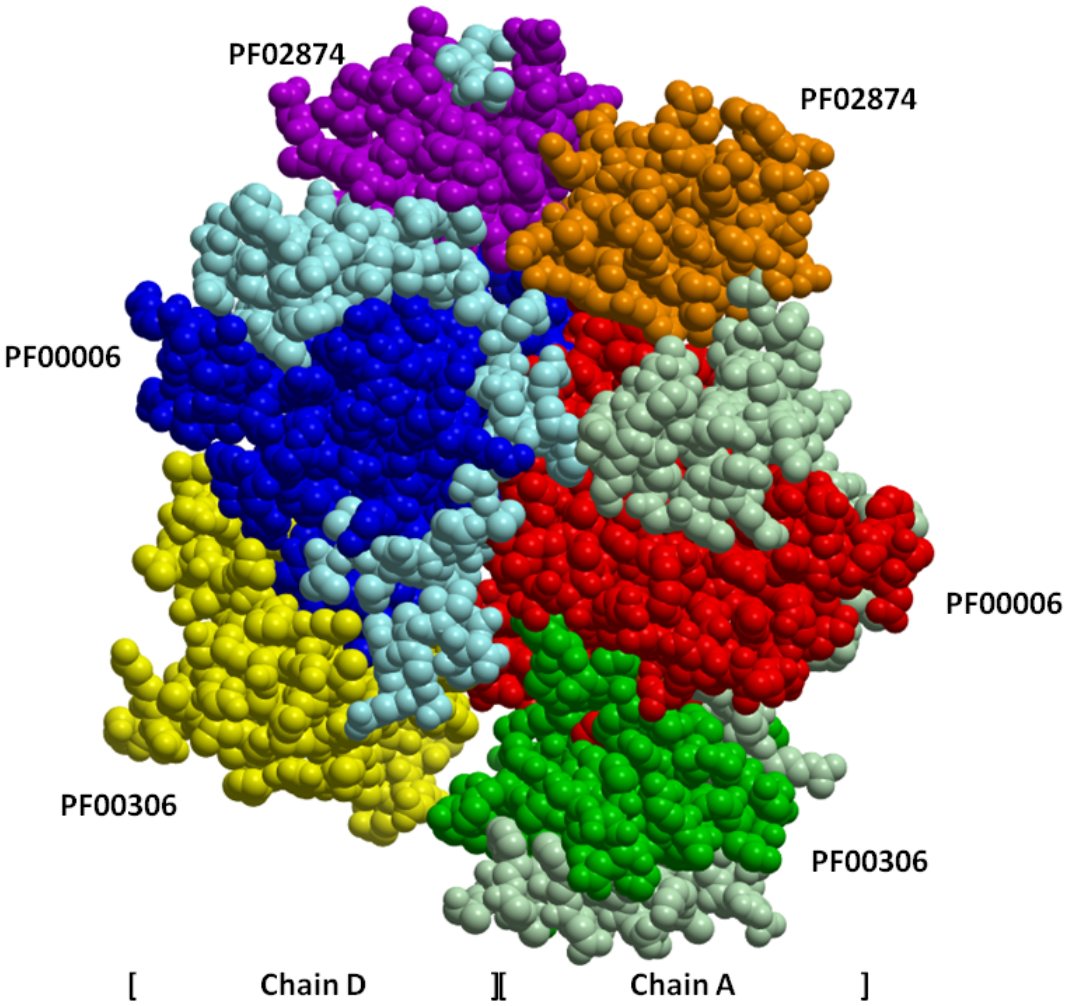


Figure 1.2: Quaternary structure of an obligate complex, PDB-ID *1h8e*, along with its interacting chains A and D and containing Pfam domains of each chain. Chains A and D are shown in light green and light blue respectively. Chain A has three domains of *PF02874* (orange), *PF00006* (red), and *PF00306* (green). Similarly, chain D has the same number and types of Pfam domains represented in purple, blue and yellow. The figure was generated using ICM browser [34].

CATH hierarchy.

In this thesis, both sequence Pfam and structural CATH domains are considered as the basis for our predictions.

1.5 Contributions

The main focus of this thesis is to predict obligate and non-obligate PPIs using different types of physicochemical and domain-based properties. The main contributions are as follows:

- Providing a generic computational framework for prediction of PPI types.
- Proposing different prediction properties for classifying obligate and non-obligate protein complexes including:
  - Physicochemical features of desolvation and electrostatic energies.
  - Domain-based features using structural CATH and sequence Pfam domains.
- Considering different interacting partners for each PPI to extend alternative representation of the same data for classification such as atom, amino acid and domain pairs present in the interface of interacting complexes.
- Proposing a feature selection method based on mRMR, which is used for selecting the most discriminative and relevant properties to distinguish between these two types of complexes.
- Performing a comprehensive comparison by computing some of the existing and currently used features for prediction of PPI types to compare with our proposed features in order to demonstrate the strength of the proposed features.

- Considering different datasets to verify the efficiency of the proposed features.
- Developing an automatic tool for extracting features for the classification. Downloading the tertiary and quaternary structures of the complexes from different databases, extracting and modifying the required information to calculate the features and calculating a wide range of features by considering different interacting partners are some of the capabilities of this tool.
- Performing a broad biological and visual analysis to yield a better insight of the PPI types and in order to analyze PPIs from a different perspective.

1.6 Thesis Organization

The thesis is organized in three parts with 9 chapters, including 7 selected papers out of 12 papers of the author [35–46] that have been previously published/submitted for publication in peer reviewed conferences and journals.

Part I, with two journal and one conference papers, covers the topics related to the proposed physicochemical features of desolvation energy (Chapters 2 and 3) and electrostatic energy (Chapter 4) as follows:

Chapter 2: Md. Aziz, M. Maleki, L. Rueda, M.Raza, S. Banerjee, "Prediction of Biological Protein-protein Interactions using Atom-type and Amino Acid Properties," Wiley-VCH Proteomics, vol. 11, no. 19, pp. 3802-10, Aug. 2011.

Chapter 3: M. Maleki, Md. Aziz, L. Rueda, "Analysis of Relevant Physicochemical Properties in Obligate and Non-obligate Protein-protein Interactions," in Workshop on

Computational Structural Bioinformatics in conjunction with BIBM 2011, GA, USA, Nov. 2011.

Chapter 4: M. Maleki, G. Vasudev, L. Rueda, "The Role of Electrostatic Energies in Prediction of Obligate Protein-Protein Interactions," Journal of BMC Proteome Science, 2013.

Parts II and III of the thesis are related to the proposed domain-based features to predict obligate and non-obligate complexes. The following two papers that consider structural CATH domains as the basis for our predictions are included in Part II:

Chapter 5: M. Maleki, M. Hall, L. Rueda, "Using Desolvation Energies of Structural Domains to Predict Stability of Protein Complexes," Journal of Network Modeling Analysis in Health Informatics and Bioinformatics (NetMAHIB), vol. 2, no. 4, pp. 267275, Dec. 2013.

Chapter 6: M. Maleki, M. Hall, L. Rueda, "Using Structural Domain to Predict Obligate and Non-obligate Protein-protein Interactions," in 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB2012), California, USA, May 2012.

Similarly, analysis of the role of sequence Pfam domain interactions in determining obligate and non-obligate PPIs is presented in Part III.

Chapter 7: M. Maleki, Md. Aziz, L. Rueda, "Analysis of Obligate and Non-obligate Complexes using Desolvation Energies in Domain-domain Interactions," in 10th International Workshop on Data Mining in Bioinformatics (BIOKDD2011) in conjunction with ACM SIGKDD 2011, San Diego, USA, Aug. 2011.

Chapter 8: M. Maleki, M. Dezfulian, W. Crosby, L. Rueda, "Computational Analysis of the Stability of SCF Ligases Employing Domain Information," in 5th ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACMBCB), CA, 2014. (submitted)

Finally, Chapter 9 concludes the thesis and identifies problems arising from this work and relevant future work.

Bibliography

[1] I. Kurareva and R. Abagyan, "Predicting molecular interactions in structural proteomics," in *Computational Protein-Protein Interactions*, R. Nussinov and G. Shreiber, Eds. CRC Press, 2009, ch. 10, pp. 185–209.

[2] Z. Liu and L. Chen, "Proteome-wide prediction of protein-protein interactions from high-throughput data," *Protein Cell*, vol. 3, no. 7, pp. 508–520, 2012.

[3] Q. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, and et al., "Structure-based prediction of protein-protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[4] H. M. Berman, G. J. Kleywegt, H. Nakamura, and J. L. Markley, "The Protein Data Bank at 40: reflecting on the past to prepare for the future," *Structure*, vol. 20, no. 3, pp. 391–396, 2012.

[5] L. Skrabanek, H. Saini, G. Bader, and A. Enright, "Computational prediction of protein-protein interactions," *Molecular Biotechnology*, vol. 38, no. 1, pp. 1–17, 2008.

[6] I. Nooren and J. Thornton, "Diversity of protein-protein interactions," *EMBO Journal*, vol. 22, no. 14, pp. 3846–3892, 2003.

[7] SEA. Ozbabacan, HB. Engin, A. Gursoy, and O. Keskin, "Transient protein-protein interactions," *Protein EngDes Sel.*, vol. 24, no. 9, pp. 635–48, 2011.

[8] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc. Natl Acad. Sci, USA*, vol. 93, no. 1, pp. 13–20, 1996.

[9] S. De, O. Krishnadev, N. Srinivasan, and N. Rekha, "Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different." *BMC Structural Biology*, vol. 5, no. 15, doi:10.1186/1472–6807–5–15, 2005.

[10] H. Zhu, F. Domingues, I. Sommer, and T. Lengauer, "NOXclass: Prediction of protein-protein interaction types," *BMC Bioinformatics*, vol. 7, no. 27, doi:10.1186/1471-2105-7-27, 2006.

[11] J. Mintseris and Z. Weng, "Atomic contact vectors in protein-protein recognition," *PROTEINS: Structure, Function and Genetics*, vol. 53, pp. 629–639, 2003.

[12] W. H. Y. C. D. La, M. Kong and D. Kihara, "Predicting permanent and transient protein-protein interfaces," *Proteins*, vol. 81, no. 5, pp. 805–18, 2013.

[13] J. L. Q. Liu Q, "Propensity vectors of low-ASA residue pairs in the distinction of protein interactions," *Proteins*, vol. 78, no. 3, pp. 589–602, 2010.

[14] M. C. Lawrence and P. M. Colman, "Shape complementarity at protein/protein interfaces," *J. Mol Biol*, vol. 234, no. 4, pp. 946–950, 1993.

[15] J. Mintseris and Z. Weng, "Structure, function, and evolution of transient and obligate protein-protein interactions," *Proc Natl Acad Sci, USA*, vol. 102, no. 31, pp. 10 930–10 935, 2005.

[16] R. Liu, W. Jiang, and Y. Zhou, "Identifying protein-protein interaction sites in transient complexes with temperature factor, sequence profile and accessible surface area," *Amino Acids*, vol. 38, pp. 263–270, 2010.

[17] S. G. J. V. Eichborn and R. Preissner, "Structural features and evolution of protein-protein interactions," *Genome Inform*, vol. 22, pp. 1–10, 2010.

[18] S. H. Park, J. Reyes, D. Gilbert, J. W. Kim, and S. Kim, "Prediction of protein-protein interaction types using association rule based classification," *BMC Bioinformatics*, vol. 10, no. 36, doi:10.1186/1471-2105-10-36, 2009.

[19] L. L. Q. Dong, X. Wang and Y. Guan, "Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins," *BMC Bioinformatics*, vol. 8, no. 147, 2007.

[20] J. Novakovic, P. Strbac, and D. Bulatovic, "Toward optimal feature selection using ranking methods and classification algorithms." *Yugoslav J. of Operations Research*, vol. 21, no. 1, pp. 119–135, 2011.

[21] R. Kohavi and G. H. John, "Wrappers for feature subset selection." *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[22] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Elsevier Academic Press, 2008.

[23] C. Camacho and C. Zhang, "FastContact: rapid estimate of contact and binding free energies," *Bioinformatics*, vol. 21, no. 10, pp. 2534–2536, 2005.

[24] L. Chen, R. Wang, and X. Zhang, *Biomolecular Networks: Methods and Applications in Systems Biology*. John Wiley and Sons, 2009.

[25] V. D. Dwivedi, S. Arora, and A. Pandey, "Computational analysis of physico-chemical properties and homology modeling of carbonic anhydrase from cordyceps militaris," *Network Modeling Analysis in Health Informatics and Bioinformatics*, pp. 1–4, 2013.

[26] N. Zaki, "Protein-protein interaction prediction using homology and inter-domain linker region information." *Advances in Electrical Engineering and Computational Science, Springer*, vol. 39, pp. 635–645, 2009.

[27] N. Zaki and P. C. S. Lazarova-Molnar, W. El-Hajj, "Protein-protein interaction based on pairwise similarity." *BMC Bioinformatics*, vol. 10, no. 150, doi:10.1186/1471–2105–10–150, 2009.

[28] M. Singhal and H. Resat, "A domain-based approach to predict protein-protein interactions." *BMC Bioinformatics*, vol. 8, no. 199, doi:10.1186/1471–2105–8–199, 2007.

[29] T. Akutsu and M. Hayashida, "Domain-based prediction and analysis of protein-protein interactions." *Biological data mining in protein interaction networks, Medical Information Science Reference, chapter 3*, pp. 29–44, 2009.

[30] P. Chandrasekaran, C. Doss, J. Nisha, R. Sethumadhavan, V. Shanthi, K. Ramanathan, and R. Rajasekaran, "In silico analysis of detrimental mutations in add domain of chromatin remodeling protein atrx that cause atr-x syndrome: X-linked disorder," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 2, no. 3, pp. 123–135, 2013.

[31] D. B. Singh, M. K. Gupta, R. K. Kesharwani, and K. Misra, "Comparative docking and admet study of some curcumin derivatives and herbal congeners targeting -amyloid," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 2, no. 1, pp. 13–27, 2013.

[32] M. Punta, P. Coggill, R. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. Sonnhammer, S. Eddy, A. Bateman, and R. Finn, "The Pfam protein families database." *Nucleic Acids Res.*, vol. 40, no. D1, pp. D290–D301, 2012.

[33] A. Cuff, I. Sillitoe, T. Lewis, O. Redfern, R. Garratt, J. Thornton, and C. Orengo, "The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies." *Nucleic Acids Res.*, vol. 37, pp. 310–314, 2009.

[34] "ICM browser user guide," available at : http://www.molsoft.com/icm_browser.html.

[35] Md. Aziz, M. Maleki, L. Rueda, M. Raza, and S. Banerjee, "Prediction of biological protein-protein interactions using atom-type and amino acid properties," *Wiley-VCH Proteomics 2011*, vol. 11, no. 19, pp. 3802–10, 2011.

[36] M. Maleki, Md. Aziz, and L. Rueda, "Analysis of relevant physicochemical properties in obligate and non-obligate protein-protein interactions," in *IEEE International Conference in Bioinformatics and Biomedicine Workshops (BIBMW), 2011*, pp. 345–351, 2011.

[37] M. Maleki and L. Rueda, "Domain-domain Interactions in Obligate and Non obligate Protein-protein Interactions," in *IEEE International Conference on Bioinformatics & Biomedicine Workshops (BIBMW 2011)*, GA, USA, pp. 907–908, 2011.

[38] M. Maleki, Md. Aziz, and L. Rueda, "Analysis of Obligate and Non-obligate Complexes using Desolvation Energies in Domain-domain Interactions," in *10th International Workshop on Data Mining in Bioinformatics (BIOKDD 2011) in conjunction with ACM SIGKDD 2011*, San Diego, USA, pp. 21–26, Aug. 2011.

[39] M. Maleki, M. Hall, and L. Rueda, "Using structural domain to predict obligate and non-obligate protein-protein interactions," in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB2012)*, San Diego, USA, pp. 9–15, May 2012.

[40] S. Banerjee, L. Rueda, and M. Maleki, "Prediction of Crystal Packing and Biological Protein-protein Interactions," in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB2012)*, San Diego, USA, pp. 16–20, May 2012.

[41] M. Hall, M. Maleki, and L. Rueda, "Multi-level structural domain-domain interactions for prediction of obligate and non-obligate protein-protein interactions," in *ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM-BCB)*, Florida, USA, pp. 518–520, Oct. 2012.

[42] M. Maleki, M. Hall, and L. Rueda, "Using desolvation energies of structural domains to predict stability of protein complexes," *Journal of Network Modeling Analysis in Health Informatics and Bioinformatics (NetMahib)*, vol. 2, no. 4, pp. 267–275, Nov. 2013.

[43] M. Maleki, G. Vasudev, and L. Rueda, "The role of electrostatic energy in prediction of obligate protein-protein interactions," *BMC Proteome Science*, vol. 11, 2013.

[44] M. Maleki, M. Dezfulian, W. Crosby, and L. Rueda, "Computational Analysis of Domain Interactions between Components of the SCF Ligase," in *18th Annual International Conference on Research in Computational Molecular Biology*, Pittsburgh, Pennsylvania, April 2014.

[45] M. Maleki, M. Dezfulian, W. Crosby, and L. Rueda, "Computational Analysis of the Stability of SCF Ligases Employing Domain Information," in *5th ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACMBCB)*, CA, Sept. 2014. (submitted)

[46] M. Maleki, M. Hall, and L. Rueda, *Structural Domains in Prediction of Biological Protein-protein Interactions*. Pattern Recognition in Computational Molecular Biology: Techniques and Approaches. Wiley, 2014. (submitted)

PART 1

PHYSICOCHEMICAL FEATURES

Chapter 2: Md. Aziz, M. Maleki, L. Rueda, M.Raza, S. Banerjee, "Prediction of Biological Protein-protein Interactions using Atom-type and Amino Acid Properties," Wiley-VCH Proteomics, vol. 11, no. 19, pp. 3802-10, Aug. 2011.

Chapter 3: M. Maleki, Md. Aziz, L. Rueda, "Analysis of Relevant Physicochemical Properties in Obligate and Non-obligate Protein-protein Interactions," in Workshop on Computational Structural Bioinformatics in conjunction with BIBM 2011, GA, USA, Nov. 2011.

Chapter 4: M. Maleki, G. Vasudev, L. Rueda, "The Role of Electrostatic Energies in Prediction of Obligate Protein-Protein Interactions," Journal of BMC Proteome Science, Nov. 2013.

Chapter 2

Prediction of Biological Protein-protein Interactions using Atom-type and Amino Acid Properties

2.1 Introduction

Protein-protein interactions (PPIs), binding of two or more proteins, are of prime importance in essential biological processes in living cells [1]. As a consequence of this, more attention has been drawn to this field of study, in particular, for identification and analysis of interacting proteins. Traditionally, the detection of protein-protein interactions was limited to labor-intensive experimental techniques such as co-immunoprecipitation or affinity chromatography. However, because of the possibility of introducing systematic errors for the large-scale prediction of PPIs, these methods have been recently replaced with various computational approaches. These new methods have been developed based on many different properties such as protein sequence, structure and evolutionary relationships in complete genomes.

Some studies in PPI consider geometric properties, e.g., shape complementarity of the protein structures [2], recognition of sites [3] or analysis of the conservation of residues [4]

present in the interaction surface of protein-protein complexes [5]. In another study, the role of hydrogen bonds and saline bridges appearance on the surface of proteins has been considered [5], while the study of the loss of surface accessible to solvent was presented in [6].

In [7], the amino acid composition of protein-protein interfaces in sequence level were studied and six different types of interfaces of intra and inter domains, homo and hetero-oligomers, and permanent and transient complexes were found. According to that study the there is only 1.5% of similarity between the internal and external surfaces, and 0.2% similarity between hetero surfaces of obligate homo complexes and transient homo complexes.

In general, transient interactions are more difficult to study and understand because of their short life, while obligate interactions are more stable [8]. This is one of the main reasons for which it is important to distinguish between obligate and transient complexes. In [9], the behavior of transient and obligate interactions was studied and a prediction method of these two types of interactions was proposed. In [10], it was shown that transient complexes are rich in aromatic residues and arginine, while depleted in other charged residues. Traditionally, the interfaces of some transient complexes were found to be hydrophobic [11]. Additionally, in [12], it was proposed that interfaces of obligate complexes are within clusters of hydrophobic residues. However, hydrophobicity in the interfaces of transient complexes from the remainder of the surface is not as distinguishable as in the obligate complexes [10]. The study of [13] suggested that mobility differences of amino acids are more significant for obligate and large interface complexes than for transient and medium-sized ones. As a consequence of this, making an accurate prediction of transient and obligate complexes using a single parameter of residue interface propensity is difficult. To study PPIs, in [14], each interaction was analyzed in physical interaction, co-complex rela-

tionship and co-member of the pathway. Also, in [15], three different types of interactions, namely crystal packing, obligate and non-obligate interactions, were studied by considering solvent accessible surface area (SASA), conservation scores, and the shapes of the interfaces. In [16], after classifying permanent and transient protein interactions based on 300 different interface attributes, the difference in molecular weight between interacting chains was reported as the best single feature to distinguish transient from permanent interactions. Based on their results, interactions with the same molecular weight or large interfaces are permanent. Although there are many studies that consider a wide range of interface parameters, including desolvation energies, amino acid composition, conservation, electrostatic energies, and hydrophobicity to predict protein-protein interactions, a prediction accuracy of 70% has been independently achieved by different research groups [17–20].

In a recent works [21], an approach to classify obligate and non-obligate complexes has been proposed in which only 20 minimum and maximum values of desolvation and electrostatic energies for two interacting chains, calculated by the FastContact tool [22], were considered as the input features of the classifiers. The results of that study show that desolvation energies are better discriminant than solvent accessibility and conservation properties. However, one of the drawbacks of that work is that the properties for prediction are considered at the residue level, and not at the amino acid or atom level, and these values are limited to only the 20 larger (or smaller) values for the energies, ignoring intermediate values that could be important in the classification scheme. In this paper, which is an extended version of [23], we present an analysis of PPIs that uses desolvation energies of all atom type and amino acid pairs present in the interface of interacting complexes to predict obligate and non-obligate protein-protein interactions by using linear dimensionality reduction (LDR) and support vector machines (SVM) methods. Ten-fold cross validation of the

proposed scheme on our newly-compiled BPPI dataset demonstrates that using desolvation energies of atom type features (76.94% prediction accuracy) are slightly better than amino acid properties (76.16% prediction accuracy) and much better than the features used in [15] (75% prediction accuracy) for predicting obligate and non-obligate complexes.

Furthermore, we have also presented a numerical and visual analysis on the desolvation energies of atom type and amino acid pairs present in these two types of interactions. A heatmap is used as a visual tool to achieve a closer view and find appropriate properties for prediction. Although a little decrease in prediction accuracy (1%–5%) is noticed, this decrease is acceptable because of the less time and space complexity required for prediction.

2.2 Materials and Methods

2.2.1 Dataset

We have compiled a new dataset by merging two existing, pre-classified datasets of obligate and non-obligate protein complexes obtained from the studies of Zhu et al., [15], and Mintseris and Weng [24]. The former contains 75 obligate (permanent) and 62 transient interactions, while the latter contains 115 obligate and 212 non-obligate interactions. There are 39 common interactions in these two datasets and hence the redundant complexes were removed. In addition, we carefully examined all the interactions and removed complexes with contradicting class labels. For example "*1eg9,A:B*" is classified as both obligate and non-obligate in [15] and [24]. In total, seven complexes (*1eg9, 1hsa, 1i1a, 1raf, 1d09, 1jkj* and *1cqi*) showed this contradiction and were then removed from the new dataset. After this pre-processing stage, the new dataset resulted in 417 complexes from which 182 were obligate and 235 were non-obligate. In this study, each complex is considered as the interaction

Table 2.1: Datasets used in this study

Name	# Complexes	# Obligate	# Non-obligate
Dataset of [15]	137	75	62
Dataset of [24]	327	115	212
BPPI	516	213	303

of two chains (two single sub-units). Since the dataset of [24] considers the interaction of two sub-units in which each may contain more than one chain, e.g., "*1qfu,AB:HL*", all these complexes were converted to interactions between two single chains (binary interactions). For this, all binary interactions of each of the 93 multiple-chain complexes were identified, obtaining 289 interactions, and each of these was converted into a separate complex in the new dataset. For example, the multiple-chain of *1qfu* was transformed to four binary chains as follows: *A:H*, *A:L*, *B:H* and *B:L*. The final step involves filtering binary complexes with non-interacting pairs. Using the interface definition of [25], complexes with interacting chains with less than five interface residues were removed. Two residues (from different chains) are considered to be interacting, if at least one pair of atoms from these residues is 5Å or less apart from each other. This resulted in our final dataset that contains 516 complexes, from which 303 are non-obligate and 213 are obligate complexes. We call this dataset binary protein-protein interactions (BPPI), as detailed in Table 2.1. The PDB IDs of these complexes and the interacting chains are shown in Table 2.2.

2.2.2 Prediction Properties

In our approach, we have introduced the use of desolvation energies as physicochemical properties to predict obligate and non-obligate complexes. For comparison purposes, we have also used various interface and non-interface properties such as solvent accessibility.

Table 2.2: BPPI dataset containing 213 obligate and 303 non-obligate binary complexes.

Obligate Complexes								
1a0f , A:B	1be3 , E:A	1dor , A:B	1go3 , E:F	1jb0 , B:D	1k8k , B:F	1lti , C:E	1qfe , A:B	1ytf , B:D
1a4i , A:B	1be3 , G:A	1dtw , A:B	1gpe , A:B	1jb0 , A:E	1k8k , C:G	1luc , A:B	1qfh , A:B	1ytf , C:D
1a6d , A:B	1bjn , A:B	1dxt , A:B	1gpw , A:B	1jb0 , A:E	1k8k , A:E	1m2v , A:B	1qla , A:B	1yve , I:J
1afw , A:B	1bo1 , A:B	1e50 , A:B	1gux , A:B	1jb0 , A:C	1k8k , C:F	1mjg , B:M	1qlb , B:C	2aai , A:B
1ahj , A:B	1brm , A:B	1e6v , A:B	1h2a , L:S	1jb0 , C:E	1k8k , D:F	1mjg , A:M	1qor , A:B	2ae2 , A:B
1aj8 , A:B	1byf , A:B	1e8o , A:B	1h2r , L:S	1jb0 , B:C	1kfu , L:S	1mro , A:B	1qu7 , A:B	2ahj , A:B
1ajs , A:B	1byk , A:B	1e9z , A:B	1h2v , C:Z	1jb0 , A:D	1kpe , A:B	1mro , B:C	1req , A:B	2hdh , A:B
1aom , A:B	1c3o , A:B	1eex , A:B	1h32 , A:B	1jb0 , A:D	1kqf , B:C	1mro , A:C	1sgf , A:B	2hhm , A:B
1aq6 , A:B	1c7n , A:B	1eex , A:G	1h4i , A:B	1jb0 , C:D	1kqf , A:B	1msp , A:B	1sgf , A:Y	2kau , A:C
1at3 , A:B	1ccw , A:B	1efv , A:B	1h8e , A:D	1jb7 , A:B	1ktd , A:B	1n98 , A:B	1smt , A:B	2kau , B:C
1aui , A:B	1cmb , A:B	1ep3 , A:B	1hcn , A:B	1jk0 , A:B	1l7v , A:C	1nbw , C:B	1sox , A:B	2min , A:B
1b34 , A:B	1cnz , A:B	1exb , A:E	1hfe , L:S	1jk8 , A:B	1l9j , C:L	1nbw , A:B	1spp , A:B	2mta , A:H
1b3a , A:B	1coz , A:B	1ezv , D:H	1hgx , A:B	1jkm , A:B	1l9j , C:M	1nse , A:B	1spu , A:B	2nac , A:B
1b4u , A:B	1cp2 , A:B	1ezv , C:F	1hjr , A:C	1jmx , A:G	1ld8 , A:B	1one , A:B	1tbg , A:E	2pfl , A:B
1b5e , A:B	1cpc , A:B	1f3u , A:B	1hr6 , A:B	1jmz , A:B	1ldj , A:B	1pnk , A:B	1tco , A:B	2utg , A:B
1b7b , A:C	1dce , A:B	1f6y , A:B	1hss , A:B	1jmz , G:B	1li1 , A:C	1poi , A:B	1trk , A:B	3gtu , A:B
1b7y , A:B	1dii , A:C	1fcd , A:C	1hxm , A:B	1jnr , A:B	1li1 , B:C	1pp2 , L:R	1vcb , A:B	3pce , A:M
1b8a , A:B	1dj7 , A:B	1ffu , A:C	1hzz , A:C	1jro , A:B	1lti , A:H	1prc , C:H	1vkx , A:B	3tmk , A:B
1b8j , A:B	1dkf , A:B	1ffv , A:B	1ihf , A:B	1jv2 , A:B	1lti , C:G	1prc , C:L	1vlt , A:B	4mdh , A:B
1b8m , A:B	1dm0 , A:D	1fm0 , D:E	1ir1 , A:S	1jwh , A:C	1lti , A:F	1prc , C:M	1vok , A:B	4rub , D:T
1b9m , A:B	1dm0 , A:B	1fs0 , E:G	1isa , A:B	1jwh , A:D	1lti , A:G	1qae , A:B	1wgj , A:B	4rub , A:T
1be3 , D:A	1dm0 , A:F	1fxw , A:F	1jb0 , B:E	1k28 , A:D	1lti , C:H	1qax , A:B	1xik , A:B	
1be3 , K:A	1dm0 , A:E	1g8k , A:B	1jb0 , B:E	1k3u , A:B	1lti , C:D	1qbi , A:B	1xso , A:B	
1be3 , C:A	1dm0 , A:C	1gka , A:B	1jb0 , B:D	1k8k , A:B	1lti , C:F	1qdl , A:B	1ypi , A:B	

Non-obligate Complexes								
1a14 , L:N	1bi7 , A:B	1dn1 , A:B	1f3v , A:B	1gaq , A:B	1ib1 , A:E	1k5d , A:C	1nf5 , A:B	1uea , A:B
1a14 , H:N	1bi8 , A:B	1doa , A:B	1f51 , A:E	1gc1 , C:G	1ibr , A:B	1k5d , A:B	1noc , A:B	1ugh , E:I
1a2k , B:C	1bj1 , H:V	1dow , A:B	1f51 , B:E	1gcq , B:C	1icf , B:I	1k90 , A:D	1nsn , H:S	1wej , F:H
1a4y , A:B	1bj1 , L:W	1dpj , A:B	1f80 , A:E	1gh6 , A:B	1icf , A:I	1kac , A:B	1nsn , L:S	1wej , F:L
1acb , E:I	1bj1 , H:W	1dtd , A:B	1f83 , A:C	1ghq , A:B	1iis , B:C	1kcg , A:C	1o6s , A:B	1wq1 , G:R
1agr , E:A	1bkd , R:S	1du3 , A:D	1f83 , A:B	1gl1 , A:I	1iis , A:C	1kcg , B:C	1o94 , A:C	1www , V:X
1ahw , A:C	1bml , A:C	1du3 , A:F	1f93 , A:E	1gla , F:G	1ijk , A:B	1kkl , A:H	1osp , L:O	1www , W:X
1ahw , B:C	1bqh , A:G	1dx5 , M:I	1f93 , B:F	1go4 , A:G	1ijk , A:C	1kkl , C:H	1osp , H:O	1xdt , R:T
1ak4 , A:D	1buh , A:B	1e6e , A:B	1f93 , B:E	1gp2 , A:B	1im3 , A:D	1kmi , Y:Z	1pdk , A:B	1ycs , A:B
1akj , B:D	1buv , M:T	1e6j , L:P	1f93 , A:F	1grn , A:B	1iod , B:G	1kxp , A:D	1qbk , B:C	1zbd , A:B
1akj , A:E	1bvn , P:T	1e6j , H:P	1fak , H:T	1gvn , A:B	1iod , A:G	1kxq , H:A	1qfu , A:L	2btc , E:I
1akj , A:D	1bzq , A:L	1e96 , A:B	1fak , L:T	1gxd , A:C	1is8 , C:M	1kxt , A:B	1qfu , A:H	2btf , A:P
1ao7 , A:E	1c0f , S:A	1eai , A:C	1fbi , L:X	1gzs , A:B	1is8 , B:L	1kyo , O:W	1qfw , A:M	2hmi , B:C
1ao7 , C:E	1c1y , A:B	1eay , A:C	1fbi , H:X	1h2k , A:S	1is8 , E:O	1l0o , A:C	1qfw , B:M	2hmi , B:D
1ao7 , C:D	1c4z , A:D	1ebd , A:C	1fc2 , C:D	1h59 , A:B	1is8 , D:N	1l0o , B:C	1qfw , B:I	2jel , L:P
1ao7 , A:D	1cc0 , A:E	1ebd , B:C	1fg9 , B:C	1he1 , A:C	1is8 , A:K	1l6x , A:B	1qgw , A:C	2jel , H:P
1ar1 , B:C	1cgi , E:I	1ebp , A:D	1fg9 , A:C	1hez , A:E	1is8 , D:O	1lb1 , A:B	1qkz , A:L	2mta , A:L
1ar1 , B:D	1clv , A:I	1ebp , A:C	1fin , A:B	1hlu , A:P	1is8 , A:L	1lfd , A:B	1qkz , A:H	2mta , A:C
1aro , L:P	1cmx , A:B	1eer , A:B	1fle , E:I	1hwg , A:C	1is8 , E:K	1lk3 , A:L	1qo0 , A:E	2mta , H:L
1atn , A:D	1cs4 , A:C	1efu , A:B	1flt , V:X	1hwg , A:B	1is8 , C:N	1lk3 , A:H	1qo0 , A:D	2pcb , A:B
1ava , A:C	1cs4 , B:C	1efx , C:D	1flt , W:X	1hx1 , A:B	1is8 , B:M	1lpb , A:B	1rlb , A:E	2pcc , A:B
1avg , H:I	1cse , I:E	1efx , A:D	1fns , A:L	1hzz , B:C	1itb , A:B	1m10 , A:B	1rlb , C:E	2prg , B:C
1avw , A:B	1cvs , A:C	1eja , A:B	1fns , A:H	1i2m , A:B	1jch , A:B	1m1e , A:B	1rlb , B:E	2ptc , E:I
1avx , A:B	1cxz , A:B	1emv , A:B	1fq1 , A:B	1i3o , A:E	1jiw , I:P	1m2o , A:B	1rrp , A:B	2sic , E:I
1avz , B:C	1d2z , A:B	1es7 , C:B	1fqj , A:C	1i3o , D:E	1jma , A:B	1m4u , A:L	1sbb , A:B	2tec , E:I
1awc , A:B	1d4x , A:G	1es7 , A:B	1fqv , A:B	1i3o , B:E	1jsu , B:C	1mah , A:F	1smf , E:I	3hhr , A:B
1ay7 , A:B	1d5x , A:C	1eth , A:B	1frv , A:B	1i4d , B:D	1jsu , A:C	1mbu , A:C	1smp , I:A	3sgb , E:I
1azz , A:D	1de4 , C:A	1euv , A:B	1fsk , A:B	1i4d , A:D	1jtd , A:B	1ml0 , A:D	1stf , E:I	3ygs , C:P
1azz , A:D	1dee , D:G	1evt , A:C	1fsk , A:C	1i7w , A:B	1jtg , A:B	1mr1 , A:D	1t7p , A:B	4htc , H:I
1b6c , A:B	1dev , A:B	1ezv , E:Y	1fss , A:B	1i85 , B:D	1jw9 , B:D	1n2c , A:F	1tab , E:I	4sgb , E:I
1b9y , A:C	1df9 , B:C	1ezv , E:X	1g0y , I:R	1i8l , A:C	1k3z , B:D	1n2c , B:E	1tgs , I:Z	7cei , A:B
1bdj , A:B	1dfj , E:I	1ezx , A:C	1g4y , B:R	1i9r , A:L	1k3z , A:D	1n2c , A:E	1tmq , A:B	
1bgx , L:T	1dhk , A:B	1f02 , I:T	1g73 , A:C	1i9r , A:H	1k4c , A:C	1n2c , B:F	1toc , B:R	
1bgx , H:T	1dkg , A:D	1f34 , A:B	1g73 , B:C	1ib1 , B:E	1k4c , B:C	1nbf , A:D	1tx4 , A:B	

Desolvation Energy

Different approaches have been developed to group different types of proteins, based on their different properties. Among them, desolvation energies are very efficient for prediction [23]. Knowledge-based contact potential that accounts for hydrophobic interactions, self-energy change upon desolvation of charged and polar atom groups, and side-chain entropy loss is defined as desolvation energy. In [22], the binding free energy, ΔG_{bind}, is defined by the following equation:

$$\Delta G_{bind} = \Delta E_{elec} + \Delta G_{des}, \tag{2.1}$$

where ΔE_{elec} is the total electrostatic energy and ΔG_{des} is the total desolvation energy, which for a protein is defined as follows [22]:

$$\Delta G_{des} = g(r)\Sigma\Sigma e_{ij}. \tag{2.2}$$

If we are considering the interaction between the i^{th} atom of a ligand and the j^{th} atom of a receptor, then e_{ij} is the atomic contact potential (ACP) [26] between them, and $g(r)$ is a smooth function based on their distance. The value of $g(r)$ is 1 for atoms that are less than 5 Å apart [22]. For simplicity, we consider the smooth function to be linear within the range of 5 and 7 Å, and the value of $g(r)$ is $(7-r)/2$.

We collected the structural data from the Protein Data Bank (PDB) [27] for each complex in the BPPI dataset. From each PDB file two chains (ligand and receptor) were extracted. We have considered 18 different atom types as in [26]. Thus, for each protein complex a feature vector with 18^2 values were obtained, where each feature contains the cumulative sum of desolvation energies of a pair of atom types, computed using Eq. (2.2).

As the order of interacting atom pairs is not important, the final length of the feature vector for each complex is 171 that corresponds to the number of unique pairs.

We have also considered pairs of amino acids, and for this, we computed 20^2 desolvation energy values for each pair of atoms using Eq. (2.2), and accumulated the values for each pair of amino acids. Avoiding repeated pairs resulted in 210 different features (unique pairs of amino acids).

After computing all properties, some feature vectors contain zeros in most of their values, which filtered by applying principal component analysis (PCA). By using desolvation energies for different types of features, two subsets of features for prediction and evaluation were generated, as listed in Table 2.3. The names of the subsets are BPPI-X where X is DEAT for atom type and DEAA for amino acid pairs.

Interface Properties

To analyze the power of desolvation energy in prediction of obligate and non-obligate complexes and for comparison, we have also considered other properties, mainly for those atoms and amino acids in the interface. A residue is defined as being part of the interface, if its solvent accessible surface area decreases by more than 1 $\AA^2$. upon the formation of the complex.

The following four interface properties of NOXclass [15] were extracted from the BPPI dataset, since these properties were identified in [15] as the best ones for prediction of different types of PPIs:

- Interface Area (IA)

$$IA = \frac{1}{2}(SASA_a + SASA_b - SASA_{ab}). \tag{2.3}$$

Table 2.3: Description of the subsets of features used in this study.

