THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

PreDiZ: a PDZ domain-peptide interaction prediction method

Zhangyan Dai

Bachelor of Science (Honours)

A thesis submitted for the degree of Master of Philosophy at

The University of Queensland in 2015

School of Chemistry and Molecular Biosciences

Abstract

PDZ domains are one of the most well studied peptide binding domains. These domains usually bind short peptides at the C-terminus of their target proteins and play a crucial role in cellular signalling processes. Computational approaches have been published to determine the interaction specificity of PDZ domains, but these prediction methods often limit their predictions on a limited subset of PDZ domains. In this research work, we developed PreDiZ, a computational method for PDZ domain-peptide interaction prediction based on the SDR approach. The SDR approach was originally created to predict specificity of protein kinases. In this work, improvements have been made to apply the SDR approach to the PDZ domains, including using a more sophisticated strategy to determine SDRs, and using both positive and negative interactions in the prediction. As a result, PreDiZ is able to work on a wide range of PDZ domains, including novel PDZ domains. In cross-validations, PreDiZ scored AUCs range from 0.82 to 0.94. In the comparison against published methods, PreDiZ showed competitive performance on making prediction to distantly related PDZ domains, but not as good as other recently published methods on mouse test set. However, the results also suggested that PreDiZ could be improved by optimising SDRs. We also conducted proteome-wide predictions on *A. thaliana*, *C. elegans*, *D. rerio*, *M. musculus* and *H. sapiens* and showed PDZ domains were evolved relatively late in eukaryotic cells. Network studies on human PDZ domain interaction revealed the enriched GO terms and KEGG pathways of PDZ domain binding proteins. Lastly, we studied how H7N9's NA and H5N1's NS1 protein regulate human biological processes using the human PDZ interaction network. Human proteins that regulated via PDZ domain interactions by these two kind of viral proteins, were enriched in similar biological processes. Therefore, we concluded that the function of PBM in the NS1 proteins of H5N1 was replaced by the PBM in the NA proteins of H7N9.

Declaration by author

(All candidates to reproduce this section in their thesis verbatim)

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

No publications

Publications included in this thesis

No publications included

Contributions by others to the thesis

No contributions by others.

Statement of parts of the thesis submitted to qualify for the award of another degree

None

Acknowledgements

I would like to express my very great appreciation to *Prof.* Bostjan Kobe and *Prof.* Mikael Boden, my supervisors, for their guidance, encouragement and advice of this research project. I would also like to thank *Prof.* Yu Xue and Dr. Jonathan Ellis for their valuable and constructive suggestions at different stage of this work. My grateful thanks are also extended to *Prof.* Alan Mark for his useful critiques.

I would also like to extend my thanks to Dr. Lei Li for sharing unpublished PDZ interaction data.

Finally, I dedicate this work to my parents, for their love and encouragement, without whom I would never have the opportunity to finish this work.

Keywords

PDZ domain, protein-peptide interaction, PreDiZ, the SDR approach, prediction

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060102 Bioinformatics, 100%

Fields of Research (FoR) Classification

FoR code: 0601, Biochemistry and Cell Biology, 100%

Table of Contents

List of Figures

List of Figures

List of Abbreviations used in the thesis

ROC: Receiver operating characteristic

AUC: Area under ROC curve

GO: Gene ontology

KEGG: Kyoto encyclopedia of genes and genomes

HPAI: Highly pathogenic avian influenza

MCD: Maximum contacting distance

ACS: Average column similarity

PIR: Proportion of interacting residues

BLOSUM: BLOSUM matrix

PBM: PDZ domain binding motif

SDR: Specificity-determining residues

ACC: Accuracy

TPR: True positive rate

SPC: Specificity

MCC: Matthew’s correlation coefficient

1 Introduction

1.1 Peptide-protein interactions

Proteins often interact with each other to form functional complexes and these interactions are important in all biological pathways and signalling mechanisms (1). Understanding protein-protein interactions is crucial to the study of protein functions and reconstructing biological pathways. It will also help in discovering the functions of new proteins by identifying their interaction partners.

Many proteins have well-structured globular domains. The functions of proteins are largely dependent upon these domains. However, disordered regions of proteins also play an important role in protein functions (2, 3). Recently, increasing evidence shows that linear motifs inherent in these disordered regions play important roles in protein-protein interactions. Linear motifs are short peptide sequences containing key residues for function or binding (4). They usually form transient complexes with their interaction partners. Peptide-protein interactions usually have smaller interacting interfaces and weaker interaction affinities compared to interactions between globular domains. The transient nature makes linear motifs very good candidates for signalling pathways, which require fast response to stimuli (5).

Experimental methods such as the yeast two-hybrid approach (6, 7), affinity purification-mass spectrometry (8) and oriented peptide libraries (9) have been widely used to perform large-scale analyses of interactions in different organisms. These experimental methods have produced a massive quantity of interaction information, however they are also known to be expensive, labour-intensive and time-consuming. Moreover, the high-throughput methods may generate false positive results, and therefore particular caution is necessary when using these data (10, 11).

With the rise of bioinformatics in recent years, computational methods can be used to complement certain limitations of high-throughput experiments (12). Generally, these bioinformatics algorithms can be very helpful for selecting potential targets for experimental screening or for validating experimental data. Sometimes, they can even provide detailed binding information which might not be found easily using the experimental techniques. Hence, they reduce significantly the time and cost required to determining interactions experimentally (13). A number of domains such as SH2, SH3, WW, 14-3-3 and PDZ domains have been found to interact with their partners via peptide-protein interactions.

1.2 Interaction studies on PDZ domains

PDZ domain, also named GFGL domain or DHR domain, was first identified in three proteins: postsynaptic density protein-95, disks large tumor suppressor and zonula occludens-1; and named using the first letter of each of these proteins (14-16). PDZ domains are involved in cell signalling and polarity, and mostly found in multi-cellular organisms. Hence, it has been suggested that PDZ domains co-evolved with multi-cellularity (17). PDZ domains usually consist of 80 to 90 amino acids and fold into a globular structure comprising six beta-strands (βA to βF) and two alpha-helices (αA and αB). The N- and C-termini of PDZ domains are mostly found close in space. There are over 400 structures of PDZ domains in the Protein Data Bank (PDB) (18). These protein domains share only around 30% sequence identity on average, but the core structure remains the same (19). Furthermore, some PDZ domains contain variable loop regions and extension regions that affect their structures and functions (19). Experiments show that PDZ domain structures are robust to extensive mutagenesis (20).

PDZ domains are able to bind short peptides at the C-terminus of their target proteins (21). Short C-terminal peptides are recognised by a carboxylate binding loop (βA – βB loop), containing the conserved GLGF motif, and the αB helix (22-25). The binding peptide binds to the PDZ domain as an anti-parallel extension of the β-sheet of the PDZ domain and the ligand residues in positions -1 and -3 point towards to the solvent, while residues in positions 0 and -2 point towards to the binding pocket, with the last residue at the C-terminus as position 0 (12, 26). PDZ domains are quite promiscuous with their specificity; one PDZ domain can interact with multiple peptides and the same peptide can be recognised by multiple PDZ domains (27). The specificity of PDZ domains can be changed by mutagenesis, which make PDZ domains adapt quickly during evolution (27). The PDZ-peptide interactions are regulated by various factors, such as buffer conditions (28, 29), allosteric differences (30) and phosphorylation in PDZ-domain binding motives (PBM) (31). In addition, some PDZ domains are also able to bind to internal (non-C-terminal) motifs (32-39) and membrane phospholipids (40-43).

Early studies classified PDZ domains into classes based on the last few amino acids in their binding peptides (21). Class 1 PDZ domains bind the motif Ser/Thr-X-Φ-COOH and class 2 PDZ domains bind Φ-X-Φ-COOH, where X is any residue and Φ is a hydrophobic residue (26). Less common classes of PDZ domains recognize a different motif, Asp/Glu-X-Φ-COOH (22). Later research showed that it was not appropriate to classify PDZ domains into three simple classes (44, 45). Tonikian*, et al.* (46) demonstrated that PDZ domains can be classified into up to 16 different classes and the selection specificity depended on up to 7 C-terminal residues in their binding peptides. Bezprozvanny and Maximov (44) classified PDZ domains into 25 different sub-classes based on amino acids in two positions of PDZ domains. Further research on PDZ binding specificity showed that every position of the last five C-terminal amino acid of the binding peptide affects the binding specificity (47).

PDZ domain interactions have been experimentally characterised using methods such as immunoprecipitation experiments (48-50), mass spectrometry (51), the yeast two-hybrid approach (52), and oriented peptide libraries (9, 21). However, high throughput methods are known to be expensive, labour-intensive and time-consuming. With the help of computational methods, the time and cost of determining interactions experimentally can be reduced.

1.3 Computational predictive methods of PDZ-peptide interactions

A number of computational approaches have been published to determine the interaction specificity of PDZ domains, as they are one of the most well studied interaction domains. These approaches have used sequence information, structure information or both to predict specificity of PDZ domains.

Structure-based methods rely on the information provided from 3D structures of PDZ domain-peptide complexes. With the 3D structures available, researchers can study in detail how the two proteins interact, as well as the physicochemical properties of the two interacting partners (53). On the other hand, these methods are often limited by the lack of available 3D structures. There are over 400 PDZ domain structures in the PDB (54), compared to 58998 PDZ domain-containing proteins available in the SMART nrdb database (55). The lack of coverage of PDZ domains will make it difficult for this kind of approaches to predict PDZ domains without known structures. Therefore, most structure-based methods can only predict interactions with the PDZ domain structure available.

Encinar*, et al.* (56) published a structure-based interaction prediction tool, called prediADAN, as part of the ADAN database. It provides position-specific scoring matrices for 212 PDZ domains calculated from their high quality structures using a protein design algorithm called FoldX (57). Their benchmarking showed an area under receiver operating characteristic (AUC) value from 0.48 to 0.96.

Smith and Kortemme (58) analysed 17 human PDZ domain structures. They used Rosetta to simulate and score interactions between large numbers of peptides with five residues against these PDZ domain proteins. Then the position specific scoring matrices (PSSMs) of PDZ domains were calculated for each domain. They evaluated their method on mutated Erbin PDZ domains. For Erbin single point mutation, the method reported an AUC value of 0.90 and 0.72 for Erbin with 10 mutations.

Both structure-based methods described above provide web access, and require a protein structure to perform interaction prediction. However, it is not always the case that the structure is available for the protein of interest. This will limit the usage of these methods. On the other hand, sequence-based methods analyse the amino acid sequences of the PDZ domains and their binding partners. Some methods also consider extra parameters in their predictions, such as structures, co-localisation and phylogenetic profiles. In general, these methods benefit from the large amount of available data for PDZ domain interactions (53).

Stiffler*, et al.* (59) performed a large scale study on mouse PDZ domains and developed a prediction model called multidomain selectivity model (MDSM). They identified both positive and negative interactions between 157 mouse PDZ domains against 217 peptides using yeast-two-hybrid experiments. By using the positive interactions identified, they generated a PSSM for each mouse PDZ domain based on the probability of amino acid on last five positions of the positive binding peptides. They were able to build PSSMs for 74 mouse PDZ domains. They benchmarked MDSM on interactions that weren't included in the training data and successfully predicted 48% of the positive interactions and 88% of the negative interactions. It is worth noting that the interaction data published in this study were widely used as training data in almost every PDZ domain prediction approach.

Chen*, et al.(60)* used a statistical model to predict interactions of mouse PDZ domains and peptides by using both sequences of PDZ domains and peptides. They constructed a multiple sequence alignment of mouse PDZ domains with available structures in the PDB. Using the structures of α1-syntrophin PDZ and heptapeptide GVKESLV as a reference, the model chooses position pairs in close proximity (<5.0 Å). They excluded any residue position in the PDZ domains that was not perfectly aligned. They obtained 38 interacting pairs involving 16 PDZ-domain binding pocket residues and 5 peptide ligand residues. They then generated a scoring matrix for each residue pair. This model was then fit with affinities or binary interaction data. Cross-validation tests (randomly assigning 12% of the domains and 8% of the peptides as the test set) of binary interaction showed AUC scores of 0.84, 0.91 and 0.87 for extrapolations to novel mouse peptides, novel mouse PDZ domains or both. They also tested the method on PDZ domains from other species. The AUC was 0.77 for *D. melanogaster* domains and 0.68 for *C. elegans* domains. Chen's method successfully used PDZ structure to determine the key binding sites, then used sequence information to build a predictor that performs at a high level on the cross-validations. The method has high accuracy predicting mouse PDZ domain interactions. On the other hand, further tests showed that this method was general for mouse PDZ domains, but performance for domains derived from more distantly related species was not very good.

