Critical Assessment of Protein-Protein Interaction Databases and Features towards Prediction of Interactions

by

Mehmet Cengiz ULUBAŞ

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science
in
Electrical and Computer Engineering

Koç University

June, 2009

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Mehmet Cengiz ULUBAŞ

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

____

Attila GÜRSOY, Ph. D. (Advisor)

____

Özlem KESKİN, Ph. D.

____

Elif ÖZKIRIMLI, Ph. D.

Date: ____________________

ABSTRACT

Protein-protein interactions (PPI) are of crucial importance at all levels of biological processes. The experimentally identified PPI are deposited in several databases. These databases contain diverse information about PPI; but their coverage is low when we consider full processes in cells. Thus, reliable, accurate computational methods are needed to improve the coverage. Many research groups have developed PPI prediction algorithms with varying accuracies based on different data and methods. However, to develop a new PPI prediction method with high accuracy is challenging.

This study aims to assess existing sequence based PPI prediction methods and to propose a new algorithm with improved accuracies. The predictions are made via Support Vector Machines (SVM), which is a machine learning algorithm. SVM creates models based on training sets and predicts interactions via those models. In this study, positive training sets contain experimental PPI and negative training sets contain computational non-interacting proteins. In order to represent interaction data in SVM, n-gram frequencies of proteins are calculated according to their amino acid sequences. It is shown that SVM performance is strongly affected by interactions in training datasets, amino acid categorization techniques, n-gram frequencies, and γ values used. SVM models are created for eight datasets and the critical assessment of those datasets is made via their SVM scores. Based on those scores, combined training datasets are created that make accurate prediction of interactions in every dataset. Then, the best feature set that leads to the highest SVM scores is found. Finally, the best SVM models are utilized to eliminate false positives in putative protein interactions predicted by PRISM (Protein Interactions by Structural Matching) algorithm.

ÖZET

Protein-protein etkileşimleri (PPE) biyolojik süreçlerin her seviyesinde çok önemlidir. Deneysel olarak kanıtlanmış PPE farklı veritabanlarına koyulmaktadır. Bu veritabanları PPE hakkında çeşitli bilgiler içermektedir, fakat hücrelerdeki tüm süreçler göz önüne alındığında, kapsamları düşüktür. Bu yüzden, PPE kapsamını genişletmek için güvenilir, daha doğru hesaplamalı metotlar gerekmektedir. Birçok araştırma grubu farklı bilgi ve metotlara dayanan çeşitli doğrulukta PPE tahmin algoritmaları geliştirmiştir. Ancak, yüksek doğrulukta bir PPE tahmin etme metodu geliştirmek ilgi çekicidir.

Bu çalışma, var olan dizilim tabanlı PPE tahmin etme metotlarını değerlendirmeyi ve doğruluk oranları geliştirilmiş yeni bir metot önermeyi hedeflemektedir. Tahminler bir makine öğrenimi algoritması olan Destek Vektör Makineleri (DVM) ile yapılmaktadır. DVM, öğrenim etkileşim veri kümelerine göre kalıplar oluşturur ve etkileşimleri bu kalıplar ile tahmin eder. Bu çalışmada, pozitif öğrenim veri kümeleri deneysel PPE'leri, negatif öğrenim veri kümeleri hesaplanmış etkileşmeyen proteinleri içermektedir. Etkileşim bilgisini DVM'de betimlemek için, proteinlerin amino asit dizilim sıralarına göre n-gram frekansları hesaplanmıştır. DVM performansının, öğrenim veri kümelerindeki etkileşimlerden, farklı amino asit sınıflandırması tekniklerinden, n-gram frekanslarından ve γ değerlerinden fazlaca etkilendiği gösterilmiştir. Sekiz öğrenim veri kümesi için DVM kalıpları oluşturulmuştur ve DVM skorları ile detaylı karşılaştırmaları yapılmıştır. Bu skorlara göre, her veri kümesindeki etkileşimleri iyi tahmin eden birleştirilmiş öğrenim veri kümeleri oluşturulur. Daha sonra, en yüksek DVM skorunu elde etmeyi sağlayan en belirleyici nitelikler kümesi bulunur. Son olarak, en iyi DVM kalıpları, YUPE (Yapısal Uyumlu Protein Etkileşimleri) algoritması tarafından tahmin edilen PPE içindeki yanlış pozitiflerin elenmesi için kullanılır.

ACKNOWLEDGEMENTS

I would like to thank **Assoc. Prof. Attila Gürsoy** (my advisor) and **Assoc. Prof. Özlem Keskin** for the chance of being a member of their research group, supplying us a productive and comfortable research environment and their support during my work. Also, I would like to thank **Asst. Prof. Elif Özkırımlı** for her critical reading and useful comments.

I would like to thank also **the Scientific and Technological Research Council of Turkey (TUBITAK)** for their financial support during my M.S. study. I would like to thank my formal and informal home mates (**Osman N. Yoğurtçu, M. Fatih Toy, Sefer Baday, Özkan Eğri, Ramazan Sancak**); my former and recent office mates (**Nurcan Tunçbağ, Emre Güney, Ekin Tüzün, Gözde Kar**); and my friends **Bahar Öndül, Özge Engin, Besray Ünal, Gülay Ergül, Aslıhan Arslan, Yasemin Demir, Zeynep Kaymak, Abdullah M. Turan, Hüseyin Seren, Selçuk Dağ, Koray Mutlu, and Serhan Ö. & Elif Işıkman** and also my all other friends for their support, encouragement and good times during these 2 years.

And finally, I would like to thank **my family** for their continuous support and patience during every step of my education.

To my nephew Abdullah who moved to Heaven when he was only 4…

TABLE OF CONTENTS

Chapter 3

Chapter 4

Chapter 5

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 4

LIST OF FIGURES

Chapter 3

Chapter 4

NOMENCLATURE

BIND	Biomolecular Interaction Network Database
CA	Carbon Alpha
DIP	Database of Interacting Proteins
DT	Decision Trees
HPRD	Human Protein Reference Database
MIPS	Munich Information Center for Protein Sequences
MULTIPROT	Multiple Protein Structural Alignment Algorithm
PDB	Protein Data Bank
PPI	Protein-protein interactions
PRISM	Protein Interactions by Structural Matching
RBF	Radial Basis Function kernels
SVM	Support Vector Machines
WEKA	Waikato Environment for Knowledge Analysis
YPD	Yeast Proteome Database

Chapter 1

INTRODUCTION

Almost all biological processes are controlled by interactions between proteins [1]. To elucidate the protein interactions, several experiments have been performed and their results are deposited into protein interaction databases. Although these databases contain diverse information about protein – protein interactions, they are still not complete [2]. Thus, besides the experimental methods, there is a need for computational methods to predict new protein interactions. Several groups have developed protein interaction prediction algorithms that perform well but the most important drawback of these methods is their high false positive rates, besides the true positive predictions. Thus, a novel PPI prediction method is required for minimizing the number of false positives [3].

In recent years, learning based methods get popular for PPI prediction. Especially, Support Vector Machines (SVM), Bayesian Networks, and Decision Trees are commonly used [4]. However, SVM is found to be the best performing machine learning algorithm in several studies [3, 4]. In order to use SVM for PPI prediction algorithm, a reliable, diverse and non-redundant training set should be supplied, which contains positive and negative protein interactions. So, construction of gold standard positive and negative sets is crucial to develop high performance prediction algorithms. For positive interaction set, there are several resources which contain verified PPI. However, to generate these sets, different experimental techniques are used on different organisms. Thus, it is hard to generate a gold standard dataset for positive interactions which is reliable and sufficiently diverse that covers every organism. For negative interaction set, this problem is more complicated. In literature, an experimental non-interacting protein set is not available [5]. Therefore, negative interaction sets are created via computational methods. But there are many different computational methods and each of them has some drawbacks [6]. Therefore, a careful analysis of the datasets in literature should be done in order to select the best training set for SVM and make reliable predictions.

In this thesis work, we aimed to present a new, accurate, SVM-based method to predict protein interactions. For this purpose, the questions: which features are important for PPI prediction to obtain best prediction performance; what are the optimum parameters and functions in SVM; how to construct a reliable, diverse and non-redundant training set with high coverage to make accurate predictions are challenging. To address these questions, we present the comparative assessment of eight datasets in literature via their SVM performances. Our prediction method is solely sequence-based. The sequence information is stored in vectors of the same size. As a result, the performance of the algorithm is directly related with the information stored in these vectors. The main focus of the work is to discover the best features for representing residue sequences of PPI in vectors and to find the optimum parameters and functions in SVM. The effects of these features are found by the comparison of SVM performances. The goal is to find the features leading to the best SVM performance and to create combined datasets that predict interactions in every dataset accurately. Two combined datasets are created from the eight datasets in literature and SVM models are generated for all these datasets. The assessment of those models is made via the prediction scores across datasets. Following that, the best features are used for structure and sequence information based PPI prediction on PRISM (Protein Interactions by Structural Matching) server [7].

The outline of this thesis study is as follows:

In Chapter 2, related studies in the literature are demonstrated. Initially, the computational techniques to predict protein – protein interactions are introduced. Following that, n-gram representation, categorization of amino acids, and the machine learning tool WEKA are reviewed which are utilized also in Chapter 4 to develop our method. Then, protein interaction databases are introduced. Next, computational negative PPI creation methods are illustrated. Finally, statistical measures used for SVM performance evaluation are explained.

Chapter 3 contains the materials of this study and the methods to find the best features for SVM. The description of eight datasets selected from literature and two combined sets created from them are introduced. Following that, the features affecting the SVM performance are analyzed in detail. For each feature, the SVM results are given and the comparative assessment of datasets is done via those results. Finally, the SVM model generated from each training set is used to predict the PPI in other datasets.

Chapter 4 includes the improved version of the PRISM algorithm which predicts PPI by spatial similarity. Here, the prediction steps of new PRISM are explained. First, the prediction algorithm by structural matching is presented. Then, the procedure to check the collision in the predicted protein complexes is explained. In the next step, the biological relevancy and the interaction type of the predicted protein complexes are found by NOXclass [8]. Finally, the best SVM model generation technique found in Chapter 3 is integrated into the last step of PRISM algorithm to label the predicted interactions as positive or negative.

This thesis ends with a chapter which includes the summary of the presented work, discussion of the results, future directions and concluding remarks.

The Appendix chapter includes the in depth analysis of datasets used during this study. For each dataset, the PFAM families of all interacting proteins are found [9]. A detailed analysis of all interactions based on PFAM families is made in order understand the effect of families in SVM classification. Following that, the n-gram frequencies are analyzed for each dataset. The most and least frequent n-gram frequencies in positive and negative sets are compared to each other and also with other datasets in order to find the most important ones. Finally, the classification of interactions using energy parameters and different types of amino acid categorizations are presented.

Chapter 2

BACKGROUND

This chapter summarizes the detailed literature review about the studies and methods used in this thesis. In Section 2.1, the works related to PPI prediction are presented. In this section, the general aspects of these works, the prediction methods and their advantages and drawbacks are illustrated. In the next section, Section 2.2, amino acids representation with n-gram frequencies is explained in detail. Several n-gram frequencies might be generated based on the categorization of amino acids. Section 2.3 lists the categorization techniques based on specific amino acid properties.

In Section 2.4, the machine learning tool, WEKA, is introduced which is utilized to perform SVM on n-gram frequency vectors. Following that, the features affecting the performance of SVM are explained. This study uses the interactions extracted from several databases, which contain experimental PPI. Section 2.5 gives detailed information about those databases. Afterwards, in Section 2.6 some computational methods are introduced related to negative interaction sets construction. This part consists of purely computational techniques because an experimental noninteracting set of interactions does not exist. Finally in Section 2.7, the statistical measures used for evaluating the performance of SVM are explained. In this section, the formulas and classification terms used in assessments are given.

2.1 Prediction of Protein-Protein Interactions

PPI are important to elucidate the cellular machinery [10]. The data extracted from these interactions are used for various studies such as drug discovery and understanding cellular operations [3, 11]. There are several experimental techniques developed for discovering PPI such as yeast two-hybrid based methods and mass spectrometry. But the number of interactions found via those methods is limited and covers only a small part of all interactions [2]. In addition, these are time and labor demanding techniques [12]. Thus, computational PPI prediction methods are needed. Several PPI prediction methods are established, which are used for determining the PPI networks of the organisms [13]. The common drawback of these computational methods is the prediction of many false positive interactions besides the true positive ones [11]. This section summarizes the detailed literature review about several PPI prediction methods, most of which use machine learning tools. The detail of these methods, their drawbacks, and performance are presented. In general, there are two methods used for PPI prediction: (i) structure based methods and (ii) sequence based methods. During this study, we performed purely sequence based analysis for prediction of PPI while making assessment of datasets. Several distinct approaches, used for protein sequence based predictions, are given in order to get a better understanding of the methods used in this thesis study and compare their performances.

The most remarkable work about sequence based PPI prediction is preformed by Shen et al. (2007) [3]. In this study, a method is established for PPI prediction that uses only sequence information of proteins. First, the amino acids are grouped into seven classes based on the dipoles and volumes of side chains. Following that, “conjoint triplet method” is used in order to create input vectors for SVM. The positive interactions used in training set are human PPI taken from HPRD. The negative interaction set is created by a new method, which is introduced in following sections. 16.000 positive and 16.000 negative interactions are classified via SVM with 5-fold cross validation technique. It is stated that the obtained positive precision is about 84%. In addition, the model created by SVM is used for PPI network prediction and satisfactory results are achieved.

In the study of Bock et al. (2001), PPI are predicted using SVM. Their method is based solely on protein sequence information and associated physicochemical residue properties [14]. The positive interaction set is taken from DIP database and the diversity of interactions is shown by mapping to PFAM. For negative interaction set, a computational approach is followed which uses randomized protein sequences similar to DIP proteins. Positive and negative interactions are represented by adding charge, hydrophobicity, and surface tension of each residue combined with sequential charge and surface tension. As a result of the analysis, an average accuracy about 80% is achieved.

Yan et al. (2003) shows that, SVM might be also used to predict binding site residues [15]. In this study, the surface residues on interaction sites are predicted with 67% sensitivity using the identity of target residues and their 10 neighbors in sequence. It is claimed that the prediction results would be improved by creating model for different protein-protein complexes. It is also stated that the models created by this method would be a way to predict PPI in future work. There is a similar study to this one that also tries to predict interaction sites from residue sequence for identification of pharmacological targets and drug discovery [16]. In that study, the success rate of classification is about 59-80% for the two datasets used.

In another work, Yan et al (2004) predict binding site residues using two classifiers based on protein sequences [17]. Besides the SVM classification in their previous study, Bayesian classifier is used at the second step. The probability of being an interface residue is calculated for the neighbors of interface residues. Two step classification is used based on the information that interface residues are likely to form clusters, which is claimed in another study [18]. The generated model is tested on CAPRI (Critical Assessment of PRedicted Interactions) targets [19]. In addition, the predictions are verified by comparing

3D structures of proteins. This two-step classification is compared to previous work and it is pointed out that the accuracy of prediction increases from 66% to 72%.

MULTIPROSPECTOR uses the threading algorithm PROSPECTOR (Protein Structure Predictor Employing Combined Threading to Optimize Results), interfacial energies of PPI, some empirical indicators, and sequence information to make PPI prediction [20]. The prediction success rate is over 90% for the test cases composed of homodimers, heterodimers, and monomers. On the other hand, when a test is made on 2.457 MIPS interactions between 1.872 proteins, only the 15.7% is predicted. It is claimed that the low prediction rate is due to interaction creation method of MIPS and to the limited dimer database used. Marcotte et al. (1999) aims to predict PPI using the sequence similarity of protein interactions in different organisms [21]. In this study, the main focus is to find proteins whose sequences are similar to the sequence of a protein in different organisms. The possible interactions found by domain fusion analysis are tested. As a result of the study, 6.809 Escherichia Coli and 45.502 yeast interactions are found. In Escherichia Coli, the predictions result in 47% true positive and 65% false positive rate.