Dataset	Description
BPPI-DEAT	desolvation energies for atom type features
BPPI-DEAA	desolvation energies for amino acid type features
BPPI-NOXclass	NOXclass features [15]

- Interface Area Ratio (IAR)

$$IAR = \frac{IA}{min(SASA_a, SASA_b)}. \tag{2.4}$$

- Amino acid composition of the interface
- Correlation between amino acid compositions of interface and protein surface

SASA values for the residues were calculated using NACCESS [28] with a probe sphere of radius 1.4 $\AA^2$. These derived features were computed as the methods described in [15].

2.2.3 Prediction Methods

To predict complexes on the basis of desolvation energies (210 features for amino acid type and 171 features for atom type), we first applied PCA as a pre-processing step to eliminate ill-conditioned matrices involved in the LDR techniques. To obtain the principal components, we used different threshold values, and selected the threshold that leads to the highest accuracy. After obtaining those principal components, the complexes are classified via LDR methods. For classifying on the basis of different number of physicochemical interface properties, LDR methods were compared to SVMs.

Linear Dimensionality Reduction

LDR methods have been successfully used in pattern recognition due to their easy implementation and high classification speed [29]. In LDR the objects (protein complexes in our study) are represented by two normally distributed random vectors $\mathbf{x}_1 \sim N(\mathbf{m}_1, \mathbf{S}_1)$ and $\mathbf{x}_2 \sim N(\mathbf{m}_2, \mathbf{S}_2)$, respectively, with p_1 and p_2 being the *a priori* probabilities. The aim of LDR is to apply a linear transformation to project large dimentional data onto a lower dimensional space, $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$, in such a way that the classification is as efficient as possible, if not better, in the new space. To obtain the underlying transformation matrix $\mathbf{A}$, $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$ and $\mathbf{S}_E = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$, the within-class and between-class scatter matrices respectively, are first computed.

Three LDR criteria are considered in this study:

(a) Fisher's discriminant analysis (FDA) [30], which aims to maximize:

$$J_{FDA}(\mathbf{A}) = tr\left\{(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_E\mathbf{A}^t)\right\}. \tag{2.5}$$

The optimal $\mathbf{A}$ is found by considering the eigenvector corresponding to the largest eigenvalue of $\mathbf{S}_{FDA} = \mathbf{S}_W^{-1}\mathbf{S}_E$.

(b) The heteroscedastic discriminant analysis (HDA) approach [29], which aims to obtain the matrix $\mathbf{A}$ that maximizes the following function:

$$J_{HDA}(\mathbf{A}) = tr\left\{(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\left[\mathbf{A}\mathbf{S}_E\mathbf{A}^t \quad -\mathbf{A}\mathbf{S}_W^{\frac{1}{2}}\frac{p_1\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}})+p_2\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1p_2}\mathbf{S}_W^{\frac{1}{2}}\mathbf{A}^t\right]\right\} \tag{2.6}$$

This criterion is maximized by obtaining the eigenvectors, corresponding to the largest

eigenvalues, of the matrix:

$$\mathbf{S}_{HDA} = \mathbf{S}_W^{-1}\left[\mathbf{S}_E - \mathbf{S}_W^{\frac{1}{2}}\frac{p_1\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}})+p_2\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1p_2}\mathbf{S}_W^{\frac{1}{2}}\right]. \quad (2.7)$$

(c) The Chernoff discriminant analysis (CDA) approach [29], which aims to maximize:

$$J_{CDA}(\mathbf{A}) = tr\{p_1p_2\mathbf{A}\mathbf{S}_E\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1} + \log(\mathbf{A}\mathbf{S}_W\mathbf{A}^t) - p_1\log(\mathbf{A}\mathbf{S}_1\mathbf{A}^t) - p_2\log(\mathbf{A}\mathbf{S}_2\mathbf{A}^t)\}. \quad (2.8)$$

To solve this problem, a gradient-based algorithm is used [29]. This iterative algorithm needs a learning rate, α_k, which is maximized by using the secant method to ensure that the gradient algorithm converges. The initialization of the matrix $\mathbf{A}$ is also an important issue in the gradient-based algorithm. In this study, ten different initializations were preformed and the solution for $\mathbf{A}$ that yields the maximum Chernoff distance in transformed space were selected. More details about this algorithm, the CDA approach and LDR can be found in [21, 29].

Once the dimension reduction takes place, the vectors in the new space of dimension d can be classified using any classification technique. To achieve the reduction, the linear transformation matrix $\mathbf{A}$, which corresponds to the one obtained by one of the LDR criteria, is found independently for every fold in the cross-validation process. In this work, two classifiers are considered to classify the vectors in the lower dimensional space: Quadratic Bayesian (QB) classifier [29], which is the optimal classifier for normal distributions, and a linear Bayesian (LB) classifier obtained by deriving a Bayesian classifier with a common covariance matrix, $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$.

Support Vector Machines

SVMs are well known machine learning techniques used for classification, regression and other tasks. The aim of the SVM is to find the support vectors (most difficult vectors to be classified), and derive a linear classifier, which ideally separates the space into two regions. Classification is normally inefficient when using a linear classifier, because the data is not linearly separable, and hence the use of kernels is crucial in mapping the data onto a higher dimensional space in which the classification is much more efficient. There are a number of kernels that can be used in SVM models. In this study, polynomial, radial basis function (RBF) and sigmoid are used.

2.3 Results and Discussions

2.3.1 Experimental Settings

The three above-mentioned LDR schemes, including FDA, HDA and CDA combined with a QB or LB classifier were applied. In a 10-fold cross validation process, reductions to dimensions $d = 1,\ldots,20$ were performed, followed by QB and LB, and the maximum average classification accuracy was recorded for each classifier. The SVM was also trained in a 10-fold cross validation process with three kernels: RBF, polynomial and sigmoid. The training phase was carried out with the LIBSVM package [31]. A grid search was performed on the parameters gamma and C, choosing the ones that yield the maximum average accuracy for all kernels. For the polynomial kernel, the degree of the polynomial was set to 3.

PCA was used as a pre-processing step to eliminate ill-conditioned matrices present

in the LDR. To select the principal components, we used different threshold values (from $\lambda_{max}10^{-2}$ to $\lambda_{max}10^{-7}$), where λ_{max} is the largest eigenvalue of the scatter matrix. The results for the threshold that achieves the highest accuracy are reported. In this study, LDR was applied with default parameters. However, the parameters for CDA and HDA could be optimized to obtain even better results. This is an open problem that is worth investigating.

The subsets of features shown in Table 2.3 were used for prediction. After running the classifiers in a 10-fold cross validation procedure for all subsets of features, the average accuracies were computed as follows: $acc = (TP + TN)/N$, where TP is the number of true positive (obligate), TN is the number of true negative (non-obligate), and N is the total number of complexes in the test sets of all 10 folds. In the subsequent tables, the best accuracy for each method in each subset of features is bolded.

2.3.2 Analysis of Prediction

The results of SVM and LDR with different subsets of features are depicted in Table 2.4. For SVM, it is clear that the RBF kernel performs better that polynomial and sigmoid kernels for all subsets of features. The atom type features (BPPI-DEAT) are best classified with SVM and the RBF kernel, achieving an average accuracy of 76.94%, while accuracy for amino acid type features (BPPI-DEAA) in the best case is 76.16%. Furthermore, the subset based on NOXclass features (BPPI-NOXclass) with best accuracy of 75% classification accuracy yields less efficient predictions than the subsets based on desolvation energy properties (BPPI-DEAT and BPPI-DEAA) with the SVM classifier.

For LDR, the best accuracy, 74.38%, is achieved by CDA with the linear classifier, which is still lower than the best accuracy achieved by SVM. Note that both of them are on the BPPI-DEAT subset. Additionally, as in SVM, the subset of atom type features perform

Table 2.4: Prediction results for SVM and LDR on the BPPI dataset.

	SVM			LDR					
	RBF	Poly.	Sig.	Linear			Quadratic		
				FDA	HDA	CDA	FDA	HDA	CDA
BPPI-DEAT	**76.94**	74.81	75.19	71.85	74	**74.38**	72.43	74.15	73.79
BPPI-DEAA	**76.16**	73.45	74.61	68.88	73.22	**74.17**	69.47	73.57	73.55
BPPI-NOXclass	**75.00**	71.51	72.67	**73.06**	72.48	71.71	72.48	71.32	71.32

slightly better than amino acid ones with 74.17% accuracy and also much better than the NOXclass properties (73.06% accuracy).

Generally, it can be concluded that for the BPPI dataset:

(a) The SVM with RBF and optimized parameters outperforms the LDR methods in all subsets of features.

(b) Amino acid type feature yield lower accuracies than atom type features for both LDR and SVM.

(c) Desolvation energies are more powerful than the four properties of NOXclass (interface area, interface area ratio, amino acid composition of the interface and correlation between amino acid compositions of interface and protein surface) in predicting obligate and non-obligate complexes, achieving 76.94% prediction accuracy in comparison to 75% achieved when using SVM; also, 74.36% prediction accuracy in comparison to 73.06% by using LDR.

2.3.3 Visual Analysis of Desolvation Energy

To visually observe the effect of desolvation energies among obligate and non-obligate complexes within the BPPI dataset, we have used heatmaps of interacting atom and amino acid pairs. For this, all complexes in the BPPI dataset were used to generate the feature

vector by atom and amino acid type properties for obligate and non-obligate complexes, separately. Then, we computed the column-wise sum, obtaining four sums of feature vectors (obligate/non-obligate atom type pairs and obligate/non-obligate amino acid pairs). Since the atomic contact potential (ACP) matrix is symmetric, we represented each sum of feature vector as an 18x18 or 20x20 matrix for atom type or amino acid, respectively. Then, each matrix was condensed as an upper triangular matrix. Summing all indices for obligate and non-obligate, heatmaps were generated (red and blue respectively). Heatmaps of interacting chains of obligate and non-obligate complexes are shown in Figure 2.1. In the heatmaps, the lighter the color is, the larger the desolvation energy value of that interacting atom or amino acid pair is. Combined heatmaps of obligate and non-obligate complexes for atom and amino acid type features were obtained.

In the combined heatmaps, red/light or blue/light blocks indicate that the pair shows strong obligate or non-obligate behavior, and those are our pairs of interest. Thus, we selected 17 interacting pairs of atom type features (out of 171 features) and 27 interacting pairs of amino acid features (out of 210 features) for prediction of obligate and non-obligate complexes (white-border blocks in the combined heatmaps). The LDR methods were applied, again, for prediction of obligate and non-obligate complexes on the selected pairs. The performance of LDR is shown in Table 2.5. By comparing the new results with those of Table 2.4, reductions of 1% to 5% in prediction accuracies are noticed. Because of the smaller number of features, the classification performed faster, and hence this decrease can be acceptable. Also, this analysis shows that a few atom and amino acid pairs are good descriptors for prediction of the two types of complexes.

Table 2.5: Prediction results for LDR classifiers on the BPPI dataset after using visual pair selection.

dataset	# features	Linear			Quadratic		
		FDA	HDA	CDA	FDA	HDA	CDA
BPPI-DEAT	17	**70.87**	70.28	70.81	70.47	70.86	70.68
BPPI-DEAA	27	66.98	68.73	68.14	67.97	**68.93**	68.14

2.3.4 Analysis of Interacting Sub-units

As explained earlier, a particular atom pair has a pre-determined desolvation energy value in the ACP matrix. Thus, during the interaction, the actual energy depends on the number of interactions and the distances between their interacting pairs. As a result, different complexes with the same interacting pairs may have different desolvation energy values, because of the different distances in their interacting pairs. Moreover, the value for $g(r)$ in Eq. (2.2) varies in [0,1], which is equivalent to 0–100% of the desolvation energy value of the atom pairs. Since interactions between proteins are weak, we expect many pairs of atoms at distances between 5 and 7 Å. Thus, we expect that $g(r)$ is influenced by the distance between interacting pairs. On the other hand, the closer interactions (i.e. stronger) lead to higher values of desolvation energy for those interacting pairs.

The 3D structures of the interacting sub-units of three obligate (*1efv*, *1req* and *1luc*) and three non-obligate complexes (*1ava*, *1ak4* and *1atn*) of the BPPI dataset are shown in Figure 2.2. It is clear that obligate complexes have more interacting atom pairs than non-obligate ones. The numbers of atoms in each sub-unit of these complexes, the number of interacting atoms (less than 5Å), and the average of interacting pairs are shown in Table 2.6. While obligate complexes have more than 50% interacting atom pairs, non-obligate complexes have less than 35% interacting pairs. Thus, the larger the number of interacting

Table 2.6: Analysis of interacting sub-units in obligate and non-obligate complexes.

Obligate Complexes				
Complex	# atom-chain 1	# atom-chain 2	# interacting pairs	% Interacting Average
1efv , A:B	2296	1928	1880	89.69%
1req , A:B	5563	4695	3093	60.74%
1luc , A:B	2577	2516	1303	51.18%
Non-obligate Complexes				
1ak4 , A:D	1553	1375	504	34.55%
1ava , A:C	3184	1404	630	32.33%
1atn , A:D	2897	2034	584	24.44%

pairs is, the higher the value for desolvation energy for that complex is expected to be.

2.4 Conclusion

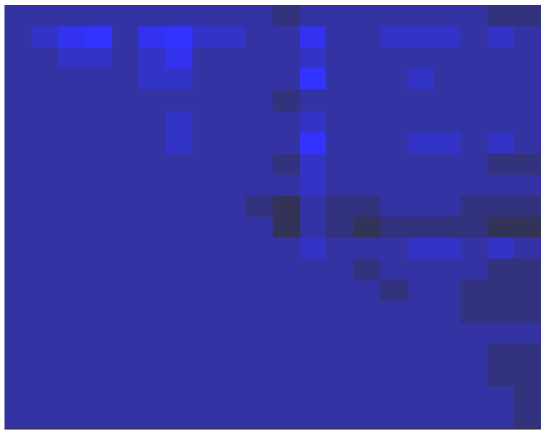
We have presented a prediction approach that uses desolvation energy properties to distinguish between obligate and non-obligate protein complexes. We have investigated various interface properties of these interactions including the NOXclass properties and desolvation energies for atom and amino acid type pairs. The prediction results on the BPPI dataset, which is a compiled version of two well-known datasets, show that the SVM classifier with 76.94% accuracy performs much better than LDR schemes coupled with quadratic and linear classifiers for all subset of features. The results also demonstrate that desolvation energy outperforms solvent accessible surface properties [15], namely interface area, interface area ratio, amino acid composition of the interface and correlation between amino acid compositions of interface and protein surface. Also, the proposed method reveals that the use of desolvation energies for atom type properties are better discriminants than SASA features for obligate and non-obligate complexes on the BPPI dataset.

Furthermore, a visual analysis (heatmaps) indicates that a few pairs of atoms/amino acids are proper for prediction. While using these features, prediction accuracies decreases a little, this decrease in performance can be acceptable because of less required time and space complexity. The approach proposed here can also be used for prediction of other types of complexes, including intra and inter domains, homo and hetero-oligomers. Other features can also be used, including geometric (e.g., shape, planarity, roughness or others), and other statistical and physicochemical properties such as residue and atom vicinity, secondary structure elements and domains, hydrophobicity, salt bridges, among others.

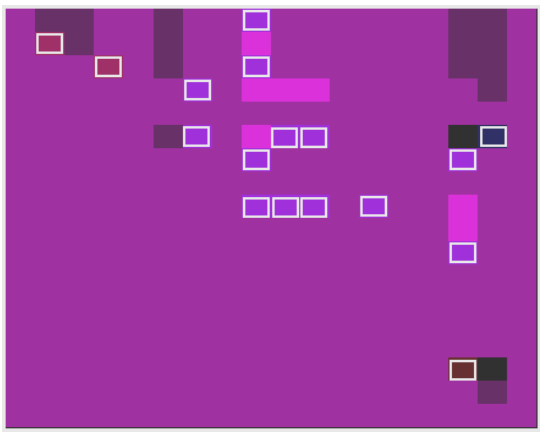
Atom pairs of obligate complexes

Amino acid pairs of obligate complexes

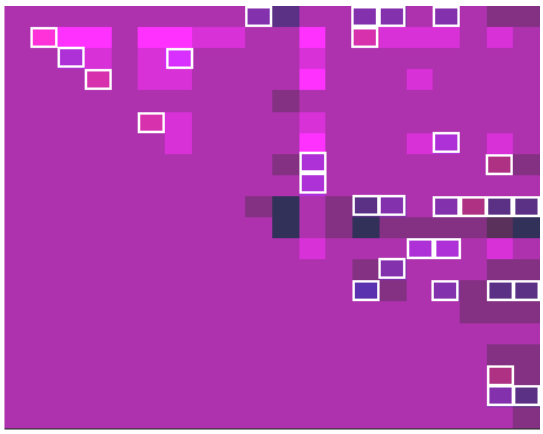
Atom pairs of non-obligate complexes



Amino acid pairs of non-obligate complexes



Atom pairs of all complexes



Amino acid pairs of all complexes

(a) Interacting atom pairs

(b) Interacting amino acid pairs

Figure 2.1: Heatmaps of desolvation energies of (a) interacting atom pairs and (b) interacting amino acid pairs.

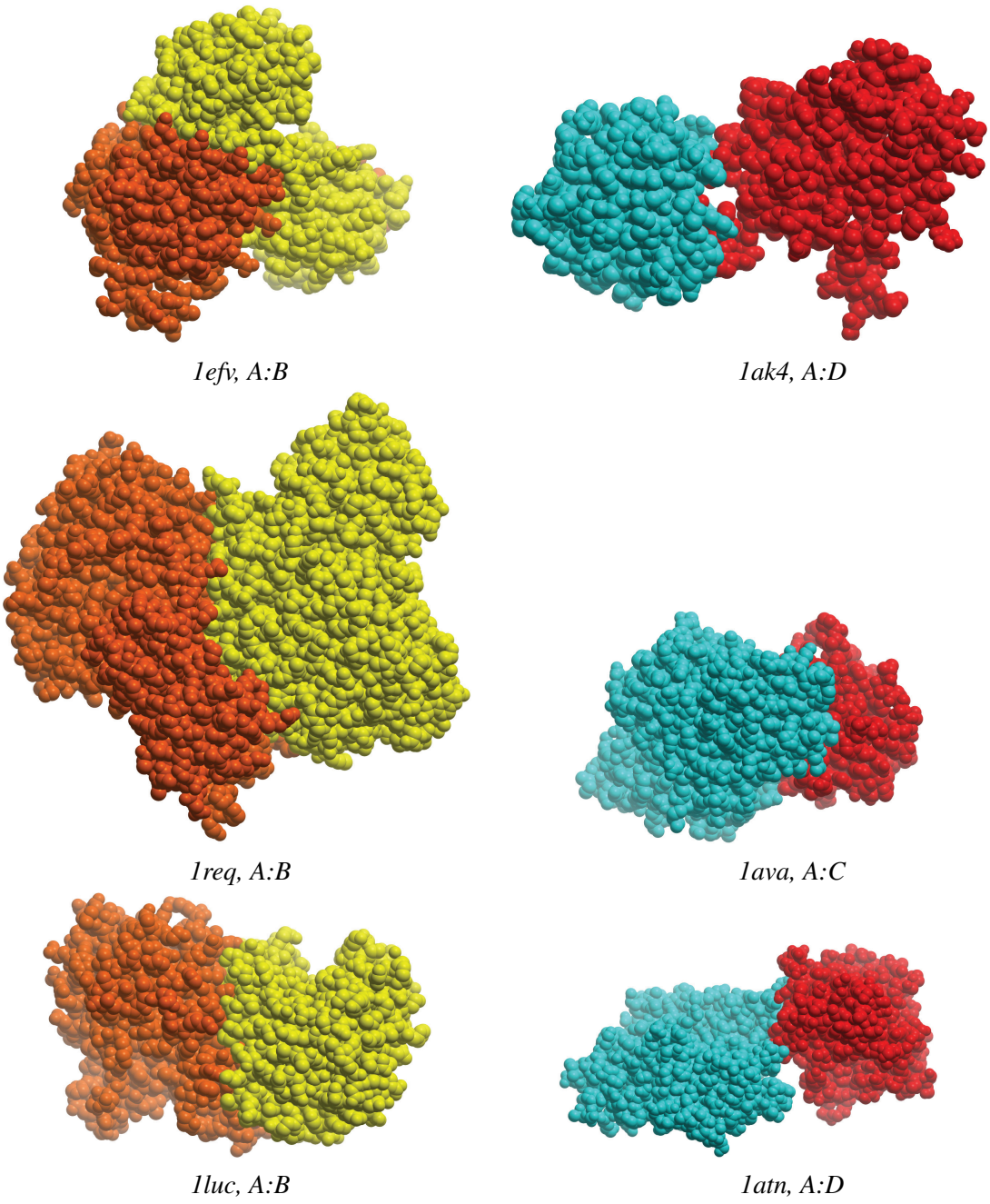


Figure 2.2: 3D structure of (a) obligate complexes and (b) non-obligate complexes visualized using ICM Browser [32].

Bibliography

[1] A. Mendelsohn and R. Brent, "Protein interaction methods-toward an endgame." *Science*, vol. 284(5422), pp. 1948–1950, 1999.

[2] M. C. Lawrence and P. M. Colman, "Shape complementarity at protein/protein interfaces," *J. Mol Biol*, vol. 234, no. 4, pp. 946–950, 1993.

[3] P. Chakrabarti and J. Janin, "Dissecting protein-protein recognition sites," *Proteins*, vol. 47, no. 3, pp. 334–343, 2002.

[4] B. Ma, T. Elkayam, H. Wolfson, and R.Nussinov, "Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces," *Proc Natl Acad Sci, USA*, vol. 100, no. 10, pp. 5772–5777, 2003.

[5] D. Xu, C. Tsai, and R. Nussinov, "Hydrogen bonds and salt bridges accross protein-protein interfaces," *Protein Eng*, vol. 10, no. 9, pp. 999–1012, 1997.

[6] H. Shanahan and J. Thornton, "Amino acid architecture and the distribution of polar atoms on the surfaces of proteins," *Biopolymers*, vol. 78, no. 6, pp. 318–328, 2005.

[7] Y. Ofran and B. Rost, "Analysing six types of protein-protein interfaces," *J Mol Biol*, vol. 325, no. 2, pp. 377–387, 2003.

[8] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc. Natl Acad. Sci, USA*, vol. 93, no. 1, pp. 13–20, 1996.

[9] I. Nooren and J. Thornton, "Diversity of protein-protein interactions," *EMBO Journal*, vol. 22, no. 14, pp. 3846–3892, 2003.

[10] L. LoConte, C. Chothia, and J. Janin, "The atomic structure of protein-protein recognition sites," *J Mol Biol*, vol. 285, no. 5, pp. 2177–2198, 1999.

[11] J. Young, "A role for surface hydrophobicity in protein protein recognition," *Protein Sci*, vol. 3, pp. 717–729, 1994.

[12] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal, "Residue frequencies and pairing preferences at protein-protein interfaces," *Proteins*, vol. 43, no. 2, pp. 89–102, 2001.

[13] O. K. A. Zen, C. Micheletti and R. Nussinov, "Comparing interfacial dynamics in protein-protein complexes: an elastic network approach," *BMC Structural Biology*, vol. 10, no. 26, 2010, doi: 10.1186/1472-6807-10-26.

[14] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction," *Proteins*, vol. 63, no. 3, pp. 490–500, 2006.

[15] H. Zhu, F. Domingues, I. Sommer, and T. Lengauer, "NOXclass: Prediction of protein-protein interaction types," *BMC Bioinformatics*, vol. 7, no. 27, 2006, doi:10.1186/1471-2105-7-27.

[16] M. S. S. Kottha, "Classifying permanent and transient protein interactions," *German Conference on Bioinformatics*, vol. 83GI, pp. 54–63, 2006.

[17] A. J. Bordner and R. Abagyan, "Statistical analysis and prediction of protein-protein interfaces," *Proteins*, vol. 60, no. 3, pp. 353–366, 2005.

[18] H. J. Caffrey and S. Somaroo, "Are protein protein interfaces more conserved in sequence than the rest of the protein surface?" *Protein Science*, vol. 13, pp. 190–202, 2004.

[19] S. Neuvirth and R. Raz, "ProMate. a structure based prediction program to identify the location of protein protein binding sites," *J Mol Biol*, vol. 338, pp. 181–199, 2004.

[20] H. Zhou and Y. Shan, "Prediction of protein interaction sites from sequence profile and residue neighbor list," *Proteins*, vol. 44, no. 3, pp. 336–343, 2001.

[21] L. Rueda, C. Garate, Banerjee, and M. M. Aziz, "Biological protein-protein interaction prediction using binding free energies and linear dimensionality reduction," proceedings of the 5th. IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2010), Nijmegen, The Netherlands, LNBI 6282, pp. 383–394, 2010.

[22] C. Camacho and C. Zhang, "FastContact: rapid estimate of contact and binding free energies," *Bioinformatics*, vol. 21, no. 10, pp. 2534–2536, 2005.

[23] L. Rueda, S. Banerjee, M. M. Aziz, and M. Raza, "Protein-protein interaction prediction using desolvation energies and interface properties," proceedings of the 2nd. IEEE International Conference on Bioinformatics & Biomedicine (BIBM 2010), pp. 17–22, 2010.

[24] J. Mintseris and Z. Weng, "Structure, function, and evolution of transient and obligate protein-protein interactions," *Proc Natl Acad Sci, USA*, vol. 102, no. 31, pp. 10 930–10 935, 2005.

[25] S.G.J.V. Eichborn and R. Preissner, "Structural features and evolution of protein-protein interactions," *Genome Inform*, vol. 22, pp. 1–10, 2010.

[26] C. Zhang, G. Vasmatzis, J. L.Cornette, and C. DeLisi, "Determination of atomic desolvation energies from the structures of crystallized proteins," *J. Mol. Biol.*, vol. 267, pp. 707–726, 1997.

[27] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[28] S. Hubbard and J. Thornton, "Naccess," 1993. [Online]. Available: http://www.bioinf.manchester.ac.uk/naccess/

[29] L. Rueda and M. Herrera, "Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space," *Pattern Recognition*, vol. 41, no. 10, pp. 3138–3152, 2008.

[30] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York, NY: John Wiley and Sons, Inc., 2000.

[31] C. L. C. Chang, "LIBSVM: a library for support vector machines," 2011, available at : http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.pdf.

[32] "ICM browser user guide," available at : http://www.molsoft.com/icm_browser.html.

Chapter 3

Analysis of Relevant Physicochemical Properties in Obligate and Non-obligate Protein-protein Interactions

3.1 Introduction

Biological protein-protein interaction (PPI) has been studied for a long time because of its fundamental role in many essential biological and cellular processes, including gene regulation, drug development, signaling, among others. Prediction of interaction types between two proteins and analyzing relevant properties involved in the interface have been studied from different perspectives. Some studies in PPI consider identifying amino acids, atoms, domains, motifs or other components of the molecules which are crucial in understanding how proteins interact with each other. These studies have been carried out mostly by relying on biological knowledge about the atoms or molecules, which, normally, are selected manually by observing groups of complexes or prediction results. Another important aspect in studying PPIs is to predict different types of complexes, including similarities between subunits (homo/hetero-oligomers), number of subunits involved in the interaction (dimers, trimers, etc.), duration of the interaction (transient vs. permanent) [1], stability of the inter-

action (non-obligate vs. obligate) [2], among others; we focus on the latter problem.

Obligate interactions are usually considered as permanent, while non-obligate interactions can be either permanent or transient [3]. Non-obligate and transient interactions are more difficult to study and understand due to their instability and short life, while obligate and permanent interactions last for a longer period of time, and hence are more stable [4]. For these reasons, an important problem is to distinguish between obligate and non-obligate complexes. To study the behavior of obligate and non-obligate interactions, in [5], it was shown that non-obligate complexes are rich in aromatic residues and arginine, while depleted in other charged residues. The study of [6] suggested that mobility differences of amino acids are more significant for obligate and large interface complexes than for transient and medium-sized ones. Traditionally, the interfaces of some non-obligate complexes were also found to be with clusters of hydrophobic residues [7]. Additionally, in [8], it was proposed that interfaces in obligate complexes are inherently hydrophobic. However, hydrophobicity at the interfaces of non-obligate complexes is not as distinguishable from the remainder of the surface as hydrophobicity at the interfaces of obligate complexes [5].

For successful prediction, using the relevant features is very important. Features are the observed properties of each sample that is used for prediction. There are a wide range of properties that can be used for PPI prediction such as analysis of solvent accessibility [2, 9], geometry [10], hydrophobicity [7, 8], sequence based features [11] and desolvation energy [12–14]. In this study, we used desolvation energies which have already been shown to be very efficient for PPI prediction [12, 13].

Different studies have claimed that only a few highly conserved residues are crucial for protein interactions [15, 16]. Moreover, the computational cost for predictions may increase substantially with more features; thus, feature selection methods can be applied to

obtain more relevant and discriminating features for prediction, and to remove the irrelevant and redundant ones that lead to greater computational cost [17]. As a consequence of this, automatic feature selection algorithms have been studied for long time in pattern recognition and prediction, and have been successfully used in obtaining relevant properties in many problems [18, 19]. One of the most efficient feature selection methods, which is based on mutual information, is Minimum Redundancy Maximum Relevance (MRMR) [20]. Recently, MRMR, which discards the redundant features from the feature vector and uses maximum relevance score as the class separability criterion, has been applied in many biological problems such as prediction of tyrosine sulfation [21] or lysine ubiquitination [22], prediction of protein-protein interactions [23] or protein-nucleic acids interactions [24], and gene selection [25, 26]. Determining the optimal number of features is one of the main challenges in all feature selection methods such as MRMR.

In this paper, a prediction approach that uses desolvation energies of atom and amino acid pairs in the interface to identify obligate and non-obligate interactions is proposed. Automatically selecting relevant properties useful in prediction of PPI is a scheme that we propose in this study and that we endeavor to do using a well-known feature selection algorithm, MRMR. Using linear dimension reduction (LDR) as the classification scheme, we demonstrate that prediction results are improved by applying feature selection and identifying relevant features for two well-known datasets of [1] and [2]. We also discuss a biologically-guided feature selection analysis, which selects features based on their polarity properties. We conclude that hydrophobic amino acids are better discriminating features for obligate and non-obligate prediction.

3.2 Materials and Methods

3.2.1 Datasets and Properties

Two pre-classified datasets of protein complexes were obtained from the studies of Zhu et al., [2], and Mintseris and Weng [1], namely the ZH and MW datasets respectively. The former contains 75 obligate (permanent) and 62 transient interactions, while the latter contains 115 obligate and 212 non-obligate interactions.

Different approaches have been developed to group different types of proteins, based on their properties for prediction. Among them, desolvation energies have been found very efficient for prediction [12] and [13]. Knowledge-based contact potential that accounts for hydrophobic interactions, self-energy change upon desolvation of charged and polar atom groups, and side-chain entropy loss is defined as desolvation energy. We have followed the equation used in [27] for the binding free energy, ΔG_{bind}, which is defined by using Eq. (3.1).

$$\Delta G_{bind} = \Delta E_{elec} + \Delta G_{des}. \tag{3.1}$$

where ΔE_{elec} is the total electrostatic energy and ΔG_{des} is the total desolvation energy, which for a protein is defined as follows [27]:

$$\Delta G_{des} = g(r)\Sigma\Sigma e_{ij}. \tag{3.2}$$

We consider the interaction between the i^{th} atom of a ligand and the j^{th} atom of a receptor. Then, e_{ij} is the atomic contact potential (ACP) [28] between them, and $g(r)$ is a smooth function based on inter-atom distance. The value of $g(r)$ is 1 for atoms that are less

Table 3.1: Description of datasets used in this study.

Name	dataset	Atom Type	Amino Acid
MW-AT	[1]	√	-
MW-AA	[1]	-	√
ZH-AT	[2]	√	-
ZH-AA	[2]	-	√

than 5 Å apart [27]. For simplicity, we consider the smooth function to be linear within the range of 5 and 7 Å, and the value of $g(r)$ is $(7-r)/2$ where r is the distance between the i^{th} atom of the ligand and the j^{th} atom of the receptor.

For each complex in our datasets, structural data from the Protein Data Bank (PDB) [29] was collected. Two chains (ligand and receptor) were extracted from each PDB file. We consider 18 different atom types as in [28]. Thus, for each protein complex a feature vector with 18^2 values were obtained, where each feature contains the cumulative sum of desolvation energies of a pair of atom types, computed using Eq. (3.2). Since the order of interacting atom pairs is not important, the final length of the feature vector for each complex is 171 which corresponds to the number of unique pairs. We have also considered pairs of amino acids, and for this, we computed 20^2 desolvation energy values for all pairs of atoms using Eq. (3.2), by accumulating the values of the corresponding atoms for each pair of amino acids. Avoiding repeated pairs resulted in 210 different features (unique pairs of amino acids).

By using desolvation energies for different types of features and different datasets, four subsets of features for prediction and evaluation were generated, as listed in Table 3.1.

3.2.2 The Prediction Methods

For prediction, we have used LDR. The basic idea of LDR is to represent an object of dimension n as a lower-dimensional vector of dimension d, achieving this by performing a linear transformation. We consider two classes, obligate and non-obligate, represented by two normally distributed random vectors $\mathbf{x}_1 \sim N(\mathbf{m}_1, \mathbf{S}_1)$ and $\mathbf{x}_2 \sim N(\mathbf{m}_2, \mathbf{S}_2)$, respectively, with p_1 and p_2 the *a priori* probabilities. After the LDR is applied, two new random vectors $\mathbf{y}_1 = \mathbf{A}\mathbf{x}_1$ and $\mathbf{y}_2 = \mathbf{A}\mathbf{x}_2$, where $\mathbf{y}_1 \sim N(\mathbf{A}\mathbf{m}_1; \mathbf{A}\mathbf{S}_1\mathbf{A}^t)$ and $\mathbf{y}_2 \sim N(\mathbf{A}\mathbf{m}_2; \mathbf{A}\mathbf{S}_2\mathbf{A}^t)$ with $\mathbf{m}_i$ and $\mathbf{S}_i$ being the mean vectors and covariance matrices in the original space, respectively. The aim of LDR is to find a linear transformation matrix $\mathbf{A}$ in such a way that the new classes ($\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$) are as separable as possible. Let $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$ and $\mathbf{S}_E = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$ be the within-class and between-class scatter matrices respectively. Various criteria have been proposed to measure this separability [30]. We consider the following two LDR methods:

(a) The heteroscedastic discriminant analysis (HDA) approach [30], which aims to maximize the following function, optimized via eigenvalue decomposition:

$$J_{HDA}(\mathbf{A}) = tr\{(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}[\mathbf{A}\mathbf{S}_E\mathbf{A}^t - \mathbf{A}\mathbf{S}_W^{\frac{1}{2}}\frac{p_1\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}}) + p_2\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1p_2}\mathbf{S}_W^{\frac{1}{2}}\mathbf{A}^t]\}. \tag{3.3}$$

(b) The Chernoff discriminant analysis (CDA) approach [30], which aims to maximize the following function, maximized via a gradient-based algorithm:

$$J_{CDA}(\mathbf{A}) = tr\{p_1 p_2 \mathbf{A}\mathbf{S}_E\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1} + \log(\mathbf{A}\mathbf{S}_W\mathbf{A}^t) - p_1\log(\mathbf{A}\mathbf{S}_1\mathbf{A}^t) - p_2\log(\mathbf{A}\mathbf{S}_2\mathbf{A}^t)\}. \tag{3.4}$$

In order to classify each complex, first, a linear algebraic operation $\mathbf{y} = \mathbf{A}\mathbf{x}$ is applied to the n-dimensional vector, obtaining $\mathbf{y}$, a d-dimensional vector, where d is ideally much smaller than n. The linear transformation matrix $\mathbf{A}$ corresponds to the one obtained by one of the LDR methods, namely HDA or CDA. The resulting vector $\mathbf{y}$ is then passed through a quadratic Bayesian (QB) classifier [30], which is the optimal classifier for normal distributions and a linear Bayesian (LB) which is obtained by deriving a Bayesian classifier with a common covariance matrix, $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$.

3.2.3 The Feature Selection Methods

Feature selection is a process used to select the best subset of features that represents the whole feature set efficiently. This process not only reduces the size of the feature vector and helps to find discriminating features for prediction among all the features but also reduces the prediction time.

Feature Selection based on MRMR

MRMR is a commonly used feature selection method which uses mutual information to compute relevancy and redundancy among features [20, 31]. In MRMR, the features that have minimal redundancy and are highly relevant to the classes are recursively selected and scored to generate a ranking list of the features.

Determining the optimal number of features is a main challenge in all feature selection methods, because it has significant effects on classification performance. Furthermore,

this number is different from one dataset or subset of features to another and is a time-consuming process. This is what will be demonstrated in the next section. Moreover, in the original version of MRMR ($MRMR^{org}$), if we want to select i features from the whole n features ($i < n$), the top i features in the list of scored features in MRMR will be selected. In this case, we may encounter some zero-samples (complexes) that do not have any values for the selected features. Zero-samples may lead to misclassification errors.

To solve this problem, in this study, we applied an efficient method that we call $MRMR^{pro}$, for determining the optimal number of features. For this, in the first step, MRMR is applied to sort the features based on to their relevance scores. By starting from the top of the MRMR feature list, each feature is examined to decide about its existence in the final list of selected features. A feature can be selected if (a) it has a high MRMR score and (b) it has values for more new complexes. For instance, if m complexes have non-zero values for the first selected feature, the second feature will be selected if at least one new complex (not in the m previously seen complexes) has any non-zero values for it. Having at least one feature for each complex and selecting at least four features are used as a criterion to determine the final subset of selected features in $MRMR^{pro}$. By applying our feature selection method, four features for ZH-AT, 12 features for ZH-AA, four features for MW-AT, and 14 features for MW-AA datasets have been found.

Biological Feature Selection

According to the authors of [32], by considering polarity, amino acids can be divided into the following three groups:

- Hydrophobic: Alanine, Isoleucine, Valine, Leucine, Phenylalanine and Proline are hydrophobic amino acids and avoid contact with water.

- Hydrophilic: Arginine, Asparagine, Aspartic acid, Cysteine, Glutamic acid, Glutamine, Histidine and Serine are hydrophilic amino acids and like to interact with water.

- Amphipathic: Lysine, Methionine, Threonine, Tryptophan and Tyrosine are amphipathic amino acids and have both polar and non-polar behavior and hence a tendency to form interfaces between hydrophobic and hydrophilic molecules.

As Glycine is a neutral amino acid, generally, we have 19 amino acids according to this classification. Using this information, the desolvation energy values can be grouped by amino acid type into three subsets of features with only hydrophobic, only hydrophilic and only amphipathic type. In each subset, only the interaction between two pairs of amino acids in the same group is possible. These three subsets of features were applied to our classifiers to check the accuracy and decide the most relevant features for distinguishing obligate and non-obligate complexes.

