PDZ Domains: Interaction Prediction, Classification and Peptide Library Construction

by

Sibel Kalyoncu

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

Master of Science

in

Chemical and Biological Engineering

Koç University

November 2010

Koç University
Graduate School of Sciences and Engineering

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

Sibel Kalyoncu

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:

Özlem Keskin, Ph. D. (Advisor)

Attila Gürsoy, Ph. D. (Advisor)

I. Halil Kavaklı, Ph. D.

Ceyda Oğuz, Ph. D.

Serdar Kozat, Ph. D.

Date:

To my family and my fiancé

ABSTRACT

PDZ domain is a well-conserved, structural protein-protein interaction domain found in hundreds of signaling proteins that are otherwise unrelated. PDZ domains can bind to the C-terminal peptides of different proteins and they cluster different protein complexes together, target specific proteins and route these proteins in many signaling pathways. PDZ domains are classified into Class I, II and III, depending on their binding partners and the nature of bonds formed. Binding specificities of PDZ domains are very crucial in order to understand the complexity of signaling pathways. It is still an open question how these domains recognize and bind their partners.

The focus of this thesis is three folds: 1) predicting to which peptides a PDZ domain will bind, 2) classification of PDZ domains as Class I, II or I-II and 3) construction of peptide libraries for PDZ domains using genetic algorithm. For the first two parts, trigram and bigram amino acid frequencies are used as features in machine learning methods. Using 85 PDZ domains and 181 peptides, our model reaches high prediction accuracy (91.4%) for binary interaction prediction which outperforms previously investigated similar methods. Also, we can predict classes of PDZ domains with an accuracy of 90.7%. We propose three critical amino acid sequence motifs that could have important roles on specificity pattern of PDZ domains. For the last part, we implemented genetic algorithm to generate possible binding peptides for PDZ domains by using the sequence of experimentally verified binding peptides of PDZ domains. Then, the performance of this generated peptide library is evaluated by PDZ interaction prediction model constructed in the first part.

ÖZETÇE

PDZ yapısal bölgeleri, birbirinden farklı birçok sinyal iletim proteinlerinde bulunan, iyi korunmuş yapısal protein etkileşim bölgeleridir. PDZ yapısal bölgeleri proteinlerin karboksil ucuna bağlanarak, farklı protein komplekslerini bir araya getirir, belli proteinleri hedef alır ve bu proteinleri sinyal iletim yollarına yönlendirir. PDZ yapısal bölgeleri, bağlandığı hedef peptitlere ve oluşturduğu bağların niteliğine göre Sınıf I, II, III olmak üzere üç sınıfa ayrılır. PDZ yapısal bölgelerinin bağlanma özgünlüğü, sinyal iletimlerinin karmaşıklığını anlamak adına çok önemlidir. Bu yapısal bölgelerin, hedeflerini nasıl tanıdığı ve hedeflerine nasıl bağlandığı hala açık bir sorudur.

Bu tez, üç odak noktasından oluşmaktadır: 1) PDZ yapısal bölgelerinin hangi peptitlere bağlanabileceğini tahmin etmek, 2) PDZ yapısal bölgelerini Sınıf I, II, I-II olarak sınıflandırabilmek, 3) genetik algoritma kullanılarak PDZ yapısal bölgeleri için peptit veri tabanı oluşturmak. İlk iki kısım için, trigram ve bigram amino asit frekansları hesaplanarak, bunlar oluşturulan otomatik öğrenme metodunda özellik olarak kullanılmıştır. 85 PDZ yapısal bölgesi ve 181 peptit kullanılarak, modelimiz ikili etkileşim tahmininde yüzde 91.4 doğruluğa ulaşarak benzer diğer metotlarından daha üstün olmuştur. Aynı zamanda, bu metotla PDZ yapısal bölgelerinin sınıfları yüzde 90.7 doğrulukla tahmin edilmiştir. Ve PDZ yapısal bölgelerinin özgünlüğünde önemli roller üstlenebilecek üç kiritk amino asit sekans motifi önerilmiştir. Son kısım için, genetik algoritma uygulamasıyla, PDZ yapısal bölgelerine bağlandığı deneysel olarak kanıtlanmış peptitlerin sekansları kullanılarak, PDZ yapısal bölgelerine bağlanabilecek olası peptitler oluşturulmuştur. Daha sonra, bu oluşturulmuş peptit veri tabanlarının performansları, ilk kısımda oluşturulan PDZ etkileşimi tahmin modeli ile test edilmiştir.

ACKNOWLEDGEMENTS

I would like to express my gratitude and thanks to my advisors, Dr. Ozlem Keskin, Dr. Attila Gursoy for their guidance and support through my master education. I would like to thank Dr. I. Halil Kavakli, Dr. Burak Erman, Dr. Ceyda Oguz, Dr. Serdar Kozat, Dr. Banu Ozkan and Dr. Can Erkey for their generous guidance and kindness.

I would like to thank my friends in our research group, COSBI (Computational Systems Biology), Dr. Nurcan Tuncbag, Gozde Kar, Billur Engin, S. Ece Ozbabacan, Guray Kuzu, Serap Beldar, Engin Cukuroglu, Beytullah Ozgur, Omer An, Emine Guven, Halil Peynirci and Bora Karasulu for their supports and friendships. I would like to thank Dr. Ibrahim Baris for his invaluable help and guidance through my days in Molecular Biology and Biochemistry Laboratory.

I am grateful to all my friends in Koc University, Bogazici University, Istanbul Technical University, to name some, Deniz Sanli, Onur Oztas, Izzet Yildiz, Evren Fatih Arkan, Seyda Ipek, Bengisu Seferoglu, Seda Giray, Nil Ezgi Dincer, Selmi Bozbag, Bilal Cakir, Enis Demir, Zeynep Ulker, R. Oguz Caniaz, Selimcan Azizoglu, Onur Dagliyan, Mehmet Ali Ozturk, Ceren Tuzmen, Ayse Kucukyilmaz, Mert Gur, Hande Asimgil, Mehmet Tardu, Isil Tulum, M. Salih Kilic, Musa Ozboyaci, Ozlem Ulucan, Tugba Arzu Ozal and Ozdem Sezgin.

Finally and most importantly, I would like to express my gratitude to my family, my father Aydin Kalyoncu, my mother Nurhan Kalyoncu, my brothers Tarik Kalyoncu, Tayfun Kalyoncu, and my brother's wife Merve Kalyoncu for their invaluable support and love. I would specially thank my only love, my fiancé, Erdal Uzunlar for his never-ending help, support and love.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1

1 Introduction

1.1 Literature Review

1.1.1 Protein-Protein Interaction Domains

An increasing body of data suggests that proteins involved in many cellular mechanisms are regulated in a modular manner that a protein could contain functionally or structurally independent regions (domains). The networks and pathways that connect receptors to their targets usually involve a series of protein-protein interactions. Many different cellular mechanisms are regulated by protein interaction domains [1]. They organize the association of proteins with one another, small molecules, nucleic acids or phospholipids. Protein interaction domains can route other proteins to specific cellular locations, form signaling multi-complex proteins, secure recognition of post-translational modifications, control activity, formation and specificity of enzymes [2]. Therefore, protein interaction networks are heavily investigated due to their potential applications in drug discovery. They can give key insights about the mechanisms of human diseases.

Protein-protein interactions play fundamental roles in signal transduction, formation of functional protein complexes and protein modification [3]. Many biological processes are regulated through the dynamic interactions of modular protein domains (e.g., WW, SH3, SH2, PH, and PDZ) and their corresponding binding targets. Investigation of the selectivity, specificity, and regulatory mechanisms involved in these protein-protein interactions can therefore provide important insights into biological activities.

Protein-protein interaction domains can usually be expressed independently from their main proteins, namely they can provide their intrinsic function of binding to their targets without the host protein. Binding pocket and N- and C- terminus of these domains are usually on the opposite face [2]. This structural arrangement let these domains insert easily into the host protein without blocking their binding pockets.

Protein interaction domains can form hetero/homo-typic domain interactions and they can also bind to short peptide motifs or small molecules. There is not a clear distinction between these types of interactions. For example, PDZ domains generally bind to short peptide motifs at the C termini of their target proteins, but they can also form PDZ-PDZ domain interactions [4-6].

1.1.2 PDZ Domains

One of the most common protein interaction domains in the cell is PDZ domain which is a central signaling protein of most species [68]. The PDZ domains, among other nearly 70 distinct recognition domains, are crucial because they are involved in development of multi-cellular organisms by constructing cell polarity, coordination of intercellular signaling system and directing the specificity of signaling proteins [9]. They consist of 80 to 90 amino acids and have a compact globular module composed of a core of six β strands (βA - βF) and two α helices (αA, αB). By binding the C-terminal motifs of their target proteins, PDZ domains target, cluster and route these proteins [10]. However, some PDZ domains also can bind to the internal motifs of target proteins, lipids and other PDZ domains [6,11].

C-terminus of the peptides recognizes and binds to a pocket between carboxylate-binding loop (βA – βB loop) that contains the conserved GLGF motif, and αB helix of the PDZ domain [12-15], this is also called the canonical binding. The ligand binds to the PDZ domain as an anti-parallel extension of the β-sheet of the domain and while ligand positions -1 and -3 head towards to the solvent, the positions 0 and -2 point towards to the binding pocket [16] (Figure 1.1). Therefore, it can be suggested that ligand positions 0 and -2 are very crucial for recognition and binding to target proteins. The importance of these two positions also lead to the general classification of PDZ domains into three classes according to short peptide motifs of the last three residues at the extreme C-termini of their peptide ligands. Class I PDZ domains bind to C terminal motifs with the sequence of [Ser/Thr-X-Φ COOH], Class II PDZs bind to the sequence of [Φ-X-Φ-COOH] and Class III PDZs prefer the sequence of [Asp/Glu-X-Φ-COOH] where Φ is any hydrophobic amino acid and X is any amino acid. However, some PDZ domain interactions do not satisfy these restrictive types of recognition and so additional classes and additional important residues are proposed to exist for ligand specificity of PDZ domains [16-18]. For example, Songyang *et al.* investigated the binding specificities of nine PDZ domains by using an oriented peptide library and concluded that additional selection specificities, depending on up to -8 position of the peptide ligand, were observed beside the 0 and -2 positions [19].

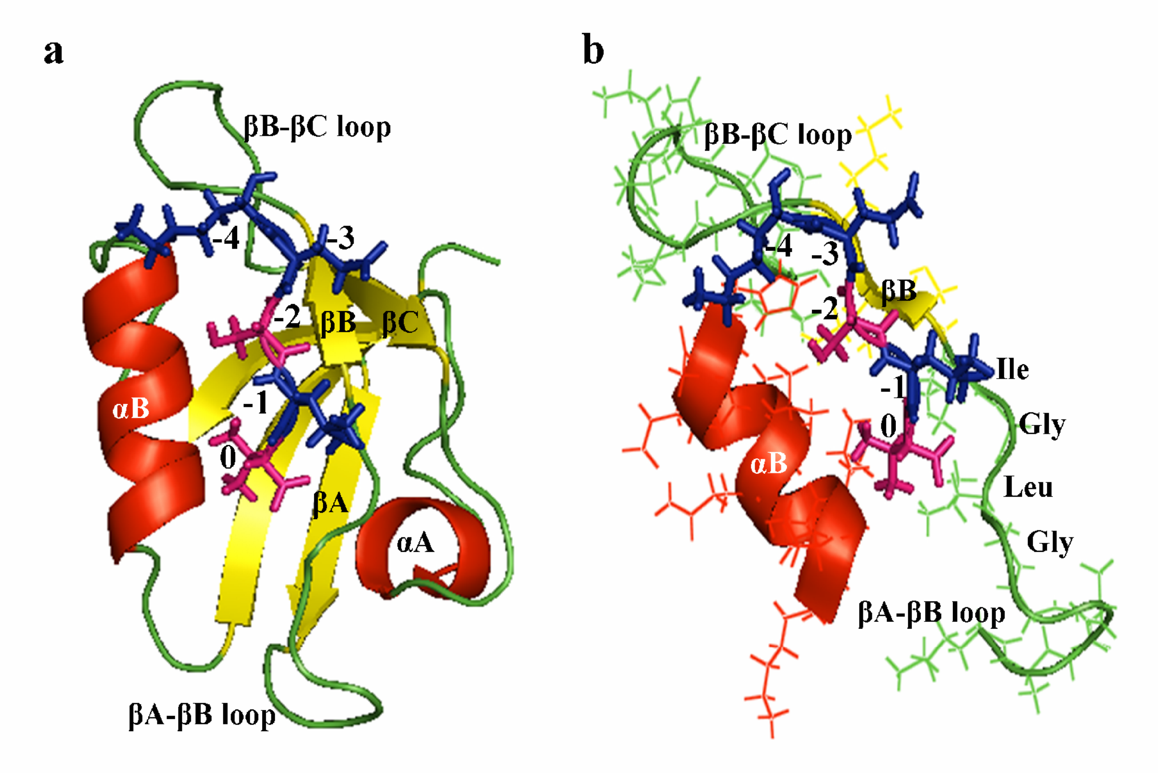


Figure 1.1: Representative structure of a PDZ domain in complex with its ligand (a) The common representation of a PDZ domain (α-1 syntrophin) with a peptide (in its stick form) in its binding pocket. Peptide positions -1 and -3 (blue) point towards to the solvent, the positions 0 and -2 (pink) head towards to the binding pocket (b) The interaction of the peptide with αB helix and conserved GLGF segment (here it is GLGI) of the βA-βB loop (PDB ID:2PDZ).

Although PDZ domains show selectivity toward their target ligands, they also display promiscuity, binding to more than one ligand, and degenerate specificity [2023], so interaction prediction of these domains can be challenging. Several studies aimed to classify and predict interaction specificity of PDZ domains that could save time-consuming and expensive experiments. Chen *et al.* predicted PDZ domain-peptide interactions from primary sequences of PDZ domains and peptides by using a statistical model and reported an area under curve (AUC) value of 0.87 for extrapolations to both novel mouse peptides and PDZ domains [24]. Bezprozvanny and Maximov used a classification method based on the two critical positions of 249 PDZ domains and they presented 25 different classes of PDZ domains [17]. Stiffler *et al.* also tried to characterize the binding selectivity of PDZ domains by training multi-domain selectivity model for 157 mouse PDZ domains with respect to 217 peptides and they indicated that PDZ domains are distributed throughout the selectivity space contrary to discrete specificity classes [25]. Schillinger *et al.* used a new approach, Domain Interaction Footprint (DIF), to predict binding peptides of SH3 and PDZ domains by using only the sequence of the peptides, they reported an AUC value of 0.89 for PDZ multi-domain model by using the sequence information of binding and non-binding peptides of four different PDZ domains [26]. Tonikian *et al.* constructed a specificity map consisting of 16 unique specificity classes for 72 PDZ domains and this lead to the prediction of PDZ domain interactions [27]. Wiedemann *et al.* tried to quantify specificity of three PDZ domains by relating the last four C-terminal motifs of their ligands to the corresponding dissociation constants which can provide selectivity pattern of PDZ domains and design of super-binding peptides [23]. Eo *et al.* used an SVM classifier by adapting amino acid contact matrices and physiochemical distance matrix as a feature encoding in order to identify PDZ domain ligand interactions [28].