A different approach for prediction of protein-protein binding sites is proposed in the study of Bradford at al. [22]. The starting point is the discovery of common properties of binding sites that are different from rest of the protein. Six surface properties are used while classifying non-homolog interactions extracted from PDB. One of those properties is based on the sequence of proteins. As a result of SVM classification with surface patch analysis, interacting and non-interacting surface patches of transient and obligate interfaces are predicted with 76% accuracy.

A comparative study of learning methods for PPI prediction is in introduced in another report [23]. A method called “Mixture-of-Feature-Effects” (MFE) is used in order to predict interactions. In this method, the features of yeast are grouped into four and the features of human are grouped into three. While representing yeast protein interactions, 162 features taken from 17 sources are used. For human protein interactions, 17 features found from 8 sources are used. Sequence similarity is a feature of both human and yeast protein interactions. In order to have a gold standard dataset, 2.900 yeast interactions are extracted from DIP and 15.000 interactions are extracted from HPRD. For training the set, a computational negative interaction list is created. While creating the negative set, randomly selected two proteins are selected as a negative interaction if that interaction does not exist in the verified interaction list. In order to test the performance of MFE, the results are compared with 4 commonly used learning methods. Those methods are Logistic Regression, Naïve Bayes, Support Vector Machines, and Random Forest. In most of the evaluation criteria, MFE performs better than the other 4 learning methods. Thus, in further study, this method is desired to be applied on important human proteins whose interactions are not identified experimentally yet.

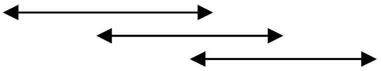
Espadalar et al. (2005) present a study which uses the similarity of both structure and sequence patches of proteins to predict PPI [24]. The basic idea of this prediction method is that close homologs interact using similar patterns. Using sequence search method, 12.225 sequences are found, 8.552 of which is also defined in DIP. In addition 132.627 interactions are predicted by SSIP method. When structural similarity is used, 2.636 human proteins are found and 74.598 interactions are predicted by SSIP. HPRD interactions are used in order to validate the predictions but only <5% of the interactions are verified. On the other hand, the analysis shows that this method increases the probability of predicting a human protein interaction from 0.09% to 0.17%.

2.2 N-gram Representation of Residue Sequences

The protein-protein interactions should be represented with some numeric features for classification via SVM. In this study, protein-protein interactions are characterized by the sequence information of the proteins. There are several methods to describe the properties of the protein sequences. One of them is “conjoint triplet” method that records three consecutive amino acids in the protein as a unit [3].

Conjoint triplet method can be explained best with an example. In the example given below, amino acids are assumed to be clustered into 7 groups based on their common properties. The first line lists the amino acids and the second line lists the groups of the amino acids on the first line. The last line shows the triplets resulting from the sample sequence.

Amino Acids:	Ala	Leu	Tyr	His	Met	Asp	Cys	Glu
Groups:	1	5	3	4	3	2	7	6



Triplets: 1-5-3, 5-3-4, 3-4-3, 4-3-2, 3-2-7, 2-7-6

In this method, every triplet in the sequence is found. Triplets are found by sliding the window containing three residues one residue at each step. When a triplet is recorded, the second amino acid of that triplet is taken as the first amino acid of the next triplet. Thus, each time the first amino acid of the triplets is shifted to the right by one. That means, the triplets are composed of the amino acids with indexes 1-3 in the first triplet; 2-4 in the second triplet; 3-5 in the third triplet; k-(k+2) in the kth triplet where k is the index of the amino acid in protein sequence. In this way, m-2 triplets are created where m is the number of amino acids in the protein sequence.

Conjoint amino acids technique might generate different number of possible combinations based on the number of amino acids groups. For instance, if triplets are created according to 7-group clustering, the number of possible combinations is 7*7*7 = 343, where there are 8*8*8 = 512 combinations for 8-group clustering. That is because each group might be represented on each slot of a triplet. In brief, the number of possible combinations can be generalized in this way: if n is the number of amino acid groups, the number of possible combinations is n^3. Notice that n^3 is only an upper bound on the number of possible combinations and that many triplets might not exist in each sequence.

When classifying protein-protein interactions via SVM, each protein is represented by a frequency vector that is created based on the conjoint amino acids. Triplets might generate at most n^3 possible combinations and all these combinations are elements of a vector of size n^3 where each element is initially set to 0. For each triplet in residue sequence, the element of the vector corresponding to that triplet is incremented by one. When all the triplets in the whole sequence are recorded, each element in the vector has a value between 0 and m-2. The sum of all triplet frequencies in the vector will be m-2, since there are m-2 triplets in the protein sequences. But in this way, the frequencies will be high for proteins with long sequence and low for proteins with short sequence. In order to have vectors of same scale while classifying interactions, the frequencies of amino acids are normalized. Assume the case when there are n^3 triplets and the frequencies of those triplets are represented by $\boldsymbol{f_i}$, where $\boldsymbol{i}$ stands for the index of the triplet. The normalized values of triplets, represented by $\boldsymbol{d_i}$, is found by decrementing the minimum $\boldsymbol{f}$ value from current $\boldsymbol{f}$ value and then dividing by the maximum $\boldsymbol{f}$ value in the vector. That normalizes the frequencies in range 0 and 1 and makes two proteins of different sequence lengths comparable [3]. The formula of normalization is:

$$\mathbf{d_i} = (\mathbf{f_i} - \min \{\mathbf{f_1}, \mathbf{f_2}, \ldots\ldots, \mathbf{f_{n^\wedge 3}}\}) / \max \{\mathbf{f_1}, \mathbf{f_2}, \ldots\ldots, \mathbf{f_{n^\wedge 3}}\} \qquad \textbf{(2.1)}$$

When classifying an interaction, the normalized vectors of conjoint amino acids are generated for both proteins. As a result, there are two vectors of length n^3. In addition, the type of the interaction, either positive or negative, is added to the end of those vectors as a separate element. Thus, the feature vector size of a protein-protein interaction is $2*n^3+1$ for triplets where n is the number of amino acid groups. The calculation of feature vectors of different lengths is given below for a residue sequence of length m and amino acid categorization of group count n:

Singlets: Singlets are computed by adding up the frequency of each amino acid. There are normalized vectors of length n. Thus, the size of the feature vector is;

$$\mathbf{2*n+1} \qquad \textbf{(2.2)}$$

Since the amino acids are taken one by one, the sum of frequencies would be m.

Doublets: While calculating doublet frequencies, the same approach used in triplet calculation is followed. The only difference is doublets are created from two consecutive amino acids where triplets created from three. As a result, the size of the feature vector turns out to be;

$$\mathbf{2*n^2+1} \qquad \textbf{(2.3)}$$

The first doublet is generated from the amino acids at indexes 1-2 and the last doublet is generated from the amino acids at indexes (m-1)-m. Thus, the sum of doublet frequencies is m-1.

Triplets: The calculation of triplets is explained above with examples. Each time consecutive three amino acids are taken. Thus, the size of the feature vector is;

$$\mathbf{2*n^3+1} \qquad \textbf{(2.4)}$$

The first triplet is created from the first three amino acids (indexes 1-3) and the last triplet is created from the last three amino acids (indexes (m-2)-m). As a result, the sum of triplet frequencies is m-2.

Quadruplets: Quadruplets record consecutive four amino acids as a unit. There are n possibilities for each amino acid slot in and there are four amino acids in each quadruplet. That results in a feature vector of size;

$$\mathbf{2*n^4+1} \tag{2.5}$$

The quadruplets start from the amino acids at indexes 1-4 and end with the amino acids at indexes (m-3)-m. Thus, there are m-3 quadruplets in total.

2.3 Amino Acid Categorization Techniques

In this study, 20 natural amino acids are used and the rest of the amino acids are labeled as “Unknown”. So, there are 21 amino acid types to be used in classification. If those amino acids are recorded as triplets for SVM classification, a vector space of dimension $2*21^3+1=18.523$ is generated according to Equation (2.4). This large dimension creates sparse graph which is hard to classify. In order to reduce the size of the vectors generated, three amino acid categorization methods from literature are used. Those methods cluster amino acids based on their chemical and physical properties. The methods, which are used during this study, are given below.

2.3.1 Shen Categorization:

This categorization groups amino acids into 7 based on their dipoles and side chain volumes [3]. The groups of amino acids are shown in below in Table 2.1.

Table 2.1 Shen Categorization

[a] Dipole scale (Debye): -, Dipole<1.0; +, 1.0<Dipole<2.0; ++, 2.0<Dipole<3.0; ++, Dipole>3.0; +'+'+', Dipole>3.0 with opposite orientation.

[b] Volume scale (Å^3): -, Volume<50; +, Volume> 50.

[c] Cys is separated from class 3 because of its ability to form disulfide bonds.

Group #	**Dipole scale**[a]	**Volume scale**[b]	**Amino Acids**
1	-	-	ALA, GLY, VAL
2	-	+	ILE, LEU, PHE, PRO
3	+	+	TYR, MET, THR, SER
4	++	+	HIS, ASN, GLN, TPR
5	+++	+	ARG, LYS
6	+'+'+'	+	ASP, GLU
7	+ [c]	+	CYS

2.3.2 Sandberg Categorization:

In this categorization, amino acids are represented by their 26 physicochemical descriptor variables. These 26 variables are composed of experimentally verified properties (i.e. nuclear magnetic resonance) and basic properties of amino acids (i.e. molecular weight) [25]. Unlike Shen Categorization, “z-scales” are given for amino acids instead of group ids. In order to have comparable classes with Shen Categorization, the amino acids are clustered into 7 for each z-scale based on the z-scale values. Initially, the minimum and maximum z-scale values are found in order to group the amino acids. The difference of these values gives the overall range of z-scales. Then, z-scale interval for each group is found by dividing the difference value by the number of groups desired. For instance, for z_1 scale the minimum scale is -4.36 and maximum scale is 3.98 when the 20 natural amino acids are considered. So, the range of z-scale is 3.98 - (-4.36) = 8.34. If 10 groups will be created, the interval of groups will be 8.34 / 10 = 0.834. Thus, the amino acids with z-scale between -4.36 and (-4.36+0.834) are in the first group, the ones with z-scale between (-4.36+0.834) and (-4.36+2*0.834) are in the second group and so on. In order to assign a group to unnatural amino acids, the average of natural amino acid z-scales is used. The average is found by adding z-scales of the 20 natural amino acids and then dividing the sum by 20. The group that the average value falls in is the group of the unnatural amino acids. The 5 different z-scales are explained below and the categorization of amino acids based on these scales is given in Table 2.2.

I. z_1 scale is a descriptor of lipophilicity. The scales range between -4.36 and 3.98. Lipophilic amino acids have large negative values where polar and hydrophilic amino acids have large positive values [25].

II. z_2 scale describes molecular weight and surface area of amino acids. Negative z_2 corresponds to amino acids with low molecular weight and small surface area [25].

III. z_3 is the descriptor of polarity.

IV. z_4 and z_5 scales are created according to electro negativity, heat of formation, electrophilicity, hardness, and NMR.

Table 2.2 Categorization of Amino Acids based on z-scales

	z-scale categorization				
Group #	**z1**	**z2**	**z3**	**z4**	**z5**
1	PHE, ILE, LEU, TRP	GLY	LYS, ARG	GLU, ASP	CYS
2	MET, VAL, TYR	ALA, ARG, THR, VAL	ILE, LEU, GLN, VAL	ASN, GLN, SER, THR	TRP, TYR
3	PRO	CYS, ILE, LEU, SER	THR	ALA, GLY, ILE, LEU, VAL	MET
4	ALA, UNK	GLU, MET, GLN, PRO, UNK	ALA, GLU, GLY, HIS, MET, UNK, TRP, TYR	CYS, PHE, PRO, UNK, TYR	GLU, GLY, PHE, ARG, UNK, THR, VAL
5	CYS, THR	ASP, LYS, ASN	PHE, ASN, SER	LYS	ILE, HIS, LYS, GLN, SER
6	GLY, HIS, LYS, GLN, SER	PHE, HIS, TYR	ASP, PRO	MET, ARG	ALA, ASP, LEU
7	GLU, ASP, ASN, ARG	TRP	CYS	HIS, TRP	ASN, PRO

2.3.3 Murphy Categorization:

This categorization is made based on the physiochemical properties of amino acids. The categorization is made in 5 steps. The number of categories generated is 15, 10, 8, 4, and 2 [26]. During this study two of them are used:

2.3.3.1 8-Group Categorization

Previously given two categorizations use 7 groups. So, in order to have a similar size feature vector, 8 group clustering is used while creating triplet frequencies. These 8 groups are found by detecting structural homology using sequence alignments [26]. The amino acids and their groups are below given in Table 2.3.

Table 2.3 Murphy 8-group Amino Acid Categorization

Group #	Amino Acids
1	LEU, VAL, ILE, MET, CYS
2	ALA, GLY
3	SER, THR
4	PRO
5	PHE, TYR, TRP
6	GLU, ASP, ASN, GLN
7	LYS, ARG
8	HIS

2.3.3.2 2-Group Categorization:

While creating quadruplet frequencies, if clustering into 7 or 8 were used, the size of the feature vector would have been very large ($2*7^4+1 = 4803$). Thus, in order to reduce the size of the feature vector, grouping into 2 is used that is introduced as the most basic clustering [26]. In this method, amino acids are separated as hydrophobic/small and hydrophilic. The list of amino acids in these groups is given in Table 2.4.

Table 2.4 Murphy 2-group Amino Acid Categorization

Group	Amino Acids
Hyrophobic/Small	LEU, VAL, ILE, MET, CYS, ALA, GLY, SER, THR, PRO, PHE, TYR, TRP
Hyrophilic	GLU, ASP, ASN, GLN, LYS, ARG, HIS

2.4 A Machine Learning Tool: WEKA

WEKA (Waikato Environment for Knowledge Analysis) is the software implemented at University of Waikato in New Zealand. It contains libraries for machine learning and data mining algorithms. The libraries are developed in Java in order to enable platform independent usage [27]. WEKA provides packages for association, attribute selection, classification, clustering, etc. In this study, the learning scheme part, which contains classification methods, is mostly used. There are several classification methods given in learning scheme part such us Naïve Bayes, Decision Table, SMO, j48, and Linear Regression. SMO implements "Sequential Minimal Optimization" algorithms for SVM, which are introduced as a tool for binary classification problems [27, 28]. Currently there are many studies for extending classification to multiclasses [29]. In SVM, each input instance is represented in a high dimensional space nonlinearly using the selected features [30]. Then, the best surface is created that separates the two groups in classification problem, which is called optimal hyperplane [31]. Separation via optimal hyperplane is simple when the number of dimensions is few. The classification is harder for non-separable data in high dimensions. In order to classify instances successfully, Radial Basis Function kernels (RBF) that handles nonlinear polynomials can be used [32]. While using RBF, selection of parameters can improve the performance of classification [33]. A common technique is to make adjustments on the γ (gamma) parameter that determines the RBF width [34]. For instance, in the study of Shen et al. (2007), the best results are reached when γ parameter is set to 0.25. Due to its high performance, SMO will be used throughout this study.

2.5 Protein Interaction Databases

Several experimental techniques are used for identification of protein interactions in different organisms. The PPI found via those techniques are deposited in many different biological interaction databases. In this study, six different databases are used, which are commonly cited in literature. Descriptions of these databases are given below.