3.3 Results and Discussions

3.3.1 Experimental Settings

As discussed earlier, for the LDR schemes, four different classifiers were implemented and evaluated, namely the combinations of HDA and CDA with QB and LB classifiers. In a 10-fold cross validation process, reductions to dimensions $d = 1, \ldots, 20$ were performed, followed by QB and LB. The maximum average classification accuracy was taken into account for each classifier, which is the one that is reported for each subset of datasets.

For feature selection, we applied MRMR^{org} and MRMR^{pro}, which used the list of scored

features obtained from the online mRMR tool [1]. Furthermore, we applied biologically guided feature selection methods to our datasets to find the most relevant subset of features.

After running the classifiers in a 10-fold cross validation procedure, the accuracy was computed as follows: $acc = (TP+TN)/N$, where TP and TN are the total numbers of true positive (obligate) and true negative (non-obligate) counters over the 10 folds, respectively, and N is the total number of complexes in the dataset.

3.3.2 Analysis of MRMR-based Feature Selection

The performances of LDR for amino acid type features of the MW and ZH datasets are plotted against the number of selected features in Figure 3.1, respectively. The order of the selected features is based on the order of features scored by MRMR^{org}. It is clear that the best number of features for the MW-AA dataset is 21, achieving 79.75% prediction accuracy while this number for the ZH-AT dataset is 28 with 86.86% prediction accuracy. This results demonstrate that the best number of features is different from one dataset or subset of features to another.

The results of the LDR classifier for the MW and ZH datasets with atom and amino acid type features with/without using feature selection are depicted in Table 3.2. After finding the optimal number of features by MRMR^{pro}, the prediction accuracies for the same number of features by using MRMR^{org} are also reported – the numbers of selected features are the same. The only difference between MRMR^{org} and MRMR^{pro} in this analysis is based on the selected features. For instance, the top four features of $\{1,2,3,4\}$ were selected for the MW-AT using MRMR^{org} while the selected features using MRMR^{pro} were $\{1,3,4,6\}$.

It is clear that for all datasets, except the ZH-AT, the predictions show better perfor-

[1] Available at http://penglab.janelia.org/proj/mRMR/

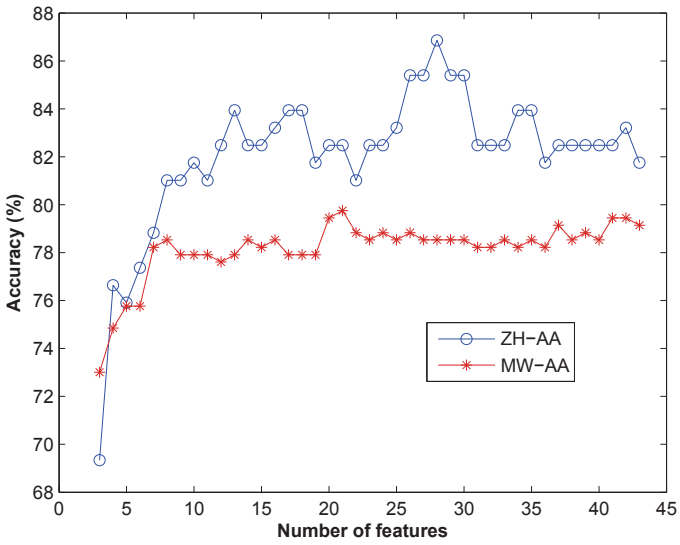


Figure 3.1: Prediction accuracy of the ZH-AA and MW-AA datasets using MRMR feature selection method.

mance by using feature selection methods. The best accuracy, 82.13%, for LDR methods without using feature selection is achieved on the ZH-AA dataset, which is still lower than the best accuracy achieved by MRMR^{org} and MRMR^{pro} (82.48% and 83.21% respectively). The most notable difference between accuracies for with and without feature selection is approximately 3%, which observed in the MW dataset. While there is a slight decrease in prediction accuracy for atom type features in the ZH dataset (ZH-AT) after applying feature selection, this decrease in performance can be acceptable considering that only four features instead of the 171 original features are used for prediction. This also implies savings in the required classification time and space complexity. In general, it can be concluded that a few pairs of atoms/amino acids are appropriate for prediction.

Furthermore, it is clearly observable that our MRMR^{pro} feature selection method performs better than MRMR^{org} for all datasets. For the ZH-AT dataset, the prediction results for MRMR^{org} and MRMR^{pro} are the same, because the subset of selected features for both

Table 3.2: Prediction results for LDR classifier by using different MRMR-based feature selection methods.

Feature subset	No Feature selection		$MRMR^{org}$		$MRMR^{pro}$	
	# features	accuracy	# features	accuracy	# features	accuracy
MW-AT	171	77.91	4	78.22	4	78.53
MW-AA	210	75.77	14	78.53	14	78.83
ZH-AT	171	78.39	4	77.37	4	77.37
ZH-AA	210	82.13	12	82.48	12	83.21

of them are also the same. This indicates that $MRMR^{pro}$ is more accurate in finding relevant features for prediction of obligate and non-obligate features.

3.3.3 Analysis of Biologically-guided Feature Selection Methods

As explained earlier, amino acids can be classified in three groups of hydrophobic (6), hydrophilic (8) and amphipathic (5) amino acids. Using this information, the desolvation energy values can be grouped by amino acid type into three subsets of features with only hydrophobic, only hydrophilic and only amphipathic type. In each subset, only the interaction between two pairs of amino acids in the same group is possible, and totally, there are 21, 36 and 15 interacting hydrophobic, hydrophilic and amphipathic amino acid pairs respectively. These three subset of features were applied to our classifiers to decide the distinguishing features of obligate and non-obligate complexes. The results of LDR for the MW and ZH datasets with these three groups of features are depicted in Table 3.3.

It is clear that hydrophobic amino acid pairs achieve the highest LDR accuracies for both MW-AA (75.54%) and ZH-AA (77.07%) datasets, and hence they are the best properties for prediction. After hydrophobic amino acid pairs, hydrophilic amino acid pairs are more relevant than amphipathic pairs in classification of obligate and non-obligate complexes.

Table 3.3: Prediction results for LDR classifier by using biologically guided feature selection methods for the MW and ZH datasets.

	Hydrophobic		Hydrophilic		Amphipathic	
Datasets	# Features	Accuracy	# Features	Accuracy	# Features	Accuracy
MW-AA	21	75.54	36	71.49	15	66.36
ZH-AA	21	77.07	36	68.97	15	67.72

3.3.4 Visual Analysis of Relevant Features

To visually observe the effect of polarity of amino acids among obligate and non-obligate complexes and to find appropriate properties for prediction, we have used heatmaps of interacting amino acid pairs. For this, desolvation energies of all complexes in the ZH and MW datasets were calculated to generate the feature vectors by amino acid type properties for obligate and non-obligate complexes, separately. Then, we computed the column-wise sum, obtaining two sums of feature vectors (obligate/non-obligate). We represented each sum of feature vector as a 20x20 matrix for amino acids. Then, as explained earlier, all amino acids were sorted in three groups of amphipathic, hydrophilic and hydrophobic. Glycine which is a neutral amino acid added at the end of the list. After that, each matrix was condensed as an upper triangular matrix. Subtracting all indices for obligate and non-obligate, heatmaps were generated. Heatmaps of interacting chains in the MW and ZH datasets are shown in Figure 3.2. In the heatmaps, the lighter the green color is, the larger the desolvation energy value of that interacting amino acid pair is, while negative desolvation energies correspond to red colors.

There are many hydrophobic-hydrophobic (top right most green squares) and hydrophobic-amphipathic (bottom right most green squares) amino acid pairs in both ZH and MW datasets which indicate that these properties are more relevant for prediction. That is what was previously shown in Table 3.3 for hydrophobic-hydrophobic amino acid pairs.

In contrast, the interacting pairs of hydrophobic-hydrophilic amino acids, in general, are not suitable for prediction because of the less number of interacting pairs in the heatmaps.

Similarly, it can be concluded that hydrophilic-hydrophilic amino acid pairs are more discriminating than hydrophilic-amphipathic pairs because of the number of interacting pairs in these two groups while they are less discriminating than only hydrophobic or hydrophobic-amphipathic amino acids.

In contrast, on the bottom left side of the heatmaps (first five left amphipathic amino acids), there are only four amphipathic-amphipathic amino acid pairs for the MW dataset and six pairs for the ZH datasets. Thus, the only amphipathic amino acid pairs are not appropriate features for prediction of obligate and non-obligate complexes, as shown also in Table 3.3 and discussed previously.

To validate the extracted results from the heatmaps, we have done a post-processing analysis on the selected features by MRMR^{pro}. As explained earlier, 12 and 14 features were selected by MRMR^{pro} from the ZH-AA and MW-AA datasets respectively. A summary of types of these selected features is shown in Table 3.4. The interactions of polar amino acids with Glycine are shown in the last row of the table. In both MW-AA and ZH-AA datasets, most of the features selected by MRMR^{pro} are hydrophobic pairs and then hydrophobic-amphipathic amino acid pairs as it was shown in the heatmaps. The fact that the number of hydrophilic amino acid pairs is greater than the number of amphipathic amino acid pairs and less than two groups of hydrophobic and hydrophobic-amphipathic amino acids, is also shown in this table.

As a consequence of this, it can be concluded from the heatmaps and the numerical analysis that interacting amino acids can be sorted by relevance as follows: hydrophobic-hydrophobic, hydrophobic-amphipathic, hydrophilic-hydrophilic, hydrophilic-amphipathic,

Table 3.4: Post-processing analysis of features selected by MRMR^{pro} for the MW and ZH datasets.

Interaction Type	MW-AA	ZH-AA
Hydrophobic only	6	4
Hydrophobic-Hydrophilic	1	2
Hydrophobic-Amphipathic	3	3
Hydrophilic only	1	2
Hydrophilic-Amphipathic	1	0
Amphipathic only	0	1
Gly-Others	2	0

hydrophobic-hydrophilic, and amphipathic-amphipathic.

3.4 Conclusion

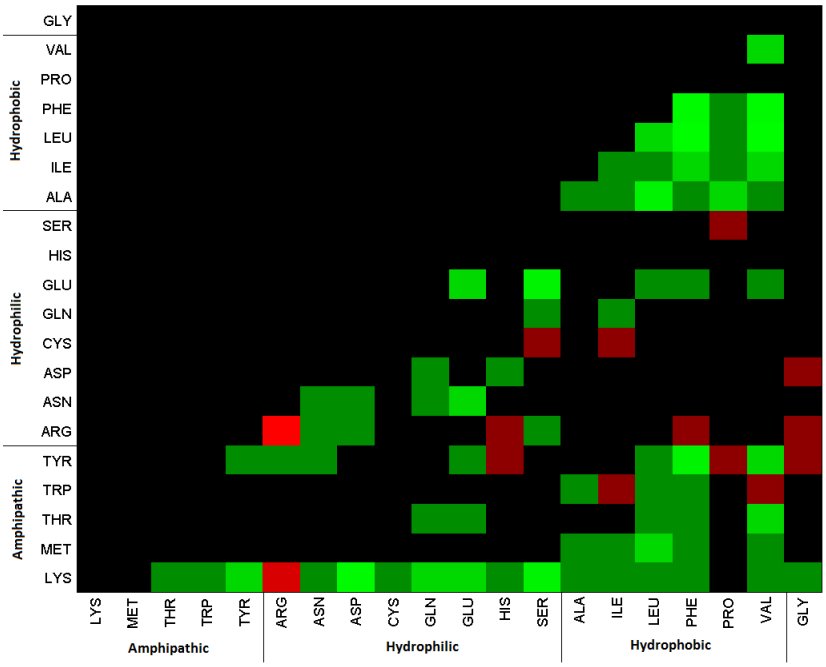
We have proposed an approach for automatically selecting relevant properties useful in prediction and analysis of obligate and non-obligate protein complexes. Our prediction approach uses desolvation energies of pairs of atoms or amino acids present in the interfaces of such complexes and the classification is performed via different LDR methods that involve heteroscedastic criteria.

The results on two well-known datasets of pre-classified complexes demonstrate that only a few subset of features are crucial in distinguishing obligate and non-obligate complexes– these relevant features can be found by using a feature selection method such as MRMR. Also, our MRMR^{pro}, which is based on MRMR, not only can find the best number of features but also can find the best relevant subset of features for prediction.

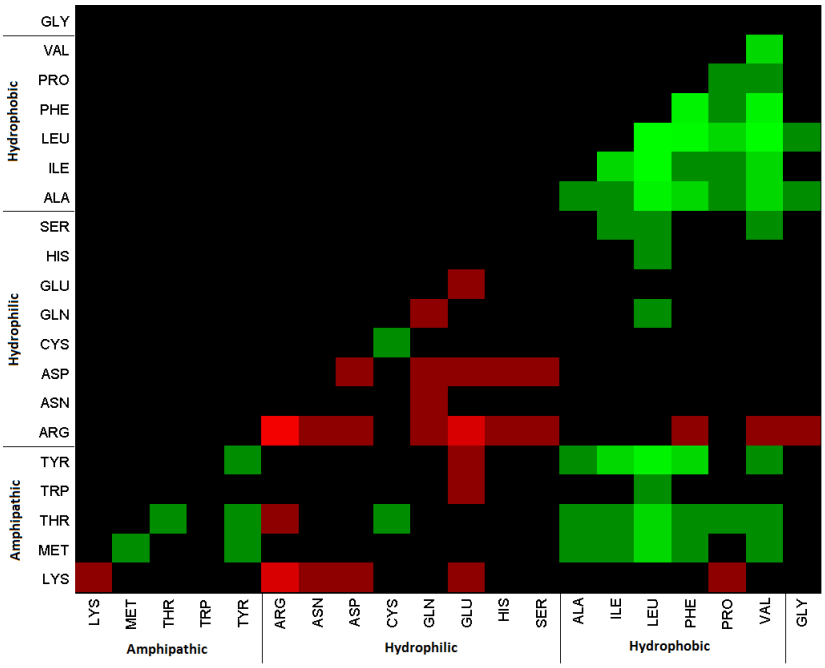
Furthermore, based on our visual (heatmaps) and numerical analysis, interacting amino acid pairs can be sorted from the most to the least relevant pairs for prediction of obligate and non-obligate complexes as follows: hydrophobic pairs, hydrophobic-amphipathic, hy-

drophilic pairs, hydrophilic-amphipathic, hydrophobic-hydrophilic, and amphipathic pairs.

The approach proposed here can also be used for prediction of other types of complexes, including intra and inter domains, homo and hetero-oligomers. Other properties can also be used including geometric (e.g., shape, planarity, roughness or others), and other statistical and physicochemical properties such as residue and atom vicinity, secondary structure elements and domains, hydrophobicity, salt bridges, among others. Applying different feature selection and feature weighting methods can also be used to find the relevant features for prediction.



(a) The MW-AA dataset



(b) The ZH-AA dataset

Figure 3.2: Heatmaps of desolvation energies of interacting amino acid pairs in (a) the MW-AA dataset and (b) the ZH-AA dataset. Amino acids are grouped based on their polarity.

Bibliography

[1] J. Mintseris and Z. Weng, "Structure, function, and evolution of transient and obligate protein-protein interactions," *Proc Natl Acad Sci, USA*, vol. 102, no. 31, pp. 10 930–10 935, 2005.

[2] H. Zhu, F. Domingues, I. Sommer, and T. Lengauer, "NOXclass: Prediction of protein-protein interaction types," *BMC Bioinformatics*, vol. 7, no. 27, 2006, doi:10.1186/1471-2105-7-27.

[3] I. Nooren and J. Thornton, "Diversity of protein-protein interactions," *EMBO Journal*, vol. 22, no. 14, pp. 3846–3892, 2003.

[4] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc. Natl Acad. Sci, USA*, vol. 93, no. 1, pp. 13–20, 1996.

[5] L. LoConte, C. Chothia, and J. Janin, "The atomic structure of protein-protein recognition sites," *J Mol Biol*, vol. 285, no. 5, pp. 2177–2198, 1999.

[6] O. K. A. Zen, C. Micheletti and R. Nussinov, "Comparing interfacial dynamics in protein-protein complexes: an elastic network approach," *BMC Structural Biology*, vol. 10, no. 26, 2010, doi: 10.1186/1472-6807-10-26.

[7] J. Young, "A role for surface hydrophobicity in protein protein recognition," *Protein Sci*, vol. 3, pp. 717–729, 1994.

[8] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal, "Residue frequencies and pairing preferences at protein-protein interfaces," *Proteins*, vol. 43, no. 2, pp. 89–102, 2001.

[9] H. Shanahan and J. Thornton, "Amino acid architecture and the distribution of polar atoms on the surfaces of proteins," *Biopolymers*, vol. 78, no. 6, pp. 318–328, 2005.

[10] M. C. Lawrence and P. M. Colman, "Shape complementarity at protein/protein interfaces," *J. Mol Biol*, vol. 234, no. 4, pp. 946–950, 1993.

[11] J. Mintseris and Z. Weng, "Atomic contact vectors in protein-protein recognition," *PROTEINS: Structure, Function and Genetics*, vol. 53, pp. 629–639, 2003.

[12] L. Rueda, S. Banerjee, Md. Aziz, and M. Raza, "Protein-protein interaction prediction using desolvation energies and interface properties." *Proceedings of the 2nd. IEEE International Conference on Bioinformatics & Biomedicine (BIBM 2010)*, pp. 17–22, 2010.

[13] L. Rueda, C. Garate, Banerjee, and Md. Aziz, "Biological protein-protein interaction prediction using binding free energies and linear dimensionality reduction." *Proceedings of the 5th. IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2010)*, pp. 383–394, 2010.

[14] Md. Aziz, M. Maleki, L. Rueda, M. Raza, and S. Banerjee, "Prediction of biological protein-protein interactions using atom-type and amino acid properties," *Wiley-VCH Proteomics 2011*, vol. 11, no. 19, pp. 3802–10, 2011.

[15] S. De, O. Krishnadev, N. Srinivasan, and N. Rekha, "Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different." *BMC Structural Biology*, vol. 5, no. 15, 2005.

[16] J. V. Eichborn, S. Gnther, and R. Preissner, "Structural features and evolution of protein-protein interactions." *Intenational Conference of Genome Informatics.*, vol. 22, pp. 1–10, 2010.

[17] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Elsevier Academic Press, 2006.

[18] L. Luo, L. Ye, M. Luo, D. Huang, H. Peng, and F. Yang, "Methods of forward feature selection based on the aggregation of classifiers generated by single attribute." *Comput Biol Med.*, vol. 41, no. 7, pp. 435–41, 2011.

[19] Y. Lee, C. Chang, and C. Chao, "Incremental forward feature selection with application to microarray gene expression data." *Journal of biopharmaceutical statistics*, vol. 18, no. 5, pp. 827–840, 2008.

[20] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data." *Journal of Bioinformatics and Computational Biology.*, vol. 3, no. 2, pp. 185–205, 2005.

[21] S. Niu, T. Huang, K. Feng, Y. Cai, and Y. Li, "Prediction of Tyrosine sulfation with MRMR feature selection and analysis." *J Proteome Res*, vol. 9, no. 12, pp. 6490–6497, 2010.

[22] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, , and Y. Li, "Prediction of Lysine ubiquitination with MRMR feature selection and analysis." *Amino Acids*, 2011.

[23] L. Liu, Y. D. Cai, W. C. Lu, C. Peng, and B. Niub, "Prediction of proteinprotein interactions based on PseAA composition and hybrid feature selection." *Biochemical and Biophysical Research Communications*, vol. 380, no. 2, pp. 318–322, 2009.

[24] Y. Yuan, x. Shi, X. Li, W. Lu, Y. Cai, L. Gu, L. Liu, M. Li, X. Kong, and M. Xing, "Prediction of interactiveness of proteins and nucleic acids based on feature selections." *Mol Divers.*, vol. 14, no. 4, pp. 627–33, 2009.

[25] P. Mundra and J. Rajapakse, "SVM-RFE with MRMR filter for gene selection." *IEEE Transactions on Nanobioscience*, vol. 9, no. 1, pp. 31–37, 2010.

[26] Y. Zhao and Z. Yand, "Improving MSVM-RFE for multiclass gene selection." *The Fourth International Conference on Computational Systems Biology (ISB2010)*, 2010.

[27] C. Camacho and C. Zhang, "FastContact: rapid estimate of contact and binding free energies," *Bioinformatics*, vol. 21, no. 10, pp. 2534–2536, 2005.

[28] C. Zhang, G. Vasmatzis, J. L.Cornette, and C. DeLisi, "Determination of atomic desolvation energies from the structures of crystallized proteins," *J. Mol. Biol.*, vol. 267, pp. 707–726, 1997.

[29] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[30] L. Rueda and M. Herrera, "Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space," *Pattern Recognition*, vol. 41, no. 10, pp. 3138–3152, 2008.

[31] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[32] G. A. Petsko and D. Ringe, *Protein structure and function*. New Science Press, 2004.

Chapter 4

The role of electrostatic energy in prediction of obligate protein-protein interactions

4.1 Background

Gene expression, cell growth, proliferation, signal transduction, cellular motion and gene regulation are some of the essential biological processes in living cells which are controlled by proteins [1]. As a consequence of this, more attention has been drawn to this field of study, in particular, for identification and analysis of interacting proteins and their relevant properties [2, 3]. Proteins bind to each other, creating protein-protein interactions (PPIs) through a combination of hydrophobic bonding, van der Waals forces and salt bridges. The strength of these interactions may depend on the size of the binding interface which can be large surfaces, small binding clefts or even a few peptides.

Prediction of PPI types is one of the main challenges when studying protein interactions. There are different types of PPIs and their associated prediction problems, including homo vs. hetero-oligomers based on the similarities between sub-units [4], dimers vs. trimers based on the number of interacting sub-units, transient vs. permanent based on the duration

of the interaction [5] and obligate vs. non-obligate based on the stability of the complex [6–9]. Despite obligate and permanent interactions, which are more stable and last for a longer period of time, studying non-obligate and transient interactions is a very difficult problem, because of their instability and short life [10]. We focus on distinguishing between obligate and non-obligate complexes.

Using relevant features or observed properties of protein complexes is essential in performing accurate predictions. As a consequence of this, previous studies in PPI have considered a wide range of relevant properties that can be used for PPI prediction including geometric properties [11], recognition of sites [12], conservation of residues present in the surface of PPIs [13, 14], hydrogen bonds and salt bridges on the surface of the proteins [13], solvent accessibility [6, 15], hydrophobicity [8, 16], sequence-based features [17], desolvation energy [18–20] and recently, electrostatic energy [21]. Electrostatic interactions are one of three types of non-covalent interactions, which occur between electrically charged atoms having both positive and negative interactions [22]. Non-covalent interactions are very common between macromolecules such as proteins. Van der Waal interactions, which occur between any pair of charged atoms that are close to each other, and non-polar interactions, which occur between atoms that do not have any charge, are other two types of non-covalent interactions.

In previous studies, it has been claimed that only a few highly conserved residues are important for protein interactions [23–25]. Moreover, removing irrelevant and redundant features not only can decrease the computational burden, but also may increase the prediction performance [26]. These are the main tasks carried out by specialized machine learning algorithms for feature selection and classification. In this regard, automatic feature selection algorithms have been used in many biological problems such as prediction

of tyrosine sulfation and lysine ubiquitination [27, 28], prediction of protein-protein interactions [25, 29], protein-nucleic acid interactions [30], gene selection [31, 32] and gene expression [33]. In this study, a few feature selection methods, including gain ratio (GR), information gain (IG), chi-square (Chi2) and minimum redundancy maximum relevance (mRMR), are applied to score and rank features based on their relevance, and select the top ranked features for prediction of obligate and non-obligate PPIs.

In one of our recent works [21], a model to predict obligate and non-obligate protein interaction types has been presented in which electrostatic energy values for both atom and amino acid pairs present in the interface were considered as the input features of the classifiers. Linear dimensionality reduction (LDR) and a support vector machine (SVM) were applied as the classifiers to predict these types. The prediction results of that study for two well-known datasets, referred to as the ZH [6] and MW [5] datasets, show an impressive accuracy in prediction. For the ZH dataset, an accuracy of 96.18% was achieved by using SVM and electrostatic energy values of amino acid type features, which is much higher than the accuracy obtained by using four interface properties including interface area, interface area ratio, conservation score and gap volume index of NOXClass [6] with 88.52% prediction accuracy (as reported by the authors), 46 solvent accessible and interface area properties of [18] with 81.83% prediction accuracy, 210 features of solvent accessible area of [34] with 92.20% prediction accuracy, and even higher than 210 desolvation energy values for amino acid type features of [18] with 83.21% prediction accuracy. Similarly, applying the proposed scheme on the MW dataset demonstrates that using electrostatic energy values of amino acid type features (95.38% prediction accuracy for SVM) is better than using the four interface features as in [6] (77.96% prediction accuracy), and also better than using 210 desolvation energy properties as in [18] (78.83% prediction accuracy). Generally, the

results reported in our previous study [21] implied an increase of at least 5% in prediction performance from previous approaches.

This paper is an extension of the work presented in [21] by incorporating a wider range of classification techniques that include LDR, SVM, naive Bayes (NB) and *k*-nearest neighbor (*k*-NN). Distance cutoff selection approaches are also used for analysis of long-range interactions (ranging from 5Å to 13Å), and feature selection algorithms for identifying relevant physicochemical properties of interacting pairs of atoms and amino acids, including GR, IG, Chi2 and mRMR, and an extended visual analysis. The results confirm that electrostatic energy with distance cutoffs ranging from 9Å to 12Å is the best property to predict obligate and non-obligate PPIs on the basis of the experimental results using different classification methods and different distance cutoffs on two well-known datasets. This is due the fact that using electrostatic energy with a long distance cutoff, atoms on the surface and some atoms buried under the surface may participate in the prediction that lead to excellent classification performance. In fact, the latter is a problem that opens an interesting research avenue in the field. Furthermore, using LDR as the classification scheme, we demonstrate that prediction results are improved by applying feature selection and identifying more relevant and discriminative features, while removing redundant and noisy ones for the two datasets.

4.2 Methods

4.2.1 Datasets

In this study, we have used the same datasets as those used in [18, 25]. The first dataset, referred to as the ZH dataset, was obtained from the study of Zhu *et al.* [6]. It originally

contained 62 non-obligate and 75 obligate complexes. Since the electrostatic energy values of some complexes (1cc0 A:E, 1qbk B:C, 1b8a A:B, 1cli A:B, 1qav A:B, 1bkd R:S and 1nse A:B) cannot be computed, they were removed from the ZH dataset. The second dataset, referred to as the MW dataset, was obtained from the study of Mintseris *et al.* [5], and originally contained 209 non-obligate and 115 obligate complexes. Similarly, 24 complexes of the original dataset (1b7y A:B, 1be3 CDEGK:A, 1jb0 AB:C, 1jb0 AB:D, 1jb0 AB:E, 1jro A:BD, 1jv2 A:B, 1k28 A:D, 1kqf A:B, 1ldj A:B, 1m2v A:B, 1mjg AB:M, 1nbw AC:B, 1prc C:HLM, 1bgx HL:T, 1de4 CF:A, 1ezv E:XY, 1is8 ABEJCIDHGF:KLOMN, 1m2o AC:B, 1o94 AB:CD, 1qfu AB:HL, 2hmi AB:CD, 4cpa I:0 and 2q33 A:B) were left out because the electrostatic energy values for all atoms in their interfaces cannot be computed.

4.2.2 Prediction properties

Different properties can be employed to predict protein interactions and, in particular, types of protein complexes. In our recent study [21], it has been demonstrated that electrostatic energy is a powerful property to predict obligate and non-obligate complexes. Moreover, we have previously shown that desolvation energy is also very effective for prediction of these types of PPIs [18, 20]. In this study, electrostatic energy properties are used for prediction of obligate and non-obligate interactions and desolvation energy properties are used for comparison purposes. Our method to obtain these prediction properties are summarized below.

Desolvation energy

Considering e_{ij} as the atomic contact potential (ACP) between the i^{th} atom of a ligand and the j^{th} atom of a receptor, the total desolvation energy for a protein (ΔG_{des}) is defined as

follows [35]:

$$\Delta G_{des} = \Sigma_{i=1}^{18} \Sigma_{j=1}^{18} e_{ij} * g(r_{ij}). \tag{4.1}$$

where all atom pairs (18 different atoms) are considered in the double summation and $g(r_{ij})$ is a smooth function based on the distance of interacting atoms i and j. For simplicity, in our comparisons, the value of $g(r_{ij})$ is 1 for pairs of atoms that are less than the selected distance cutoff apart from each other, and 0 otherwise. Using Eq. (4.1), the desolvation energy between any pair of ligand and receptor can be calculated. Thus, by following the approach of [36], it is possible to compute the desolvation energy by using different criteria. Desolvation energy values are calculated for atom and amino acid types. More details about the computation of desolvation energy values for atom and amino acid types as features can be found in [20].

Electrostatic energy

The main property that we use in this study for predicting obligate and non-obligate complexes is electrostatic energy, because of its role in charged molecules [37]. Electrostatic energy involves a long-range interaction and can occur between charged atoms of two interacting proteins or two different molecules. Moreover, these interactions can occur between charged atoms on the protein surface and charges in the environment. In order to compute electrostatic energy values, PDB2PQR [38] and APBS [39] software packages are used.

For each complex in the datasets, after extracting the structural data from the Protein Data Bank (PDB) [40], PDB2PQR is employed for preparing the structures for electrostatic calculations. Adding missing heavy atoms, placing missing hydrogen atoms and assigning charges are some of the main tasks performed by PDB2PQR. To customize the parameters

of PDB2PQR in our experiments, we consider the following parameters: (a) the AMBER forcefield is employed (b) "apbs-input" is specified to create output files with ".in" extension, and (c) "−−chain" is also specified to include the chain name in the ".pqr" files. The outputs of this package, a "pqr" file and an "in" file, are the inputs to APBS.

APBS is utilized to compute electrostatic energy values of interactions between solutes in salty and aqueous media. In APBS, the Poisson-Boltzmann equation is solved numerically and electrostatic calculations are performed in a range from ten to million atoms. Before running APBS, the parameters should be set accordingly as detailed in [21].

To compute the features for classification, first of all, a cutoff distance should be defined. While in most studies, this cutoff, which is the maximum distance between interacting atoms, is considered to be less than 7 Å we use cutoffs greater than 7 Å . Due to the long-range nature of electrostatic interactions, electrostatic forces towards the stability of the protein complex may be affected by atoms that are under the surface of the proteins. Afterwards, the distances between all atom pairs of interacting chains are computed and those that are lower than our defined cutoff distance are considered as interface atoms. The quaternary structures of chains A (shown in red) and B (shown in blue) of an obligate complex, PDB ID 1b8j, are depicted in Figure 4.1. The yellow and purple colors indicate atoms that are under the specific cutoff distance and act as interface atoms of chains A and B, respectively. It is clear that a large interface area is taken into account due to the long-range nature of electrostatic interactions.

As in [36], 18 different atom types and 20 different amino acid types were taken into account to calculate the features for prediction. Since the order of the interacting atoms and amino acid pairs is not important, we generated feature vectors for atom type features containing 171 ($\binom{18}{2}C$+18) values. Similarly, for amino acid type features, the length of the

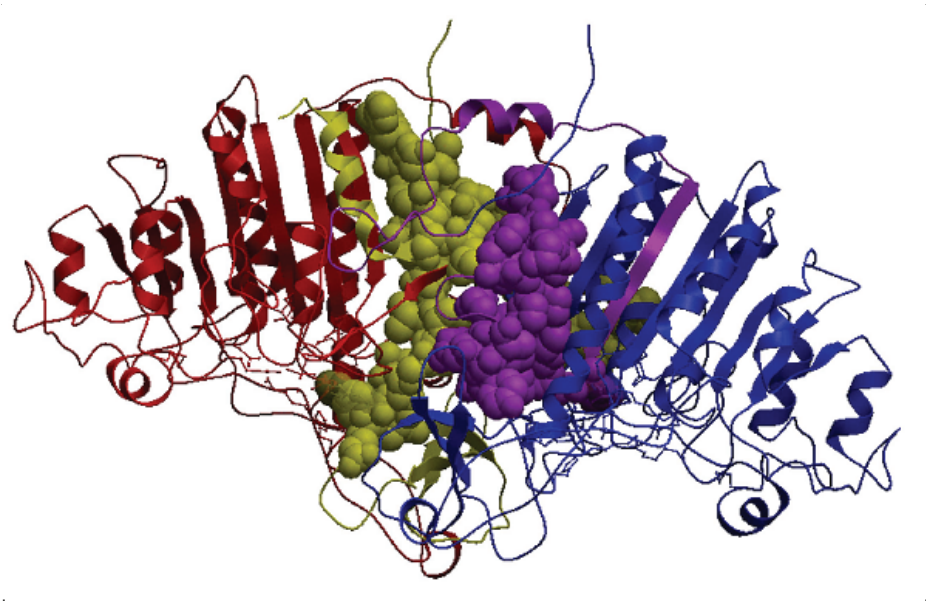


Figure 4.1: Quaternary structure of an obligate complex, PDB-ID *1b8j*, visualized with ICM Browser, along with its interacting chains A and B. Positive and negative charges are represented in red and blue respectively. Interface atoms of the interacting chains are represented in yellow and purple, respectively.

feature vector 210 ($\binom{20}{2}C$+20). Each feature contains the cumulative sum of electrostatic energy values for all pairs of atoms or amino acids of the same type. More details about the computation of electrostatic energy values for atom and amino acid type features are described in [21].

For the ZH and MW datasets, the names of the generated subsets of features for prediction using different feature types (interacting atoms or amino acids) are listed in Table 4.1.

4.2.3 Prediction methods

After finding the features of the complexes of the MW and ZH datasets, a prediction method should be applied to them. In this paper, the prediction is performed via several commonly

Table 4.1: Description of datasets used in this study

Authors	Reference	Atom type	Amino acid type
Mintseris *et al.*	[5]	MW-AT	MW-AA
Zhu *et al.*	[6]	ZH-AT	ZH-AA

used classification methods, including LDR, SVM, NB and k-NN. More details regarding the applied prediction methods are discussed below.

Linear Dimensionality Reduction

The main goal of LDR is to use linear combinations of the original features to generate new features in a lower dimensional space in which classification is, hopefully, more efficient than in the original space. There are different supervised LDR methods, and in this study, the following are considered [41]:

1. Fisher's discriminant analysis (FDA): FDA is a homoscedastic criterion that maximizes the Mahalanobis distance between the means assuming that the covariance matrices are equal.

2. Heteroscedastic discriminant analysis (HDA): HDA is a criterion that starts from the Chernoff distance in original space and takes correlations between random variables to project the data onto a lower dimensional space.

3. Chernoff discriminant analysis (CDA): CDA is a heteroscedastic criterion and aims to maximize the Chernoff distance between random vectors in the transformed space.

LDR is followed by a Bayesian classifier (linear or quadratic). More details about these LDR methods and the corresponding classification tasks can be found in [41].

Support Vector Machine

SVMs are well known machine learning techniques used for classification, regression and other tasks. The main goal of the SVM is to find a hyperplane that classifies all the feature vectors into two regions. In most cases, the separating hyperplane is not unique, and hence the SVM chooses the hyperplane that leaves the maximum margin from that hyperplane to the support vectors. Since most classification problems are not linearly separable, using a linear classifier is inefficient. Thus, in order to achieve a more efficient classification, using kernels to map the data onto a higher dimensional space can be useful. There are a number of kernels that can be used in SVM models such as polynomial, radial basis function (RBF) and sigmoid. The effectiveness of the SVM depends on the selection of the kernel, the selection parameters and the soft margin [42]. In addition, sequential minimal optimization (SMO), is a fast learning algorithm that has been widely applied to the training phase of a SVM classifier to solve the underlying optimization problem. In this study, the SMO module of the Waikato Environment for Knowledge Analysis (WEKA) with a polynomial kernel, default parameter settings and 10-fold cross validation is used for performing classification via the SVM [43].

k-Nearest Neighbor

k-NN is one of the simplest classification methods in which the class of each test sample can be easily found by voting on the class labels of its neighbors. To achieve this, after computing and sorting the distances between the test sample and each training sample, the most frequent class label in the first k train samples (nearest neighbors) is assigned to the class of the test sample. Determining the appropriate number of neighbors is one of the challenges of this method. In this study, the IBK module of WEKA with Euclidean distance,

default parameter settings, and 10-fold cross validation is used for k-NN classification [43].

Naive Bayes

One of the simplest probabilistic classifiers is NB. Assuming independence of the features, the class of each test samples can be found by applying Bayes' theorem. The basic mechanism of NB is rather simple. The reader is referred to [26] for more details. In this study, the NaiveBayes module of WEKA with default parameters and 10-fold cross validation is used [43].

4.2.4 Feature selection methods

Feature selection is the process of selecting the best subset of relevant features that represents the whole dataset efficiently and removing redundant and/or irrelevant ones. Applying feature selection before running a classifier is useful in reducing the dimensionality of the data and, thus, reducing the prediction time, while improving the prediction performance by eliminating irrelevant, redundant and noisy features. There are two different ways of doing feature selection: wrapper methods and filter methods [44]. In this study filter-based methods are used in which the quality of the selected features are scored and ranked independently of the classification algorithm and by using some criteria based on their relevance. The following filter-based feature selection methods are used in this study.

Minimum Redundancy Maximum Relevance

One of the most widely-used feature selection methods based on mutual information is mRMR [45, 46]. In this method, the features are selected and scored based on their relevance and redundancy among other features. A feature with minimum redundancy and

maximum relevance and with respect to the class concept is assigned a high score. After assigning a significance score to each feature, a ranking list of all features is generated. In this study, the online mRMR tool [47] with default parameters is used to obtain a complete list of all scored features by mRMR.

Information Gain

Information gain (IG) is based on the concept of entropy [44]. The IG value of a feature X with respect to class attribute Y is calculated as follows:

$$IG(Y,X) = H(Y) - H(Y|X). \tag{4.2}$$

Here, $H(Y)$ is the entropy of class Y and $H(Y|X)$ is the conditional entropy of Y given X, which are calculated by means of the following formulas:

$$H(Y) = -\sum_{y \in Y} p(y) \log_2(p(y)), \tag{4.3}$$

and

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2(p(y|x)), \tag{4.4}$$

where $p(y)$ is the marginal probability density function for random variable Y and $p(y|x)$ is the conditional probability of Y given X. In this study, the InfoGainAttributeEval module of WEKA is used for feature ranking based on the score of features by measuring the information gain with respect to the class.

Gain Ratio

GR attribute evaluation is a well-known feature selection method that is based on the concept of IG and entropy [44]. The GR value of a feature X with respect to class attribute Y is calculated as follows:

$$GR(Y,X) = \frac{(H(Y) - H(Y|X))}{H(X)} = \frac{IG(Y,X)}{H(X)}, \tag{4.5}$$

where $H(Y)$, the entropy of class Y, and $H(Y|X)$, the conditional entropy of Y given X, are calculated using Eqs. (4.3) and (4.4) respectively. A value of GR = 1 indicates that feature X is highly relevant and one of the best features to predict class Y, while GR = 0 means that feature X is not relevant at all. In this study, the GainRatioAttributeEval module of WEKA is used for feature ranking based on the relevance of each feature by measuring its gain ratio with respect to the class.

