2011

PREDICTION OF OBLIGATE AND NON-OBLIGATE PROTEIN-PROTEIN INTERACTIONS

Muhammad Aziz
University of Windsor

PREDICTION OF OBLIGATE AND NON-OBLIGATE PROTEIN-PROTEIN INTERACTIONS

by
Muhammad Mominul Aziz

A Thesis
Submitted to the Faculty of Graduate Studies
through Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada
2011

PREDICTION OF OBLIGATE AND NON-OBLIGATE PROTEIN-PROTEIN INTERACTIONS

by

Muhammad Mominul Aziz

APPROVED BY:

Dr. James Gauld, External Reader
Chemistry and Biochemistry

Dr. Alioune Ngom, Internal Reader
Computer Science

Dr. Luis Rueda, Advisor
Computer Science

Dr. Subir Bandyopadhyay, Chair of Defense
Computer Science

August, 2011

Declaration of Co-Authorship

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's 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 and have included copies of such copyright clearances to my appendix.

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.

Abstract

Protein-protein interactions are very important for many biological processes as this often leads to a particular protein complex to perform particular function. Thus, to identify different protein interactions helps to understand the function performed by that protein. The interaction between obligate and non-obligate complexes with each other is a particular problem that has drawn the attention of the research community in the past few years. In this thesis, we discuss this classification problem and show an efficient model to distinguish these two types of protein complexes correctly. We used new features such as desolvation energies for atom and amino acid type to compare with some other features which have already been used to validate and evaluate our model and test the strength of our newly selected features. We also used some well-known feature selection techniques to perform classification with almost the same or higher accuracy but in time efficient manner. To achieve a better insight of this classification, we also performed some visual and post-analysis, and biochemically driven feature selection to achieve a better perspective about the reasons for interaction of these types of complexes.

Dedication

I would like to dedicate this thesis to my parents.

It was my father's dream that brought me here in pursuit of higher studies, his ideology that helped me get through the tough times. It was my mother, who never stopped believing in me, gave me strength all the way through and never stopped praying for my success. I do not know what I would do without you. Thank you.

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to Dr. Luis Rueda, my supervisor, for his steady encouragement, patient guidance and enlightening discussions throughout my graduate studies. Without his help, the work presented here could not have been possible.

I also wish to express my appreciation to Dr. Alioune Ngom, School of Computer Science and Dr. James Gauld, Department of Chemistry and Biochemistry for being in the committee and spending their valuable time and Dr. Subir Bandyopadhyay, School of Computer Science for serving as the chair of the defense.

Finally, I would like to thank Numanul Subhani, Sridip Banerjee, Mina Maleki and all my friends on and off campus in Windsor for their consistent moral support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Molecular biology is a branch of biology that overlaps biology and chemistry. It studies different biological activities with respect to molecules. This branch mainly deals with different types of interactions between various cell systems such as different types of DNA, RNA and protein complexes. Understanding these interactions is also included in this branch of biology. Molecular biology revealed the original convergence of geneticists, physicists and structural biochemists on a common problem; the structure and function of a biological complex. Key concepts of molecular biology include mechanisms, information and genes. The history of molecular biology provides the importance of the discovery of macromolecular mechanisms [12].

1.1 Bioinformatics

Bioinformatics can be seen as an application of statistics and computer science. It is actually, a combination of computer science techniques applied to molecular biology [28], and an indispensable field for modern genomics. The growth of biological information made this branch very important for researchers. Rapid development in molecular biology is producing huge amounts of data every day which needs efficient computational and mathematical approaches. The most common problems in bioinformatics are to analyze DNA and protein sequences aligning and comparing different DNA and proteins, viewing and studying the structure of proteins, among others. The goal of this section is to understand different biological processes by applying computationally intensive methods such as pattern recognition, machine learning and data mining. Drug development and evolution needs information about biological processes, and hence this area of research is very important for health.

1.2 Protein-protein Interaction Prediction

Protein is an organic compound made of a chain of amino acids that forms a globular form. They have different types and four level of structure (details discussed in Chapter 2). In our thesis, we focus only on specific types of proteins, namely obligate and non-obligate protein complexes. Precise collision in different proteins often leads to the final complex which is known as obligate complex, and when this collision forms an encounter complex that may finally lead to a different one is known as non-obligate complex [20]. In Figure 1.1 (a) complex 1B4U is an obligate complex with chains A-B and in (b) complex 1AVA is a non-obligate complex with chains A-C. Our goal is to predict these types of protein complexes based on their structural and interaction data.

Multiple cellular processes such as signal transduction, immune response, regulation of gene expression and different biological processes that needs oligomerization are involved in protein-protein interaction (PPI). These interactions can be attractive or repulsive. Though PPI has dependency on protein surfaces and the environment conditions,

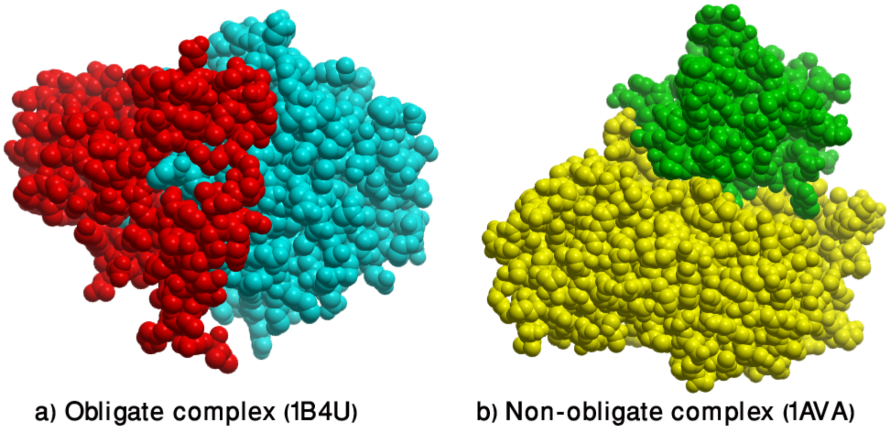


Figure 1.1: Examples of obligate and non-obligate protein complexes.

there have been many efforts made to identify and understand the responsible factors for different types of interactions between proteins at different levels such as atomic and amino acid level [13, 15, 25]. The study of PPI depends on purposes and perspectives. The key problems in PPI are [16]:

- Predicting interfaces involved in the interaction
- Predicting spatial arrangement of the interacting chains or molecules
- Predicting the identity of the molecules involved in the interaction

Different types of protein-protein interactions provide different levels of information on different biological processes [20]. Specific types of PPI are obligate and non-obligate interactions [19, 30]. These types of interactions are mostly based on the lifetime and stability of the protein complexes. Non-obligate interactions are usually less stable which makes the prediction and discriminating this type from obligate very hard. In vivo, the structural units of obligate complexes do not exist as stable, whereas in non-obligate complexes, structure may stay as stable as functional units. In our study, we focus on this problem of predicting obligate and non-obligate interactions with different prediction methods.

1.3 Feature Generation and Selection

To predict the class types, every prediction method needs observed properties of the known class samples called features. Features are generally nominal or numeric values, and the process of calculating the features for each sample from the input dataset is called feature generation. To reduce the size of the generated features from the input we use feature extraction methods. Feature extraction [10] is a popular pattern recognition method. It is a special form of dimensionality reduction. When the length of the feature vector is very large, we may need to apply feature extraction methods to find the lower dimensional representation of that original feature vector. The transformation of high dimensional data to lower dimensional data is called feature extraction [10]. There are many feature extraction methods and for our thesis we use principal component analysis (PCA) [14] and three linear dimensionality reduction (LDR) methods [24] (details discussed in Chapter 3) which are:

- Fisher's discriminant analysis (FDA) [5, 6]
- Heteroscedastic discriminant analysis (HDA) [17]
- Chernoff discriminant analysis (CDA) [24]

Lower dimensional data obtained by these three LDR methods are then passed through quadratic Bayesian (QB) and linear Bayesian (LB) classifiers [5] for final prediction. For each classifier, prediction accuracy and time taken for that prediction are very important. Thus, to reduce the computational time for classification, in our thesis we use some feature selection algorithms to find the best subset of the original feature set that produces almost the same accuracy level, if not better. Feature selection methods select some of the features that are strong enough for prediction. Thus, using feature selection methods we can have less number of features for our classifiers.

1.4 Motivation and Objectives

Many researchers are working to understand different biological functions based on protein sequence or secondary structure. They are conducting their research work either by following labor intensive experimental or computational approaches. These approaches gather the required information from different types of interactions that happen within the protein complex or between different protein complexes. Protein-protein interaction may change the shape of the complex or modify another protein complex, which means that its functionality might also change. Thus, by understanding protein-protein interactions, we could provide a plausible mechanism for complex formation and explain different biological processes such as signal transduction. These information might also help researchers understand different diseases better which can lead to effective drug development and open a new way for treatment. Obtaining information from protein composition, stability and interaction duration which is the key factor to differentiate two specific type of protein groups namely obligate and non-obligate complexes, can also help this research work greatly. Thus, studying these two types of protein-protein interactions will help the researchers gather more information for understanding and explaining biological processes and mechanism of complex formation.

Among different types of protein-protein interactions, we focus on obligate and non-obligate protein-protein interactions [19, 30]. During their life span, proteins interact with each other or even within themselves to change their shape or other complexes to perform a specific biological function. Determination of obligate and non-obligate interactions via experimental approaches such as co-immunoprecipitation or affinity chromatography is often labor expensive and might suffer from system errors [1, 7, 23]. Thus, efficient computational approaches are necessary to solve this problem successfully. In our thesis, we study this problem of determining the type of interaction based on the stability of the protein complexes which is a two class prediction problem where the classes are obligate and non-obligate.

Different feature and prediction methods can be used to solve this specific problem. In our thesis we use desolvaion energies [4]. Using these properties we show that it is better than recently used properties such as NOXclass features (interface area, interface area ratio, amino acid composition of the interface, correlation between amino acid compositions of interface and protein surface, gap volume index and conservation score of the interface) [30]. With our proposed new properties, we also include some grouping methods such as grouping by amino acids or by atom types. We use some feature selection algorithms such as forward/backward feature selection [26] and minimum redundancy maximum relevance (mRMR) [8] to select the best feature set that can reduce the prediction time while still achieving good prediction performance.

In this thesis, we also use some biological groups such as hydrophobicity, hydrophilicity and amphipathicity which provides a better insight of the solution. To perform visual post-analysis we use heatmaps that can help us choose the atom pairs or amino acid pairs responsible for obligate and non-obligate interactions visually. Finally, we compile a dataset of obligate and non-obligate complexes by using our computational approach. Merging all available data helps the classifier train better for other new features. In out thesis we also propose a general model to solve similar kinds of problems.

1.5 Contribution

In this thesis, we focus on prediction of two types of protein-protein interactions, namely obligate and non-obligate and evaluate the results efficiently. Our main contributions in this thesis are:

- Propose a new prediction scheme that combines LDR classification methods and desolvation energy as properties.
- Compile a new dataset by combining Mintseris *et al.* dataset [19] and Zhu *et al.* dataset [30].
- The use of desolvation energies with different grouping criteria such as by atom types and amino acid types.
- Post analysis by using visual tools (heatmaps) and feature selection algorithms for pattern recognition [26].

In this thesis, we are proposing a new method to solve this type of problems efficiently. The development of an automatic tool is also important, which downloads the structural information from PDB [2] and calculates desolvation energies for the classifiers. In order to provide a large number of samples for prediction and future use a new dataset is created. Furthermore, to achieve a better insight about the results, biologically meaningful grouping such as by hydrophobicity, hydrophilicity and amphipathicity is also applied to find out the pairs of amino acid that are mainly involved in these types of interactions. Post-analysis with heatmaps and different feature selection algorithms are used to evaluate the results of our experiments and draw some valid and interesting biological points from this study.

1.6 Thesis Organization

The thesis is organized in six chapters. Chapter II provides a survey of obligate and non-obligate protein-protein interaction and the prediction methods used to determine those types. Chapter III presents different feature extraction and selection methods that can be used for prediction. Chapter IV describes the proposed model and features and all required methods for the experiments. Chapter V discusses the experimental results with the proposed approach and a comparison with some existing methods. Finally, Chapter VI concludes the thesis and identifies the problems arising from this work and relevant future works.

Chapter 2

Obligate and Non-obligate PPI Prediction

2.1 Proteins

In 1838, Dutch chemist Gerhardus Johannes Mulder first described proteins, which were named by Swedish chemist J*ö*ns Jakob Berzelius. Protein is an organic compound, made of an arranged chain of amino acids which forms a globular or fibrous form [27]. All the amino acids in a protein are joined by peptide bonds between carboxyl and amino groups of adjacent amino acids.

2.2 Protein Structures

A protein is a polymer of amino acid that has four levels of structure [22]:

1. Primary

2. Secondary
3. Tertiary
4. Quaternary

In its primary structure, Figure 2.1 (a), the protein is expressed with its amino acid sequence of its polypeptide chain. This expression contains either one letter (or three letter) abbreviations for the amino acid of that specific protein. Secondary structure, Figure 2.1 (b), describes the regions of the chains of a protein that are organized into regular shapes known as alpha-helices, beta-sheets and others. These secondary structures are held together by hydrogen bonds. Adding the folding information to the secondary structure, the tertiary structure, Figure 2.1 (c), describes the three dimensional shape of the protein. When a protein has more then one polypeptide chain (also called subunit), the structure of that protein complex is the quaternary structure, Figure 2.1 (d). In our thesis, we focus on tertiary and quaternary structures of the proteins and use the three dimensional shape data for our calculations.

2.3 Protein-protein Interactions

To perform many biological functions, one or more proteins must bind with each other to react. Protein-protein interaction involves [20]:

- Direct contact association of molecules, which means that different molecules belonging to specific amino acids within a protein may interact with each other if they are close to each other. Generally, for direct association molecules, it should be within 7$\AA$ distance [4].

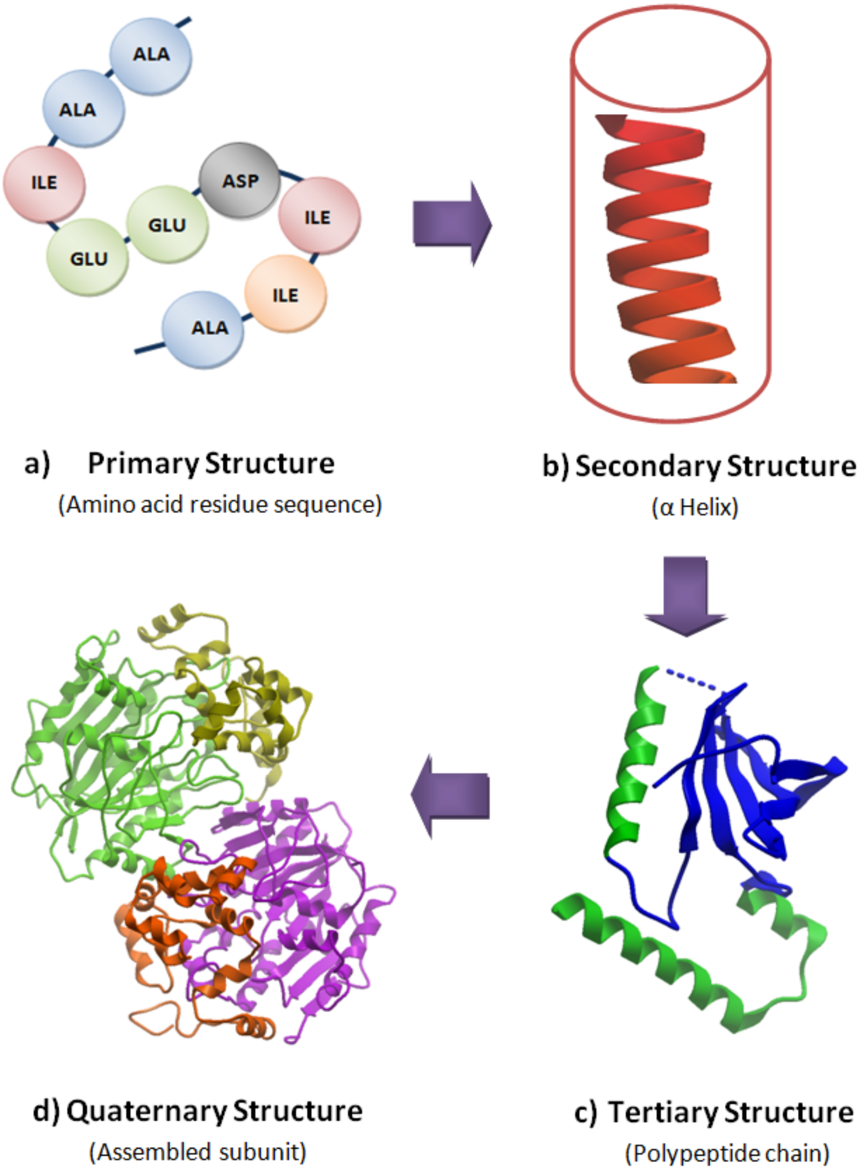


Figure 2.1: Schematic representation of protein structure.