2.5.1 Human Protein Reference Database (HPRD)

HPRD contains the interactions of health and disease related human proteins. Nonredundant human proteins taken out from hundred thousands of articles gives information about experimental PPI, posttranslational modifications, enzyme/substrate relationships, disease associations, tissue expression, and subcellular localizations [35, 36].

2.5.2 Database of Interacting Proteins (DIP)

DIP provides experimentally verified protein-protein interactions of various organisms. The goal of DIP database is to merge the result of more than 20 experimental techniques. The protein interaction networks are also given in the database [37, 38].

2.5.3 Munich Information Center for Protein Sequences (MIPS)

MIPS contains a combined set of interactions of several organisms such as mammals, fungi, plants and microorganisms. Separate interaction lists for each organism is also provided in the database. In addition, MIPS presents a functional classification of proteins used in the database. All the interactions listed in the database are experimental [39, 40].

2.5.4 Biomolecular Interaction Network Database (BIND)

BIND contains interaction, molecular pathways, and pathway descriptions. BIND provides the experimental interactions that belong to various organisms. The interacting molecules might be proteins, nucleic acids, or small molecules. It is a useful database for learning protein interaction networks [41, 42].

2.5.5 Yeast Proteome Database (YPD)

YPD contains experimental protein-protein interactions taken from scientific resources. In addition to the interaction information of proteins, YPD gives biochemical function, localization, regulation, domain, and motif information. The proteins given in YPD belong to yeast organism [43, 44].

2.5.6 Interface Dataset

This dataset contains a list of interactions of proteins whose structural data are known. That means the binding sites of interacting proteins are identified. The interaction list is derived from PDB [45]. There are 49.000 interactions in the dataset. 17.210 of those interactions are classified as biological interactions and 10.545 are classified as crystal interactions [46].

2.6 Negative Dataset Creation Techniques

There are many experimental interaction sets are available but the difficulty appears at selection of negative dataset because a set of experimental non-interacting proteins does not exist [5]. Since verified negative set is not available, a gold standard negative set can be used instead in order to get a high precision of classification via SVM [47]. There are many studies in which computational negative sets are used. But even though those negative sets lead to high accuracy in classification, they have some drawbacks. Four of these methods, Jansen, Benhur, Shen, and Gough are given below.

2.6.1 Jansen Method:

This is one of the most common techniques used for creating negative interactions. In this method, interactions are created from proteins that have different subcellular locations [47]. While using this approach, first the locations of the proteins in the cell are found. Then, two proteins are randomly selected and if the selected proteins are from different locations in the cell, they are recorded as a negative interaction. For instance; a protein from nucleus and a protein from mitochondria is a sample negative interaction. But this is a biased approach since location constraint is applied on the interactions [6].

2.6.2 Benhur Method:

This approach is based on selecting random non-interacting protein pairs [48]. In this method, if randomly selected two proteins do not exist in positive interaction set, they are taken as a negative interaction. In another study, where a similar approach is used, it is claimed that only 1 of 600 possible yeast interactions is an actual interaction [23]. The ratio of real interactions to possible interactions is lower while using human proteins: only 1 of several thousands. The drawback of this approach is that since the protein-protein interaction networks are not complete, the protein pairs that are taken as negative interaction indeed might be interacting but has not been discovered yet [6].

2.6.3 Shen Method:

The third technique is used by Shen et al. (2007) and the main idea is to get the complement of positive interactions. In this method, two positive interactions are taken randomly and the cross list of the proteins in these interactions are taken as possible negative interactions. Then, if those possible negative interactions do not exist in the positive list, they are recorded as negative interactions. For instance, if AB and CD are positive interactions; AC, AD, BC, and BD are taken as negative interactions if they do not appear in positive set. The drawback is; the created interactions might be interacting in real. This method and its drawback are similar to Benhur method in a sense.

2.6.4 Gough Method

The last method for creating a negative list is introduced by Bock et al. (2001) [14]. In this method, random residue sequences are taken from DIP database and negative interactions are created by conserving the amino acid composition and n-gram frequencies of residue sequences. Random residue sequences are created via Shufflet method [49]. This method is different from the previous ones since artificial protein sequences are used.

2.7 Performance Evaluation of SVM Classification

The result of classification done by SVM is given using a few classification terms. Those classification terms are;

True Positive (TP): the interaction is positive and classified as positive.

True Negative (TN): the interaction is negative and classified as negative.

False Positive (FP): the interaction is negative but classified as positive.

False Negative (FN): the interaction is positive but classified as negative.

These terms are also summarized in Table 2.5.

When all the interactions are classified using the terms given in the table, the success of the classification is given using some statistical measures whose definitions are given below. Those terms are accuracy, precision, sensitivity, and f-measure.

Table 2.5 Classification Terms

		Condition Given	
		TRUE	**FALSE**
Test Outcome	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{2.6}$$

$$Positive\ \mathrm{Pr}\,ecision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2.7}$$

$$Negative\ \mathrm{Pr}\,ecision = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \tag{2.8}$$

$$Positive\ Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.9}$$

$$Negative\ Sensitivity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{2.10}$$

$$F - Measure = \frac{2 * \mathrm{TP}}{2 * \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2.11}$$

Where *positive precision* represents the ratio of the number of correctly classified positive interactions to the number of all interactions classified as positive interaction; *negative precision* is the ratio of the number of correctly classified negative interactions to the number of all interactions classifies as negative interaction; *positive sensitivity* is the proportion of number of correctly classified positive interactions to the all positive interactions and finally, *negative sensitivity* is the ratio of the number of correctly classified negative interactions to the number of all negative interactions. Here, *F-Measure* checks the balance between positive precision and positive sensitivity.

Chapter 3

ASSESSMENT of DATASETS

Currently, there are several experimental methods for discovering PPI such as yeast two-hybrid and mass spectrometry. These techniques are time and resource demanding. Therefore, the complete protein interaction network has not been formed yet [2]. Computational PPI prediction methods have been developed in order to use the predictions in studies where the verified interactions are not sufficient. While making predictions, firstly SVM, a machine learning algorithm, is applied on positive and negative interaction sets. Then, the interaction pattern learned from the training sets is used to make PPI predictions. The main focus of this chapter is to find the features that make SVM perform the best in terms of accuracy, by critical assessment of datasets selected from literature. Then, these features will be used to create combined datasets that predict interactions in every dataset accurately.

Section 3.1 introduces the ten datasets that are used for assessment in this study. For each dataset, the database, from which the positive interactions are taken, is given. Following that, the methods that are used while creating negative interactions and the total number of interactions are given.

The assessment of the ten datasets is made in Section 3.2. The features affecting the SVM performance are analyzed in detail. Several n-gram analyses, amino acid categorization methods, complete and partial residue sequences, datasets of different sizes, and a few negative set creation techniques are analyzed in order to achieve the best performance in SVM. Following that, the effects of using turn around interaction lists in training sets, different RBF and γ parameters, and combination of n-gram frequencies are questioned in order to improve the performance of SVM.

The features, which are found to be the best for SVM, are used to create models from each dataset. In Section 3.3, these models are used to predict the interactions in other datasets. By comparing the prediction scores of these models, the model that makes the most accurate PPI prediction is found.

3.1 Datasets

During this research, ten different datasets are used for comparison of SVM performances. The six verified interaction databases, which are previously introduced, are the source of positive interaction sets in these datasets. A dataset might contain verified interactions taken from one or more of those databases. On the other hand, the negative interactions in the datasets are obtained via computational methods. The datasets are composed as following:

1. **Shen Dataset:** The positive interactions of Shen dataset are extracted from HPRD version 2005_0913 [3, 35]. It contains 16.443 nonredundant experimental interactions in the positive set. The negative set also contains 16.443 interactions which are created computationally via Shen method.

2. **Jansen Dataset:** The Jansen dataset uses the positive interactions taken from MIPS database published in 2002 [50, 51]. The negative interactions are created computationally using the Jansen method. The positive set contains 8.617 interactions where as negative set contains 2.705.844 interactions. That many negative interactions are created in order to have a nonbiased interaction set.

3. **Peri Dataset:** Peri dataset contains positive interactions extracted from HPRD version 2007_0901 [35]. There are 37.107 interactions in the positive set. The negative interactions of Jansen dataset are used also in this dataset since no negative interactions are provided in HPRD [51].

4. **Benhur - BIND Dataset:** In the positive set, there are 10.517 interactions taken from MIPS dataset [52]. The computational negative interactions are created using the Benhur method and contains 10.517 interactions [48].

5. **Benhur - DIP-MIPS Dataset:** This dataset contains a combined list of experimental interactions taken from DIP and MIPS databases [39, 53]. There are 4.837 interactions in the positive set. The negative set is created computationally via Jansen method and contains 9.674 interactions [48].

6. **Deane Dataset:** Deane dataset contains the last published experimental interaction set in DIP database [37]. There are 6.459 PPI in the positive set. The negative interactions are the ones in Jansen dataset since there are no negative interactions in DIP [51].

7. **von Mering Dataset:** The positive set of this dataset is a combination of the interactions taken from MIPS and YPD. There are 80.000 interactions in the positive set and they are listed in sorted order of confidence [1, 44, 50]. The article does not provide any negative interaction set. Thus, the negative Jansen dataset is used also in this dataset [51].

8. **Tuncbag Dataset:** There are 17.210 interactions in Structural Interface Dataset but there are many redundant interactions in it [46]. So, redundancy elimination is made on those interactions resulting in a set of 2.607 biological nonredundant interactions, which are used in positive interaction list. Instead of using the crystal interactions, which are not proven to be non-interacting, a new computational negative dataset is created. This negative set is created using the proteins in positive set. Similar to Benhur Method, two random proteins are selected from positive interaction list and recorded as a negative interaction if that interaction does not exist in positive interaction sets of Deane and Peri Datasets. In this way, 250.000 interactions are created and after redundancy elimination, 203.495 negative interactions are left.

9. **Combined Dataset – 2 Clusters:** This dataset is a combination of some of the datasets given above. It is called *2 Clusters* because positive set is composed using two datasets. Positive and negative sets of this dataset is created as following:
    a. **Positive Set:** In this set, half of the interactions are randomly selected from Shen positive set and the other half is randomly selected from the Deane positive set.
    b. **Negative Set:** As in positive set, half of the interactions are randomly selected from the Shen dataset. Then, the intersection set of Jansen, Benhur - BIND, and Benhur - DIP-MIPS negative datasets is found. The second half of the negative interactions is randomly selected from the intersection set.

10. **Combined Dataset – 4 Clusters:** This dataset is a combination of all the datasets given above. It is labeled as *4 Clusters* because the datasets are clustered into four

as shown in Figure 3.1 based on their average prediction scores. Positive and negative sets of this dataset is created as following:

a. **Positive Set:** In this set, equal number of interactions is taken from each positive dataset cluster. If there are more than one dataset in a cluster, equal number of interaction is taken from each dataset in that cluster. For instance 1/8 of the interactions are taken from Shen dataset, 1/8 from Peri dataset, 1/4 from Jansen dataset and so on.

b. **Negative Set:** As in positive set, equal number of interaction is taken from each cluster. For the datasets where no negative set exists, the negative set of the other datasets in the cluster is used. For instance, von-Mering dataset is single in its cluster, so the negative Jansen dataset is used for that cluster.

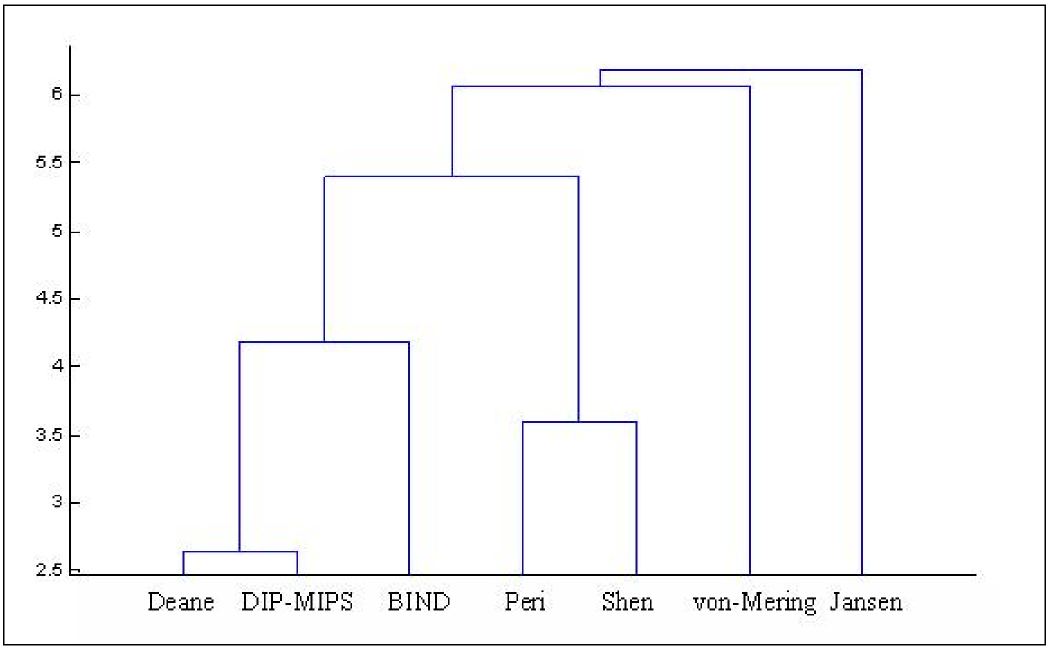


Figure 3.1 Clustering of Datasets

Table 3.1 Datasets Used

Name of the Dataset	# of Positive Interactions	# of Negative Interactions	Reference
Shen	16.443	16.443	[3]
Jansen	8.617	2.705.844	[51]
Peri	37.107	2.705.844	[36, 51]
Ben-Hur, BIND	10.517	10.517	[48, 52]
Ben-Hur, DIP-MIPS	4.837	9.674	[39, 48, 53]
Deane	6.549	2.705.844	[37, 51]
von Mering	80.000	2.705.844	[1, 37, 44, 50, 51]
Tuncbag	2.607	203.495	[46]
Combined - 2 Clusters	5.000	5.000	-
Combined - 4 Clusters	5.000	5.000	-

3.2 Features Affecting SVM Classification Performance

In this section, the aim is to find out the best features for SVM in order to generate a model from datasets that make predictions with minimum number of false positives. To achieve that, the effect of training sets and several key features on SVM performance is analyzed. In each subsection, a feature or property is studied in detail and the improvement on SVM performance is examined. Firstly, the best value for n in n-gram analysis is questioned. SVM classification is applied on several datasets and performances are compared. Following that, the outcome of using different amino acid categorizations is analyzed. The categorization that performs the best is desired to be selected for the rest of the study. Another important feature is the sequences of amino acids. PDB gives the sequence of structurally known part of the proteins where Swiss-Prot gives the complete sequence. In order to understand which one works better, SVM classification is applied on several datasets for both sequences. Then, the effect of the dataset size on SVM performance is examined. It is checked whether large or small size datasets are better classified.

Following that, new negative datasets are created via several computational methods. SVM is applied on positive and negative sets in order to find which computational method improves SVM performance the most. Then, the turn around of each interaction (i.e. B-A is the turn around interaction of A-B) is added to training set and the change of SVM performance for the new dataset is analyzed. The output of this analysis might be useful for small datasets where the number of interactions is desired to be incremented. In another section, the outcome of combining n-gram frequencies is analyzed. It is checked if using a single n-gram frequency works better or worse then using the combination of n-gram frequencies. The last section is different from previous sections because it focuses on SVM rather than inputs of SVM. In that section, the performance of the SVM algorithm is tried to be improved by using RBF and different γ parameters. Finally, the best function and parameter values are presented.

3.2.1 N-gram Frequencies

The sequence of proteins can be represented using n-gram frequency vectors. In order to find the best value for n, SVM classifications are made on Shen, Jansen, Benhur – BIND, and Benhur – DIP-MIPS datasets. These datasets are selected for testing because they are taken from articles where positive and negative sets are published together.