1.1.3 Roles of PDZ Domains in Diseases

Some members of PDZ domain family play considerable roles in neurological diseases. They interact with pre-synaptic and post-synaptic proteins and they have crucial roles on synaptic neurotransmission and plasticity [29]. It is shown that PICK1, one of the PDZ domains, interacts with Glutamate (Glu) receptor family [30]. These Glutamate receptors have roles in excitatory neurotransmission and synaptic functions and it is shown that they are related to some neurological diseases such as stroke, neurodegeneration, depression, anxiety, epilepsy and schizophrenia. Also, other PDZ domains such as GRIP, ABP and PSD95 are observed to have interactions with Glu receptors. In addition, PICK1 interacts with monoamine plasma membrane transporters (dopamine, neuroepinephrine, serotonin) [31]. Any destruction to monoamine neurotransmission could cause neurodegenerative diseases (depression, lack of attention, hyperactivity, schizophrenia).

PICK1 also interacts with three proteins which have roles in cancer generation and cell growth. These proteins are Neurolignin/ErbB receptors (ErbB/Rs-breast, lung and liver cancer) [32], tetradecanoyl phorbol ester-induced main receptor sequences (TIS21-cell growth inhibition) [33] and Coxsackie and adenovirus receptors (CAR-its over-expression decreases cell growth speed) [34]. PICK1 probably affects PKC-phosphorylation states and/or surface expressions and distribution of these proteins.

1.1.4 Methods Overview

Two methods are used in this study. For the first part, interactions and classes of PDZ domains are predicted through a machine learning approach. Then, a new method is developed by using genetic algorithm to generate peptide libraries specific to PDZ domains.

1.1.4.1 Machine Learning and Interaction Prediction

Machine learning methods are used for pattern recognition tasks where data is massive and a set of rules can not discriminate the patterns. The main idea behind these methods is to learn to discriminate experimentally verified data and obtain learned complex rules to predict probable solutions. Although there are so many different machine learning algorithms, they are all driven by the data used to train them [35].

There are several machine learning approaches to predict domain interactions [36-38]. We chose five classifiers, SVM (Support Vector Machine), Nearest Neighbor, Naïve Bayes, J48 and Random Forest which have been commonly used in protein-protein interaction prediction problems. In SVM algorithm, feature vectors are non-linearly mapped on a high dimensional feature space and a set of hyperplanes are constructed to be used for classification or regression [39]. The simplest one among used classifiers is Nearest Neighbor which classifies instances according to their closeness to the training examples [40]. The basic idea behind Naïve Bayes is to predict the class of an instance by learning conditional probability of each attribute [41]. J48, also known as C4.5 grows an initial tree by using divide-and-conquer algorithm and then rank test instances [42]. Random Forest developed by Breiman generates many classification trees simultaneously where each node uses a random subset of the features and outputs the classification based on majority voting over all trees in the forest [43].

1.1.4.2 Peptide Library Construction

There are many biochemical and structural studies trying to develop small molecules to regulate protein-protein interactions. These studies are generally based on the modification of existing binding peptides [44,45] or random peptide sequence design [46-50]. Because many of these studies are experimental, they usually deal with small amount of proteins and ligands. The most common problem in these studies is uncharacterized binding or regulatory regions of corresponding proteins. PDZ domains are very advantageous because binding regions of most of the PDZ domains are known in detail [51]. The development of new methods to regulate PDZ domains is very crucial due to the important functions of these domains in cancer and other diseases. Therefore, construction of PDZ domain specific peptide database is one of the aims of this study. These constructed peptide database could efficiently contribute to novel drug design studies.

1.1.4.2.1 PDZ Domain Peptide Libraries

Peptide library approaches are used for PDZ domain-peptide interactions because PDZ domains recognize short linear motifs (C-termini) of their target proteins. There are two commonly used experimental peptide library approaches: phage display and SPOT synthesis [52].

Phage display is a high-throughput screening of protein-protein interactions. Protein of interest is expressed on phage surface to be exposed to short randomized peptide sequences and if any peptide interact with the protein, it can be analyzed after washing and elution [53]. SPOT synthesis allows the parallel synthesis and screening of thousands of cellulose membrane-bound peptides. These approaches have been applied to study PDZ-mediated interactions [44,54,55].

Studies with PDZ peptide libraries and microarrays is usually conducted in order to have more information about PDZ-peptide interactions, these studies could lead to generation of a key resource to investigate signaling pathways within cells [25,56]. This information needs to be comprehensively deposited in publicly available repositories, such as iSPOT, DOMINO, and PDZBase [57-60] in order to maximally accelerate the discovery of novel PDZ-mediated interactions in cells. PDZBase is a unique database that contains information extracted from the literature of all known PDZ domain-mediated protein-protein interactions obtained from in vivo or in vitro experiments [60].

1.1.4.2.2 Genetic Algorithm

Genetic algorithm is a population based method which gives a set of solutions instead of one solution. It applies a random search with controlled selection. The techniques used in genetic algorithm are generally originated from mechanisms of evolutionary biology (mutation, cross-over, selection) [61].

Genetic algorithm generally consists of four main steps: 1) initialization, 2) selection, 3) reproduction and 4) termination. In the first step, the problem is defined as a genetic representation and an initial population with size N is generated. In the second step, fitness value of each individual is calculated through an objective function and parents of next populations are generated according to their fitness values. In the reproduction step, selected parents are paired and genetic operators, crossover and mutation, are applied to form a new population. The whole process is repeated until the last step, when the termination criterion is met, the algorithm gives the best solution [62]. A representative scheme of genetic algorithm is shown in Figure 1.2.

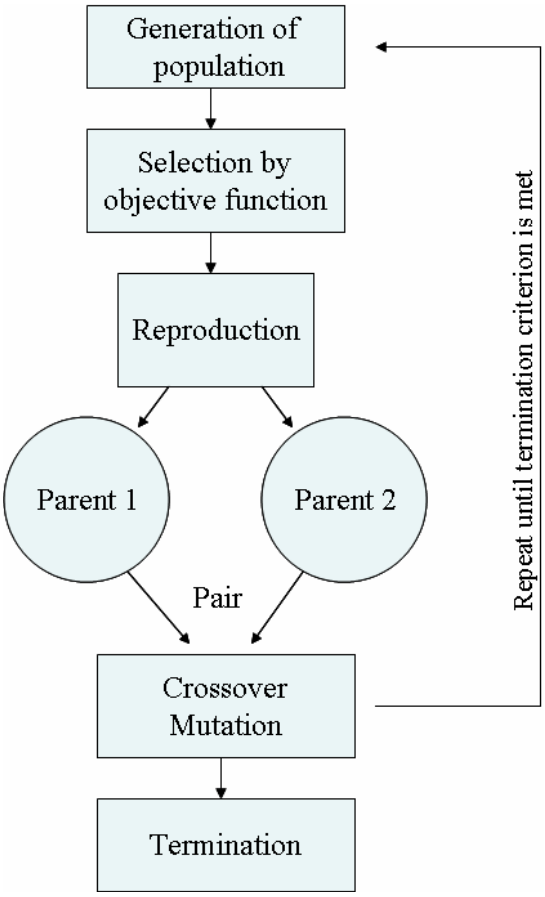


Figure 1.2: A representative scheme of the genetic algorithm.

1.2 Statement of the Problem

As stated in the literature review section, there are some experimental and computational studies investigating the interactions of PDZ domains which play important roles in human disease pathways. In this study, we try to figure out interaction specificity and different classes of PDZ domains by means of machine learning methods and we also try to generate peptide database which consists of potential binding partners for PDZ domains by using genetic algorithm.

We propose a method to predict domain-peptide interactions and classes of PDZ domains by using only the sequence information of PDZ domains and their experimentally verified binding/nonbinding ligands. In order to construct a numerical feature vector for each interaction, trigram and bigram frequencies of each primary sequence of PDZ domains and peptides are calculated. We make use of the most commonly used classifiers (SVM, Nearest Neighbor, Naïve Bayes, J48, Random Forest). Moreover, we show that our method can be efficiently used to distinguish between Class I, Class II and Class I-II PDZ domains. After completion of models, we try to find some critical amino acid motifs on PDZ domains which contribute their specificity more than other amino acid sequences by reducing the dimensions of features. Last, we implemented genetic algorithm to construct possible binding peptides for PDZ domains. Generated peptide libraries are trained in our interaction prediction model in order to see their performance. These models can be used to reduce the search space of experimental studies by giving the most probable candidate binding partners of PDZ domains.

1.3 Outline

This thesis is composed of five chapters:

In Chapter 2, construction of PDZ interaction and class prediction models is explained with detail. Structure of datasets, feature encoding method, machine learning methods and dimensionality reduction process are explained.

In Chapter 3, the genetic algorithm and its implementation to our problem are presented. The generation of initial population by constructing amino acid probabilities of binding peptides is described in detail.

In Chapter 4, results obtained from PDZ interaction and class prediction models are presented. The selection of used classifier for both models is explained by showing performance comparison of five different classifiers. Accuracy and AUC (Area under curve) values of interaction prediction and classification models is shown for both trigram/bigram frequency encodings. Then, critical sequence motifs for PDZ domains found after feature reduction are presented by demonstrating them on the three dimensional structures of PDZ domains. Lastly, the results of parameter tuning and method selection for genetic algorithm is explained and obtained peptide library for a PDZ domain, α1-syntrophin, is demonstrated.

In Chapter 5, the findings in this study are summarized, and the future work with suggestions is presented.

Chapter 2

2 Prediction Methods

2.1 Datasets Used in Prediction Methods

The aim of this study is two-fold. First, we try to predict interactions and classes of PDZ domains. Second, we try to construct peptide libraries for PDZ domains. For both parts, we need experimental data in order to give the input for our algorithm and construct our predictive model. The structure of our dataset is explained in the following sections and datasets are given in Appendix A.1.

2.1.1 PDZ Domain Interaction Dataset

For interaction prediction part, a positive (binding) and a negative (non-binding) dataset are needed in our machine learning model. The PDZ interaction dataset is retrieved from the study of Stiffler et al., which is composed of interaction data of 85 mouse PDZ domains with respect to 217 mouse genome-encoded peptides [25,26]. They used the combination of protein microarrays and fluorescence polarization (FP) methods to identify biological interactions of PDZ domains. In the current study, only binding and non-binding information that were confirmed by FP is used as the training set due to the fidelity of FP. After selection of FP confirmed interactions, we obtained 731 binding and 1361 non-binding interactions between 85 PDZ domains and 181 peptides.

2.1.2 PDZ Domain Classification Dataset

For class prediction part, 86 PDZ domains are categorized, resulting in 45 Class I, 20 Class II and 21 Class I-II PDZ domains. These are retrieved from our interaction dataset and PDZBase [60] by looking at their interactions with different classes of peptides. PDZ domains are annotated as Class I and Class II according to the C terminus sequence of the interacting peptides, [Ser/Thr-X-Φ-COOH] for Class I peptides and [Φ-X-Φ-COOH] for Class II peptides, respectively. Class I-II PDZ domains are determined if they bind to both Class I and Class II peptides. (See Appendix A.1, Table A.1.1)

2.1.3 Validation Dataset

An independent validation dataset is also used in interaction prediction part in order to test the predictive performance of our model. The validation dataset is extracted from the previous study of Stiffler et al. and it is composed of 27 binding and 62 non-binding interactions of 16 PDZ domains and 20 peptides [56].

In order to be consistent in our prediction model, we took the last 10 residues of each peptide sequence due to the selection specificities of PDZ domains up to -10 positions of peptides. The sequence data of PDZ domains and peptides can be seen in Table A.1.2, and Table A.1.3 respectively (Appendix A.1).

2.2 Feature Encoding

2.2.1 Bigram/Trigram Frequency Model

Frequencies of consecutive three amino acids (trigram) and two amino acids (bigram) in the primary sequences are used as features. For instance, a sequence of “ABCDE” results in trigram set of “ABC”, “BCD”, “CDE” and bigram set of “AB”, “BC”, ”CD” and “DE”. In order to reduce the dimension of the features, 20 amino acids are clustered into 7 different classes (Table 2.1) according to their dipoles and volumes of the side chains which reflect their interaction specificity by giving an insight about their electrostatic and hydrophobic natures [63].

Table 2.1: Seven amino acid classes used in our model.

Class	Amino acid(s)	Volume (Å^3)	Dipole (Debye)
1	Ala, Gly,Val	<50	0
2	Ile, Leu, Phe, Pro	>50	0
3	Tyr, Met, Thr, Ser	>50	<1.0
4	His, Asn, Gln, Trp	>50	1.0<Dip.<2.0
5	Arg, Lys	>50	2.0<Dip.<3.0
6	Asp, Glu	>50	>3.0
7	Cys*	>50	<1.0

*Cys is differentiated from class 3 because it can form disulfide bonds

To calculate trigram frequency of the PDZ domains and corresponding peptide sequences, the number of occurrence of each subsequent trigram in the sequence is counted, and this number is divided by the total number of trigrams in the sequence which is (n-2), where n is the sequence length. At the end, we obtain 343 (7x7x7) features for each sequence because amino acids are clustered into seven classes resulting in 7x7x7 different combination of trigrams. For the interaction prediction part, feature vector space is constructed by combining trigram frequency sets of both PDZ domain and corresponding peptide which gave 686 features for each interaction (343 for PDZ domain, 343 for peptide). For bigram frequency calculation, the same procedure was applied and we obtained 49 (7x7) features for each sequence and a total of 98 features (49 for PDZ domain, 49 for peptide) were constructed for each interaction. Therefore, we obtained a feature vector space (X,Y,W) to represent an interaction:

$$\left(X,Y,W\right)=\left\{\left(x_1,x_2,\dots,x_{343}\right),\left(y_1,y_2,\dots,y_{343}\right),\left(w_1,w_2\right)\right\}$$

Here, X is the feature vector space of the PDZ sequence, and each feature x_i represents the frequency of each trigram where i=1,2,….,343 or each bigram where i=1,2,…,49, Y is the feature vector space of peptide sequence, each feature y_i represents the frequency of each trigram or bigram, and W is the corresponding label that contains binary data (w_1:binding, w_2:non-binding). Thus, a 686 dimensional vector for trigram part and a 98 dimensional vector for bigram part have been constructed to represent each binding/non-binding interaction.

For the class prediction part, the peptide sequences are discarded and only the sequences of PDZ domains are used to construct the feature vector space, because peptide sequences are used as the label of the dataset. Therefore, a 343 dimensional vector space for trigram part and 49 for bigram part with three labels (w_1:ClassI, w_2:ClassII, w_3:ClassI-II) have been built to represent each class of PDZ domain.

2.3 Machine Learning Classifiers

2.3.1 Used Classifiers

Five commonly used machine learning classifiers (SVM, Nearest Neighbor, Naïve Bayes, J48, Random Forest) are trained for both interaction prediction and classification models. After comparison of these different classifiers by using Weka 3.6 [64], it has been indicated that Random Forest algorithm outperforms other classifiers which were previously shown as the best classification algorithm such as SVM [65].