Chi Square

Feature selection via the Chi square test is another, very commonly used method [44]. This method evaluates the relevance of a feature with respect to a class by computing the value of the Chi square statistic. In this study, the ChiSquaredAttributeEval module of WEKA is used to obtain the scored feature vector.

4.3 Results and discussion

To test our proposed method and perform an in-depth analysis of the strength of electrostatic energy as the prediction property, four different classification methods including SMO, k-NN, LDR and NB and also four different feature selection methods including IG, GR, Chi2

and mRMR have been used. The performances of the prediction methods are compared in terms of their accuracies, which are computed as follows: $acc = (TP + TN)/N$, where TP and TN are the total numbers of true positive (obligate) and true negative (non-obligate) counters over the 10-fold cross-validation procedure, respectively, and N is the total number of complexes in the dataset.

4.3.1 Analysis of prediction properties

In previous works [18–20], it has been shown that desolvation energy is very efficient for prediction of obligate and non-obligate complexes in comparison with solvent accessible and interface area properties. However, in our recent study of [21] and in this work, it has been shown that employing electrostatic energy deliver impressive prediction accuracy.

To validate our previous results and compare the strength of electrostatic and desolvation energies as properties for prediction, SMO, k-NN, NB and LDR have been applied for prediction on these two types of features. For the LDR schemes, six different classifiers were implemented and evaluated, namely the combinations of FDA, HDA and CDA with quadratic and linear classifiers; the maximum average classification accuracy for each classifier is reported for each dataset. For SVM, k-NN and NB, the classification modules of WEKA have been used with default parameters in a 10-fold cross-validation process. The distance cutoffs between atom pairs of interacting chains are 9 Å and 7 Å for electrostatic and desolvation energies as properties respectively.

The prediction results of SMO, NB, k-NN and LDR for atom and amino acid type properties for the ZH and MW datasets with desolvation and electrostatic energies as properties are shown in Table 4.2. For ZH-AT, the best accuracy by using electrostatic energy is 96.95% with SMO, while by using desolvation energy, accuracy is much lower, 74.34%,

Table 4.2: Comparison of accuracies for electrostatic and desolvation energies as properties

Dataset	SMO		NB		*k*-NN		LDR	
	DE	EE	DE	EE	DE	EE	DE	EE
ZH-AT	72.52%	**96.95%**	72.52%	94.65%	64.12%	95.42%	74.34%	95.42%
ZH-AA	66.42%	**97.70%**	75.91%	92.37%	54.74%	96.18%	72.13%	93.89%
MW-AT	77.30%	96.04%	77.96%	89.44%	74.43%	95.71%	78.95%	**96.30%**
MW-AA	73.93%	**98.68%**	72.39%	90.10%	57.36%	98.68%	75.15%	92.08%

with LDR. Also, for ZH-AA, using electrostatic energy leads to 97.70% accuracy with SMO, being more efficient than using desolvation energy with NB, 75.91%. Similarly, the best accuracies for MW-AT, 96.30%, and MW-AA, 98.68%, are obtained using electrostatic energy in comparison with accuracies of 78.95% and 75.15% for both MW-AT and MW-AA respectively by using desolvation energy.

Generally, from the table, it can be concluded that electrostatic energy yields much more efficient prediction than desolvation energy, on the basis of the experimental results shown here using different classification methods. In addition, for most subsets of features, SMO performs better than *k*-NN, NB and LDR, for both desolvation and electrostatic energies.

Figure 4.2 shows the receiver operating characteristic (ROC) curves for the MW-AT and ZH-AT datasets using electrostatic and desolvation energies as properties for prediction by LDR. These ROC curves are plotted based on the true positive rate (TPR), aka "sensitivity", vs. the false positive rate (FPR), known as "1 - specificity", at various threshold settings. For both datasets, ZH-AT and MW-AT, the prediction performances of LDR using electrostatic energy are clearly much better than using desolvation energy for prediction. In addition, the area under the curve (AUC) for each of the above ROC curves was computed. The AUC for ZH-AT using electrostatic energy is 0.90 while using desolvation energy is 0.73. Similarly, the AUC for MW-AT using electrostatic energy is 0.91 while using desolvation energy is 0.72. By comparing the AUC values, it can be concluded that electrostatic

Table 4.3: Prediction accuracies using desolvation energy and different distance cutoffs

Dataset	Inter-atom distance cutoffs								
	5Å	6Å	7Å	8Å	9Å	10Å	11Å	12Å	13Å
ZH-AT	71.75%	74.04%	72.52%	71.75%	70.99%	69.46%	68.70%	67.93%	67.93%
MW-AT	75.99%	76.32%	77.96%	76.32%	75.99%	73.02%	73.02%	72.36%	71.38%

energy clearly shows much better prediction accuracy than desolvation energy.

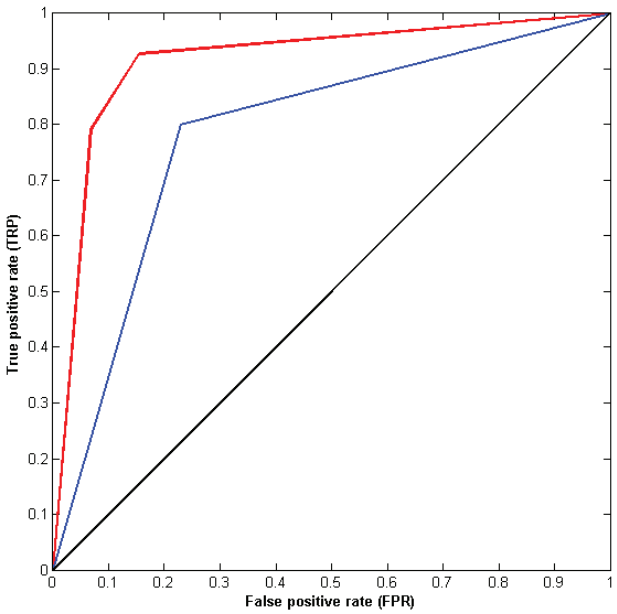
4.3.2 Analysis of distance cutoffs

In order to obtain a better insight into the classification results by using desolvation and electrostatic energies as properties, different experiments were performed by varying the distance cutoff between atom pairs of interacting chains.

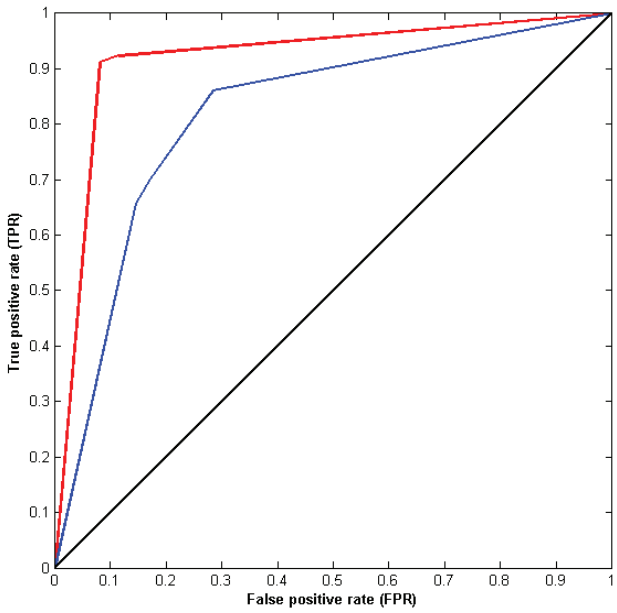
Table 4.3 shows the prediction results for all cutoff values ranging from 5Å to 13Å for atom-type datasets, namely ZH-AT and MW-AT. For this analysis, desolvation energy values are used as the prediction properties and the NB classifier is applied for classification. The best distance cutoff for the ZH-AT dataset is 6Å, achieving 74.04% prediction accuracy, while for MW-AT the highest prediction accuracy, 77.96%, is achieved for 7Å.

In Figure 4.3, the performances of NB for atom type features for the MW and ZH datasets, when using desolvation energy, are plotted against the interaction distances. From the plots, it is observable that for both datasets, the best prediction accuracies are obtained for distance cutoffs between 5Å and 8Å. Moreover, for both datasets the performances decrease gradually by increasing the distance cutoffs. These results demonstrate that the best distance cutoffs for prediction by using desolvation energy is less than 8Å.

Similarly, Table 4.4 shows the prediction results for the ZH-AT and MW-AT datasets for distance cutoffs from 7Å to 13Å. Here, electrostatic energy is used as the prediction property and NB for classification. For the ZH-AT dataset, the best accuracy, 96.95%,



(a)MW-AT dataset



(b)ZH-AT dataset

Figure 4.2: ROC curves for the (a) MW-AT and (b) ZH-AT datasets using desolvation energy (blue line) and electrostatic energy (red line) as properties for prediction by using LDR.

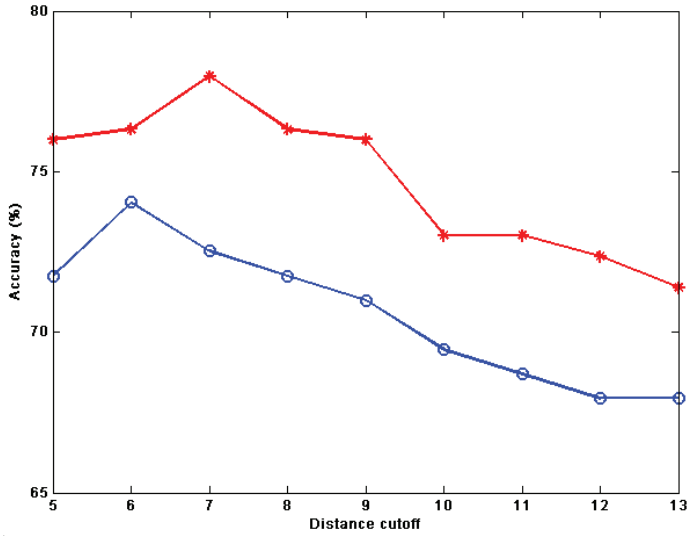


Figure 4.3: Prediction accuracy for NB on MW-AT (red line) and ZH-AT (blue line) using desolvation energy as the prediction property and different distance cutoffs ranging from 5Å to 13Å.

is obtained for a distance cutoff of 12Å, while for the MW-AT dataset the best accuracy, 90.42%, is achieved for a distance cutoff of 11Å.

The classification accuracies for the atom type datasets, MW-AT and ZH-AT, when using electrostatic energy, are plotted in Figure 4.4. The x-axis shows the distance cutoff between atom pairs of interacting chains (ranging from 7Å to 13Å) while the y-axis shows the prediction accuracy. For ZH-AT, the best accuracies are achieved for distance cutoffs in the range 10Å to 12Å, and these accuracies are all close to 96%. By increasing the distance cutoff to 13Å, the accuracy decreases rather quickly. Also, for MW-AT, the prediction accuracies in the range 9Å to 12Å are almost the same, around 90%. As in the ZH-AT, the performance decreases when the distance cutoff is increased to 13Å.

As a general remark, it can be concluded that the best distance cutoffs for prediction of obligate and non-obligate complexes using electrostatic energy range from 9Å to 12Å, while by using desolvation energy the best distance cutoffs range from 5Å to 7Å. These

Table 4.4: Prediction accuracies for electrostatic energy and different distance cutoffs

	Inter-atom distance cutoffs						
Dataset	7Å	8Å	9Å	10Å	11Å	12Å	13Å
ZH-AT	94.65%	94.65%	94.65%	96.15%	96.18%	96.95%	90%
MW-AT	84.44%	84.16%	89.44%	89.44%	90.42%	89.85%	82.83%

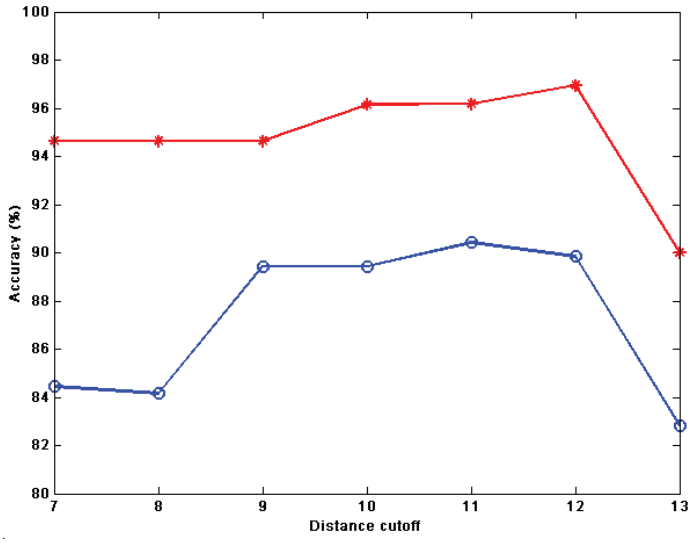


Figure 4.4: Prediction accuracy for NB on MW-AT (red line) and ZH-AT (blue line) using electrostatic energy as the prediction property and different distance cutoffs ranging from 7Å to 13Å.

distance cutoffs for desolvation energy are reasonable and are in agreement with all previous studies [5, 6, 36]. In most studies, a distance cutoff of 6Å is typically used to determine whether or not two atoms from different chains interact with each other. Moreover, in [20, 35, 36], a function g is used to compute the distance between two atoms. These approaches consider a smooth function for inter-atom distances between 5Å and 7Å, while g evaluates to 0 if the distance is greater than 7Å. On the other hand, electrostatic energy is considered to be long-range [21, 48], extending inter-atom interactions up to a 10Å distance or more, and hence covering a much broader and deeper area of the interface. In other words, this suggests that using electrostatic energy with a long distance cutoff, the atoms in the surface and some atoms buried under the surface may participate in the prediction that led to outstanding classification performance. This is a topic of interest for further studies.

4.3.3 Analysis of feature selection

Determining the minimum number of features while keeping, or even improving, classification performance is the main challenge in all feature selection methods. To demonstrate this, the accuracies of LDR for atom type features of the MW and ZH datasets are plotted against the number of selected features in Figure 4.5. The order of the selected features for prediction is based on the order of features scored by GR. The best number of features for MW-AT is 20, achieving 99.67% while for ZH-AT, 15 features are found with 97.69% accuracy. From the plot, it can be concluded that (a) a few features are good descriptors for prediction of obligate and non-obligate complexes; (b) the best number of features is different from one dataset or subset of features to another; (c) prediction accuracy for the MW-AT dataset is much higher, achieving almost perfect prediction.

To compare the performance of feature selection methods and their effects on the pre-

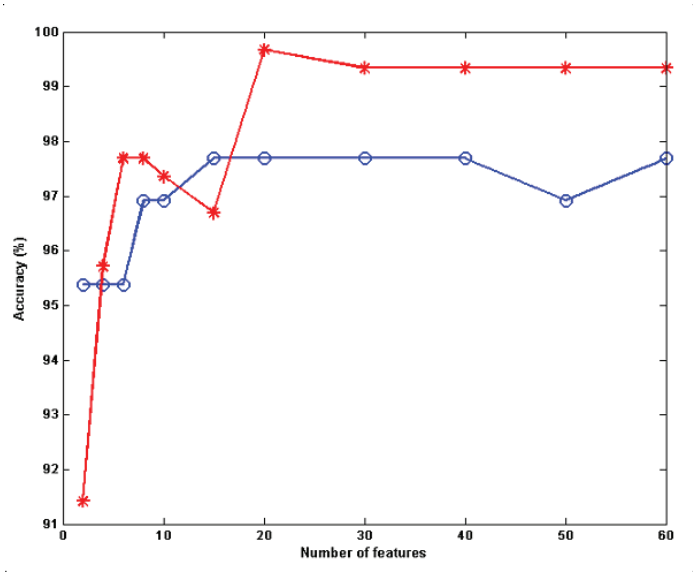


Figure 4.5: Prediction accuracy for LDR on MW-AT (red line) and ZH-AT (blue line) using electrostatic energy plotted against the number of features selected by GR.

diction algorithms from a different perspective, the features of all datasets were scored and ranked by GR, IG, Chi2 and mRMR, separately. Then, LDR methods were applied for prediction by selecting a subset of the top-ranked features. In this experiment, the number of selected features was less than 10, starting from two features and then adding two more features at each subsequent step. The results of the LDR classifier with atom and amino acid type features with and without feature selection are depicted in Table 4.5. For all datasets, except MW-AA, the predictions show better performance by using feature selection methods. The most significant increase in prediction performance is for ZH-AT, for which by using only two of the top-ranked features scored by Chi2, yields 97.69% accuracy, which is much higher than using no feature selection at all (95.42%). While the most notable decrease in prediction accuracy between with and without feature selection is approximately 3%, which is observed in MW-AA by applying GR, this decrease can be acceptable considering that only four out of the 210 original features are used for prediction. This also

Table 4.5: Prediction accuracies for electrostatic energy and different feature selection methods

FS method	ZH-AA		ZH-AT		MW-AA		MW-AT	
	n	accuracy	n	accuracy	n	accuracy	n	accuracy
No FS	210	93.89%	171	95.42%	210	92.08%	171	96.30%
Chi2	8	97.69%	2	97.69%	10	91.09%	6	97.69%
GR	4	96.92%	8	96.92%	4	86.80%	6	97.69%
IG	8	97.69%	2	97.69%	8	88.78%	10	96.37%
mRMR	10	96.15%	10	97.69%	10	90.94%	10	96.10%

implies savings in the required classification time and space resources.

In general, it can be concluded that a few pairs of atoms/amino acids are appropriate for prediction. Also, feature selection increases the performance of classification models by eliminating redundant, irrelevant and noisy features and selecting the more discriminative features. Moreover, by comparing the performance of the applied feature selection methods, Chi2 is the best method for ranking features. In contrast, mRMR is the worst ranking method because it used more features and achieved lower performance for all datasets.

4.3.4 Visual analysis

To show the effect of using electrostatic energy for prediction of PPI types from a different perspective, a visual analysis is presented. In this analysis, an obligate complex, PDB ID *2min*, and a non-obligate complex, PDB ID *1a2k*, both from the MW dataset are considered. For these protein complexes the solvent accessible surfaces by electrostatic potential are generated with the help of Jmol embedded in APBS. In the plots, positive electrostatic potentials are shown in blue, while negative electrostatic potentials are shown in red.

The electrostatic potentials of the sub-units corresponding to chains A and B of *2min* are shown in Figures 4.6(a) and (b), respectively. The whole complex (chains A and B together)

is shown in Figure 4.6(c). By observing Figure 4.6(c), it is clear that the interaction between chains A and B of *2min* takes place at regions of the two chains (highlighted in yellow) that have different electrostatic potentials; the highlighted region of chain A has positive charge (Figure 4.6(a)), while for chain B has negative charge (Figure 4.6(b)). It other words, the positive and negative potentials on the interface areas of chains A and B cause them to interact with each other.

Similarly, Figure 4.7 shows a non-obligate complex, PDB ID *1a2k AB:C*, along with the electrostatic potential for three different cases: chains AB as a sub-unit (Figure 4.7(a)), chain C as a sub-unit (Figure 4.7(b)) and the whole complex including chains AB and chain C (Figure 4.7(c)). From the plots, it is clear that the region highlighted in yellow in Figure 4.7(a) shows negative electrostatic potential (shown in red), while in Figure 4.7(b), the highlighted yellow region shows positive electrostatic potential (shown in blue). The interaction between the two chains takes place at these regions is shown in Figure 4.7(c).

Similarly, the positive and negative potentials on the interface areas of chains AB and chain C yield very high affinity and cause them to interact with each other. However, the interface area of complex *1a2k*, which is non-obligate, is smaller than the interface area of complex *2min*, which is obligate. Electrostatic energy is a very good property in the sense that it captures the size of the interface area and the complementarity of the sub-units participating in the interaction. Observing Figure 7(b), it is clear that the concavity of the sub-unit corresponding to chain B will match very well the salient part on the right of the sub-unit of chain A. These features are well captured by electrostatic energy and this is, indeed, the main aspect that we exploit to predict the stability of protein complexes, which is corroborated in the experimental results.

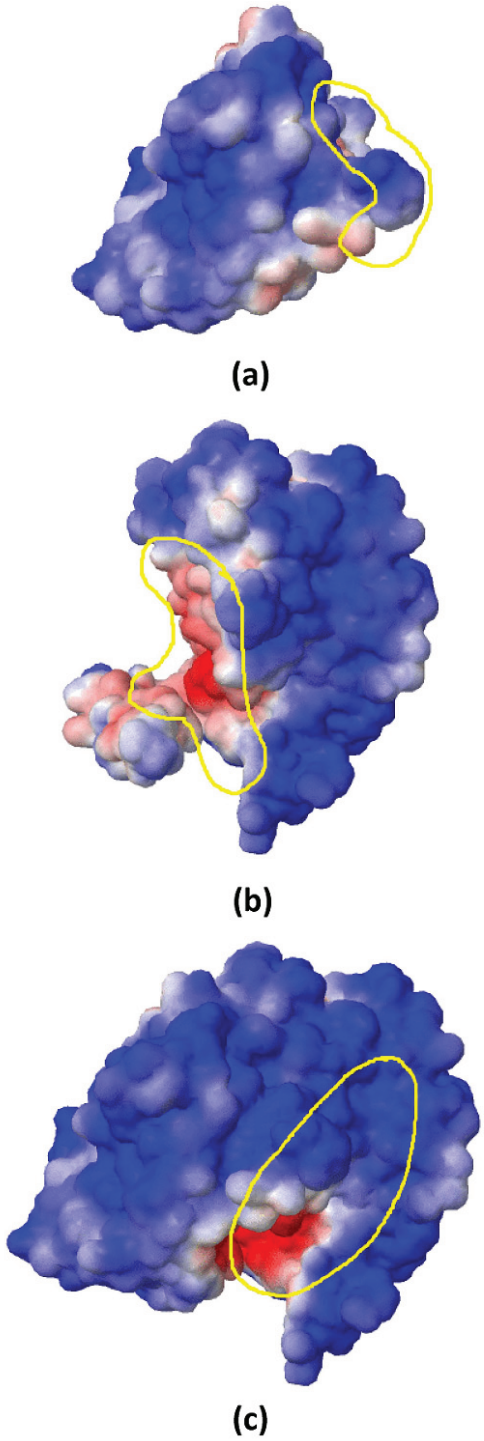


Figure 4.6: Plot of solvent accessible surface by electrostatic potential of an obligate complex, PDB-ID *2min*, before and after the interaction takes place; (a) Electrostatic potential of chain A of *2min*, (b) Electrostatic potential of chain B of *2min*, (c) Electrostatic potential of chains A and B of *2min*.

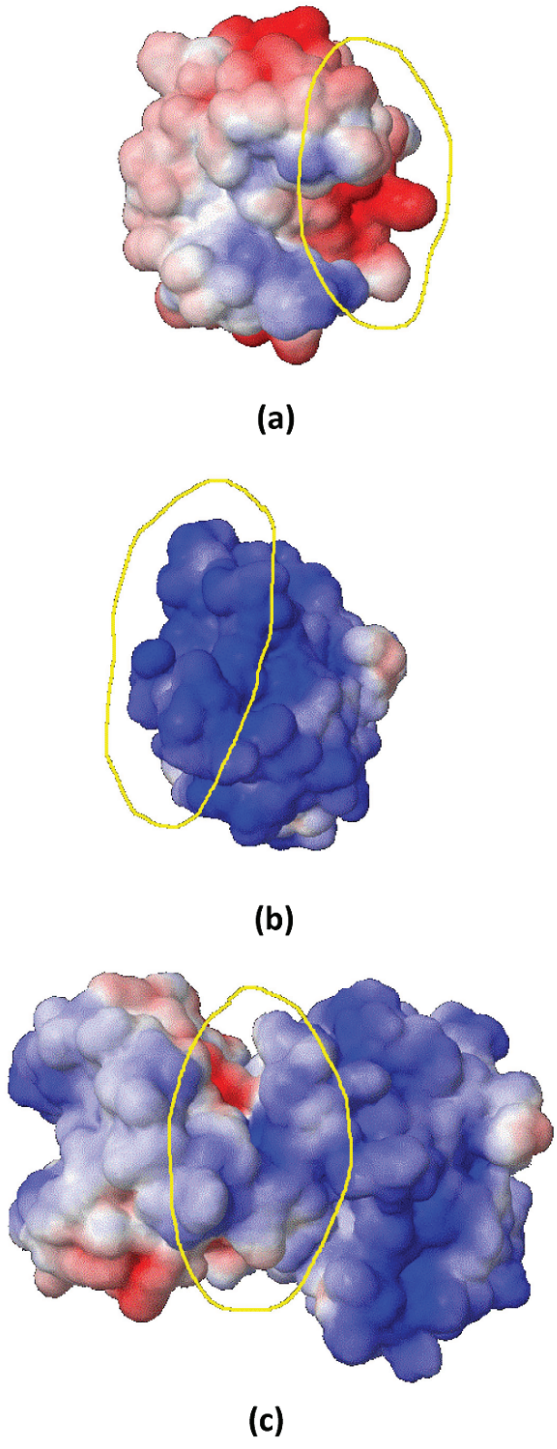


Figure 4.7: Plot of solvent accessible surface area by electrostatic potential of a non-obligate complex, PDB-ID *1a2k*, before and after the interaction takes place. (a) Electrostatic potential of chains AB of *1a2k*, (b) Electrostatic potential of chain C of *1a2k*, (c) Electrostatic potential of chains AB and C of *1a2k*.

4.4 Conclusions

The proposed prediction model works exceptionally well for distinguishing protein interaction types. Our prediction approach uses electrostatic energy values for pairs of atoms or amino acids present in the interfaces of obligate and non-obligate complexes. The classification is performed via various classification techniques including LDR, SVM, *k*-NN and NB.

We observe that electrostatic energy values with distance cutoffs in the range 9Å to 12Å turn out to be the best ones for prediction of interaction types on the basis of our experimental results. The reason for why electrostatic energy yields better prediction results is because electrostatic interactions are long-range. Thus, by using electrostatic energy with a large distance cutoff, not only the atoms in the surface but also some atoms which are buried under the surface may participate in the interaction, and this leads to excellent prediction results. Therefore, among various types of molecular interactions, electrostatic interactions play a special role. The proposed features then exploit the high affinity of proteins to interact with each other (in terms of negative and positive potentials). Furthermore, applying several feature selection algorithms on the MW and ZH datasets demonstrates that removing irrelevant and noisy pairs of atom type/amino acid type features and selecting the most relevant pairs improve the prediction results.

From this study, various open questions remain to be answered. One of these is to investigate domains and motifs present in the interface in order to achieve a better insight on proteins, their interactions, and function. Another problem that deserves attention is to investigate the role of buried atoms and their influence in obligate interactions. This study could consider atoms that are 10 Å (or more) apart from each other, but one of these atoms

may not be on the surface of the protein.

Bibliography

[1] A. Mendelsohn and R. Brent, "Protein interaction methods-toward an endgame." *Science*, vol. 284, no. 5422, pp. 1948–1950, 1999.

[2] S. Park, J. Reyes, D. Gilbert, J. Kim, and S. Kim, "Prediction of protein-protein interaction types using association rule based classification," *BMC Bioinformatics*, vol. 10, no. 1, p. 36, 2009.

[3] Q. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, and et al., "Structure-based prediction of protein-protein interactions on a genome-wide scale," *Nature*, vol. 490, no. 7421, pp. 556–560, 2012.

[4] J. Qiu, X. Sun, S. Suo, S. Shi, S. Huang, P. Liang, and L. Zhang, "Predicting homo-oligomers and hetero-oligomers by pseudo-amino acid composition: an approach from discrete wavelet transformation." *Biochimie*, vol. 93, no. 7, pp. 1132–1138, 2011.

[5] J. Mintseris and Z. Weng, "Structure, function, and evolution of transient and obligate protein-protein interactions," *Proc Natl Acad Sci*, vol. 102, no. 31, pp. 10 930–10 935, 2005.

[6] H. Zhu, F. Domingues, I. Sommer, and T. Lengauer, "NOXclass: Prediction of protein-protein interaction types," *BMC Bioinformatics*, vol. 7, no. 27, 2006, doi:10.1186/1471-2105-7-27.

[7] L. LoConte, C. Chothia, and J. Janin, "The atomic structure of protein-protein recognition sites," *J Mol Biol*, vol. 285, no. 5, pp. 2177–2198, 1999.

[8] J. Young, "A role for surface hydrophobicity in protein protein recognition," *Protein Sci*, vol. 3, pp. 717–729, 1994.

[9] O. K. A. Zen, C. Micheletti and R. Nussinov, "Comparing interfacial dynamics in protein-protein complexes: an elastic network approach," *BMC Structural Biology*, vol. 10, no. 26, 2010, doi: 10.1186/1472-6807-10-26.

[10] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc. Natl Acad. Sci, USA*, vol. 93, no. 1, pp. 13–20, 1996.

[11] M. C. Lawrence and P. M. Colman, "Shape complementarity at protein/protein interfaces," *J. Mol Biol*, vol. 234, no. 4, pp. 946–950, 1993.

[12] P. Chakrabarti and J. Janin, "Dissecting protein-protein recognition sites," *Proteins*, vol. 47, no. 3, pp. 334–343, 2002.

[13] D. Xu, C. Tsai, and R. Nussinov, "Hydrogen bonds and salt bridges accross protein-protein interfaces," *Protein Eng*, vol. 10, no. 9, pp. 999–1012, 1997.

[14] B. Ma, T. Elkayam, H. Wolfson, and R.Nussinov, "Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces," *Proc Natl Acad Sci, USA*, vol. 100, no. 10, pp. 5772–5777, 2003.

[15] H. Shanahan and J. Thornton, "Amino acid architecture and the distribution of polar atoms on the surfaces of proteins," *Biopolymers*, vol. 78, no. 6, pp. 318–328, 2005.

[16] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal, "Residue frequencies and pairing preferences at protein-protein interfaces," *Proteins*, vol. 43, no. 2, pp. 89–102, 2001.

[17] J. Mintseris and Z. Weng, "Atomic contact vectors in protein-protein recognition," *PROTEINS: Structure, Function and Genetics*, vol. 53, pp. 629–639, 2003.

[18] L. Rueda, S. Banerjee, M. Aziz, and M. Raza, "Protein-protein interaction prediction using desolvation energies and interface properties," in *Bioinformatics and Biomedicine (BIBM)*, 2010, pp. 17–22.

[19] L. Rueda, C. Garate, Banerjee, and Md. Aziz, "Biological protein-protein interaction prediction using binding free energies and linear dimensionality reduction." *Proceedings of the 5th. IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2010)*, pp. 383–394, 2010.

[20] Md. Aziz, M. Maleki, L. Rueda, M. Raza, and S. Banerjee, "Prediction of biological protein-protein interactions using atom-type and amino acid properties," *Proteomics*, vol. 11, pp. 3802–3810, 2011.

[21] G. Vasudev and L. Rueda, "A model to predict and analyze protein-protein interaction types using electrostatic energies," in *5th IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012)*, 2012, pp. 543–547.

[22] A. Kessel and N. Ben-Tal, *Introduction to Proteins: Structure, Function, and Motion*. CRC Press, 2010.

[23] S. De, O. Krishnadev, N. Srinivasan, and N. Rekha, "Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different." *BMC Structural Biology*, vol. 5, no. 15, 2005.

[24] J. V. Eichborn, S. Gunther, and R. Preissner, "Structural features and evolution of protein-protein interactions." *Intenational Conference of Genome Informatics.*, vol. 22, pp. 1–10, 2010.

[25] M. Maleki, M. Aziz, and L. Rueda, "Analysis of relevant physicochemical properties in obligate and non-obligate protein-protein interactions," in *IEEE International Conference in Bioinformatics and Biomedicine Workshops (BIBMW), 2011*, 2011, pp. 345–351.

[26] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Elsevier Academic Press, 2006.

[27] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, , and Y. Li, "Prediction of Lysine ubiquitination with MRMR feature selection and analysis." *Amino Acids*, 2011.

[28] S. Niu, T. Huang, K. Feng, Y. Cai, and Y. Li, "Prediction of Tyrosine sulfation with MRMR feature selection and analysis." *J Proteome Res*, vol. 9, no. 12, pp. 6490–6497, 2010.

[29] L. Liu, Y. Cai, W. Lu, C. Peng, and B. Niub, "Prediction of protein-protein interactions based on PseAA composition and hybrid feature selection," *Biochemical and Biophysical Research Communications*, vol. 380, no. 2, pp. 318–322, 2009.

[30] Y. Yuan, x. Shi, X. Li, W. Lu, Y. Cai, L. Gu, L. Liu, M. Li, X. Kong, and M. Xing, "Prediction of interactiveness of proteins and nucleic acids based on feature selections." *Mol Divers.*, vol. 14, no. 4, pp. 627–33, 2009.

[31] P. Mundra and J. Rajapakse, "SVM-RFE with MRMR filter for gene selection." *IEEE Transactions on Nanobioscience*, vol. 9, no. 1, pp. 31–37, 2010.

[32] Y. Zhao and Z. Yand, "Improving MSVM-RFE for multiclass gene selection." *The Fourth International Conference on Computational Systems Biology (ISB2010)*, 2010.

[33] Y. Lee, C. Chang, and C. Chao, "Incremental forward feature selection with application to microarray gene expression data," *biopharmaceutical statistics*, vol. 18, no. 5, pp. 827–840, 2008.

[34] Q. Liu and J. Li, "Propensity vectors of low-ASA residue pairs in the distinction of protein interactions," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 3, pp. 589–602, 2010.

[35] C. Camacho and C. Zhang, "FastContact: rapid estimate of contact and binding free energies," *Bioinformatics*, vol. 21, no. 10, pp. 2534–2536, 2005.

[36] C. Zhang, G. Vasmatzis, J. L.Cornette, and C. DeLisi, "Determination of atomic desolvation energies from the structures of crystallized proteins," *J. Mol. Biol.*, vol. 267, pp. 707–726, 1997.

[37] N. Baker, "Continuum models for biomolecular solvation." 2008, pacific Northwest National Laboratory.

[38] T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe, and N. A. Baker, "PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations." *Nucleic Acids Research*, vol. 35, pp. 522–525, 2007.

[39] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. Mccammon, "Electrostatics of nanosystems: Application to microtubules and the ribosome," *Proceedings of the National Academy of Sciences*, vol. 98, no. 18, pp. 10 037–10 041, 2001.

[40] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[41] L. Rueda and M. Herrera, "Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space," *Pattern Recognition*, vol. 41, no. 10, pp. 3138–3152, 2008.

[42] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York, NY: John Wiley and Sons, Inc., 2000.

[43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[44] J. Novakovic, P. Strbac, and D. Bulatovic, "Toward optimal feature selection using ranking methods and classification algorithms." *Yugoslav J. of Operations Research*, vol. 21, no. 1, pp. 119–135, 2011.

[45] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data." *Journal of Bioinformatics and Computational Biology.*, vol. 3, no. 2, pp. 185–205, 2005.

[46] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[47] "minimum redundancy maximum relevance feature selection (mrmr)." [Online]. Available: http://penglab.janelia.org/proj/mRMR/

[48] E. Fadrná, K. Hladecková, and J. Koca, "Long-range electrostatic interactions in molecular dynamics: An endothelin-1 case study." *Journal of Biomolecular Structure and Dynamics*, vol. 23, no. 2, pp. 151–162, 2005.

PART 2

DOMAIN-BASED FEATURES-CATH

Chapter 5: M. Maleki, M. Hall, L. Rueda, "Using Desolvation Energies of Structural Domains to Predict Stability of Protein Complexes," Journal of Network Modeling Analysis in Health Informatics and Bioinformatics (NetMAHIB), vol. 2, no. 4, pp. 267275, Dec. 2013.

Chapter 6: M. Maleki, M. Hall, L. Rueda, "Using Structural Domain to Predict Obligate and Non-obligate Protein-protein Interactions," in 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB2012), California, USA, May 2012.

Chapter 5

Using Desolvation Energies of Structural Domains to Predict Stability of Protein Complexes

5.1 Introduction

Domains can be considered as the minimal and fundamental units of proteins, which have a clear biological role and act as basic functional units within cells [1, 2]. Recent studies focus on employing domain knowledge to predict protein-protein interactions [3–9]. This is based on claims that only a few highly conserved residues are crucial for protein-protein interactions [10, 11], and most domains and domain-domain interactions (DDIs) are evolutionarily conserved [12]. As a consequence, it has been observed that proteins interact if a domain in one protein interacts with a domain in the other protein [13, 14]. There are a number of domain family resources that can be applied for this purpose such as Pfam [15] and CATH – Class, Architecture, Topology and Homologous superfamily – databases [16].

On the other hand, an important problem surrounding PPIs is the identification and prediction of different types of complexes, which are characterized by properties such as similarities between subunits (homo/hetero-oligomers), number of subunits involved in the

interaction (dimers, trimers, etc.), duration of the interaction (transient vs. permanent), stability of the interaction (non-obligate vs. obligate), among others. We focus on the prediction of obligate and non-obligate complexes. It is important to be able to distinguish between obligate and non-obligate complexes, since non-obligate interactions are more difficult to study and understand due to their instability and short life, while obligate interactions are more stable [17].

Using the relevant features is very important for successful prediction. Features are the observed properties of each sample that is used for prediction. Some studies in PPI consider the analysis of a wide range of parameters for predicting obligate and non-obligate complexes, including analysis of solvent accessibility [18, 19], geometry [20], hydrophobicity [21, 22], sequence-based features [23], desolvation energy [3, 4, 24–26] and, more recently, electrostatic energies [27]. In this study, we use desolvation energies, which have been shown to be very efficient for PPI prediction [24, 25].

To study the behavior of obligate and non-obligate interactions using domain knowledge, in [10], interactions between residues were used for finding obligate and non-obligate residue contacts of PPIs. The study concluded that non-obligate interfaces occupy less than 2% of the area of the domain surfaces, while the area occupied by obligate interfaces is between 0–6%. In [11], the interface of 750 transient DDIs (interactions between domains that are part of different proteins) and 2,000 obligate DDIs were studied. The interactions between domains of one amino acid chain were analyzed to obtain a better understanding of molecular recognition and identify frequent amino acids in the interfaces and on the surfaces of protein complexes. Also, in [28], the domain information from protein complexes was used to predict four different types of interactions, including transient enzyme inhibitor/non enzyme inhibitor and permanent homo/hetero obligate complexes. In this

way, the physical interaction between proteins can be better analyzed in terms of interactions among their structural domains. In [29], a prediction model was proposed in which Pfam domains were used to predict obligate and non-obligate PPIs. The results demonstrated that desolvation energies are more efficient and powerful than interface area and composition properties for prediction. Moreover, a visual and numerical analysis of the DDIs present in these two types of complexes showed that different pairs of DDIs can be identified in obligate and non-obligate complexes, and highlighted that homo-DDIs are more likely to be present in obligate interactions.