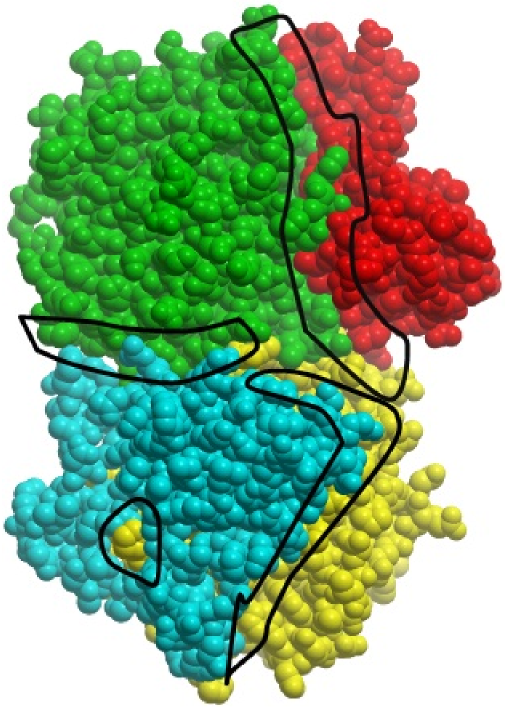


Figure 2.2: Protein-protein interaction for complex 1B4U.

- Long range interactions through the surrounding neighborhood. Interaction may take place if the molecules are more than $7\AA$ apart. But this is possible when the surrounding neighborhood such as water helps molecules to interact with each other.

If we need to understand why proteins interact with each other, we first need to know that proteins perform different biological functions and that is one of the main reasons why they interact. The protein becomes stable when the molecules correlate mutation across its interface. In our thesis, we consider only direct contact association of molecules. If we look at Figure 2.2, we can see that, complex 1B4U has 4 chains (different colors are used for atoms in different chains) and they might have direct contacts among the atoms that are in the marked area.

2.3.1 Protein-protein Interaction types

The structural and functional diversity of protein-protein interactions (PPIs) depends on the protein family and their three dimensional structures. PPIs play diverse roles in different biological processes. PPIs can give ideas about the function performed by different proteins and that is the main reason why researchers are very much interested in understanding PPIs. Based on physiological functions, specificity and evolution, PPIs can be divided into three broad non-mutually exclusive categories [20]:

1. Homo and hetero-oligomeric complexes
2. Non-obligate and obligate complexes
3. Transient and permanent complexes

PPIs can take place between identical or non-identical chains which are based on their structural similarity. If the interacting chains of an oligomer has structural symmetry then it is called homo-oligomeric PPI, otherwise it is called hetero-oligomeric PPI. Based on the composition, there are two types: obligate and non-obligate PPIs. In an obligate PPI, the proteomers are not found as stable structures on their own, which are generally functionally obligate. In our thesis we focus on this type of PPI. If we consider the life time of a complex then we can have transient and permanent PPI. Permanent PPI outputs a stable complex, but transient PPIs are less stable and they tend to continue changing their shapes until they dissociate or result in permanent complexes. To find out about the biological functions performed by different type of complexes, it is important to know about the PPIs. In our thesis we study the type that considers the composition, that is obligate and non-obligate PPI.

2.4 Protein-protein Interaction Prediction

Many researchers have been working on predicting different types of PPI with different perspective. These approaches are mainly divided into two categories, namely:

1. Experimental Approaches
2. Computational Approaches

Traditionally, the detection of PPI prediction was limited to experimental techniques such as co-immunoprecipitation or affinity chromatography [1, 7, 23]. They are labor-intensive and often the results of these approaches contain systematic errors. Since the amount of data for prediction is getting larger, these experimental processes become less applicable. This is why computational approaches are in demand. In our thesis, in order to predict the obligate and non-obligate PPIs we use a computational approach which is described in the next few subsections.

2.4.1 Prediction types

To predict obligate and non-obligate PPIs, many classifiers including random forest (RF), Bayes, decision trees, logistic regression, and support vector machines (SVM) can be used. Different classifiers achieve different performances based on the type of data and properties used. Some research in [5, 6, 24] achieved a classification accuracy of 70% for the problem of distinguishing obligate and non-obligate interactions with a wide range of parameters and different types of features such as desovation energy, amino acid composition, conservation scores, electrostatic energies, and hydrophobicity. In our thesis, we focus on the same classifiers used in [5, 6, 24] and improve the classification accuracy with our proposed features. For this, first we use principal component analysis (PCA) as a pre-processing step. PCA, though an unsupervised method, is applied to eliminate ill-conditioned matrices involved in the linear dimensionality reduction (LDR) techniques. Applying different threshold values, the desired number of principal components from a dataset can be achieved with PCA. We use different threshold values and select the threshold that leads to the highest classification accuracy. After obtaining those principal components, we classify complexes with quadratic Bayesian (QB) and linear Bayesian (LB) classifiers [5] combined with LDR methods including the well-known Fisher's discriminant analysis [5, 6] and two heteroscedastic approaches [17, 24].

If we consider groups by 18 unique atom type (N, CA, C, O, GCA, CB, KNZ, KCD, DOD, RNH, NND, RNE, SOG, HNE, YCZ, FCZ, LCD, CSG) pairs then the length of feature vector is 171 ($^{18}C_2 + 18$) and group by 20 unique amino acid (Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val) pairs then the length of feature vector is 210 ($^{20}C_2 + 20$) [29]. The basic idea of LDR is to represent a desolvation energy object of dimension n (171 or 210) as a lower-dimensional vector of dimension d, achieving this by performing a linear transformation. We consider two classes, obligate as ω_1 and non-obligate as 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 separability between atoms [24]. In our thesis, we focus on the following three LDR methods:

Fisher's discriminant analysis (FDA) [5, 6] : Its optimization criterion is as follows.

$$J_{FDA}(\mathbf{A}) = tr\left\{(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_E\mathbf{A}^t)\right\} . \tag{2.1}$$

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

Heteroscedastic discriminant analysis (HDA) [17] : It aims to obtain the matrix **A** that maximizes the function:

$$\begin{aligned} J_{HDA}(\mathbf{A}) = tr\Big\{&(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\Big[\mathbf{A}\mathbf{S}_E\mathbf{A}^t \\ &-\mathbf{A}\mathbf{S}_W^{\frac{1}{2}}\frac{p_1\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}})+p_2\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1p_2}\mathbf{S}_W^{\frac{1}{2}}\mathbf{A}^t\Big]\Big\} . \end{aligned} \tag{2.2}$$

This criterion is maximized by obtaining the eigenvectors, corresponding to the largest eigenvalues, of the matrix:

$$\begin{aligned} &\mathbf{S}_{HDA} = \mathbf{S}_W^{-1} \\ &\left[\mathbf{S}_E - \mathbf{S}_W^{\frac{1}{2}}\frac{p_1\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}})+p_2\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1p_2}\mathbf{S}_W^{\frac{1}{2}}\right] . \end{aligned} \tag{2.3}$$

Chernoff discriminant analysis (CDA) [24] : It aims to maximize the following function:

$$\begin{aligned} J_{CDA}(\mathbf{A}) = tr\{&p_1p_2\mathbf{A}\mathbf{S}_E\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1} \\ &+\log(\mathbf{A}\mathbf{S}_W\mathbf{A}^t) - p_1\log(\mathbf{A}\mathbf{S}_1\mathbf{A}^t) - p_2\log(\mathbf{A}\mathbf{S}_2\mathbf{A}^t)\}. \end{aligned} \tag{2.4}$$

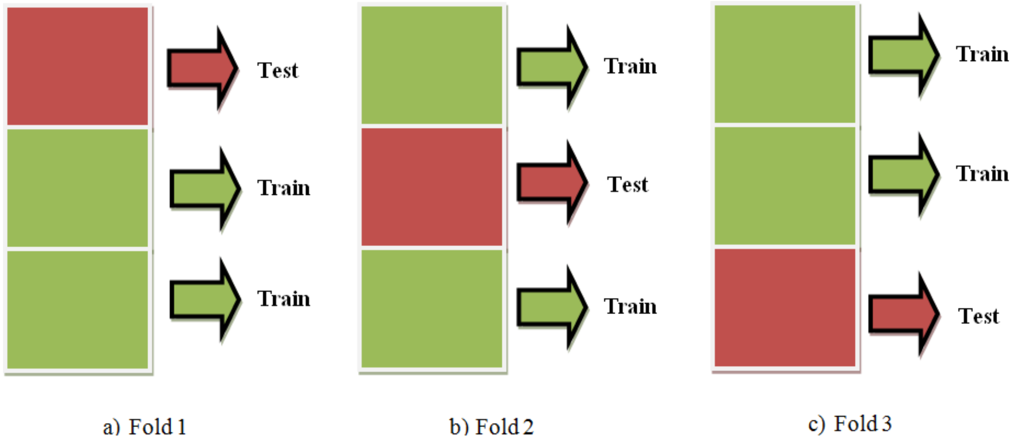


Figure 2.3: Schematic of 3-fold cross validation.

In [24], a gradient-based algorithm was proposed, which maximizes the function in an iterative way. For this gradient algorithm, a learning rate, α_k needs to be computed. In order to ensure that the gradient algorithm converges, α_k is maximized by using the secant method. One of the keys in this algorithm is the random initialization of the matrix $\mathbf{A}$, and in this work, we have performed ten different initializations and then chosen the solution for $\mathbf{A}$ that gives the maximum Chernoff distance.

2.4.2 Prediction Evaluation

K-fold cross validation is a commonly used technique which takes a set of m samples and partitions them into K sets (folds) of size m/K. For each fold, a classifier is trained on the other folds and then tested on that fold. For our experiments we use 10-fold cross validation that means, first the dataset with desolvation energy is partitioned into 10 equal sets (if possible) and then in each iteration 9 sets are chosen as training and one set is used for testing. Figure 2.3 shows an example of 3-fold cross validation.

To compute the accuracy for each classifier we use the following equation:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.5}$$

Here, TP is number of correctly identified obligate complex samples and TN is number of correctly identified non-obligate complex samples and FP and FN are number of incorrectly classified obligate and non-obligate complexes respectively.

Chapter 3

Feature Generation and Selection Methods

3.1 Features

In machine learning and pattern recognition, one of the key factors is to include and select the right features for successful prediction. They are the observed properties of each sample that is used for the prediction. The value of the features are usually numeric, but other types such as strings and graphs are also used as features.

3.2 Features used for PPI Prediction

There are many properties of PPI that can be used for PPI prediction. Some of them are [21]:

β-factor : It is the flexibility of the protein complexes during the interaction.

Solvent Accessibility : It is the exposed surface area that affects contact of atoms during the interaction.

Geometric features : The shape index, planarity or curvedness of the interacting complexes can also be used as features.

Evolutionary features : They include conservation scores or sequence profiling information.

Physicochemical features : They include hydrophobicity, electrostatic potential and desolvation energy.

3.3 Proposed Features

According to [4], knowledge-based contact potential that accounts for hypdophobic interactions, self-energy change upon desolvation of charged and polar atom groups and side-chain entropy loss is called desolvation free energy. In our thesis, we propose the use of this desolvation energy for PPI prediction as features. The desolvation energy of an atom pair *i* and *j* of a complex is defined as [4] :

$$g(r)\Sigma\Sigma e_{ij} \tag{3.1}$$

where, e_{ij} is the atomic contact potential (ACP) [18, 29] between i and j and $g(r)$ is the smooth function score, based on their distance. For simplicity, we consider the smooth function to be linear for a distance within $5\AA$-$7\AA$. We also consider the criteria that for a successful interaction, atoms should be within $7\AA$ [4]. Within $5\AA$-$7\AA$, the value of $g(r)$ varies from 1 to 0 which is equivalent to $100\% - 0\%$. We know that for a particular atom pair it has a predetermined (approximated) desolvation energy value which we can obtain

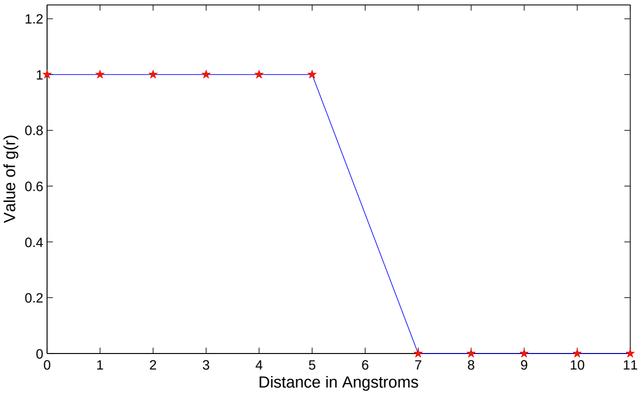


Figure 3.1: Smooth function behavior of $g(r)$.

from the ACP matrix [18, 29]. Thus, during the interaction the actual energy depends only on their distances. If the atoms are within 5$\AA$, then their actual energy will also be closer to that fixed value for those particular atom pairs in the ACP matrix in Table 3.1 [29]. If the distance is within 5$\AA$-7$\AA$, then we use the following equation to calculate smooth function value,

$$x = 7 - 2y \tag{3.2}$$

where x is the distance between an atom pair in Angstroms and the value of the smooth function for that pair is y. The behavior of the smooth function is shown in Figure 3.1.

3.4 Feature Generation

Using Equation 3.1, we can approximate the desolvation energy between two atoms. There are 18 unique type of atoms and 20 amino acids. The ACP matrix of Table 3.1 [29] is an 18×18 matrix, where all possible combinations of 18 unique atom types are represented. If we know the ligand and receptor of a complex, our first job is to find the interacting atoms. Then, we need to convert all atoms to 1 of 18 unique atom types. The method of conversion is discussed in [29]. Based on our experiments and the study of [29], we use Tables 3.2 and 3.3 for atom type conversion. In these conversion tables, the first row of each amino acid contains the original atoms that are inside and the second row contains the converted unique types. When we have the unique types we simply find that pair in the ACP matrix and obtain the value of e_{ij} of Equation 3.1 and for $g(r)$ we need to find the Euclidian distance between the two atoms, which we compute from their structural data that can be found from Protein Data Bank (PDB) [2]. Considering atom type and amino acid type, we obtain 18^2 and 20^2 features respectively. For our thesis, we use only unique pairs of atom and amino acid features which leads to feature vectors of length 171 (${}^{18}C_2 + 18$) and 210 (${}^{20}C_2 + 20$) respectively. To find the solvent accessible surface area (SASA) we use the NACCESS [9] program. This program gives atom and residue-wise information that how much of that atom/residue is exposed to solvent. To strengthen our approach, we use this information to weight our generated feature vector.

3.5 Feature Selection Methods

Feature selection involves selecting best subset of features that represents the whole feature set efficiently. If the feature vector is too large, it is wise to use feature selection to reduce the size to improve the prediction time while keeping good performance. In our thesis, we have feature vectors of length 171 and 210 respectively. We use some computational approaches [26] and visual analysis for feature selection which are discussed below.

3.5.1 Sequential Forward/Backward Search Selection

We explain this method through an example. Let us consider, the length of the original feature vector, $m = 4$ which is x_1, x_2, x_3, x_4 and our target is to select the best subset of size $l = 2$. Then, backward search selection works as follows:

- Adopt a class separability criterion, C, and compute its value for the feature vector.
- Eliminate one feature and for each of the possible resulting combinations, that is, $[x_1, x_2, x_3]$, $[x_1, x_2, x_4]$, $[x_1, x_3, x_4]$, $[x_2, x_3, x_4]$. Compute the value of the corresponding criterion C for each subset. Select the combination with the best value, say $[x_1, x_2, x_3]$.
- From the selected feature vector in the previous step eliminate one feature, and for each of the resulting combinations, $[x_1, x_2]$, $[x_1, x_3]$, $[x_2, x_3]$, compute the criterion value and select the one with the best value, say $[x_1, x_3]$.
- This process will continue until the length of the current best selected vector is equal to l.

The forward search selection can be seen as the reverse of sequential backward search selection which can be explained for the same example as follows:

- Calculate the criterion value for each of the features. Select the feature with the best class separability criterion value, say x_1.
- Form all possible next level vectors that contain the winner from the previous step, that is, $[x_1, x_2]$, $[x_1, x_3]$, $[x_1, x_4]$. Calculate the criterion value for each of them and select the best one, say $[x_1, x_3]$.
- This process will continue until the length of the current best selected vector is equal to l.