2.3.2 Comparison of Each Classifiers

Each classifier is trained by using 10-fold cross-validation. Cross-validation measures the prediction performance in a stable way by leaving out a few instances (about 10% for 10-fold cross-validation) to be used as test set during the training process. The exclusion is repeated until every instance in the dataset is once among those left-out. In comparison to using an independent test set, using cross-validation provide less bias and better predictive performance. Parameter selection for each classifier is done by varying their parameters step-by-step and their accuracy and AUC values are compared to obtain best parameters with highest performance (See Appendix A.2, Table A.2.1). At the end, the classifier with the best performance was chosen as a model classifier.

In order to determine classification statistics of used models, the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are used to calculate true positive rate (also named as recall or sensitivity), TPR=TP/(TP+FN), false positive rate, FPR=FP/(FP+TN) and precision, P=TP/(TP+FP). We measure the performance of each classifier by using a ROC curve. The ROC curve is drawn as TPR

(Sensitivity) versus FPR (1-Specificity). The area under the ROC curve, referred as AUC, represents the predictive power: while a random predictive model has AUC=0.5, a perfect one has AUC=1.0 so that a larger AUC shows better predictive power. However, ROC curves can sometimes be misleading while dealing with highly unbalanced datasets. Therefore, Precision versus Recall (PR) curves are also constructed to interpret the performance of our model in a more informative manner [66]. PR curves show how many true positives are likely to be obtained in a prediction system.

2.4 Prediction Models

2.4.1 PDZ Domain Interaction Prediction

Random forest was chosen to build our model due to the highest AUC and accuracy values of this algorithm (see the result section for comparison of classifiers for trigram and bigram models). To adjust the parameters of Random Forest algorithm, we evaluated the effect of changes in parameters on its prediction performance by measuring out-of-bag (OOB) error rate of each model tree. There are two parameters: number of trees (numTree) and number of randomly selected features (numFeature). The number of features to be used in random selection is rather sensitive and it must be much lower than the total number of features [43]. On the other hand, the changes in the number of trees can only result in small decreases in OOB error rate). Also, resampling was applied as a pre-process in order to balance our imbalanced dataset which could be overwhelmed by the major class otherwise and to derive robust estimates of standard errors. Resampling is a supervised filter producing a random subset of the dataset. In our study, class distribution was left as-is and sampling was done with replacement by adjusting the parameters.

2.4.2 PDZ Domain Class Predictions

There is a multi-classification problem for class prediction because we do not only want to discriminate between PDZ domains which bind to Class I or Class II, but also we want to classify Class I-II domains whose interaction specificity reflects the promiscuous pattern of PDZ domains. All five classifiers are trained on these classification dataset and again Random Forest gives the best predictive performance with highest AUC and accuracy values (see the result section for comparison of classifiers for trigram and bigram models). The parameters of Random Forest algorithm is also adjusted in class prediction part as done in interaction prediction model.

2.5 Critical Sequence Motifs

2.5.1 Dimensionality Reduction

In order to make the resulting model faster and extract important features, dimensionality of our dataset is reduced by using feature extraction and selection methods. Selection of important features can help us to get rid of redundant and/or irrelevant data. At this point, there is a need for the correct selection of the features [67]. Feature extraction is used to evaluate important features and it precedes feature selection which was used as a search method. Correlation-based feature subset selection method is used as a feature extraction method which considers the worth of a subset of features by evaluating the individual predictive performance of each feature. In this selection method, performance of individual features for predicting labels (w_i) as well as the level of intercorrelation among all features are considered, successful feature subsets include features highly correlated with the label, but uncorrelated with each other [68]. For feature selection part, several search algorithms are performed and results of all of them are considered in order to reduce features carefully. Used search methods are presented in Appendix A.2, Table A.2.2.

2.5.2 Extraction of Important Sequence Motifs

We analyzed the selected features to understand why they are important to distinguish different PDZ interactions and classes. These extracted features might correspond to some critical amino acid motifs which may be important for PDZ domain interaction specificity. Indeed, one of the obtained features point out the GLGF (Gly-Leu-Gly-Phe) repeat of PDZ domains which is an important conserved region for peptide interaction [19]. This conserved GLGF sequence is located between βA-βB loop and αB helix and directly involved in peptide selectivity and binding. In order to obtain these important sequence regions from our models, we firstly compare extracted features of trigram and bigram models. Consequently, we present these sequence motifs from trigrams and bigrams positioned at the similar secondary structure regions of PDZ domains.

Chapter 3

3 Peptide Library Construction

3.1 Genetic Algorithm

Genetic algorithm is a search heuristic technique which uses natural evolutionary process of most organisms. For example, it uses mutation and crossover mechanisms to reach better results as most organisms do in their evolution and reproduction. It usually starts with a random population and end with a population of potential solutions. Namely, it simulates the nature to reach better results.

3.1.1 Implementation

The population used in this study contains the sequences of binding peptides of a PDZ domain ($D_{desired}$). Each individual in the population consists of 10 amino acid sequences. Initial population ($D_{generated}$) is generated randomly by the algorithm and the objective function below (least square function) is used to quantify the difference between desired and generated amino acid probability distributions. The construction of amino acid distributions and generation of populations is shown in Figure 3.1.

$$S = \sum_{n} wt(n)[D_{desired}(n) - D_{generated}(n)]^2$$

Here, n represents each amino acid positions of the individuals in the population, $D_{desired}$ (7x10) is the desired amino acid probability distribution matrix, $D_{generated}$ (7x10) is the generated amino acid probability distribution matrix, wt is the weight function and S gives the optimized fitness values [69].

Weight function is used to discriminate the importance of each amino acid position of PDZ domain. As stated in literature review part, peptide positions 0 and -2 are very crucial for recognition and binding to target proteins because the positions 0 and -2 point towards to the binding pocket of PDZ domain during PDZ domain-peptide interaction [16]. Therefore, a weight function which gives more weight to these two positions (0 and -2) among other 8 amino acid positions is used in our objective function. Here, wt is a row-matrix which gives double weight to these two positions compared to other remaining positions in the peptide of 10 amino acids. Namely, a change in these two positions affect obtained fitness value more than others.

$$\begin{tikzcd}
\mathrm{D'}_{\text{desired}} = \begin{bmatrix} \mathit{RSLDRIETSG} \\ \vdots \qquad\qquad \vdots \\ \vdots \qquad\qquad \vdots \end{bmatrix}_{(Nx10)} \qquad \mathrm{D'}_{\text{generated}} = \begin{bmatrix} \mathit{LDGSMQWEIS} \\ \vdots \qquad\qquad \vdots \\ \vdots \qquad\qquad \vdots \end{bmatrix}_{(Nx10)} \arrow[d] \\
\mathrm{D'}_{\text{desired}} = \begin{bmatrix} 5326526331 \\ \vdots \qquad\qquad \vdots \\ \vdots \qquad\qquad \vdots \end{bmatrix}_{(Nx10)} \qquad \mathrm{D'}_{\text{generated}} = \begin{bmatrix} 2613744623 \\ \vdots \qquad\qquad \vdots \\ \vdots \qquad\qquad \vdots \end{bmatrix}_{(Nx10)} \arrow[d, "\frac{\mathrm{P_a(b)}}{\sum_{\mathrm{b}} \mathrm{P_a(b)}}"] \\
\mathrm{D}_{\text{desired}} = \begin{bmatrix} 0.15\ 0.30\ ...\ 0.08\ 0.11 \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ \vdots \qquad\qquad\qquad\qquad \vdots \end{bmatrix}_{(7x10)} \qquad \mathrm{D}_{\text{generated}} = \begin{bmatrix} 0.25\ 0.12\ ...\ 0.04\ 0.00 \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ \vdots \qquad\qquad\qquad\qquad \vdots \end{bmatrix}_{(7x10)} \arrow[d] \\
\mathrm{D}_{\text{distance}} = \mathrm{D}_{\text{desired}} - \mathrm{D}_{\text{generated}}
\end{tikzcd}$$

Figure 3.1: The construction of amino acid distributions and generation of populations for genetic algorithm.

Here, we try to minimize $D_{distance}$ which is the difference between $D_{desired}$ and $D_{generated}$. Firstly, we have sequences of amino acids in string form as a population; the string form has to be converted into the numeric form in order to be processes in the genetic algorithm. Therefore, 20 amino acid types are represented in 7 amino acid classes as seen in Table 2.1 and amino acid probability distribution of these sequences are calculated by the equation of $P_a(b)/\sum_b P_a(b)$. Here, a is the amino acid position, b is seven amino acid classes and $P_a(b)$ is the number of occurrence of amino acid b in position a. Therefore, by this equation, we calculate frequency of each amino acid in a specific position and we name it as amino acid probability distribution. At the end, an amino acid distribution matrix with a size of 7x10 is obtained to be used as populations for genetic algorithm.

3.1.2 Parameter Tuning

Parameter adjustment of the genetic algorithm is very crucial in order to obtain optimum results and the parameters have to be customized according to the nature of the problem.

First, there are many methods for selection (best, roulette, tournament, percent and random) which have to be selected according to the problem. We tried some different selection methods and compared them according to their fitness and accuracy values.

Second, crossover and mutation rates must be optimized carefully. To obtain a successful result, crossover rate is usually larger than the mutation rate. In this study, we tuned crossover and mutation rates to obtain best solution.

Third, features of the individuals has to be consistent with their importance in the problem, namely some weight function could be applied to the objective function. In this study, a weight function is used to highlight the importance of positions 0 and -2 of the peptide sequences which play crucial roles on PDZ domain specificity as stated in the introduction. Therefore, weight of these positions are kept higher than other positions and results of weighted and unweighted objective functions are compared to see the difference.

Lastly, termination criteria and generated population size which could considerably effect the performance of the algorithm is also tuned. All parameter tunings and found results are explained and shown in the result section.

3.1.3 Construction of Binding Peptides for a Given PDZ Domain

This implemented genetic algorithm can be used for each PDZ domain. The only necessary thing is to change desired amino acid probability matrix which consists of interested PDZ domain binding peptide sequence information.

Here, we analyzed and obtained results for α1-syntrophin as a representative manner. The desired amino acid probability matrix is generated with 20 peptide sequences which bind to α1-syntrophin. Genetic algorithm with optimized parameters is applied to this population and the resultant population is trained with PDZ interaction prediction model. The population with lowest fitness value and highest prediction accuracy is obtained. This population can be used for further experimental studies.

Chapter 4

4 Results and Discussion

4.1 Performance Evaluation for Each Classifier

All five classifiers (SVM, Nearest Neighbor, Naïve Bayes, J48 and Random Forest) are trained for both interaction prediction and classification trigram/bigram models and Random forest was chosen to build our models due to the highest AUC and accuracy values of this algorithm. The high performance of Random Forest algorithm for interaction prediction and classification trigram models can be seen in Figure 4.1 and Figure 4.3 respectively. Random forest also gives the best performance for interaction prediction and classification bigram models as seen in Figure 4.2 Figure 4.4 respectively.

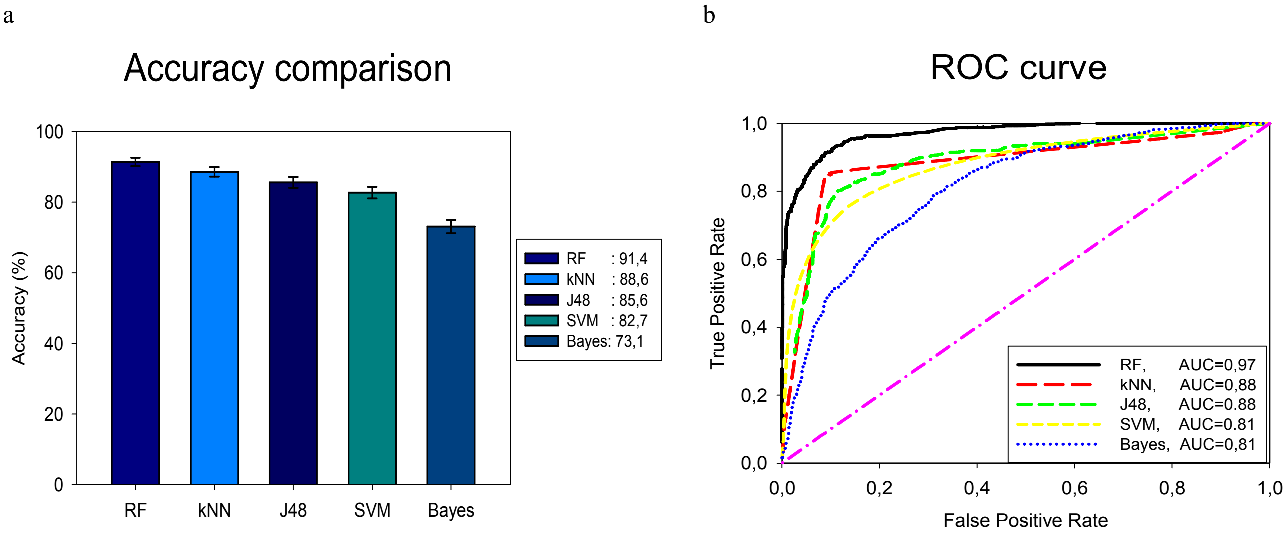


Figure 4.1: Comparison of all classifiers used in interaction prediction trigram model (a) Accuracy values with 95% confidence intervals (b) ROC curves and AUC values.

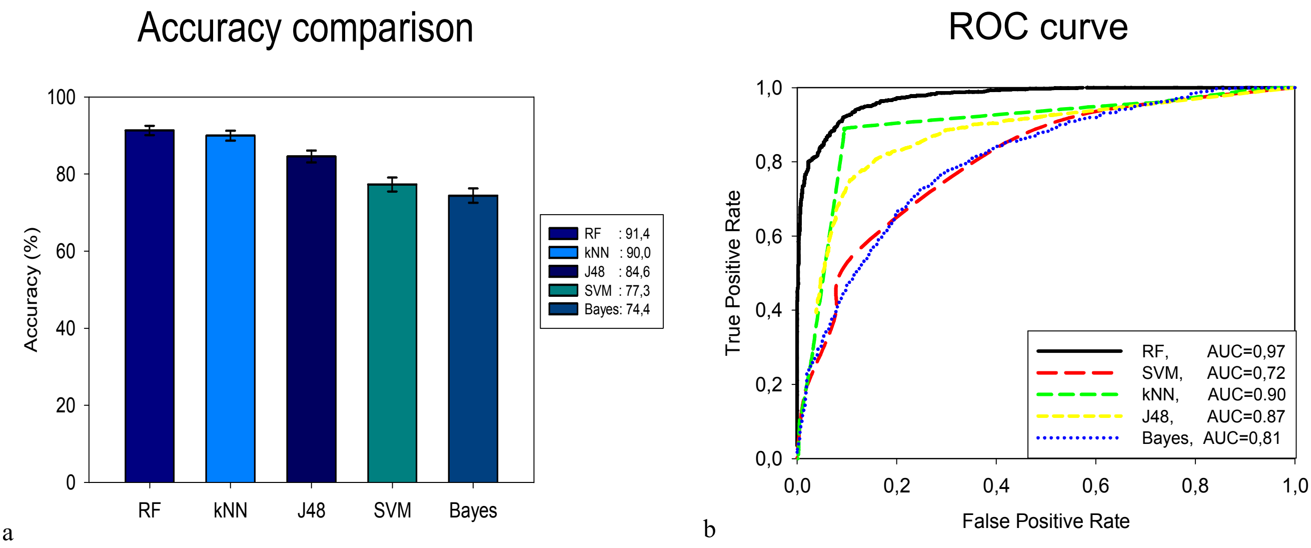


Figure 4.2: Comparison of all classifiers used in interaction prediction bigram model (a) Accuracy values with 95% confidence intervals (b) ROC curves and AUC values.