Schillinger, *et al.* (61) developed the domain interaction footprint (DIF) method. This method was designed to predict protein-peptide interactions for SH3 and PDZ domains. For the PDZ domain part, it used experimentally tested data from four different PDZ domains, AF6, SNA1, ERBIN and N1P1. The properties of both binding and non-binding peptide sequences of PDZ domains were studied using a machine learning method. Parameters such as logP, Verloop parameters for volume, parameters for hydrophobicity, polarization, frequency of occurrence in elements of secondary structure, flexibility, and surface description were studied using correlation-based feature selection to select the best subset of features to classify the binding and non-binding peptides. The selected subset was then used to create DIFs for PDZ domains. For a PDZ domain, one DIF for the binding and another DIF for the non-binding peptides were created. A peptide is allocated to the DIF with the best score. On the ten-fold cross-validation, the DIF method scored an average AUC of 0.89, whereas the AUC scores of the single classifiers ranged from 0.84 to 0.93.

Kalyoncu, *et al.* (12) published a PDZ domain interaction prediction and classification method, which is based on the sequence features from mouse PDZ domains. The training dataset consisted of interaction information from 85 mouse PDZ domains and

181 peptides (59). They calculated the frequencies of consecutive two amino acids and three amino acids in the amino acid sequences of PDZ domains. Amino acids were arranged into seven groups according to their dipoles and volumes of the side-chains. Frequencies of consecutive two amino acids (bigram) and three amino acids (trigram) in the PDZ domain sequences were used as features to predict their binding partners. Five machine learning approaches, which were support vector machine (SVM), nearest neighbour, naïve Bayes, J48 and random forest, were trained using 10-fold cross-validation. In the end, the random forest method was chosen as it out-performed the other methods. For the trigram part, this method scored an AUC of 0.97 and 91.4% accuracy on their cross-validation. The method was also tested on an unpublished validation dataset and scored an accuracy of 79.8%. Kalyoncu's method showed a very good performance on predicting mouse PDZ domain interactions. However, on an unpublished dataset the accuracy of the prediction dropped to 79.8%. The tests performed on this method were all based on mouse PDZ domains. Performance for PDZ domains on other species was not tested.

The method of Shao, *et al. (62)* is another sequence-based PDZ domain interaction prediction approach. It used a novel regression framework that considers both positive and negative interaction data available for mouse PDZ domains. Shao's method, called semi-quantitative support vector regression (SVR), predicts the binding affinity of PDZ-peptide interactions from quantitative binding data and qualitative non-binding data. SVR is a well-known machine learning method of non-linear regression. They modified this method to take advantage of negative information they had available. They scored an average AUC of 0.86 in the benchmarking.

DomPep developed by Li*, et al.* (63) is a sequence-based prediction model for domain-mediated protein-protein interactions using SVM. They demonstrated their approach by building prediction models for PDZ and SH2 domains. They clustered PDZ domains in pairs by their positive binding peptides using a parameter called ligand-binding similarity (LBS). LBS is positively correlated to the number of shared binding peptides between two PDZ domains. Domain sequence identity (DSI) and PWM distance for each domain pair was also calculated. PDZ domains with LBS above an arbitrary set cut-off, in this case 0.7, were used to determine thresholds of DSI and PWM distance. PDZ domains with PWM distance and DSI greater than the corresponding thresholds were considered similar in specificity. A SVM algorithm was used to build prediction models for each PDZ domain using binding data from PDZ domains with similar specificity. DomPep scored an average AUC of 0.81 on a test set of 52 mouse PDZ domains and an average AUC of 0.84 on a test set of PDZ domains in other species than mouse. They also compared DomPep predictions for three PDZ domains in Scrib, PDZ-1, -2, and -3 with experimental data and scored AUC values of 0.90, 0.85 and 0.89 respectively. A web server is available for DomPep.

Kundu and Backofen (64) developed PdzPepInt, a cluster-based prediction method for PDZ-peptide interactions. They used the Markov clustering algorithm (MCL) to cluster human, mouse, fly and worm PDZ domains based on their sequence identify. PDZ domains with sequence identity greater than 50% were considered to have similar binding specificity and were therefore grouped together. Then they built a predictive model using a Gaussian kernel support vector machine with sequence-based and contact-based feature encoding. For the sequence-based approach, they used a strategy similar to DomPep (63), considering last five C-terminal residues of binding peptides of each PDZ domain group. For the contact-based approach, their approach was similar to Chen's method (60). The sequence-based approach covered 136 PDZ domains and scored an average AUC of 0.92 on a 5-fold cross-validation test. The contact-based approach covered 70 PDZ domains and scored an average AUC of 0.89. A web server is available for the public to access this method (65).

Overall, computational PDZ-peptide prediction methods show high-level performances. Some prediction methods limit their predictions on a limited number of species such as mouse and human. Other methods show a decrease in performance in test sets from other species or PDZ domains without binding data. Some of these prediction methods did not provide any convenient way, such as a web interface, for other researchers to use them without bioinformatics or programing background. Stiffler's MSDM method (59) provided PSSMs for 74 mouse PDZ domains. DomPep (63) and PdzPepInt (64, 65) both have web servers for users to submit their queries and make predictions. However, although DomPep claimed it can make predictions for any user-submit PDZ domain, this function wasn't working on their web server at the time of writing.

1.4 The SDR approach

The SDR approach was first developed by Brinkworth *et al.* (66) to predict substrate specificities of protein kinases. They identified specificity determining residues (SDR) by analysing the crystal structures of protein kinases. With the assumption of “proteins with similar SDR have similar specificity”, protein kinases that contain SDRs similar to the query protein were clustered. The binding peptides of these protein kinases are used to build PSSMs for the query protein and predict potential ligands. Later on, this approach has been proven successful not only for the prediction of phosphorylation sites (67, 68), but also prediction of peptide binding to MHC class II proteins (69, 70). In this work, we developed PreDiZ, which used the SDR approach to predict PDZ-peptide interactions using both structure and sequence information.

2 Datasets and Methodology

2.1 Datasets

2.1.1 PDZ domain structures (PDB)

The PDZ domain structures were obtained from the RCSB PDB (54). Structures that contained at least one PDZ domain and one ligand were extracted. In order to avoid low resolution and redundant structures, these structures were further filtered with the criteria of resolution higher than 2.4 Å and no more than 95% overall identity. PDZ complex structures that represent the canonical PDZ-peptide interaction, where the PDZ domain binds to a C-terminal peptide at the GLGF motif, were manually selected. After filtering, 22 complex structures were extracted (Table 1).

Table 1 list of PDZ domain structures extracted from the PDB.

PDB ID	Name
1BE9	The third PDZ domain from the synaptic protein PSD-95 in complex with a C-terminal peptide derived from CRIPT.
1IHJ	Crystal structure of the N-terminal PDZ domain of InaD in complex with a NorpA C-terminal peptide.
1KWA	Human CASK/LIN-2 PDZ domain.
1L6O	Xenopus Dishevelled PDZ domain.
1MFG	The structure of ERBIN PDZ domain bound to the carboxy-terminal tail of the ErbB2 Receptor.
1N7F	Crystal structure of the sixth PDZ domain of GRIP1 in complex with liprin C-terminal peptide.
1OBX	Crystal structure of the complex of PDZ2 of syntenin with an interleukin 5 receptor alpha peptide.
1OBY	Crystal structure of the complex of PDZ2 of syntenin with a syndecan-4 peptide.

1Q3P	Crystal structure of the Shank PDZ-ligand complex reveals a class I PDZ interaction and a novel PDZ-PDZ dimerization.
1RZX	Crystal Structure of a Par-6 PDZ-peptide complex.
1TP3	PDZ3 domain of PSD-95 protein complexed with KKETPV peptide ligand.
1TP5	Crystal structure of PDZ3 domain of PSD-95 protein complexed with a peptide ligand KKETWV.
1V1T	Crystal structure of the PDZ tandem of human syntenin with the TNEYKV peptide.
1W9E	Crystal structure of the PDZ tandem of human syntenin in complex with the TNEFYF peptide.
1W9O	Crystal structure of the PDZ tandem of human syntenin in complex with the TNEYYV peptide.
1W9Q	Crystal structure of the PDZ tandem of human syntenin in complex with the TNEFAF peptide.
1YBO	Crystal structure of the PDZ tandem of human syntenin with the syndecan peptide.
2AWW	Synapse associated protein 97 PDZ2 domain variant C378G with C-terminal GluR-A peptide.
2FNE	The crystal structure of the 13th PDZ domain of MPDZ.
2HE2	Crystal structure of the 3rd PDZ domain of human discs large homologue 2, DLG2.
2I04	X-ray crystal structure of MAGI-1 PDZ1 bound to the C-terminal peptide of HPV18 E6.
2I0L	X-ray crystal structure of Sap97 PDZ2 bound to the C-terminal peptide of HPV18 E6.
2I1N	Crystal structure of the 1st PDZ domain of human DLG3.
2IWP	12th PDZ domain of multiple PDZ domain protein MPDZ (CASP target).
2JIL	Crystal structure of 2nd PDZ domain of glutamate receptor interacting protein-1 (GRIP1).
2OPG	The crystal structure of the 10th PDZ domain of MPDZ.
2QBW	The crystal structure of PDZ-fibronectin fusion protein.
2QT5	Crystal structure of GRIP1 PDZ12 in complex with the Fras1 Peptide.
2R4H	Crystal structure of a C1190S mutant of the 6th PDZ domain of human membrane associated guanylate kinase.
2V90	Crystal structure of the 3rd PDZ domain of intestine- and kidney-enriched PDZ domain IKEPP (PDZD3).
2VPH	Crystal structure of the human protein tyrosine phosphatase, non-receptor type 4, PDZ domain.
2VRF	Crystal structure of the human beta-2-syntrophin PDZ domain.
3B76	Crystal structure of the third PDZ domain of human ligand-of-numb protein-X (LNX1) in complex with the C-terminal peptide from the coxsackievirus and adenovirus receptor.
3CBY	The Dvl2 PDZ domain in complex with the N1 inhibitory peptide.
3CH8	The crystal structure of PDZ-fibronectin fusion protein.

2.1.2 PDZ domain-peptide interactions

PDZ domain-peptide interactions were extracted from published studies, including three large-scale studies of PDZ domain interactions in *M. musculus* (59), *H. sapiens* (46) and *C. elegans* (71), and an online interaction database (72).

Stiffler*, et al.* (59) studied PDZ domain-peptide interactions from 157 mouse PDZ domains and 217 genome-encoded peptides using protein microarray screening. Equilibrium dissociation constant (K_D value) cut-off of 100 μM was used to determine positive and negative interactions. There were 726 positive interactions and 16142 negative interactions extracted, involving 84 PDZ domains and 217 peptides.

Tonikian*, et al.* (46) conducted a study on human PDZ domain specificity using phage-display experiments. We extracted 1473 positive PDZ domain-peptide interactions involving 54 human PDZ domains and 1283 peptides from this study. This study only provided positive interaction data, therefore no negative interaction data was collected.

Lenfant*, et al.* (71) performed yeast two-hybrid screens to study PDZ domain interactions in *C. elegans*. We collected 396 positive PDZ domain-peptide interactions comprising 47 *C. elegans* PDZ domains and 327 peptides.

The DOMINO database (72) is a manually curated database of protein interactions. Interactions involving a PDZ domain and that were not from the three large scale studies above were extracted from the database. We collected 1947 PDZ domain-peptide interactions from the DOMINO database.

2.1.3 Proteome-wide datasets

Reference proteomes are complete non-redundant proteome sets for selected well-studied organisms. In this study, reference proteomes of *A. thaliana*, *C. elegans*, *D. rerio*, *M. musculus* and *H. sapiens* were downloaded from the UniProt database (release 2014_04) (73) for proteome-wide analyses.

2.1.4 GO database

The GO database is a database of Gene Ontologies and annotation of genes and gene products (7). In this study, GO database version 2013-03-13 and the gene annotation file of version 3/5/2013 were used.

2.1.5 Kyoto Encyclopedia of Genes and Genomes pathway database

The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database is a database of manually curated biological pathway maps (74, 75). It is widely used in studies such as metabolism, genetic information processing and cellular processes.

2.1.6 Influenza A virus datasets

Highly pathogenic avian influenza (HPAI) A virus protein sequences were downloaded from the NCBI Influenza Virus Resource database (76) with search conditions of “Type:A, Host: Human, Full-length only and collapse identical sequences”. There were a total of 33604 virus protein sequences, including 1182 protein sequences from H5N1 and 218 protein sequences from H7N9.

2.2 SDR selection based on structural alignments

2.2.1 PDZ structure-based alignments

PDZ domain structures were extracted from the PDZ domain-peptide complexes from the PDB. These PDZ domain structures were then aligned using PROMALS3D (77), a multiple sequence alignment tool using 3D structural information to improve sequence alignment quality. Then an HMM profile (referred to as the PDZ HMM profile) was built from the structure alignments using the HMMER 3.0 package. The HMM profile of PDZ domain structures were later used to determine SDRs.