In one of our recent works [3], a domain-based model to predict obligate and non-obligate PPIs was presented, in which structural domains from the CATH database were considered as the input features. That model used desolvation energies of amino acid pairs present in the interface of DDIs as features for prediction. The results show that DDIs at higher levels in the CATH hierarchy, especially those at the architecture and topology levels (levels 2 and 3), provide the best prediction performance. Whereas our previous efforts in [3] focused on the case in which all DDIs were taken from the same level of the CATH hierarchy, we have extended our approach in [4] to cover the more general case in which each domain can be represented at one of a number of possible levels. We restricted our efforts to levels 2 and 3 of the CATH hierarchy, which have been shown to be very efficient for prediction.

This work is an extension of the work presented in [4], by incorporating a wider range of classification techniques that include LDR, SVM-SMO, NB, and k-NN and also a numerical analysis focused on selecting relevant structural properties. The results on two pre-classified datasets from [19] and [30] confirm that using DDIs from different levels of the CATH hierarchy as prediction properties yields better performance than using DDIs of

individual levels to predict obligate and non-obligate PPIs based on the obtained prediction results using different classification methods. Furthermore, by grouping the DDI feature vectors of the second level of the CATH hierarchy based on their secondary structures, it is shown that most of the interactions are between domains that have mainly-beta structures. Also, the prediction results for each group of DDIs from level 2 demonstrate that DDIs related to *mainly-beta*, especially DDIs of mainly-beta with *alpha-beta* are the most discriminative properties for predicting obligate and non-obligate PPIs using SVM-SMO and k-NN.

5.2 Datasets and Prediction Properties

Two pre-classified datasets of obligate and non-obligate protein complexes were obtained from the studies of Zhu et al. [19], and Mintseris and Weng [30].The first dataset contains 75 permanent (obligate) and 62 non-obligate interactions, while the second dataset contains 115 obligate and 212 non-obligate interactions. These datasets were obtained from the literature and manually curated by the authors of [30] and [19] by removing inconsistent complex types and homologous protein sequences.

5.2.1 Desolvation Energy

In this study, desolvation energies are used as the prediction properties, which have been shown to be very efficient for the prediction of obligate and non-obligate complexes [24, 29]. Desolvation energy is defined as knowledge-based contact potential (accounting for hydrophobic interactions), self-energy change upon desolvation of charged and polar atom groups, and side-chain entropy loss. As in [31], the binding free energy ΔG_{bind} is defined

as follows:

$$\Delta G_{bind} = \Delta E_{elec} + \Delta G_{des}, \tag{5.1}$$

where ΔE_{elec} is the total electrostatic energy and ΔG_{des} is the total desolvation energy. For a protein, ΔG_{des} is defined as follows:

$$g(r)\Sigma\Sigma e_{ij}. \tag{5.2}$$

If we consider the interaction between the i^{th} atom of a ligand and the j^{th} atom of a receptor, then e_{ij} is the atomic contact potential (ACP) between them and $g(r)$ is a smooth function based on their distance [32]. For simplicity, we consider the smooth function to be linear. We also consider the criterion that for a successful interaction, atoms should be within 7 Å distance; between 5 and 7 Å, the value of $g(r)$ varies from 0 to 1 based on a smooth function. For atoms that are less than 5 $\mathring{A}$ apart, the value of $g(r)$ is 1 [31].

5.2.2 Domain-based Properties

We consider structural CATH domains [16] in this study. The CATH database is organized in a hierarchical fashion, which can be visualized as a tree with levels numbered from 1 to 8, hereafter referred to as L1 to L8 [16]. Domains at upper levels of the tree represent more general classes of structure than those at lower levels. For example, domains at level 1 represent mainly-alpha (*c1*), mainly-beta (*c2*), mixed alpha-beta (*c3*) and few secondary structures (*c4*), whereas those at level 2 represent more specific structures. As shown in Figure 5.1, roll, beta barrels and 2-layer sandwich are three different sample architectures of domains in class *c3*. Domains at level 3 are even more specific, and so on.

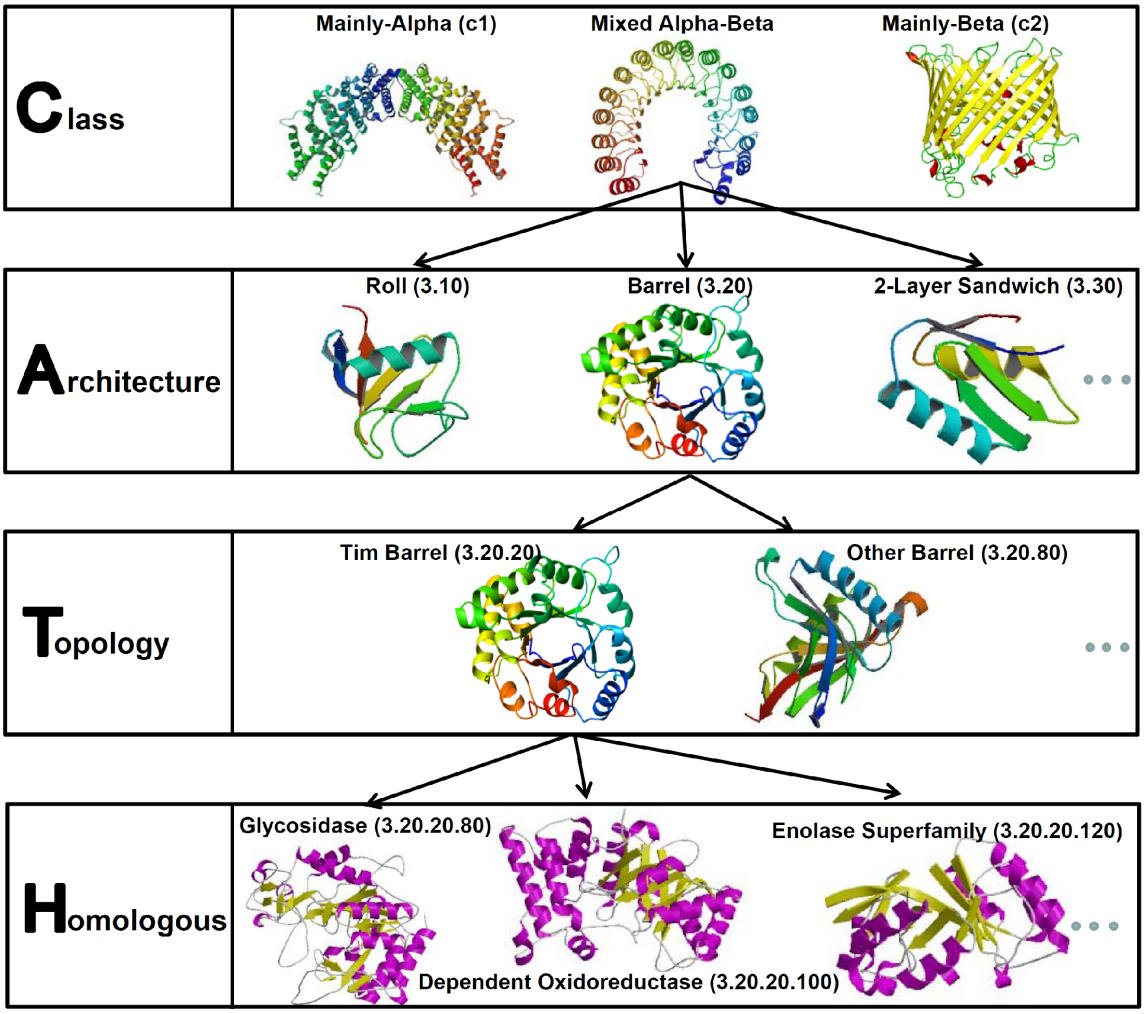


Figure 5.1: Four levels of the CATH hierarchy (Class, Architecture, Topology and Homologous superfamily).

To extract domain-based properties, we first collected the 3D structures of each complex in our datasets from the Protein Data Bank (PDB) [33]. Then, we collected the domain information for each complex from CATH database and added this information to each atom present in the chain. Complexes that did not have domain information in at least one of their subunits were discarded. We refer to these two new datasets as the MW and ZH datasets. The new MW dataset contains 100 permanent (obligate) and 161 non-obligate interactions, while the new ZH dataset contains 72 obligate and 55 non-obligate interactions.

After identifying all the unique domains present in the interface of at least one complex in the datasets, the desolvation energies for all pairs of domains (DDIs) were calculated using Eq. (5.2). For each ligand-receptor (protein-protein) pair, if we found any duplicate DDIs during calculation, we simply computed the cumulative desolvation energy across all occurrences of that DDI. A domain is considered to be in the interface, if it has at least one residue interacting with a domain in the other chain. In this study, two types of domain-based properties are considered.

Domain-based Properties at the Individual Level

Since the CATH database is organized in a hierarchical scheme, in [3], a separate dataset of feature vectors was created for each level of the CATH hierarchy. After calculating the desolvation energies for all DDIs in level 8, for each DDI in higher levels, the desolvation energy was calculated by taking the sum of the desolvation energies of the corresponding DDIs at the next lowest level. After pre-processing the datasets, all zero-columns, which represent DDIs that were not present in any complexes, were removed. More details about the generation of domain-based feature vectors for each level are given in [3].

Each of these subsets of features was used for classification separately, in order to de-

termine the predictive power of a specific level in the CATH hierarchy. In [3], it was shown that domain-based features taken from level 2 (L2) and level 3 (L3) of CATH are more predictive than the features of other levels.

Domain-based Properties at the Combined Level

To generate these types of feature vectors, instead of considering each level in the CATH hierarchy separately, we consider combinations of levels. Thus, we do not obtain only one set of feature vectors per level. Indeed, by allowing arbitrary combinations of nodes, the total number of feature vectors would be exponential, with each feature vector corresponding to a sequence of nodes chosen to represent the domains found in the dataset. In order to maintain computational tractability and eliminate any redundancy in the feature vectors, the following constraints have been imposed: (a) there can be no overlap between nodes. That is, there cannot exist a pair of nodes in a sequence such that one node is an ancestor of the other; (b) only combinations of nodes taken from levels 2 and 3 of the hierarchy have been considered. Based on the results of our previous study [3], it is pertinent to conclude that the optimal combination of nodes will be found somewhere between these two levels; (c) nodes at level 3 which are the sole child of their parent node at level 2 have been discarded.

However, the number of node sequences to be evaluated is still exponential with respect to the number of nodes at level 2. Though, even an exhaustive enumeration of the entire search space is still computationally tractable given the size of our datasets (a conservative estimate would be about 30 days and 60 days for the ZH and MW datasets, respectively, on a single-core machine), this would be a poor choice, in general. Accordingly, a method based on sequential floating forward search (SFFS) [34] has been implemented to find a reasonable approximation to the best combination of nodes between levels 2 and 3. For

this, SFFS was initialized at the sequence of nodes consisting of the set of all nodes at level 2, as this sequence showed the greatest promise in our previous study [3]. Then, the search proceeded downward through the CATH tree towards the sequence of nodes corresponding to level 3.

A complete list of domain-based features in the individual and combined levels for ZH and MW datasets are shown in Tables 5.3 to 5.8 of the supplementary material, respectively.

5.3 Prediction Methods

After finding the domain-based features of the complexes of the MW and ZH datasets, we applied several prediction methods to them. In this work, the prediction is performed via commonly used classification methods, including LDR, SVM-SMO, NB and k-NN. More detailed explanation of each prediction method is given below.

5.3.1 Linear Dimensionality Reduction

The basic idea of LDR, which has become popular in pattern recognition due to its relatively easy implementation and high classification speed, is to represent an object of dimension n onto a lower-dimensional vector of dimension d, achieving this by performing a linear transformation. Each class, obligate or non-obligate, is represented by a random vector $\mathbf{x}_1 \sim N(\mu_1, \mathbf{S}_1)$ or $\mathbf{x}_2 \sim N(\mu_2, \mathbf{S}_2)$ respectively, with p_1 or p_2 as *a priori* probabilities. Each random vector is distributed normally with its mean μ and covariance $\mathbf{S}$. The aim of LDR is to find a linear transformation matrix $\mathbf{A}$ in such a way that the new classes $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$ are as separable as possible.

In this work, we use a generalization of the Chernoff discriminant analysis (CDA) crite-

rion proposed in [35], by relaxing the constraint that $p_1 = \beta$ and $p_2 = 1-\beta$. The generalized formula for the CDA criterion can be stated starting from the Chernoff distance as given in [36]. As in [35], we take the trace of the resulting matrix in the transformed space as follows:

$$J_{CDA}(\mathbf{A}) = tr\{p_1 p_2 (\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\mathbf{A}\mathbf{S}_E\mathbf{A}^t + \log(\mathbf{A}\mathbf{S}_W\mathbf{A}^t) - p_1\log(\mathbf{A}\mathbf{S}_1\mathbf{A}^t) - p_2\log(\mathbf{A}\mathbf{S}_2\mathbf{A}^t)\}. \tag{5.3}$$

where $\mathbf{S}_E = (\mu_1-\mu_2)(\mu_1-\mu_2)^t$ and $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$. Also, the most accurate error bound given in [36] is for a value of β ($\beta\varepsilon[0,1]$) that maximizes the Chernoff distance.

The aim of the CDA approach is to maximize the above equation. To solve this problem, a gradient-based algorithm is used [35]. This iterative algorithm needs a learning rate, α_k, which is maximized using the secant method to ensure that the gradient algorithm converges. The initialization of the matrix $\mathbf{A}$ is also an important issue in the gradient-based algorithm.

In this study, ten different initializations were performed and the solution for $\mathbf{A}$ that yielded the maximum Chernoff distance in the transformed space was selected. Since the best value of β is unknown in advance, an exhaustive search over all possible values of β, ranging from 0 to 1 with steps of 0.05, is applied in this study. This search gives a more accurate bound for the classification error, and hence we expect higher classification accuracy than other LDR methods. Note that the optimization of β is performed over all ten cross-validation folds, in order to avoid any bias in selecting the parameter for a particular fold. The resulting vectors $\mathbf{y}_i$ are then input to a quadratic Bayesian (QB) classifier and a linear Bayesian (LB) classifier, which is obtained by deriving a Bayesian classifier with a common covariance matrix. The maximum of the average classification accuracies from

these classifiers is reported. More details about the CDA approach and LDR methods can be found in [35].

5.3.2 Support Vector Machines Based on SMO

The aim of the SVM is to find the support vectors, and derive a linear classifier, which ideally separates the space into two regions. Classification using a linear classifier is not possible when the data is not linearly separable, and hence kernels are used to map the data into a higher dimensional space in which the classification boundary can be found much more efficiently. Sequential minimal optimization (SMO) is a fast learning algorithm which is widely applied in the training phase of an SVM classifier as one possible way to solve the underlying quadratic programming problem. In this study, the SMO module of the Waikato Environment for Knowledge Analysis (WEKA) with a normalized polynomial kernel, default parameter settings, and 10-fold cross-validation is used [37].

5.3.3 *k*-Nearest Neighbor

k-NN is one of the simplest classification methods, in which the class of each test sample can be easily found by a majority vote of the class labels of its neighbors. To achieve this, after computing and sorting the distances between the test sample and each training sample, the most frequent class label in the first "*k*" training samples (nearest neighbors) is assigned as the class of the test sample. Determining the appropriate number of neighbors is one of the challenges of this method. In this study, the IBK module of WEKA with Euclidean distance, default parameter settings, and 10-fold cross-validation is used [37].

5.3.4 Naive Bayes

One of the simplest probabilistic classifiers is naive Bayes. Assuming independence of features, the class of each test sample can be found by applying Bayes' theorem. The basic mechanism of NB is rather simple. The reader is referred to [38] for more details. In this study, the NaiveBayes module of WEKA with kernel estimator, default parameters, and 10-fold cross-validation is used [37].

5.4 Results and Discussion

To test our proposed method and perform an in-depth analysis of the domain-based prediction properties, the four classification methods outlined above have been used. The performance of these prediction methods is compared in terms of their classification accuracies, which are computed as follows: $acc = (TP+TN)/N$, where TP and TN are the total numbers of true positive (true obligate) and true negative (true non-obligate) predictions over the 10 cross-validation folds, respectively, and N is the total number of complexes in the dataset.

5.4.1 Analysis of the Prediction Properties

The prediction results of the LDR, NB, SVM-SMO and k-NN classifiers with individual and combined domain-based features for the MW and ZH datasets are shown in Table 5.1.

For the domain-based subsets of features at each level, extracted from the MW dataset, the MW-L2 subset achieves the best classification accuracy of 71.65% with SVM-SMO, while for MW-L3 the best obtained performance with SVM-SMO is 68.97%. However, by combining the feature vectors from levels 2 and 3 of MW (MW-L2+L3), the prediction

accuracy improves to 73.56%, which is much better than using features from individual levels (MW-L2 and MW-L3). This trend can be seen for all of the applied classifiers, which shows that using domain-based features by combining levels is better that using only features of level 2 and much better than using features of level 3 of the CATH hierarchy for prediction. Also, by comparing the classification accuracies, we can see that for all subsets of features extracted from the MW dataset, SVM-SMO performs better than other classifiers.

Similarly, the best accuracy for the ZH dataset, 78.74%, is obtained using combined domain-based properties (ZH-L2+L3) with SVM-SMO, compared to the best accuracies of 77.17% for ZH-L2 with the SVM-SMO classifier and 66.14% for ZH-L3 with LDR. Also, the performances of other classifiers for all subsets of features of the ZH dataset show the same trend: using the feature vector generated by combining features from levels 2 and 3 (ZH-L2+L3) is more efficient than using features from individual levels. Moreover, from the results, it is clear that after ZH-L2+L3, domain-based features of level 2 (ZH-L2) are more powerful for prediction than domain-based features of level 3.

Generally, it can be concluded for both the MW and ZH datasets that (a) domain-based properties at the combined level yield higher accuracies than domain-based properties on the individual levels; (b) domain-based features related to level 2 of CATH are more powerful than the features from level 3; (c) SVM-SMO is the most powerful classifier for all subsets of features; (d) SVM-SMO, LDR, NB and *k*-NN classifiers, however, show a similar trend. For all classifiers, DDIs from L2 are better than those of L3, while DDIs from a combination of L2 and L3 are much better than those of both L2 and L3 individually.

The receiver operating characteristic (ROC) curves for the MW and ZH datasets using different DDI properties for prediction are shown in Figs. 5.2(a) and 5.2(b), respectively.

Table 5.1: Prediction accuracies of SVM-SMO, NB, *k*-NN and LDR for all domain-based subsets of features of the ZH and MW datasets.

Subset Name	# Features	LDR	SVM-SMO	*k*-NN	NB
MW-L2	96	70.01	**71.65**	68.96	69.39
MW-L3	291	67.05	**68.97**	67.43	67.05
MW-L2+L3	133	70.11	**73.56**	69.73	70.15
ZH-L2	64	74.80	**77.17**	66.14	71.65
ZH-L3	150	**66.14**	58.27	59.04	56.70
ZH-L2+L3	70	75.59	**78.74**	66.93	72.44

These ROC curves are plotted based on the true positive rate (TPR), aka "sensitivity", vs. the false positive rate (FPR), or "1 - specificity", at various threshold settings. To generate the ROC curves, the sensitivity and specificity of each subset of features were determined for different values of d and β values in the CDA classifier. Then, by applying a simple algorithm, the FPR and TPR points were filtered as follows: (a) for the same FPR values, the greatest TPR value (top point) was chosen, and (b) for the same TPR values, the smallest FPR value (left point) was chosen. A polynomial function with degree 2 was then fitted to the selected points. From the ROC curves, it is clear that for both datasets, the prediction performances of LDR using DDI properties on the combined level (ZH-L2+L3 and MW-L2+L3) are clearly better than using DDI properties of level 2 (ZH-L2 and MW-L2) and much better than those of level 3 (ZH-L3 and MW-L3).

In addition, the area under the curve (AUC), is computed for each of the above ROC curves using the trapezoid rule. The AUC values are also shown in Figure 5.2. The AUC for ZH-L2+L3 is 0.68 which is greater than AUC of both ZH-L2 (0.66) and ZH-L3 (0.63). Similarly, the AUC for the MW dataset using DDI properties on the combined level (MW-L2+L3) is 0.65 while for MW-L2 is 0.60. Also, the AUC of MW-L2 is greater than that of MW-L3. Generally, by comparing the AUC values, it can be concluded that DDI properties

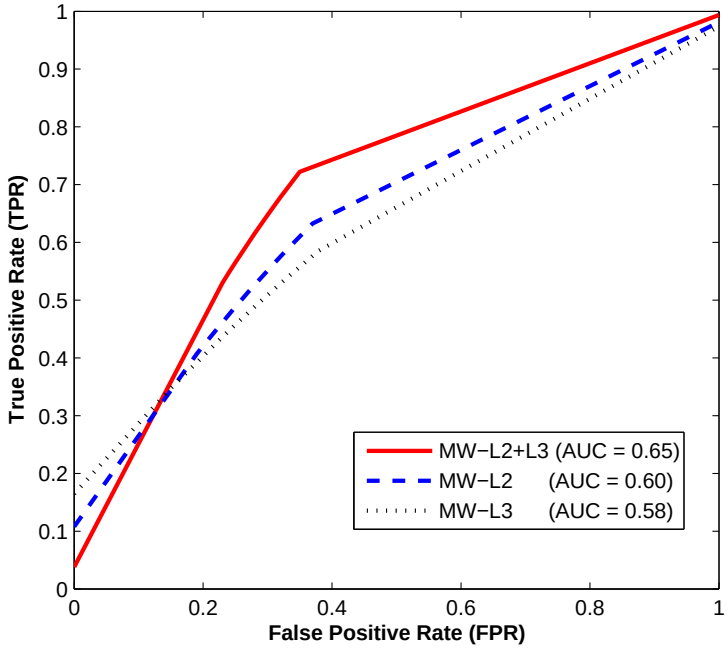
from the combined levels show much better predictive power than DDI properties from the individual levels.

5.4.2 Analysis of Structural Properties

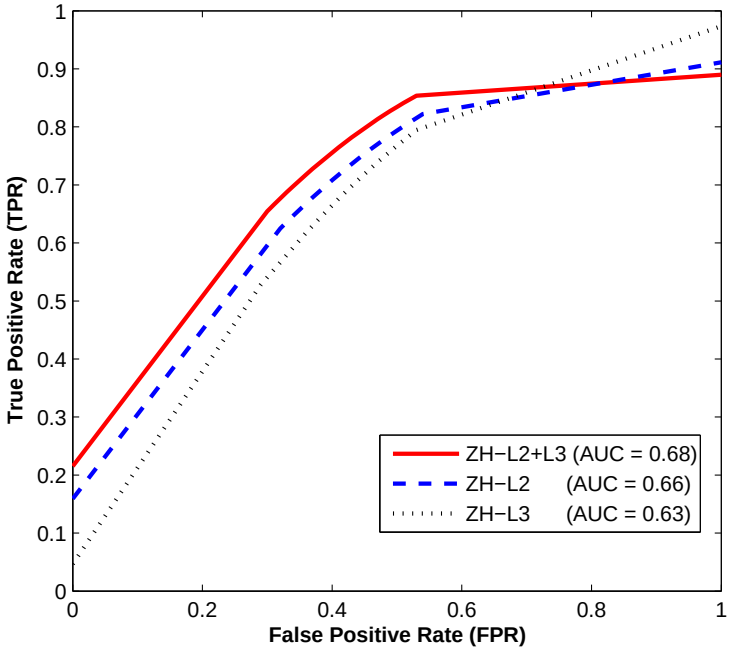
As discussed earlier, in level 1 of the CATH hierarchy, the "class" of each complex is defined. The four classes of CATH, which are determined based on the secondary structure composition of the complexes, are mainly-alpha (*c1*), mainly-beta (*c2*), mixed alpha-beta (*c3*) and secondary structure content (*c4*) [16]. A summary of the number of DDIs present in both the ZH and MW datasets, categorized by class type, *c1* to *c4*, is shown in Table 5.2. From the table, it is clear that most of the DDIs are between domains of *c2* and other classes in which *c2:c2* and *c2:c3* have the highest ranks. However, domains of *c4* have no interactions (with *c1* and *c4*) or the least number of interactions with the domains of other levels. This indicates that DDIs taken from *c4* are less important and could be ignored for achieving a faster, yet still accurate, prediction. In contrast, DDIs taken from *c2* are more powerful for prediction.

To investigate this hypothesis, a structural feature selection scheme has been applied on the MW-L2 and ZH-L2 datasets. For this, DDI feature vectors from level 2 have been grouped based on their class (secondary structure) type interactions such as *c1-c1*, *c1-c2*, and so on. Then, each group of features was classified with SVM-SMO and *k*-NN classifiers, individually. The classification results are shown in Table 5.2.

For the MW-L2 subset, the feature vector of *c2-c3* achieves the best prediction with 67.43% and 64.75% accuracies by SVM-SMO and *k*-NN, respectively. The other DDIs from *c2* also achieve better performance than using DDIs of other classes. The most notable feature vectors are *c2-c4* and *c1-c1*, because they achieve acceptable prediction accuracies



(a) MW



(b) ZH

Figure 5.2: ROC curves and AUC values for all subsets of features of (a) MW and (b) ZH datasets.

Table 5.2: A summary of the number of CATH DDIs from level 2 present in the ZH and MW datasets, categorized by their class types.

		MW-L2			ZH-L2		
Domain1	Domain2	#DDIs	SVM-SMO	*k*-NN	#DDIs	SVM-SMO	*k*-NN
c1	*c1*	5	63.98	63.68	3	56.69	56.69
c1	*c2*	9	63.98	63.68	10	56.69	58.27
c1	*c3*	5	62.07	61.68	5	56.69	56.69
c1	*c4*	0	0	0	0	0	0
c2	*c2*	24	63.98	62.07	18	60.63	61.42
c2	*c3*	32	**67.43**	**64.75**	17	**62.99**	**61.42**
c2	*c4*	6	64.75	63.98	2	56.69	56.69
c3	*c3*	13	59	61.68	7	56.69	56.69
c3	*c4*	2	55.17	61.3	2	55.9	55.9
c4	*c4*	0	0	0	0	0	0

with less features. As expected, the worst prediction results were achieved using DDI feature vectors from *c4* and other classes.

Similarly, for the ZH-L2 subset, it is clear that while the most discriminative feature vector for prediction is *c2-c3*, obtaining accuracies of 62.99% by SVM-SMO and 61.42% by *k*-NN, the worst feature vectors are DDIs taken from *c4* (*c1-c4*, *c3-c4* and *c4-c4*). Moreover, the feature vector of *c2-c2* is the second most powerful for prediction. All other subsets of features yield almost the same performance. Some notable DDIs are *c2-c4* and *c1-c1*, as they achieve reasonable performance with fewer features.

Furthermore, using structural feature selection, a decrease of 4%-8% in prediction accuracy compared to the original subset of features from MW-L2 and ZH-L2 (Table 5.1) are observed. However, these decreases in performance can be acceptable given that there are fewer features than in the original feature vectors, leading to a reduction in time and space requirements.

5.5 Conclusion

The idea of employing a structural domain-based approach for predicting obligate and non-obligate protein complexes, which were presented in our previous studies, is extended in this paper. Different interface properties, including domain-based properties on the individual levels and on the combined levels of the CATH hierarchy are used for prediction. The classification is performed using various techniques, including LDR, SVM-SMO, *k*-NN, and NB, for two well-known datasets of pre-classified complexes.

The prediction results demonstrate a significant improvement by combining nodes from different levels in the CATH hierarchy, rather than considering DDI features of each level separately. Also, it has been shown that DDIs at upper levels are more powerful than those at lower levels for prediction. The plotted ROC curves and calculated AUC values corroborate the prediction results.

Furthermore, a numerical analysis shows that while there are fewer interactions between domains of *c4* and domains of other classes, most of the interactions are between domains of *c2* and domains of other classes of level 2 of the CATH hierarchy. Also, the prediction results on the structurally selected features of the MW-L2 and ZH-L2 datasets confirm that DDIs taken from the mainly-beta class (*c2*), especially DDIs between the mainly-beta and alpha-beta classes (*c2-c3*) are the best properties for predicting obligate and non-obligate PPIs.

5.6 Supplementary Materials

Table 5.3: List of feature vectors for the ZH-L2 dataset.

1.1 : 1.1	1.5 : 3.1	2.14 : 2.3	2.3 : 3.9
1.1 : 1.25	1.5 : 3.9	2.14 : 2.6	2.4 : 2.4
1.1 : 1.5	2.1 : 2.1	2.14 : 2.7	2.4 : 2.6
1.1 : 2.1	2.1 : 3.4	2.14 : 2.8	2.6 : 3.3
1.1 : 2.14	2.102 : 2.4	2.14 : 3.1	2.7 : 2.7
1.2 : 2.1	2.102 : 2.6	2.14 : 3.2	2.8 : 3.1
1.2 : 2.7	2.102 : 3.1	2.14 : 3.3	2.8 : 3.9
1.2 : 3.8	2.102 : 3.6	2.14 : 3.5	3.1 : 4.1
1.25 : 2.3	2.12 : 2.14	2.14 : 3.9	3.2 : 3.6
1.25 : 3.6	2.12 : 2.6	2.15 : 2.3	3.2 : 3.9
1.25 : 3.8	2.12 : 4.1	2.15 : 2.4	3.2 : 4.1
1.5 : 2.1	2.13 : 2.6	2.15 : 2.8	3.3 : 3.4
1.5 : 2.4	2.13 : 2.7	2.15 : 3.1	3.3 : 3.6
1.5 : 2.6	2.13 : 3.1	2.15 : 3.2	3.3 : 3.8
1.5 : 2.7	2.13 : 4.1	2.15 : 3.3	3.5 : 3.8
1.5 : 2.8	2.14 : 2.14	2.15 : 3.4	3.6 : 3.6

Table 5.4: List of feature vectors for the ZH-L3 dataset.

1.10.10 : 1.10.10	1.10.238 : 3.30.62	2.10.69 : 3.90.380	2.40.30 : 3.40.720
1.10.10 : 1.10.140	1.10.238 : 3.40.30	2.10.69 : 3.90.830	2.40.50 : 4.10.480
Continued on next page			

1.10.10 : 1.10.510	1.10.287 : 1.20.870	2.120.10 : 3.90.330	2.40.70 : 3.10.120
1.10.10 : 1.25.40	1.10.400 : 1.20.1250	2.130.10 : 3.30.10	2.40.70 : 3.20.20
1.10.10 : 3.10.450	1.10.400 : 3.30.428	2.140.10 : 2.30.29	2.60.40 : 3.10.50
1.10.10 : 3.30.350	1.10.400 : 3.30.560	2.140.10 : 3.40.420	2.60.90 : 3.30.810
1.10.10 : 3.40.20	1.10.420 : 1.20.870	2.140.10 : 3.40.50	2.60.90 : 3.90.1150
1.10.10 : 3.60.21	1.10.420 : 3.40.190	2.150.10 : 3.90.1150	2.70.70 : 3.90.340
1.10.10 : 3.80.10	1.10.439 : 3.30.70	2.30.26 : 3.30.350	2.70.98 : 3.30.560
1.10.1040 : 1.10.150	1.10.472 : 1.10.760	2.30.26 : 3.30.390	2.70.98 : 3.30.62
1.10.1040 : 1.25.40	1.10.472 : 2.40.10	2.30.26 : 3.30.470	2.80.10 : 3.90.420
1.10.1040 : 3.30.800	1.10.472 : 3.30.1330	2.30.26 : 3.30.540	3.10.120 : 3.60.20
1.10.120 : 1.10.494	1.10.494 : 3.20.20	2.30.26 : 3.30.560	3.10.130 : 3.10.50
1.10.120 : 2.40.70	1.10.494 : 3.90.110	2.30.26 : 3.40.192	3.10.450 : 3.30.350
1.10.120 : 3.40.20	1.10.506 : 1.10.510	2.30.26 : 3.40.20	3.10.50 : 3.90.640
1.10.120 : 3.40.532	1.10.506 : 3.30.62	2.30.26 : 3.40.420	3.20.16 : 3.20.20
1.10.1200 : 1.20.870	1.10.510 : 3.40.47	2.30.26 : 3.40.532	3.20.20 : 3.90.180
1.10.1200 : 3.10.120	1.10.520 : 3.30.200	2.30.26 : 3.40.640	3.30.10 : 3.30.572
1.10.1320 : 2.80.10	1.10.555 : 3.90.70	2.30.26 : 3.40.80	3.30.10 : 3.40.570
1.10.1320 : 3.90.1150	1.10.580 : 1.20.90	2.30.26 : 3.50.50	3.30.10 : 3.90.380
1.10.1320 : 3.90.650	1.10.580 : 3.30.1330	2.30.26 : 3.60.20	3.30.1120 : 3.30.572
1.10.140 : 1.50.10	1.10.620 : 2.10.60	2.30.26 : 3.90.1150	3.30.170 : 3.30.572
1.10.1400 : 3.40.718	1.10.760 : 2.140.10	2.30.26 : 3.90.340	3.30.170 : 3.30.70
1.10.150 : 2.30.30	1.10.760 : 2.40.50	2.30.26 : 3.90.540	3.30.200 : 3.90.80
1.10.150 : 3.10.450	1.10.840 : 3.30.470	2.30.29 : 2.40.30	3.30.350 : 3.40.390

Continued on next page

1.10.150 : 3.90.440	1.10.840 : 3.40.720	2.30.29 : 2.60.90	3.30.390 : 3.40.570
1.10.150 : 3.90.650	1.20.1050 : 3.30.450	2.30.29 : 2.70.98	3.30.40 : 3.40.80
1.10.167 : 3.30.200	1.20.120 : 2.70.70	2.30.29 : 3.10.50	3.30.428 : 3.90.770
1.10.196 : 2.30.120	1.20.120 : 3.90.640	2.30.29 : 3.20.16	3.30.560 : 3.40.720
1.10.210 : 1.10.510	1.20.58 : 3.40.718	2.30.29 : 3.20.70	3.30.60 : 3.40.47
1.10.210 : 2.70.70	1.20.870 : 3.30.70	2.30.29 : 3.30.170	3.30.70 : 3.90.830
1.10.210 : 2.80.10	1.20.90 : 3.30.350	2.30.29 : 3.30.40	3.40.190 : 3.90.1150
1.10.230 : 2.40.70	1.20.90 : 3.90.770	2.30.29 : 3.30.70	3.40.192 : 3.80.10
1.10.230 : 3.40.80	1.25.10 : 3.80.10	2.30.30 : 3.20.20	3.40.390 : 3.90.830
1.10.238 : 2.150.10	1.25.40 : 3.40.80	2.30.36 : 3.90.540	3.40.640 : 3.80.10
1.10.238 : 3.30.1330	2.10.25 : 2.120.10	2.40.10 : 2.80.10	3.90.110 : 4.10.480
1.10.238 : 3.30.350	2.10.60 : 3.30.800	2.40.10 : 3.40.532	3.90.370 : 3.90.80
1.10.238 : 3.30.360	2.10.60 : 4.10.410		

Table 5.5: List of feature vectors for the ZH-L2+L3 dataset.

1.1 : 1.1	1.5 : 2.30.42	2.15 : 3.6	2.30.26 : 3.2
1.1 : 1.20.120	1.5 : 2.8	2.15 : 3.9	2.30.30 : 2.4
1.1 : 1.20.870	1.5 : 3.1	2.15 : 4.1	2.30.30 : 2.7
1.1 : 3.3	2.1 : 2.102	2.30.120 : 2.30.120	2.30.30 : 2.8
1.1 : 3.4	2.1 : 2.15	2.30.120 : 2.30.26	2.30.42 : 2.6
1.1 : 3.9	2.102 : 2.15	2.30.120 : 2.30.29	2.4 : 2.4
1.20.1050 : 2.102	2.12 : 2.14	2.30.120 : 2.30.42	2.6 : 3.2
1.20.120 : 2.13	2.12 : 2.15	2.30.120 : 2.4	2.6 : 4.1
1.20.120 : 3.2	2.12 : 2.30.29	2.30.120 : 2.7	2.8 : 3.6
1.20.1250 : 1.5	2.12 : 3.3	2.30.120 : 3.1	3.1 : 3.4
1.20.1250 : 2.8	2.13 : 2.30.26	2.30.120 : 3.4	3.1 : 3.6
1.20.58 : 3.2	2.13 : 2.30.42	2.30.120 : 4.1	3.1 : 3.8
1.20.90 : 1.20.90	2.14 : 2.30.26	2.30.26 : 2.30.36	3.1 : 4.1
1.25 : 2.30.42	2.14 : 3.2	2.30.26 : 2.6	3.2 : 3.3
1.25 : 2.4	2.14 : 3.3	2.30.26 : 2.7	3.2 : 3.4
1.25 : 3.6	2.15 : 2.15	2.30.26 : 2.8	3.5 : 3.8
1.5 : 2.30.26	2.15 : 3.1	2.30.26 : 3.1	3.6 : 3.6
1.5 : 2.30.29	2.15 : 3.4		

Table 5.6: List of feature vectors for the MW-L2 dataset.

1.1 :1.1	2.1 : 2.7	2.15 : 2.7	2.6 : 2.6
1.1 : 1.2	2.1 : 2.8	2.15 : 2.8	2.6 : 3.4
1.1 : 1.25	2.1 : 3.2	2.15 : 3.2	2.7 : 3.4
1.1 : 1.5	2.1 : 3.5	2.15 : 3.3	2.7 : 3.8
1.1 : 2.1	2.1 : 3.8	2.15 : 3.4	2.7 : 4.1
1.1 : 2.14	2.1 : 3.9	2.15 : 3.5	2.8 : 2.8
1.1 : 2.4	2.1 : 4.1	2.15 : 3.6	2.8 : 3.4
1.1 : 2.6	2.102 : 2.15	2.15 : 3.8	2.8 : 3.5
1.2 : 1.5	2.11 : 2.6	2.15 : 4.1	2.8 : 3.9
1.2 : 2.102	2.11 : 2.7	2.17 : 2.17	3.1 : 3.8
1.25 : 2.8	2.11 : 3.2	2.17 : 2.3	3.2 : 3.4
1.5 : 2.11	2.11 : 3.3	2.17 : 3.1	3.2 : 3.6
1.5 : 2.13	2.11 : 3.8	2.17 : 3.2	3.2 : 3.8
1.5 : 2.17	2.13 : 2.3	2.17 : 3.3	3.2 : 4.1
1.5 : 3.2	2.13 : 2.4	2.17 : 3.4	3.3 : 3.4
1.5 : 3.3	2.13 : 2.8	2.17 : 3.6	3.3 : 3.5
1.5 : 3.4	2.13 : 3.5	2.17 : 3.8	3.4 : 3.6
1.5 : 3.8	2.14 : 3.1	2.17 : 3.9	3.4 : 3.8
1.5 : 3.9	2.14 : 3.2	2.17 : 4.1	3.5 : 3.8
2.1 : 2.17	2.14 : 3.3	2.2 : 2.2	3.5 : 3.9
2.1 : 2.2	2.14 : 3.4	2.3 : 2.6	3.6 : 3.8
2.1 : 2.3	2.14 : 4.1	2.3 : 2.8	3.6 : 4.1

Continued on next page

2.1 : 2.4	2.15 : 2.4	2.3 : 3.3	3.8 : 3.8
2.1 : 2.6	2.15 : 2.6	2.4 : 4.1	3.8 : 3.9

Table 5.7: List of feature vectors for the MW-L3 dataset.