3.5.2 Floating Forward/Backward Search Selection

Both forward and backward search selection algorithms suffer from nesting-effect [26]. That is, when we discard one feature from backward selection, there is no possibility to consider this feature again throughout the process. Similarly, when we add one feature from forward selection, there is no way to discard this feature later. This problem can be solved through floating search selection. With this method, a memory is used to save the best criterion value among all the combinations and at each step it is updated, if possible. Thus, we have a process to ”backtrack” and select a different subset which gives us the best subset of features. With this backtracking technique, at every stage, the method will save more than one best subsets. If at a future stage the selected best subset is not providing any improvement, the process will discard that selection and try another from best subset from the previous step.

3.5.3 mRMR Selection

Minimum redundancy maximum relevance feature selection (mRMR) [8] is a tool that discards the redundant features from the feature vector and uses maximum relevance score as the class separability criterion. If the selected feature set is S with m features x_i which jointly have the largest dependency on the target class c and I is the criterion function for each selected subset. Then, for maximum relevance,

$$max(S,c), D = \frac{1}{S}\sum_{x_i \varepsilon S} I(x_i;\ c) \tag{3.3}$$

for minimum redundancy,

$$min(S), R = \frac{1}{S^2} \sum_{x_i, x_j \varepsilon S} I(x_i, x_j) \tag{3.4}$$

and finally, the mRMR is:

$$max\Phi(D,R), \Phi = D - R \tag{3.5}$$

This tool ranks each feature in the feature vector with Equation 3.5 and selects the best number of features based on the user's desired length for the feature vector.

3.5.4 Heatmaps

Graphical representation is a very good way to visualize and analyze data. A heatmap is a kind of graphical representation of data in which the values of a two dimensional matrix are represented with different shades of color. In our work we generate feature vectors with desolvation energies for all samples. Then, if we sum along the same type for each pair, we can have a single feature vector for each dataset of each type. For the heatmap, we consider the two dimensional vector of size 18×18 that represents 171 features and of size 20×20 that represents 210 features, and fill the values of the matrix in the upper diagonal. Figures 3.2 and 3.3 show examples of heatmaps that are used for our analysis and discussion.

3.5.5 Biological Feature Selection

According to [22], if we consider the polarity of amino acids, they can be of the following three types:

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

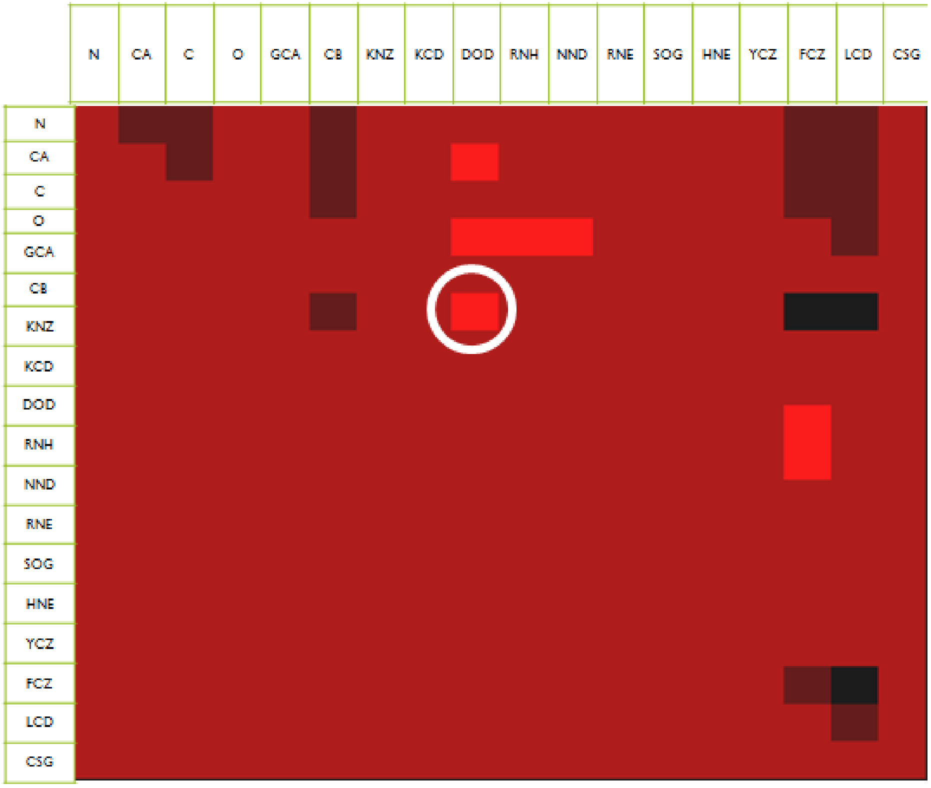


Figure 3.2: Obligate samples (in red color representation).

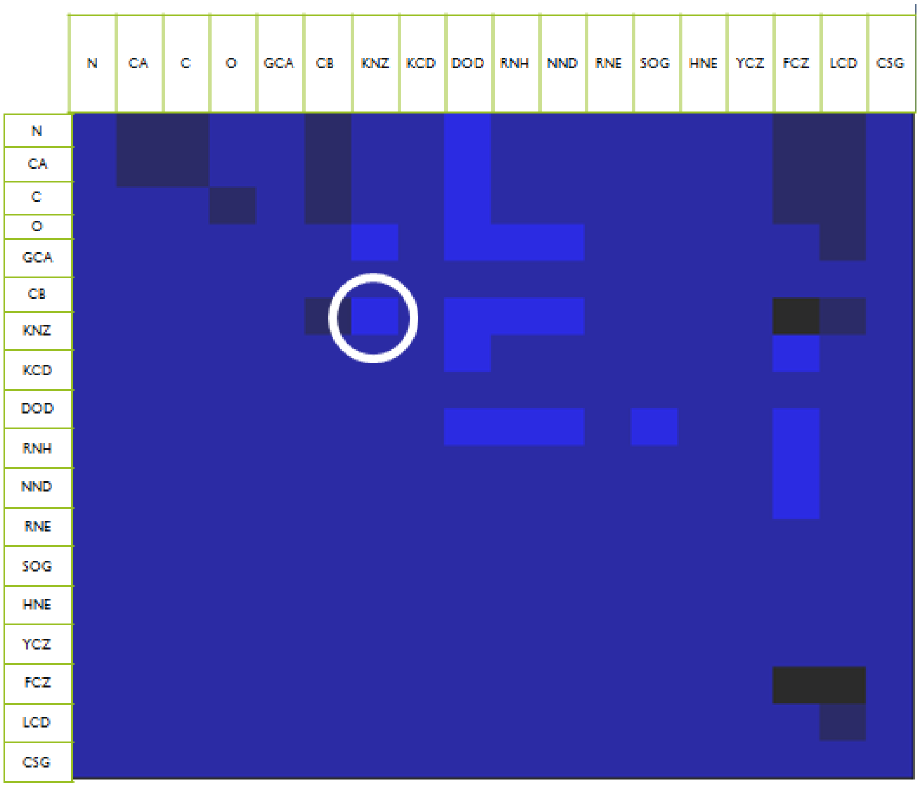


Figure 3.3: Non-obligate samples (in blue color representation).

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

Amphipathic : It has both polar and nonpolar behavior and therefore a tendency to form interfaces between hydrophobic and hydrophilic molecules. Lysine, Tyrosine, Methionine, Tryptophan and Threonine are amphipathic amino acids.

Using this information we can group the desolvation energy values by amino acid type into three sub categories with only hydrophobic, only hydrophilic and only amphipathic type. Then, we can use our classifiers to check the accuracy to find the distinguishing features of obligate and non-obligate complexes.

Table 3.1: ACP matrix.

	N	CA	C	O	GCA	CB	KNZ	KCD	DOD	RNH	NND	RNE	SOG	HNE	YCZ	FCZ	LCD	CSG
N	-0.724	-0.903	-0.722	-0.322	-0.331	-0.603	1.147	0.937	0.752	0.502	0.405	0.495	-0.116	-0.092	-0.412	-0.499	-1.005	-2.06
CA	-0.903	-0.842	-0.85	-0.332	-0.365	-0.531	1.139	0.918	0.852	0.452	0.445	0.436	-0.044	-0.165	-0.539	-0.59	-1.123	-2.028
C	-0.722	-0.85	-0.704	-0.371	-0.308	-0.499	1.14	0.846	0.788	0.49	0.408	0.418	-0.045	-0.089	-0.415	-0.478	-0.963	-2.033
O	-0.322	-0.332	-0.371	-0.016	0.205	-0.059	1.414	1.075	1.152	0.831	0.758	0.649	0.383	0.241	-0.073	-0.166	-0.65	-1.65
GCA	-0.331	-0.365	-0.308	0.205	0.182	-0.009	1.502	1.385	1.192	0.912	0.872	0.943	0.434	0.392	-0.062	-0.021	-0.342	-1.212
CB	-0.603	-0.531	-0.499	-0.059	-0.009	-0.469	1.31	1.01	0.859	0.671	0.591	0.565	0.122	0	-0.438	-0.533	-1.034	-1.8
KNZ	1.147	1.139	1.14	1.414	1.502	1.31	3.018	2.911	1.157	2.635	1.848	2.699	1.587	1.557	1.004	1.34	1.252	0.7
KCD	0.937	0.918	0.846	1.075	1.385	1.01	2.911	2.811	0.989	2.439	1.622	2.525	1.361	1.277	0.685	0.938	0.814	0.411
DOD	0.752	0.852	0.788	1.152	1.192	0.859	1.157	0.989	1.978	0.695	1.386	0.641	1.002	0.642	0.618	0.894	0.792	-0.029
RNH	0.502	0.452	0.49	0.831	0.912	0.671	2.635	2.439	0.695	1.589	1.395	1.515	1.022	0.948	0.457	0.707	0.586	0.021
NND	0.405	0.445	0.408	0.758	0.872	0.591	1.848	1.622	1.386	1.395	1.3	1.283	0.931	0.829	0.484	0.633	0.389	-0.046
RNE	0.495	0.436	0.418	0.649	0.943	0.565	2.699	2.525	0.641	1.515	1.283	1.309	0.921	0.777	0.265	0.484	0.274	-0.253
SOG	-0.116	-0.044	-0.045	0.383	0.434	0.122	1.587	1.361	1.002	1.022	0.931	0.921	0.559	0.319	0.301	0.212	-0.155	-0.949
HNE	-0.092	-0.165	-0.089	0.241	0.392	0	1.557	1.277	0.642	0.948	0.829	0.777	0.319	-0.301	-0.159	-0.132	-0.338	-1.324
YCZ	-0.412	-0.539	-0.415	-0.073	-0.062	-0.438	1.004	0.685	0.618	0.457	0.484	0.265	0.301	-0.159	-0.314	-0.478	-0.964	-1.41
FCZ	-0.499	-0.59	-0.478	-0.166	-0.021	-0.533	1.34	0.938	0.894	0.707	0.633	0.484	0.212	-0.132	-0.478	-0.687	-1.24	-1.784
LCD	-1.005	-1.123	-0.963	-0.65	-0.342	-1.034	1.252	0.814	0.792	0.586	0.389	0.274	-0.155	-0.338	-0.964	-1.24	-1.873	-2.402
CSG	-2.06	-2.028	-2.033	-1.65	-1.212	-1.8	0.7	0.411	-0.029	0.021	-0.046	-0.253	-0.949	-1.324	-1.41	-1.784	-2.402	-3.742

Table 3.2: Atom conversion table (a).

Amino acid	Conversion											
ALA	N	CA	CB	C	O	OXT						
	N	CA	CB	C	O	O						
ARG	N	CA	CB	CG	CD	NE	CZ	NH1	NH2	C	O	OXT
	N	CA	CB	FCZ	RNE	RNE	RNH	RNH	RNH	C	O	O
ASN	N	CA	CB	CG	OD1	ND2	C	O	OXT			
	N	CA	CB	NND	NND	NND	C	O	O			
ASP	N	CA	CB	CG	OD1	OD2	C	O	OXT			
	N	CA	CB	DOD	DOD	DOD	C	O	O			
CYS	N	CA	CB	SG	C	O	OXT					
	N	CA	CB	CSG	C	O	O					
GLN	N	CA	CB	CG	CD	OE1	NE2	C	O	OXT		
	N	CA	CB	FCZ	NND	NND	NND	C	O	O		
GLU	N	CA	CB	CG	CD	OE1	OE2	C	O	OXT		
	N	CA	CB	FCZ	DOD	DOD	DOD	C	O	O		
GLY	N	CA	C	O	OXT							
	N	GCA	C	O	O							
HIS	N	CA	CB	CG	ND1	CD2	NE2	CE1	C	O	OXT	
	N	CA	CB	HNE	HNE	HNE	HNE	HNE	C	O	O	
ILE	N	CA	CB	CG2	CG1	CD	C	O	OXT	CD1		
	N	CA	CB	LCD	FCZ	LCD	C	O	O	LCD		

Table 3.3: Atom conversion table (b).

Amino acid	Conversion														
LEU	N	CA	CB	CG	CD1	CD2	C	O	OXT						
	N	CA	CB	FCZ	LCD	LCD	C	O	O						
LYS	N	CA	CB	CG	CD	CE	NZ	C	O	OXT					
	N	CA	CB	FCZ	KCD	KNZ	KNZ	C	O	O					
MET	N	CA	CB	CG	SD	CE	C	O	OXT						
	N	CA	CB	FCZ	FCZ	LCD	C	O	O						
PHE	N	CA	CB	CG	CD1	CD2	CE1	CE2	CZ	C	O	OXT			
	N	CA	CB	FCZ	FCZ	FCZ	FCZ	FCZ	FCZ	C	O	O			
PRO	N	CA	CB	CG	CD	C	O	OXT							
	N	CA	CB	CB	CB	C	O	O							
SER	N	CA	CB	OG	C	O	OXT								
	N	CA	SOG	SOG	C	O	O								
THR	N	CA	CB	OG1	CG2	C	O	OXT							
	N	CA	CB	SOG	FCZ	C	O	O							
TRP	N	CA	CB	CG	CD2	CE2	CE3	CD1	NE1	CZ2	CZ3	CH2	C	O	OXT
	N	CA	CB	FCZ	FCZ	FCZ	FCZ	FCZ	HNE	FCZ	FCZ	FCZ	C	O	O
TYR	N	CA	CB	CG	CD1	CE1	CD2	CE2	CZ	OH	C	O	OXT		
	N	CA	CB	FCZ	FCZ	YCZ	FCZ	YCZ	YCZ	SOG	C	O	O		
VAL	N	CA	CB	CG1	CG2	C	O	OXT							
	N	CA	CB	LCD	LCD	C	O	O							

Chapter 4

Methodology

To predict the types of obligate and non-obligate protein complexs with good accuracy, we follow the proposed model as depicted in Figure 4.1. This process starts from property selection and continue to the post-analysis step to evaluate the behavior of the proposed features to improve the prediction.

4.1 Procedure

A description of the procedure follows:

Step 1: (Preparing the dataset)

Merge the Obligate dataset obtained from Mintseris *et al.* [19] and Zhu *et al.* [30].

Merge the Non-obligate dataset obtained from Mintseris *et al.* [19] and Zhu *et al.* [30].

Remove redundant complexes and complexes with contradicting labels.

Convert all the complexes with multiple chains into two-chain by applying a suitable distance threshold.

Add all the converted single chain complexes and remove those multiple chain com-

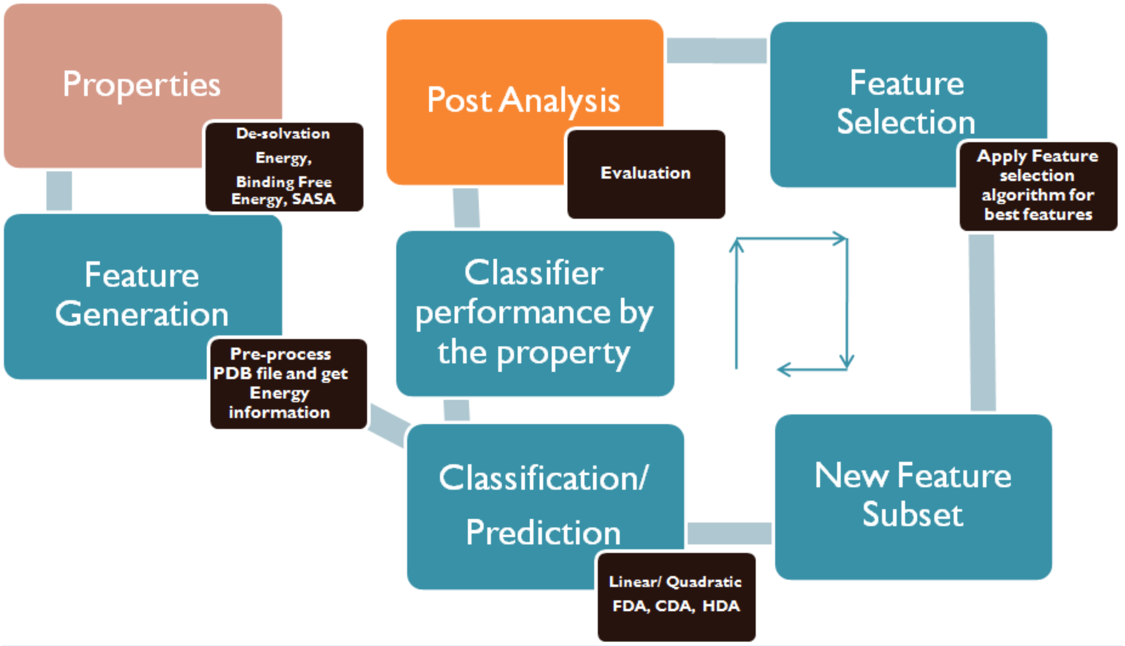

Properties
De-solvation Energy, Binding Free Energy, SASA
Feature Generation
Pre-process PDB file and get Energy information
Post Analysis
Evaluation
Classifier performance by the property
Classification/ Prediction
Linear/ Quadratic FDA, CDA, HDA
Feature Selection
Apply Feature selection algorithm for best features
New Feature Subset

Figure 4.1: Proposed model to classify obligate and non-obligate interactions.

plexes to obtain the binary protein-protein interaction dataset (BPPI).