2.2.2 Four-parameter-test to identify SDRs

SDRs were determined for each position of the PDZ binding motif (PBM) independently. We designed a four-parameter-test to look at every column of the consensus sequence of the PDZ HMM profile. The four parameters were: maximum contacting distance (MCD), average column similarity (ACS), proportion of interacting residues (PIR) and BLOSUM matrix (BLOSUM). Maximum contacting distance is the distance between PDZ domain residue’s side-chain atoms to the closest side-chain atoms of their binding peptides. Residues with distance below or equal to the MCD were considered as contacting residues. BLOSUM corresponds to the BLOSUM substitution matrix used to determine similar residues. Residues with a positive score on the selected BLOSUM matrix were considered similar. ACS is the proportion of similar residues on an aligned column of the alignment. PIR is the proportion of contacting residues on an aligned column of the alignment. For each position of the PBM (i.e. the last five residues at the C-terminus), if a column satisfied the condition of all four parameters, it was selected as a SDR.

The SDRs were named based on their positions on the consensus sequence of the PDZ HMM profile. Two conserved motifs on the consensus sequence of the PDZ HMM profile were selected as marker motifs. They were the GLGF motif located in the βB strand and the GD motif located on the βD. The position of a SDR was marked using a marker + offset format. For example, if a SDR was on the column two residues after GD, it was marked as GD+2.

2.3 Implementation of the SDR approach

2.3.1 Customised PDZ interaction database

Similar to the SDR method on Predikin, a purpose-built protein-peptide interaction database of PDZ domain interactions was needed. The database was structured in a fashion that the relationship between SDRs and PBM residues could be easily obtained (Figure 1). Hence, each PDZ domain in the database was scanned with the PDZ HMMER profile and each residue on the PDZ domain was encoded into the motif + offset format. There were in total 20494 interactions in this database.

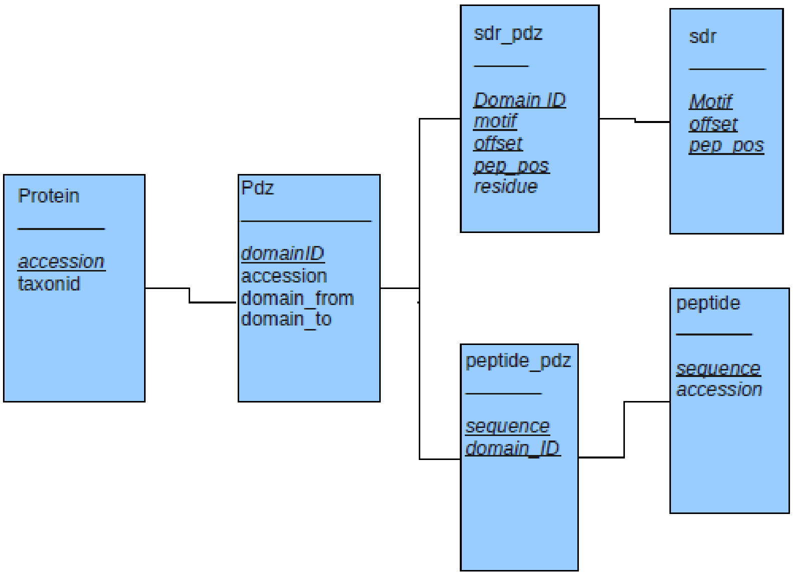

Figure 1 Database schema of the customised PDZ interaction database

2.3.2 PreDiZ prediction module

In general, PreDiZ predicts binding between PDZ domains and five-residue PBMs by establishing a correlation between the SDRs in the query PDZ domains and the SDRs associated with known PDZ domain interactions. After the user submits a query, which includes at least one PDZ protein and one five-residue peptide sequence (if the submitted peptide is longer than 5-residues, only the last 5 residues are used), PreDiZ executes the following steps.

2.3.2.1 Locate SDRs from query protein

Firstly, PreDiZ identifies all the SDRs for each of the positions of the PBM. The PDZ HMM profile built from structural alignments is used to locate PDZ domain(s) on the query protein sequence by hmmsearch function from HMMER with the condition of E-value less than 0.01. If a PDZ domain is found, SDRs are identified from the sequence alignment of the HMM consensus sequence and the query protein. Residues that align to the SDR positions of the consensus sequence, correspond to the SDRs for the query PDZ domain.

2.3.2.2 Query the customised database

Secondly, PreDiZ queries the customised PDZ interaction database for known PBM sequences associated with PDZ domains that share similar SDRs with the query protein at each binding position. The term similar SDR is defined as amino acid residues that score positively in the selected BLOSUM substitution matrix. Because PreDiZ treats each position of the PBM independently, five sets of amino acids are collected for five positions of PBM and are used to build a PSSM or a naive Bayesian classifier.

2.3.2.3 Scoring methods

2.3.2.3.1 PSSM scoring method

To construct a PSSM, PreDiZ first calculates the amino-acid frequencies of positive interactions for each position from the previous step. A 20 columns (amino acids) x 5 rows (5 PBM positions) PSSM based on the amino acid frequency results is built using equation (i). The query peptide is scored with the PSSM using equation (ii).

$$W_{i,j} = \frac{F_{i,j}+\sqrt{\frac{n}{20}}}{\mathrm{n}+\sqrt{n}} \qquad \text{(i)}$$

$$score = \frac{\sum w_{i,j} - \sum w_{min,j}}{\sum w_{max,j} - \sum w_{min,j}} \quad \text{(ii)}$$

where F is the frequency of amino-acid i for PBM position j and n is the number of sequences used to calculate the frequency. Note that the PSSM scoring method only uses positive binding data to make predictions.

2.3.2.3.2 Naive Bayesian classifier scoring method

To utilise both positive and negative interaction data in PreDiZ's prediction, the naive Bayesian classifier is used to build the classification models for query PDZ domains from the results of the previous step (2.3.2.2). For each position of the PBM, amino acids are categorized into binding or non-binding class, and a vector x = (x_1, …, x_n) where x_1, …, x_n are the amino acids. Based on Bayesian's theorem, we get

$$P(C|X_1, \dots, X_n) = \frac{P(C)P(X_1, \dots X_n|C)}{P(X_1, \dots X_n)}$$

where C refer to the binding or non-binding class and $P(C|X_1, \dots, X_n)$ is the posterior probability of the class C. The posterior probability corresponds to $P(C|X_1, \dots, X_n)$ when C = binding is used as the score of the query peptide.

2.4 Performance evaluation and optimisation

The performance of PreDiZ was evaluated extensively by 4-, 6-, 8- and 10-fold cross-validation. Four standard measurements, including sensitivity (*Sn*), specificity (*Sp*), accuracy (*Ac*), and the Matthew's correlation coefficient (*MCC*) were defined as follows:

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$Ac = \frac{TP + TN}{TP + FP + TN + FN} \quad MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

The receiver operating characteristic (ROC) curves were drawn for each cross-validation. The area under the ROC curve (AUC) was calculated as a measurement of the performance.

2.4.1 GO and KEGG enrichment analysis of PDZ-binding proteins

In this study, a statistical approach based on hypergeometric distribution was employed to perform enrichment analysis of PDZ-binding proteins in *H. sapiens* (78). We compared the predicted PDZ-binding proteins (group S) against the proteome (group W) to find out if a GO/KEGG term t is enriched in the predicted PDZ-binding proteins. The following terms were defined: N was the total number of proteins in group W annotated by GO/KEGG; n was the number of proteins in group W annotated by GO/KEGG term t; M was the total number of proteins in group S annotated by GO/KEGG; m was the number of proteins in group S annotated by GO/KEGG term t. Hence,

$$\text{Enrichment ratio (E} - \text{ratio)} \ = \frac{m}{M} / \frac{n}{N}$$

$$p - value = \sum_{m'=m}^{n} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} (\text{Enrichment ratio} \ \geq \ 1)$$

or

$$p - value = \sum_{m'=0}^{m} \frac{\binom{M}{m\prime}\binom{N-M}{n-m\prime}}{\binom{N}{n}} \ (\text{Enrichment ratio} < \ 1)$$

Here, we considered only the over representation of GO/KEGG groups with enrichment ratio greater than 1.

3 Results

3.1 Computational approach to predict PDZ domain-peptide interactions (PreDiZ)

3.1.1 Construction of PreDiZ for the prediction of PDZ domain interactions

In this work, we developed a computational tool called PreDiZ to predict peptide-protein interactions involving PDZ domains based on the SDR approach (66-68, 79). The procedures used for data preparation and prediction making are shown in Figure 2. PreDiZ is able to predict the binding between any PDZ domain and any 5-residue-long peptide. It will return a score between 0 and 1 for each predicted binding pair as the result.

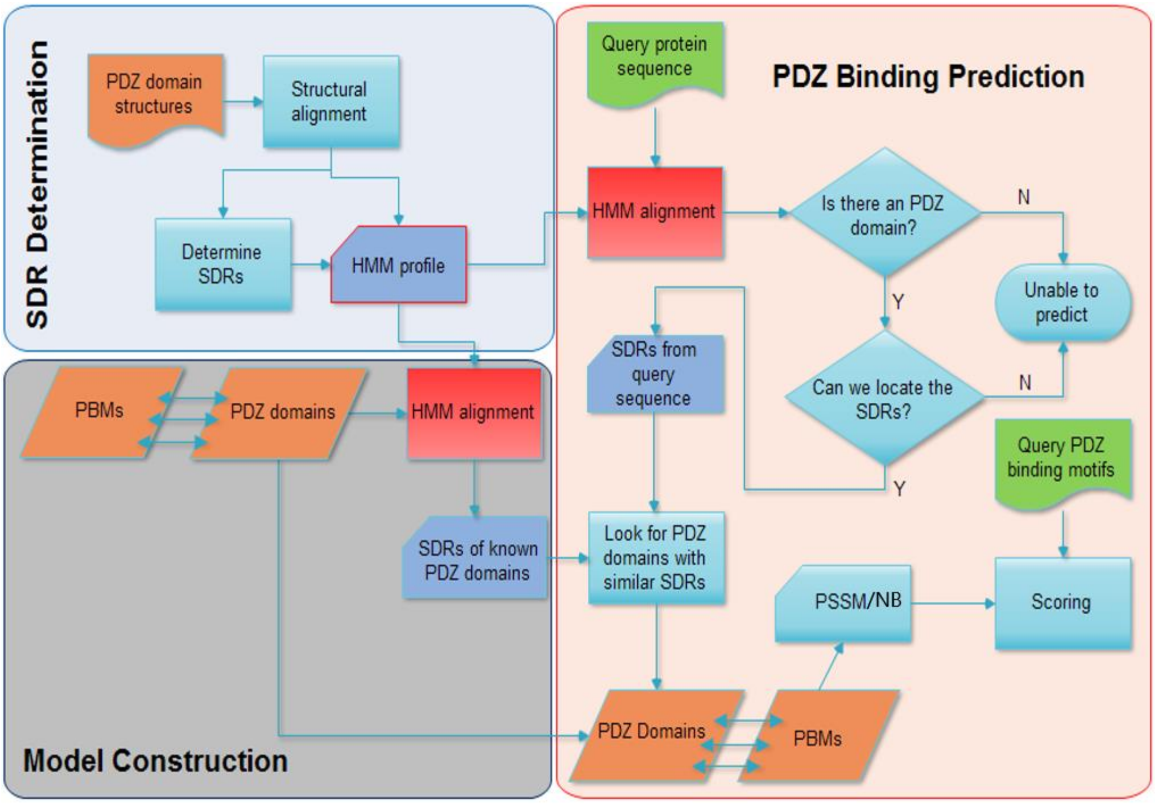

Figure 2 A schematic representation of PreDiZ. ***SDR determination****: PDZ domain structures were then aligned using PROMALS3D. Then an HMM profile was built from the structure alignments using the HMMER 3.0 package. The HMM profile of PDZ domain structures were later used to determine SDRs.* ***PDZ binding prediction****: PreDiZ identifies all the SDRs for each of the positions of the PBM. The PDZ HMM profile built from structural alignments is used to locate PDZ domain(s) on the query protein sequence. If a PDZ domain is found, SDRs are identified from the sequence alignment of the HMM consensus sequence and the query protein. Residues that align to the SDR positions of the consensus sequence, correspond to the SDRs for the query PDZ domain.* ***Model construction****: PreDiZ queries the customised PDZ interaction database for known PBM sequences associated with PDZ domains that share similar SDRs with the query protein at each binding position. Because PreDiZ treats each position of the PBM independently, five sets of amino acids are collected for five positions of PBM and are used to build a PSSM or a naive Bayesian classifier. The PSSM or the naive Bayesian classifier is used to score PDZ domain-binding motifs.*

3.1.2 Performance evaluation and optimisation

3.1.2.1 SDR selection and cross-validations

SDR selection is critical for the performance of PreDiZ. In this work we performed extensive tests to select the best SDR set for PDZ binding prediction. Firstly, structural alignment was performed with 22 PDZ domain structures from the PDB. From this alignment, we determined 128 groups of SDRs according to combinations of four parameters: average column similarity (ACS) of 0.2, 0.4, 0.6, 0.8, proportion of interacting residues (PIR) of 0.2, 0.4, 0.6, 0.8, maximum contacting distance (MCD) of 5 Å, 6 Å, 8 Å, 10 Å, and BLOSUM matrix (BLOSUM) of BLOSUM62, BLOSUM80. The SDR set with the best area-under-ROC curves (AUC) from the cross-validation tests was used in the final version of PreDiZ. Both the position-specific scoring matrix scoring method (PreDiZ-PSSM) and the naive Bayesian scoring method (PreDiZ-NB) were tested. For the PreDiZ–PSSM method, PreDiZ used only the positive binding data of PDZ interactions and performed best when used the SDR selected using from ACS = 0.8, PIR = 0.8, MCD = 10 Å and BLOSUM = BLOSUM62 parameters (Table 2, Figure 3). The resulting AUC's values were 0.812, 0.811, 0.812 and 0.812 for 4-, 6-, 8- and 10-fold cross-validation, respectively (Figure 4). For the PreDiZ–NB method, which using both positive and negative binding data of PDZ interactions, the same SDR set gave the best results with AUC of 0.885, 0.883, 0.879 and 0.886 for 4-, 6-, 8- and 10-fold cross-validation, respectively (Figure 5). The performance of the 10-fold validations was used for determining the cut-offs. If a pair of PDZ protein and PBM gave a score higher than the cut-off value, it was considered positive, otherwise negative. Three levels of thresholds were selected (Table 3). The High threshold was set as default and was used in the following analyses unless stated otherwise.