1.10.10 : 3.40.420	1.10.468 : 4.10.1030	1.20.950 : 3.90.380	2.170.240 : 3.30.365
1.10.10 : 4.10.720	1.10.468 : 4.10.40	1.20.950 : 4.10.410	2.170.40 : 3.30.1340
1.10.10 : 4.10.800	1.10.468 : 4.10.410	1.20.950 : 4.10.820	2.20.25 : 3.30.1650
1.10.100 : 1.10.520	1.10.468 : 4.10.740	1.25.10 : 2.10.22	2.20.25 : 3.30.70
1.10.100 : 1.10.620	1.10.468 : 4.10.980	1.25.10 : 2.10.25	2.20.25 : 3.90.760
1.10.100 : 3.30.30	1.10.472 : 1.10.506	1.25.10 : 2.40.200	2.30.36 : 2.70.50
1.10.100 : 3.40.470	1.10.472 : 1.10.530	1.25.10 : 2.60.15	2.30.36 : 3.30.1130
1.10.1030 : 1.10.472	1.10.472 : 1.10.565	1.25.10 : 2.60.200	2.30.40 : 3.90.440
1.10.1030 : 3.30.1450	1.10.472 : 1.10.645	1.25.10 : 2.70.50	2.30.42 : 2.40.250
1.10.1030 : 3.90.380	1.10.472 : 1.20.1060	1.25.10 : 3.10.10	2.30.42 : 3.40.462
1.10.1060 : 1.10.1170	1.10.472 : 1.20.1130	1.25.10 : 3.30.280	2.30.42 : 3.90.830
1.10.1060 : 2.60.90	1.10.472 : 3.30.505	1.25.10 : 3.30.370	2.40.128 : 2.60.120
1.10.1060 : 2.70.230	1.10.472 : 4.10.470	1.25.40 : 2.70.70	2.40.200 : 3.40.50
1.10.1060 : 3.10.130	1.10.494 : 3.30.830	1.25.40 : 3.10.20	2.40.200 : 3.90.470
1.10.1060 : 3.30.10	1.10.494 : 3.40.710	1.50.10 : 3.30.420	2.40.30 : 3.30.720
1.10.1090 : 1.10.286	1.10.510 : 1.20.930	2.10.150 : 2.10.22	2.40.30 : 3.90.440
1.10.1140 : 2.10.90	1.10.510 : 4.10.40	2.10.150 : 3.30.1120	2.40.30 : 3.90.700
1.10.1170 : 1.10.1820	1.10.520 : 2.170.240	2.10.150 : 3.30.1410	2.40.40 : 3.10.320
1.10.1170 : 1.10.555	1.10.520 : 2.30.40	2.10.150 : 3.30.1490	2.40.40 : 3.30.1470
1.10.1170 : 1.20.210	1.10.520 : 3.10.390	2.10.22 : 3.50.30	2.40.40 : 3.30.62
1.10.1170 : 1.20.810	1.10.520 : 3.30.10	2.10.22 : 4.10.820	2.40.50 : 3.30.1340
1.10.1170 : 2.60.40	1.10.520 : 3.30.170	2.10.50 : 2.70.70	2.40.70 : 4.10.160

Continued on next page

1.10.120 : 1.10.167	1.10.530 : 1.20.150	2.10.50 : 3.90.640	2.60.120 : 3.90.1170
1.10.120 : 1.10.468	1.10.530 : 1.20.930	2.10.60 : 3.30.1410	2.60.130 : 3.90.550
1.10.1200 : 1.20.89	1.10.533 : 2.170.240	2.10.60 : 3.40.950	2.60.15 : 3.40.970
1.10.1200 : 2.60.90	1.10.533 : 3.40.630	2.10.60 : 3.40.970	2.60.40 : 4.10.740
1.10.1200 : 3.10.320	1.10.555 : 1.10.8	2.10.69 : 3.30.420	2.60.90 : 3.90.470
1.10.1200 : 3.40.950	1.10.555 : 1.20.190	2.10.70 : 2.102.10	2.70.230 : 3.30.40
1.10.1200 : 3.90.70	1.10.555 : 3.30.30	2.10.70 : 2.170.240	2.70.230 : 3.90.470
1.10.1200 : 4.10.980	1.10.555 : 3.90.550	2.102.10 : 3.30.365	3.10.10 : 3.90.1170
1.10.1320 : 2.170.40	1.10.555 : 3.90.70	2.102.10 : 3.90.20	3.10.100 : 4.10.320
1.10.168 : 1.10.565	1.10.565 : 2.10.10	2.102.10 : 3.90.470	3.10.130 : 3.10.20
1.10.168 : 3.90.175	1.10.565 : 3.30.62	2.102.10 : 3.90.640	3.10.130 : 4.10.320
1.10.168 : 3.90.210	1.10.620 : 2.30.42	2.102.10 : 3.90.830	3.10.20 : 3.30.350
1.10.1760 : 3.30.530	1.10.645 : 3.30.530	2.102.10 : 4.10.40	3.10.380 : 3.90.440
1.10.1760 : 3.90.20	1.10.645 : 3.60.21	2.102.10 : 4.10.410	3.10.390 : 3.90.760
1.10.1780 : 4.10.820	1.10.8 : 1.20.58	2.102.10 : 4.10.720	3.30.10 : 3.90.190
1.10.1820 : 3.90.380	1.10.8 : 2.60.30	2.102.10 : 4.10.800	3.30.1130 : 3.30.830
1.10.196 : 1.20.1250	1.10.8 : 3.30.500	2.102.10 : 4.10.980	3.30.1390 : 3.40.810
1.10.238 : 1.20.1270	1.10.840 : 1.20.840	2.110.10 : 2.110.10	3.30.1450 : 3.30.62
1.10.238 : 2.30.42	1.10.840 : 2.10.10	2.110.10 : 2.170.240	3.30.1450 : 3.90.330
1.10.238 : 2.70.50	1.20.1060 : 3.60.120	2.110.10 : 2.20.25	3.30.1450 : 4.10.940
1.10.246 : 1.10.510	1.20.1060 : 4.10.630	2.110.10 : 2.30.36	3.30.1460 : 3.40.420
1.10.246 : 3.30.420	1.20.1130 : 2.10.70	2.110.10 : 2.30.40	3.30.1460 : 3.50.50
1.10.246 : 3.90.209	1.20.1130 : 3.30.420	2.110.10 : 2.40.200	3.30.1460 : 4.10.540

Continued on next page

1.10.246 : 3.90.760	1.20.120 : 1.20.1270	2.110.10 : 2.40.250	3.30.1490 : 3.40.1470
1.10.246 : 3.90.970	1.20.120 : 2.150.10	2.110.10 : 2.40.30	3.30.1650 : 3.30.70
1.10.246 : 4.10.160	1.20.120 : 3.30.479	2.110.10 : 2.40.70	3.30.170 : 3.90.830
1.10.246 : 4.10.365	1.20.1250 : 3.90.330	2.110.10 : 2.60.120	3.30.190 : 3.90.540
1.10.286 : 3.10.110	1.20.1270 : 3.10.20	2.110.10 : 2.60.90	3.30.370 : 3.90.210
1.10.287 : 2.40.30	1.20.1310 : 2.10.90	2.110.10 : 3.10.110	3.30.40 : 3.90.330
1.10.287 : 2.70.50	1.20.1310 : 2.60.90	2.110.10 : 3.20.20	3.30.465 : 3.90.760
1.10.287 : 3.30.1460	1.20.1310 : 3.90.830	2.110.10 : 3.30.1470	3.30.500 : 3.40.532
1.10.287 : 3.40.970	1.20.150 : 2.60.210	2.110.10 : 3.30.365	3.30.505 : 3.40.462
1.10.3470 : 1.10.468	1.20.150 : 3.90.760	2.110.10 : 3.30.390	3.30.505 : 3.60.120
1.10.3470 : 1.20.150	1.20.150 : 4.10.160	2.110.10 : 3.30.43	3.30.505 : 3.90.20
1.10.3470 : 2.150.10	1.20.190 : 3.10.50	2.110.10 : 3.30.479	3.30.559 : 3.90.700
1.10.3470 : 2.30.42	1.20.190 : 3.30.1130	2.110.10 : 3.30.530	3.30.56 : 3.40.50
1.10.3470 : 3.20.20	1.20.190 : 3.30.43	2.110.10 : 3.30.60	3.30.565 : 3.30.60
1.10.3470 : 3.40.1080	1.20.190 : 3.40.420	2.110.10 : 3.30.70	3.30.62 : 3.90.330
1.10.400 : 3.30.1130	1.20.190 : 3.90.370	2.110.10 : 3.30.710	3.30.830 : 3.90.700
1.10.420 : 1.20.1130	1.20.5 : 2.30.22	2.110.10 : 3.30.930	3.40.1380 : 4.10.720
1.10.420 : 2.10.10	1.20.5 : 2.40.70	2.110.10 : 3.40.30	3.40.228 : 3.60.21
1.10.468 : 3.10.380	1.20.5 : 3.30.1390	2.110.10 : 3.40.630	3.40.532 : 3.90.210
1.10.468 : 3.30.559	1.20.5 : 3.90.470	2.110.10 : 3.40.830	3.40.830 : 3.90.470
1.10.468 : 3.40.20	1.20.810 : 3.10.390	2.110.10 : 3.50.50	3.50.40 : 4.10.980
1.10.468 : 3.40.810	1.20.840 : 2.60.15	2.110.10 : 3.60.10	3.80.10 : 3.90.380
1.10.468 : 3.60.10	1.20.840 : 2.70.230	2.110.10 : 3.90.209	3.90.20 : 4.10.40

Continued on next page

1.10.468 : 3.90.1170	1.20.85 : 4.10.630	2.110.10 : 3.90.640	3.90.380 : 3.90.640
1.10.468 : 3.90.190	1.20.89 : 4.10.75	2.130.10 : 3.40.420	3.90.510 : 4.10.470
1.10.468 : 3.90.340	1.20.930 : 3.30.1390	2.150.10 : 3.40.532	3.90.540 : 3.90.70
1.10.468 : 3.90.640	1.20.950 : 2.60.90	2.150.10 : 4.10.540	3.90.830 : 4.10.940
1.10.468 : 3.90.70	1.20.950 : 3.50.50	2.170.240 : 3.30.1460	

Table 5.8: List of feature vectors for the MW-L2+L3 dataset.

1.1 : 1.1	2.10.240 : 3.10.390	2.13 : 4.10.470	2.2 : 4.10.480
1.1 : 1.2	2.10.240 : 3.4	2.13 : 4.10.630	2.2 : 4.10.540
1.1 : 1.25	2.10.240 : 3.6	2.13 : 4.10.75	2.2 : 4.10.75
1.1 : 1.5	2.10.240 : 3.8	2.13 : 4.10.980	2.2 : 4.10.800
1.1 : 2.10.10	2.10.240 : 3.9	2.14 : 2.150.10	2.3 : 2.60.130
1.1 : 2.10.25	2.10.240 : 4.10.365	2.14 : 2.17	2.3 : 3.10.20
1.2 : 2.10.69	2.10.240 : 4.10.40	2.14 : 2.2	2.3 : 4.10.320
1.2 : 3.2	2.10.25 : 2.60.200	2.14 : 2.3	2.3 : 4.10.470
1.2 : 3.6	2.10.25 : 3.10.130	2.14 : 2.4	2.3 : 4.10.630
1.25 : 2.10.90	2.10.50 : 2.60.120	2.14 : 2.60.40	2.3 : 4.10.720
1.25 : 2.11	2.10.50 : 2.60.90	2.14 : 3.10.100	2.3 : 4.10.75
1.5 : 2.4	2.10.60 : 3.6	2.14 : 3.10.130	2.3 : 4.10.820
1.5 : 3.3	2.10.69 : 4.10.365	2.14 : 3.10.450	2.4 : 4.10.800
1.5 : 3.4	2.10.69 : 4.10.75	2.14 : 3.10.50	2.60.130 : 4.10.140
1.5 : 3.9	2.10.70 : 2.60.250	2.14 : 3.4	2.60.15 : 4.10.540
1.5 : 4.10.820	2.10.70 : 3.10.110	2.14 : 3.8	2.60.200 : 3.4
1.5 : 4.10.940	2.10.77 : 2.150.10	2.14 : 4.10.1030	2.60.200 : 3.5
1.5 : 4.10.980	2.10.77 : 2.60.210	2.14 : 4.10.160	2.60.210 : 4.10.470
2.10.10 : 2.10.25	2.10.90 : 3.10.380	2.14 : 4.10.365	2.60.210 : 4.10.540
2.10.10 : 2.10.60	2.102 : 2.11	2.14 : 4.10.470	2.60.250 : 2.60.250
2.10.10 : 2.10.77	2.102 : 3.2	2.14 : 4.10.720	2.60.30 : 3.3
2.10.10 : 2.11	2.11 : 2.60.40	2.14 : 4.10.75	2.60.90 : 3.6

Continued on next page

2.10.10 : 2.13	2.11 : 3.10.320	2.14 : 4.10.800	2.8 : 2.8
2.10.10 : 2.60.130	2.11 : 3.10.390	2.14 : 4.10.820	3.10.10 : 3.10.100
2.10.10 : 3.10.10	2.13 : 2.60.15	2.14 : 4.10.940	3.10.130 : 4.10.410
2.10.150 : 2.10.60	2.13 : 2.60.200	2.150.10 : 4.10.320	3.10.380 : 4.10.365
2.10.150 : 2.10.69	2.13 : 2.60.210	2.150.10 : 4.10.630	3.10.50 : 3.4
2.10.240 : 2.10.77	2.13 : 2.60.250	2.150.10 : 4.10.75	3.3 : 3.6
2.10.240 : 2.60.200	2.13 : 3.10.50	2.17 : 4.10.480	3.5 : 4.10.940
2.10.240 : 2.60.210	2.13 : 3.6	2.17 : 4.10.540	4.10.1030 : 4.10.820
2.10.240 : 2.60.250	2.13 : 3.8	2.17 : 4.10.820	4.10.260 : 4.10.800
2.10.240 : 2.60.30	2.13 : 4.10.260	2.2 : 3.3	4.10.320 : 4.10.40
2.10.240 : 3.10.10	2.13 : 4.10.40	2.2 : 4.10.140	4.10.410 : 4.10.75
2.10.240 : 3.10.380			

Bibliography

[1] L. Chen, R. Wang, and X. Zhang, *Biomolecular Networks: Methods and Applications in Systems Biology*. John Wiley and Sons, 2009.

[2] V. D. Dwivedi, S. Arora, and A. Pandey, "Computational analysis of physico-chemical properties and homology modeling of carbonic anhydrase from cordyceps militaris," *Network Modeling Analysis in Health Informatics and Bioinformatics*, pp. 1–4, 2013.

[3] M. Maleki, M. Hall, and L. Rueda, "Using structural domain to predict obligate and non-obligate protein-protein interactions," in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB2012)*, San Diego, USA, May 2012, pp. 9–15.

[4] M. Hall, M. Maleki, and L. Rueda, "Multi-level structural domain-domain interactions for prediction of obligate and non-obligate protein-protein interactions," in *ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM-BCB)*, Florida, USA, October 2012, pp. 518–520.

[5] N. Zaki, "Protein-protein interaction prediction using homology and inter-domain linker region information." *Advances in Electrical Engineering and Computational Science, Springer*, vol. 39, pp. 635–645, 2009.

[6] N. Zaki and P. C. S. Lazarova-Molnar, W. El-Hajj, "Protein-protein interaction based on pairwise similarity." *BMC Bioinformatics*, vol. 10, no. 150, pp. doi:10.1186/1471–2105–10–150, 2009.

[7] M. Singhal and H. Resat, "A domain-based approach to predict protein-protein interactions." *BMC Bioinformatics*, vol. 8, no. 199, pp. doi:10.1186/1471–2105–8–199, 2007.

[8] T. Akutsu and M. Hayashida, "Domain-based prediction and analysis of protein-protein interactions." *Biological data mining in protein interaction networks, Medical Information Science Reference, chapter 3*, pp. 29–44, 2009.

[9] P. Chandrasekaran, C. Doss, J. Nisha, R. Sethumadhavan, V. Shanthi, K. Ramanathan, and R. Rajasekaran, "In silico analysis of detrimental mutations in add domain of chromatin remodeling protein atrx that cause atr-x syndrome: X-linked disorder," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 2, no. 3, pp. 123–135, 2013.

[10] S. De, O. Krishnadev, N. Srinivasan, and N. Rekha, "Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different." *BMC Structural Biology*, vol. 5, no. 15, pp. doi:10.1186/1472–6807–5–15, 2005.

[11] S. G. J. V. Eichborn and R. Preissner, "Structural features and evolution of protein-protein interactions," *Genome Inform*, vol. 22, pp. 1–10, 2010.

[12] D. B. Singh, M. K. Gupta, R. K. Kesharwani, and K. Misra, "Comparative docking and admet study of some curcumin derivatives and herbal congeners targeting -amyloid," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 2, no. 1, pp. 13–27, 2013.

[13] D. Caffrey, S. Somaroo, J. Hughes, J. Mintseris, and E. Huang, "Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?" *Protein Sci.*, vol. 13, no. 1, pp. 190–202, 2004.

[14] J. Park and D. Bolser, "Conservation of protein interaction network in evolution." *Genome Inform*, vol. 12, pp. 135–140, 2001.

[15] M. Punta, P. Coggill, R. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. Sonnhammer, S. Eddy, A. Bateman, and R. Finn, "The Pfam protein families database." *Nucleic Acids Res.*, vol. 40, no. D1, pp. D290–D301, 2012.

[16] A. Cuff, I. Sillitoe, T. Lewis, O. Redfern, R. Garratt, J. Thornton, and C. Orengo, "The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies." *Nucleic Acids Res.*, vol. 37, pp. 310–314, 2009.

[17] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc. Natl Acad. Sci, USA*, vol. 93, no. 1, pp. 13–20, 1996.

[18] H. Shanahan and J. Thornton, "Amino acid architecture and the distribution of polar atoms on the surfaces of proteins," *Biopolymers*, vol. 78, no. 6, pp. 318–328, 2005.

[19] H. Zhu, F. Domingues, I. Sommer, and T. Lengauer, "NOXclass: Prediction of protein-protein interaction types," *BMC Bioinformatics*, vol. 7, no. 27, 2006, doi:10.1186/1471-2105-7-27.

[20] M. C. Lawrence and P. M. Colman, "Shape complementarity at protein/protein interfaces," *J. Mol Biol*, vol. 234, no. 4, pp. 946–950, 1993.

[21] J. Young, "A role for surface hydrophobicity in protein protein recognition," *Protein Sci*, vol. 3, pp. 717–729, 1994.

[22] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal, "Residue frequencies and pairing preferences at protein-protein interfaces," *Proteins*, vol. 43, no. 2, pp. 89–102, 2001.

[23] J. Mintseris and Z. Weng, "Atomic contact vectors in protein-protein recognition," *PROTEINS: Structure, Function and Genetics*, vol. 53, pp. 629–639, 2003.

[24] L. Rueda, S. Banerjee, Md. Aziz, and M. Raza, "Protein-protein interaction prediction using desolvation energies and interface properties," *Proceedings of the 2nd. IEEE International Conference on Bioinformatics & Biomedicine (BIBM 2010)*, pp. 17–22, 2010.

[25] L. Rueda, C. Garate, Banerjee, and Md. Aziz, "Biological protein-protein interaction prediction using binding free energies and linear dimensionality reduction." *Proceedings of the 5th. IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2010)*, pp. 383–394, 2010.

[26] Md. Aziz, M. Maleki, L. Rueda, M. Raza, and S. Banerjee, "Prediction of biological protein-protein interactions using atom-type and amino acid properties," *Proteomics 2011*, vol. 11, no. 19, pp. 3802–10, 2011.

[27] G. Vasudev and L. Rueda, "A model to predict and analyze protein-protein interaction types using electrostatic energies," in *5th IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012)*, 2012, pp. 543–547.

[28] S. H. Park, J. Reyes, D. Gilbert, J. W. Kim, and S. Kim, "Prediction of protein-protein interaction types using association rule based classification," *BMC Bioinformatics*, vol. 10, no. 36, 2009, doi:10.1186/1471-2105-10-36.

[29] M. Maleki, Md. Aziz, and L. Rueda, "Analysis of Obligate and Non-obligate Complexes using Desolvation Energies in Domain-domain Interactions," in *10th International Workshop on Data Mining in Bioinformatics (BIOKDD 2011) in conjunction with ACM SIGKDD 2011*, San Diego, USA, August 2011, pp. 21–26.

[30] J. Mintseris and Z. Weng, "Structure, function, and evolution of transient and obligate protein-protein interactions," *Proc Natl Acad Sci, USA*, vol. 102, no. 31, pp. 10 930–10 935, 2005.

[31] C. Camacho and C. Zhang, "FastContact: rapid estimate of contact and binding free energies," *Bioinformatics*, vol. 21, no. 10, pp. 2534–2536, 2005.

[32] C. Zhang, G. Vasmatzis, J. L.Cornette, and C. DeLisi, "Determination of atomic desolvation energies from the structures of crystallized proteins," *J. Mol. Biol.*, vol. 267, pp. 707–726, 1997.

[33] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[34] P. Pudil and F. F. et al., "Floating search methods for feature selection with nonmonotonic criterion functions," in *Proc. of the 12th international Conference on Pattern Recognition*, vol. 2, 1994, pp. 279–283.

[35] L. Rueda and M. Herrera, "Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space," *Pattern Recognition*, vol. 41, no. 10, pp. 3138–3152, 2008.

[36] R. O. Duda, D. G. Stork, and P. E. Hart, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[37] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[38] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Elsevier Academic Press, 2008.

Chapter 6

Using Structural Domains to Predict Obligate and Non-obligate Protein-protein Interactions

6.1 Introduction

Due to the fundamental role in many essential biological processes, the identification of protein-protein interactions (PPIs) is a key research topic. Prediction of PPIs has been studied using various computational approaches and from many different perspectives. Prediction of interfaces or interactions between subunits in large molecules involves analysis of patches, sites, amino acids, or even specific atoms, while the physicochemical and geometric arrangement of subunits in protein complexes is best known as docking. An important problem that has recently drawn the attention of the research community is the prediction of "when" the interactions will occur – this is mostly studied at the level of protein interaction networks. Another important problem surrounding PPIs is the identification of different types of complexes, which are characterized by properties such as similarities between subunits (homo/hetero-oligomers), number of subunits involved in the interaction (dimers, trimers, etc.), duration of the interaction (transient vs. permanent), stability of the

interaction (non-obligate vs. obligate), among others; we focus on the latter problem.

Obligate interactions are usually considered to be permanent, while non-obligate interactions can be either permanent or transient [1]. Non-obligate and transient interactions are more difficult to study and understand due to their instability and short life, while obligate and permanent interactions are more stable [2]. For this reason, it is important to be able to distinguish between obligate and non-obligate complexes.

Some studies in PPIs consider the analysis of a wide range of parameters for predicting obligate and non-obligate complexes, including analysis of solvent accessibility [3, 4], geometry [5], hydrophobicity [6, 7], sequence-based features [8] and desolvation energy [9–11]. In this study, we use desolvation energies, which have already been shown to be very efficient for PPI prediction [9, 10].

Recent studies of PPIs focus on employing domain knowledge to predict the protein-protein interactions [12–15]. The motivation behind these approaches is that: (i) domains are the minimal and fundamental units of proteins, which have a clear biological role and act as basic functional units within proteins [16]; (ii) it has been claimed that only a few highly conserved residues are crucial for protein interactions [17, 18]; (iii) it has been shown that most domains and domain interactions are evolutionary conserved, and consequently, proteins will interact if a domain in one protein interacts with a domain in the other protein [19, 20]. In [17], interactions between residues were used for finding obligate and non-obligate residue contacts of PPIs. The study concluded that non-obligate interfaces occupy less than 2% of the area of the domain surfaces, while the area occupied by obligate interfaces is between 0–6%. In [18], the interface of 750 transient DDIs (interactions between domains that are part of different proteins) and 2,000 obligate DDIs were studied. The interactions between domains of one amino acid chain were analyzed to obtain a better

understanding of molecular recognition and identify frequent amino acids in the interfaces and on the surfaces of PPIs. Also, in [21], the domain information from protein complexes was used to predict four different types of PPIs including transient enzyme inhibitor/non enzyme inhibitor, and permanent homo/hetero obligate complexes. Thus, the physical interaction between proteins can be better analyzed in terms of the interaction between their structural domains. There are a number of domain family resources that can be applied for this purpose such as Pfam [22] and CATH [23].

In this paper, we propose a domain-based approach which uses CATH- Class, Architecture, Topology and Homologous superfamily- domain information to predict obligate and non-obligate protein-protein interactions. Desolvation energies of amino acid pairs present in the interface of DDIs as well as desolvation energies of all amino acid pairs present in the interface of interacting complexes are used in the prediction. The prediction approach relies on two state-of-the-art classification techniques of linear dimensionality reduction (LDR) [24] and support vector machines (SVM) [25]. Ten-fold cross validation of the proposed scheme on two well-known datasets of [4] and [26] shows that: (i) DDI features of the first three levels of CATH, especially level 2, are more powerful than features of other levels in predicting obligate and non-obligate complexes; (ii) prediction accuracies using DDI features for levels 5 to 8 of CATH are lower than those of features of upper levels; (iii) although the prediction accuracies achieved by considering amino acid pairs present in the interacting domains instead of all interacting amino acid pairs of two chains for both LDR and SVM are relatively low, they are still acceptable and provide additional information about the specific domains.

We have also performed a visual and numerical analysis on the DDIs present in obligate and non-obligate interactions of levels 1 to 3 of CATH domains. The analysis shows that

homo-DDIs are mostly present in obligate interactions. In addition, by grouping the DDIs into three main groups of more obligate, more non-obligate and non-interaction groups considering the distribution of DDIs, some important DDI features for the prediction of complex types can be easily found.

6.2 Prediction Methods

6.2.1 Linear Dimensionality Reduction

One of the approaches we use for prediction is LDR. The basic idea of LDR is to represent an object of dimension n as a lower-dimensional vector of dimension d, achieving this by performing a linear transformation. We consider two classes, ω_1 and ω_2, represented by two normally distributed random vectors $\mathbf{x}_1 \sim N(\mathbf{m}_1, \mathbf{S}_1)$ and $\mathbf{x}_2 \sim N(\mathbf{m}_2, \mathbf{S}_2)$, respectively, with p_1 and p_2 the *a priori* probabilities. After the LDR is applied, two new random vectors $\mathbf{y}_1 = \mathbf{A}\mathbf{x}_1$ and $\mathbf{y}_2 = \mathbf{A}\mathbf{x}_2$, where $\mathbf{y}_1 \sim N(\mathbf{A}\mathbf{m}_1; \mathbf{A}\mathbf{S}_1\mathbf{A}^t)$ and $\mathbf{y}_2 \sim N(\mathbf{A}\mathbf{m}_2; \mathbf{A}\mathbf{S}_2\mathbf{A}^t)$ with $\mathbf{m}_i$ and $\mathbf{S}_i$ being the mean vectors and covariance matrices in the original space, respectively. The aim of LDR is to find a linear transformation matrix $\mathbf{A}$ in such a way that the new classes ($\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$) are as separable as possible. Let $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$ and $\mathbf{S}_E = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$ be the within-class and between-class scatter matrices respectively. Various criteria have been proposed to measure this separability [24]. We consider the following two LDR methods:

(a) The heteroscedastic discriminant analysis (HDA) approach [24], which aims to obtain the matrix $\mathbf{A}$ that maximizes the following function and which is optimized via eigenvalue decomposition:

$$J_{HDA}(\mathbf{A}) = tr\left\{(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\left[\mathbf{A}\mathbf{S}_E\mathbf{A}^t - \mathbf{A}\mathbf{S}_W^{\frac{1}{2}}\frac{p_1\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}})+p_2\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1p_2}\mathbf{S}_W^{\frac{1}{2}}\mathbf{A}^t\right]\right\} \quad (6.1)$$

(b) The Chernoff discriminant analysis (CDA) approach [24], which aims to maximize the following function and which is maximized via a gradient-based algorithm:

$$\begin{aligned} J_{CDA}(\mathbf{A}) = tr\{&p_1p_2\mathbf{A}\mathbf{S}_E\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1} \\ &+\log(\mathbf{A}\mathbf{S}_W\mathbf{A}^t) - p_1\log(\mathbf{A}\mathbf{S}_1\mathbf{A}^t) - p_2\log(\mathbf{A}\mathbf{S}_2\mathbf{A}^t)\}. \end{aligned} \quad (6.2)$$

In order to classify each complex, first a linear algebraic operation $\mathbf{y} = \mathbf{A}\mathbf{x}$ is applied to the n-dimensional vector, obtaining $\mathbf{y}$, a d-dimensional vector, where d is ideally much smaller than n. The linear transformation matrix $\mathbf{A}$ corresponds to the one obtained by one of the LDR methods, namely HDA or CDA. The resulting vector $\mathbf{y}$ is then passed through a quadratic Bayesian (QB) classifier [24], which is the optimal classifier for normal distributions. For additional tests, a linear Bayesian (LB) classifier is considered by deriving a Bayesian classifier with a common covariance matrix: $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$.

6.2.2 Support Vector Machines

SVMs are well known machine learning techniques used for classification, regression and other tasks. The aim of the SVM is to find the support vectors (most difficult vectors to be classified) to derive a decision boundary that separates the feature space into two regions. While a more detailed description of the SVM can be found in standard machine learning textbooks (cf. [27]), for the sake of clarity, we provide a brief description below.

Let $\{\mathbf{x}_i\}$ where $i = 1,2,3.....n$, be the feature vectors of the training dataset $\mathbf{X}$. These vectors belong to one of two classes ω_1 and ω_2, which are assumed to be linearly separable.

The goal of the SVM is to find a hyperplane that classifies all the training vectors as follows:

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0 \tag{6.3}$$

This kind of hyperplane is not unique. The SVM chooses the hyperplane that leaves the maximum margin from that hyperplane to the *support vectors*.

The distance from a point to a hyperplane is given by:

$$z = \frac{|g(\mathbf{x})|}{||\mathbf{w}||} \tag{6.4}$$

If for each $\mathbf{x}_i$ we denote the corresponding class label by y_i (+1 for ω_1, -1 for ω_2), the SVM finds the best hyperplane by computing the parameters $\mathbf{w}$ and w_0 of the hyperplane so that the following is minimized:

$$J(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2 \tag{6.5}$$

subject to

$$y_i(\mathbf{w}^t x_i + w_0) \geq 1, i = 1, 2, ...n \tag{6.6}$$

The classification by using the SVM is usually inefficient when using a linear classifier, because in general, the data is not linearly separable, and hence the use of kernels is crucial in mapping the data onto a higher dimensional space in which the classification is more efficient. The effectiveness of the SVM depends on the selection of the kernel, the selection parameters and the soft margin [28]. There are a number of different kernels that can be used in SVMs. In our model, we use polynomial, radial basis function (RBF) and sigmoid. In addition to this, these kernels require some parameters, which are discussed in the Results

section.

6.3 Datasets and Prediction Properties

Two pre-classified datasets of obligate and non-obligate protein complexes were obtained from the studies of Zhu et al. [4], and Mintseris and Weng [26], which we refer to as the MW and ZH datasets respectively. The first dataset contains 75 permanent (obligate) and 62 non-obligate interactions, while the second dataset contains 115 obligate and 212 non-obligate interactions.

6.3.1 Desolvation Energy

Different approaches have been developed to group different types of proteins, based on their different properties. Among them, desolvation energies have been found to be very efficient for prediction [9]. Desolvation energy is defined as knowledge-based contact potential (accounting for hydrophobic interactions), self-energy change upon desolvation of charged and polar atom groups, and side-chain entropy loss. As in [29], the binding free energy ΔG_{bind} is defined a follows:

$$\Delta G_{bind} = \Delta E_{elec} + \Delta G_{des}, \tag{6.7}$$

where ΔE_{elec} is the total electrostatic energy and ΔG_{des} is the total desolvation energy. For a protein, ΔG_{des} is defined as follows:

$$g(r)\Sigma\Sigma e_{ij}. \tag{6.8}$$

If we consider the interaction between the i^{th} atom of a ligand and the j^{th} atom of a receptor, then e_{ij} is the atomic contact potential (ACP) [30] between them and $g(r)$ is a smooth function based on their distance. For simplicity, we consider the smooth function to be linear. We also consider the criteria that for a successful interaction, the atoms should be within 7 $\AA$ distance. Between 5 and 7 $\AA$, the value of $g(r)$ varies from 0 to 1 based on a smooth function. For atoms that are less than 5 $\AA$ apart, the value of $g(r)$ is 1 [29].

6.3.2 Domain-based Properties

In this study, we consider CATH domains [23]. The CATH database is organized in a hierarchical fashion, which can be visualized as a tree with levels numbered from 1 to 8, thereafter referred to as L1 to L8. Domains at upper levels of the tree represent more general classes of structure than those at lower levels. For example, domains at level 1 represent alpha helices, beta sheets, and combinations thereof, whereas those at level 2 represent more specific structures such as beta barrels and rolls. Domains at level 3 are even more specific, and so on.

To extract domain-based properties, we first collected the structural files of each complex in our datasets from the Protein Data Bank (PDB) [31]. Then, we collected the domain information of each complex from the CATH[1] database and added the collected domain information to each atom present in the chain. Complexes that did not have domain information in at least one of their subunits were discarded. A summary of the number of obligate and non-obligate complexes before and after filtering these complexes by considering CATH domains is given in Table 6.1, where MW-CATH and ZH-CATH refer to as the MW and ZH datasets after removing the complexes without CATH domains in their

[1] www.cathdb.info

Table 6.1: Datasets and their number of complexes used in this study.

Dataset Name	# Complexes	# Obligate	# Non-obligate
MW	327	115	212
MW-CATH	287	106	181
ZH	137	75	62
ZH-CATH	127	72	55

interaction.

After identifying all the unique domains present in the interface of at least one complex in the datasets, the desolvation energies for all pairs of domains (DDIs) were calculated using Eq. (6.8). For each ligand-receptor pair, if we found any duplicate DDIs during calculation we simply computed the cumulative desolvation energy across all occurrences of that DDI. A domain is considered as being in the interface if it has at least one residue interacting with a domain in the other chain.

Since the CATH database is organized in a hierarchical scheme, we created a separate dataset of feature vectors for each level of the hierarchy. Each of these datasets were used for classification separately, in order to observe the prediction power of a specific level in the CATH hierarchy. To speed up computations, after calculating the desolvation energies for all DDIs in level 8, for each DDI in higher levels the desolvation energy was calculated by taking the sum of the desolvation energies of the corresponding DDIs at the next lowest level. For each node in the CATH tree, the set of DDIs associated with it are completely disjoint (with the exception of reflexive pairs, which have been accounted for in postprocessing). Thus, when we combine the desolvation energies of DDIs from one level to find the desolvation energies for the respective parent nodes at the next highest level, we do not introduce any redundancy into the corresponding features.

Since there are a large number of possible DDIs $(0.5[n(n+1)]+n$, where n is the

Table 6.2: Subsets of features used in this study.

Subset Name	# Domains	# Non-zero DDIs
(a) The Mintseris and Weng dataset[26]		
MW-CATH-L1	4	9
MW-CATH-L2	26	106
MW-CATH-L3	237	342
MW-CATH-L4	386	403
MW-CATH-L5	740	563
MW-CATH-L6	803	573
MW-CATH-L7	864	576
MW-CATH-L8	899	576
(b) The Zhu et al. dataset [4]		
ZH-CATH-L1	4	9
ZH-CATH-L2	24	67
ZH-CATH-L3	136	154
ZH-CATH-L4	186	180
ZH-CATH-L5	272	228
ZH-CATH-L6	278	230
ZH-CATH-L7	287	230
ZH-CATH-L8	301	236

number of unique domains), after pre-processing the datasets we removed all zero-columns. Zero-columns represent DDIs that were not present in any complex. A summary of the number of features after removing zero-columns for each subset of features is given in Table 6.2. The names of these subsets show the dataset name (MW or ZH) and the level of the domains in the CATH hierarchy (L1 to L8).

6.4 Results and Discussions

6.4.1 Experimental Settings

For the LDR schemes, four different classifiers were implemented and evaluated, namely the combinations of HDA and CDA, and QB and LB classifiers. Within 10-fold cross validation, reductions to dimensions $d = 1, \ldots, 20$ were performed, followed by QB and LB, and the maximum of the average classification accuracies for each classifier was recorded. Of these, the maximum for the four LDR schemes is reported.

The SVM was also trained with 10-fold cross validation for three kernels: RBF, polynomial and sigmoid. The training was carried out with the LIBSVM package [25]. A grid search was performed on the parameters gamma and C, both in the range $[2^{-20}..2^{20}]$ choosing the ones that give the maximum accuracy for a specific kernel. For the polynomial kernel, the degree of the polynomial was set to 2 and 3. For each subset of features the maximum average classification accuracy obtained from these three kernels is reported. The average accuracy for the 10 folds was computed as follows: $acc = (TP + TN)/n$, where TP and TN are the true positive (obligate) and true negative (non-obligate) counters respectively, and n is the total number of complexes.

In addition to domain-based features (Table 6.2), we have considered pairs of amino acids present in the interface of PPIs. For this, we computed 20^2 desolvation energy values for all pairs of atoms using Eq. (6.8), and accumulated the values for each pair of amino acids. Avoiding repeated pairs resulted in 210 different features (unique pairs of amino acids). We refer to these new subsets of features as the MW-aa for the MW dataset and ZH-aa for the ZH dataset.

6.4.2 Analysis of Prediction

The results of the SVM and LDR classifiers with amino acid and DDI-type features for the MW and ZH datasets are depicted in Table 6.3. The maximum accuracies for LDR and SVM variants are shown in the table.

For the subsets of features extracted from the MW dataset, the MW-aa subset is best classified with SVM, achieving an accuracy of 77.70%, while MW-CATH-L2 achieves the best performance of 77.35% – these two are almost identical. It is also clear that after DDIs of level 2, DDIs of levels 1 and 3 are more powerful for prediction than DDIs of other levels. Moreover, subsets of the last four levels of the CATH hierarchy (MW-CATH-L5 to MW-CATH-L8) with the same number of features yield approximately the same low prediction accuracy for both SVM and LDR. This could be due to the fact that for higher levels a large number of zero-features is generated, and in that case, the classification would rely only on the non-zero features.