Step 2: (Initialization of the properties)

Set the desolvation energy equation to find the desolvation energy between two atoms.

Set the equation for calculating the surface area for each of the complexes.

Step 3: (Downloading structure information of each complex)

Download the structural files for all the complexes from the PDB [2].

Step 4: (Gathering information required for calculation)

Remove everything except the information about the ATOM of the complexes

(Atom name, atom number, chain name, residue name, residue number, x-y-z coordinates and occupancy factor)

Combine all the occupancy factors to add up to 1, if it is less than 1 for all the same atoms within the same amino acid.

Seperate ligand and receptor for each complex based on to their chain information.

Step 5: (Calculating SASA)

For each complex in the dataset, run NACCESS [9] program separately for each chain in the input dataset.

Step 6: (Calculate the feature vector with desolvation energy)

```
For each complex in the dataset
    For each atom in the ligand
        Calculate the Euclidean distance to all others atoms in the receptor
        If atom pair distance is less than or equal to $7\AA$ then do
            Map Ligand/Receptor atom type to one of 18 unique type atoms [29]
            Find the atomic contact potential from the ACP matrix [29]
            Find the value of $g(r)$ using Equation 3.2
            Find SASA values from Naccess for that ligand atom
            Calculate the desolvation energy value using Equation 3.1
            Find the position of the unique atom-atom pair in the feature vector
            Accumulate the desolvation energy value in that position
            Multiply this value by the SASA value to obtain a weighted feature vector
            If the input complex list belongs to Obligate then
                set 1 to the class label
            Else
```

add 2 to the class label

Step 7: (Feature extraction)

(Principal component analysis)

If the final dataset (with/without SASA or by amino acid/atom) for prediction has a large number of zeros then apply PCA with different thresholds to reduce the feature vector with fewer zeros

Step 8: (Feature selection algorithms)

Apply different feature selection algorithms (Forward/backward/floating and mRMR)

Find the best feature subset that gives highest value for the objective function

Step 9: (Prediction)

Apply different LDR (HDA, FDA, CDA) combined with quadratic Bayesian (QB) and linear Bayesian (LB) classifiers [5], to test the quality of the features (desolvation energies).

Step 10: (Post analysis)

Generate the Heatmap of obligate and non-obligate complexes for both atom type and amino acid type

Combine the Heatmaps to find the best pair(s) that can predict the complexes

Find the common amino acid pairs from the heatmaps and the feature selection methods

Perform biological analysis

4.2 Flow Diagram

These steps can also be easily understood with a graphical tool called data flow diagram (DFD) in Figure 4.2.

4.3 Dataset Preparation

We worked on two well-known datasets, namely Mintseris *et al.* [19] and Zhu *et al.* [30] which contain obligate and non-obligate complex names with the corresponding chain information. All the complexes in Zhu *et al.* dataset have the characteristics that one chain is interacting with only one chain, while in Mintseris *et al.* dataset there are complexes which have more than two chains in the interaction.

We have compiled a new dataset by merging these two datasets. Zhu *et al.* dataset contains 75 obligate and 62 non-obligate complex names with interacting chains, and Mintseris *et al.* dataset contains 115 obligate and 212 non-obligate complexes. There are 39 redundant complexes in those two datasets and 7 complexes (1eg9, 1hsa, 1i1a, 1raf, 1d09, 1jkj and 1cqi) had contradicting class labels. For example if we find a complex ”A” is defined as obligate in one dataset and in other dataset it is defined as non-obligate then we conclude the complex ”A” as contradicting complex. Thus, in the first step we removed all the contradicting and redundant complexes and generated the merged dataset which contains 182 obligate and 235 non-obligate complexes.

The second step is a pre-processing stage. After counting, we found that the merged dataset from the first step contains 93 complexes which have multiple chain interactions. Now, to make the dataset with similar characteristics, those 93 complexes were removed and copied to a new dataset so that we can convert them and finally add them to the merged

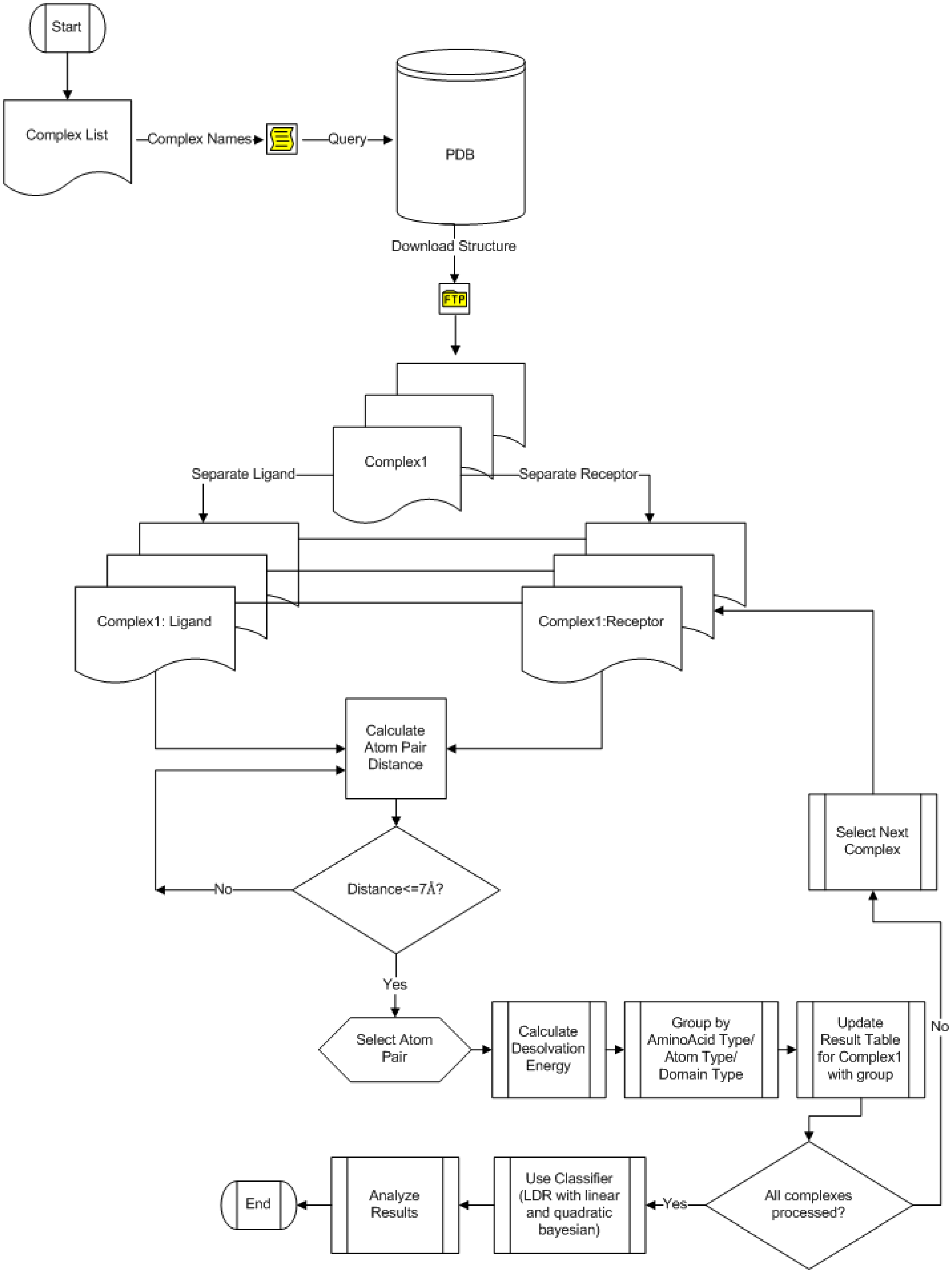


Figure 4.2: Flow diagram for processing the data.

datasets. In this pre-processing stage, we first convert each multiple chain complex into a set of binary chain complexes. For example, if a complex has multiple chain information such as $A\ B : C\ D$ then it is converted into binary chain complexes: $A : C$, $A : D$, $B : C$ and $B : D$. After this we did an experiment to find out the number of interacting residues on the surface of the interacting chain of those converted complexes for different distance thresholds ($4\AA$, $4.5\AA$, $5\AA$ and $6\AA$). The experimental results for this conversion are listed in Table A.1 in Appendix A.

From the study of [11], using the interface definition we specified the following two criteria for filtering:

- Each pair of interacting residues from different chains must be within $5\AA$ distance for successful interaction.
- For each complex there must be five interacting residues on the interface.

We removed all the complexes which did not satisfy both of these criteria. After considering all these complexes, we obtain the merged dataset that has 516 complexes in which 303 are non-obligate and 213 are obligate complexes. We call this final merged dataset binary protein-protein interaction (BPPI) dataset. This BPPI dataset of obligate and non-obligate complexes with their interacting binary chain information used for all the later experiment are listed in Tables 4.1 and 4.2

Table 4.1: Obligte BPPI dataset (213 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	

Table 4.2: Non-obligate BPPI dataset (303 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	

Chapter 5

Results and Discussion

5.1 Dataset Description

We tested our prediction model on two well-known data sets and on our compiled BPPI dataset. The datasets used for the experiments are as follows:

Table 5.1: Dataset description.

Dataset name	No of obligate complex	No of non-obligate complex
Mintseris *et al.* [19] dataset	115	212
Zhu *et al.* [30] dataset	75	62
BPPI dataset	213	303

These datasets includes predefined class label of obligate and non-obligate interaction. For our experiments we used different types of combinations such as:

One against one : All possible combinations of binary interactions are considered. For example, if we have a complex with chains AB:CD, then we use A:C, A:D, B:C and B:D for calculation.

All against all : Only the original combination is considered. For example, if we have a complex with ABC:DE, then we use X:Y for calculation where X is equal to all atoms in chain A, B, C and Y is equal to all atoms in chain D and E.

With SASA : The energy calculations are weighted by SASA values.

Without SASA : The energy calculations do not include SASA values.

We use the following acronyms listed in Table 5.2 for datasets with different combination for our experiments.

Table 5.2: Acronyms used for the datasets.

Acronym	Dataset Description
MAS	Mintseris *et al.* [19] dataset all against all with SASA
MAW	Mintseris *et al.* [19] dataset all against all without SASA
MOS	Mintseris *et al.* [19] dataset one against one with SASA
MOW	Mintseris *et al.* [19] dataset one against one without SASA
ZS	Zhu *et al.* [30] with SASA
ZW	Zhu *et al.* [30] without SASA
BPPI-S	Merged dataset with SASA
BPPI-W	Merged dataset without SASA

5.2 Experimental Results

5.2.1 Prediction with proposed features

Based on our prediction model in Figure 4.1, first we use desolvation energies and SASA values as properties. In the next step, we calculated desolvation energy features with unique pairs of amino acids and unique pairs of atom types using the datasets mentioned in Table 5.1. To reduce the possibility of singularity during prediction, we use PCA with threshold values : 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6} and 10^{-7}. Then, we applied different LDR (HDA, FDA, CDA) combined with quadratic Bayesian (QB) and linear Bayesian (LB) classifiers to achieve maximum 82.13%, 80.86% and 74.38% accuracies on Zhu *et al.*, Mintseries *et al.* and BPPI dataset respectively. The details of the prediction accuracies with different combinations of desolvation energies are listed in Tables 5.3, 5.4 and 5.5.

Table 5.3: Classification results for desolvation properties with unique pairs on Zhu *et al.* dataset.

	Quadratic			Linear		
with atom type	FDA	HDA	CDA	FDA	HDA	CDA
ZW	66.05	74.75	76.29	66.05	82.13	71.85
ZS	64.62	73.42	72.76	66.16	80.66	74.51
	Quadratic			Linear		
with amino acid type	FDA	HDA	CDA	FDA	HDA	CDA
ZW	65.38	71.97	72.08	64.61	78.39	55.45
ZS	56.27	40.71	73.62	50.19	53.96	57.04

Table 5.4: Classification results for desolvation properties with unique pairs on Mintseris *et al.* dataset.

	Quadratic			Linear		
with atom type	FDA	HDA	CDA	FDA	HDA	CDA
MAW	70.58	77.96	78.88	69.94	78.26	77.03
MAS	75.49	78.53	77.00	73.94	74.26	74.84
MOW	73.16	80.09	80.86	69.52	78.54	75.11
MOS	77.00	79.70	78.36	77.19	77.40	75.10
	Quadratic			Linear		
with amino acid type	FDA	HDA	CDA	FDA	HDA	CDA
MAW	69.56	76.40	73.62	68.68	77.65	65.65
MAS	68.69	76.13	71.80	68.40	73.67	65.92
MOW	76.43	79.31	76.82	73.16	77.39	72.22
MOS	75.86	78.93	76.43	75.08	77.39	72.60

Table 5.5: Classification results for desolvation properties with unique pairs on the BPPI dataset.

	Quadratic			Linear		
With atom type	FDA	HDA	CDA	FDA	HDA	CDA
BPPI-W	71.85	72.05	74.38	72.43	73.58	73.79
BPPI-S	64.77	72.25	71.65	63.02	73.55	71.04
	Quadratic			Linear		
with amino acid type	FDA	HDA	CDA	FDA	HDA	CDA
BPPI-W	68.88	72.25	74.17	69.47	73.57	73.55
BPPI-S	69.67	73.02	71.45	69.48	72.19	68.84

5.2.2 Prediction with Related Works

In order to compare the results with some related works [19, 30], we computed the NOXclass features (interface area, interface area ratio, amino acid composition of the interface, correlation between amino acid compositions of interface and protein surface) [30] for all three datasets, and used our classification methods to find the accuracies listed in Table 5.6. The BPPI dataset achieved maximum accuracy of 74.20%, Mintseries *et al.* dataset achieved maximum of 77.32% and Zhu *et al.* dataset achieved maximum accuracy of 77.92% with our method and NOXclass features. The details of the classification results are listed in Table 5.6.

Table 5.6: Classification results for NOXclass properties with different datasets.

	Quadratic			Linear		
	FDA	HDA	CDA	FDA	HDA	CDA
BPPI	74.20	69.90	71.27	72.65	70.10	70.88
Mintseris *et al.* [19]	77.32	76.45	75.20	77.32	76.42	74.90
Zhu *et al.* [30]	77.15	67.99	60.47	77.92	65.79	62.16

5.2.3 Prediction with Biochemical Groups

Based on study of [22], to perform a biological analysis, we separated desolvation energies of amphipathic-amphipatic, hydrophobic-hydrophobic and hydrophilic-hydrophilic pairs and measured the performance of our classification methods on those biochemical groups. If we need to pick two out of n elements, the total number of possible combinations is $\frac{n(n-1)}{2}+n$. Thus, by taking 5 amphipatic amino acids, 8 hydrophilic amino acids and 6 hydrophobic amino acids we selected 15, 36 and 21 combinations of amino acid pairs respectively. The classification results are listed in Tables 5.7, 5.8 and 5.9. Among all three datasets, the highest accuracy achieved by amphipatic, hydrophilic and hydrophobic groups are 72.05%, 76.43% and 77.60% respectively.

Table 5.7: Classification results for desolvation energy grouped by amphipatic amino acids.