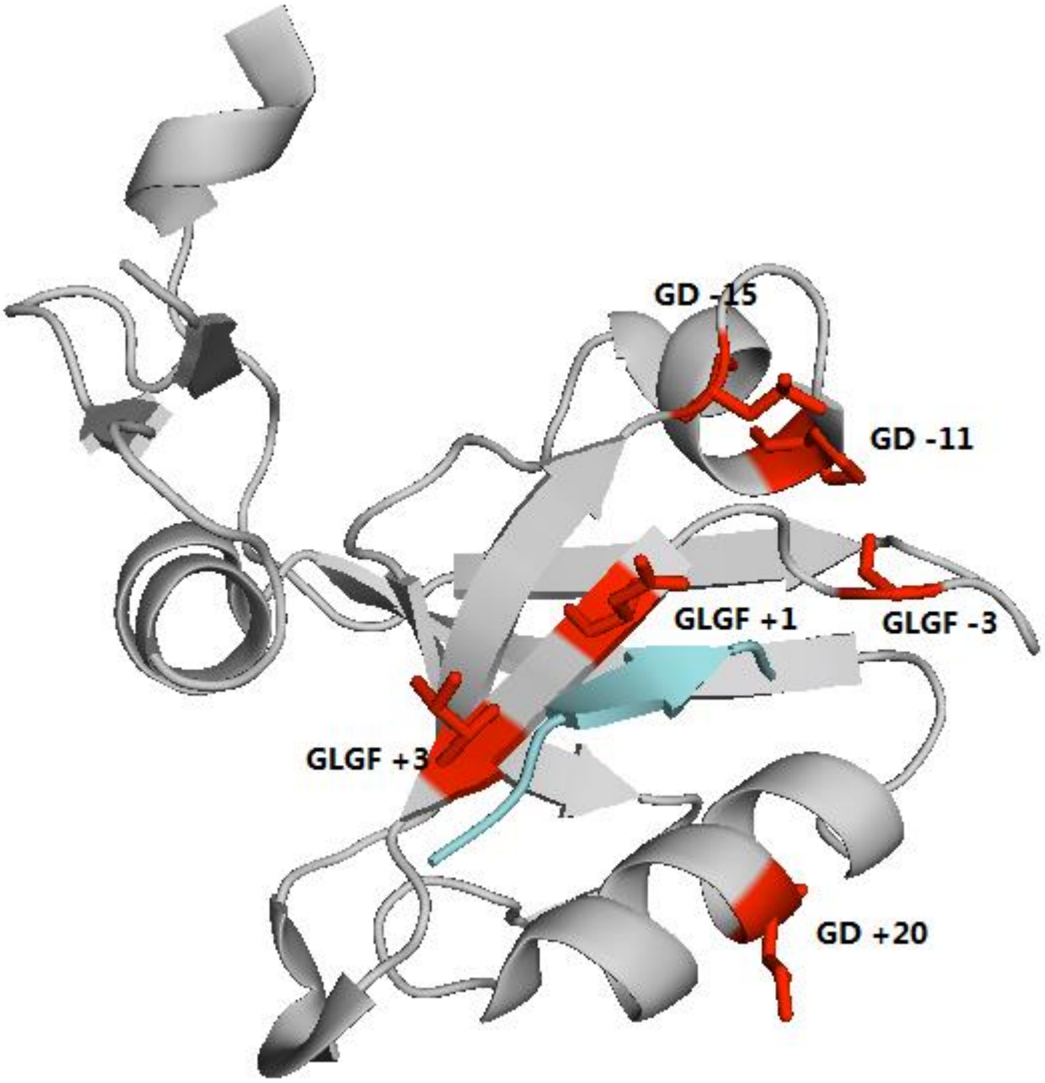

Figure 3 3D structure of a representative PDZ domain (PDB ID 1TP5) with SDRs highlighted in red. The PDZ domain is shown in grey and the peptide is shown in cyan.

Table 2 SDR set used in the PreDiZ

Position	SDR in motif + offset format				
0	GLGF -3	GLGF 1	GD -15	GD -11	GD 20
-1	GLGF -3	GLGF 1	GD -15		
-2	GLGF 1	GLGF 3			
-3	GLGF 1	GLGF 3			
-4	GLGF 1	GLGF 3			

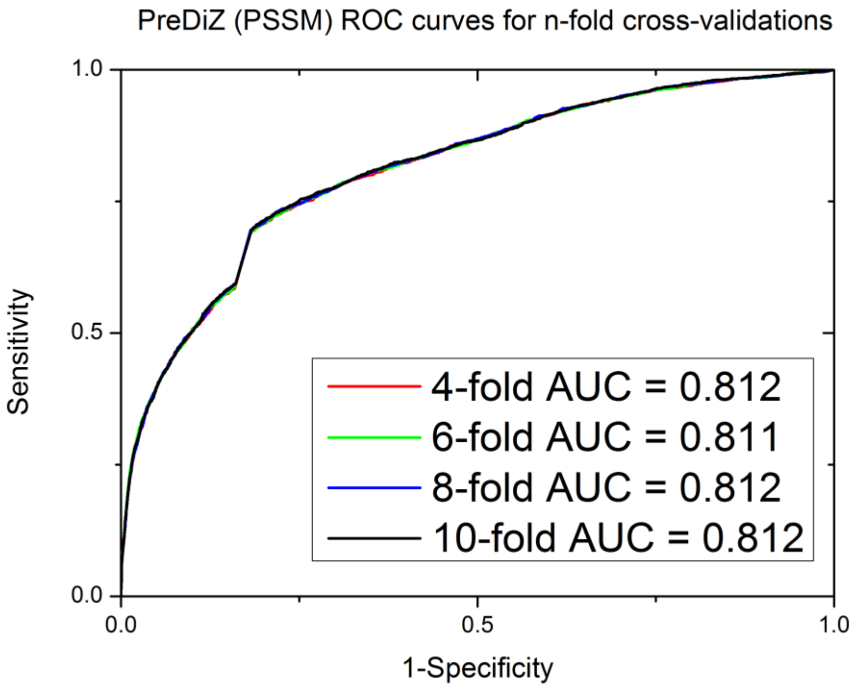

Figure 4 ROC curves of cross-validations for PreDiZ-PSSM

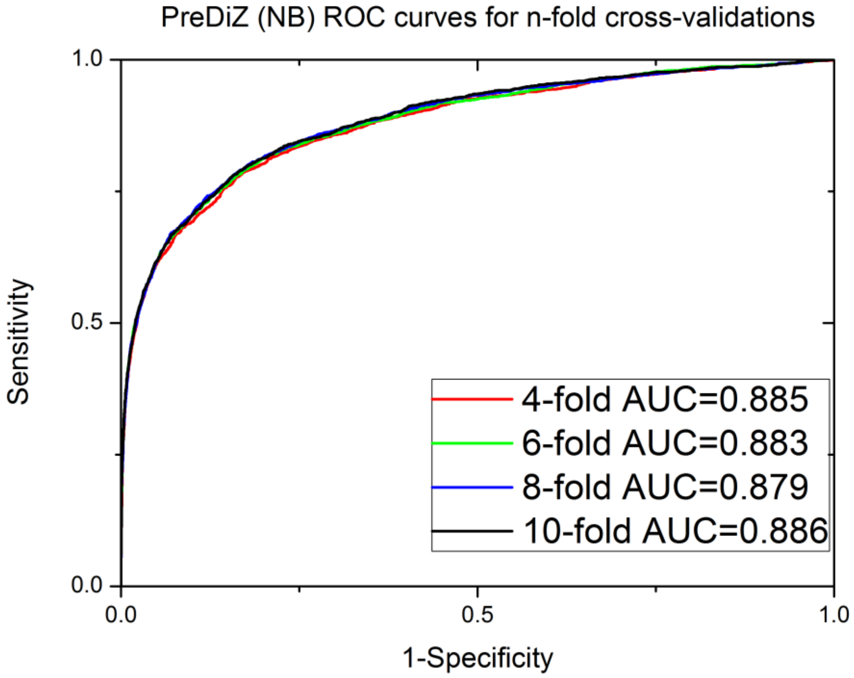

Figure 5 ROC curves of cross-validations for PreDiZ-NB.

Table 3 Performance evaluations of PreDiZ

Predictor	Threshold	*ACC* (%)	*TPR* (%)	*SPC* (%)	*MCC*
PreDiZ - PSSM	High	88.35	33.14	97.06	0.4047
	Medium	84.60	50.34	90.00	0.3825
	Low	80.23	69.54	81.91	0.4079
PreDiZ - NB	High	91.60	55.98	96.89	0.5919
	Medium	87.48	70.45	90	0.5291
	Low	83.96	76.9	85	0.4925

ACC stands for accuracy. TPR stands for true positive rate or sensitivity. SPC stands for specificity. MCC stands for the Matthew's correlation coefficient.

3.1.2.2 Performance comparison with published methods

3.1.2.2.1 Performance evaluation on unpublished mouse interaction data

We compared PreDiZ with three state-of-the art PDZ interaction prediction tools, MDSM, DomPep and PdzPepInt. An independent test set consisting of interactions between 74 mouse PDZ domains and 48 peptides was used in this test. There were in total 493 positive interactions and 3059 negative interactions. From this test set, interactions involving 50 PDZ domains that were shared in all the prediction methods were selected. These interactions were carefully excluded from the training data of PreDiZ. The comparison results (*Table 4*) showed that the performance of PreDiZ-PSSM achieved an AUC of 0.78, which was better than MDSM's AUC of 0.74, but not as good as PdzPepInt and DomPep which scored 0.85 and 0.84, respectively. PreDiZ-NB performed worst among all the prediction methods in the test, with an AUC of 0.67. PreDiZ-NB also scored the worst true-positive/false-positive (TP/FP) ratios of 0.36, 0.25 and 0.24 for high, medium and low threshold respectively, which suggested this method made more false positive predictions than the others. PreDiZ-PSSM achieved the best TP/FP ratio of 1.12 at high threshold, but the true positive rate (TPR) was 0.36, which was around half of the TRP of PdzPepInt or DomPep.

Table 4 Performance evaluation on the independent mouse test set

Predictor	Threshold	TPR	FPR	TP/FP	AUC
PreDiZ - NB	High	0.43	0.2	0.36	0.67
	Medium	0.69	0.45	0.25	
	Low	0.8	0.54	0.24	
PreDiZ - PSSM	High	0.36	0.05	1.11	0.78

	Medium	0.57	0.21	0.51	
	Low	0.79	0.32	0.45	
MDSM		0.55	0.17	0.55	0.74
PdzPepInt		0.67	0.14	0.87	0.85
DomPep		0.66	0.15	0.79	0.84

TRP is true positive rate, FPR is false positive rate, TP/FP is true-positive/false-positive ratio.

In another experiment, we tested PreDiZ on a test set of 20 mouse PDZ domain-peptide interactions derived from PDZbase (80). This test set has been used to compare performance of MDSM and PdzPepInt (64). Both PreDiZ-PSSM and PreDiZ-NB were tested using the high threshold. The results showed PdzPepInt performed best by successfully predicting 14 out of 20 interactions. While PreDiZ-NB was the worst predictor, which only managed to predict one interaction. PreDiZ-PSSM predicted 8 interactions, which was better than MDSM's 4 interactions, but still not as good as the PdzPepInt's result. All prediction scores for the validated set were listed in Table 5.

Table 5 PdzPepInt, MDSM and PreDiZ scores for validated set.