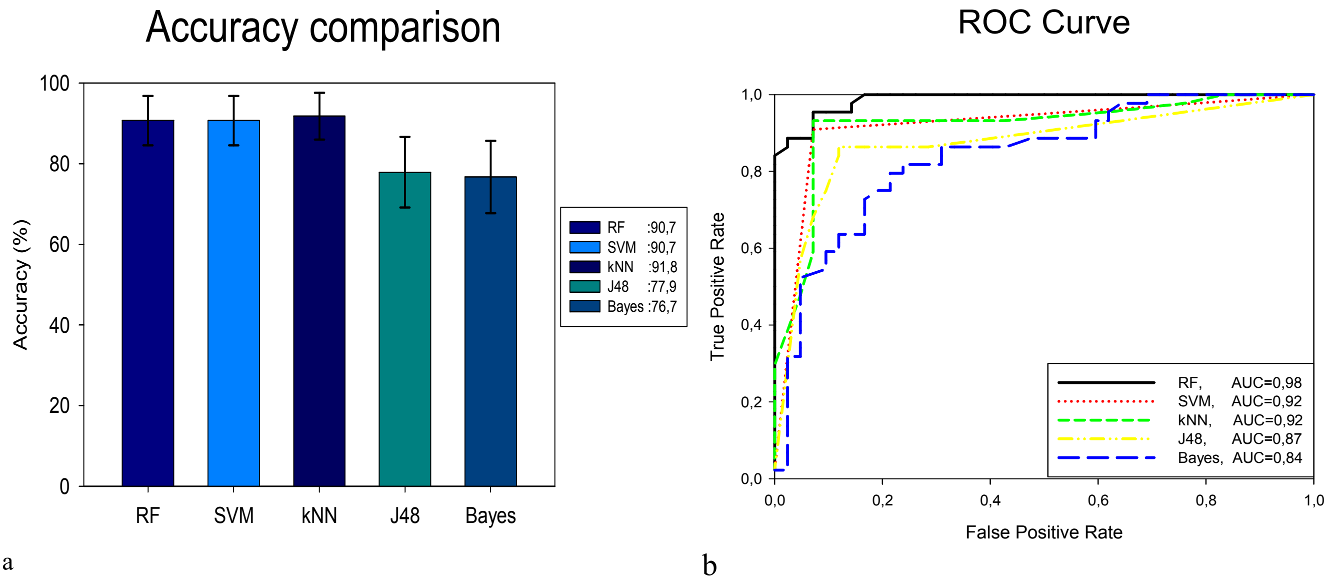


Figure 4.3: Comparison of all classifiers used in classification trigram model (the result of multi-classification to discriminate between Class I, Class II and Class I-II). (a) Accuracy values with 95% confidence intervals (b) ROC curves and corresponding AUC values.

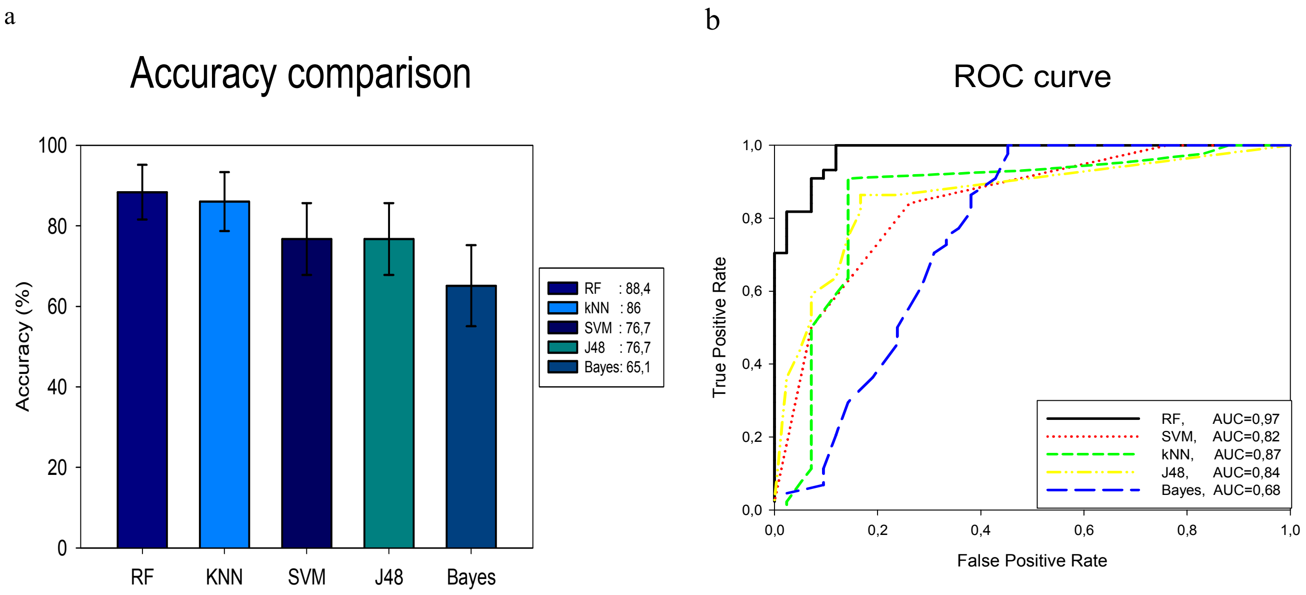


Figure 4.4: Comparison of all classifiers used in classification bigram model (the result of multi-classification to discriminate between Class I, Class II and Class I-II). (a) Accuracy values with 95% confidence intervals (b) ROC curves and corresponding AUC values.

Random Forest grows many classification trees. Each tree gives a classification, and votes for a class. The forest chooses the classification having the most votes over all the trees in the forest. Each tree is grown as follows: (i) if the number of instances in the training set is N, sample N is selected randomly with replacement from the original data. This sample will be the training set for growing the tree, (ii) if there are M features in an instance, a number m<<M is specified such that at each node, m features are selected randomly out of M features and the best split on these m is used to split the node, (iii) each tree is grown to the largest extent possible, namely there is no pruning.

4.2 PDZ Domain Interaction Prediction Statistics

To tune the parameters of Random Forest algorithm, we evaluated the effect of changes in parameters on its prediction performance by measuring out-of-bag (OOB) error rate of each model tree. There are two parameters for Random Forest: number of trees (numTree) and number of randomly selected features (numFeature). The lowest OOB error rate was obtained when numTree=200 and numFeature=30 as seen in Figure 4.5.

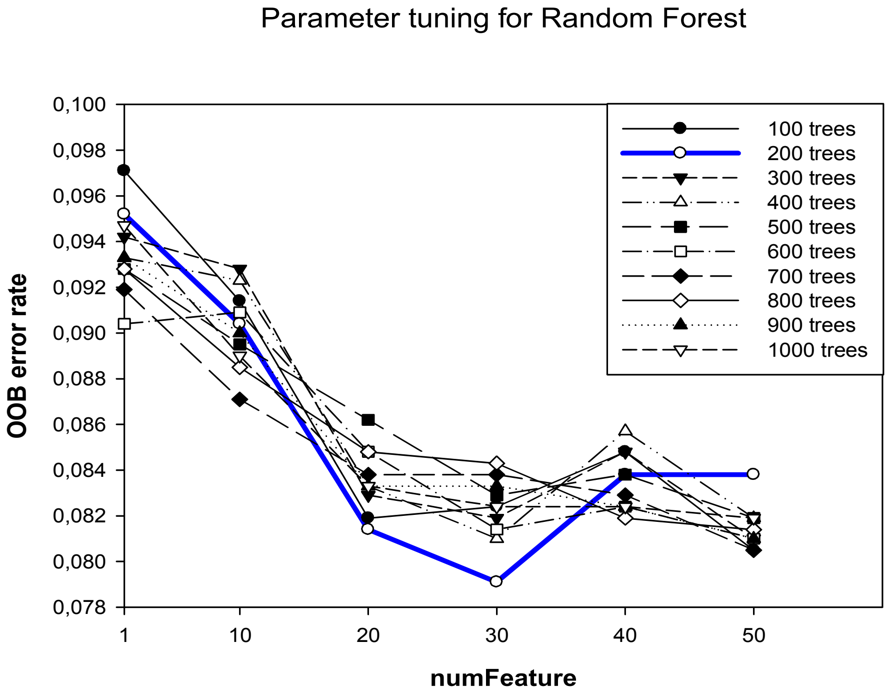


Figure 4.5: Parameter selection of Random Forest algorithm for interaction prediction.

After parameter tuning for Random forest, we trained our interaction model with numTree=200 and numFeature=30. The accuracy of trigram part (91.4%) is slightly higher than the bigram part (91.2%) (Table 4.1).

Table 4.1: Prediction results for interaction prediction of PDZ domains for both trigram and bigram models.

	Training set (10-fold cross validation)				Validation set			
	TPR	FPR	Precision	Accuracy	TPR	FPR	Precision	Accuracy
Trigram	0.89	0.075	0.85	91.4	0.61	0.042	0.92	79.8
Bigram	0.844	0.053	0.89	91.2	0.889	0.323	0.545	74.2

As seen in Table 4.1, although precision of trigram model is lower than the precision of bigram model, its other values (TPR, FPR and accuracy) are better. So, we design our model according to trigram frequency feature space. In order to see the performance of trigram model, ROC and PR curves are plotted (Figure 4.6a, b). Our result of AUC=0.97 for trigram part is high enough to be able to characterize PDZ interaction specificity.

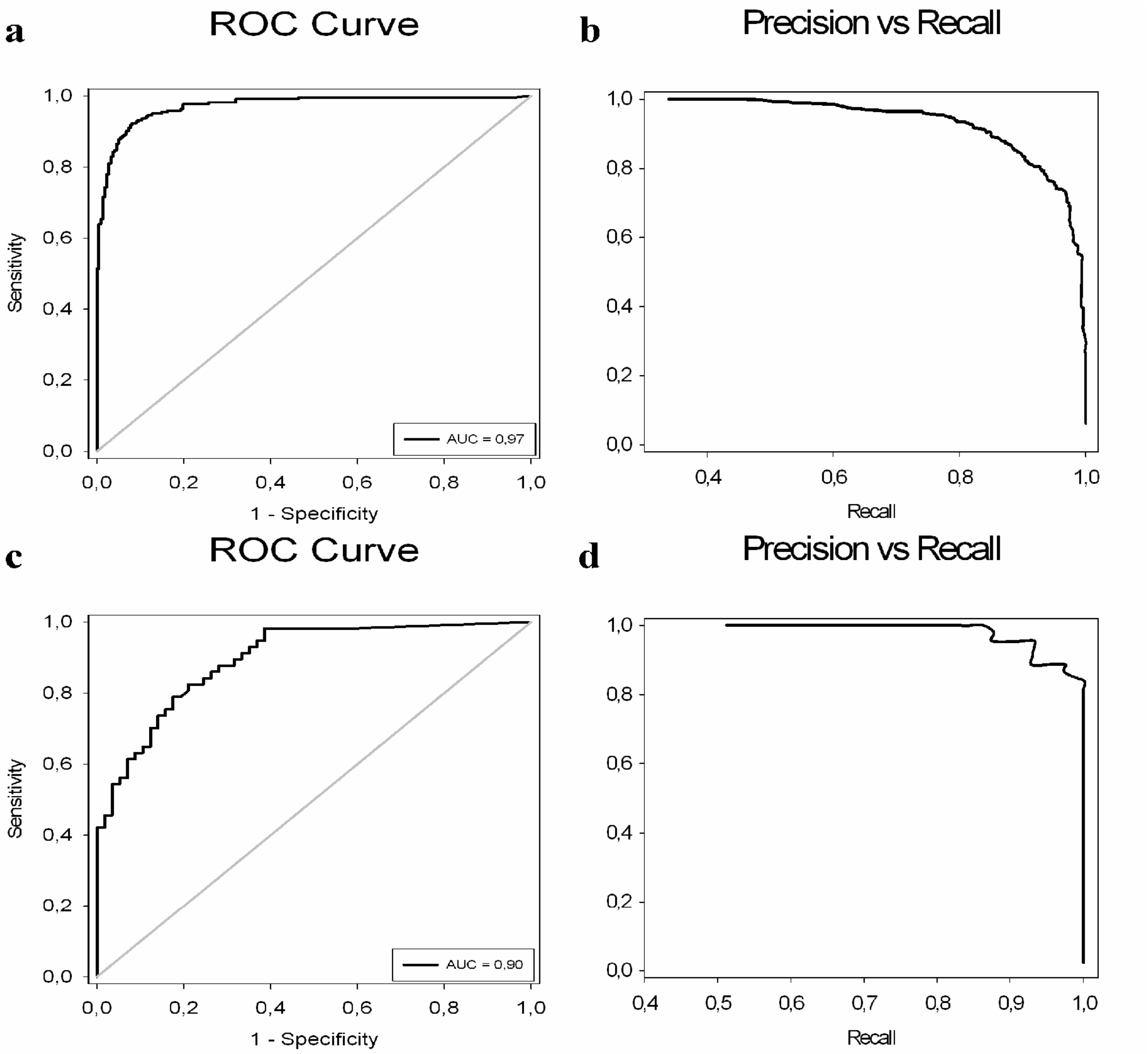


Figure 4.6: Performance evaluation of Random Forest trigram model. (a) ROC curve, (b) precision versus recall curve for interaction prediction part (c) ROC curve, (d) precision versus recall curve for classification part.

4.2.1 Interaction Validation Statistics

Also, we validate the power of our model by predicting the interaction of an unseen validation dataset. After training the model with complete set of PDZ interaction database, an unseen interaction validation dataset is sent to be classified according to the rules that are learned from the trained model. The model performs well on the validation set with an accuracy of 79.8% that it correctly classifies 25 of 27 binding and 46 of 62 non-binding interactions (Table 4.1). The performance of bigram model is somehow lower in validation dataset compared to trigrams. This can be due to the fact that bigrams assign more common features of most of the interactions by masking the discriminative features.

4.3 PDZ Domain Class Prediction Statistics

The Random Forest model is used to discriminate both multi-classes (Class I/Class II/Class I-II) and binary classes (Class I/Class II, Class I/Class I-II or Class II/Class I-II) in order to have an insight about their pair wise classifications.

As seen in Table 4.2, the predictive performance for multi-class learning is a bit lower than binary-class ones which are very close with each other and the results for trigram model has still better performance than bigram model. As we are trying to distinguish all three classes of PDZ domains, we obtained the performance results of trigram model for multi-class learning (Figure 4.6c, d). The model correctly classified 43 of 45 Class I, 16 of 20 Class II and 19 of 21 Class I-II PDZ domains. The results of binary comparisons in Table 4.2 show that the highest accuracy is for differentiating Class II PDZs from Class I-II PDZs and the least successful one is between Class I and Class I-II. This means that amino acid distribution of Class I-II PDZs is slightly more similar to Class

I PDZs. To ensure this similarity, amino acid frequency distribution histogram for Class I/II/I-II PDZ domains is plotted (Figure 4.7)

Table 4.2: Prediction results for class prediction of PDZ domains for trigram and bigram models.