	Quadratic			Linear		
	FDA	HDA	CDA	FDA	HDA	CDA
BPPI-W	61.86	64.80	64.81	62.45	64.81	65.37
BPPI-S	62.44	63.42	64.00	63.23	64.98	63.81
MAW	64.80	65.76	66.36	64.52	65.47	66.02
MOW	71.47	71.85	72.05	70.89	71.28	71.86
MAS	66.59	66.30	67.53	65.68	66.67	65.68
MOW	72.23	72.81	73.39	72.81	70.89	70.90
ZW	61.03	67.72	66.24	61.75	66.14	63.26
ZS	59.48	66.19	62.44	59.49	61.73	62.56

Table 5.8: Classification results for desolvation energy grouped by hydrophilic amino acids.

	Quadratic			Linear		
	FDA	HDA	CDA	FDA	HDA	CDA
BPPI-W	66.92	68.70	$\underline{69.09}$	66.93	68.11	68.32
BPPI-S	67.49	$\underline{68.48}$	67.70	67.11	67.11	68.47
MAW	68.40	$\underline{71.49}$	$\underline{71.49}$	68.39	70.28	68.11
MOW	74.71	$\underline{\underline{76.43}}$	76.05	73.94	73.56	72.99
MAS	67.79	$\underline{70.86}$	69.65	67.16	68.75	65.36
MOS	73.94	74.90	$\underline{75.10}$	72.59	74.71	73.94
ZW	56.17	59.71	60.37	57.60	$\underline{68.97}$	60.10
ZS	61.34	58.78	62.31	63.55	$\underline{64.77}$	58.88

Table 5.9: Classification results for desolvation energy grouped by hydrophobic amino acids.

	Quadratic			Linear		
	FDA	HDA	CDA	FDA	HDA	CDA
BPPI-W	72.41	72.61	72.80	$\underline{74.74}$	73.34	74.12
BPPI-S	71.85	71.66	71.07	$\underline{72.82}$	71.79	72.38
MAW	73.36	75.51	$\underline{75.54}$	73.67	75.53	75.52
MOW	76.06	$\underline{\underline{77.60}}$	77.40	76.07	77.39	77.02
MAS	73.67	$\underline{75.51}$	75.23	74.58	74.29	74.93
MOS	75.86	$\underline{77.21}$	76.63	76.25	77.20	75.86
ZW	$\underline{72.61}$	76.23	73.38	74.04	75.78	77.07
ZS	68.05	73.27	75.42	70.30	72.66	$\underline{76.85}$

5.2.4 Prediction with Feature Selection Methods

5.2.4.1 With Floating Forward/backward Feature Selection

Using the floating forward/backward feature selection algorithms described in Section 3.5, we did two experiments to select the best subsets of features. In the first experiment, we let the algorithm choose the best subset of minimum size upto 1 and in the second one, we fixed the size to 60 (one third of the length of unique amino acid type features). The selected features achieved maximum accuracies of 71.05%, 77.83% and 77.39% for BPPI, Zhu *et al.* and Mintseries *et al.* datasets respectively. The results are listed in Table 5.10.

Table 5.10: Classification results for desolvation energy with floating forward/backward feature selection (minimum length 1).

	Quadratic			Linear		
with atom type	FDA	HDA	CDA	FDA	HDA	CDA
BPPI-S	70.86	70.47	70.66	71.05	69.30	69.88
MAW	69.71	70.31	70.93	69.98	70.31	69.39
MAS	72.75	72.44	72.14	72.44	72.14	71.52
MOS	76.07	76.45	75.87	77.39	76.63	76.43
ZW	67.99	73.37	75.36	67.23	77.83	75.06
	Quadratic			Linear		
with amino acid type	FDA	HDA	CDA	FDA	HDA	CDA
BPPI-W	66.54	66.75	66.54	66.34	66.32	66.91
BPPI-S	62.24	61.85	61.45	61.85	61.85	61.26
MOS	71.85	71.66	72.42	70.89	71.47	70.12
MOW	73.57	74.33	74.14	72.99	73.94	72.99
ZW	67.13	65.44	64.71	65.49	65.48	64.14
ZS	61.93	60.55	60.55	53.99	53.22	53.22

In experiment one, some combinations resulted the best subsets of length 1 by the selection algorithm. Thus, for those combinations, there are no classification results in Table 5.10. All selected amino acid pairs for this experiment are listed in Table 5.11.

Table 5.11: Selected amino acid pairs with floating forward/backward feature selection (minimum length 1).

Dataset	Selected amino acid pairs
BPPI-W	GLU-GLY, ILE-VAL, LEU-TYR, PRO-THR
BPPI-S	SER-THR, TYR-VAL
ZW	PHE-TYR, PRO-PRO, THR-THR, THR-VAL
ZS	TRP-TRP, TYR-TYR, VAL-VAL
MAS	PHE-TYR
MOS	GLU-TYR, LYS-TYR, PHE-SER, PRO-TRP, THR-TRP, THR-TYR, VAL-VAL
MAW	VAL-VAL
MOW	SER-THR, SER-VAL, THR-TRP, TRP-TRP, TRP-TYR, TRP-VAL, TYR-TYR

In experiment two, we selected 60 pairs of amino acid with floating forward/backward feature selection algorithm.

5.2.4.2 With mRMR Feature Selection

With mRMR, we selected the top 60 amino acid pairs. Then, we merged all the amino acid pairs selected by mRMR and floating forward/backward feature selection with length 60 and counted frequency of different amino acid pairs in that list. The frequency histogram is shown in Figure 5.1. We also tested the strength of the mRMR selected features with our prediction method. The classification results with 60 mRMR features are listed in Tables 5.12 and 5.13. For those classifications, the highest accuracy achieved by BPPI, Mintseris *et al.* and Zhu *et al.* datasets are 74.89%, 80.27% and 82.18% respectively.

5.2.4.3 Analysis With Heatmaps

To generate heatmaps we created a 19×19 matrix with row labels [LYS, MET, THR, TRP, TYR, ARG, ASN, ASP, CYS, GLN, GLU, HIS, SER, ALA, ILE, LEU, PHE, PRO, VAL] and column labels [LYS, MET, THR, TRP, TYR, ARG, ASN, ASP, CYS, GLN, GLU,

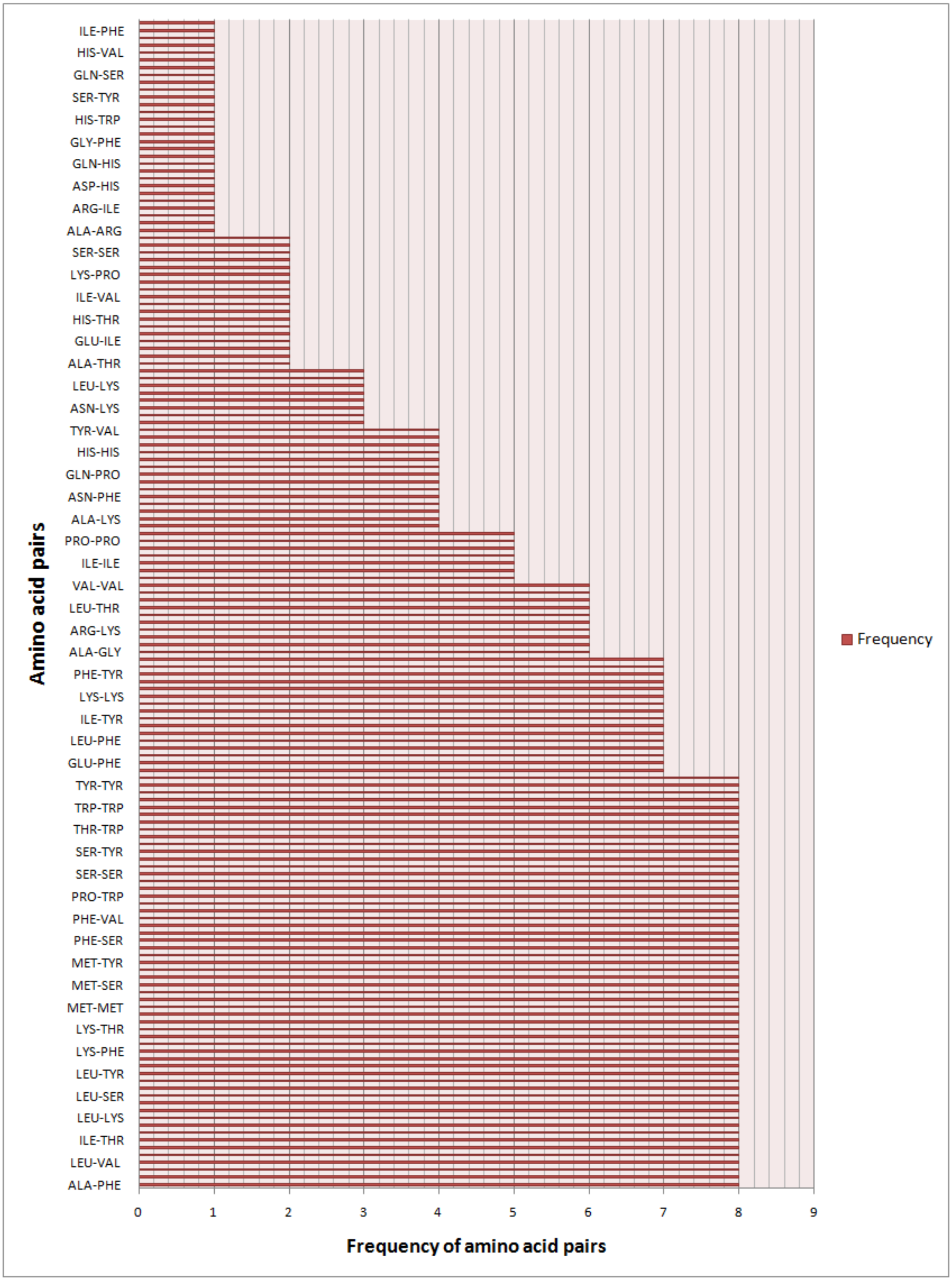


Figure 5.1: Frequency histogram of amino acid pairs selected by mRMR and floating forward/backward feature selection.

Table 5.12: Classification results for desolvation energy of amino acid pairs selected with mRMR.

	Quadratic			Linear		
	FDA	HDA	CDA	FDA	HDA	CDA
BPPI-W	71.46	73.81	71.85	72.05	74.53	74.71
BPPI-S	71.44	73.61	72.43	72.42	73.93	74.89
MAW	74.94	78.55	77.35	73.08	78.59	76.12
MAS	74.60	77.38	76.76	73.65	76.44	74.26
MOW	77.40	79.70	77.97	77.02	78.15	76.06
MOS	77.39	80.08	77.21	77.00	79.30	75.09
ZW	68.36	79.05	76.29	67.64	82.18	72.70
ZS	66.36	76.91	78.43	67.80	82.07	70.05

Table 5.13: Classification results for desolvation energy of atom type pairs selected with mRMR.

	Quadratic			Linear		
	FDA	HDA	CDA	FDA	HDA	CDA
BPPI-W	68.90	71.66	72.05	69.09	74.72	74.71
BPPI-S	68.90	71.66	72.04	69.29	72.93	72.94
MAW	73.93	77.03	78.9	73.91	77.67	77.36
MAS	73.34	77.35	77.64	72.09	77.03	77.91
MOW	79.70	80.08	80.27	78.17	79.12	78.54
MOS	77.02	79.89	80.08	76.82	78.55	77.59
ZW	69.04	74.91	76.40	67.65	81.31	77.22
ZS	70.57	73.93	75.26	70.57	78.95	75.58

HIS, SER, ALA, ILE, LEU, PHE, PRO, VAL$]^T$. First, we took the column-wise sums for all the feature vectors, and then sorted desolvation energies for amino acid pairs with amphipatic-amphipatic, hydrophilic-hydrophilic and hydrophobic-hydrophobic pairs and filled the matrix (we set zero for rest of the entries in the matrix). We created two separate matrices for obligate and non-obligate for each datasets. We plotted obligate features in the red color heatmap and non-obligate in the blue color heatmap. To perform a visual analysis, we combined obligate and non-obligate heatmaps into a single heatmap to see the blue and red colored points in the heatmap. When we have two individual matrices of obligate and non-obligate, we also calculated their difference matrix. We used this difference matrix to generate the difference heatmap. The difference heatmap is plotted with standardized values. Thus, it is represented in red-green color. We generated the heatmaps for 3×3, 4×4, 5×5, 6×6, 7×7 and 8×8 color resolutions. The best heatmaps are shown in Figures 5.2, 5.3 and 5.4. Other heatmaps are shown in Appendix B (Figures B.1, B.2, B.3, B.4, B.5 and B.6).

For all three datasets, in the combined heatmap, we can see that the bottom right corners (which represents hydrophobic-hydrophobic pairs) are different and higher energies than other two groups. In the difference heatmap, we can see that this same group is expressed with red color (red means higher energies).

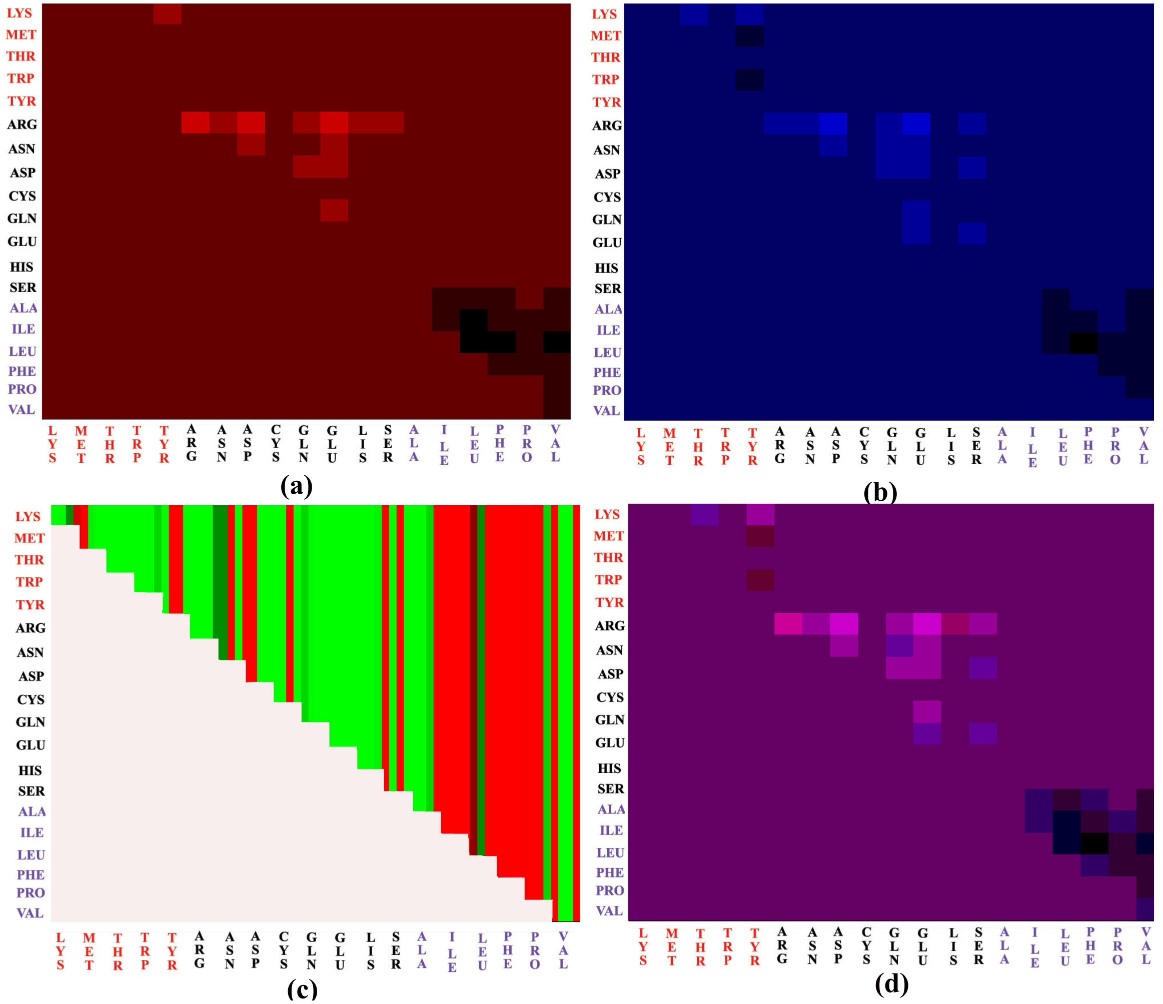
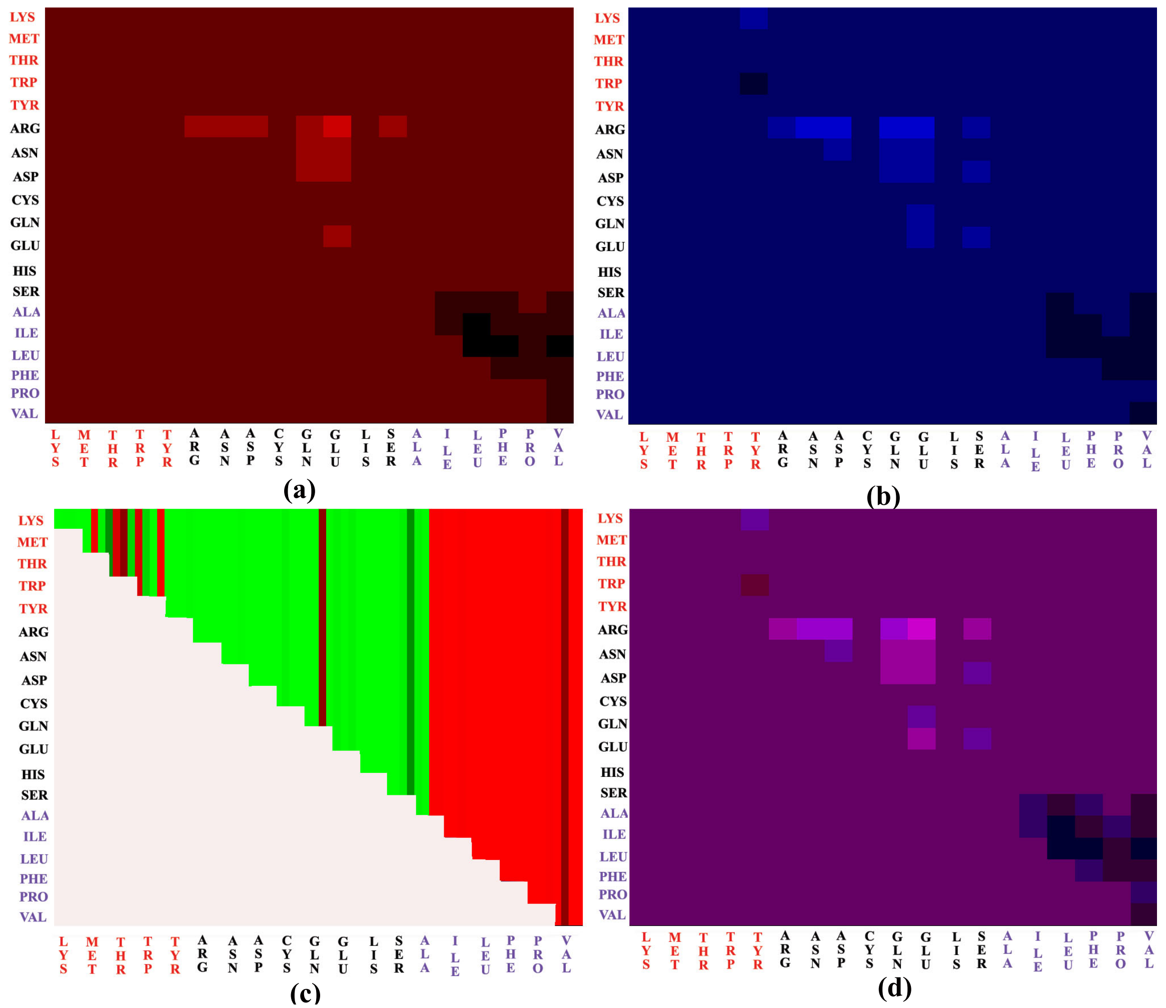