Table 3.2 shows the results of 5-fold cross validated SVM classification for each dataset. The results show that triplet and doublet frequencies are much better than singlet and quadruplet frequencies when amino acids are grouped via Shen Categorization. That is because the size of singlet frequency vectors is too small and the size of quadruplet frequency vectors is too large for classification. In singlet frequencies, there are 15 (2*7+1) elements in each vector and each interaction has very similar frequencies. So, the classification is very hard. In quadruplet frequencies, the frequency vector sizes are too large; thus sparse vector spaces of dimension 4.803 ($2*7^4+1$) are created. Classification of such a large space happens to be very hard. The accuracies for each dataset are a little higher than 50% when quadruplet frequencies are used. This is almost the same as randomly classifying the interactions because random classification is expected to have 50% accuracy in theory according to probability rules.

The difficult task is to make a selection between doublet and triplet frequencies. In Shen and Benhur – BIND datasets, triplet frequency vectors are better classified by SVM where in Benhur – DIP-MIPS dataset doublet frequency vectors work better. In Jansen dataset, triplet and doublet frequency vectors classification results are very close to each other. Thus, to find which one performs better, the number of interactions is increased and a new test is done to have a better idea about doublets and triplets.

Table 3.2 SVM Results of Different N-Gram Frequency Vectors of 2000 Interactions

Dataset	Freqs	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	Singlet	57,35%	0,58	0,57	0,53	0,62	2000
	Doublet	60,30%	0,61	0,60	0,58	0,63	2000
	Triplet	63,05%	0,71	0,59	0,44	0,82	2000
	Quadruplet	54,95%	0,95	0,53	0,11	0,99	2000
Jansen	Singlet	76,00%	0,89	0,69	0,59	0,93	2000
	Doublet	81,90%	0,91	0,76	0,71	0,93	2000
	Triplet	81,35%	0,94	0,74	0,67	0,96	2000
	Quadruplet	56,45%	0,95	0,53	0,14	0,99	2000
Benhur - BIND	Singlet	60,75%	0,61	0,61	0,61	0,61	2000
	Doublet	62,80%	0,64	0,62	0,60	0,66	2000
	Triplet	66,85%	0,71	0,64	0,57	0,77	2000
	Quadruplet	52,80%	0,80	0,51	0,08	0,98	2000
Benhur - DIP-MIPS	Singlet	54,70%	0,55	0,55	0,54	0,56	2000
	Doublet	59,40%	0,59	0,60	0,60	0,59	2000

	Triplet	56,45%	0,59	0,55	0,43	0,70	2000
	Quadruplet	51,20%	0,53	0,51	0,24	0,79	2000

Table 3.3 shows that as the number of interactions increase, triplet frequency vectors are classified more accurately than doublet frequency vectors for all datasets except Jansen. In Jansen dataset, accuracies are very similar but positive precision is better when triplet frequency vectors are used. As positive precision increases, the number of false positives decreases in predictions. Thus, triplet frequency vectors can also be used in Jansen dataset. So, it can be concluded that triplet frequency vectors work better than other frequency vectors and they will be used for the rest of the analyses.

Table 3.3 SVM Results of Different N-Gram Frequency Vectors of 4000 Interactions

Dataset	Freqs	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	Doublet	62,80%	0,63	0,62	0,61	0,65	4000
	Triplet	65,93%	0,74	0,62	0,49	0,83	4000
Jansen	Doublet	85,03%	0,93	0,80	0,76	0,94	4000
	Triplet	84,30%	0,97	0,77	0,71	0,97	4000
Benhur - BIND	Doublet	64,30%	0,65	0,64	0,63	0,66	4000
	Triplet	68,95%	0,75	0,65	0,57	0,81	4000
Benhur - DIP-MIPS	Doublet	60,70%	0,61	0,60	0,58	0,63	4000
	Triplet	62,38%	0,66	0,60	0,50	0,74	4000

3.2.2 Amino Acid Categorization

During this study, 21 types of amino acids are used in classification. 20 of them are the natural amino acids in nature where the 21st represents the rest of the amino acids. 21 amino acids lead to a huge vector space that is very hard and time consuming to classify when triplet frequencies are used ($2*21^3+1=18.523$). Thus, in order to reduce the size of the vectors generated for SVM, amino acids are categorized based on their pyhsico-chemical properties. To understand the effect of categorizing amino acids on SVM performances, three categorization techniques are used for testing. In addition, those three methods are also compared with the case where amino acids are not categorized.

In Shen Categorization, amino acids are grouped into 7. On the other hand, the Sandberg Categorization does not provide any groups but instead gives z-scales [25]. Initially z_1, z_2, and z_3 scales are used for testing since they are reported as the most informative scales [25]. For these three z-scales, categorization of amino acids into 5, 7, and 10 are tested in order to find the best grouping performance. Amino acids are grouped into 5 and 10 in the same way they are grouped into 7. Table 3.4 gives the results of SVM classification for z_1, z_2, and z_3 scales for groups of 5, 7, and 10 for randomly selected 1.000 positive and 1.000 negative interactions from Shen Dataset. The results show that grouping into 7 works better than grouping into 5 and 10 in all z-scales used. The performance of using 5 groups is worse than 7 groups but better than 10 groups. In addition, z_3 scale has the best performance and z_1 scale is as successful as the z_3 scale.

Table 3.4 SVM Classification Results of z_1, z_2, and z_3 scales according to groups of size 5, 7, and 10

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	z_1	5	62,15%	0,62	0,62	0,64	0,61	2000
		7	63,85%	0,61	0,68	0,75	0,53	2000
		10	59,50%	0,64	0,57	0,44	0,75	2000
	z_2	5	61,05%	0,61	0,61	0,61	0,61	2000
		7	62,90%	0,62	0,65	0,69	0,57	2000
		10	56,15%	0,57	0,55	0,49	0,64	2000
	z_3	5	60,60%	0,61	0,61	0,60	0,61	2000
		7	63,90%	0,64	0,64	0,64	0,64	2000
		10	60,20%	0,69	0,57	0,37	0,84	2000

In order to improve SVM performance, frequency vectors of these three z-scales are united for testing. In this test, the frequency vector of each z-scale is computed and added one after another, resulting in a vector of size $3*(2n^3+1)$, where n is the number of amino acid groups. The union of frequency vectors is tested via SVM for grouping into 5 and 7. This test is not done for grouping into 10 because of its poor performance in the previous test. Table 3.5 summarizes the results of united z-scale vectors. In order to make results comparable with the previous test, the interaction lists in previous test are used. The table shows that uniting the z-scale vectors is not an efficient method because in 5 group case, the performance of united vectors is almost the same as the individual performance of z_1 scale vector. In addition, in 7 group case, the performance of combined list is much worse than the individual scores of z-scales. This might be due to the large size of the frequency vectors generated by uniting the three separate vectors ($3*(2*7^3+1) = 2059$).

Table 3.5 SVM Classification Results of Union of 5 and 7 groups of Z-Scales

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	z_1-z_2-z_3	5	62,80%	0,61	0,66	0,73	0,53	2000
		7	55,05%	0,53	0,76	0,95	0,15	2000

The analyses that are done so far show the effect of z-scale categorization on SVM performance. In another study, frequency vectors are created for interactions where amino acids are not categorized into groups. Then, the same interaction lists are used to create frequency vectors based on Sandberg and Shen categorizations. Table 3.6 makes a comparison of these vectors based on SVM performances. The results show that grouping works better than the case where no categorization is applied. That is because when amino acids are not grouped, the vector space gets too large ($2*21^3+1$ = 18.523), which is hard to classify for SVM. When categorization scores are compared, it can be concluded that z_4 performs the best, and z_5 performs the worst.

Table 3.6 SVM Classification Results of Shen Dataset for Shen and Sandberg Categorizations

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	None	21	62,55%	0,63	0,62	0,61	0,65	2000
	Shen	7	63,05%	0,71	0,59	0,44	0,82	2000
	z_1	7	63,85%	0,61	0,68	0,75	0,53	2000
	z_2	7	62,90%	0,62	0,65	0,69	0,57	2000
	z_3	7	63,90%	0,64	0,64	0,64	0,64	2000
	z_4	7	64,65%	0,68	0,63	0,56	0,73	2000
	z_5	7	62,85%	0,62	0,64	0,68	0,58	2000

The tests on Shen and Sandberg categorizations are also done for Tuncbag Dataset in order to find the training set that eliminates false positive interactions the most. The tests are applied on randomly selected 1.000 positive and 1.000 negative nonredundant interactions. The results of those tests are presented in Table 3.7. Unlike the case in Shen Dataset, z_5 performs the best in Tuncbag Dataset. Two more tests are applied on Tuncbag Dataset. In the first test, a binary classification of amino acids is obtained using the first three z-scales. The binary classification is done as following: The average score for each z-scale is computed and the amino acids above the average are labeled as 1, the ones below the average are labeled as 0. When these labels are found for each amino acid based on three z-scales, each amino acid had 3 labels that might generate 8 (2*2*2) different groups since there are two possibilities for each z-scale. The group numbers are computed just like binary number computation. For instance, if an amino acid is labeled as 1, 0, 1 in z_1, z_2, z_3 respectively, it is recorded in group 5 ($1*2^2 + 0*2^1 + 1*2^0 = 5$). The result of this classification is also given in Table 3.7 and it performs much worse than individual z-scales.

The tests given in Table 3.5 show that union of the z-scale vectors do not improve performance. Thus, in the second test, instead of uniting z-scale vectors, a z-scale is combined with the Shen Categorization. In this test, z_3 is used because it is the most successful one in Table 3.4. The result of this test is also given in Table 3.7. This is the worst score for Tuncbag dataset when compared to other tests in the table. Consequently, union of the categorizations or applying binary categorization does not help to improve SVM classification accuracy.

Table 3.7 SVM Classification Results of Tuncbag Dataset for Shen and Sandberg Categorizations

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Tuncbag	Shen	7	67,65%	0,76	0,63	0,52	0,84	2000
	z_1	7	65,20%	0,76	0,61	0,45	0,86	2000
	z_2	7	68,20%	0,77	0,64	0,51	0,85	2000
	z_3	7	72,60%	0,80	0,68	0,61	0,85	2000
	z_4	7	71,90%	0,75	0,70	0,66	0,77	2000
	z_5	7	74,60%	0,78	0,72	0,68	0,81	2000
	z_1-z_2-z_3 binary	8 (7)	62,60%	0,73	0,59	0,40	0,85	2000
	Shen-z_3	7	60,35%	0,76	0,56	0,31	0,90	2000

The comparison of Sandberg and Shen Categorization for other datasets is given in Table 3.8 to Table 3.14:

Table 3.8 summarizes the SVM results for Jansen Dataset. In this dataset, z_4 performs the best where z_3 and Shen categorizations also perform well. In addition, the performance of each categorization is better in Jansen dataset than the performances in Shen and Tuncbag datasets.

Table 3.8 SVM Classification Results of Jansen Dataset for Shen and Sandberg Categorizations

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Jansen	Shen	7	81,35%	0,94	0,74	0,67	0,96	2000
	z_1	7	76,15%	0,73	0,80	0,83	0,69	2000
	z_2	7	78,10%	0,94	0,71	0,60	0,96	2000
	z_3	7	81,75%	0,91	0,76	0,71	0,93	2000
	z_4	7	83,60%	0,89	0,80	0,77	0,90	2000
	z_5	7	80,45%	0,90	0,75	0,69	0,92	2000

In Table 3.9**,** classification of randomly selected 1.000 interactions from Peri Dataset is presented. Similar to Jansen Dataset, the accuracy values are higher than Shen and Tuncbag datasets. On the other hand, in contrast to Jansen dataset, Shen categorization performs the worst. When z-scales are compared, z_2 performs the best.

Table 3.9 SVM Classification Results of Peri Dataset for Shen and Sandberg Categorizations

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Peri	Shen	7	76,60%	0,87	0,71	0,63	0,91	2000
	z_1	7	81,50%	0,78	0,87	0,89	0,75	2000
	z_2	7	83,85%	0,96	0,77	0,70	0,97	2000
	z_3	7	82,00%	0,92	0,76	0,70	0,94	2000
	z_4	7	82,75%	0,91	0,77	0,73	0,93	2000
	z_5	7	82,95%	0,88	0,79	0,76	0,90	2000

Classification of Benhur – BIND dataset is given in Table 3.10 and the results are not similar to previous datasets. Shen categorization performs the best and z_4 performs as well as that. Overall accuracy values are not as high as Peri dataset but nearly the same as the accuracy values of Shen dataset.

Table 3.10 SVM Classification Results of Benhur - BIND Dataset for Shen and Sandberg Categorizations

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Benhur - BIND	Shen	7	66,85%	0,71	0,64	0,57	0,77	2000
	z_1	7	64,80%	0,69	0,62	0,54	0,75	2000
	z_2	7	65,45%	0,72	0,62	0,50	0,81	2000
	z_3	7	63,50%	0,66	0,62	0,55	0,72	2000
	z_4	7	66,30%	0,70	0,64	0,57	0,76	2000
	z_5	7	65,65%	0,69	0,63	0,57	0,74	2000

When randomly selected 1.000 interactions from Benhur – DIP-MIPS dataset are classified via Shen and Sandberg categorizations, the worst accuracies are achieved when compared to other datasets. z_1 and z_4 performs the best and z_5 performs as well as those two. Similar to the case in Peri dataset, Shen categorization performs the worst. The results are given in Table 3.11.

Table 3.11 SVM Classification Results of Benhur – DIP-MIPS Dataset for Shen and Sandberg Categorizations

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Benhur - DIP-MIPS	Shen	7	56,45%	0,59	0,55	0,43	0,70	2000
	z_1	7	59,75%	0,61	0,59	0,53	0,67	2000
	z_2	7	58,60%	0,62	0,57	0,46	0,72	2000
	z_3	7	57,50%	0,59	0,57	0,50	0,65	2000
	z_4	7	59,75%	0,62	0,58	0,50	0,69	2000
	z_5	7	59,45%	0,61	0,58	0,51	0,68	2000

In Deane dataset, similar to Peri and Benhur – DIP-MIPS datasets, Shen Categorization performs the worst. z_4 and z5 performs the best and z_3 also has a high accuracy. The results for Deane Dataset are given in Table 3.12.

Table 3.12 SVM Classification Results of Deane Dataset for Shen and Sandberg Categorizations

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Deane	Shen	7	66,10%	0,73	0,62	0,51	0,82	2000
	z_1	7	67,05%	0,71	0,64	0,58	0,76	2000
	z_2	7	67,15%	0,73	0,64	0,54	0,80	2000
	z_3	7	68,50%	0,75	0,65	0,56	0,81	2000
	z_4	7	69,00%	0,73	0,66	0,60	0,78	2000
	z_5	7	69,00%	0,73	0,66	0,60	0,78	2000

Performance of categorizations in von Mering dataset is similar to Deane Dataset. z_3 performs the best and Shen, z_1 and z_2 Categorizations have poor accuracy values. The interesting point in this dataset is the similar positive precisions, which are given in Table 3.13.

Table 3.13 SVM Classification Results of von Mering Dataset for Shen and Sandberg Categorizations

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
von Mering	Shen	7	65,20%	0,72	0,62	0,50	0,81	2000
	z_1	7	65,85%	0,73	0,62	0,50	0,82	2000
	z_2	7	65,70%	0,72	0,62	0,51	0,81	2000
	z_3	7	68,45%	0,74	0,65	0,57	0,80	2000
	z_4	7	67,60%	0,73	0,64	0,57	0,79	2000
	z_5	7	66,80%	0,72	0,64	0,55	0,79	2000

Table 3.14 summarizes the SVM results for Combined Dataset – 2 Clusters. This dataset has high accuracies when compared to Benhur – BIND and Benhur – DIP-MIPS datasets. Different from the previous datasets, z_1 performs better than other categorizations. Z_2 performs as well as z_1, where Shen Categorization has the worst performance.