PDZ domain	Peptide	PdzPepInt	MDSM	PreDiZ - NB	PreDiZ - PSSM
Cipp-(3/10)	IESDV	**0.44**	-0.7	0.86	0.73
Cipp-(3/10)	LESEV	**0.3**	-0.62	0.68	0.71
Cipp-(3/10)	QQSNV	**0.29**	-0.78	0.67	0.51
Cipp-(3/10)	KEYYV	**0.51**	-0.34	0.42	0.66
Dvl1-(1/1)	SETSV	-1.27	-0.74	0.22	0.51
Pdlim5-(1/1)	DITSL	-0.24	-0.15	0.29	0.42
Erbin-(1/1)	LDVPV	**0.99**	0.61	0.11	0.42
Magi-2-(5/6)	KESSL	**1.76**	0.19	0.005	0.21
MUPP1-(10/13)	IATLV	**1**	0.46	0.49	0.64
MUPP1-(10/13)	GKDYV	**1**	**1.68**	0.03	0.41
NHERF-1-(1/2)	FDTPL	**1.06**	0.01	0.99	0.89
LIN-7A-(1/1)	IESDV	**0.33**	0.29	0.83	0.77
Lin7c-(1/1)	IESDV	**0.33**	1	0.83	0.77
ZO-3-(1/3)	GKDYV	**0.99**	0.09	0.35	0.4
a1-syntrophin-(1/1)	VLSSV	-1.47	0.16	0.49	0.47
PSD95-(1/3)	LQTEV	**0.38**	**1.41**	0.79	0.76
PSD95-(1/3)	NETVV	-1.35	**1.19**	0.73	0.85
PSD95-(1/3)	GETAV	-1.32	**1.23**	0.74	0.86
PSD95-(1/3)	EESSV	-2.23	0.77	0.17	0.62
PSD95-(1/3)	RTTPV	**1**	0.61	0.73	0.81

Scores marked red are the true positive interactions.

3.1.2.2.2 Prediction comparison on novel PDZ domains

Both DomPep and PdzPepInt claimed that they are able to predict interactions for unknown PDZ domains, as does PreDiZ. Therefore, we designed an experiment to test the performance of making prediction on unknown PDZ domains for these prediction methods. At the time of writing DomPep webserver returned an error on any user-submitted PDZ domains. Hence, we could only make comparisons between PdzPepInt and PreDiZ. We used 251 PDZ domain-peptide interactions from *C. elegans*, which the PDZ domains were not modelled in PdzPepInt and these interactions were excluded from the training set of PreDiZ. PdzpepInt was able to correctly predict 4 interactions, while PreDiZ-NB with high threshold predicted 8 and PreDiZ-PSSM predicted 5. Both methods therefore showed a similar level of performance on predicting unknown PDZ interactions (Table 6).

Table 6 Performance evaluation between PreDiZ and PdzPepInt on unknown PDZ domains

Predictor	**Threshold**	**True positive interactions**	**TPR**
PdzPepInt		4	0.02
PreDiZ - NB	High	8	0.03
	Medium	17	0.07
	Low	23	0.09
PreDiZ - PSSM	High	5	0.02
	Medium	8	0.03
	Low	18	0.07

3.2 Proteome-wide analysis

3.2.1 Proteome-wide PDZ domain-peptide interaction predictions

Proteome-wide prediction of PDZ domain mediated interactions in *A. thaliana*, *C. elegans*, *D. rerio*, *M. musculus* and *H. sapiens* were performed. We retrieved 26577, 20310, 25418, 21966 and 20661 proteins from *A. thaliana*, *C. elegans*, *D. rerio*, *M. musculus* and *H. sapiens* reference proteomes, respectively. We scanned these proteomes with PreDiZ to identified PDZ domains and predict their binding partners. The results showed PDZ domains exist in 0.02%, 0.24%, 0.90%, 0.60% and 0.68% of the proteins in *A. thaliana*, *C. elegans*, *D. rerio*, *M. musculus* and *H. sapiens* proteomes, respectively (Table 7). Around one third of the proteome was predicted with the ability to bind PDZ domains for *D. rerio*, *M. musculus* and *H. sapiens*, while 2.82% for *D. melanogaster* and 6.10% for *C. elegans*.

Table 7 Proteomic prediction of 5 species PDZ interactions

Organism	**Proteome**	**PDZ**		**PDZ-binding (PSSM)**	
		Num.	***Per.***	***Num.***	***Per.***

A. thaliana	26577	*6*	0.02%	*749*	2.82%
C. elegans	20310	*49*	0.24%	*1238*	6.10%
D. rerio	25418	*229*	0.90%	*9144*	35.97%
M. musculus	21966	132	0.60%	7589	34.55%
H. sapiens	20661	140	0.68%	7212	34.91%

3.2.2 GO/KEGG enrichment analysis of human PDZ domains

Using the predicted *H. sapiens* PDZ interactions, a PDZ interaction network was constructed with 140 PDZ proteins and 7212 PDZ binding proteins. The top 10 interconnected PDZ domains and PDZ binding proteins are listed (Figure 6). Hypergeometric distribution tests (p-value < 1E-3) were performed to analyse the enrichment of GO terms for proteins in the network. The results show that the network is significantly enriched with proteins locating to the membrane and involving a number of biological processes such as small GTPase-mediated signal transduction, steroid metabolic process and xenobiotic metabolic processes (Table 8). Similar statistical tests were also performed to test the enrichment of KEGG pathways. The results suggested the network proteins were associated in pathways such as metabolism of xenobiotics by cytochrome P450, drug metabolism by cytochrome P450 and epithelial cell signalling in *H. pylori* infection (Table 4). These results are consistent with our knowledge of PDZ domains. For example, PDZ domains are known to interact with xenobiotic transporters (81). The involvement of PDZ domains in other top-ranking biological processes, such as small GTPase mediated signal transduction (82, 83), steroid metabolic process (84, 85) and regulation of proteolysis (86, 87), are well documented.

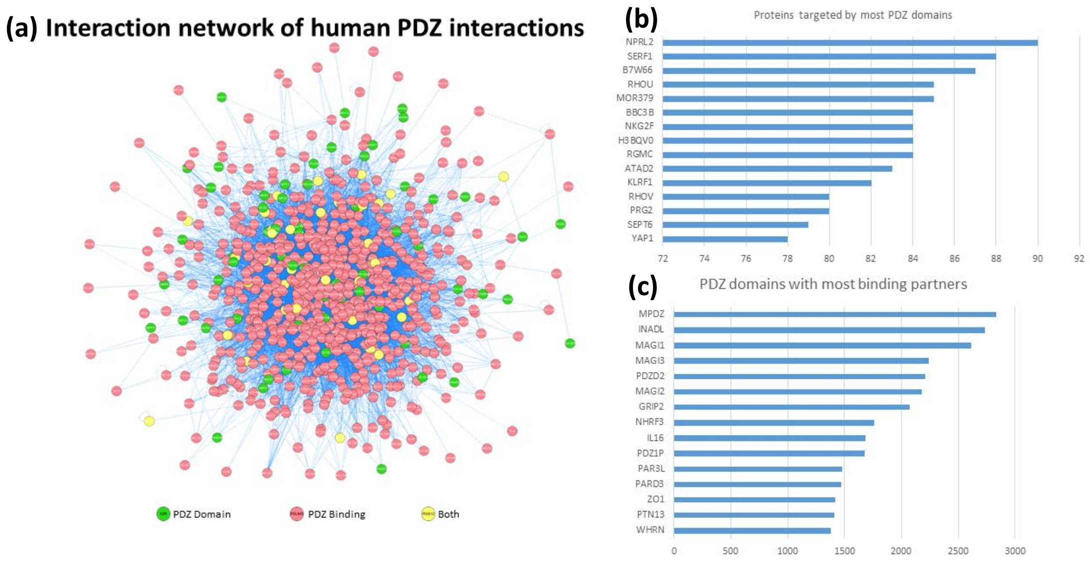

Figure 6 Interaction network of human PDZ interactions and statistics. (a) Interaction network of human PDZ interactions. (b) Proteins targeted by most PDZ domains. (c) PDZ domain proteins with most binding partners.

Table 8 The most enriched GO terms in the H. sapiens PDZ interaction network (p-value < 1E-3)

Term Description	**PDZ Domain**		**Proteome**		**E-ratio**	***p*-value**
	Num.	***Per.***	***Num.***	***Per.***		
The most enriched biological processes						
small GTPase mediated signal transduction (GO:0007264)	186	2.37%	326	1.79%	1.32	2.10E-07
regulation of small GTPase mediated signal transduction (GO:0051056)	90	1.15%	147	0.81%	1.42	6.62E-06
steroid metabolic process (GO:0008202)	42	0.54%	61	0.34%	1.60	4.13E-05
xenobiotic metabolic process (GO:0006805)	81	1.03%	136	0.75%	1.38	7.54E-05
regulation of proteolysis (GO:0030162)	29	0.37%	40	0.22%	1.68	1.55E-04
glycosaminoglycan biosynthetic process (GO:0006024)	26	0.33%	35	0.19%	1.72	1.77E-04
auditory receptor cell differentiation (GO:0042491)	10	0.13%	10	0.05%	2.32	2.19E-04
protein homooligomerization (GO:0051260)	85	1.08%	150	0.82%	1.32	5.27E-04
neural tube formation (GO:0001841)	11	0.14%	12	0.07%	2.13	6.84E-04
cellular response to zinc ion (GO:0071294)	11	0.14%	12	0.07%	2.13	6.84E-04
regulation of protein stability (GO:0031647)	20	0.26%	27	0.15%	1.72	1.08E-03
positive regulation of phosphorylation (GO:0042327)	17	0.22%	22	0.12%	1.79	1.18E-03
daunorubicin metabolic process (GO:0044597)	8	0.10%	8	0.04%	2.32	1.18E-03
sulfation (GO:0051923)	8	0.10%	8	0.04%	2.32	1.18E-03
doxorubicin metabolic process	8	0.10%	8	0.04%	2.32	1.18E-03

(GO:0044598)						
fertilization (GO:0009566)	21	0.27%	29	0.16%	1.68	1.32E-03
ATP hydrolysis coupled proton transport (GO:0015991)	21	0.27%	30	0.16%	1.63	2.62E-03
transferrin transport (GO:0033572)	21	0.27%	30	0.16%	1.63	2.62E-03
The most enriched molecular functions						
PDZ domain binding (GO:0030165)	55	0.70%	75	0.41%	1.70	9.96E-08
glutathione transferase activity (GO:0004364)	20	0.26%	25	0.14%	1.86	1.83E-04
sulfotransferase activity (GO:0008146)	21	0.27%	27	0.15%	1.81	2.58E-04
aldo-keto reductase (NADP) activity (GO:0004033)	9	0.11%	9	0.05%	2.32	5.08E-04
oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor (GO:0016655)	8	0.10%	8	0.04%	2.32	1.18E-03
alditol:NADP+ 1-oxidoreductase activity (GO:0004032)	8	0.10%	8	0.04%	2.32	1.18E-03
polysaccharide binding (GO:0030247)	10	0.13%	11	0.06%	2.11	1.47E-03
delayed rectifier potassium channel activity (GO:0005251)	24	0.31%	35	0.19%	1.59	2.04E-03
cytochrome-b5 reductase activity, acting on NAD(P)H (GO:0004128)	7	0.09%	7	0.04%	2.32	2.74E-03
cyclin-dependent protein serine/threonine kinase activity (GO:0004693)	22	0.28%	32	0.18%	1.60	2.97E-03
calcium channel regulator activity (GO:0005246)	14	0.18%	18	0.10%	1.81	2.98E-03
The most enriched cellular components						
apical plasma membrane (GO:0016324)	115	1.47%	210	1.15%	1.27	3.99E-04
clathrin adaptor complex (GO:0030131)	14	0.18%	17	0.09%	1.91	1.10E-03
plasma membrane (GO:0005886)	1556	19.85%	3436	18.88%	1.05	1.91E-03
Golgi lumen (GO:0005796)	46	0.59%	77	0.42%	1.39	2.33E-03
Golgi apparatus (GO:0005794)	321	4.09%	667	3.66%	1.12	4.14E-03
nuclear speck (GO:0016607)	81	1.03%	150	0.82%	1.25	4.41E-03
Golgi membrane (GO:0000139)	219	2.79%	447	2.46%	1.14	6.12E-03
PCAF complex (GO:0000125)	6	0.08%	6	0.03%	2.32	6.37E-03
early endosome (GO:0005769)	71	0.91%	132	0.73%	1.25	8.28E-03
endocytic vesicle (GO:0030139)	23	0.29%	36	0.20%	1.48	9.47E-03
heterotrimeric G-protein complex (GO:0005834)	23	0.29%	36	0.20%	1.48	9.47E-03
adherens junction (GO:0005912)	23	0.29%	36	0.20%	1.48	9.47E-03

Table 9 The most enriched KEGG pathways in the H. sapiens PDZ interaction network (p-value < 1E-3)

Term Description	**PDZ Domain**		**Proteome**		**E-ratio**	**p-value**
	Num.	**Per.**	**Num.**	**Per.**		
Metabolism of xenobiotics by cytochrome P450(path:hsa00980)	54	1.98%	76	1.23%	1.61	1.80E-06
Drug metabolism - cytochrome P450(path:hsa00982)	47	1.72%	68	1.10%	1.56	2.73E-05
Epithelial cell signalling in Helicobacter pylori infection(path:hsa05120)	46	1.68%	68	1.10%	1.53	7.70E-05

Regulation of autophagy(path:hsa04140)	25	0.92%	32	0.52%	1.77	9.55E-05
Endocrine and other factor-regulated calcium reabsorption(path:hsa04961)	32	1.17%	49	0.79%	1.48	2.32E-03
Synaptic vesicle cycle(path:hsa04721)	39	1.43%	64	1.04%	1.38	5.10E-03
Huntington's disease(path:hsa05016)	98	3.59%	182	2.95%	1.22	5.18E-03
Collecting duct acid secretion(path:hsa04966)	19	0.70%	27	0.44%	1.59	5.42E-03
Cardiac muscle contraction(path:hsa04260)	45	1.65%	76	1.23%	1.34	5.89E-03

3.3 PDZ-binding motifs in influenza virus A

3.3.1 Interaction predictions of human PDZ domains and virus proteins

In 2013, a novel avian-origin influenza A virus H7N9 has emerged in China, infecting over 160 patients. Highly pathogenic avian influenza (HPAI) A virus, such as H5N1, are known to contain a PDZ domain binding motif at the C-terminus of the NS1 protein. This motif is able to affect virulence, but not the replication of the virus (88, 89). However, the PDZ domain-binding motif (PBM) is not found in the recent H7N9's NS1 proteins, but is found in the NA proteins instead (90).