	TP Rate		FP Rate		Precision		Accuracy (%)	
	Trigram	Bigram	Trigram	Bigram	Trigram	Bigram	Trigram	Bigram
ClassI, ClassII, Class I-II*	0.907	0.895	0.081	0.093	0.911	0.902	90.7	89.5
ClassI, ClassII	0.918	0.956	0	0.200	1	0.915	93.8	90.8
ClassI, ClassI-II	0.900	0.955	0	0.227	1	0.894	92.4	89.4
ClassII, ClassI-II	1	0.813	0.107	0	0.812	1	92.7	92.7

*The first row shows a multi-class learning and remaining rows shows the binary-class learning for pair wise combinations of three classes. For multi-class learning, weighted average results were shown.

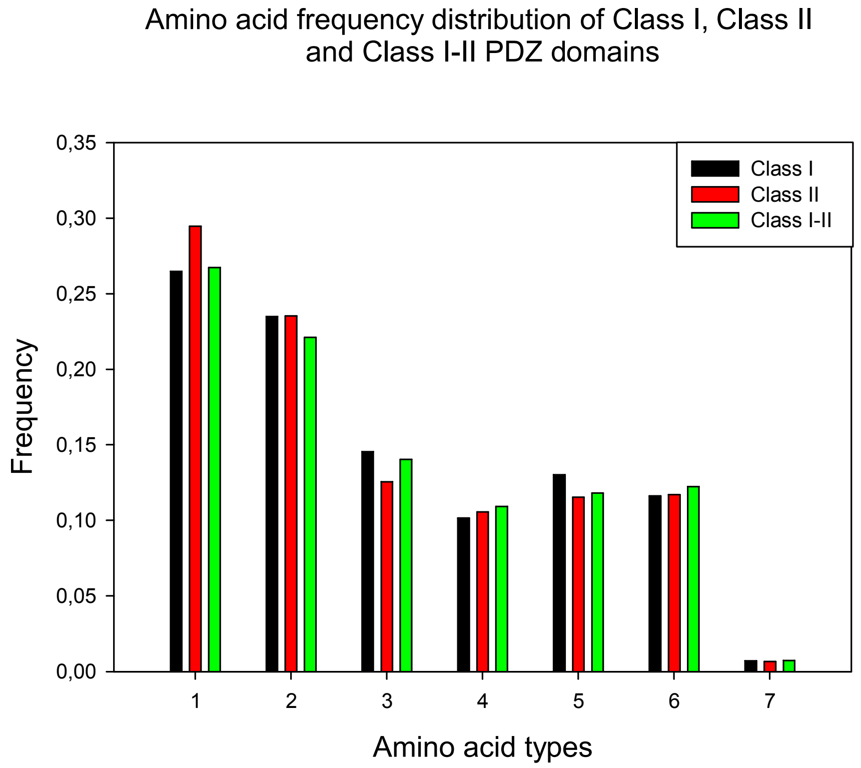


Figure 4.7: Amino acid frequency distribution of Class I/II/I-II PDZ domains.

4.4 Important Sequence Motifs of PDZ Domains

Dimension reduction is applied to both trigram and bigram models because we want to observe important common features of both models. For trigram model, we obtained 23 features for PDZ domain and 23 features for peptide spaces to represent interaction prediction part. Also, feature reduction was performed for classification part and we obtained 53 features (Data is not shown).

For bigram model, there were 11 features for PDZ domain and 12 features for peptide space in order to construct interaction prediction part and we extracted 10 features for classification part. The accuracy values of our model did not increase after feature reduction for both interaction prediction and classification parts except trigram classification model (Table 4.3). However, reduction in feature space let us analyze these extracted important features.

Table 4.3: Prediction results after feature reduction.

	TPR		FPR		Precision		AUC		Accuracy (%)	
	Trigram	Bigram	Trigram	Bigram	Trigram	Bigram	Trigram	Bigram	Trigram	Bigram
Interaction prediction	0.744	0.786	0.096	0.07	0.798	0.851	0.905	0.948	85	88.1
Classification*	0.942	0.86	0.044	0.096	0.942	0.859	0.994	0.966	94.2	86

*The first row shows a multi-class learning (ClassI, ClassII, Class I-II)

We observe that bigram parts of extracted trigrams are common to extracted bigrams, i.e. there are some highly occurring bigrams in both trigram and bigram feature sets for interaction prediction and classification parts. After the selection of the most occurring ones, we obtained sequence motifs of “12”, “16” and “25” (Figure 4.8). Types of amino acids can be seen in Table 2.1. For example, motif “12” corresponds to small hydrophobic amino acid (A, G, V) followed by large hydrophobic amino acid (I, L, F, P).

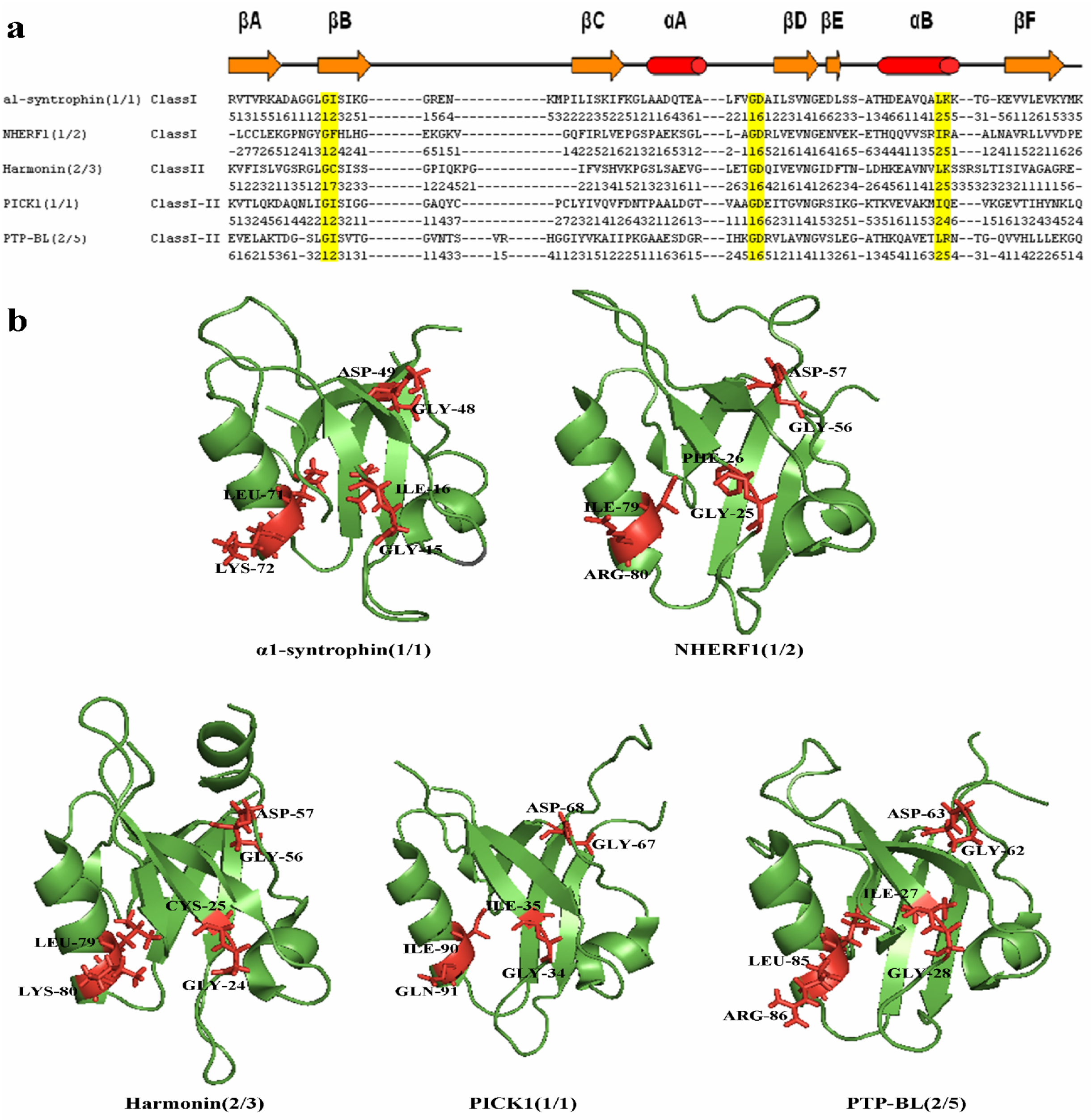


Figure 4.8: Critical sequence motifs (a) Aligned sequences of 5 representative PDZ domains: α1-syntrophin(1/1) (PDB ID:2pdz), NHERF1(1/2) (PDB ID:1i92), Harmonin(2/3) (PDB ID:2kbs), Pick1(1/1) (PDB ID:2pku) and PTP-BL(2/5) (PDB

ID:1vj6). While first row indicates the aligned sequence of corresponding PDZ domain, second row represents the sequence in seven class amino acid types. Secondary structure positions of the PDZ sequences are represented graphically at the top (αA, Ab, βA-βF). Three sequence motifs ("12", "16", "25") proposed to account for ligand specificity are indicated by yellow highlight. (b) Cartoon diagrams of these PDZ domains, motifs "12", "16" and "25" are colored in red and shown in stick form.

As seen in Figure 4.8, characteristic GLGF repeat of PDZ domains were determined by extracting sequence motif of "12" between βA-βB loop and αB helix. Other two highly occurring sequence motifs were positioned at the end of the αB ("25") and at the loop between αA and βD ("16"). When these sequence motifs are displayed on the 3D structure of PDZ domains, while motif "25" is positioned near the binding groove (at the end of the αB), motif "16" is positioned far from the binding groove (at the αA-βD loop). Extracted motif on αB helix could function in specificity of PDZ domains. Songyang *et al.* investigated the importance of αB helix on peptide selectivity of PDZ domains by showing high correlation between first residue in the αB helix and peptide position -2 [19].

4.5 Case Studies: Biologically Important Sequence Motifs

Below, we discuss important sequence motifs of some specific PDZ domains:

α1-syntrophin(1/1): The specific interaction property of α1-syntrophin PDZ domain is investigated by Schultz *et al.* and they found that Leu 14, Gly 15 and Ile 16 showed a large chemical shift upon binding of ligand [70]. PDZ domain of α1-syntrophin forms hydrophobic pocket consisting of Leu 14, Ile 18 and Leu 71 to bury the side chain of Val -2 of the peptide. Motif "12" corresponds to Gly 15, Ile 16 and "5" of motif "25" corresponds to Leu 71 which is an important part of the hydrophobic pocket.

NHERF(1/2): First PDZ domain of NHERF1 plays important role in cellular localization by binding to the cystic fibrosis transmembrane conductance regulator (CFTR) [71]. Leu 0 of the ligand forms hydrophobic contact with Phe 26 and Ile 79 and makes H-bonds with Gly 25, Phe 26 and Arg 80. These important residues were also extracted by using our method: while motif “12” in βB corresponds to Gly 25, Phe 26, motif “25” in αB exactly corresponds to Ile 79, Arg 80.

Harmonin(2/3): Pan *et al.* tried to elucidate structural basis of binding pattern of Harmonin(2/3) and found that carboxyl group of cad 23 ligand forms hydrogen bonds with Leu 222, Glu 223, Cys 224 (GLGF motif) and is stabilized by Lys 279 [72]. These important residues of Harmonin were also observed in our motifs as seen from Figure 3 (PDZ2 domain of Harmonin includes residues 208-299, but in the 3D structure it is between residues 9-100).

Pick1(1/1): The carboxyl group of ligand forms hydrogen bonds with Ile 33, Gly 34 and Ile 35 of Pick1 PDZ domain [73]. While Gly 34 and Ile 35 constitue motif “12”, we observed motif “24” on αB helix instead of motif “25”.

PTP-BL(2/5): Gianni *et al.* investigated allosteric property of PTP-BL(2/5) domain by using structural and dynamical methods and found that binding is regulated by long range interactions which showed correlation with ligand-induced structural rearrangements [74]. There is a detectable conformational change, dominantly occuring in αB-βB interface, L1 loop and hydrophobic core, upon ligand binding to PTP-BL domain. Plasticity and selectivity of PTP-BL domain are usually determined by reorientation of alpha B helix. Amides of Leu 25, Gly 26 and Ile 27 stabilize the charge of C-terminus of the ligand and there is a hydrophobic contact between C-terminal peptide valine and Leu 85, Val 82 positions. In our study, motif “12” in βB corresponds to the Gly 26, Ile 27 and “5” of motif “25” in αB corresponds to Leu 85 as seen Figure 4.8.

4.6 Constructed Peptide Libraries

4.6.1 Selection and Cross-over Methods

Selection is a genetic operator that chooses an individual from the current generation's population for inclusion in the next generation's population. Before making it into the next generation's population, selected individuals may undergo crossover and mutation in which case the offspring individual(s) are actually the ones that make it into the next generation's population. There are many different methods for selection and cross-over parts.

For selection stage, roulette (the chance of an individual selected is proportional to its fitness), tournament (uses roulette selection N times to produce a tournament subset of individuals), top percent (randomly selects an individual from the top N percent of the population), best (selects the best individuals according to their fitness value), random are commonly used. In this study, best, roulette and random methods are trained and best method gives the best fitness and accuracy values as seen in Table 4.4.