Table 3.14 SVM Classification Results of Combined Dataset – 2 Clusters for Shen and Sandberg Categorizations

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Combined Set – 2 Clusters	Shen	7	69,85%	0,82	0,64	0,51	0,88	2000
	z_1	7	74,45%	0,73	0,76	0,78	0,71	2000
	z_2	7	74,00%	0,87	0,68	0,56	0,92	2000
	z_3	7	71,20%	0,84	0,65	0,52	0,90	2000
	z_4	7	70,70%	0,80	0,66	0,55	0,86	2000
	z_5	7	72,85%	0,79	0,69	0,63	0,83	2000

When the accuracies for nine datasets are compared, it can be concluded that there is no best categorization. Shen categorization performs the best in Benhur – BIND dataset but performs the worst in Benhur – DIP-MIPS dataset. In general, one of the z-scales performs the best but the type of the best z-scale is dependent on the dataset used.

At the end of this section, Murphy categorization is applied on datasets in order to understand the effect of categorization by comparing the results with the case where no categorization is applied. For this test, the datasets used for Table 3.2 are used. The classification accuracy values of the datasets are given in Table 3.15. The table shows that Shen or Murphy categorizations always perform better than uncategorized amino acids in every dataset. Besides, the size of frequency vectors generated for uncategorized amino acids is huge ($2*21^3+1$). That causes SVM to take too much time to train data when compared to categorized cases. When all these results are considered, it can be concluded that categorizing amino acids is much better than uncategorizing.

Table 3.15 SVM Classification Results of 4 datasets for Shen, Murphy Categorizations and Uncategorized Amino Acids

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	Shen	7	63,05%	0,71	0,59	0,44	0,82	2000
	Murphy	8	65,20%	0,70	0,62	0,54	0,76	2000
	No Classes	21	62,55%	0,63	0,62	0,61	0,65	2000
Jansen	Shen	7	81,35%	0,94	0,74	0,67	0,96	2000
	Murphy	8	75,30%	0,72	0,80	0,83	0,68	2000
	No Classes	21	78,80%	0,80	0,78	0,77	0,81	2000
Benhur - BIND	Shen	7	66,85%	0,71	0,64	0,57	0,77	2000
	Murphy	8	63,95%	0,69	0,61	0,50	0,78	2000
	No Classes	21	61,10%	0,61	0,62	0,63	0,59	2000
Benhur - DIP-MIPS	Shen	7	56,45%	0,59	0,55	0,43	0,70	2000
	Murphy	8	58,65%	0,61	0,57	0,47	0,71	2000
	No Classes	21	58,40%	0,59	0,58	0,54	0,63	2000

In the two datasets given in Table 3.15 Shen Categorization performs better than Murphy Categorization and in the remaining the opposite works. In order to understand which one is better, another test is made to compare performances of Shen dataset. Randomly selected 5.000 positive and 5.000 negative interactions from Shen Dataset are classified using Shen and Murphy categorizations. The results are given in Table 3.16. The results show that although Murphy Categorization works better then Shen Categorization for a random set of 2.000 interactions, they perform almost the same when the interaction size gets larger. In addition, SVM classification takes less time for Shen Categorization since it has seven groups (frequency vector size = $2*7^3+1$) where Murphy Categorization has eight groups (frequency vector size = $2*8^3+1$). Thus, Shen Categorization is preferred over Murphy Categorization since its positive precision value is better and the frequency vector size is smaller.

Table 3.16 SVM Classification Results of Shen dataset for Shen and Murphy Categorizations

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	Shen	7	71,64%	0,80	0,67	0,58	0,86	9971
	Murphy	8	71,67%	0,77	0,68	0,62	0,81	9971

3.2.3 Complete versus Partial Residue Sequences

The interactions in Tuncbag Dataset are given with PDB ids. There are two sequence choices for those proteins. The first choice is getting the residue sequences of the structurally known part of the proteins, which are FASTA sequences provided in PDB [45]. The second choice is getting the complete residue sequence of proteins that are given in Swiss-Prot database [54] [55]. Since sequences in PDB consist of structurally known part of the proteins, they are shorter than the sequences in Swiss-Prot database.

Tuncbag Dataset is used in SVM classification in order to find which sequence is more distinctive. The tests are done using Shen Categorization on both randomly selected 1.000 interactions and full list of interactions. The results are shown in Table 3.17 and indicate that using the complete sequence of proteins provided in Swiss-Prot database works better in SVM classification. It is concluded that the structurally unknown part of the proteins is important in classification.

Table 3.17 SVM Classification Results of Tuncbag Dataset Comparing PDB and Swiss-Prot Sequences

Dataset	Sequence	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Tuncbag	Swiss-Prot Sequnce	67,65%	0,76	0,63	0,52	0,84	2000
		74,05%	0,85	0,68	0,58	0,90	5214

	Pdb Sequence	63,75%	0,74	0,60	0,43	0,85	2000
		72,80%	0,81	0,68	0,59	0,86	5214

3.2.4 Dataset Size

The SVM classification tests made in previous sections are mostly made on randomly selected sets of 1.000 positive and 1.000 negative interactions. Because, the results are taken quickly when a small size set is used in SVM. However, all datasets have more than 1.000 interactions and those interactions can be used in classification to improve performance because more interactions help SVM to learn better.

Table 3.18 summarizes the results of SVM classification for interaction sets of different sizes taken from the same dataset. In this test, Shen dataset is used in order to compare the obtained results with the published results [3]. The first 4 interaction sets are composed of randomly selected interactions from Shen dataset. In all tests, 5-fold cross validation is used with the same parameters. The results show that as the number of interactions increases, the accuracy and precision values also increase. The SVM classification on full dataset results in 88% positive precision which is better than published average result which is 84.21%.

Table 3.18 SVM results for test sets of different sizes from Shen Dataset

Dataset	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	61,90%	0,68	0,59	0,46	0,78	1995
	66,57%	0,74	0,63	0,50	0,83	3990
	71,64%	0,80	0,67	0,58	0,86	9971
	76,30%	0,85	0,71	0,64	0,89	19944
	79,58%	0,88	0,74	0,68	0,91	32336

The results in Table 3.18 show the importance of input size for classification. It can be summarized that, to obtain the best SVM performance and have the best models of each dataset; classifications should be made using maximum number of interactions. However, in all datasets, the number of positive interactions is less than the number of negative interactions. The reason is, positive interactions are experimental and there are not many verified interactions. On the other hand, infinite number of negative interactions can be generated using different computational methods. Thus, the number of positive interactions is the limiting factor in training set size.

Table 3.19 gives the results of classifications for each dataset where all interactions are used in SVM. The best accuracy values are achieved for Peri dataset. It is followed by Jansen Dataset. The worst performing dataset is Benhur – DIP-MIPS, which performs almost 30% worse than Peri Dataset in terms of accuracy. It can be noticed from the table that, the selection of training datasets is as effective as to the number of interactions in the datasets. For instance, von Mering Dataset performs better than Benhur – DIP-MIPS Dataset although its size is almost half the size of Benhur – DIP-MIPS Dataset.

Table 3.19 Overall Accuracies of Datasets

Dataset	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	79,58%	0,88	0,74	0,68	0,91	32336
Jansen	88,10%	0,98	0,83	0,77	0,98	15697
Peri	95,55%	0,99	0,93	0,93	0,99	33005
Benhur - BIND	75,52%	0,84	0,71	0,64	0,88	19308
Benhur - DIP-MIPS	66,20%	0,71	0,63	0,56	0,77	9581
Deane	78,76%	0,88	0,73	0,67	0,91	12842
von Mering	71,28%	0,77	0,67	0,61	0,81	4910
Tuncbag	74,05%	0,85	0,68	0,58	0,90	5214
Combined - 2 Clusters	69,75%	0,78	0,65	0,54	0,85	10000
Combined - 4 Clusters	78,00%	0,71	0,92	0,95	0,61	10000

3.2.5 Negative Set Creation

There are several negative set creation methods, two of which are Shen and Jansen methods. In order to analyze the effect of negative set creation method on SVM performance, negative sets are created using both methods for four datasets. Table 3.20 shows the results of SVM classification using these methods. For each dataset, the results are given for three negative sets where the positive set is kept constant. The first result gives the accuracy of using original negative set given in dataset. The second one is the result of using Shen Method for creating negative set and the third one is the result of using Jansen Method. In Shen and Jansen datasets, there are two rows instead of three because in Shen dataset, original negative set is created using Shen method and in Jansen dataset original negative set is created using Jansen method. Thus, there is no need to duplicate result rows.

The accuracies show that Jansen method works better than both the original negative sets of the datasets and the Shen method except for the Shen dataset. In Benhur - DIP-MIPS dataset, both methods work better than the original set, which shows the weakness of the original negative set. In Benhur - BIND dataset, Jansen Method and original dataset have similar performance but better than Shen method. In Jansen and Shen datasets, original sets work better than new techniques applied. The precision values make Jansen method attractive but there are some criticisms about Jansen method, which is about creating a biased interaction set [6].

Table 3.20 Affect of different negative set creation methods

Dataset	Neg. Set Creation Method	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	Original Set = Shen Method	63,05%	0,71	0,59	0,44	0,82	2000
	Jansen Method	61,00%	0,60	0,63	0,68	0,54	2000
Jansen	Original Set = Jansen Method	81,35%	0,94	0,74	0,67	0,96	2000

	Shen Method	74,05%	0,87	0,68	0,57	0,92	2000
Benhur - BIND	Original Set	66,85%	0,71	0,64	0,57	0,77	2000
	Shen Method	63,25%	0,66	0,61	0,54	0,73	2000
	Jansen Method	67,35%	0,65	0,70	0,74	0,61	2000
Benhur - DIP-MIPS	Original Set	56,45%	0,59	0,55	0,43	0,70	2000
	Shen Method	59,30%	0,61	0,58	0,51	0,67	2000
	Jansen Method	62,90%	0,62	0,63	0,65	0,61	2000

3.2.6 Turn Around of Interactions

Analysis in Section 3.2.4 proves that the increase in the number of interactions used in classification improves the SVM performance. The more interactions used in training set, the higher the accuracy. But in some datasets, there are not many experimental interactions. Thus, in order to have a better learning scheme for SVM, the turn around of the interactions can be used in classification. In databases, the interactions are given in the form A-B where A and B represent the protein ids. But there is no difference between saying A interacts with B and B interacts with A. So, that means A-B and B-A interactions are the same. As a consequence, number of interactions in datasets can be doubled using turn around of interaction lists.

In order to test the effect of using turn around of interaction lists, classifications are made on the datasets used in Table 3.2**.** The results are given in Table 3.21. The first row for each dataset is the accuracy value for randomly selected 1.000 positive and 1.000 negative interactions. The second row is the accuracy that is achieved when the turn around of the interactions in the first row are added to training set. That means the second row is the result of training 2.000 positive and 2.000 negative interactions. The table shows that using the same interaction set, the SVM performance can be increased when the turn around of the interactions are used. But the amount of performance increase in is not the same in all datasets. For instance, in Benhur – DIP-MIPS dataset performance increases 4% where the increase is only 0.23% in Jansen dataset. These results are in parallel with the results given in Section 3.2.4. Unfortunately, some databases do not have as many verified interactions as in HPRD or DIP. Therefore, in order to get a better performance, the turn around of interaction lists can be used in training sets for the databases with few interactions.

Table 3.21 Accuracies of Datasets for Turn Around of Interaction Lists

Dataset	Inter. List	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	Org. List	63,05%	0,71	0,59	0,44	0,82	2000
	Org. & Turn Arn. List	63,50%	0,72	0,60	0,44	0,83	4000
Jansen	Org. List	81,35%	0,94	0,74	0,67	0,96	2000
	Org. & Turn Arn. List	81,58%	0,94	0,75	0,67	0,96	4000
Benhur - BIND	Org. List	66,85%	0,71	0,64	0,57	0,77	2000
	Org. & Turn Arn. List	69,08%	0,74	0,66	0,59	0,80	4000
Benhur - DIP-MIPS	Org. List	56,45%	0,59	0,55	0,43	0,70	2000
	Org. & Turn Arn. List	60,48%	0,64	0,58	0,47	0,74	4000

3.2.7 Multiple Frequency Vectors

Instead of using only the triplet frequencies, union of several frequency vectors might be used to improve the SVM performance. Two tests are made in this section in order to analyze the effect of uniting frequency vectors on SVM performance. In the first test, the frequencies of amino acids without classification are united with triplet frequencies. There are 20 natural amino acids and the rest of amino acids are taken as the 21st amino acid. So, the new vector contains 42 (21*2) elements for singlets and 686 (7^3*2) for triplets which adds up to a vector of size 728 (42+686+1) where the last element classifies the interaction class (positive or negative).

In the second test, singlet, doublet, triplet, and quadruplet frequencies are united. The singlet frequencies are obtained as in the first test. The doublet and triplet frequencies are found using the Shen Categorization that creates a vector of 49 (7^2*2) for doublets and 686 for triplets. For quadruplet frequencies, 2 group classification of Murphy categorization is used where the groups are created based on hydrophobicity. Since there are two groups, a vector of size 32 (2^4*2) is added for quadruplets. While calculating the doublet and quadruplet frequencies, the same method used for triplet calculation is used. The only difference is the number of consecutive amino acids taken each time for frequency calculation. In triplets, consecutive 3 amino acids are taken where 2 amino acids are taken in doublets and 4 amino acids in quadruplets. In total, the size of the vector used in the second test is 859 (($21+7^2+7^3+2^4$)*2 = 1).

The tests explained above are performed on the datasets that are used in previous sections and the results are given in Table 3.22. The results show that classification performance is the best when only triplet frequencies are used for all the datasets tested. These results are similar to the results given in Section 3.2.2.

Table 3.22 Affect of using multiple frequency vectors in classification

Dataset	Input Vector	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	Triplet	63,05%	0,71	0,59	0,44	0,82	2000
	Singlet-Triplet	62,25%	0,69	0,59	0,45	0,80	2000
	Singlet-Doublet-Triplet-Quadruplet	61,45%	0,70	0,58	0,40	0,83	2000
Jansen	Triplet	81,35%	0,94	0,74	0,67	0,96	2000
	Singlet-Triplet	76,85%	0,91	0,70	0,60	0,94	2000
	Singlet-Doublet-Triplet-Quadruplet	76,75%	0,92	0,70	0,58	0,95	2000
Benhur - BIND	Triplet	66,85%	0,71	0,64	0,57	0,77	2000
	Singlet-Triplet	65,90%	0,70	0,63	0,55	0,77	2000
	Singlet-Doublet-Triplet-Quadruplet	59,15%	0,65	0,57	0,41	0,78	2000

Benhur - DIP-MIPS	Triplet	56,45%	0,59	0,55	0,43	0,70	2000
	Singlet-Triplet	55,80%	0,57	0,55	0,47	0,65	2000
	Singlet-Doublet-Triplet-Quadruplet	55,50%	0,58	0,54	0,39	0,72	2000

3.2.8 RBF and γ (Gamma) Parameter

In order to find the best γ value for SVM, a few tests are made on randomly selected 1.000 positive and 1.000 negative interactions from Shen and Benhur - BIND datasets. In addition, using linear polynomial kernels instead of RBF is tested in this section. The Table 3.23 and Table 3.24 show the accuracy and precision values for different γ parameters for Shen and Benhur – BIND datasets respectively. The results show that using RBF is much better than using linear polynomial kernels. Using γ = 0.25 with RBF instead of linear polynomial kernels increments classification accuracy from 56.20% to 63.05% in Shen dataset and from 61.45% to 66.85% in Benhur - BIND dataset. The comparison of different γ parameters shows that 0.25 is better than the others. For instance, using 0.25 instead of 0.01 (default parameter for RBF in WEKA) improves accuracy from 50% to 60.05% in Shen dataset and from 63.45% to 66.85% in Benhur - BIND dataset. Besides, when the γ parameter gets closer to 1, positive sensitivity gets closer to 0 except the case where γ is 1 in Benhur - BIND dataset.