Figure 5.2: Mintseris *et al.* [19] heatmaps- (a) obligate, (b) non-obligate, (c) difference of (a) and (b), (d) combined of (a) and (b), - with color resolution 5×5.



(a)

(b)

(c)

(d)

Figure 5.3: Zhu *et al.* [30] heatmaps- (a) obligate, (b) non-obligate, (c) difference of (a) and (b), (d) combined of (a) and (b), - with color resolution 5×5.

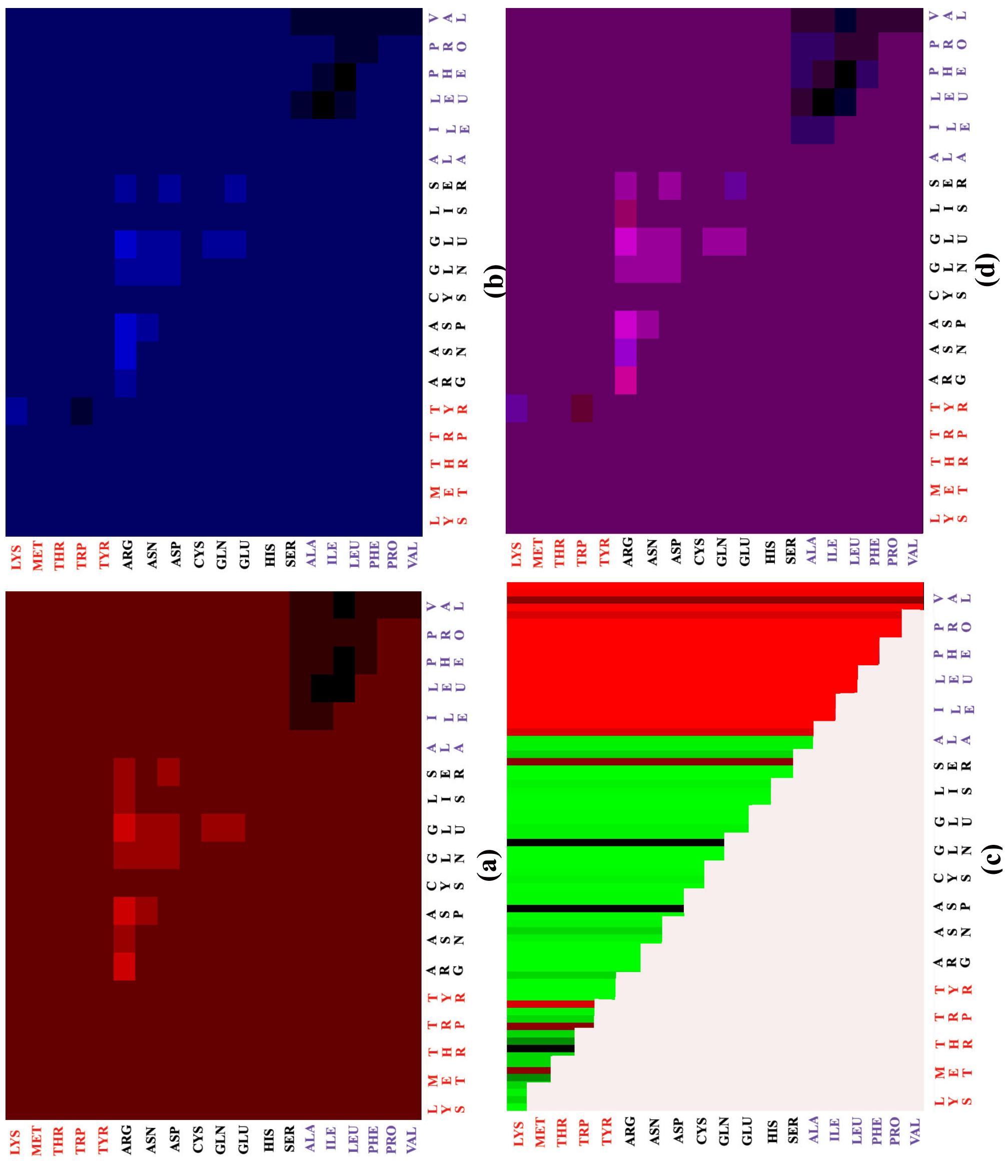


Figure 5.4: BPPI-W heatmaps- (a) obligate, (b) non-obligate, (c) difference of (a) and (b), (d) combined of (a) and (b), - with color resolution 5×5.

5.3 Discussion and Comparison

In paper [30], Zhu *et al.* predicted the type of obligate and non-obligate complexes with 70.07% accuracy. They used four NOXclass features with two staged SVM [3] classifier to achieve that performance. With our approach, that is LDR (HDA, FDA, CDA) combined with quadratic Bayesian (QB) and linear Bayesian (LB), we achieved maximum accuracy of 82.13%. Thus, we have over 12% improved result for this case. To make a fair comparison, we also used the same features with our prediction methods and it achieved maximum accuracy of 77.92% which is still lower than the accuracy of our approach by 4%. With the same four NOXclass features, Mintseris *et al.* [19] achieved 77.64% accuracy with optimized SVM [3] classifier and our approach achieved 80.86% accuracy which is over 3% improvement from their results. We also tested those features for [19] with our prediction method, in which it achieved 77.32% accuracy. We also applied feature selection (FS) algorithms for selecting the best subsets of feature. Using floating forward/backward feature selection, we achieved almost the same performance for both of the related works in [19, 30] and 3-5% lower accuracy than without feature selection results with our approach. When we used mRMR selected pairs coupled with our approach, the accuracies are almost the same as those of our original approach without feature selection. The summary of the comparison is shown in Table 5.14.

Table 5.14: Comparison with related works of [19, 30].

Dataset	NOXclass features + SVM	NOXclass features + LDR	Our approach	Our approach + mRMR	Our approach + FS
Mintseris *et al.* [19]	77.64	77.32	80.86	80.27	77.39
Zhu *et al.* [30]	70.07	77.92	82.13	82.18	77.83

Our main idea for applying feature selection was to find some biologically meaningful characteristics for predicting obligate and non-obligate complexes. When we look closely at Figure 5.1, we find that among the selected pairs the highest number of pairs belongs to hydrophobic-hydrophobic pairs. In Table 5.15, the numbers for all selected groups are listed.

Table 5.15: Summary of feature selection.

Hydrophilic pairs		Amphipatic pairs		Hydrophobic pairs	
unique pairs	total pairs	unique pairs	total pairs	unique pairs	total pairs
14	39	15	114	32	189

Analyzing (c) and (d) of Figures 5.2, 5.3 and 5.4, we can see that in all difference heatmaps hydrophobic-hydrophobic pairs have very high values (red). In the combined heatmap, the bottom right corner colors are significantly different. This area is mostly red which belongs to hydrophobic-hydrophobic pairs and other two regions (amphipatic-amphipatic pairs and hydrophilic-hydrophilic pairs) are mostly green. When we used our approach with only amphipathic-amphipatic pairs, hydrophobic-hydrophobic pairs and hydrophilic-hydrophilic pairs, we find hydrophobic-hydrophobic pairs achieve the highest accuracy of 77.60% among these three groups (see Tables 5.7, 5.8 and 5.9). Thus, for all cases hydrophobic-hydrophobic pairs are significantly better for prediction than other pairs.

Chapter 6

Conclusion

6.1 Summary of Contributions

In this thesis, we have presented a new model used to predict obligate and non-obligate complexes. The key contributions of the thesis can be summarized as follows:

- The new model presented in this thesis results in significant improvements in Zhu *et al.* dataset and moderate improvement in Mintseris *et al.* dataset for predicting obligate and non-obligate complexes with desolvation energies as properties and LDR (HDA, FDA, CDA) combined with quadratic Bayesian (QB) and linear Bayesian (LB) as classifiers.

- Including post analysis in the proposed model can help find some biologically meaningful interesting facts. In our thesis, we found that hydrophobic-hydrophobic pairs in obligate and non-obligate complexes have very high discriminating capabilities.

- Different feature selection methods can be coupled with our model and mRMR feature selection works better with our proposed model.

- By picking hydrophobic-hydrophobic amino acid pairs, it is possible to predict obligate and non-obligate protein-protein interactions efficiently.
- Our BPPI dataset can be used for obligate and non-obligate protein-protein interactions prediction.

6.2 Limitations

We could not consider small atoms in our calculations as they do not have any mapping in the atom conversion table (Tables 3.2 and 3.3). While scanning structure files for different complexes, we found some atoms also have the same mapping problem. We do not know the actual smooth function equation and using predetermined atomic contact potentials, we approximated desolvation energy values and not the actual energy values between atom pairs.

6.3 Future Work

Our future work involves the use of this model in different protein-protein interaction classification problems such as intra and inter domains, homo and hetero-oligomers and the use of other properties such as residual vicinity, shape of the structure of the interface, secondary structure, planarity, physicochemical features and others.

Appendix A

Counting numbers of amino acids

Table A.1: Counts for interacting amino acids for different distances in different complexes.

Name	Chain	4Å	4.5Å	5Å	6Å	Name	Chain	4Å	4.5Å	5Å	6Å
1a14	L:N	25	36	59	134	1is8	E:O	5	12	20	34
1a14	H:N	29	47	68	145	1is8	E:M	0	0	0	0
1a2k	A:C	1	1	3	4	1is8	E:N	0	0	0	0
1a2k	B:C	47	85	122	261	1is8	J:K	0	0	0	0
1ahw	A:C	17	28	55	105	1is8	J:L	0	0	0	0
1ahw	B:C	52	87	149	284	1is8	J:O	0	0	0	0
1akj	A:D	40	59	88	161	1is8	J:M	0	0	0	0
1akj	A:E	14	22	28	57	1is8	J:N	0	0	0	0
1akj	B:D	9	17	27	57	1is8	C:K	0	0	0	0
1akj	B:E	0	0	2	3	1is8	C:L	0	0	0	0
1ao7	A:D	24	44	68	121	1is8	C:O	0	0	0	0
1ao7	A:E	8	19	28	45	1is8	C:M	6	11	18	35

1ao7	B:D	0	0	0	0	1is8	C:N	8	14	28	62
1ao7	B:E	0	0	0	0	1is8	I:K	0	0	0	0
1ao7	C:D	14	26	41	69	1is8	I:L	0	0	0	0
1ao7	C:E	14	19	35	83	1is8	I:O	0	0	0	0
1ar1	A:C	0	0	0	0	1is8	I:M	0	0	0	0
1ar1	A:D	0	0	0	0	1is8	I:N	0	0	0	0
1ar1	B:C	20	29	53	111	1is8	D:K	0	0	0	0
1ar1	B:D	24	39	58	112	1is8	D:L	0	0	0	0
1avg	H:I	42	69	113	205	1is8	D:O	8	11	23	63
1avg	L:I	0	0	0	0	1is8	D:M	0	0	0	0
1azz	A:D	41	80	127	233	1is8	D:N	7	14	20	34
1azz	A:D	41	80	127	233	1is8	H:K	0	0	0	0
1b9y	A:C	108	198	296	592	1is8	H:L	0	0	0	0
1b9y	B:C	0	0	0	1	1is8	H:O	0	0	0	0
1be3	C:A	17	28	43	92	1is8	H:M	0	0	0	0
1be3	D:A	4	8	13	29	1is8	H:N	0	0	0	0
1be3	E:A	51	100	151	313	1is8	G:K	0	0	0	0
1be3	G:A	57	113	187	360	1is8	G:L	0	0	0	0
1be3	K:A	7	18	41	72	1is8	G:O	0	0	0	0
1bgx	H:T	74	149	252	542	1is8	G:M	0	0	0	0
1bgx	L:T	65	113	207	449	1is8	G:N	0	0	0	0
1bj1	H:V	2	2	5	10	1is8	F:K	0	0	0	0
1bj1	H:W	75	143	228	391	1is8	F:L	0	0	0	0
1bj1	L:V	0	0	0	0	1is8	F:O	0	0	0	0

1bj1	L:W	2	5	8	21	1is8	F:M	0	0	0	0
1bqh	A:G	36	56	83	168	1is8	F:N	0	0	0	0
1bqh	B:G	0	0	0	0	1jb0	A:E	37	52	84	150
1cs4	A:C	11	26	38	67	1jb0	B:E	18	31	58	111
1cs4	B:C	30	71	101	213	1jb0	B:D	23	49	83	186
1de4	C:A	55	99	146	296	1jb0	A:D	57	95	146	313
1de4	F:A	1	1	1	5	1jb0	A:E	37	52	84	150
1dee	C:G	0	0	0	0	1jb0	B:E	18	31	58	111
1dee	C:H	0	0	0	0	1jb0	B:D	23	49	83	186
1dee	D:G	8283	10537	13053	19050	1jb0	A:D	57	95	146	313
1dee	D:H	0	0	1	20	1jb0	A:C	21	53	85	170
1dkg	A:D	35	76	121	270	1jb0	B:C	42	76	114	247
1dkg	B:D	0	1	1	2	1jmz	A:B	66	130	184	363
1dm0	A:B	7	12	26	55	1jmz	G:B	63	121	192	430
1dm0	A:C	15	34	59	114	1jro	A:D	0	0	0	0
1dm0	A:F	12	26	47	89	1jro	A:B	221	379	604	1225
1dm0	A:D	8	14	23	63	1jsu	A:C	117	211	309	569
1dm0	A:E	15	40	54	103	1jsu	B:C	69	134	182	381
1du3	A:D	41	76	113	238	1jwh	A:C	6	11	19	40
1du3	A:E	0	0	0	0	1jwh	A:D	35	60	97	181
1du3	A:F	44	74	115	233	1k3z	A:D	86	155	221	413
1dx5	A:I	0	0	0	0	1k3z	B:D	42	77	131	282
1dx5	M:I	55	85	131	252	1k4c	A:C	39	57	85	155
1e6j	H:P	37	66	104	202	1k4c	B:C	34	63	98	162