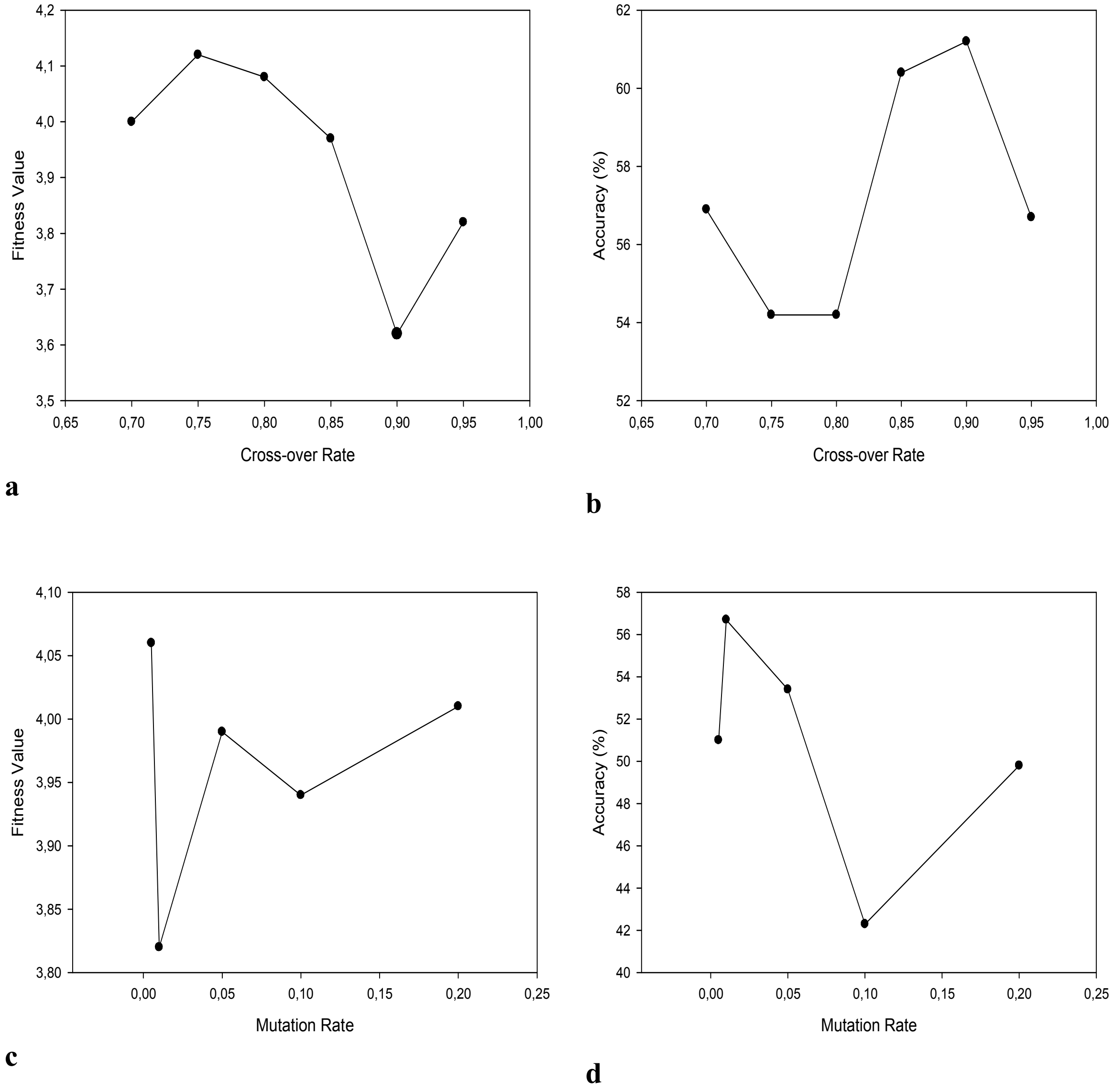
For cross-over part, there are also different methods such as one-point (a single crossover point on both parents), two point (two points on both parents) and uniform (individual bits in the parents are compared between two parents and the bits are swapped with a fixed probability) cross-overs. In this study, we tried these three cross-over methods and two-point cross-over gives the best result as seen in Table 4.4.

Table 4.4: Selection and cross-over method selection according to their fitness and accuracy values

Selection			Cross-over		
Method	Fitness value	% Accuracy	Method	Fitness value	% Accuracy
Best	3.54	63.8	One-point	3.65	63.2
Roulette	3.01	57.4	Two-point	3.54	63.8
Random	3.88	60.7	Uniform	3.87	61.5

4.6.2 Parameter Tuning

In order to construct peptide libraries specific to PDZ domains, we implemented the genetic algorithm. A distance matrix is minimized by this algorithm as explained in Chapter 4. Distance matrix calculates the difference between desired and generated population. To obtain the most successful results, parameters of the genetic algorithm is tuned according to our problem. Cross-over rate, mutation rate, iteration number (termination) and population size is tuned and obtained results are used in the algorithm. The results without weight function are not shown here because its performance is very low compared to the weighted one. Then, resultant populations are trained in our PDZ domain interaction prediction model. As seen in Figure 4.9, the parameters obtaining lowest fitness value and highest accuracy (of interaction prediction model) are selected to construct our genetic algorithm.


Fitness Value
4,2
4,1
4,0
3,9
3,8
3,7
3,6
3,5
0,65
0,70
0,75
0,80
0,85
0,90
0,95
1,00
Cross-over Rate
a
Accuracy (%)
62
60
58
56
54
52
0,65
0,70
0,75
0,80
0,85
0,90
0,95
1,00
Cross-over Rate
b
Fitness Value
4,10
4,05
4,00
3,95
3,90
3,85
3,80
0,00
0,05
0,10
0,15
0,20
0,25
Mutation Rate
c
Accuracy (%)
58
56
54
52
50
48
46
44
42
40
0,00
0,05
0,10
0,15
0,20
0,25
Mutation Rate
d

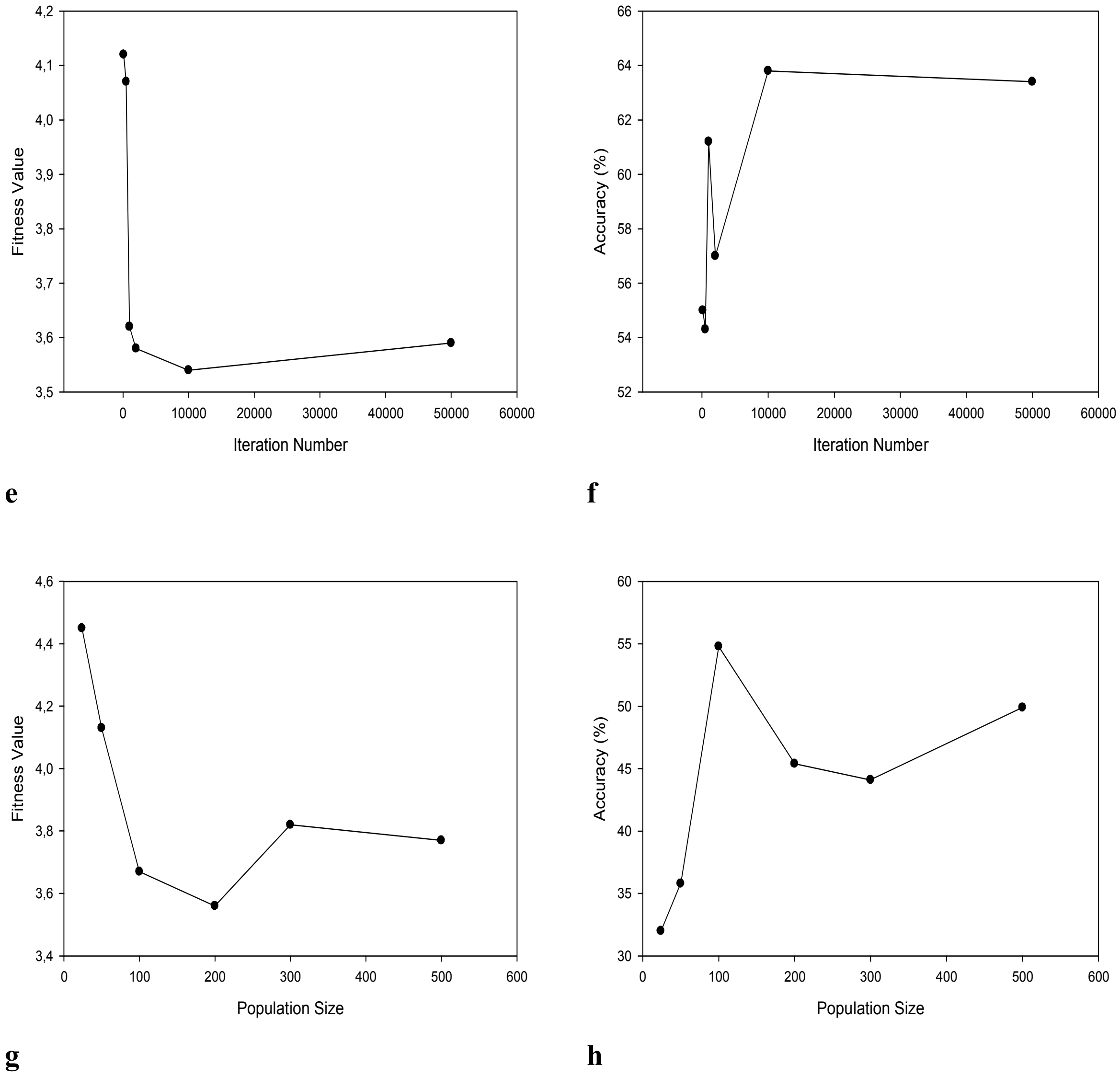


Figure 4.9: Parameter tunings of the genetic algorithm by looking at their fitness value (obtained from genetic algorithm) and accuracy values (obtained from interaction prediction model). Parameters: (a,b) cross-over rate, (c,d) mutation rate, (e,f) iteration number, (g,h) population size.

According to these results in Figure 4.9(a-f), both lowest fitness value and highest accuracy are obtained at the same point for cross-over rate, mutation rate and iteration number, so the following parameters are used for genetic algorithm: cross-over rate is 0.9, mutation rate is 0.01 and iteration number is 10000. However, for the population size tuning, while highest accuracy is obtained for population size of 100, lowest fitness value is not obtained for that population (Figure 4.9g, h). Rather population size of 200 gives the lowest fitness value. We choose the population size as 100 because the difference between fitness values of population size of 100 and 200 is smaller than the difference between their accuracy values.

4.6.3 Peptide Library for α1-syntrophin

As a representative example, peptide library for α1-syntrophin is constructed. The parameters obtaining lowest fitness value and highest model accuracy is used to obtain peptide library (cross-over rate: 0.9, mutation rate: 0.01, iteration number: 10000, and population size: 100). The resultant peptide library (population size of 100) with a fitness value of 3.54 and interaction prediction accuracy of 63.8% is shown in Appendix A.3, Table A.3.1.

Chapter 5

5 Conclusion

5.1 Conclusion

This study includes three inter-correlated aims: prediction of PDZ domain-peptide interactions, and classification of PDZ domains as Class I, II and I-II and peptide library construction for PDZ domains. A statistical learning model was constructed by using interaction dataset of PDZ domains (consist of 85 PDZ domains and corresponding 181 peptides). To convert primary sequence information into numerical feature input, trigram and bigram amino acid frequencies are calculated for each instance. We predicted binary interactions and classes of PDZ domains with accuracies of 91.4% and 90.7% respectively. After feature extraction, three critical amino acid sequence motifs are proposed to have significant roles on PDZ domain specificity. With these highly encouraging results, this study could be an important step in the automated prediction of PDZ domain interactions. Our results for PDZ interaction prediction and classification models are published on June 2010 [75]. Then, peptide libraries for PDZ domains are constructed by means of genetic algorithm and these populations are trained by PDZ domain interaction model and we obtained an accuracy of 63.8%.

Combination of these three methods can be effectively used as a virtual screening method for PDZ domains. It could predict the binding/non-binding binding partners and classes of PDZ domains by giving only the sequence information of PDZ domains and corresponding peptides and also it can generate possible target peptide sequences for PDZ domains.

5.2 Future Work

There could be more work to do in order to increase the performance of this study. First, feature encoding for primary sequences of known protein interaction pairs could be subsequently developed by using other additional features such as binding affinities, secondary/tertiary structure information in the learning model. Second, other dimensionality reduction techniques (feature extraction and selection methods) can be used to determine more critical sequence motifs and improve the performance of prediction models. Third, the results of genetic algorithm could be further improved by using different objective function(s), weight function(s), and selection/cross-over methods. Also, parameter tuning of genetic algorithm could be further improved by increasing time points for cross-over/mutation rates, population size and iteration number. By this mean, performance of peptide library construction could be increased and more specific putative peptides binding to PDZ domains could be obtained.

With this newly developed method, the door is opened to the future identification of binding partners for other PDZ domains in addition to derivation of a detailed description of their binding specificity. Further protein–peptide interactions where terminal carboxyl groups play a role will similarly be open to investigation. Moreover, variations of this method will allow the synthesis of peptides with modified C termini, facilitating ways of studying the biological function of C-terminal modifications using peptide libraries.

Further improvements on these lines may generate a powerful computational virtual screening technique that significantly reduces the search space for putative candidate target proteins of PDZ domains. Further, we plan to provide a web server that will predict binding peptide candidates for PDZ domains.

Appendix

A.1 Datasets

Table A.1.1: Classes of PDZ domains according to their binding/non-binding target peptide sequences.

PDZ domain	Organism	Class
a1-syntrophin-(1/1)	mouse	ClassI
b1-syntrophin-(1/1)	mouse	ClassI
Chapsyn-110-(2/3)	mouse	ClassI
Chapsyn-110-(3/3)	mouse	ClassI
Erbin-(1/1)	mouse	ClassI
g1-syntrophin-(1/1)	mouse	ClassI
g2-syntrophin-(1/1)	mouse	ClassI
Interleukin-16-(1/4)	mouse	ClassI
LIN-7A-(1/1)	mouse	ClassI
Lin7c-(1/1)	mouse	ClassI
Lrrc7-(1/1)	mouse	ClassI
Magi-1-(2/6)	mouse	ClassI
Magi-1-(4/6)	mouse	ClassI
Magi-2-(2/6)	mouse	ClassI
Magi-2-(5/6)	mouse	ClassI
Magi-3-(2/5)	mouse	ClassI
Mals2-(1/1)	mouse	ClassI
MUPP1-(12/13)	mouse	ClassI
MUPP1-(13/13)	mouse	ClassI
NHERF-1-(1/2)	mouse	ClassI
NHERF-2-(2/2)	mouse	ClassI
nNOS-(1/1)	mouse	ClassI
OMP25-(1/1)	mouse	ClassI
PAR6B-(1/1)	mouse	ClassI
Pdlim5-(1/1)	mouse	ClassI
Pdzk1-(1/4)	mouse	ClassI
Pdzk1-(3/4)	mouse	ClassI
Pdzk11-(1/1)	mouse	ClassI
PSD95-(1/3)	mouse	ClassI
PSD95-(2/3)	mouse	ClassI
PSD95-(3/3)	mouse	ClassI
SAP102-(2/3)	mouse	ClassI
SAP102-(3/3)	mouse	ClassI
SAP97-(1/3)	mouse	ClassI
SAP97-(2/3)	mouse	ClassI
SAP97-(3/3)	mouse	ClassI
Scrb1-(1/4)	mouse	ClassI
Scrb1-(2/4)	mouse	ClassI
Scrb1-(3/4)	mouse	ClassI
Semcap3-(1/2)	mouse	ClassI
Shank3-(1/1)	mouse	ClassI
Shroom-(1/1)	mouse	ClassI
SLIM-(1/1)	mouse	ClassI
TIP-1-(1/1)	mouse	ClassI
Whirlin-(3/3)	mouse	ClassI
ABP-(3/7)	rat	ClassII
ABP-(5/7)	rat	ClassII
ABP-(6/7)	rat	ClassII
AF-6-(1/1)	human	ClassII
ASIP/PAR3-(1/3)	mouse	ClassII
CASK-(1/1)	human	ClassII
ZO-1-(2/3)	human	ClassII
p55-(1/1)	mouse	ClassII
Cipp-(5/10)	mouse	ClassII
Cipp-(9/10)	mouse	ClassII
D930005D10Rik-(1/1)	mouse	ClassII
Dlgh3-(1/1)	mouse	ClassII
Grip1-(6/7)	mouse	ClassII