Table 3.23 Comparison of γ parameters and RBF for Shen Dataset

Dataset	RBF	Gamma	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	-	-	56,20%	0,56	0,56	0,57	0,56	2000
	+	0,01 (default)	50,00%	0,67	0,38	0,44	0,61	2000
	+	0,25	63,05%	0,71	0,59	0,44	0,82	2000
	+	0,4	61,20%	0,76	0,57	0,33	0,89	2000
	+	0,6	59,75%	0,85	0,56	0,24	0,96	2000

	+	0,8	59,85%	0,90	0,56	0,22	0,98	2000
	+	1	50,00%	0,98	0,46	0,12	1,00	2000

Table 3.24 Comparison of γ parameters and RBF for Benhur - BIND Dataset

Dataset	RBF	Gamma	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Benhur - BIND	-	-	61,45%	0,61	0,62	0,63	0,60	2000
	+	0,01 (default)	63,45%	0,64	0,63	0,60	0,67	2000
	+	0,25	66,85%	0,71	0,64	0,57	0,77	2000
	+	0,4	59,35%	0,76	0,56	0,27	0,92	2000
	+	0,6	54,70%	0,76	0,53	0,14	0,96	2000
	+	0,8	54,55%	0,80	0,52	0,12	0,97	2000
	+	1	52,50%	0,55	0,52	0,26	0,79	2000

3.3 Applying SVM Models on All Datasets

When SVM classification is applied on a training set, it creates a model based on the positive and negative interactions in the training set. That model can be applied on other interaction sets. When a model is applied on another dataset, it categorizes the given interactions as positive or negative and then compares its predictions with the actual type of the interactions given in that dataset. **Error! Reference source not found.** gives the prediction results achieved by applying the models of each dataset on others. The first ten datasets given in the table are the ones introduced at the beginning. The last two datasets are created using the cross datasets of *Combined – 2 Clusters* and *Combined – 4 Clusters* datasets in order to evaluate the combination of datasets. *Combined 2 Clusters Pos – 4 Clusters Neg* dataset is created using the positive set of *Combined – 2 Clusters* and the negative set of *Combined – 4 Clusters*. On the other hand, *Combined 4 Clusters Pos – 2 Clusters Neg* dataset is created using the positive set of *Combined – 4 Clusters* and the negative set of *Combined – 2 Clusters*.

Table 3. 25 Applying dataset models on each other

Model Set	Test Set	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Shen	Jansen	51,53%	0,43	0,52	0,09	0,89	15697
	Peri	69,00%	0,82	0,63	0,50	0,89	33365
	Benhur - BIND	51,28%	0,54	0,51	0,19	0,83	19308
	Benhur - DIP-MIPS	53,53%	0,59	0,52	0,25	0,83	9581
	Deane	57,14%	0,69	0,54	0,26	0,89	12842
	von Mering	51,38%	0,56	0,50	0,20	0,84	4910
	Tuncbag	62,50%	0,71	0,59	0,43	0,82	5214
	Combined - 2 Clusters	77,85%	0,90	0,71	0,63	0,93	10000
	Combined - 4 Clusters	58,29%	0,67	0,55	0,32	0,84	10000
	Average	59,17%	65,70%	56,55%	31,77%	86,28%	
Jansen	Shen	50,71%	0,78	0,50	0,02	0,99	32336
	Peri	49,14%	0,51	0,49	0,02	0,98	33365
	Benhur - BIND	51,05%	0,60	0,51	0,06	0,96	19308
	Benhur - DIP-MIPS	58,53%	0,81	0,55	0,23	0,95	9581
	Deane	55,36%	0,88	0,53	0,12	0,98	12842
	von Mering	54,32%	0,71	0,52	0,17	0,93	4910
	Tuncbag	53,61%	0,75	0,52	0,11	0,96	5214
	Combined - 2 Clusters	52,86%	0,87	0,51	0,07	0,99	10000
	Combined - 4 Clusters	66,07%	0,95	0,60	0,34	0,98	10000
	Average	54,63%	0,76	0,53	0,13	0,97	
Peri	Shen	53,90%	0,52	0,69	0,94	0,14	32336
	Jansen	57,16%	0,91	0,55	0,10	0,99	15697
	Benhur - BIND	47,28%	0,40	0,48	0,11	0,84	19308
	Benhur - DIP-MIPS	51,16%	0,56	0,50	0,13	0,89	9581
	Deane	59,32%	0,94	0,55	0,20	0,99	12842
	von Mering	49,19%	0,50	0,49	0,19	0,81	4910

	Tuncbag	50,44%	0,50	0,51	0,71	0,30	5214
	Combined - 2 Clusters	56,72%	0,57	0,57	0,57	0,56	10000
	Combined - 4 Clusters	54,04%	0,56	0,53	0,38	0,70	10000
	Average	53,24%	0,61	0,54	0,37	0,69	
Benhur - BIND	Shen	51,22%	0,79	0,51	0,03	0,99	32336
	Jansen	52,72%	0,49	0,53	0,11	0,90	15697
	Peri	45,77%	0,23	0,47	0,03	0,90	33365
	Benhur - DIP-MIPS	62,01%	0,69	0,59	0,45	0,80	9581
	Deane	62,14%	0,78	0,58	0,34	0,90	12842
	von Mering	50,31%	0,54	0,50	0,15	0,87	4910
	Tuncbag	51,98%	0,78	0,51	0,06	0,98	5214
	Combined - 2 Clusters	56,07%	0,74	0,53	0,19	0,93	10000
	Combined - 4 Clusters	54,24%	0,61	0,53	0,23	0,85	10000
	Average	54,05%	0,63	0,53	0,18	0,90	
Benhur - DIP-MIPS	Shen	55,02%	0,75	0,53	0,15	0,95	32336
	Jansen	57,50%	0,63	0,56	0,23	0,88	15697
	Peri	49,85%	0,53	0,49	0,14	0,87	33365
	Benhur - BIND	58,93%	0,70	0,56	0,32	0,86	19308
	Deane	70,72%	0,81	0,66	0,54	0,87	12842
	von Mering	57,80%	0,67	0,55	0,34	0,82	4910
	Tuncbag	57,73%	0,71	0,55	0,27	0,89	5214
	Combined - 2 Clusters	62,35%	0,77	0,58	0,35	0,90	10000
	Combined - 4 Clusters	59,97%	0,67	0,57	0,39	0,81	10000
	Average	58,87%	0,69	0,56	0,30	0,87	
Deane	Shen	56,64%	0,63	0,54	0,32	0,81	32336
	Jansen	47,08%	0,70	0,43	0,19	0,89	15697
	Peri	59,93%	0,78	0,56	0,30	0,91	33365
	Benhur - BIND	57,36%	0,60	0,56	0,46	0,69	19308
	Benhur - DIP-MIPS	64,24%	0,63	0,66	0,71	0,57	9581

	von Mering	58,15%	0,60	0,57	0,53	0,63	4910
	Tuncbag	58,36%	0,66	0,56	0,34	0,83	5214
	Combined - 2 Clusters	75,12%	0,80	0,71	0,67	0,84	10000
	Combined - 4 Clusters	64,35%	0,70	0,61	0,51	0,78	10000
	Average	60,14%	0,68	0,58	0,45	0,77	
von Mering	Shen	50,55%	0,68	0,50	0,02	0,99	32336
	Jansen	55,63%	0,65	0,55	0,12	0,94	15697
	Peri	46,86%	0,22	0,48	0,02	0,94	33365
	Benhur - BIND	51,65%	0,62	0,51	0,09	0,95	19308
	Benhur - DIP-MIPS	53,87%	0,64	0,52	0,20	0,89	9601
	Deane	56,54%	0,76	0,54	0,19	0,94	12842
	Tuncbag	51,42%	0,91	0,51	0,03	1,00	5214
	Combined - 2 Clusters	52,54%	0,67	0,51	0,10	0,95	10000
	Combined - 4 Clusters	64,89%	0,88	0,59	0,34	0,95	10000
	Average	53,77%	0,67	0,52	0,12	0,95	
Tuncbag	Shen	57,43%	0,67	0,55	0,28	0,86	32336
	Jansen	49,31%	0,41	0,51	0,18	0,77	15697
	Peri	51,09%	0,54	0,50	0,26	0,77	33365
	Benhur - BIND	51,96%	0,53	0,51	0,31	0,73	19308
	Benhur - DIP-MIPS	52,47%	0,54	0,52	0,36	0,69	9581
	Deane	56,50%	0,61	0,55	0,36	0,77	12842
	von Mering	52,02%	0,55	0,51	0,32	0,73	4910
	Combined - 2 Clusters	56,05%	0,62	0,54	0,32	0,81	10000
	Combined - 4 Clusters	54,07%	0,58	0,53	0,28	0,80	10000
	Average	53,43%	0,56	0,52	0,30	0,77	
Combined - 2 Clusters	Shen	73,75%	0,80	0,69	0,63	0,85	32336
	Jansen	55,11%	0,18	0,88	0,58	0,55	15697
	Peri	66,21%	0,80	0,61	0,45	0,88	33365
	Benhur - BIND	54,09%	0,56	0,53	0,36	0,73	19308

	Benhur - DIP-MIPS	61,12%	0,64	0,59	0,53	0,69	9581
	Deane	80,45%	0,86	0,76	0,73	0,88	12842
	von Mering	53,89%	0,57	0,52	0,39	0,70	4910
	Tuncbag	59,30%	0,65	0,57	0,41	0,78	5214
	Combined - 4 Clusters	60,41%	0,65	0,58	0,45	0,76	10000
	Average	62,70%	0,63	0,64	0,50	0,76	
Combined - 4 Clusters	Shen	59,87%	0,56	0,70	0,86	0,34	32336
	Jansen	56,41%	0,52	0,93	0,98	0,19	15697
	Peri	52,92%	0,52	0,56	0,85	0,19	33365
	Benhur - BIND	57,22%	0,55	0,64	0,82	0,32	19308
	Benhur - DIP-MIPS	56,09%	0,54	0,65	0,87	0,25	9581
	Deane	52,55%	0,52	0,58	0,86	0,19	12842
	von Mering	56,50%	0,54	0,87	0,98	0,13	4910
	Tuncbag	53,26%	0,52	0,72	0,96	0,11	5214
	Combined - 2 Clusters	56,42%	0,54	0,65	0,85	0,28	10000
	Average	55,69%	0,53	0,70	0,89	0,22	
Combined 2 Clusters Pos - 4 Clusters Neg	Shen	56,34%	0,53	0,97	1,00	0,13	32336
	Jansen	49,88%	0,48	0,81	0,98	0,07	15697
	Peri	53,92%	0,52	0,92	0,99	0,07	33365
	Benhur - BIND	52,82%	0,52	0,63	0,92	0,13	19308
	Benhur - DIP-MIPS	54,18%	0,52	0,71	0,95	0,13	9581
	Deane	52,56%	0,51	0,81	0,98	0,07	12842
	von Mering	51,81%	0,51	0,65	0,98	0,04	4910
	Tuncbag	50,44%	0,50	1,00	1,00	0,01	5214
	Average	52,23%	0,51	0,79	0,97	0,07	
Combined 4 Clusters Pos - 2 Clusters Neg	Shen	61,39%	0,86	0,57	0,27	0,96	32336
	Jansen	69,78%	0,84	0,65	0,44	0,92	15697
	Peri	55,82%	0,73	0,53	0,21	0,92	33365
	Benhur - BIND	57,67%	0,65	0,55	0,33	0,83	19308

	Benhur - DIP-MIPS	61,35%	0,67	0,59	0,46	0,76	9581
	Deane	67,45%	0,85	0,62	0,42	0,92	12842
	von Mering	70,53%	0,70	0,71	0,74	0,67	4910
	Tuncbag	58,50%	0,69	0,55	0,30	0,87	5214
	Average	63,01%	0,73	0,60	0,42	0,84	

The results show that the accuracies are around 60% or less for most of the datasets. On the other hand, the positive precision values go up to 90% and most of the times positive precisions are higher than the negative precisions. In the table, the number of false positives is very low where positive negative numbers are very high. This means the negative sensitivity is much better than positive sensitivity. When the average accuracy values are considered, it is seen that the best performance is achieved for *Combined 4 Clusters Pos – 2 Clusters Neg* dataset. In addition, when the models created by similar datasets are applied on each other, the predictions are more accurate. For example, when Benhur - BIND Dataset model is applied on Deane dataset, the obtained accuracy is about 62% and positive precision is 78%. These are the highest values obtained for Benhur - BIND dataset model. Both datasets contains yeast interactions and this might be the reason of high accuracy. In contrast, when Benhur - BIND Dataset model is applied on Peri Dataset, the accuracy is about 45% and positive precision is 23%. These are the lowest values obtained for Benhur - BIND dataset model application. Peri dataset contains human protein interactions where Benhur - BIND Dataset contains yeast and some other organisms. This might be the reason of low accuracy in the test. In conclusion, when the highest and lowest accuracies are taken into consideration, it is easy to say that dataset models work better when they are applied on datasets coming from similar organisms.

Chapter 4

Structure and Sequence Based Prediction of Protein-Protein Interactions

The website PRISM (Protein Interactions by Structural Matching) is used for analysis of protein interfaces and putative protein-protein interactions [7, 56]. This chapter explains the steps of PPI prediction done by PRISM server. First, the structural prediction method is given briefly. Following that, the collision detection method for predicted interactions is explained. Then, the prediction details of noncolliding interactions via NOXclass is given [8]. Finally, the method of creating a model for sequential prediction of interactions and its application on noncolliding interactions is described.

4.1 Prediction of Protein-Protein Interactions

The protein-protein interactions are predicted for a target protein dataset based on the template protein dataset interfaces. First, these two datasets and interfaces are introduced in order to explain the algorithm.

4.1.1 Interfaces and Hotspots

Proteins interact through their interfaces. Interfaces can be defined as the region where two polypeptide chains are linked via non-covalent interactions. Energies are not homogenously distributed along these regions. Some critical residues account for the majority of the binding energy, called “hot spots” [57].

4.1.2 Template Dataset

Template dataset contains a subset of non-redundant unique protein-protein interfaces architectures. These interfaces are found by applying redundancy elimination of all interfaces extracted. Using these interfaces as template, new putative protein interactions are predicted. Currently, 158 interfaces are available in the template dataset.