1e6j	L:P	5	11	24	67	1kcg	A:C	16	25	54	119
1ebd	A:C	14	21	31	60	1kcg	B:C	23	41	59	129
1ebd	B:C	22	43	66	120	1kkl	A:H	23	49	69	147
1ebp	A:C	29	43	65	117	1kkl	B:H	0	0	0	0
1ebp	A:D	6	16	25	64	1kkl	C:H	22	37	76	157
1efx	A:D	38	70	114	207	1l0o	A:C	25	43	70	140
1efx	B:D	0	0	0	0	1l0o	B:C	34	65	110	220
1efx	C:D	6	6	8	18	1l7v	A:C	41	72	124	273
1es7	A:B	46	100	157	305	1l7v	B:C	0	0	0	0
1es7	C:B	15	33	55	103	1l9j	C:H	0	0	0	0
1ezv	E:X	44	69	100	232	1l9j	C:L	7	15	22	54
1ezv	E:Y	12	18	28	63	1l9j	C:M	12	20	48	104
1ezx	A:C	25	44	66	142	1li1	A:C	166	284	446	845
1ezx	B:C	0	0	0	0	1li1	B:C	184	333	516	958
1f51	A:E	211	362	577	1326	1lk3	A:H	65	82	115	218
1f51	B:E	362	548	780	1392	1lk3	A:L	33	57	85	149
1f83	A:C	41	65	102	224	1lti	A:D	0	2	2	2
1f83	A:B	99	148	220	449	1lti	A:E	1	1	1	3
1f93	A:E	14	16	22	36	1lti	A:H	3	4	7	17
1f93	A:F	16	28	57	107	1lti	A:F	3	6	15	26
1f93	B:E	19	34	55	113	1lti	A:G	12	16	22	42
1f93	B:F	12	17	24	36	1lti	C:D	14	28	38	74
1fak	H:T	48	70	106	200	1lti	C:E	21	41	67	149
1fak	L:T	57	119	176	344	1lti	C:H	6	12	22	53

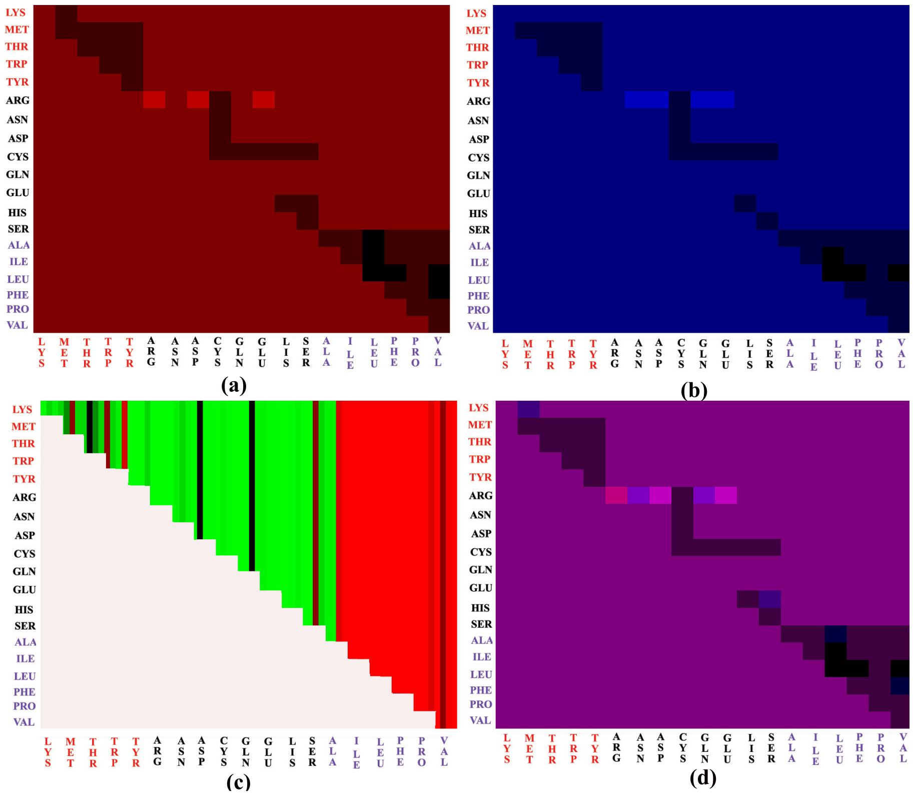
1fbi	H:X	41	77	139	262	1lti	C:F	19	33	50	105
1fbi	L:X	7	14	29	76	1lti	C:G	3	3	8	26
1fg9	A:C	43	80	141	261	1m2o	A:B	91	162	266	526
1fg9	B:C	14	24	36	77	1m2o	C:B	0	0	0	0
1flt	V:X	9	16	36	72	1mjg	A:M	85	170	264	529
1flt	W:X	24	42	76	159	1mjg	B:M	21	39	67	155
1fns	A:L	9	11	22	31	1n2c	A:E	24	47	79	167
1fns	A:H	48	78	118	196	1n2c	A:F	24	48	70	149
1fsk	A:C	48	83	128	227	1n2c	B:E	31	52	72	170
1fsk	A:B	19	31	48	97	1n2c	B:F	34	64	101	206
1g73	A:C	27	41	67	127	1nbw	A:B	38	60	91	197
1g73	B:C	44	75	103	161	1nbw	C:B	17	43	65	146
1gvn	A:B	78	122	166	389	1nsn	H:S	21	35	50	127
1gvn	C:B	0	0	0	0	1nsn	L:S	19	38	57	135
1hez	A:E	23	48	86	191	1o94	A:C	35	45	72	118
1hez	B:E	0	0	0	0	1o94	A:D	0	0	0	0
1hr6	A:B	95	193	287	539	1o94	B:C	0	0	0	0
1hr6	E:B	0	0	0	0	1o94	B:D	0	0	0	0
1hwg	A:B	69	134	225	448	1osp	H:O	45	87	131	215
1hwg	A:C	49	90	137	272	1osp	L:O	47	65	98	187
1hzz	A:C	21	38	64	122	1prc	C:H	13	23	39	65
1hzz	B:C	43	68	100	234	1prc	C:L	97	190	287	576
1i3o	A:E	18	25	36	88	1prc	C:M	167	313	502	985
1i3o	B:E	53	105	167	311	1qfu	A:H	37	71	116	229

1i3o	C:E	0	0	0	0
1i3o	D:E	25	45	63	104
1i4d	A:D	39	73	109	221
1i4d	B:D	7	13	18	44
1i9r	A:H	22	44	84	200
1i9r	A:L	17	28	46	90
1i9r	B:H	0	0	0	1
1i9r	B:L	0	1	1	1
1i9r	C:H	0	0	0	0
1i9r	C:L	0	0	0	0
1ib1	A:E	62	104	162	357
1ib1	B:E	1	1	5	8
1icf	A:I	74	133	220	414
1icf	B:I	12	22	31	54
1iis	A:C	19	45	74	160
1iis	B:C	24	42	70	129
1ijk	A:B	13	28	43	81
1ijk	A:C	17	32	55	115
1im3	A:D	33	62	105	222
1im3	B:D	0	0	0	0
1iod	A:G	8	20	31	56
1iod	B:G	10	20	29	70
1iqd	A:C	0	0	0	0
1iqd	B:C	0	0	0	0
1qfu	A:L	4	6	8	44
1qfu	B:H	0	0	0	0
1qfu	B:L	0	0	0	0
1qfw	A:I	0	0	0	0
1qfw	A:M	4	8	15	33
1qfw	B:I	34	55	74	127
1qfw	B:M	24	43	64	126
1qkz	A:L	5	5	17	37
1qkz	A:H	37	85	116	212
1qo0	A:D	34	54	80	155
1qo0	A:E	29	47	74	136
1rlb	A:E	6	10	21	44
1rlb	B:E	23	41	61	113
1rlb	C:E	9	18	23	57
1rlb	D:E	0	0	0	0
1sgf	A:B	10	21	38	94
1sgf	A:Y	36	60	108	193
1toc	A:R	0	0	0	0
1toc	B:R	164	271	420	792
1wej	F:H	26	36	53	106
1wej	F:L	30	48	68	118
1www	V:X	21	32	47	88
1www	W:X	44	74	120	256
1ytf	B:D	51	101	157	338

1is8	A:K	7	14	21	33
1is8	A:L	8	11	26	61
1is8	A:O	0	0	0	0
1is8	A:M	0	0	0	0
1is8	A:N	0	0	0	0
1is8	B:K	0	0	0	0
1is8	B:L	6	12	19	34
1is8	B:O	0	0	0	0
1is8	B:M	9	13	31	68
1is8	B:N	0	0	0	0
1is8	E:K	9	13	27	67
1is8	E:L	0	0	0	0
4rub	D:T	21	49	70	147
1ytf	C:D	122	241	386	708
2hmi	A:C	0	0	0	0
2hmi	A:D	0	0	0	0
2hmi	B:C	9	17	23	49
2hmi	B:D	28	52	74	161
2jel	H:P	34	67	92	183
2jel	L:P	14	28	50	89
2mta	A:H	10	14	31	85
2mta	A:L	13	28	48	144
4htc	H:I	86	163	270	522
4htc	L:I	0	0	0	0
4rub	A:T	44	79	131	259

Appendix B

Heatmaps



(a)

(b)

(c)

(d)

Figure B.1: BPPI-W heatmaps- (a) obligate, (b) non-obligate, (c) difference of (a) and (b), (d) combined of (a) and (b), - with color resolution 4×4.

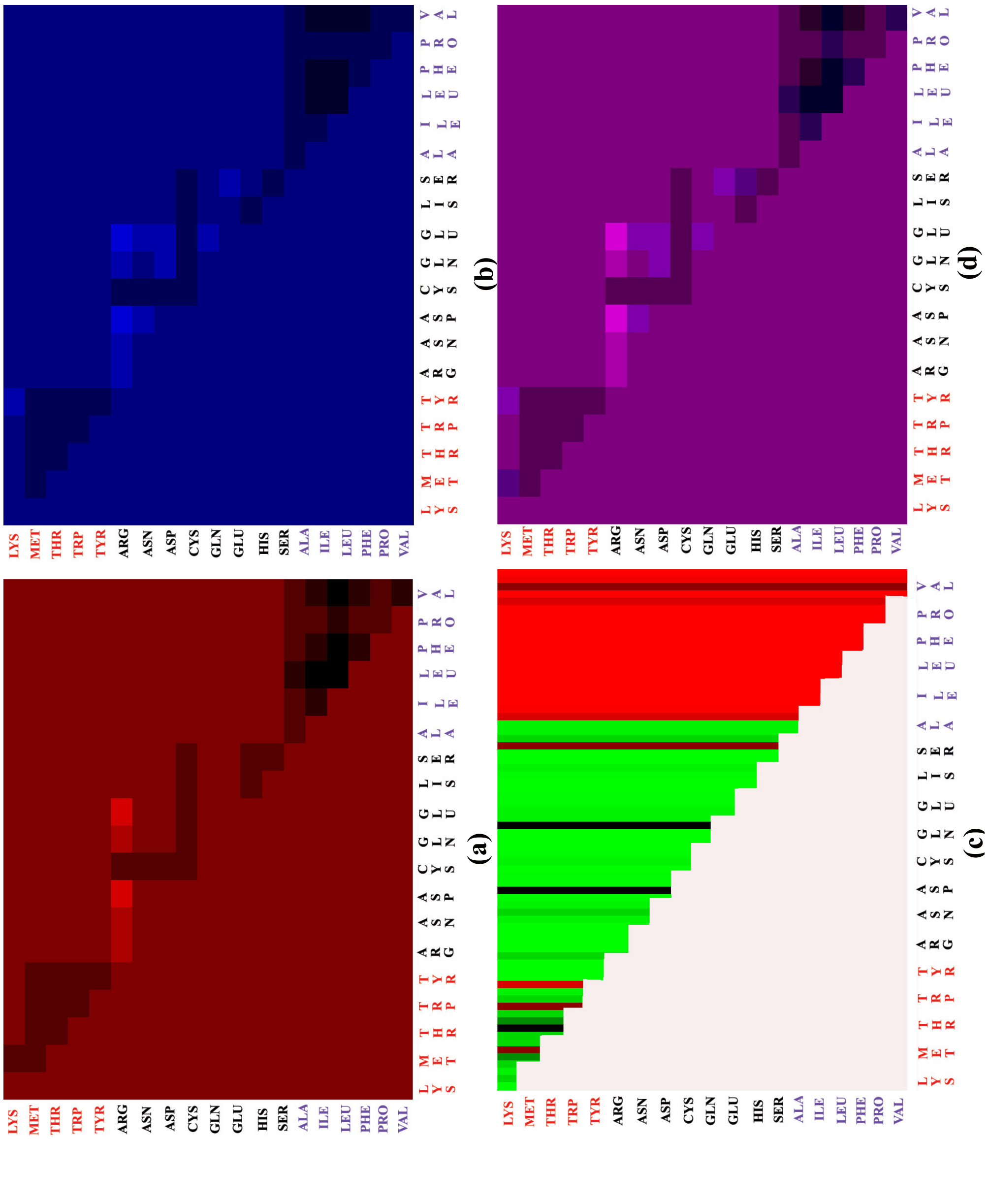


Figure B.2: BPPI-W heatmaps- (a) obligate, (b) non-obligate, (c) difference of (a) and (b), (d) combined of (a) and (b), - with color resolution 6×6.