Harmonin-(2/3)	mouse	ClassII
Mpp7-(1/1)	mouse	ClassII
MUPP1-(1/13)	mouse	ClassII
MUPP1-(5/13)	mouse	ClassII
MUPP1-(11/13)	mouse	ClassII
PAR-3-(3/3)	mouse	ClassII
ZO-1-(2/3)	mouse	ClassII
Cipp-(3/10)	mouse	ClassI-II
Cipp-(8/10)	mouse	ClassI-II
Cipp-(10/10)	mouse	ClassI-II
Dvl1-(1/1)	mouse	ClassI-II
Dvl2-(1/1)	mouse	ClassI-II
Dvl3-(1/1)	mouse	ClassI-II
GRASP55-(1/1)	mouse	ClassI-II
HtrA1-(1/1)	mouse	ClassI-II
HtrA3-(1/1)	mouse	ClassI-II
Lnx1-(2/4)	mouse	ClassI-II
Magi-1-(6/6)	mouse	ClassI-II
Magi-2-(6/6)	mouse	ClassI-II
Magi-3-(1/5)	mouse	ClassI-II
Magi-3-(5/5)	mouse	ClassI-II
MUPP1-(10/13)	mouse	ClassI-II
PDZ-RGS3-(1/1)-	mouse	ClassI-II
PTP-BL-(2/5)	mouse	ClassI-II
ZO-1-(1/3)	mouse	ClassI-II
ZO-2-(1/3)	mouse	ClassI-II
PICK1-(1/1)	human	ClassI-II
Syntenin-(2/2)	mouse	ClassI-II

Table A.1.2: PDZ domain sequence IDs

PDZ domain	**Sequence ID**
a1-syntrophin-(1/1)	Q61234
b1-syntrophin-(1/1)	Q99L88
Chapsyn-110-(2/3)	Q91XM9
Chapsyn-110-(3/3)	Q91XM9
Cipp-(3/10)	Q63ZW7
Cipp-(5/10)	Q63ZW7
Cipp-(8/10)	Q63ZW7
Cipp-(9/10)	Q63ZW7
Cipp-(10/10)	Q63ZW7
D930005D10Rik-(1/1)	Q69Z89
Dlgh3-(1/1)	O88910
Dvl1-(1/1)	P51141
Dvl2-(1/1)	Q60838
Dvl3-(1/1)	Q61062
Erbin-(1/1)	Q80TH2
g1-syntrophin-(1/1)	Q925E1
g2-syntrophin-(1/1)	Q925E0
Gm1582-(2/3)	UPI0000D670BC (196-264)
GRASP55-(1/1)	Q99JX3
Grip1-(6/7)	Q925T6
Grip2-(5/7)	UPI00001E3EA7 (431-504)
Harmonin-(2/3)	Q9ES64
HtrA1-(1/1)	Q9R118
HtrA3-(1/1)	Q9D236
Interleukin-16-(1/4)	Q9QZP6
LARG-(1/1)	UPI0000D63612 (296-364)
LIN-7A-(1/1)	Q8JZS0
Lin7c-(1/1)	O88952
Lnx1-(2/4)	O70263
Lrrc7-(1/1)	Q80TE7
Magi-1-(2/6)	Q6RHR9
Magi-1-(4/6)	Q6RHR9
Magi-1-(6/6)	Q6RHR9
Magi-2-(2/6)	Q9WVQ1
Magi-2-(5/6)	Q9WVQ1
Magi-2-(6/6)	Q9WVQ1
Magi-3-(1/5)	Q9EQJ9
Magi-3-(2/5)	Q9EQJ9
Magi-3-(5/5)	Q9EQJ9
Mals2-(1/1)	O88951
Mpp7-(1/1)	Q8BVD5

MUPP1-(1/13)	Q8VBX6
MUPP1-(5/13)	Q8VBX6
MUPP1-(10/13)	Q8VBX6
MUPP1-(11/13)	Q8VBX6
MUPP1-(12/13)	Q8VBX6
MUPP1-(13/13)	Q8VBX6
NHERF-1-(1/2)	P70441
NHERF-2-(2/2)	Q9JHL1
nNOS-(1/1)	Q9Z0J4
OMP25-(1/1)	Q8K4F3
PAR-3-(3/3)	Q99NH2
PAR3B-(1/3)	Q8TEW8
PAR6B-(1/1)	Q9JK83
Pdlim5-(1/1)	Q8CI51
Pdzk1-(1/4)	Q9JIL4
Pdzk1-(3/4)	Q9JIL4
Pdzk11-(1/1)	Q9CZG9
Pdzk3-(1/1)	ENSMUSP00000074788
Pdzk3-(2/2)	ENSMUSP00000043100
PDZ-RGS3-(1/1)	P49796
PSD95-(1/3)	Q62108
PSD95-(2/3)	Q62108
PSD95-(3/3)	Q62108
PTP-BL-(2/5)	Q64512
SAP102-(2/3)	P70175
SAP102-(3/3)	P70175
SAP97-(1/3)	Q811D0
SAP97-(2/3)	Q811D0
SAP97-(3/3)	Q811D0
Scrb1-(1/4)	Q80U72
Scrb1-(2/4)	Q80U72
Scrb1-(3/4)	Q80U72
Semcap3-(1/2)	Q69ZS0
Shank1-(1/1)	Q9Y566
Shank3-(1/1)	Q4ACU6
Shroom-(1/1)	Q9QXN0
SLIM-(1/1)	Q8R1G6
Tiam2-(1/1)	ENSMUSP00000024562
TIP-1-(1/1)	Q9DBG9
Whirlin-(3/3)	Q80VW5
ZO-1-(1/3)	P39447
ZO-1-(2/3)	P39447
ZO-2-(1/3)	Q9Z0U1
ZO-3-(1/3)	UPI00005652A2
ABP-(3/7)	Q9WTW1
ABP-(5/7)	Q9WTW1
ABP-(6/7)	Q9WTW1
AF-6-(1/1)	P55196
ASIP/PAR3-(1/3)	Q99NH2
CASK-(1/1)	O14936
ZO-1-(2/3)	Q07157
p55-(1/1)	P70290
PICK1-(1/1)	Q9NRD5
Syntenin-(2/2)	O08992

Table A.1.3: Peptide sequences (up to -10 amino acid position)

Peptide	**Sequence**
AcvR1	NSLDKLKTDC
AcvR2	VDFPPKESSL
AcvR2b	VDLLPKESSI
AN2	PALRNGQYWV
APC	HSGSYLVTSV
Aquaporin-4	DSSGEVLSSV
AXL	PAPPGQEDGA
Cacna1a	AYSESEDDWC
Caspr2	IDESKKEWLI
Caspr4	VGENQKEYFF
Cav1.2	ADSRSYVSNL
Cav2.2	YHHPDQDHWC
Cav2.3	LSDTEEDDKC
Cav3.2	APDDSGDEPV
Cftr	TEEEVQETRL
c-KIT	TQPLLVHEDA
Claudin-1	PTPSSGKDYV

Claudin-10	SKQFDKNAYV
Claudin-11	SPTHAKSAHV
Claudin-14	HSGYRLNDYV
Claudin-15	FGKYGKNAYV
Claudin-16	AKMYAVDTRV
Claudin-18	QSHPTKYDYV
Claudin-19	GPSTAAREYV
Claudin-2	FNSYSLTGYV
Claudin-22	LELKQANPEI
Claudin-23	QNSLPCDSDL
Claudin-3	GTAYDRKDYV
Claudin-4	ARSVPASNYV
Claudin-5	NGDYDKKNYV
Claudin-6	PSEYPTKNYV
Claudin-7	PKSNSSKEYV
Claudin-8	PSIYSKSQYV
Claudin-9	ASGLDKRDYV
CNGA2	INTPEPAVAE
CNGA3	ENSEDASKTD
Cnksr2	HTHSYIETHV
Connexin-43	SRPRPDDLEI
CRIPT	DTKNYKQTSV
CSF-1R	LLQPNNYQFC
CtBP1	ADRDHTSDQL
DDR1	FLADDALNTV
Dlgap1/2/3	IYIPEAQTRL
EGFR	APPSSEFIGA
EphA2	DQVNTVGIPI
EphA3	TQSKNGPVPV
EphA4	QQMHGRMVPV
EphA5	VQMVNGMVPV
EphA6	MHIQEKGFHV
EphA7_1	LHLHGTGIQV
EphA7_2	LVTNEHLSVL
EphB2	QMNQIQSVEV
EphB3	QMNQTLPVQV
EphB6_1	HLRQPGSVEV
Ephrin-B1/2	QSPANIYYKV
Ephrin-B3	QSPPNIYYKV
ErbB4	VAQGATAEMF
FGFR3	GPPSNGGPRT
FGFR4	PFPFSDSQTT
Frizzled	TNSKQGETTV
GluR1	SGMPLGATGL
GluR2_1	NVYGIESVKI
GluR2_2	GMNVSVTDLS
GluR2_3	PKGTSLGWVE
GluR3	NVYGTESVKI
GluR5_1	RRTQRKETVA
GluR5_2	IRTQPSVHTV
GluRdelta1	ALDTSHGTSI
GluRdelta2	GNDPDRGTSI
Glycophorin-C	GDTSKKEYFI
GRK6	DSEEELPTRL
Htr2c	NVVSERISSV
JAM-1	EFKQTSSFLV
KA-2	TGPRELTEHE
KCNAB2	KPYSKKDYRS
KCNE4_1	RQAEGLVSIC
KCNE4_2	GSSENIHQNS
KCNH1	ESDRDIFGAS
KCNK3	RGLMKRRSSV
KCNK4_1	LEDFIKAMAI
KCNK4_2	GRLRDKAVPV
KCNK5	YNKADNPRGT
KCNK6	GPEREAPRSA
KCNQ2	PGTPRVTSQL
KIF17	SKNSFGGEPL
KIF1B	NLKAGRETTV
Kir2.1	PRPLRRESEI
Kir2.2	VRPYRRESEI
Kir3.2_2	VANLENESKV
Kir3.2_3	NPEELTERNG
Kir3.3	LPPPESESKV
Kir4.1	SALSVRISNV
Kir4.2	RSLLLQQSNV
Kir5.1	LNRISMESQM
Kir6.1	PEGNQCPSES

Kir6.2	KFSISPDSLS
Kv1.1	VNKSKLLTDV
Kv1.2	VNITKMLTDV
Kv1.3	VNIKKIFTDV
Kv1.4	SNAKAVETDV
Kv1.5	CLDTSRETDL
Kv1.6	YAEKRMLTEV
Kv1.7	PAGKHMVTEV
Kv2.1	AHGSTRDQSI
Kv3.1	GRKPLRGMSI
Kv3.3_1	RAPPTLPSIL
Kv3.3_2	FGERDSETQV
Kv4.1	LPETVKISSL
Kv4.2	GGNIVRVSAL
L-glutaminase	LSKENLESMV
Liprin-a2	DNSTVRTYSC
Megalin	ANLVKEDSDV
Mel1a/b	NNNLIKVDSV
mGluR1	RDYKQSSSTL
mGluR3	EVLDSTTSSL
Na/Pi-cotransporter	LPAHHNATRL
Nav1.4	VRPGVKESLV
Nav1.5	SPDRDRESIV
Nav1.6	RQKEVRESKC
Nav2	EEKASIQTQI
Neurexin-1/2	KKNKDKEYYV
Neurexin-3	QKNKDKEYYV
Neurexin-4	PQILEESRSE
Neuroligin-2	LPHPHSTTRV
NHE1	EGEPFIPKGQ
NMDAR2A	KKMPSIESDV
NMDAR2B	EKLSSIESDV
NMDAR2C	RRISSLESEV
NMDAR2D	AHFSSLESEV
P2Y1	EFKQNGDTSL
Parkin	ACMGDHWFDV
PDGFR	PLAEAEDSFL
PDGFRa_1	SSDLVEDSFL
PDGFRa_2	HSGKYDLSVV
PFK-M	SRKRSGEAAV
PIX	NDPAWDETNL
PKC	FVHPILQSAV
PMCA1	SPLHSLETSL
PTK7	LGDSPADSKQ
Ril	VYPNAKVELV
ROR1	HTESMISAEV
ROR2	TEAAHVQLEA
RYK	EFHAALGAYV
Sapk3	GARVPKETAL
Sema3a	HEFERAPRSV
Sema3b	ERGPRSAAHW
Sema3f	RNRRHHPPDT
Sema4a-	DNNHLGAEVA
Sema4b-	LGSEIRDSVV
Sema4c	PDSNPEESSV
Sema4f	PLATCDETSI
Sema5a	FTDLNNYDEY
Sema6b-	TGERTAPPVP
Sema6c	PAPHGGHFNF
SERCA2A	NYLEQPAILE
SERCA3	RGESPVWPSD
SSTR2	SGAEDIIAWV
Stargazin	NTANRRTTPV
Syndecan-1	KPTKQEEFYA
Syndecan-2	QKAPTKEFYA
Syndecan-3	KPDKQEEFYA
TAZ	NKSEPFLTWL
TIE1	AGIDATAEEA
TPC1	GSRQRSQTVT
Trip6	ELSATVTTDC
TRPC1	SKYAMFYPRN
TRPC2	EGDLETKGES
TRPC3	KLNPSVLRCE
TRPC4	AHEDYVTTRL
TRPC5	GQEEQVTTRL
TRPC6	LEPKLEESRR
TRPM3	DPAEHPFYSV

TRPM5	SQPLLETGST	TRPV3	ELDEFPETSV
TRPM6	RSSLEDHTRL	TRPV4	PKWRTDDAPL
TRPM7	EATNSVRLML	TRPV6	EDGEGWEYQI
TRPM8	LLKEIANNIK	TYRO3	QQGLLPHSSC
TRPP2	SGNGSANVHA		

A.2 Classifiers

Table A.2.1: Parameter values for each classifier used in trigram interaction prediction model.

Classifier	Parameters
SVM	Complexity parameter: 1.0
	Tolerance parameter: 0.001
	Epsilon: 10-12
	Kernel: Linear kernel (Exponent: 1.0)
Nearest Neighbor	k (number of neighbors to use): 1
	Distance function: Eucledian distance
Naïve Bayes	Alpha: 0.5 (Simple Estimator)
J48	Confidence factor: 0.25
	Min number of objects per leaf: 2
Random Forest	Number of trees: 200
	Number of Features:30

Table A.2.2: Search methods that are used to reduce dimensionality.

Search Method	Description
Best First	Searches the space of attribute subsets by greedy hillclimbing augmented with a backtracking facility
Greedy Stepwise	Performs a greedy forward or backward search through the space of attribute subsets
Linear Forward Selection	Extension of BestFirst. Takes a restricted number of k attributes into account. Fixed-set selects a fixed number k of attributes, whereas k is increased in each step when fixed-width is selected
Rank Search	From the ranked list of attributes, subsets of increasing size are evaluated, ie. The best attribute, the best attribute plus the next best attribute, etc.
Subset Size Forward Selection	The search performs an interior cross-validation (here 5-fold is used). A Linear Forward Selection is performed on each fold to determine the optimal subset-size

A.3 Generated Population

Table A.3.1: Peptide Library for α1-syntrophin with 7 amino acid class notation.