To investigate how these changes affect the virus regulating human cells through PDZ binding, PrediZ was used to predict the binding between human PDZ proteins from the previously generated human PDZ interaction network and the HPAI H5N1's NS1 proteins and HPAI H7N9's NA proteins. The PDZ domains, that were predicted to bind the viral proteins, were mapped to the human PDZ interaction network to find their binding partners. The results revealed that all H7N9's NA proteins and 96.72% H5N1's NS1 proteins were predicted to interact with at least one PDZ domain. H7N9's NA proteins are targeted by 20 different human PDZ domains, while H5N1's NS1 are targeted by 18. The number of human proteins that interact with these PDZ domains is similar, with 3366 proteins for H7N9's NA and 4238 for H5N1's NS1 (Table 10).

Table 10 Predicted PDZ interaction in HAPI influenza A viral proteins

Protein	Hits	Total	Percentage	PDZ domains	PDZ binding proteins
H7N9 NA	26	26	100.00%	20	3366
H5N1 NS1	118	122	96.72%	18	4238

** hit means the protein is predicted to interact with at least one PDZ domain.*

To further validate how H5N1's NS1 proteins and H7N9's NA proteins affect the virus regulation human cells by the virus, we analysed the PDZ binding proteins related to these two viral proteins by comparing their enrichment of biological process of GO ontology. The results showed two sets of protein enriched in similar biological processes by sharing 6 out of 10 most enriched biological processes (Table 11, Table 12).

Table 11 GO enrichment analyse of H5N1 NS1 related PDZ binding proteins

Term Description	**PDZ**		**Proteome**		**E-ratio**	**p-value**
	Num.	**Per.**	**Num.**	**Per.**		
Top 10 most enriched biological processes						
transmembrane transport (GO:0055085)	22	9.36%	449	2.47%	3.80	9.44E-08
glutamate receptor signaling pathway (GO:0007215)	5	2.13%	13	0.07%	29.79	4.07E-07
actin crosslink formation (GO:0051764)	4	1.70%	8	0.04%	38.73	1.82E-06
Wnt receptor signaling pathway, calcium modulating pathway (GO:0007223)	4	1.70%	11	0.06%	28.17	8.32E-06
muscular septum morphogenesis (GO:0003150)	3	1.28%	4	0.02%	58.09	8.42E-06
glossopharyngeal nerve morphogenesis (GO:0021615)	3	1.28%	4	0.02%	58.09	8.42E-06
G-protein coupled receptor signaling pathway (GO:0007186)	15	6.38%	301	1.65%	3.86	9.12E-06
brain development (GO:0007420)	11	4.68%	170	0.93%	5.01	1.34E-05
membranous septum morphogenesis (GO:0003149)	3	1.28%	5	0.03%	46.47	2.08E-05
startle response (GO:0001964)	3	1.28%	6	0.03%	38.73	4.13E-05

Table 12 GO enrichment analyse of H7N9 NA related PDZ binding proteins

Term Description	**PDZ Domain**		**Proteome**		**E-ratio**	**p-value**
	Num.	***Per.***	***Num.***	***Per.***		
Top 10 most enriched biological processes						
transmembrane transport (GO:0055085)	30	10.56%	449	2.47%	4.28	2.15E-11
glutamate receptor signaling pathway (GO:0007215)	5	1.76%	13	0.07%	24.65	1.04E-06
ion transmembrane transport (GO:0034220)	8	2.82%	56	0.31%	9.16	2.36E-06
actin crosslink formation (GO:0051764)	4	1.41%	8	0.04%	32.05	3.87E-06
potassium ion transport (GO:0006813)	9	3.17%	83	0.46%	6.95	5.78E-06
muscular septum morphogenesis (GO:0003150)	3	1.06%	4	0.02%	48.07	1.49E-05
Wnt receptor signaling pathway, calcium modulating pathway (GO:0007223)	4	1.41%	11	0.06%	23.31	1.76E-05
regulation of small GTPase mediated signal transduction (GO:0051056)	11	3.87%	147	0.81%	4.80	1.99E-05
synaptic transmission (GO:0007268)	18	6.34%	381	2.09%	3.03	3.26E-05
membranous septum morphogenesis (GO:0003149)	3	1.06%	5	0.03%	38.45	3.67E-05

4 Discussion

4.1 New developments on the SDR approach

We have presented a PDZ domain-peptide interaction prediction tool based on the SDR approach, called PreDiZ. The SDR approach was first developed to predict specificity of protein kinases and showed the potential of working on other peptide recognition proteins (66). It was later applied in Predivac to predict CD4+ T-cell epitopes (69). In this work, we successfully adapted the SDR approach on another peptide recognition domain, the PDZ domain. Unlike other methods that build models for individual domains or a group of domains, PreDiZ analyses the SDRs on each PDZ domain and constructs a specific model according to these SDRs. Hence, PreDiZ isn't limited to predicting interactions for known PDZ domains, but also able to make prediction for novel PDZ domains.

4.1.1 Using a four-parameter-test as the SDR selection strategy

A number of improvements have been made to the SDR approach to apply it to the PDZ domains. First of all, a four-parameter-test (described in 2.2.2) was introduced to select SDRs. In previous tools using the SDR approach, i.e. Predikin and Predivac, residues within 5 Å distance between PDZ domain residue's side-chain atoms to the closest side-chain atoms of their binding peptides were selected as SDRs. Here, we used the four-parameter-test that determines SDRs from structural alignments, by filtering each column of the alignment with four parameters, including MCD, ACS, PIR and BLOSUM. Therefore, SDRs were selected not only based on the contact distance, but also the amino acid similarity of each column. This test was designed based on the theory that SDR is likely to be located at the position that is in close proximity of the binding peptide but not conserved across the PDZ domains. Hence, SDRs that affect the binding specificity either through direct contact or in indirect ways can all be considered. Furthermore, this test can be easily applied to other protein domains. In this study, the SDR set with the best performance was used MCD = 10 Å (3.1.2.1), as a typical Van der Waal interaction's contacting distance is normally less than 5 Å. This result supported our opinion of the specificity of PDZ domain isn't only affected by residues making direct contact. A recent study analysing 28 ligand-bound PDZ structures revealed structural determinants of peptide binding specificity for each of the last four residue of the binding peptide (91). Comparing these results and SDRs from PreDiZ, most of the SDRs we used were also found in this study (Table 2, Table 13). Another study used molecular dynamics simulations to analyse PDZ domain-peptide complexes with known binding affinities (92). The study identified 13 binding pocket residues on the PDZ domain structure 1TP3. Four out of six SDRs we used were in these binding pocket residues. The SDRs identified for PreDiZ were therefore consistent with the findings in the literature.

Table 13 SDRs presented in Ernst's study

Position	SDRs	SDRs converted into motif + offset format
0	β1:β2-7, β2-1, β2-3, α2-5, α2-8	GLGF-2, GLGF0, GLGF2, GD20, GD23
-1	β2-2, β2-4, β3:α1-1, β3-5	GLGF1, GLGF3, GD-15, GD-13
-2	β2-2, α2-1, α2-5	GLGF1, GD16, GD20
-3	β2-2, β2-4, β3-4, β3-5	GLGF1, GLGF3, GD-16, GD-15

4.1.2 Using both positive and negative interactions in the prediction

We employed the naive Bayesian classifier as an alternative scoring method in PreDiZ, to take the advantage of having both positive and negative interactions available for the PDZ domains. Both Predikin and Predivac used PSSM scoring methods, where only positive interactions were taken into account. In our case, negative interaction data were available for PDZ domains, therefore the naive Bayesian classifier scoring method (PreDiZ-NB) was employed as an alternative scoring method, in addition to the PSSM method (PreDiZ-PSSM). As expected, the performance of PreDiZ-NB was better than the PreDiZ-PSSM in the cross-validation tests, improving the AUC from 0.82 to 0.88. Despite the better scores in the cross-validation tests, PreDiZ-NB's AUC was only 0.67 using on the mouse test set of 50 mouse PDZ domains and 48 peptides, compared to PreDiZ-PSSM's AUC of 0.78 (3.1.2.2.1). This result was due to PreDiZ-NB predicting a large number of false positives. The predictions for individual PDZ domain showed PreDiZ did better in some PDZ domains than others. It is suggested that PreDiZ could be trained to perform well on the training set, but the selected SDRs are not specific enough to distinguish the interactions in some PDZ domains from this specific mouse test set. This mouse available test set is the only test data that with negative interactions. There is no any other test data to further validate either of these contradictory results. However, it is clear that there is still room for improvement for PreDiZ.

4.1.3 Using a wide range of interaction data

To the best of our knowledge, PreDiZ uses the widest range of experimentally verified PDZ interactions among the published PDZ interaction prediction algorithms. The PreDiZ's interaction data was extracted from three proteome-wide large scale studies on *C. elegans*, *M. musculus* and *H. sapiens*, as well as the manually curated interaction database Domino. Using a wide range of data enables PreDiZ to make predictions for more PDZ domains. However, including *C. elegans* interactions into our dataset lowered the results in cross-validation in our case. The reason is that including PDZ interactions from *C. elegans* introduced a large number of false negative results in the cross-validations due to the poor performance on predicting *C. elegans* PDZ interactions. To validate this assumption, we excluded all *C. elegans* data from our database and used only mouse and human PDZ proteins to perform 10-fold cross-validation. The AUCs showed significant improvement from 0.82 to 0.89 and from 0.88 to 0.94 for PreDiZ-PSSM and PreDiZ-NB, respectively (Figure 7, Table 14). The AUC value of PreDiZ-NB is one of the best score among all published methods. We also performed tests on the unpublished mouse test sets, and the performance showed only a slight improvement on PreDiZ-NB, but not much change for PreDiZ-PSSM (

Table 15, Table S3, Table S4). This result suggested that PreDiZ didn't use many C. elegans interactions to make prediction for PDZ domains in the mouse test set, as the PDZ domains from these two species do not share many similar SDRs. Overall, the results suggested the improvement on cross-validation tests were mainly contributed from excluding the *C. elegans* data, but the performance on making prediction on mouse PDZ domains was remain unchanged.

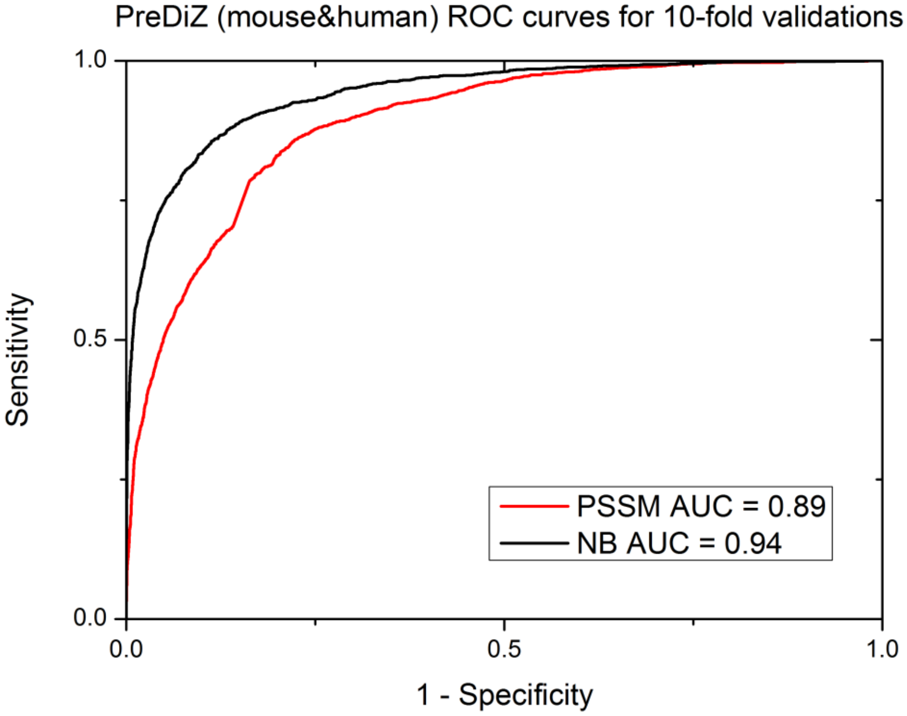

Figure 7 ROC curves of 10-fold cross-validations for PreDiZ-PSSM and PreDiZ-NB using mouse and human interaction data

Table 14 Cross-validation results of PreDiZ using only mouse and human interaction data

Predictor	***ACC* (%)**	***TPR* (%)**	***SPC* (%)**	**MCC**	**AUC**
PreDiZ-PSSM	90.16	51.21	94.87	0.4742	0.89
PreDiZ-NB	93.98	67.38	97.06	0.6664	0.94

Table 15 Performance of PreDiZ using only mouse and human interaction data on the unpublished mouse test set

Predictor	**TPR**	**FPR**	**TP/FP**	**AUC**
PreDiZ-PSSM	0.44	0.11	0.74	0.78
PreDiZ-NB	0.40	0.18	0.38	0.70

4.2 Comparisons between the PreDiZ and other methods

We compared the PreDiZ with three state-of-the-art PDZ interaction prediction tools. The results presented in 3.1.2.2.1 suggested that the PreDiZ performed better than MDSM but not as good as the other recently published methods in the mouse test sets. In the test of using *C. elegans* PDZ domains, both PreDiZ and PepPePInt scored rather low true positive rates, with the PreDiZ performing slightly better. Unfortunately, there is no negative interaction data on these *C. elegans* PDZ domains for us to do a comprehensive comparison. From the available information, it is suggested that PreDiZ is one of the best methods for predicting interactions for novel PDZ domains.