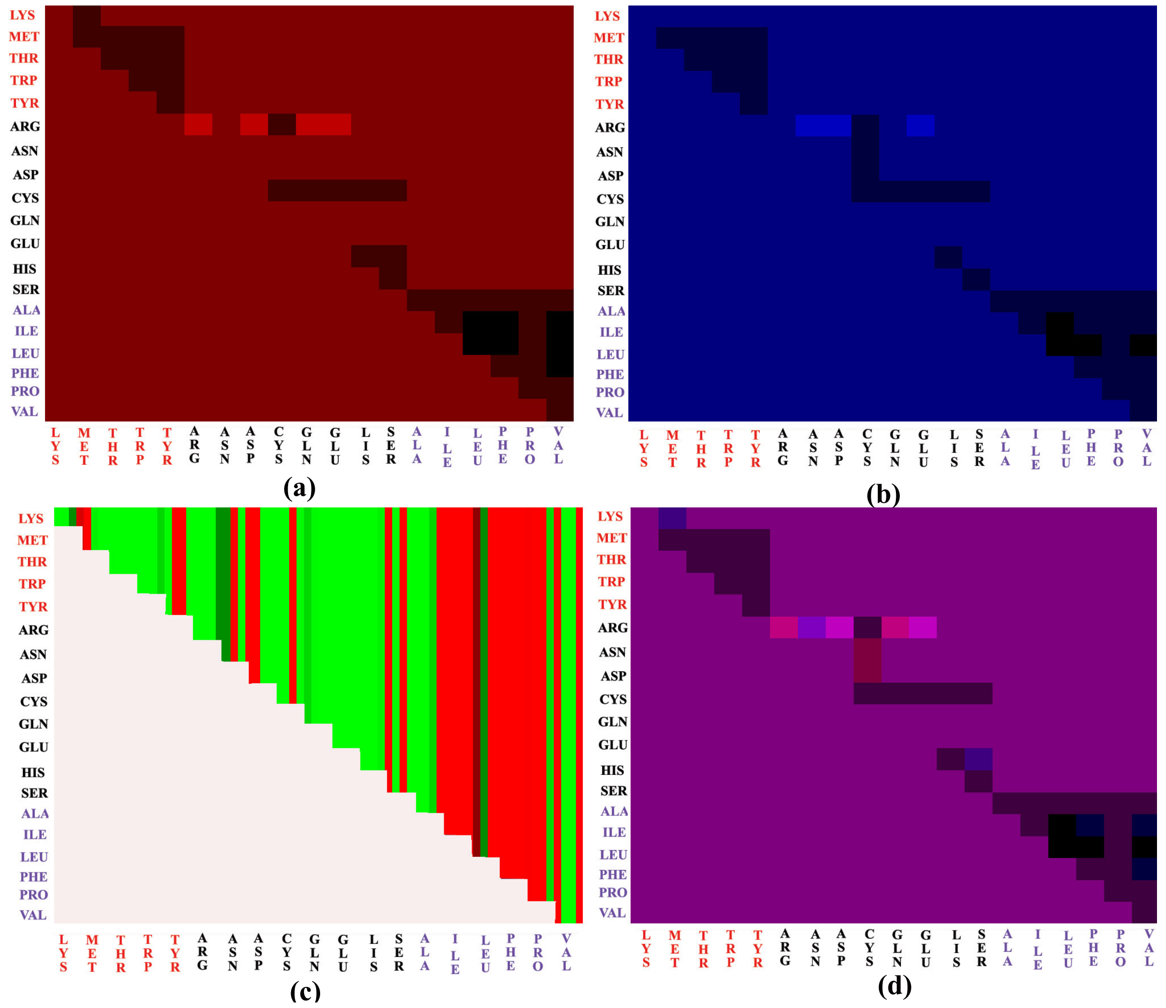
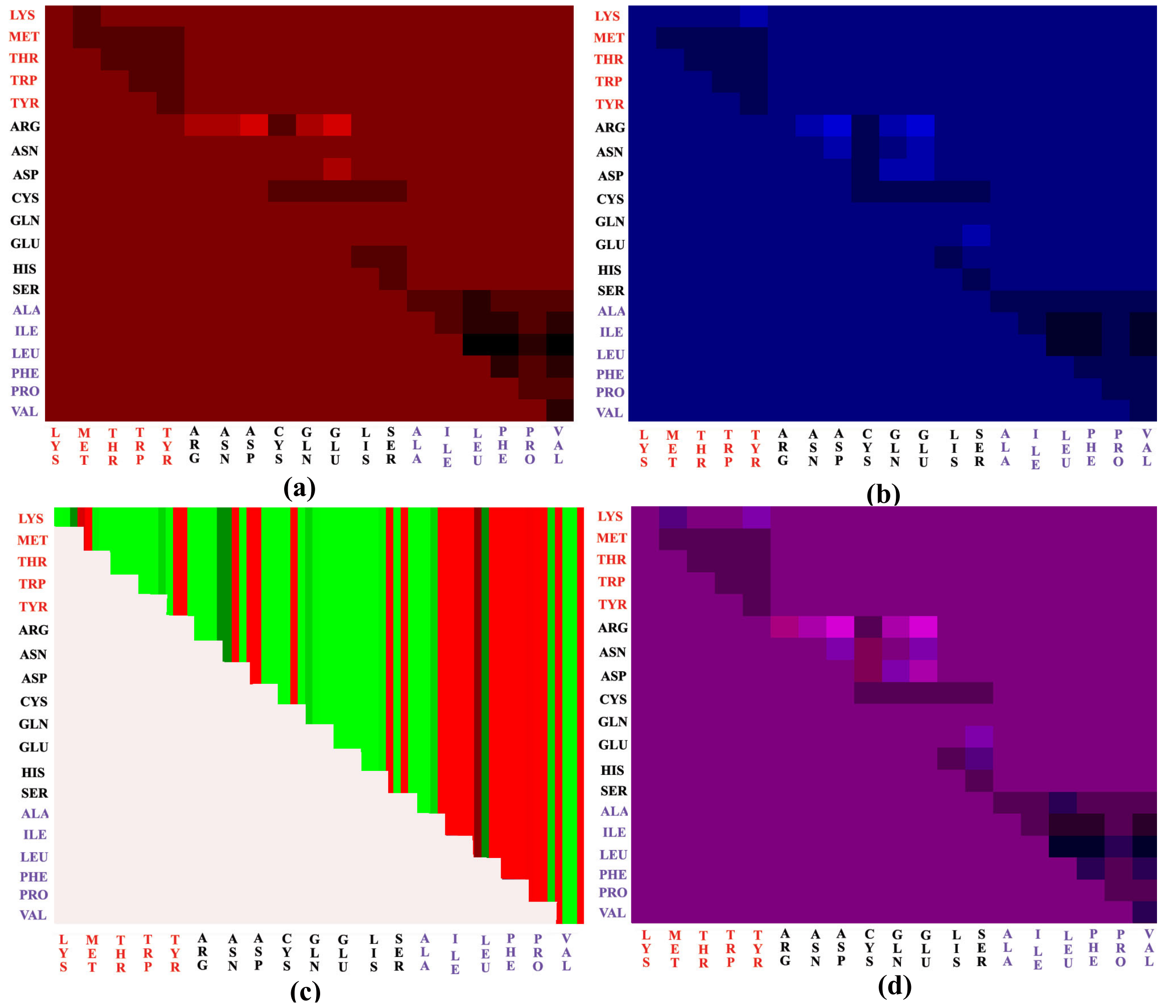


Figure B.3: Mintseris *et al.* [19] heatmaps- (a) obligate, (b) non-obligate, (c) difference of (a) and (b), (d) combined of (a) and (b), - with color resolution 4×4.



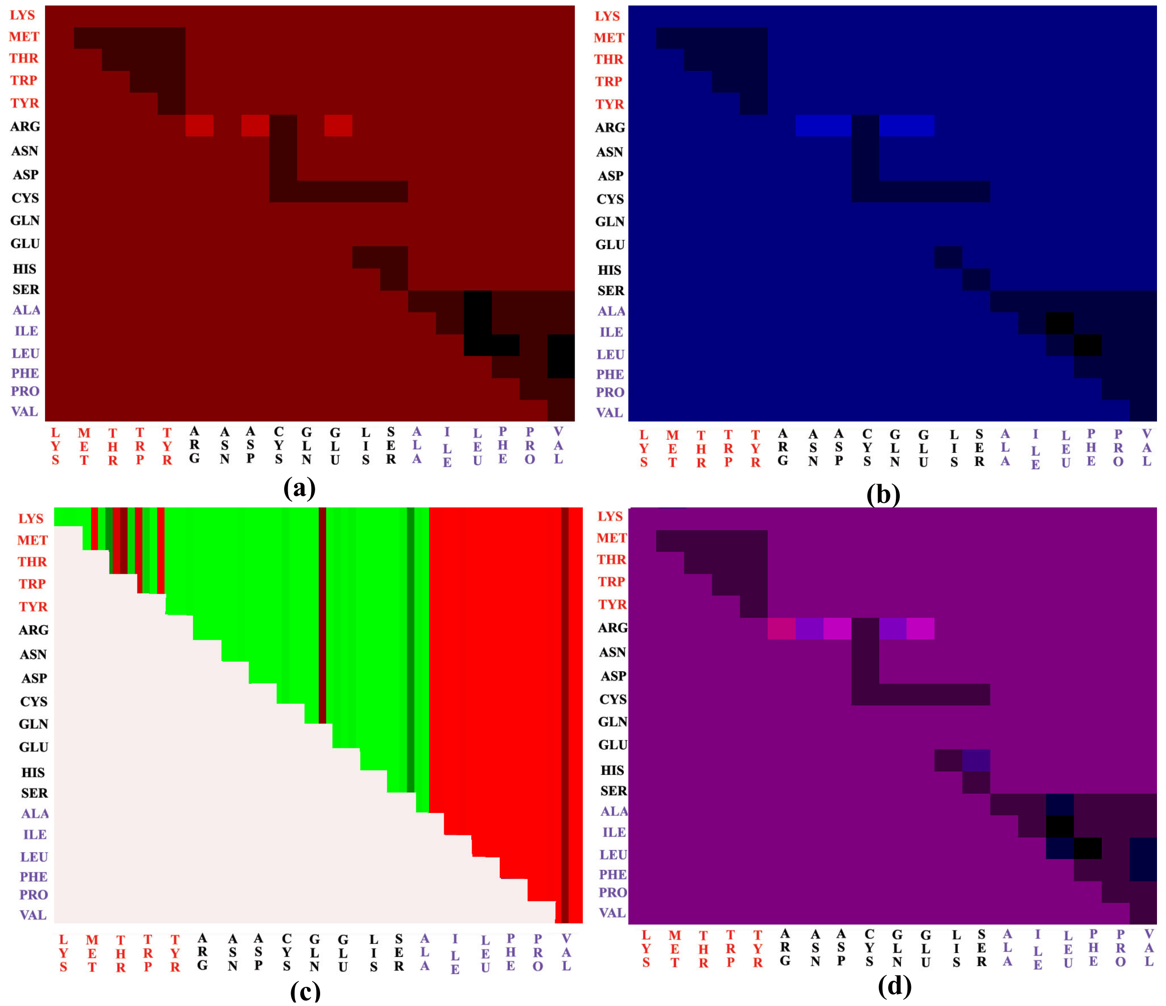
(a)

(b)

(c)

(d)

Figure B.4: Mintseris *et al.* [19] heatmaps- (a) obligate, (b) non-obligate, (c) difference of (a) and (b), (d) combined of (a) and (b), - with color resolution 6×6.



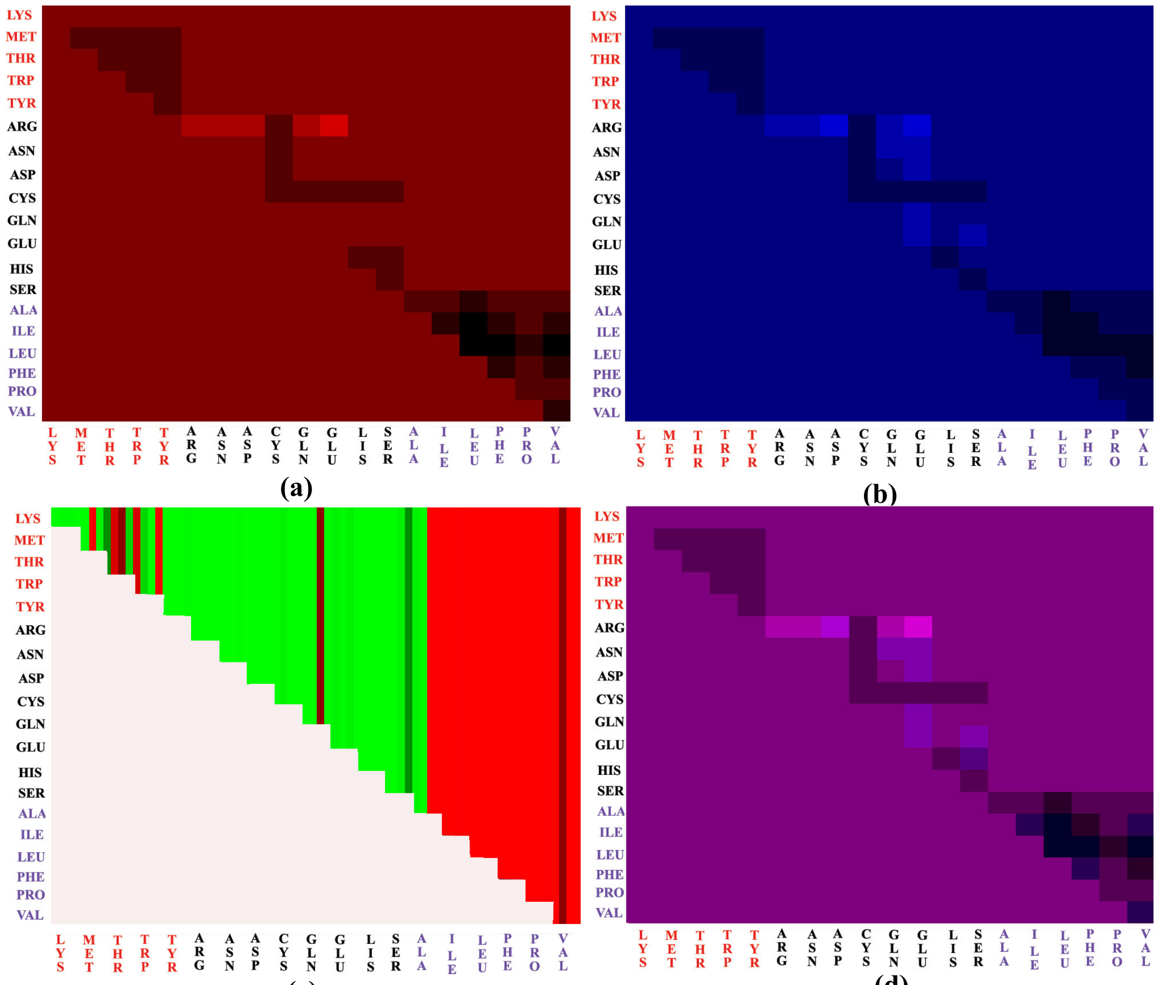
(a)

(b)

(c)

(d)

Figure B.5: Zhu *et al.* [30] heatmaps- (a) obligate, (b) non-obligate, (c) difference of (a) and (b), (d) combined of (a) and (b), - with color resolution 4×4.



(a)

(b)

(c)

(d)

Figure B.6: Zhu *et al.* [30] heatmaps- (a) obligate, (b) non-obligate, (c) difference of (a) and (b), (d) combined of (a) and (b), - with color resolution 6×6.

Bibliography

[1] Affinity Chromatography Principles and Methods. Amersham Pharmacia Biotech, ac edition.

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

[3] C. L. C. Chang. Libsvm: a library for support vector machines, 2011. Available at : http://www.csie.ntu.edu.tw/ cjlin/papers/libsvm.pdf.

[4] C. Camacho and C. Zhang. FastContact: rapid estimate of contact and binding free energies. Bioinformatics, 21(10):2534–2536, 2005.

[5] R. Duda, P. Hart, and D. Stork. Pattern Classification. John Wiley and Sons, Inc., New York, NY, 2nd edition, 2000.

[6] R. Fisher. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7:179–188, 1936.

[7] G. Geva and R. Sharan. Identifcation of protein complexes from co-immunoprecipitation data. BMC Systems Biology, 27(1):111–117, 2010.

[8] F. L. Hanchuan Peng 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, 27(8):1226–1238, 2005.

[9] S. Hubbard and J. Thornton. Naccess, 1993. URL `http://www.bioinf.manchester.ac.uk/naccess/` .

[10] M. N. Isabelle Guyon, Steve Gunn and L. Zadeh. Feature Extraction, Foundations and Applications. Springer, first edition, 2006.

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

[12] S. P. B. A. G. M. L. James D. Watson, Tania A. Baker and R. Losick. Molecular Biology of the Gene. Benjamin Cummings, sixth edition, 2008.

[13] J. Janin. Kinetics and thermodynamics of protein-protein interactions from a structural perpective. Oxford University Press, 2000. Protein-Protein Recognition, pp. 344.

[14] I. Jolliffe. Principal Component Analysis. Springer, second edition, 2002.

[15] S. Jones and J. M. Thornton. Protein-Protein Recognition, chapter Analysis and classification of protein-protein interactions from a structural perspective. Oxford University Press, 2000.

[16] I. Kurareva and R. Abagyan. Predicting molecular interactions in structural proteomics. In R. Nussinov and G. Shreiber, editors, Computational Protein-Protein Interactions, chapter 10, pages 185–209. CRC Press, 2009.

[17] M. Loog and P. Duin. Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(6):732–739, 2004.

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

[19] J. Mintseris and Z. Weng. Structure, Function, and Evolution of Transient and Obligate Protein-protein Interactions. Proceedings of the National Academy of Sciences, USA, 102(31):10930–10935, 2005.

[20] I. Nooren and J. Thornton. Diversity of protein-protein interactions. EMBO Journal, 22(14):3846–3892, 2003.

[21] Y. Ofran. Prediction of protein interaction sites. In R. Nussinov and G. Shreiber, editors, Computational Protein-Protein Interactions, chapter 9, pages 167–184. CRC Press, 2009.

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

[23] L. A. H. R. Benjamin Free and D. R. Sibley. Identifying Novel Protein-Protein Interactions Using Co-Immunoprecipitation and Mass Spectroscopy. Current Protocols in Neuroscience, page DOI: 10.1002/0471142301.ns0528s4, 2009.

[24] L. Rueda and M. Herrera. Linear Dimensionality Reduction by Maximizing the Chernoff Distance in the Transformed Space. Pattern Recognition, 41(10):3138–3152, 2008.

[25] R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and

A. Sali. A structural perspective on protein-protein interactions. Curr Opin Struct Biol, 14(3):313–324, 2004.

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

[27] E. Tropp and D. Freifelder. Molecular Biology: Genes to Proteins. Jones and Bartlett Learning, forth edition, 2008.

[28] L. Wong. The Practical Bioinformatician. World Scientific Publishing Co. Pte. Ltd., first edition, 2004.

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

[30] H. Zhu, F. Domingues, I. Sommer, and T. Lengauer. Noxclass: Prediction of protein-protein interaction types. BMC Bioinformatics, 7(27), 2006. doi:10.1186/1471-2105-7-27.

Vita Auctoris

Muhammad Mominul Aziz was born in 1985 in Dhaka, Bangladesh. He received his Bachelors degree in Computer Science and Engineering from Khulna University of Engineering and Technology, Dhaka, Bangladesh in 2007. His research interests include pattern recognition, protein-protein interaction, image processing, machine learning and bioinformatics.