Similarly, for the subsets of features extracted from the ZH dataset, it is observable that the best accuracy of 85.83% is achieved by SVM when using desolvation energies for amino acid type features (ZH-aa). Although it seems that ZH-CATH-L2 with 79.30% accuracy has less performance than ZH-aa, when we consider the number of features of these two subsets (67 for ZH-CATH-L2 and 210 for ZH-aa) this decrease is acceptable. Moreover, from the results, it is clear that only DDIs of the first few levels of CATH are good enough to predict obligate and non-obligate complexes and other levels, from levels 4 to 8, could be ignored because of their low prediction performance.

Generally, it can be concluded that for both MW and ZH datasets and all subsets of features that: (a) amino acid type features yield higher accuracies than DDI type features for both LDR and SVM; (b) domain-based features related to L2 and L1 of CATH are more

Table 6.3: Prediction results for LDR and SVM classifiers for the MW and ZH datasets.

Subset Name	# Features	LDR	SVM
(a) The Mintseris and Weng dataset[26]			
MW-aa	210	75.17	**77.70**
MW-CATH-L1	9	67.59	72.82
MW-CATH-L2	106	71.03	**77.35**
MW-CATH-L3	342	66.9	71.25
MW-CATH-L4	403	64.83	70.73
MW-CATH-L5	536	64.48	70.38
MW-CATH-L6	573	64.14	69.69
MW-CATH-L7	576	63.79	69.69
MW-CATH-L8	576	63.45	69.69
(b) The Zhu et al. dataset [4]			
ZH-aa	210	73.23	**85.83**
ZH-CATH-L1	9	68.5	74.80
ZH-CATH-L2	67	73.23	**79.30**
ZH-CATH-L3	154	64.57	69.76
ZH-CATH-L4	180	62.99	69.69
ZH-CATH-L5	228	59.84	67.93
ZH-CATH-L6	230	57.48	67.72
ZH-CATH-L7	230	57.48	67.72
ZH-CATH-L8	236	57.48	67.72

powerful than the features of other levels in the prediction of obligate and non-obligate complexes – this difference is only marginal for the MW dataset; (c) the prediction performance using DDI features for CATH levels from 5 to 8 are almost the same and could be safely ignored; (d) SVM with optimized parameters is the most powerful predictor for all subsets of features; (e) SVM and LDR classifiers, however, show a similar trend. For both classifiers, DDIs of L1 are better than those of L3, while DDIs of L2 are much better than those of both L1 and L3.

Table 6.4: A summary of the number of CATH DDIs of level 1 present in the ZH and MW datasets.

		ZH-CATH-L1			MW-CATH-L1		
Domain1	Domain2	# Ob.	# Non-ob.	Total	# Ob.	# Non-ob.	Total
c1	*c1*	17	8	25	29	18	47
c1	*c2*	6	4	10	18	41	59
c1	*c3*	12	18	30	63	59	122
c1	*c4*	1	1	2	1	7	8
c2	*c2*	15	18	33	28	92	120
c2	*c3*	7	22	29	27	77	104
c2	*c4*	2	2	4	10	6	16
c3	*c3*	101	25	126	88	70	158
c3	*c4*	0	0	0	4	8	12
c4	*c4*	1	0	1	0	0	0

6.4.3 Analysis of DDIs

As discussed earlier, DDI features related to levels 1 to 3 of CATH are more powerful than those of other levels when used for prediction of obligate and non-obligate complexes. In level 1, the "class" of each complex is defined. Four classes of CATH, which are determined based on the secondary structure composition of the complexes, are mainly-alpha (*c1*), mainly-beta (*c2*), alpha-beta (*c3*) and secondary structure content (*c4*) [23]. A summary of the number of DDIs present in both the ZH and MW datasets, categorized by complex type, obligate and non-obligate, is shown in Table 6.4. There are ten unique DDIs from these four classes of domains in CATH level 1. By observing the table, it is clear that most of the DDIs are between domains of *c3* and other classes in which *c3:c3* has the highest rank among all levels. However, domains of *c4* have the least number of interactions with the domains of other levels. In addition, the number of obligate DDIs is larger than the number of non-obligate DDIs, when considering the interactions of homo-DDIs such as *c1:c1* and *c3:c3*.

To provide a visual insight of the distribution of DDIs present in the complexes of the MW and ZH datasets, a schematic view of the DDIs of levels 2 and 3 is shown in Figure 6.1. In each figure, DDIs that are obligate, non-obligate or common are shown as blue, red and pink dots respectively. In the plots, the domains have been grouped by category and then arbitrarily numbered within each category (the numbers correspond to the x and y axes).

In the figures related to level 2 (a and b), the horizontal and vertical green lines indicate the boundaries of classes in L1 of CATH. It is clear that there are fewer dots in the right column of both a and b and these dots show the interaction of domains of *c4* and domains of other classes. This indicates that DDIs related to *c4* are not important and could be ignored for achieving a faster, yet still accurate, prediction. In contrast, DDIs in the diagonal, especially *c3:c3* and *c2:c2*, have the most number domain interactions. Moreover, there are fewer common DDIs in *c1:c1* and *c1:c2* in comparison to other DDIs.

In the figures for level 3 (c and d), the boundaries of level 1 are shown with blue lines while level 2 boundaries are shown with green lines. Some concentration of blue or red dots can be seen in some parts of the plots in both datasets. This suggests that some domains are more likely to occur in obligate, while others in non-obligate complexes. This is more noticeable along the diagonal, in which most of the obligate DDIs (blue dots) are concentrated, which indicates that the largest number of homo-domain pairs are in obligate complexes. By considering this distribution, all DDIs can be divided into three main groups of more obligate (columns 1-20, 2-30 and 3-90), more non-obligate (columns 2-60 and 3-40) or non-interaction (columns 2-102 and 2-150). By analyzing this grouping scheme, some DDIs in level 3 can be considered less important (non-interaction group), while the others are more important (more obligate and more non-obligate groups) to predict obligate

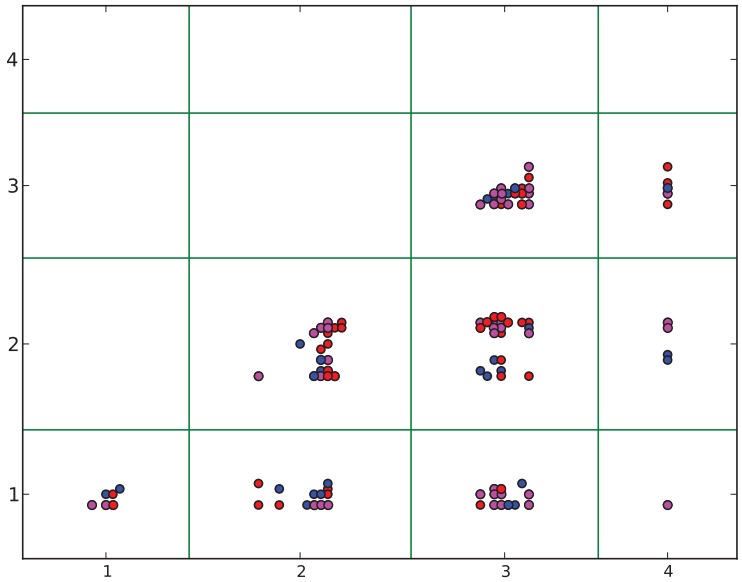
and non-obligate complexes.

6.5 Conclusion

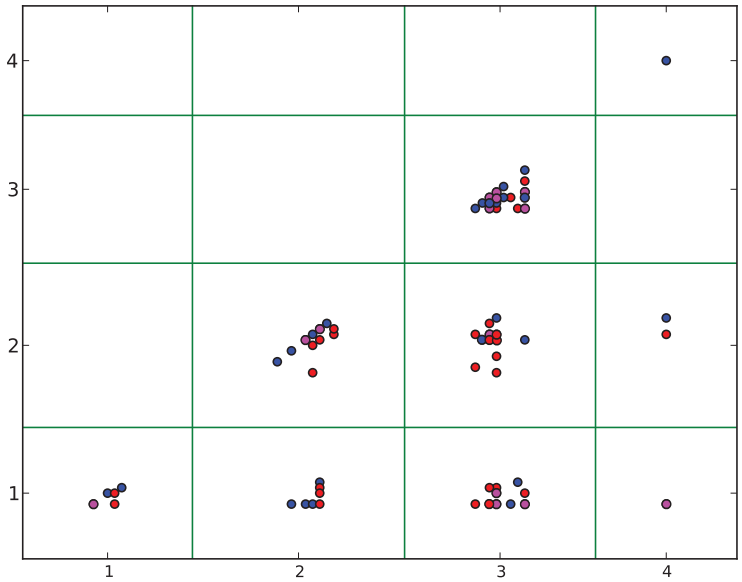
We have presented a structural, domain-based approach to predict obligate and non-obligate protein complexes. We have also investigated various interface properties of these interactions including amino acid type and DDI features for different levels of the CATH hierarchy. The classification is performed via LDR methods with heteroscedastic criteria and an SVM with RBF, polynomial, and sigmoid kernels.

The results for two well-known datasets of pre-classified complexes demonstrate that classification using DDI features from level 2 of the CATH hierarchy achieves better performance than using DDIs from levels 1 and 3, and much better performance than using DDIs of levels 5 to 8 for both LDR and SVM classifiers. This suggests that DDIs from level 2 of CATH are more powerful and discriminative than DDIs from other levels for predicting obligate and non-obligate PPIs. Also, the SVM classifier with optimized parameters outperforms the LDR methods for all subsets of features.

Furthermore, a visual and numerical analysis of DDIs shows that: (i) in both datasets, most homo-domain pairs are in obligate interactions; (ii) while there are fewer interactions between domains of *c4* and domains of other classes, most of the interactions are between domains of *c3* and domains of other classes.

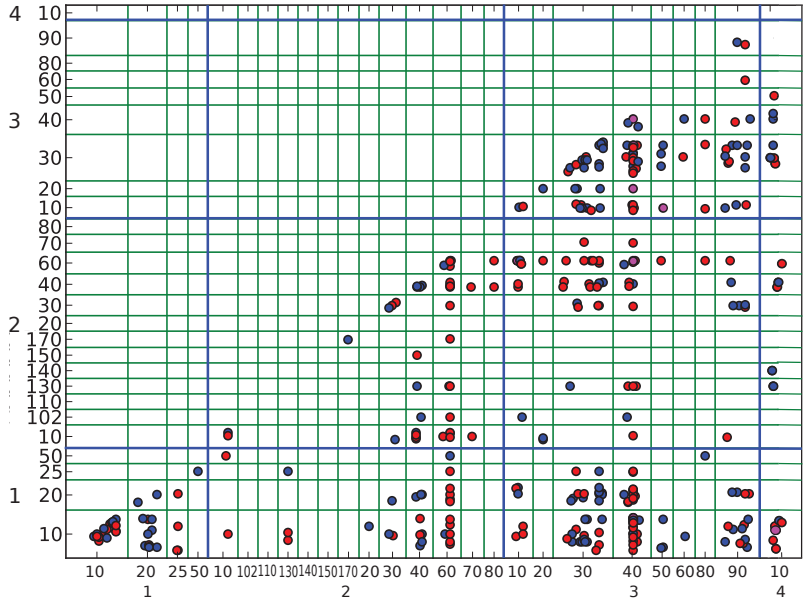


(a) MW-CATH-L2

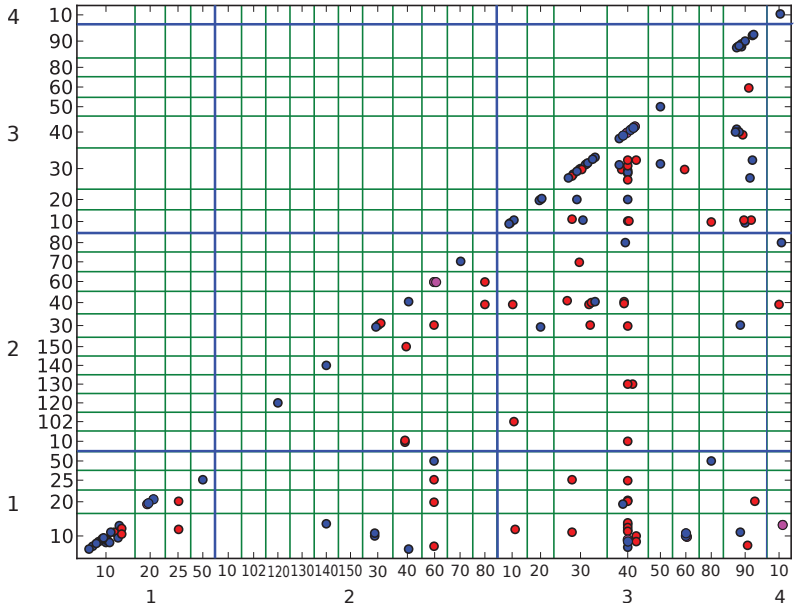


(b) ZH-CATH-L2

Figure 6.1: Schematic view of levels 2 and 3 of CATH DDIs present in the MW and ZH datasets.



(c) MW-CATH-L3



(d) ZH-CATH-L3

Figure 6.1: Schematic view of levels 2 and 3 of CATH DDIs present in the MW and ZH datasets.

Bibliography

[1] I. Nooren and J. Thornton, "Diversity of protein-protein interactions," *EMBO Journal*, vol. 22, no. 14, pp. 3846–3892, 2003.

[2] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc. Natl Acad. Sci, USA*, vol. 93, no. 1, pp. 13–20, 1996.

[3] H. Shanahan and J. Thornton, "Amino acid architecture and the distribution of polar atoms on the surfaces of proteins," *Biopolymers*, vol. 78, no. 6, pp. 318–328, 2005.

[4] H. Zhu, F. Domingues, I. Sommer, and T. Lengauer, "NOXclass: Prediction of protein-protein interaction types," *BMC Bioinformatics*, vol. 7, no. 27, 2006, doi:10.1186/1471-2105-7-27.

[5] M. C. Lawrence and P. M. Colman, "Shape complementarity at protein/protein interfaces," *J. Mol Biol*, vol. 234, no. 4, pp. 946–950, 1993.

[6] J. Young, "A role for surface hydrophobicity in protein protein recognition," *Protein Sci*, vol. 3, pp. 717–729, 1994.

[7] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal, "Residue frequencies and pairing preferences at protein-protein interfaces," *Proteins*, vol. 43, no. 2, pp. 89–102, 2001.

[8] J. Mintseris and Z. Weng, "Atomic contact vectors in protein-protein recognition," *PROTEINS: Structure, Function and Genetics*, vol. 53, pp. 629–639, 2003.

[9] L. Rueda, S. Banerjee, M. M. Aziz, and M. Raza, "Protein-protein interaction prediction using desolvation energies and interface properties," *Proceedings of the 2nd. IEEE International Conference on Bioinformatics & Biomedicine (BIBM 2010)*, pp. 17–22, 2010.

[10] L. Rueda, C. Garate, Banerjee, and M. M. Aziz, "Biological protein-protein interaction prediction using binding free energies and linear dimensionality reduction." *Proceedings of the 5th. IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2010)*, pp. 383–394, 2010.

[11] M. M. Aziz, M. Maleki, L. Rueda, M. Raza, and S. Banerjee, "Prediction of biological protein-protein interactions using atom-type and amino acid properties," *Proteomics 2011*, vol. 11, no. 19, pp. 3802–10, 2011.

[12] N. Zaki, "Protein-protein interaction prediction using homology and inter-domain linker region information." *Advances in Electrical Engineering and Computational Science, Springer*, vol. 39, pp. 635–645, 2009.

[13] N. Zaki and P. C. S. Lazarova-Molnar, W. El-Hajj, "Protein-protein interaction based on pairwise similarity." *BMC Bioinformatics*, vol. 10, no. 1, 2009.

[14] M. Singhal and H. Resat, "A domain-based approach to predict protein-protein interactions." *BMC Bioinformatics*, vol. 8, no. 1, 2007.

[15] T. Akutsu and M. Hayashida, "Domain-based prediction and analysis of protein-protein interactions." *Biological data mining in protein interaction networks, Medical Information Science Reference, chapter 3*, pp. 29–44, 2009.

[16] L. Chen, R. Wang, and X. Zhang, *Biomolecular Networks: Methods and Applications in Systems Biology*. John Wiley and Sons, 2009.

[17] S. De, O. Krishnadev, N. Srinivasan, and N. Rekha, "Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different." *BMC Structural Biology*, vol. 5, no. 15, 2005.

[18] J. V. Eichborn, S. Gnther, and R. Preissner, "Structural features and evolution of protein-protein interactions." *Intenational Conference of Genome Informatics.*, vol. 22, pp. 1–10, 2010.

[19] D. Caffrey, S. Somaroo, J. Hughes, J. Mintseris, and E. Huang, "Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?" *Protein Sci.*, vol. 13, no. 1, pp. 190–202, 2004.

[20] J. Park and D. Bolser, "Conservation of protein interaction network in evolution." *Genome Inform*, vol. 12, pp. 135–140, 2001.

[21] S. H. Park, J. Reyes, D. Gilbert, J. W. Kim, and S. Kim, "Prediction of protein-protein interaction types using association rule based classification," *BMC Bioinformatics*, vol. 10, no. 36, 2009, doi:10.1186/1471-2105-10-36.

[22] R. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. Pollington, O. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. Sonnhammer, S. Eddy, and A. Bateman, "The Pfam protein families database." *Nucleic Acids Res.*, vol. 38, pp. 211–222, 2010.

[23] A. Cuff, I. Sillitoe, T. Lewis, O. Redfern, R. Garratt, J. Thornton, and C. Orengo, "The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies." *Nucleic Acids Res.*, vol. 37, pp. 310–314, 2009.

[24] L. Rueda and M. Herrera, "Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space," *Pattern Recognition*, vol. 41, no. 10, pp. 3138–3152, 2008.

[25] C. L. C. Chang, "LIBSVM: a library for support vector machines," last date accessed: May 31, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.pdf

[26] J. Mintseris and Z. Weng, "Structure, function, and evolution of transient and obligate protein-protein interactions," *Proc Natl Acad Sci, USA*, vol. 102, no. 31, pp. 10 930–10 935, 2005.

[27] S. Abe, *Support Vector Machines for Pattern Classification*. Springer, 2005.

[28] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York, NY: John Wiley and Sons, Inc., 2000.

[29] C. Camacho and C. Zhang, "FastContact: rapid estimate of contact and binding free energies," *Bioinformatics*, vol. 21, no. 10, pp. 2534–2536, 2005.

[30] C. Zhang, G. Vasmatzis, J. L.Cornette, and C. DeLisi, "Determination of atomic desolvation energies from the structures of crystallized proteins," *J. Mol. Biol.*, vol. 267, pp. 707–726, 1997.

[31] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

PART 3

DOMAIN-BASED FEATURES-PFAM

Chapter 7: M. Maleki, Md. Aziz, L. Rueda, "Analysis of Obligate and Non-obligate Complexes using Desolvation Energies in Domain-domain Interactions," in 10th International Workshop on Data Mining in Bioinformatics (BIOKDD2011) in conjunction with ACM SIGKDD 2011, San Diego, USA, Aug. 2011.

Chapter 8: M. Maleki, M. Dezfulian, W. Crosby, L. Rueda, "Computational Analysis of the Stability of SCF Ligases Employing Domain Information," in 5th ACM Conference on Bioinformatics, Computational Biology and Health Informatics (ACMBCB), CA, 2014. (submitted)

Chapter 7

Analysis of Obligate and Non-obligate Complexes using Desolvation Energies in Domain-domain Interactions

7.1 Introduction

Protein interactions are important in many essential biological processes in living cells, including signal transduction, transport, cellular motion and gene regulation. As a consequence of this, the identification of protein-protein interactions (PPIs) is a key topic in life science research. Prediction of PPIs has been studied mostly using computational approaches and from many different perspectives. Prediction of interfaces (interactions between subunits) in different molecules includes analysis of patches, sites, amino acids, or even specific atoms. The physicochemical and geometric arrangement of subunits in protein complexes is best known as docking. An important aspect that has recently drawn the attention of the research community is to predict "when" the interactions will occur – this is mostly studied at the protein interaction network level. Another important aspect in studying PPIs is the identification of different types of complexes, including similarities between subunits (homo/hetero-oligomers), number of subunits involved in the interaction

(dimers, trimers, etc.), duration of the interaction (transient vs. permanent), stability of the interaction (non-obligate vs. obligate), among others; we focus on the latter problem.

Obligate interactions are usually considered as permanent, while non-obligate interaction can be either permanent or transient [1]. Non-obligate and transient interactions are more difficult to study and understand due to their instability and short life, while obligate and permanent interactions last for a longer period of time, and hence are more stable [2]. For these reasons, an important problem is to distinguish between obligate and non-obligate complexes. To study the behavior of obligate and non-obligate interactions, in [3], it was shown that non-obligate complexes are rich in aromatic residues and arginine, while depleted in other charged residues. The study of [4] suggested that mobility differences of amino acids are more significant for obligate and large interface complexes than for transient and medium-sized ones.

Some studies in PPI consider the analysis of a wide range of parameters, including desolvation energies, amino acid composition, conservation, electrostatic energies, and hydrophobicity for predicting obligate and non-obligate complexes. In [1], a classification of obligate and non-obligate interactions was proposed where interactions are classified based on the lifetime of the complex. In [5], three different types of interactions were studied, namely crystal packing, obligate and non-obligate interactions. That study was based on using solvent accessible surface area, conservation scores, and the shapes of the interfaces. After classifying obligate and transient protein interactions based on 300 different interface attributes in [6], the difference in molecular weight between interacting chains was reported as the best single feature to distinguish transient from obligate interactions. Based on their results, interactions with the same molecular weight or large interfaces are obligate.

Different studies have claimed that only a few highly conserved residues are crucial for

protein interactions [7, 8]. Moreover, it has been shown that physical interactions between proteins are mostly controlled by their domains, as a domain is often the minimal and fundamental module corresponding to a biochemical function [7, 8]. Thus, in previous studies, the physical interaction between proteins is analyzed in terms of the interaction between residues of their structural domains. For example, in [7], interactions between residues were used for finding obligate and non-obligate residue contacts of PPIs. That study concluded that non-obligate interfaces occupy less than 2% of the area of the domain surfaces, while the number of obligatory interfaces is between 0–6%. In [8], the interface of 750 transient DDIs, interactions between domains that are part of different proteins, and 2,000 obligate interactions were studied. The interactions between domains of one amino acid chain were analyzed to obtain a better understanding of molecular recognition and identify frequent amino acids in the interfaces and on the surfaces of PPIs. Also, in [9], the domain information from protein complexes was used to predict four different types of PPIs including transient enzyme inhibitor/non enzyme inhibitor and permanent homo/hetero obligate complexes.

In a recent work [10], an approach to distinguish between obligate and non-obligate complexes has been proposed in which desolvation energies of amino acids and atoms present in the interfaces of PPIs are considered as the input features of the classifiers. The results of that classifier show that desolvation energies are better discriminant than solvent accessibility and conservation properties. In this paper, we present an analysis of PPIs that uses properties of DDIs present in the interface to predict obligate and non-obligate protein-protein interactions. Desolvation energies of atom and amino acid pairs present in the interface of DDIs as well as desolvation energies of all atom and amino acid pairs present in the interface of interacting complexes are used in the prediction. We have also performed

an analysis on the DDIs present in the two types of interactions. A visual analysis shows that unique pairs can be identified for both types of interactions, and highlight the presence of homo-DDIs in obligate interactions. The prediction approach resorts on two state-of-the-art classification techniques of linear dimensionality reduction (LDR) and support vector machines (SVM). Ten-fold cross validation of the proposed scheme on our binary-PPID dataset, which is an extended dataset that we compiled from two well-known datasets of [5] and [11], demonstrates that (a) using desolvation energies of atom type features are better than the features used in [5] for predicting obligate and non-obligate complexes, achieving 77.78% classification accuracy in comparison to 71.80% (b) atom type features are better than amino acid type features for prediction of these two types of complexes (c) although the prediction accuracies by considering atom and amino acid pairs present in the interacting domains instead of all interacting atom and amino acid pairs of two chains are low, they are still acceptable and provide additional information about the specific domains.

7.2 Materials and Methods

7.2.1 Dataset

We have compiled a new dataset by merging two existing, pre-classified datasets of protein complexes obtained from the studies of Zhu et al. [5], and Mintseris and Weng [11]. The former dataset contains 75 obligate and 62 non-obligate interactions while the latter contains 115 obligate and 212 transient interactions. There are 39 common interactions between these two datasets and hence the redundant complexes were removed. In addition, we carefully examined all the interactions and removed complexes with contradicting class labels. For example "*1eg9,A:B*" is classified as both obligate and non-obligate in [5] and

[11]. In total, seven complexes: *1eg9, 1hsa, 1i1a, 1raf, 1d09, 1jkj* and *1cqi*, showed this contradiction and were then removed from the new dataset. After this pre-processing stage, the new dataset resulted in 417 complexes from which 182 were obligate and 235 were non-obligate. In this study, each complex is considered as the interaction of two chains (two single sub-units). Since the dataset of [11] considers the interaction of two units in which each may contain more than one chain, e.g., "*1qfu,AB:HL*", all these complexes were converted to interactions between two single chains (binary interactions). For this, all binary interactions of each of the 93 multiple-chain complexes were identified, obtaining 289 interactions, and each of these was converted into a separate complex in the new dataset. For example, the multiple-chain of *1qfu* was transformed to four binary chains as follows: *A:H*, *A:L*, *B:H* and *B:L*. Another step involves taking the whole dataset of binary complexes and filtering non-interacting pairs. Using the interface definition of [8], complexes with interacting chains with less than five interface residues were removed. Two residues (from different chains) are considered to be interacting if at least one pair of atoms from these residues is 5Å or less apart from each other. This resulted in a dataset that contains 516 complexes, from which 303 are non-obligate and 213 are obligate binary interactions. In a final step, we collected the domains contained in each interacting chain from the Pfam database [12]. The complexes that do not have any domains in at least one of their subunits were discarded in the analysis. This resulted in our final dataset of 315 complexes, from which 146 are obligate complexes and 169 are non-obligate complexes - we call this dataset binary protein-protein interactions by considering domain definitions (binary-PPID). The PDB IDs of these complexes and the interacting chains are shown in Table 7.1.

Table 7.1: Binary-PPID dataset (146 obligate and 169 non-obligate binary complexes).

Obligate Complexes							
1a0f , A:B	1byk , A:B	1eex , A:B	1hcn , A:B	1jk0 , A:B	1li1 , A:C	1qbi , A:B	2hdh , A:B
1a6d , A:B	1c3o , A:B	1eex , A:G	1hfe , L:S	1jk8 , A:B	1li1 , B:C	1qdl , A:B	2hhm , A:B
1ahj , A:B	1c7n , A:B	1efv , A:B	1hgx , A:B	1jkm , A:B	1lti , C:G	1qfe , A:B	2kau , A:C
1aj8 , A:B	1ccw , A:B	1ep3 , A:B	1hjr , A:C	1jmx , A:G	1lti , C:H	1qfh , A:B	2kau , B:C
1ajs , A:B	1cmb , A:B	1ezv , D:H	1hr6 , A:B	1jnr , A:B	1lti , C:D	1qu7 , A:B	2min , A:B
1aq6 , A:B	1cnz , A:B	1ezv , C:F	1hss , A:B	1jro , A:B	1lti , C:F	1sgf , A:B	2mta , A:H
1b34 , A:B	1coz , A:B	1f6y , A:B	1ihf , A:B	1jwh , A:C	1lti , C:E	1sgf , A:Y	2nac , A:B
1b3a , A:B	1cpc , A:B	1ffu , A:C	1jb0 , B:E	1jwh , A:D	1luc , A:B	1spp , A:B	2pfl , A:B
1b4u , A:B	1dce , A:B	1ffv , A:B	1jb0 , B:E	1k3u , A:B	1mro , A:B	1spu , A:B	2utg , A:B
1b5e , A:B	1dii , A:C	1fm0 , D:E	1jb0 , B:D	1k8k , A:B	1mro , B:C	1trk , A:B	3gtu , A:B
1b7b , A:C	1dj7 , A:B	1g8k , A:B	1jb0 , B:D	1k8k , B:F	1mro , A:C	1vcb , A:B	3pce , A:M
1b7y , A:B	1dkf , A:B	1gka , A:B	1jb0 , A:E	1k8k , A:E	1msp , A:B	1vlt , A:B	3tmk , A:B
1b8j , A:B	1dm0 , A:D	1go3 , E:F	1jb0 , A:E	1k8k , C:F	1poi , A:B	1wgj , A:B	4rub , A:T
1b8m , A:B	1dm0 , A:E	1gpe , A:B	1jb0 , A:C	1k8k , D:F	1pp2 , L:R	1xso , A:B	
1b9m , A:B	1dor , A:B	1gpw , A:B	1jb0 , C:E	1kpe , A:B	1prc , C:H	1ypi , A:B	
1be3 , G:A	1dtw , A:B	1gux , A:B	1jb0 , B:C	1kqf , B:C	1prc , C:L	1ytf , C:D	
1bjn , A:B	1dxt , A:B	1h2a , L:S	1jb0 , A:D	1ktd , A:B	1prc , C:M	2aai , A:B	
1brm , A:B	1e8o , A:B	1h2r , L:S	1jb0 , A:D	1l7v , A:C	1qae , A:B	2ae2 , A:B	
1byf , A:B	1e9z , A:B	1h8e , A:D	1jb0 , C:D	1ld8 , A:B	1qax , A:B	2ahj , A:B	

Non-obligate Complexes							
1a14 , L:N	1bml , A:C	1eai , A:C	1fq1 , A:B	1i4d , B:D	1jsu , B:C	1n2c , B:E	2btc , E:I
1a14 , H:N	1buh , A:B	1eay , A:C	1fqj , A:C	1i4d , A:D	1jsu , A:C	1n2c , A:E	2btf , A:P
1a2k , B:C	1c1y , A:B	1ebd , A:C	1frv , A:B	1i7w , A:B	1jtg , A:B	1n2c , B:F	2mta , A:L
1a4y , A:B	1c4z , A:D	1ebd , B:C	1fss , A:B	1i85 , B:D	1jw9 , B:D	1nbf , A:D	2mta , A:C
1acb , E:I	1cc0 , A:E	1eer , A:B	1gaq , A:B	1i8l , A:C	1k5d , A:B	1nf5 , A:B	2mta , H:L
1agr , E:A	1cgi , E:I	1efu , A:B	1gcq , B:C	1ib1 , B:E	1kcg , A:C	1noc , A:B	2pcb , A:B
1akj , B:D	1cmx , A:B	1efx , A:D	1gh6 , A:B	1ib1 , A:E	1kcg , B:C	1pdk , A:B	2pcc , A:B
1akj , A:D	1cs4 , A:C	1eja , A:B	1gl1 , A:I	1icf , B:I	1kkl , A:H	1qbk , B:C	2prg , B:C
1ar1 , B:D	1cs4 , B:C	1es7 , C:B	1gla , F:G	1ijk , A:B	1kkl , C:H	1qgw , A:C	2sic , E:I
1avg , H:I	1cse , I:E	1es7 , A:B	1gp2 , A:B	1ijk , A:C	1kmi , Y:Z	1rlb , A:E	2tec , E:I
1avw , A:B	1cvs , A:C	1eth , A:B	1grn , A:B	1is8 , C:M	1kxp , A:D	1rlb , C:E	3hhr , A:B
1avx , A:B	1d4x , A:G	1euv , A:B	1gvn , A:B	1is8 , B:L	1kyo , O:W	1rlb , B:E	3sgb , E:I
1avz , B:C	1d5x , A:C	1evt , A:C	1gzs , A:B	1is8 , E:O	1lb1 , A:B	1rrp , A:B	3ygs , C:P
1awc , A:B	1de4 , C:A	1f02 , I:T	1h2k , A:S	1is8 , D:N	1lpb , A:B	1stf , E:I	4htc , H:I
1ay7 , A:B	1dev , A:B	1f34 , A:B	1h59 , A:B	1is8 , A:K	1m10 , A:B	1t7p , A:B	4sgb , E:I
1azz , A:D	1df9 , B:C	1f3v , A:B	1hlu , A:P	1is8 , D:O	1m1e , A:B	1tab , E:I	
1azz , A:D	1dfj , E:I	1f80 , A:E	1hwg , A:C	1is8 , A:L	1m4u , A:L	1tgs , I:Z	
1b6c , A:B	1doa , A:B	1fak , H:T	1hwg , A:B	1is8 , E:K	1mah , A:F	1toc , B:R	
1b9y , A:C	1du3 , A:D	1fg9 , B:C	1hzz , B:C	1is8 , C:N	1mbu , A:C	1uea , A:B	
1bdj , A:B	1du3 , A:F	1fg9 , A:C	1i2m , A:B	1is8 , B:M	1ml0 , A:D	1wq1 , G:R	
1bi8 , A:B	1dx5 , M:I	1fin , A:B	1i3o , D:E	1itb , A:B	1mr1 , A:D	1ycs , A:B	
1bkd , R:S	1e6e , A:B	1fle , E:I	1i3o , B:E	1jch , A:B	1n2c , A:F	1zbd , A:B	

7.2.2 Features

We use desolvation energies as the predicting properties, which are shown to be very efficient for prediction of obligate and non-obligate complexes [10]. Knowledge-based contact potential that accounts for hydrophobic interactions, self-energy change upon desolvation of charged and polar atom groups, and side-chain entropy loss compose the so-called binding-free energy. In [13], the total desolvation energy is defined as follows:

$$\Delta G_{des} = g(r)\Sigma\Sigma e_{ij}. \tag{7.1}$$

If we are considering the interaction between the i^{th} atom of a ligand and the j^{th} atom of a receptor then e_{ij} is the atomic contact potential (ACP) [14] between them, and $g(r)$ is a smooth function based on their distance. The value of $g(r)$ is 1 for atoms that are less than 5 Å apart [13]. For simplicity, we consider the smooth function to be linear. Within the range of 5 and 7 Å, the value of $g(r)$ is $(7-r)/2$.

We collected the structural data from the Protein Data Bank (PDB) [15] for each complex in our dataset. After adding domain information obtained from Pfam to each atom present in the chain, each PDB file was divided into two different ligand and receptor files based on its side chains. From [14], we know that there are 18 atom types. Thus, for each protein complex a feature vector with 18^2 values was obtained, where each feature contains the desolvation energy of a pair of atom types. As the order of interacting atom pairs is not important, the final length of feature vector for each complex was 171 that correspond to unique pairs. We also considered pairs of amino acids, and for this, we computed desolvation energy values for each pair of atoms using Eq. (7.1) and accumulated the values for each pair of amino acids. Avoiding repeated pairs resulted in 210 different features (unique

Table 7.2: Description of the subsets of features used in this study.

Name	Feature Type	Interacting Chains	DDIs
PPID-AT	atom type	✓	-
PPID-AA	amino acid	✓	-
PPID-ATD	atom type	-	✓
PPID-AAD	amino acid	-	✓

pair of amino acids).

A posterior step was to identify the 317 unique domains present in the interface of at least one complex in the dataset. Considering all pairs of domains, the desolvation energies for all atoms and amino acids present in each interacting domains were calculated using Eq. (7.1) and finally each complex had 171 atom type and 210 amino acid type features. By using desolvation energies for different types of features, four subsets of features for prediction and evaluation were generated (Table 7.2). The names of the subsets are PPID-X where X is AT for atom type, AA for amino acid pairs, ATD for atoms in interacting domains (DDIs) or AAD for amino acid pairs in interacting domains.

7.2.3 Prediction Methods

Linear Dimensionality Reduction

One of the approaches we have used for prediction is LDR. The basic idea of LDR is to represent an object of dimension n as a lower-dimensional vector of dimension d, achieving this by performing a linear transformation. We consider two classes, ω_1 and ω_2, represented by two normally distributed random vectors $\mathbf{x}_1 \sim N(\mathbf{m}_1, \mathbf{S}_1)$ and $\mathbf{x}_2 \sim N(\mathbf{m}_2, \mathbf{S}_2)$, respectively, with p_1 and p_2 the *a priori* probabilities. After the LDR is applied, two new random vectors $\mathbf{y}_1 = \mathbf{A}\mathbf{x}_1$ and $\mathbf{y}_2 = \mathbf{A}\mathbf{x}_2$, where $\mathbf{y}_1 \sim N(\mathbf{A}\mathbf{m}_1; \mathbf{A}\mathbf{S}_1\mathbf{A}^t)$ and $\mathbf{y}_2 \sim N(\mathbf{A}\mathbf{m}_2; \mathbf{A}\mathbf{S}_2\mathbf{A}^t)$

with $\mathbf{m}_i$ and $\mathbf{S}_i$ being the mean vectors and covariance matrices in the original space, respectively. The aim of LDR is to find a linear transformation matrix $\mathbf{A}$ in such a way that the new classes ($\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$) are as separable as possible. Let $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$ and $\mathbf{S}_E = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$ be the within-class and between-class scatter matrices respectively. Various criteria have been proposed to measure this separability [16]. We consider the following two LDR methods:

(a) The heteroscedastic discriminant analysis (HDA) approach [16], which aims to obtain the matrix $\mathbf{A}$ that maximizes the following function, which is optimized via eigenvalue decomposition:

$$J_{HDA}(\mathbf{A}) = tr\left\{(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\left[\mathbf{A}\mathbf{S}_E\mathbf{A}^t - \mathbf{A}\mathbf{S}_W^{\frac{1}{2}}\frac{p_1\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}}) + p_2\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1p_2}\mathbf{S}_W^{\frac{1}{2}}\mathbf{A}^t\right]\right\}. \tag{7.2}$$

(b) The Chernoff discriminant analysis (CDA) approach [16], which aims to maximize the following function, which is maximized via a gradient-based algorithm:

$$\begin{aligned} J_{CDA}(\mathbf{A}) = tr\{&p_1p_2\mathbf{A}\mathbf{S}_E\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1} \\ &+\log(\mathbf{A}\mathbf{S}_W\mathbf{A}^t) - p_1\log(\mathbf{A}\mathbf{S}_1\mathbf{A}^t) - p_2\log(\mathbf{A}\mathbf{S}_2\mathbf{A}^t)\}. \end{aligned} \tag{7.3}$$

In order to classify each complex, first a linear algebraic operation $\mathbf{y} = \mathbf{A}\mathbf{x}$ is applied to the n-dimensional vector, obtaining $\mathbf{y}$, a d-dimensional vector, where d is ideally much smaller than n. The linear transformation matrix $\mathbf{A}$ corresponds to the one obtained by one of the LDR methods, namely HDA or CDA. The resulting vector $\mathbf{y}$ is then passed through a Quadratic Bayesian (QB) classifier [16], which is the optimal classifier for normal distri-

butions. For additional tests, a linear Bayesian (LB) classifiers is considered, by deriving a Bayesian classifier with a common covariance matrix: $\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2$.

Support Vector Machines

SVMs are well known machine learning techniques used for classification, regression and other tasks. The aim of SVM is to find the support vectors (most difficult vectors to be classified), and derive a linear classifier, which ideally separates the space into two regions. Classification is normally inefficient when using a linear classifier, because the data is not linearly separable, and so the use of kernels is crucial in mapping the data onto a higher dimensional space in which the classification is much more efficient. There are number of kernels that can be used in SVM models. In our model, we use polynomial, radial basis function (RBF) and sigmoid.