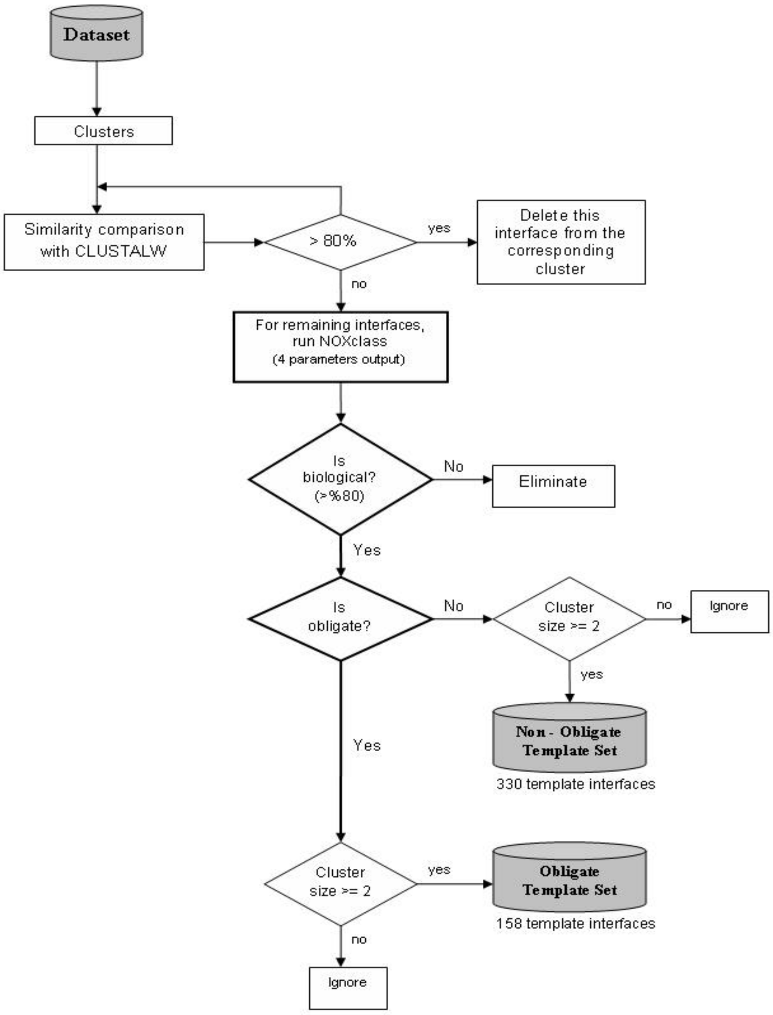


Figure 4.1 Flowchart of Template Dataset Generation

In Figure 4.1 template set construction steps are illustrated. Here, the structurally clustered interface dataset is considered [58] which contains 49512 interfaces clustered into 8205 clusters. These structurally distinct interface clusters contain some homologous complexes. These are next eliminated in each cluster (homology cutoff = 80%, using the ClustalW [59] sequence alignment) following the procedure applied by Keskin et al [60]. Then, crystal contacts are eliminated from each cluster using NOXclass. To construct a non-obligate template set we considered non-obligate interfaces: if the size of a cluster is at least two, the representative interface of that cluster joins the non-obligate set, leading to 158 template interfaces. A similar procedure is followed for the generation of the obligate template set, leading to 330 interfaces which are not used in current PRISM.

4.1.3 Target Dataset

Target dataset is the set of monomers and complexes that will be used in prediction algorithm for structural comparison with the interfaces in template dataset. %50 sequence identity elimination is applied on all proteins in PDB [57]. That results in a set of 10.193 nonredundant proteins. In prediction algorithm, the surfaces of these proteins extracted via NACCESS are used to find the structural similarity with interfaces in template dataset [61]. Residues having a relative surface accessibility greater than 5% are recorded in residues list of surfaces [62].

4.1.4 Structural Prediction Algorithm

The structural prediction algorithm is based on finding the targets similar to the complementary partners of the interfaces. Firstly, the surface of each protein in the target dataset is found. Then, the similarity of those surfaces with the interfaces in the template set is computed. The ones satisfying the residue and hotspot match thresholds (these thresholds are explained in the coming sections) are recorded. In this way, all the targets that are similar to the left or right chains of the interfaces are found. The cross list of the targets similar to the left chain and the targets similar to the right chain construct the structural PPI prediction set. Figure 4.2 shows an outline of the structural prediction algorithm [57]. In the previous algorithm used in PRISM, the target surfaces are eliminated based on only the similarity score. However, in this version a few new constraints are applied on the target proteins for similarity such as residue match and hotspot match percentages.

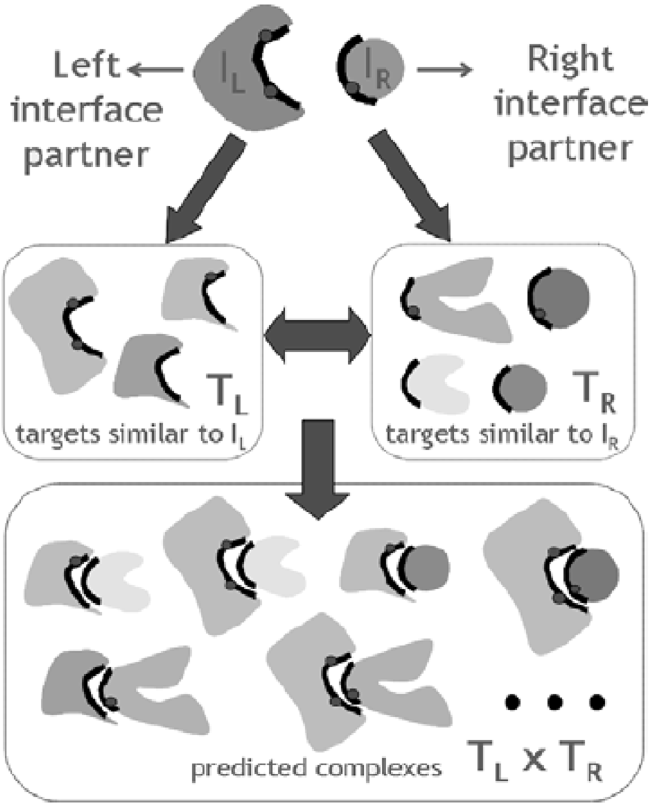


Figure 4.2 Outline of Prediction Algorithm

4.1.4.1 Residue Match Percentage

Target proteins and the interface partners are compared via MULTIPROT (Multiple Protein Structural Alignment Algorithm) software using geometric hashing on the coordinates of atoms of the proteins compared [63]. MULTIPROT gives the 10 best alignments of the proteins independent of protein sequences [57]. In this study, the first alignment is taken that has the largest number of aligned residues. In addition, different from the previous PRISM algorithm, the template interface is taken as the reference molecule in alignment. For a target to be regarded as similar to the interface chain, the similarity ratio should be at least 50%. That means the residues of the target should match with at least half of the interface residues. If the size of the interface is less than 20, targets should match minimum 10 residues of the interface.

4.1.4.2 Complex Target Match

The second constraint for residue matches is the one applied for complex protein structures. Besides the individual chains, the overall protein structure is also considered in the structural matching. But sometimes only one chain of this protein complex might be aligned with the interface. In order to eliminate this redundancy, when protein complexes are aligned with interface, each result of MULTIPROT is analyzed. If the alignment includes only the residues of one chain of the protein complex, it is not taken because it is already recorded. On the other hand, if the aligned complex contains the residues of more than one chain and the residues of those chains include at least 30% of aligned ones, that complex is recorded as a similar target.

Consider the example given in Figure 4.3. In this example, the template interface chain is 1a93A (shown in yellow in the middle) and the protein complex aligned is 1v37. The chains A (shown in blue on the left) and B (shown in red on the right) of 1v37 are aligned to the interface. In total, 19 residues are aligned, 11 residues (58%) from A chain, 8 residues (42%) from B chain. Since the chains fit minimum residue contribution constraint, 1v37 is recorded in the similar target list of the interface 1a93A.

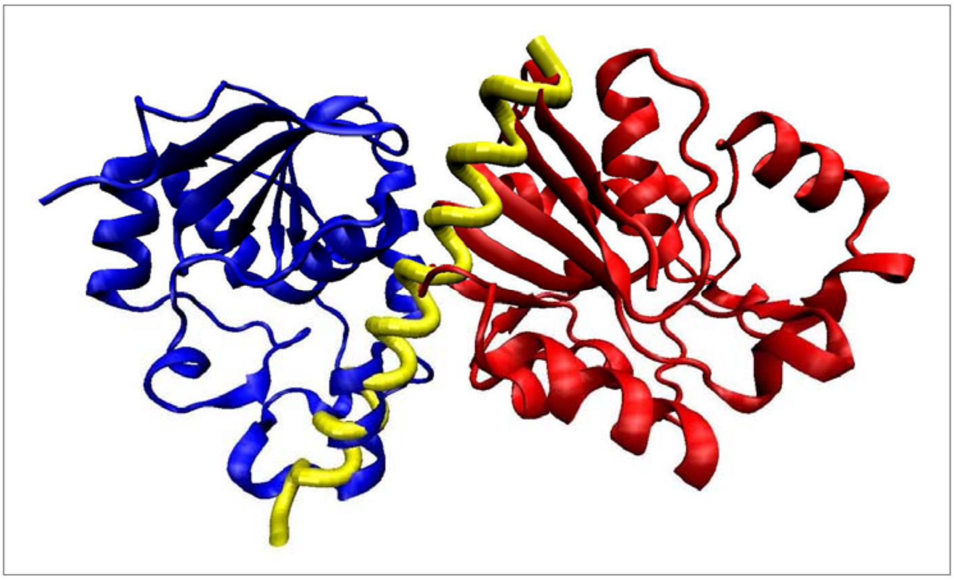


Figure 4.3 Alignment of 1a93A with complex 1v37

4.1.4.3 Hotspot Match

In addition to the residue matches, another constraint is applied on the alignments. The matching target and the interface should have at least one hotspot in common, which helps to eliminate dissimilar proteins. The hotspots in template interfaces are predicted by Hotsprint [ref].

4.2 Collision Detection

The cross list of the targets similar to the left and right chain forms the predicted protein-protein interaction set. But although the targets are similar to the interfaces, they might interpenetrate in each other when transformed on the partner chains of the template interfaces. In order to check this possibility, each target is rotated based on the MULTIPROT alignments, which are recorded while determining similarity. The output of MULTIPROT gives the rotation matrix for the target protein. That matrix is used to rotate the whole protein which is similar to the interface. Following that, all carbon alpha (CA) atoms of the target proteins are extracted. For each PPI prediction, the distances between the CA atoms are computed. If there are more than 5 colliding CA atoms between the rotated left and right targets, those interactions are directly eliminated. In this study, the CA atoms whose coordinates are closer than 3 Å are taken as colliding.

Consider the example given in Figure 4.4. This is an example of collision between the similar targets of 1as4AB interface. The target on the left (shown in blue), 1hleA, is similar to interface chain 1as4A and the target on the right (shown in red), 2ix2C, is similar to interface chain 1as4B. This interaction is not recorded because as shown in the figure, there are many colliding CA atoms.

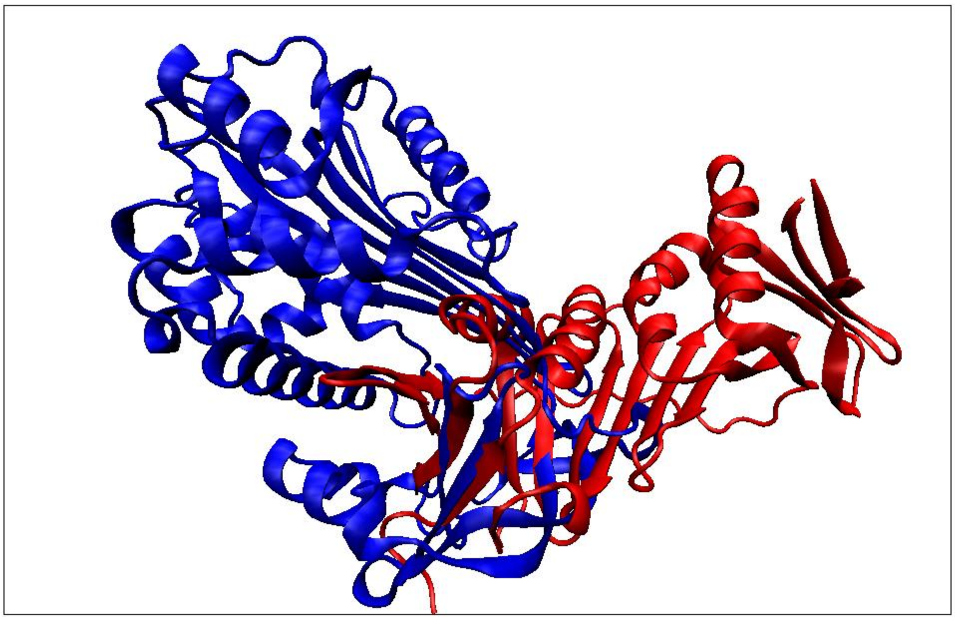


Figure 4.4 Collusion of 1hleA with 2ix2C

4.3 NOXclass Prediction

NOXclass is a support vector machine algorithm based classifier enabling the discrimination of biological and crystal interactions, also, obligate and non-obligate interactions. The classifier is created from a training set composed of nonredundant obligate, non-obligate, and crystal interactions. Six properties of interactions are used in feature vectors: interface area, interface area ratio, amino acid composition, correlation of surface and interface regions, gap volume index, and conservation score of the interface. 91.8% accuracy is achieved for the classification of training set when leave one-out cross-validation procedure is used. As a result, the software distinguishing obligate, non-obligate, and crystal packing interactions is generated [8].

NOXclass is used to make predictions for the interactions that pass the collusion detection step**.** The noncolliding interactions found by rotating the targets and computing the distances of CA atoms are recorded for each interface. NOXclass is applied on those recorded interactions. The inputs of NOXclass are the rotated forms of the target proteins generated. The interactions that are more than 80% biological are recorded in NOXclass prediction set. In addition, the output file of each interface also contains obligate percentage of the interactions.

4.4 Sequence Based Prediction

In this step, PPI predictions will be made for the noncolliding interactions based on the SVM models created using nonredundant interactions. Sandberg Categorization is found to be the best categorization technique for different datasets. There are 5 different z-scales used in SVM algorithm. The false positive and false negative predictions made by SVM models differ based on the z-scale used. Thus, in order to classify the PPI predictions, majority voting method based on 5 z-scale categorizations is applied on noncolliding interactions. In this method, the class of the interaction is found based on the majority of the predictions made by models. For instance, in the case where there are 5 models, an interaction that is classified as positive by 3 or more models is recorded as a positive interaction. In other cases, that interaction is recorded as negative.

In order to test the performance of majority voting, two tests are done on Shen and Tuncbag datasets. The first test is made on Shen dataset using the model generated from random 2.000 interactions. Then, a new random set of 1.000 positive and 1.000 negative interactions are selected from Shen Dataset and they are classified using the models built from 5 z-scale categorizations. Table 4.1 lists the classification results based on those models.

Table 4.1 Classification Results for Shen Test Set

Dataset	Categorization	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	TP	FP	TN	FN	# interactions
Shen	z1	63,40%	0,61	0,67	0,74	0,53	741	473	527	259	2000
	z2	63,45%	0,61	0,66	0,72	0,55	724	455	545	276	2000
	z3	64,05%	0,64	0,64	0,65	0,63	649	368	632	351	2000
	z4	64,80%	0,68	0,63	0,56	0,73	562	266	734	438	2000
	z5	62,80%	0,62	0,64	0,67	0,59	668	412	588	332	2000

The majority voting is done by combining the prediction result of each categorization method for each interaction. Table 4.2 lists the percentages of number of positive votes for the positive interactions and negative votes for the negative interactions. For instance, the first column in Table 4.2 (labeled as 5) for positive interactions means that 31.7% of positive interactions are classified as positive by 5 categorization methods. The last column of positive interaction row (labeled 0) means that 7.5% of positive interactions are not classified as positive by any of the categorization methods.

Table 4.2 Voting of Randomly Selected Shen Dataset Interactions

Dataset	Class \ # of Votes	5	4	3	2	1	0	# interactions
Shen	Positive	31,7%	22,8%	17,0%	12,7%	8,3%	7,5%	1000
	Negative	26,4%	18,2%	18,0%	15,6%	12,6%	9,2%	1000

The same test is also done for Tuncbag Dataset. Random 1.000 positive and 1.000 negative interactions are tested using the model generated by another random 1.000 positive and 1.000 negative interactions. Table 4.3 gives the classification results of each categorization technique and Table 4.4 gives the vote percentages for the interactions.