4.3 Applications of PDZ domain-peptide interaction prediction

4.3.1 Proteome-wide analyses

PDZ domains are known to have evolved relatively late in eukaryotic cells. The results of proteome-wide analyses of five different species supported this theory. A larger number of PDZ domains and PDZ binding proteins were found in species higher up in the evolutional tree such as *H. sapiens*, *M. musculus* and *D. rerio*. The GO and KEGG analyses revealed a number of common features of human PDZ domains. Not surprisingly, the PDZ domain involvement in most of the enriched GO terms and KEGG pathways is well documented.

4.3.2 PDZ binding motifs in the influenza A virus

The mortality rate for H5N1 and H7N9 virus infections in human is much higher compared to that of seasonal influenza infections (93, 94). There are reports showed the PBM in NS1 proteins contribute to virulence of HPAI H5N1 virus (95-98). Previous studies also revealed that compared to HPAI H5N1 virus, the HPAI H7N9 virus did not have PBM in the NS1 proteins, instead having a PBM in the NA proteins. Hence, it was suspected that the loss of function from missing the PBM in H5N1's NS1 proteins were regained by H7N9 virus through the PBM in their NA proteins (90). Our studies supported that the PBM on H5N1's NS1 proteins and H7N9's NA proteins regulate similar human biological process by PDZ-domain interactions. This finding may form the basis for further studies of the PBM in influenza A virus.

5 Conclusions and future directions

In this work, the PDZ domain-peptide interaction prediction tool PreDiZ has been developed. It is based on the SDR approach, which has been successfully applied in predicting protein-peptide interaction involving protein kinases and MHC class II proteins. A number of modifications has been made to apply the SDR approach to the PDZ domains, including using a more sophisticated strategy to determine SDRs, and using both positive and negative interactions in the prediction. One of the advantages of PreDiZ is that it is able to work on a wide range of PDZ domains, including novel PDZ domains. However, there is room to improve the tool, especially to reduce the false positive rate. We also conducted proteome-wide analyses in human PDZ binding proteins. Lastly, we studied how H7N9's NA and H5N1's NS1 protein regulate human biological processes, and found the PBM in the NS1 proteins of H5N1 was replaced by the PBM in the NA proteins of H7N9.

Improvements need to be made to PreDiZ to especially to reduce false positive predictions, in order for it to match the performance of the best available PDZ interaction prediction tool. At the time I started this project, there were only 22 PDZ complex structures that satisfied our selection criteria. However, the number of published PDZ structures have been increasing rapidly. Using more high quality structures will definitely have a positive impact on PreDiZ's performance. In the future update of PreDiZ, using more available structures to derive SDRs will be a very important part of the work. We also believe that the SDR selection could be further optimised to improve accuracy of our predictions. Due to time constraints, we couldn't employ a more sophisticate strategies of SDR identification, which could improve prediction accuracy. For example, a mutagenesis study conducted by Tonikian, *et al.* (46) showed that PDZ domain's specificity for residues at position 0 and -2 of the PBM were affected by direct side chain interaction. Specificity for position -1 was affected by both direct and indirect side chain interactions. Specificity for position -3 and -4 were affected by indirect ways. Different cut-offs of the four-parameter-test could be adjusted according these information independently for each peptide position. The prediction result from unpublished mouse data set showed PreDiZ didn't perform well on some PDZ domains. A potential solution to this problem would be using specific prediction models for each of the known PDZ domains and a general model for others. There are over 300 known PDZ domains, building prediction models for all of them is labour intensive. However, the four-parameter-test developed in this study can speed up the process by selecting SDRs automatically. For the proteome-wide analyses, including extra contextual information, such as secondary structure of binding peptides and co-cellular localisation of PDZ domain and their binding partners, may improve the quality of predictions. Finally, we expect that the SDR method, along with the four-parameter-test, can be used for other protein-peptide interaction domains.

6 Supplementary materials

Table S1 Prediction statistics of individual PDZ domains from mouse test set using PreDiZ-NB (high threshold)

PDZ Domain Name	TP	FN	FP	TN	TPR	FPR	TP/FP	Accuracy
a1-syntrophin_1-1	10	6	8	24	0.63	0.25	1.25	0.71
b1-syntrophin_1-1	9	7	9	23	0.56	0.28	1.00	0.67
Cipp_03-10	0	8	0	40	0.00	0.00	-	0.83
Cipp_05-10	0	2	12	34	0.00	0.26	0.00	0.71
Cipp_08-10	0	1	8	39	0.00	0.17	0.00	0.81
Cipp_09-10	3	2	7	36	0.60	0.16	0.43	0.81
Cipp_10-10	1	4	13	30	0.20	0.30	0.08	0.65
Dvl1_1-1	0	2	6	40	0.00	0.13	0.00	0.83
Dvl2_1-1	1	2	5	40	0.33	0.11	0.20	0.85
Dvl3_1-1	1	5	5	37	0.17	0.12	0.20	0.79
Erbin_1-1	0	1	8	39	0.00	0.17	0.00	0.81
g2-syntrophin_1-1	3	10	4	31	0.23	0.11	0.75	0.71
Gm1582_2-3	2	7	11	28	0.22	0.28	0.18	0.63
LIN-7A_1-1	6	4	11	27	0.60	0.29	0.55	0.69
Lin7c_1-1	6	3	11	28	0.67	0.28	0.55	0.71
Lrrc7_1-1	0	1	10	37	0.00	0.21	0.00	0.77
Magi-1_2-6	7	17	6	18	0.29	0.25	1.17	0.52
Magi-1_4-6	0	2	0	46	0.00	0.00	-	0.96
Magi-1_6-6	9	5	10	24	0.64	0.29	0.90	0.69
Magi-2_5-6	0	4	0	44	0.00	0.00	-	0.92
Magi-2_6-6	6	3	13	26	0.67	0.33	0.46	0.67
Magi-3_5-5	9	6	10	23	0.60	0.30	0.90	0.67
MUPP1_01-13	2	0	10	36	1.00	0.22	0.20	0.79
MUPP1_05-13	0	1	12	35	0.00	0.26	0.00	0.73
MUPP1_10-13	0	2	7	39	0.00	0.15	0.00	0.81
MUPP1_11-13	0	0	8	40	-	0.17	0.00	0.83
MUPP1_12-13	0	0	2	46	-	0.04	0.00	0.96
MUPP1_13-13	2	7	6	33	0.22	0.15	0.33	0.73
NHERF-1_1-2	1	2	4	41	0.33	0.09	0.25	0.88
NHERF-2_2-2	7	10	2	29	0.41	0.06	3.50	0.75
Pdlim5_1-1	0	0	14	34	-	0.29	0.00	0.71
Pdzk1_1-4	7	3	4	34	0.70	0.11	1.75	0.85
Pdzk3_1-1	0	1	4	43	0.00	0.09	0.00	0.90
PDZ-RGS3_1-1	0	13	0	35	0.00	0.00	-	0.73
PSD95_1-3	3	5	6	34	0.38	0.15	0.50	0.77
PTP-BL_2-5	2	2	7	37	0.50	0.16	0.29	0.81
SAP102_3-3	2	3	16	27	0.40	0.37	0.13	0.60
SAP97_1-3	3	5	6	34	0.38	0.15	0.50	0.77
SAP97_3-3	3	4	15	26	0.43	0.37	0.20	0.60

Scrb1_1-4	5	0	18	25	1.00	0.42	0.28	0.63
Scrb1_2-4	2	0	17	29	1.00	0.37	0.12	0.65
Scrb1_3-4	13	7	7	21	0.65	0.25	1.86	0.71
Shank3_1-1	10	4	4	30	0.71	0.12	2.50	0.83
SLIM_1-1	0	0	7	41	-	0.15	0.00	0.85
ZO-1_1-3	2	6	13	27	0.25	0.33	0.15	0.60
ZO-2_1-3	1	4	14	29	0.20	0.33	0.07	0.63
ZO-3_1-3	1	1	14	32	0.50	0.30	0.07	0.69

TP is true positive. FN is false negative. FP is false positive. TN is true negative. TPR is true positive rate. FPR is false positive rate.

Table S2 Prediction statistics of individual PDZ domains from mouse test set using PreDiZ-PSSM (high threshold)

PDZ Domain Name	**TP**	**FN**	**FP**	**TN**	**TPR**	**FPR**	**TP/FP**	**Accuracy**
a1-syntrophin_1-1	9	7	0	32	0.56	0.00	-	0.85
b1-syntrophin_1-1	9	7	0	32	0.56	0.00	-	0.85
Cipp_03-10	0	8	0	40	0.00	0.00	-	0.83
Cipp_05-10	1	1	5	41	0.50	0.11	0.20	0.88
Cipp_08-10	0	1	5	42	0.00	0.11	0.00	0.88
Cipp_09-10	4	1	3	40	0.80	0.07	1.33	0.92
Cipp_10-10	0	5	5	38	0.00	0.12	0.00	0.79
Dvl1_1-1	0	2	0	46	0.00	0.00	-	0.96
Dvl2_1-1	0	3	0	45	0.00	0.00	-	0.94
Dvl3_1-1	0	6	0	42	0.00	0.00	-	0.88
Erbin_1-1	0	1	0	47	0.00	0.00	-	0.98
g2-syntrophin_1-1	5	8	0	35	0.38	0.00	-	0.83
Gm1582_2-3	1	8	5	34	0.11	0.13	0.20	0.73
LIN-7A_1-1	5	5	1	37	0.50	0.03	5.00	0.88
Lin7c_1-1	5	4	1	38	0.56	0.03	5.00	0.90
Lrrc7_1-1	0	1	0	47	0.00	0.00	-	0.98
Magi-1_2-6	5	19	1	23	0.21	0.04	5.00	0.58
Magi-1_4-6	0	2	0	46	0.00	0.00	-	0.96
Magi-1_6-6	6	8	2	32	0.43	0.06	3.00	0.79
Magi-2_5-6	0	4	0	44	0.00	0.00	-	0.92
Magi-2_6-6	5	4	3	36	0.56	0.08	1.67	0.85
Magi-3_5-5	7	8	1	32	0.47	0.03	7.00	0.81
MUPP1_01-13	0	2	1	45	0.00	0.02	0.00	0.94
MUPP1_05-13	0	1	6	41	0.00	0.13	0.00	0.85
MUPP1_10-13	1	1	4	42	0.50	0.09	0.25	0.90
MUPP1_11-13	0	0	6	42	-	0.13	0.00	0.88
MUPP1_12-13	0	0	0	48	-	0.00	-	1.00
MUPP1_13-13	5	4	0	39	0.56	0.00	-	0.92
NHERF-1_1-2	1	2	12	33	0.33	0.27	0.08	0.71
NHERF-2_2-2	3	14	1	30	0.18	0.03	3.00	0.69
Pdlim5_1-1	0	0	0	48	-	0.00	-	1.00

Pdzk1_1-4	5	5	2	36	0.50	0.05	2.50	0.85
Pdzk3_1-1	0	1	1	46	0.00	0.02	0.00	0.96
PDZ-RGS3_1-1	0	13	0	35	0.00	0.00	-	0.73
PSD95_1-3	5	3	0	40	0.63	0.00	-	0.94
PTP-BL_2-5	2	2	5	39	0.50	0.11	0.40	0.85
SAP102_3-3	4	1	3	40	0.80	0.07	1.33	0.92
SAP97_1-3	5	3	0	40	0.63	0.00	-	0.94
SAP97_3-3	4	3	3	38	0.57	0.07	1.33	0.88
Scrb1_1-4	3	2	6	37	0.60	0.14	0.50	0.83
Scrb1_2-4	1	1	4	42	0.50	0.09	0.25	0.90
Scrb1_3-4	8	12	1	27	0.40	0.04	8.00	0.73
Shank3_1-1	5	9	0	34	0.36	0.00	-	0.81
SLIM_1-1	0	0	1	47	-	0.02	0.00	0.98
ZO-1_1-3	0	8	5	35	0.00	0.13	0.00	0.73
ZO-2_1-3	0	5	5	38	0.00	0.12	0.00	0.79
ZO-3_1-3	0	2	5	41	0.00	0.11	0.00	0.85

TP is true positive. FN is false negative. FP is false positive. TN is true negative. TPR is true positive rate. FPR is false positive rate.