7.3 Results and Discussions

7.3.1 Experimental Settings

For the LDR schemes, four different classifiers were implemented and evaluated, namely the combinations of HDA and CDA, and QB and LB classifiers. In a 10-fold cross validation setup, reductions to dimensions $d = 1, \ldots, 20$ were performed, followed by QB and LB, and the maximum average classification accuracy was recorded for each classifier. The best accuracy for each method for each dataset is bolded to indicate the classifier that performed the best for that dataset. Principal component analysis (PCA) was used as a pre-processing step to eliminate ill-conditioned matrices present in the LDR step. To select the principal components, we used different threshold values (from $\lambda_{max}10^{-2}$ to $\lambda_{max}10^{-7}$), where λ_{max}

is the largest eigenvalue of the scatter matrix. The results for the threshold that achieves the highest accuracy are reported.

The SVM was also trained in a 10-fold cross validation setup with three kernels: RBF, polynomial and sigmoid. The training was carried out with the LIBSVM package [17]. A grid search was performed on the parameters gamma and C, choosing the ones that gives the maximum average accuracy for all kernels. For the polynomial kernel, the degree of the polynomial was set to 3.

The subsets of features shown in Table 7.2 were used for prediction. To analyze the power of desolvation energy in discriminating obligate and non-obligate complexes, NOX-class [5] was also applied to our binary-PPID dataset. The following four interface properties were analyzed, since in [5], these properties were recognized as the best ones for prediction of different types of protein protein interactions:

- Interface area
- Interface area ratio
- Amino acid composition of the interface
- Correlation between amino acid compositions of interface and protein surface

We used NACCESS [18] to calculate solvent accessible surface area (SASA). After running the classifiers in a 10-fold cross validation procedure for all subsets of features, the average accuracies were computed. The accuracy for each individual fold was computed as follows: $acc = (TP + TN)/N_f$, where TP and TN are the true positive (obligate) and true negative (non-obligate) counters respectively, and N_f is the total number of complexes in the test set of the corresponding fold.

7.3.2 Analysis of Prediction

The results of SVM and LDR classifiers with different subsets of features are depicted in Table 7.3. For SVM, it is clearly seen that the RBF kernel performs better that polynomial and sigmoid kernels for all subsets of features. The atom type features present in interacting chains (PPID-AT) are best classified with SVM and a RBF kernel, achieving an average accuracy of 77.78%, while accuracy for atom type features present in interacting domains (PPID-ATD) is 70.30%. Similarly, the subset of amino acid type features present in interacting chains (PPID-AA) with 75.56% classification accuracy yields more efficient predictions than using the subset of amino acid type features present in DDIs (PPID-AAD) with 69.84% classification accuracy. Furthermore, the subset based on NOXclass features (with best accuracy of 72.38%) perform worse than the best subset based on desolvation energy properties (PPID-AT) on a SVM classifier.

For LDR, the best accuracy, 74.55%, is achieved by CDA with the quadratic classifier, which is still lower than the best accuracy achieved by SVM. Note that both of them are on the PPID-AT subset. Additionally, as in SVM, subsets of atom and amino acid type features present in interacting chains perform better than those in DDIs. Also, the NOXclass subset of features (PPID-NOXclass) yields lower accuracy (71.80%) than PPID-AT, which is based on calculation of desolvation energies only, and also DDI subsets.

Generally, it can be concluded that in our binary-PPID dataset:

(a) SVM with RBF kernel performs better than LDR methods in all subsets of features

(b) Amino acid type features (for both PPID-AA and PPID-AAT subsets) yeild lower accuracies than atom type features (PPID-AT and PPID-ATD) for both LDR and SVM classifiers

(c) Although the performance of both SVM and LDR classifiers are lower for subsets

Table 7.3: Prediction results for SVM and LDR classifiers on binary-PPID dataset.

	SVM			LDR			
	RBF	Polynomial	Sigmoid	Linear		Quadratic	
				HDA	CDA	HDA	CDA
PPID-AT	**77.78**	76.83	72.70	71.76	74.08	72.73	**74.55**
PPID-AA	**75.56**	71.43	71.11	71.46	**71.81**	71.46	65.07
PPID-ATD	**70.30**	67.62	67.43	68.66	68.06	**70.25**	68.97
PPID-AAD	**69.84**	67.62	66.35	67.34	66.12	**68.32**	62.80
PPID-NOXclass	**72.38**	69.84	69.52	68.89	**71.80**	67.71	68.97

of DDI features (PPID-ATD and PPID-AAD) than subsets of interacting chain features (PPID-AT and PPID-AA), they are acceptable results.

(d) Desolvation energy properties are more powerful than four properties of NOXclass (interface area, interface area ratio, amino acid composition of the interface and correlation between amino acid compositions of interface and protein surface) in predicting obligate and non-obligate complexes.

7.3.3 Analysis of DDIs

As discussed earlier, the total number of DDIs among 317 existing domains of our binary-PPID dataset is $100,489$. After preprocessing and removing all zero-columns, we obtain only 256 DDI pairs of which 125 are obligate and 131 are non-obligate DDIs.

The most salient feature in our binary-PPID dataset is the fact that all DDIs are presented in either obligate or non-obligate complexes and there are no DDIs in both obligate and non-obligate. This clearly implies that the type of complex could just be predicted by the DDIs present in the interactions, achieving nearly perfect prediction rate of 100%. One could design a simple classifier that contains binary features and indicates the presence or

absence of the DDI in the complex, and then a simple rule that checks those binary flags. However, this would not be the case when predicting new unknown complexes (not in this dataset). That is, when using the training data to test the classifier. When cross-validation is applied, as it is done in this paper, presence of a DDI in the training set may not imply its presence or absence in the test set. In addition, it is expected, though it would not be the case, that the DDI desolvation properties are much more informative than simply binary features indicating the presence or absence of the DDI in the complex.

We performed a visual analysis on our DDIs and discovered that from 317 existing domains in our binary-PPID dataset, 135 are present only in obligate DDIs, 158 are present only in non-obligate DDIs and 21 domains are in both obligate and non-obligate DDIs. We re-ordered the domain IDs based on their types (obligate, both and non-obligate). To provide a visual insight of the distribution of the DDIs in the different complexes, a schematic view of the DDIs in the dataset is shown in Figure. 7.1. It is clearly seen that the most homo-domain pairs are in obligate complexes (i.e. they lie on the diagonal line ($x = y$) of the plot). Only a small part of the domain IDs are common. This also implies we can achieve a reasonable prediction only by finding the domains of each unknown complex. This is an interesting issue that deserves a lot of attention, and that we are currently investigating.

7.4 Conclusion

We have proposed an approach for prediction and analysis of obligate and non-obligate protein complexes. We have investigated various interface properties of these interactions including atom and amino acid types present in interacting chains or domains. Various features are extracted from each complex, including the desolvation energies for atom and

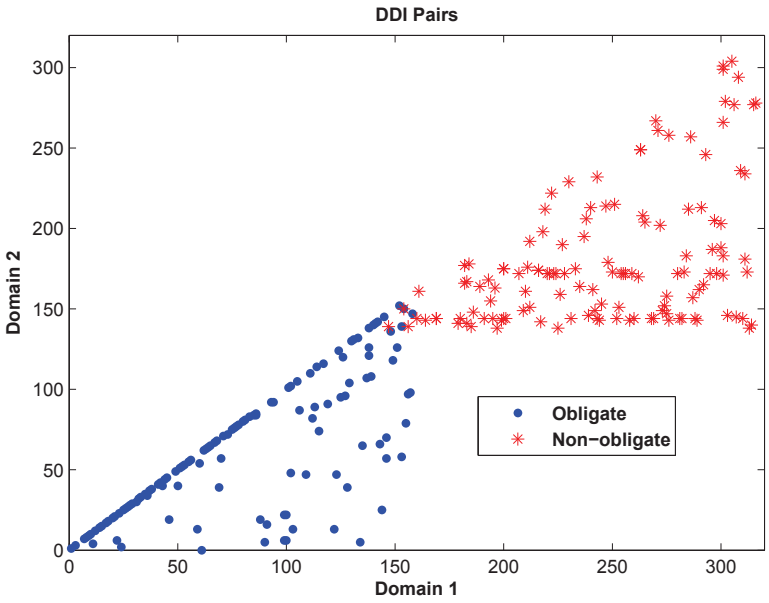


Figure 7.1: Schematic view of the DDI pairs in obligate and non-obligate interactions.

amino acid type pairs and also NOXclass properties. The classification is performed via different LDR methods and SVM with different kernels, namely RBF, polynomial and sigmoid.

The results on our binary-PPID dataset, which is a joint and modified version of two well-known datasets, show that the SVM classifier with 77.78% accuracy achieves much better classification performance, even better than LDR schemes coupled with quadratic and linear classifiers for all subset of features. The results also demonstrate that desolvation energy is better than interface area and composition for predicting obligate and non-obligate complexes.

Furthermore, visual and numerical analysis on DDIs show that (i) most homo-domain pairs are in obligate interactions and (ii) no common DDI is present in obligate and non-obligate complexes and all DDIs are present in either obligate or non-obligate complexes.

Our future work involves the use of other features such as residual vicinity, shape of the structure of the interface, secondary structure, planarity, physicochemical features, hydrophobicity, structure of domains and many others in our dataset, and also identifying pseudo-domains and motifs present in interacting proteins.

Bibliography

[1] I. Nooren and J. Thornton, "Diversity of protein-protein interactions," *EMBO Journal*, vol. 22, no. 14, pp. 3846–3892, 2003.

[2] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc. Natl Acad. Sci, USA*, vol. 93, no. 1, pp. 13–20, 1996.

[3] L. LoConte, C. Chothia, and J. Janin, "The atomic structure of protein-protein recognition sites," *J Mol Biol*, vol. 285, no. 5, pp. 2177–2198, 1999.

[4] O. K. A. Zen, C. Micheletti and R. Nussinov, "Comparing interfacial dynamics in protein-protein complexes: an elastic network approach," *BMC Structural Biology*, vol. 10, no. 26, 2010, doi: 10.1186/1472-6807-10-26.

[5] H. Zhu, F. Domingues, I. Sommer, and T. Lengauer, "NOXclass: Prediction of protein-protein interaction types," *BMC Bioinformatics*, vol. 7, no. 27, 2006, doi:10.1186/1471-2105-7-27.

[6] M. S. S. Kottha, "Classifying permanent and transient protein interactions," *German Conference on Bioinformatics*, vol. 83GI, pp. 54–63, 2006.

[7] S. De, O. Krishnadev, N. Srinivasan, and N. Rekha, "Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different." *BMC Structural Biology*, vol. 5, no. 15, 2005.

[8] J. V. Eichborn, S. Gnther, and R. Preissner, "Structural features and evolution of protein-protein interactions." *Intenational Conference of Genome Informatics.*, vol. 22, pp. 1–10, 2010.

[9] S. H. Park, J. Reyes, D. Gilbert, J. W. Kim, and S. Kim, "Prediction of protein-protein interaction types using association rule based classification," *BMC Bioinformatics*, vol. 10, no. 36, 2009, doi:10.1186/1471-2105-10-36.

[10] L. Rueda, S. Banerjee, Md. Aziz, and M. Raza, "Protein-protein interaction prediction using desolvation energies and interface properties," proceedings of the 2nd. IEEE

International Conference on Bioinformatics & Biomedicine (BIBM 2010), pp. 17–22, 2010.

[11] J. Mintseris and Z. Weng, "Structure, function, and evolution of transient and obligate protein-protein interactions," *Proc Natl Acad Sci, USA*, vol. 102, no. 31, pp. 10 930–10 935, 2005.

[12] [Online]. Available: http://pfam.sanger.ac.uk/

[13] C. Camacho and C. Zhang, "FastContact: rapid estimate of contact and binding free energies," *Bioinformatics*, vol. 21, no. 10, pp. 2534–2536, 2005.

[14] C. Zhang, G. Vasmatzis, J. L.Cornette, and C. DeLisi, "Determination of atomic desolvation energies from the structures of crystallized proteins," *J. Mol. Biol.*, vol. 267, pp. 707–726, 1997.

[15] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, 2000.

[16] L. Rueda and M. Herrera, "Linear dimensionality reduction by maximizing the Chernoff distance in the transformed space," *Pattern Recognition*, vol. 41, no. 10, pp. 3138–3152, 2008.

[17] C. L. C. Chang, "LIBSVM: a library for support vector machines," last date accessed: May 31, 2011. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.pdf

[18] S. Hubbard and J. Thornton, "Naccess," last date accessed: May 31, 2011. [Online]. Available: www.bioinf.manchester.ac.uk/naccess/

Chapter 8

Computational Analysis of the Stability of SCF Ligases Employing Domain Information

8.1 Introduction

SCF ligases are the largest class of E3 ligases and are believed to be responsible for the selection of up to 20% of the proteome for ubiquitin mediated degradation [1]. This class of E3 ligases has been shown to be minimally comprised of four subunits: RBX1, CUL1, SKP1 and an F-Box protein. The greater part of SCF-ligase function is thought to be the control of protein abundance via ubiquitination and subsequent 26S proteasome-mediated degradation. Although a substantial amount of information has been provided for regulation of the cell cycle, transcription and many other processes via proteasome-bound SCF ligase-mediated protein abundance regulation, the possibility of substrate ubiquitination events mediated by the SCF-ligase which are not subjected to proteasome bound degradation cannot be excluded at this juncture [2].

On the other hand, as the compact structural and functional units of proteins, domains have a fundamental biological role in mediating the interactions of two or more proteins

and serve some specific purpose, such as signal binding or manipulation of a substrate within cells [3]. As a consequence, recent studies focus on employing domain knowledge to predict protein-protein interactions (PPIs) [4–8]. There are few domain family resources that can be applied for this purpose such as Pfam [9] and CATH – Class, Architecture, Topology and Homologous superfamily – databases [10].

Although prediction of PPIs has been studied from many different perspectives, the main aspects that are studied include [11]: sites of interfaces (where), arrangement of proteins in a complex (how), type of protein complex (what), molecular interaction events (if), and temporal and spatial trends (dynamics). Prediction of types of PPIs, in particular, the identification and analysis of obligate and non-obligate complexes and their relevant properties has been studied from different perspective [12–18]. Obligate complexes are more stable and have high-affinity interactions than non-obligate ones [19].

In this paper, we present an analysis of the role of domain interactions in determining obligate and non-obligate PPIs that are known or predicted to occur involving subunit components of the SCF-ligase complex. For this, we used a manually curated SCF-ligase dataset of 30 complexes that contains 21 obligate and 9 non-obligate complexes.

The numerical results on the number and type of interactions demonstrate that most of the PPIs are mediated by at least one domain. Also, domain-domains interactions dominate in obligate complexes whereas in non-obligate complexes, most of the interactions are mediated by one domain and a polypeptide chain. These results are in agreement with similar studies published to date [13, 20].

Furthermore, the prediction results obtained by applying a support vector machine (SVM) classifier on different extracted domain-based subset of features show that using the combinations of domain-domain type, domain-peptide chain type and no-domain fea-

ture vectors yield the best performance (80.64% prediction accuracy), in comparison to domain-domain type or single-domain type used as feature vectors, individually. Also, by employing Chi-Square for feature selection, the Pfam domain "*PF00400*" is recognized as the most discriminative feature for classification of obligate and non-obligate complexes in the dataset, achieving 77.42% prediction accuracy. This domain appears only in non-obligate complexes and does not have any interactions with other domains.

8.2 Materials and Methods

To predict complex types, initially, the prediction properties (features) of each complex in the dataset are extracted. Then, after selecting the most powerful and discriminative features for prediction by employing a feature selection method, a classifier method is applied on the selected features to predict the complex types. More explanations regarding the dataset, extracted features and also feature selection and classifier methods used in this paper are discussed below.

8.3 SCF-Ligases Dataset

As mentioned earlier, SCF-ligases are the largest class of E3 ligases which are minimally comprised of four subunits of RBX1, CUL1, SKP1 and an F-Box protein, as shown in Figure 8.1. RBX1 is responsible for the recruitment of the E2 ligase, CUL1 acts as a scaffold for the assembly of SCF ligase, SKP1 acts as an adaptor connecting CUL1 to the F-box protein, and the F-box protein dictates the target specificity of the E3 through substrate selection [21].

Our manually curated SCF-ligase dataset contains 30 complexes. Of these, 21 com-

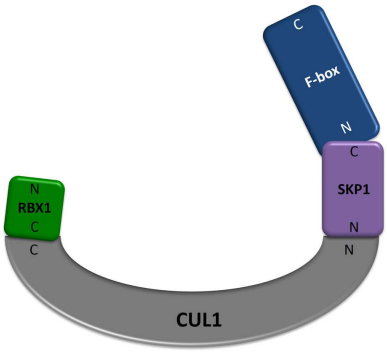


Figure 8.1: A schematic view of a SCF-ligase.

plexes have strong interactions (obligate) and 9 complexes have weak interactions (non-obligate). The Protein Data Bank (PDB) IDs of these complexes and the interacting chains are shown in Table 8.1.

8.3.1 Prediction Properties

To extract domain-based prediction properties, first, the tertiary and quaternary structures of the complexes in the dataset were downloaded from the PDB [22]. After filtering and modifying the PDB files, the sequence domain content of each subunit was gleaned from the Pfam website [9] and mapped to the corresponding amino acids in the chain.

In the dataset, 27 unique Pfam domains were present in the interface of at least one complex were identified. Of these, 17 domains were in the obligate complex class and 4 were in the non-obligate, while the remaining 7 domains were both obligate and non-obligate complexes. A domain is considered to be in the interface, if it has at least one residue interacting with a domain in the other chain.

To calculate the features for prediction, first, all pairs of interacting amino acids and their corresponding domains that are less than 7 Å apart from each other were extracted for each complex in the dataset. After that, the extracted amino acid pairs are grouped

Table 8.1: Dataset of SCF-ligase complexes.

Non-obligate Complexes (9)					
1NEX	E:B	2AST	C:D	2P1N	B:C
1P22	C:A	2E33	A:B	3DB3	A:B
2AST	B:D	2OVQ	B:C	3OGK	Q:B
Obligate Complexes (21)					
1FQV	B:A	2ASS	A:B	2P1M	A:B
1LDK	A:D	2ASS	B:C	2QHO	A:B
1LDK	B:C	2E31	A:B	3MTN	A:B
1LDK	E:D	2HYE	A:C	3NHE	A:B
1NEX	A:B	2HYE	D:C	3OGK	A:B
1P22	A:B	2HYE	A:B	3OLM	A:D
1U6G	A:C	2OVP	A:B	3PT2	A:B

into two-domain, single-domain and also no-domain groups based on their corresponding domains. For instance, an amino acid pair is a member of single-domain group if only one of the interacting amino acids belongs to at least one domain. To generate a domain-domain type (*DDT*) feature vector for each complex, all pairs of domains were considered. Since the order of the interacting domain pairs is not important, generated feature vectors for domain-domain type features contain $378 = \binom{27}{2} C + 27$ values. The value of each domain pair in the *DDT* feature vector is the cumulative frequency across all occurrences of their corresponding amino acid pairs present in the group of two-domain. Finally, after pre-processing and finding domain-domain type feature vectors for all the complexes of the dataset, all zero-columns, which represent domain pairs that were not present in any complexes, were removed.

To generate a single-domain type (*SDT*) feature vector for each complex, all 27 identified unique domains in the dataset were considered, individually. Each feature contains the sum of the frequencies for all amino acid pairs present in the group of single-domain with

the same domain-peptide chain interactions. Similarly, all the zero columns were removed after preprocessing all complexes, yielding 19 single-domain interactions for the dataset.

The no-domain (*noD*) feature vector has only one feature that shows the number of amino acid pairs for each complex in the group of no-domain.

8.3.2 Prediction Method

After finding the properties of the complexes of the SCF-ligase dataset, a prediction method is applied to them. In this work, the prediction is performed via a SVM. The main goal of SVM is to find the support vectors, and derive a linear classifier, which ideally separates all the feature vectors into two regions. Using a linear classifier is inefficient in most cases when the data is not linearly separable. Hence, kernels, such as polynomial, radial basis function (RBF) and sigmoid, can be used to map the data onto a higher dimensional space in which the classification boundary can be found much more efficiently. The effectiveness of the SVM depends on the selection of the kernel and optimizing its parameters [23]. In addition, sequential minimal optimization (SMO) is a fast learning algorithm that has been widely applied in the training phase of a SVM classifier as one possible way to solve the underlying quadratic programming problem. In this work, the SMO module of the Waikato Environment for Knowledge Analysis (WEKA) with a normalized polynomial kernel, default parameter settings, and 10-fold cross-validation is used to perform classification via the SVM [24].

8.3.3 Feature Selection

Feature selection is the process of choosing the best subset of relevant features that represents the whole dataset efficiently after removing redundant and/or irrelevant ones. Ap-

plying feature selection before running a classifier is useful in reducing the dimensionality of the data and, thus, reducing the prediction time while improving the prediction performance. In this paper, Chi Square (χ^2) is employed for feature selection. This method measures the degree of independence of each feature to the classes by computing the value of the chi-square statistic [25]. The χ^2 value of a feature X with respect to class attribute Y is calculated as follows:

$$\chi^2(Y,X) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}, \tag{8.1}$$

where A is the number of times feature X and class Y co-occur, B in the number of times X occurs without Y, C is the number of times Y occurs without X, D is the number of times neither X and Y occurs, and N is the total number of samples. In this work, the ChiSquaredAttributeEval module of the WEKA is used for ranking the features.

8.4 Results and Discussions

8.4.1 Analysis of Interaction Types

After identifying all the unique domains present in the interface of at least one complex in the dataset, for each complex, the number of domain-domain interactions (*DDIs*), domain-peptide chain interactions (*DIs*) and also the number of interactions that none of the interacting amino acids belong to any domains (*noD*) were calculated. In Figure 8.2, the number and type of the interactions for non-obligate (left) and obligate (right) complexes are shown in different colors: blue for *DDIs*, red for *DIs* and green for *noD* interactions.

From the histogram, it is clear that obligate complexes have more number of interactions

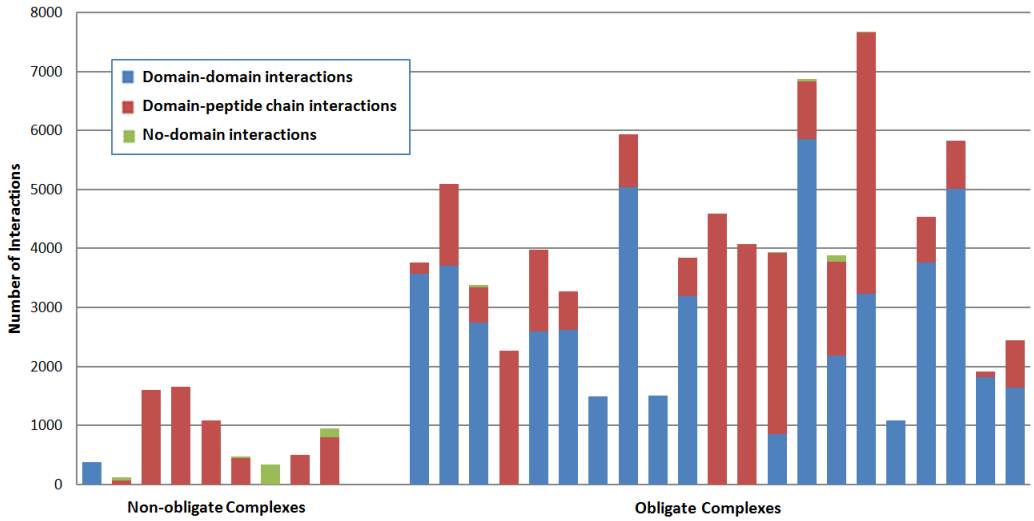


Figure 8.2: Number and type of interactions for two groups of non-obligate (left) and obligate (right) SCF-ligase complexes.

and most of them are domain-domain interactions. In contrast, non-obligate complexes have less interactions in comparison to obligate ones and most of their interactions are single-domain (*DIs*). Also, most of the interactions, are mediated by at least one domain.

Similarly, the statistical results of the average number of interactions for each obligate and non-obligate complexes of the SCF-ligase dataset categorized by their interaction types shown in Table 8.2, confirm the results demonstrated in Figure 8.2. It means that in both obligate and non-obligate complexes, less than 1% of the interactions mediated by no domains. Also, the average number of interactions of obligate complexes (3,879 pairs) is approximately five times greater than the number of interactions of non-obligate complexes (794 pairs) in the dataset. In addition, more interactions of non-obligate complexes (greater than 86%) are *DIs* while for obligate complexes, more than 63% of interactions are *DDIs*.

Table 8.2: A summary of the average number of interactions for obligate and non-obligate complexes of the SCF-ligase dataset categorized by their interaction types.

Complex Type	Type of interactions			
	# DDIs	# DI	# noD	Total
Obligate	43	686	64	794
Non-obligate	2475	1394	10	3879

8.4.2 Analysis of the Prediction Properties

After running the SVM in a 10-fold cross validation procedure for all subsets of features, the average accuracies were computed as follows: $acc = (TP + TN)/N$, where TP is the number of true positive (obligate), TN is the number of true negative (non-obligate), and N is the total number of complexes in the test sets of all 10 folds.

The prediction results of the SVM classifier with different domain-based subsets of features are depicted in Table 8.3. Although, as explained earlier, 27 unique domains were identified in our SCF-ligase dataset, only 19 of them had interactions with peptide chains (*SDT* feature vector). Similarly, from the 378 features of *DDT* features, only 21 domain pairs were present in at least one of the complexes of the dataset.

From the table, it is obvious that the subset of "*DDT* + *SDT* +*noD*" with 41 features achieves the best classification accuracy of 80.64%. Also, by combining the *SDT* feature vector with *noD* feature, the prediction accuracy improved to 77.42%, which is better than using *SDT* features individually. This improvement can also be seen by combining feature vectors of *DDT* with *noD*. Hence, it can be concluded that *noD* is one of the best features for prediction of obligate and non-obligate complexes in the dataset. Furthermore, the subset based on domain-domain interaction pairs (*DDT*) with accuracy of 70.96% classification accuracy yields less efficient predictions than other subset of features. Also, by comparing the classification accuracies of *SDT* and *DDT*, it can be seen that the features of *SDT* are

Table 8.3: Prediction accuracies of SVM-SMO for all domain-based subsets of features of the SCF-ligase datase.

Subset Name	Number of features	Accuarcy
DDT	21	70.96%
DDT + noD	22	74.20%
SDT	19	74.19%
SDT + noD	20	77.42%
DDT + SDT + noD	41	80.64%

more powerful than the features of *DDT* to classify these two types of complexes.

8.4.3 Analysis of the Feature Selection

As explained earlier, we employed the filter method of χ^2 in WEKA for feature selection. The χ^2 value of all features except a single-domain type feature "*PF00400*" is zero. Applying the SVM classifier using this single feature, achieving 77.42% prediction accuracy, confirms that this single feature is the most powerful and discriminating feature for classification of obligate and non-obligate complexes in the dataset.

Pfam domain of "*PF00400*" appears in the following complexes: chain B of *2OVP*, *2OVQ* and *1NEX* and chain A of *1P22*. But after finding the feature vectors of *SDT*, *DDT* and *noD*, "*PF00400*" can only seen in *SDT* and mediated single-peptide chain interactions of non-obligate complexes.

8.5 Conclusion

We have presented a domain-based approach to predict types of interactions in SCF-ligases. The model uses the frequencies of amino acid pairs present in the interface of domain-domain, domain-peptide chain and no-domain interactions as the prediction properties. χ^2

is applied for selecting the most powerful features and a SVM is used for prediction of our manually pre-classified SCF-ligase dataset.

The numerical results on the number and type of interactions demonstrated that (a) more than 99% of the PPIs are mediated by at least one domain, (b) the average number of interactions of obligate complexes is greater than those of non-obligate complexes, and (c) domain-domains interactions dominate in obligate complexes whereas non-obligate complexes exhibit more domain-peptide chain interactions. Also, the prediction results show 80.64% accuracy by combining all feature vectors. Furthermore, a little decrease in prediction accuracy (3%) using χ^2 feature selection is acceptable because of the less time and space complexity required for prediction.

The results presented here are limited by the current state of domain-definition and content of the PDB and Pfam databases. The utility of our study will benefit from ongoing enrichment of domain information in the public databases, resulting in further enhancements in the predictive power of our approach.

From this study, various open questions remain to be answered. One of these is to perform biological analysis on the domains and motifs present in the interface of obligate and non-obligate SCF-ligase complexes in order to achieve a better insight on these complexes, their interactions, and functions.

Bibliography

[1] M. H. Dezfulian, D. M. Soulliere, R. K. Dhaliwal, M. Sareen, and W. L. Crosby, "The SKP1-like gene family of arabidopsis exhibits a high degree of differential gene expression and gene product interaction during development," *PLOS ONE*, vol. 7, no. 11, 2012.

[2] D. M. Duda, D. C. Scott, M. F. Calabrese, E. S. Zimmerman, N. Zheng, and B. A. Schulman, "Structural regulation of cullin-RING ubiquitin ligase complexes," *Current Opinion in Structural Biology*, vol. 21, no. 2, pp. 257–264, 2012.

[3] L. Chen, R. Wang, and X. Zhang, *Biomolecular Networks: Methods and Applications in Systems Biology*. John Wiley and Sons, 2009.

[4] M. Maleki, M. Hall, and L. Rueda, "Using desolvation energies of structural domains to predict stability of protein complexes," *Journal of Network Modeling Analysis in Health Informatics and Bioinformatics (NetMahib)*, vol. 2, no. 4, pp. 267–275, 2013.

[5] M. Hall, M. Maleki, and L. Rueda, "Multi-level structural domain-domain interactions for prediction of obligate and non-obligate protein-protein interactions," in *ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM-BCB)*, Florida, USA, pp. 518–520, Oct. 2012.

[6] N. Zaki, "Protein-protein interaction prediction using homology and inter-domain linker region information." *Advances in Electrical Engineering and Computational Science, Springer*, vol. 39, pp. 635–645, 2009.

[7] T. Akutsu and M. Hayashida, "Domain-based prediction and analysis of protein-protein interactions." *Biological data mining in protein interaction networks, Medical Information Science Reference, chapter 3*, pp. 29–44, 2009.

[8] P. Chandrasekaran, C. Doss, J. Nisha, R. Sethumadhavan, V. Shanthi, K. Ramanathan, and R. Rajasekaran, "In silico analysis of detrimental mutations in ADD domain of chromatin remodeling protein ATRX that cause ATR-X syndrome: X-linked disorder," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 2, no. 3, pp. 123–135, 2013.

[9] M. Punta, P. Coggill, R. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. Sonnhammer, S. Eddy, A. Bateman, and R. Finn, "The Pfam protein families database." *Nucleic Acids Res.*, vol. 40, no. D1, pp. D290–D301, 2012.

[10] A. Cuff, I. Sillitoe, T. Lewis, O. Redfern, R. Garratt, J. Thornton, and C. Orengo, "The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies." *Nucleic Acids Res.*, vol. 37, pp. 310–314, 2009.

[11] I. Kurareva and R. Abagyan, "Predicting molecular interactions in structural proteomics," in *Computational Protein-Protein Interactions*, R. Nussinov and G. Shreiber, Eds. CRC Press, ch. 10, pp. 185–209, 2009.

[12] J. Mintseris and Z. Weng, "Structure, function, and evolution of transient and obligate protein-protein interactions," *Proc Natl Acad Sci, USA*, vol. 102, no. 31, pp. 10 930–10 935, 2005.

[13] H. Zhu, F. Domingues, I. Sommer, and T. Lengauer, "NOXclass: Prediction of protein-protein interaction types," *BMC Bioinformatics*, vol. 7, no. 27, 2006, doi:10.1186/1471-2105-7-27.

[14] S. H. Park, J. Reyes, D. Gilbert, J. W. Kim, and S. Kim, "Prediction of protein-protein interaction types using association rule based classification," *BMC Bioinformatics*, vol. 10, no. 36, 2009, doi:10.1186/1471-2105-10-36.

[15] R. Liu, W. Jiang, and Y. Zhou, "Identifying protein-protein interaction sites in transient complexes with temperature factor, sequence profile and accessible surface area," *Amino Acids*, vol. 38, pp. 263–270, 2010.

[16] Md. Aziz, M. Maleki, L. Rueda, M. Raza, and S. Banerjee, "Prediction of biological protein-protein interactions using atom-type and amino acid properties," *Proteomics 2011*, vol. 11, no. 19, pp. 3802–10, 2011.

[17] M. Maleki, G. Vasudev and L. Rueda, "The role of electrostatic energy in prediction of obligate protein-protein interactions," *BMC Proteome Science*, vol. 11, 2013.

[18] D. La, M. Kong, W. Hoffman, YI. Choi, and D. Kihara, "Predicting permanent and transient protein-protein interfaces," *Proteins*, vol. 81, no. 5, pp. 805–18, 2013.

[19] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc. Natl Acad. Sci, USA*, vol. 93, no. 1, pp. 13–20, 1996.

[20] SEA. Ozbabacan, HB. Engin, A. Gursoy, and O. Keskin, "Transient protein-protein interactions," *Protein EngDes Sel.*, vol. 24, no. 9, pp. 635–48, 2011.

[21] B. B. Chen and R. K. Mallampalli, "F-box protein substrate recognition-a new insight," *Cell Cycle*, vol. 12, no. 7, pp. 1009–1010, 2013.

[22] H. M. Berman, G. J. Kleywegt, H. Nakamura, and J. L. Markley, "The Protein Data Bank at 40: reflecting on the past to prepare for the future," *Structure*, vol. 20, no. 3, pp. 391–396, 2012.

[23] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York, NY: John Wiley and Sons, Inc., 2000.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[25] J. Novakovic, P. Strbac, and D. Bulatovic, "Toward optimal feature selection using ranking methods and classification algorithms." *Yugoslav J. of Operations Research*, vol. 21, no. 1, pp. 119–135, 2011.

Chapter 9

Conclusions and Future Work

9.1 Conclusion

In this thesis, we have presented a prediction model to analyze protein-protein interaction types, namely obligate and non-obligate. This model uses the proposed physicochemical features of desolvation and electrostatic energies for pairs of atoms, amino acids or domains present in the interfaces of such complexes as prediction properties. Moreover, the idea of employing a domain-based approach for predicting obligate and non-obligate protein complexes is also proposed in this thesis in order to achieve a better insight on proteins and their interacting domains. For this purpose, both sequence domains of Pfam and structural domains of CATH are considered.

After extracting the main features from the complexes, prediction is performed via several state-of-the-art classification techniques, including LDR, SVM, NB and k-NN for several well-known datasets of pre-classified complexes. Also, for an in-depth analysis of classification results, some other experiments were also performed such as varying the distance cutoffs between atom pairs of interacting chains, or performing some visual and/or numerical analysis. Moreover, several feature selection algorithms including GR, IG, Chi2

and mRMR are also applied on the available datasets in some of our studies to obtain more discriminative pairs of atom, amino acid, and domain types as features for prediction.

A summary of the experiments that we have performed in some of our previous studies (those reported in the thesis) including the dataset name, extracted feature types, names of applied classification and/or feature selection methods and also types of the performed analysis is shown in Table 9.1.

In the table, DE and EE are the abbreviations of desolvation energy and electrostatic energy, respectively. For the features, AT is a vector of 171 atom pairs, AA is a vector of 210 amino acid pairs, DDIs is a vector of interacting domain pairs, SDT is a vector of single domains (domain-peptide chain interactions), and noD is the number of amino acid pairs for each complex with no-domain interactions. For more details regarding the type of features, the reader is referred to the corresponding chapters of the papers.

A summary of the prediction results obtained using different types of features as well as the biological, numerical and visual analysis and discussions can be found at the end of each chapter, separately.

Table 9.1: Experimental settings employed in our different studies for prediction of obligate and non-obligate complexes.

Ch.	Dataset			Extracted Features		Process		Analysis	
	Name	# Ob.	# Nob.	Partners	Properties	Feature Sel.	Classification	Numerical	Visual
Ch. 2	BPPI	213	303	AT pairs AA pairs	DE	AA-based	LDR SVM	Accuracy # interactions	Heatmaps
Ch. 3	MW ZH	115 75	211 62	AT pairs AA pairs	DE	mRMR $mRMR^{pro}$ Biological FS.	LDR	Accuracy AA pairs	Heatmaps
Ch. 4	MW ZH	101 73	201 58	AT pairs AA pairs	DE EE	mRMR χ^2 IG GR	LDR SMO NB k-NN	Accuracy Distance cutoffs	ROC
Ch. 5	MW-CATH ZH-CATH	100 72	161 55	CATH-DDIs (levels 2 and 3) (level2+level3)	DE	DDI-based	LDR SMO NB k-NN	Accuracy AUC # DDIs	ROC
Ch. 6	MW-CATH ZH-CATH	106 72	181 55	AA pairs CATH-DDIs (levels 2 to 8)	DE	-	LDR SVM	Accuracy # DDIs	DDI types
Ch. 7	BPPI-Pfam	146	169	AT pairs AA pairs Pfam-AT Pfam-AA	DE	-	LDR SVM	Accuracy	DDI types
Ch. 8	SCF-ligase	21	9	Pfam-DDIs Pfam-SDT Pfam-noD	Freq.	χ^2	SMO	Accuracy # interactions	Histogram

9.2 Future Work

The future work involves various extensions to this thesis listed as follows:

- The approach described in this thesis can also be used for prediction of other types of complexes, including intra and inter domains, homo and hetero-oligomers.
- Performing biological analysis to find the interacting domains of different types of complexes and also investigating the types of DDIs are some of the worthy research topics in this area.
- Biologically guided feature selection and interpretation combined with automatic feature selection could also be useful.
- Performing some post analysis to obtain the more discriminating and relevant pairs of atoms, amino acids, and domains present in the interface of two interacting proteins that are biologically meaningful is worth further investigation.
- Other properties can also be extracted to predict these types of interactions including geometric (e.g., shape, planarity, roughness or others), and other statistical and physicochemical properties such as residue and atom vicinity, secondary structure elements, and salt bridges, among others.
- Studying motifs present in the interface of PPIs in order to achieve a better insight on these complexes, their interactions, and functions is useful.
- Electrostatic energies can also used for the prediction of PPI of other types of interactions.

- Investigating the role of buried atoms and their influence in different types of interactions especially for electrostatic interactions that are long-range and cover a broader area in the interface is another problem that deserves attention.

Vita Auctoris

Mina Maleki was born in 1979 in Tehran, Iran. She received her Bachelors degree in Computer Engineering from Azzahra University, Tehran, Iran, in 2002, and her Master in Computer Engineering and Information Technology from Amirkabir University of Technology, Tehran, Iran, in 2006. Her research interests include pattern recognition, machine learning, bioinformatics, data mining and protein-protein interactions.