Table 4.3 Classification Results for Tuncbag Test Set

Dataset	Categorization	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	TP	FP	TN	FN	# interactions
Tuncbag	z1	67,50%	0,74	0,64	0,55	0,80	547	197	803	453	2000
	z2	68,10%	0,74	0,65	0,57	0,80	566	204	796	434	2000
	z3	74,60%	0,80	0,71	0,66	0,83	661	169	831	339	2000
	z4	71,80%	0,73	0,71	0,69	0,74	693	257	743	307	2000
	z5	77,95%	0,75	0,82	0,84	0,72	836	277	723	164	2000

Table 4.4 Voting of Tuncbag Test Set Interactions

Dataset	Class \ # of Votes	5	4	3	2	1	0	# interactions
Tuncbag	Positive	25,0%	30,2%	15,8%	13,3%	10,5%	5,2%	1000
	Negative	37,4%	31,4%	18,9%	8,7%	2,9%	0,7%	1000

The results in Table 4.3 and Table 4.4 show that if the interactions that take more than two votes are classified as positive, 72% of positive interactions in Shen Dataset and 71% of positive interactions in Tuncbag Dataset will be classified correctly. In addition, 63% of negative interactions in Shen Dataset and 88% of negative interactions in Tuncbag Dataset will also be classified correctly. Table 4.5 shows the accuracy and precision values for

Shen and Tuncbag datasets when majority voting method is applied. The overall accuracy of Tuncbag Dataset is 79% and much better than Shen Dataset, which has 67% accuracy.

Table 4.5 Results of Majority Voting on Shen and Tuncbag Test Sets

Dataset	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	TP	FP	TN	FN	# interactions
Shen	67,05%	0,66	0,69	0,72	0,63	715	374	626	285	2000
Tuncbag	79,35%	0,85	0,75	0,71	0,88	710	123	877	290	2000

The accuracy and sensitivity values show that using majority voting on interactions work very well for Tuncbag Dataset. Thus, in order to make sequential predictions on noncolliding interactions, larger training sets for each z-scale are used to generate SVM models. While creating models, all nonredundant positive interactions and same number of randomly selected negative interactions are used. FASTA sequences are recorded while generating feature vectors. The SVM results for the models created for each z-scale is given in Table 4.6.

Table 4.6 SVM Classification Results for each z-scale

Dataset	Categorization	# of Groups	Accuracy	Pos. Prec.	Neg. Prec.	Pos. Sen.	Neg. Sen.	# interactions
Tuncbag	z1	7	76,05%	0,84	0,71	0,64	0,88	6709
	z2	7	76,97%	0,86	0,71	0,64	0,90	6709
	z3	7	82,65%	0,88	0,79	0,76	0,89	6709
	z4	7	83,41%	0,82	0,85	0,85	0,82	6709
	z5	7	84,13%	0,86	0,83	0,82	0,86	6709

All noncolliding interactions from each interface is collected in a set and the input files for each z-scale is created for classification. Then, majority voting method is applied on the set interactions. The classification results for each z-scale are combined in a file. In this file, the interaction and the number of z-scales that classify that interaction as positive is given. The interactions that are classified as positive by more than 2 z-scale models are the ones that pass the majority voting test based on sequence information. In total, there are 1.276.038 interactions and 112.380 of them are classified as positive by 3 or more z-scales. Finally, those 112.380 interactions are deposited as the sequential prediction of PRISM server.

Chapter 5

CONCLUSION

There are several databases that contain experimental protein-protein interactions of various organisms. But the number of experimentally verified interactions is very few compared to total protein-protein interactions in cells. Thus, in order to predict interactions that are not discovered yet, several computational methods are established. One of those methods is to predict PPI by using the sequence information of proteins. PPI prediction using residue sequences can be done in a few ways. The approach used in this study is to calculate the triplet frequencies of residues to create feature vectors of interactions. The generated feature vectors are used for classification via SVM. SVM generates a model from training sets and that model is used to predict interactions.

In order to make accurate prediction of PPI, several datasets containing verified interactions are compared to each other via different characteristics of sequences. For each comparison multiple datasets are tested in order to find the optimum SVM parameters and residue characteristics. Firstly, several n-gram frequencies are tested. The results show that triplet frequencies work the best in most of the cases. Following that, the effect of different types of amino acid categorizations is analyzed. It is found that the best categorization depends on the dataset but generally one of the Sandberg categorizations (z-scales) works the best. To calculate triplet frequencies, there are two ways to extract sequence information. The first one is getting the complete residue sequence of proteins from Swissprot and the second choice is getting sequence records of PDB. It is found that using sequences taken from Swissprot results in better predictions. The reason is determined to be the incomplete sequences of PDB.

Then, it is verified that the SVM performance increases as the dataset size increases. Based on this result, in order to increment the number interactions in small datasets, the turn around of each interaction is also put in training set. It is confirmed that this method has a positive impact on SVM performances.

There are several computational negative set creation methods and these methods are tested via SVM. It is shown that, Jansen method works better than the other methods within the training sets. But Jansen method has the drawback of creating a biased interaction set.

Combination of different n-gram frequency vectors is tested as well. The results reveal that combining different n-grams does not improve the predictions

The last analysis is on optimizing SVM parameters and kernel functions (RBF and γ values). The tests prove that 0.25 is the optimum value for γ parameter when used with RBF.

The resulting sequence based prediction model is integrated into PRISM server. In the previous version, the cross list of the structurally similar targets are listed as the predictions In this study, new criteria are added to eliminate the unlikely predictions. a) The distances between proteins that are transformed according to the interfaces are calculated one by one in order to eliminate predictions in which proteins collide. b) Crystal contacts are eliminated via NOXclass. c) SVM models are further used to filter interactions based on sequential information.

In conclusion, the analysis of datasets show that the classification accuracies are high within the training sets but not across datasets. The analyses show that the prediction accuracies are high when the training set and test set contain interactions originating from similar organisms. In future work, detailed analysis of protein sequences for each dataset might be done in order to find why predictions across datasets have a poor performance. In addition, the residue sequence based intersection of datasets might be used to generate combined datasets that have better prediction accuracies on each interaction set.

BIBLIOGRPHY

1. von Mering, C., et al., *Comparative assessment of large-scale data sets of protein-protein interactions.* Nature, 2002. **417**(6887): p. 399-403.
2. Han, J.D., et al., *Effect of sampling on topology predictions of protein-protein interaction networks.* Nat Biotechnol, 2005. **23**(7): p. 839-44.
3. Shen, J., et al., *Predicting protein-protein interactions based only on sequences information.* Proc Natl Acad Sci U S A, 2007. **104**(11): p. 4337-41.
4. Peto, M., et al., *Use of machine learning algorithms to classify binary protein sequences as highly-designable or poorly-designable.* BMC Bioinformatics, 2008. **9**: p. 487.
5. Alashwal, H., S. Deris, and R.M. Othman, *One-Class Support Vector Machines for Protein-Protein Interactions Prediction.* BIOMEDICAL SCIENCES, 2006. **1**.
6. Ben-Hur, A. and W.S. Noble, *Choosing negative examples for the prediction of protein-protein interactions.* BMC Bioinformatics, 2006. **7 Suppl 1**: p. S2.
7. Ogmen, U., et al., *PRISM: protein interactions by structural matching.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W331-6.
8. Zhu, H., et al., *NOXclass: prediction of protein-protein interaction types.* BMC Bioinformatics, 2006. **7**: p. 27.
9. Finn, R.D., et al., *Pfam: clans, web tools and services.* Nucleic Acids Res, 2006. **34**(Database issue): p. D247-51.
10. Qi, Y., Z. Bar-Joseph, and J. Klein-Seetharaman, *Evaluation of different biological data and computational classification methods for use in protein interaction prediction.* Proteins, 2006. **63**(3): p. 490-500.
11. Mahdavi, M.A. and Y.H. Lin, *False positive reduction in protein-protein interaction predictions using gene ontology annotations.* BMC Bioinformatics, 2007. **8**: p. 262.
12. Pitre, S., et al., *PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs.* BMC Bioinformatics, 2006. **7**: p. 365.
13. Wodak, S.J. and R. Mendez, *Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications.* Curr Opin Struct Biol, 2004. **14**(2): p. 242-9.
14. Bock, J.R. and D.A. Gough, *Predicting protein--protein interactions from primary structure.* Bioinformatics, 2001. **17**(5): p. 455-60.
15. Yan, C., D. Dobbs, and V. Honavar, *Identification of Surface Residues Involved in Protein–Protein Interaction - A Support Vector Machine Approach*, in *Intelligent Systems Design and Applications*. 2003: Berlin. p. 53-62.

16. Gallet, X., et al., *A fast method to predict protein interaction sites from sequences.* J Mol Biol, 2000. **302**(4): p. 917-26.
17. Yan, C., D. Dobbs, and V. Honavar, *A two-stage classifier for identification of protein-protein interface residues.* Bioinformatics, 2004. **20 Suppl 1**: p. i371-8.
18. Ofran, Y. and B. Rost, *Predicted protein-protein interaction sites from local sequence information.* FEBS Lett, 2003. **544**(1-3): p. 236-9.
19. Janin, J., et al., *CAPRI: a Critical Assessment of PRedicted Interactions.* Proteins, 2003. **52**(1): p. 2-9.
20. Lu, L., H. Lu, and J. Skolnick, *MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading.* Proteins, 2002. **49**(3): p. 350-64.
21. Marcotte, E.M., et al., *Detecting protein function and protein-protein interactions from genome sequences.* Science, 1999. **285**(5428): p. 751-3.
22. Bradford, J.R. and D.R. Westhead, *Improved prediction of protein-protein binding sites using a support vector machines approach.* Bioinformatics, 2005. **21**(8): p. 1487-94.
23. Qi, Y., J. Klein-Seetharaman, and Z. Bar-Joseph, *A mixture of feature experts approach for protein-protein interaction prediction.* BMC Bioinformatics, 2007. **8 Suppl 10**: p. S6.
24. Espadaler, J., et al., *Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships.* Bioinformatics, 2005. **21**(16): p. 3360-8.
25. Sandberg, M., et al., *New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids.* J Med Chem, 1998. **41**(14): p. 2481-91.
26. Murphy, L.R., A. Wallqvist, and R.M. Levy, *Simplified amino acid alphabets for protein fold recognition and implications for folding.* Protein Eng, 2000. **13**(3): p. 149-52.
27. Witten, I.H., et al., *Weka: Practical Machine Learning Tools and Techniques with Java Implementations*.
28. Cristianini, N. and J. Shawe-Taylor, *An Introduction to Support Vector Machines.* 2000.
29. Hsu, C. and C. Lin, *A Comparison of Methods for Multiclass Support Vector Machines.* IEEE, 2002.
30. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine Learning, 1995(Volume 20, Number 3 / September, 1995): p. 273-297.
31. Vapnik, V.N., *Estimation ofDependences Based on Empirical Data*. 1982, New York: Springer-Verlag.

32. Chih-Wei Hsu, C.-C.C., and Chih-Jen Lin, *A Practical Guide to Support Vector Classification*. 2007.

33. Chapelle, O., et al., *Choosing Multiple Parameters for Support Vector Machines.* 2004: p. 131-159.

34. Das, R., et al., *Computational prediction of methylation status in human genomic sequences.* Proc Natl Acad Sci U S A, 2006. **103**(28): p. 10713-6.

35. HPRD. *Human Proteins Reference Database*. [cited; Available from: http://www.hprd.org/.

36. Peri, S., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.

37. DIP. *The Database of Interacting Proteins*. [cited; Available from: http://dip.doe-mbi.ucla.edu.

38. Xenarios, I., et al., *DIP: the database of interacting proteins.* Nucleic Acids Res, 2000. **28**(1): p. 289-91.

39. Mewes, H.W., et al., *MIPS: a database for genomes and protein sequences.* Nucleic Acids Res, 2000. **28**(1): p. 37-40.

40. MIPS. *Munich Information Center for Protein Sequences*. [cited; Available from: http://mips.gsf.de.

41. BIND. *Biomolecular Interaction Network Database*. [cited; Available from: http://binddb.org.

42. Bader, G.D., et al., *BIND--The Biomolecular Interaction Network Database.* Nucleic Acids Res, 2001. **29**(1): p. 242-5.

43. YPD. *Yeast Proteome Database*. [cited; Available from: http://www.proteome.com.

44. Costanzo, M.C., et al., *YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information.* Nucleic Acids Res, 2001. **29**(1): p. 75-9.

45. PDB. *Protein Data Bank*. [cited; Available from: http://www.pdb.org/pdb/home/home.do.

46. Tuncbag, N., et al., *Architectures and Functional Coverage of Protein-Protein Interfaces.* JMB, 2008.

47. Jansen, R. and M. Gerstein, *Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction.* Curr Opin Microbiol, 2004. **7**(5): p. 535-45.

48. Ben-Hur, A. and W.S. Noble, *Kernel methods for predicting protein-protein interactions.* Bioinformatics, 2005. **21 Suppl 1**: p. i38-46.

49. Coward, E., *Shufflet: shuffling sequences while conserving the k-let counts.* Bioinformatics, 1999. **15**(12): p. 1058-9.

50. Mewes, H.W., et al., *MIPS: a database for genomes and protein sequences.* Nucleic Acids Res, 2002. **30**(1): p. 31-4.

51. Jansen, R., et al., *A Bayesian networks approach for predicting protein-protein interactions from genomic data.* Science, 2003. **302**(5644): p. 449-53.

52. Bader, G.D., D. Betel, and C.W. Hogue, *BIND: the Biomolecular Interaction Network Database.* Nucleic Acids Res, 2003. **31**(1): p. 248-50.

53. Deane, C.M., et al., *Protein interactions: two methods for assessment of the reliability of high throughput observations.* Mol Cell Proteomics, 2002. **1**(5): p. 349-56.

54. Expasy. *Expert Protein Analysis System*. [cited; Available from: http://expasy.org/.

55. Gasteiger, E., et al., *ExPASy: The proteomics server for in-depth protein knowledge and analysis.* Nucleic Acids Res, 2003. **31**(13): p. 3784-8.

56. Ogmen, U. *PRISM*. 2005 [cited; Available from: http://prism.ccbb.ku.edu.tr/prism/.

57. Aytuna, A.S., A. Gursoy, and O. Keskin, *Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces.* Bioinformatics, 2005. **21**(12): p. 2850-5.

58. Tuncbag, N., et al., *Architectures and functional coverage of protein-protein interfaces.* J Mol Biol, 2008. **381**(3): p. 785-802.

59. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.

60. Keskin, O., et al., *A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications.* Protein Sci, 2004. **13**(4): p. 1043-55.

61. Hubbard, S.J. and J.M. Thornton, *NACCESS Computer Program*. 1993. p. Department of Biochemistry and Molecular Biology, University College London.

62. Jones, S. and J. Thornton, *Analysis of protein-protein interactions using surface patches.* J. Mol. Biol., 1997(272): p. 121-132.

63. Shatsky, M., R. Nussinov, and H.J. Wolfson, *A method for simultaneous alignment of multiple protein structures.* Proteins, 2004. **56**(1): p. 143-56.

VITA

Mehmet Cengiz ULUBAŞ was born in Erzincan, Turkey on 8 December, 1983. He had attended to Erzincan Anadolu Lisesi for secondary and high school education. He received his B.Sc. Double Major Degree in Computer Engineering and Electrical & Electronics Engineering from Koç University in 2006. From September 2006 to June 2008 he has worked as teaching and research assistant at Koç University, Istanbul, Turkey.

He currently lives in İstanbul and works at Vodafone Technology as Software Engineer.