4152726571	3274657131	1132432371	3223543653	5326526331
4213727373	3275126131	1232432521	2722466541	5324226331
7152726571	3745126331	1132326521	2463543371	5326526311
4252726571	3245657332	1132426321	2722356541	5324256331
4213327311	3745127331	1132326371	5442577563	5324256311
4153326571	3245657331	2132432522	5676562441	5324526331
4212726571	3275126153	1142432521	7463454531	5324257331
4152727312	3274657151	1144326321	7116426541	5324226317
7213327311	2745126351	2244326521	2463722371	7764257311
7152326511	2744657351	2144326371	6175125431	7766557311
4213327311	2274657151	2244432371	7441311761	7764257331
7113327311	2774126151	2244332571	6724354241	7766527311
7152726571	2274656151	2244432521	1571625431	7766527331
7212727311	2774126151	1244326371	4614631641	7766257311
7153327311	2744657331	2234326371	7315654641	7766526311
7213326571	2745126331	2232432571	7476634641	7766557331
4466431331	6272334452	2717455443	2233215371	3715111423
1745233253	1655347714	5615364731	4263434641	3425721342
1446336457	3151324243	7532241553	7571311761	1433134121
5631776743	7234414261	4416555175	4476634641	6214212447

Bibliography

[1] T. Pawson, Dynamic control of signaling by modular adaptor proteins, *Current Opinion in Cell Biology* **19** (2007), 112-116.

[2] T. Pawson and P. Nash, Assembly of cell regulatory systems through protein interaction domains, *Science* **300** (2003), 445-452.

[3] Z. Keskin, A. Gursoy, B. Ma*, et al.*, Principles of protein-protein interactions: What are the preferred ways for proteins to interact? *Chemical Reviews* **108** (2008), 1225-1244.

[4] Z. Songyang, A. S. Fanning, C. Fu*, et al.*, Recognition of unique carboxyl-terminal motifs by distinct PDZ domains, *Science* **275** (1997), 73-77.

[5] B. J. Hillier, K. S. Christopherson, K. E. Prehoda*, et al.*, Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex, *Science* **284** (1999), 812-815.

[6] C. Nourry, S. G. Grant and J. P. Borg, PDZ domain proteins: plug and play! *Sci STKE* **2003** (2003), RE7.

[7] K. K. Dev, PDZ domain protein-protein interactions: A case study with PICK1, *Current Topics in Medicinal Chemistry* **7** (2007), 3-20.

[8] P. Jemth and S. Gianni, PDZ domains: folding and binding, *Biochemistry* **46** (2007), 8701-8.

[9] K. K. Dev, Making protein interactions druggable: Targeting PDZ domains, *Nature Reviews Drug Discovery* **3** (2004), 1047-1056.

[10] M. Van Ham and W. Hendriks, PDZ domains-glue and guide, *Mol Biol Rep* **30** (2003), 69-82.

[11] A. Y. Hung and M. Sheng, PDZ domains: structural modules for protein complex assembly, *J Biol Chem* **277** (2002), 5699-702.

[12] N. Basdevant, H. Weinstein and M. Ceruso, Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study, *J Am Chem Soc* **128** (2006), 12766-77.

[13] D. A. Doyle, A. Lee, J. Lewis*, et al.*, Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ, *Cell* **85** (1996), 1067-76.

[14] Z. N. Gerek, O. Keskin and S. B. Ozkan, Identification of specificity and promiscuity of PDZ domain interactions through their dynamic behavior, *Proteins* **77** (2009), 796-811.

[15] A. S. Fanning and J. M. Anderson, Protein-protein interactions: PDZ domain networks, *Curr Biol* **6** (1996), 1385-8.

[16] D. L. Daniels, A. R. Cohen, J. M. Anderson*, et al.*, Crystal structure of the hCASK PDZ domain reveals the structural basis of class II PDZ domain target recognition, *Nat Struct Biol* **5** (1998), 317-25.

[17] I. Bezprozvanny and A. Maximov, Classification of PDZ domains, *FEBS Lett* **509** (2001), 457-62.

[18] E. Song, S. Gao, R. Tian*, et al.*, A high efficiency strategy for binding property characterization of peptide-binding domains, *Mol Cell Proteomics* **5** (2006), 1368-81.

[19] Z. Songyang, A. S. Fanning, C. Fu*, et al.*, Recognition of unique carboxyl-terminal motifs by distinct PDZ domains, *Science* **275** (1997), 73-7.

[20] M. Ferrer, J. Maiolo, P. Kratz*, et al.*, Directed evolution of PDZ variants to generate high-affinity detection reagents, *Protein Eng Des Sel* **18** (2005), 165-73.

[21] B. S. Kang, D. R. Cooper, Y. Devedjiev*, et al.*, Molecular roots of degenerate specificity in syntenin's PDZ2 domain: reassessment of the PDZ recognition paradigm, *Structure* **11** (2003), 845-53.

[22] J. Reina, E. Lacroix, S. D. Hobson*, et al.*, Computer-aided design of a PDZ domain to recognize new target sequences, *Nat Struct Biol* **9** (2002), 621-7.

[23] U. Wiedemann, P. Boisguerin, R. Leben*, et al.*, Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides, *J Mol Biol* **343** (2004), 703-18.

[24] J. R. Chen, B. H. Chang, J. E. Allen*, et al.*, Predicting PDZ domain-peptide interactions from primary sequences, *Nat Biotechnol* **26** (2008), 1041-5.

[25] M. A. Stiffler, J. R. Chen, V. P. Grantcharova*, et al.*, PDZ domain binding selectivity is optimized across the mouse proteome, *Science* **317** (2007), 364-9.

[26] C. Schillinger, P. Boisguerin and G. Krause, Domain Interaction Footprint: a multi-classification approach to predict domain-peptide interactions, *Bioinformatics* **25** (2009), 1632-9.

[27] R. Tonikian, Y. Zhang, S. L. Sazinsky*, et al.*, A specificity map for the PDZ domain family, *PLoS Biol* **6** (2008), e239.

[28] H. S. Eo, S. Kim, H. Koo*, et al.*, A machine learning based method for the prediction of G protein-coupled receptor-binding PDZ domain proteins, *Mol Cells* **27** (2009), 629-34.

[29] E. J. Kim and M. Sheng, PDZ domain proteins of synapses, *Nature Reviews Neuroscience* **5** (2004), 771-781.

[30] H. Hirbec, O. Perestenko, A. Nishimune*, et al.*, The PDZ proteins PICK1, GRIP, and syntenin bind multiple glutamate receptor subtypes - Analysis of PDZ binding motifs, *Journal of Biological Chemistry* **277** (2002), 15221-15224.

[31] G. E. Torres and M. G. Caron, Functional interaction between monoamine plasma membrane transporters and the synaptic PDZ domain-containing protein PICK1, *Molecular Biology of the Cell* **12** (2001), 247A-247A.

[32] B. Dziedzic, V. Prevot, A. Lomniczi*, et al.*, Neuron-to-glia signaling mediated by excitatory amino acid receptors regulates ErbB receptor function in astroglial cells of the neuroendocrine brain, *Journal of Neuroscience* **23** (2003), 915-926.

[33] W. J. Lin, Y. F. Chang, W. L. Wang*, et al.*, Mitogen-stimulated TIS21 protein interacts with a protein-kinase-Ccn-binding protein rPICK1, *Biochemical Journal* **354** (2001), 635-643.

[34] K. J. D. A. Excoffon, A. Hruska-Hageman, M. Klotz*, et al.*, A role for the PDZ-binding domain of the coxsackie B virus and adenovirus receptor (CAR) in cell adhesion and growth, *Journal of Cell Science* **117** (2004), 4401-4409.

[35] H. Nielsen, S. Brunak and G. Von Heijne, Machine learning approaches for the prediction of signal peptides and other protein sorting signals, *Protein Engineering* **12** (1999), 3-9.

[36] J. R. Bradford and D. R. Westhead, Improved prediction of protein-protein binding sites using a support vector machines approach, *Bioinformatics* **21** (2005), 1487-94.

[37] X. W. Chen and M. Liu, Prediction of protein-protein interactions using random decision forest framework, *Bioinformatics* **21** (2005), 4394-400.

[38] R. Jansen, H. Yu, D. Greenbaum*, et al.*, A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science* **302** (2003), 449-53.

[39] C. Cortes and V. Vapnik, Support-Vector Networks, *Machine Learning* **20** (1995), 273-297.

[40] P. B. Brazdil, C. Soares and J. P. Da Costa, Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results, *Machine Learning* **50** (2003), 251-277.

[41] N. Friedman, D. Geiger and M. Goldszmidt, Bayesian network classifiers, *Machine Learning* **29** (1997), 131-163.

[42] J. R. Quinlan (1993). C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers.

[43] L. Breiman, Random forests, *Machine Learning* **45** (2001), 5-32.

[44] S. C. Sharma, A. M. Memic, C. N. Rupasinghe*, et al.*, T7 Phage Display as a Method of Peptide Ligand Discovery for PDZ Domain Proteins, *Biopolymers* **92** (2009), 183-193.

[45] S. H. Joo and D. Pei, Synthesis and screening of support-bound combinatorial peptide libraries with free C-termini: Determination of the sequence specificity of PDZ domains, *Biochemistry* **47** (2008), 3061-3072.

[46] D. Saro, E. Klosi, A. Paredes*, et al.*, Thermodynamic analysis of a hydrophobic binding site: Probing the PDZ domain with nonproteinogenic peptide ligands, *Organic Letters* **6** (2004), 3429-3432.

[47] M. Sainlos, W. S. Iskenderian and B. Imperiali, A General Screening Strategy for Peptide-Based Fluorogenic Ligands: Probes for Dynamic Studies of PDZ Domain-Mediated Interactions, *Journal of the American Chemical Society* **131** (2009), 6680-+.

[48] D. G. Udugamasooriya, S. C. Sharma and M. R. Spaller, A chemical library approach to organic-modified peptide ligands for PDZ domain proteins: A synthetic, thermodynamic and structural investigation, *Chembiochem* **9** (2008), 1587-1589.

[49] G. Udugamasooriya, D. Saro and M. R. Spaller, Bridged peptide macrocycles as ligands for PDZ domain proteins, *Organic Letters* **7** (2005), 1203-1206.

[50] S. C. Sharma, C. N. Rupasinghe, R. B. Parisien*, et al.*, Design, synthesis, and evaluation of linear and cyclic peptide Ligands for PDZ10 of the Multi-PDZ domain protein MUPP1, *Biochemistry* **46** (2007), 12709-12720.

[51] C. P. Ponting, C. Phillips, K. E. Davies*, et al.*, PDZ domains: Targeting signalling molecules to sub-membranous sites, *Bioessays* **19** (1997), 469-479.

[52] H. J. Lee and J. J. Zheng, PDZ domains and their binding partners: structure, specificity, and modification, *Cell Communication and Signaling* **8** (2010), -.

[53] G. Fuh, M. T. Pisabarro, Y. Li*, et al.*, Analysis of PDZ domain-ligand interactions using carboxyl-terminal phage display, *Journal of Biological Chemistry* **275** (2000), 21486-21491.

[54] P. R. Cushing, A. Fellows, D. Villone*, et al.*, The relative binding affinities of PDZ partners for CFTR: A biochemical basis for efficient Endocytic recycling, *Biochemistry* **47** (2008), 10084-10098.

[55] U. Wiedemann, P. Boisguerin, R. Leben*, et al.*, Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of superbinding peptides, *Journal of Molecular Biology* **343** (2004), 703-718.

[56] M. A. Stiffler, V. P. Grantcharova, M. Sevecka*, et al.*, Uncovering quantitative protein interaction networks for mouse PDZ domains using protein microarrays, *J Am Chem Soc* **128** (2006), 5913-22.

[57] S. Mathivanan, B. Periaswamy, T. K. B. Gandhi*, et al.*, An evaluation of human protein-protein interaction data in the public domain, *Bmc Bioinformatics* **7** (2006), -.

[58] A. Ceol, A. Chatr-Aryamontri, E. Santonico*, et al.*, DOMINO: a database of domain-peptide interactions, *Nucleic Acids Research* **35** (2007), D557-D560.

[59] B. Brannetti and M. Helmer-Citterich, iSPOT: a web tool to infer the interaction specificity of families of protein modules, *Nucleic Acids Research* **31** (2003), 3709-3711.

[60] T. Beuming, L. Skrabanek, M. Y. Niv*, et al.*, PDZBase: a protein-protein interaction database for PDZ-domains, *Bioinformatics* **21** (2005), 827-828.

[61] D. Whitley, A Genetic Algorithm Tutorial, *Statistics and Computing* **4** (1994), 65-85.

[62] M. Srinivas and L. M. Patnaik, Adaptive Probabilities of Crossover and Mutation in Genetic Algorithms, *Ieee Transactions on Systems Man and Cybernetics* **24** (1994), 656-667.

[63] J. Shen, J. Zhang, X. Luo*, et al.*, Predicting protein-protein interactions based only on sequences information, *Proc Natl Acad Sci U S A* **104** (2007), 4337-41.

[64] I. H. Witten and E. Frank (2005). Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco.

[65] Y. Qi, Z. Bar-Joseph and J. Klein-Seetharaman, Evaluation of different biological data and computational classification methods for use in protein interaction prediction, *Proteins* **63** (2006), 490-500.

[66] J. Davis and M. Goadrich (2006). The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning (ICML).

[67] A. K. Jain, R. P. W. Duin and J. C. Mao, Statistical pattern recognition: A review, *Ieee Transactions on Pattern Analysis and Machine Intelligence* **22** (2000), 4-37.

[68] M. A. Hall and L. A. Smith, Feature subset selection: A correlation based filter approach, *Progress in Connectionist-Based Information Systems, Vols 1 and 2* (1998), 855-858

1372.

[69] W. Wang and J. G. Saven, Designing gene libraries from protein profiles for combinatorial protein experiments, *Nucleic Acids Research* **30** (2002), -.

[70] J. Schultz, U. Hoffmuller, G. Krause*, et al.*, Specific interactions between the syntrophin PDZ domain and voltage-gated sodium channels, *Nature Structural Biology* **5** (1998), 19-24.

[71] S. Karthikeyan, T. Leung and J. A. A. Ladias, Structural basis of the Na+/H+ exchanger regulatory factor PDZ1 interaction with the carboxyl-terminal region of the cystic fibrosis transmembrane conductance regulator, *Journal of Biological Chemistry* **276** (2001), 19683-19686.

[72] L. F. Pan, J. Yan, L. Wu*, et al.*, Assembling stable hair cell tip link complex via multidentate interactions between harmonin and cadherin 23, *Proceedings of the National Academy of Sciences of the United States of America* **106** (2009), 5575-5580.

[73] L. Pan, H. Wu, C. Shen*, et al.*, Clustering and synaptic targeting of PICK1 requires direct interaction between the PDZ domain and lipid membranes, *Embo Journal* **26** (2007), 4576-4587.

[74] S. Gianni, T. Walma, A. Arcovito*, et al.*, Demonstration of long-range interactions in a PDZ domain by NMR, kinetics, and protein engineering, *Structure* **14** (2006), 1801-1809.

[75] S. Kalyoncu, O. Keskin and A. Gursoy, Interaction prediction and classification of PDZ domains, *Bmc Bioinformatics* **11** (2010), -.