Table S3 Prediction statistics of individual PDZ domains from mouse test set using PreDiZ-NB with only mouse and human interaction data

Domain Name	**TP**	**FN**	**FP**	**TN**	**TPR**	**FPR**	**TP/FP**	**Accuracy**
a1-syntrophin_1-1	11	5	6	26	0.69	0.19	1.83	0.77
b1-syntrophin_1-1	10	6	7	25	0.63	0.22	1.43	0.73
Cipp_03-10	0	8	0	40	0.00	0.00	-	0.83
Cipp_05-10	0	2	14	32	0.00	0.30	0.00	0.67
Cipp_08-10	0	1	9	38	0.00	0.19	0.00	0.79
Cipp_09-10	3	2	6	37	0.60	0.14	0.50	0.83
Cipp_10-10	1	4	4	39	0.20	0.09	0.25	0.83
Dvl1_1-1	0	2	4	42	0.00	0.09	0.00	0.88
Dvl2_1-1	1	2	3	42	0.33	0.07	0.33	0.90
Dvl3_1-1	1	5	3	39	0.17	0.07	0.33	0.83
Erbin_1-1	0	1	7	40	0.00	0.15	0.00	0.83
g2-syntrophin_1-1	3	10	2	33	0.23	0.06	1.50	0.75
Gm1582_2-3	2	7	11	28	0.22	0.28	0.18	0.63
LIN-7A_1-1	5	5	10	28	0.50	0.26	0.50	0.69
Lin7c_1-1	5	4	10	29	0.56	0.26	0.50	0.71
Lrrc7_1-1	0	1	8	39	0.00	0.17	0.00	0.81
Magi-1_2-6	5	19	1	23	0.21	0.04	5.00	0.58
Magi-1_4-6	0	2	0	46	0.00	0.00	-	0.96
Magi-1_6-6	8	6	9	25	0.57	0.26	0.89	0.69
Magi-2_5-6	0	4	0	44	0.00	0.00	-	0.92
Magi-2_6-6	6	3	11	28	0.67	0.28	0.55	0.71
Magi-3_5-5	9	6	8	25	0.60	0.24	1.13	0.71

MUPP1_01-13	2	0	4	42	1.00	0.09	0.50	0.92
MUPP1_05-13	0	1	14	33	0.00	0.30	0.00	0.69
MUPP1_10-13	0	2	8	38	0.00	0.17	0.00	0.79
MUPP1_11-13	0	0	6	42	-	0.13	0.00	0.88
MUPP1_12-13	0	0	2	46	-	0.04	0.00	0.96
MUPP1_13-13	3	6	6	33	0.33	0.15	0.50	0.75
NHERF-1_1-2	0	3	4	41	0.00	0.09	0.00	0.85
NHERF-2_2-2	2	15	2	29	0.12	0.06	1.00	0.65
Pdlim5_1-1	0	0	17	31	-	0.35	0.00	0.65
Pdzk1_1-4	7	3	4	34	0.70	0.11	1.75	0.85
Pdzk3_1-1	0	1	5	42	0.00	0.11	0.00	0.88
PDZ-RGS3_1-1	0	13	0	35	0.00	0.00	-	0.73
PSD95_1-3	2	6	7	33	0.25	0.18	0.29	0.73
PTP-BL_2-5	2	2	9	35	0.50	0.20	0.22	0.77
SAP102_3-3	2	3	17	26	0.40	0.40	0.12	0.58
SAP97_1-3	2	6	7	33	0.25	0.18	0.29	0.73
SAP97_3-3	3	4	16	25	0.43	0.39	0.19	0.58
Scrb1_1-4	5	0	14	29	1.00	0.33	0.36	0.71
Scrb1_2-4	1	1	14	32	0.50	0.30	0.07	0.69
Scrb1_3-4	13	7	6	22	0.65	0.21	2.17	0.73
Shank3_1-1	8	6	2	32	0.57	0.06	4.00	0.83
SLIM_1-1	0	0	7	41	-	0.15	0.00	0.85
ZO-1_1-3	4	4	10	30	0.50	0.25	0.40	0.71
ZO-2_1-3	2	3	12	31	0.40	0.28	0.17	0.69
ZO-3_1-3	0	2	14	32	0.00	0.30	0.00	0.67

TP is true positive. FN is false negative. FP is false positive. TN is true negative. TPR is true positive rate. FPR is false positive rate.

Table S4 Prediction statistics of individual PDZ domains from mouse test set using PreDiZ-PSSM with only mouse and human interaction data

Domain Name	TP	FN	FP	TN	TPR	FPR	TP/FP	Accuracy
a1-syntrophin_1-1	6	10	0	32	0.38	0.00	-	0.79
b1-syntrophin_1-1	6	10	0	32	0.38	0.00	-	0.79
Cipp_03-10	0	8	0	40	0.00	0.00	-	0.83
Cipp_05-10	1	1	5	41	0.50	0.11	0.20	0.88
Cipp_08-10	0	1	4	43	0.00	0.09	0.00	0.90
Cipp_09-10	4	1	3	40	0.80	0.07	1.33	0.92
Cipp_10-10	0	5	0	43	0.00	0.00	-	0.90
Dvl1_1-1	0	2	0	46	0.00	0.00	-	0.96
Dvl2_1-1	0	3	0	45	0.00	0.00	-	0.94
Dvl3_1-1	0	6	0	42	0.00	0.00	-	0.88
Erbin_1-1	0	1	0	47	0.00	0.00	-	0.98
g2-syntrophin_1-1	4	9	0	35	0.31	0.00	-	0.81
Gm1582_2-3	1	8	4	35	0.11	0.10	0.25	0.75

LIN-7A_1-1	5	5	1	37	0.50	0.03	5.00	0.88
Lin7c_1-1	5	4	1	38	0.56	0.03	5.00	0.90
Lrrc7_1-1	0	1	0	47	0.00	0.00	-	0.98
Magi-1_2-6	5	19	1	23	0.21	0.04	5.00	0.58
Magi-1_4-6	0	2	0	46	0.00	0.00	-	0.96
Magi-1_6-6	3	11	2	32	0.21	0.06	1.50	0.73
Magi-2_5-6	0	4	0	44	0.00	0.00	-	0.92
Magi-2_6-6	2	7	3	36	0.22	0.08	0.67	0.79
Magi-3_5-5	4	11	1	32	0.27	0.03	4.00	0.75
MUPP1_01-13	0	2	0	46	0.00	0.00	-	0.96
MUPP1_05-13	0	1	6	41	0.00	0.13	0.00	0.85
MUPP1_10-13	0	2	4	42	0.00	0.09	0.00	0.88
MUPP1_11-13	0	0	6	42	-	0.13	0.00	0.88
MUPP1_12-13	0	0	0	48	-	0.00	-	1.00
MUPP1_13-13	4	5	0	39	0.44	0.00	-	0.90
NHERF-1_1-2	1	2	1	44	0.33	0.02	1.00	0.94
NHERF-2_2-2	0	17	1	30	0.00	0.03	0.00	0.63
Pdlim5_1-1	0	0	0	48	-	0.00	-	1.00
Pdzk1_1-4	4	6	2	36	0.40	0.05	2.00	0.83
Pdzk3_1-1	0	1	1	46	0.00	0.02	0.00	0.96
PDZ-RGS3_1-1	0	13	0	35	0.00	0.00	-	0.73
PSD95_1-3	5	3	0	40	0.63	0.00	-	0.94
PTP-BL_2-5	2	2	3	41	0.50	0.07	0.67	0.90
SAP102_3-3	4	1	3	40	0.80	0.07	1.33	0.92
SAP97_1-3	5	3	0	40	0.63	0.00	-	0.94
SAP97_3-3	4	3	3	38	0.57	0.07	1.33	0.88
Scrb1_1-4	3	2	6	37	0.60	0.14	0.50	0.83
Scrb1_2-4	1	1	4	42	0.50	0.09	0.25	0.90
Scrb1_3-4	8	12	1	27	0.40	0.04	8.00	0.73
Shank3_1-1	1	13	0	34	0.07	0.00	-	0.73
SLIM_1-1	0	0	1	47	-	0.02	0.00	0.98
ZO-1_1-3	1	7	5	35	0.13	0.13	0.20	0.75
ZO-2_1-3	0	5	6	37	0.00	0.14	0.00	0.77
ZO-3_1-3	0	2	6	40	0.00	0.13	0.00	0.83

TP is true positive. FN is false negative. FP is false positive. TN is true negative. TPR is true positive rate. FPR is false positive rate.

7 References

1. Suter B, Kittanakom S, & Stagljar I (2008) Interactive proteomics: what lies ahead? *BioTechniques* 44(5):681-691.

2. Petsalaki E & Russell RB (2008) Peptide-mediated interactions in biological systems: new discoveries and applications. *Current opinion in biotechnology* 19(4):344-350.

3. Kobe B & Boden M (2012) Computational modelling of linear motif-mediated protein interactions. *Current topics in medicinal chemistry* 12(14):1553-1561.

4. Neduva V & Russell RB (2005) Linear motifs: evolutionary interaction switches. *FEBS letters* 579(15):3342-3345.

5. Neduva V & Russell RB (2006) Peptides mediating interaction networks: new leads at last. *Current opinion in biotechnology* 17(5):465-471.

6. Fields S & Song O (1989) A novel genetic system to detect protein-protein interactions. *Nature* 340(6230):245-246.

7. Young KH (1998) Yeast two-hybrid: so many interactions, (in) so little time. *Biology of reproduction* 58(2):302-311.

8. Ho YP & Hsu PH (2002) Investigating the effects of protein patterns on microorganism identification by high-performance liquid chromatography-mass spectrometry and protein database searches. *Journal of chromatography. A* 976(1-2):103-111.

9. Songyang Z*, et al.* (1994) Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Current Biology* 4(11):973-982.

10. Hakes L, Pinney JW, Robertson DL, & Lovell SC (2008) Protein-protein interaction networks and biology--what's the connection? *Nature biotechnology* 26(1):69-72.

11. Shi TL, Li YX, Cai YD, & Chou KC (2005) Computational methods for protein-protein interaction and their application. *Current protein & peptide science* 6(5):443-449.

12. Kalyoncu S, Keskin O, & Gursoy A (2010) Interaction prediction and classification of PDZ domains. *BMC bioinformatics* 11:357.

13. Shifman JM & Mayo SL (2003) Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proceedings of the National Academy of Sciences of the United States of America* 100(23):13274-13279.

14. Cho KO, Hunt CA, & Kennedy MB (1992) The rat brain postsynaptic density fraction contains a homolog of the Drosophila discs-large tumor suppressor protein. *Neuron* 9(5):929-942.

15. Woods DF & Bryant PJ (1993) ZO-1, DlgA and PSD-95/SAP90: homologous proteins in tight, septate and synaptic cell junctions. *Mechanisms of development* 44(2-3):85-89.

16. Kim E, Niethammer M, Rothschild A, Jan YN, & Sheng M (1995) Clustering of Shaker-type K+ channels by interaction with a family of membrane-associated guanylate kinases. *Nature* 378(6552):85-88.

17. Ponting CP (1997) Evidence for PDZ domains in bacteria, yeast, and plants. *Protein science : a publication of the Protein Society* 6(2):464-468.

18. Sussman JL*, et al.* (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica Section D Biological Crystallography* 54(Pt 6 Pt 1):1078-1084.

19. Wang CK, Pan L, Chen J, & Zhang M (2010) Extensions of PDZ domains as important structural and functional elements. *Protein Cell* 1(8):737-751.

20. Ernst A*, et al.* (2009) Rapid evolution of functional complexity in a domain family. *Science signaling* 2(87):ra50.

21. Songyang Z*, et al.* (1997) Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science (New York, N.Y.)* 275(5296):73-77.

22. Doyle DA*, et al.* (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell* 85(7):1067-1076.

23. Fanning AS & Anderson JM (1996) Protein-protein interactions: PDZ domain networks. *Current Biology* 6(11):1385-1388.

24. Basdevant N, Weinstein H, & Ceruso M (2006) Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study. *Journal of the American Chemical Society* 128(39):12766-12777.

25. Gerek ZN, Keskin O, & Ozkan SB (2009) Identification of specificity and promiscuity of PDZ domain interactions through their dynamic behavior. *Proteins* 77(4):796-811.

26. Daniels DL, Cohen AR, Anderson JM, & Brunger AT (1998) Crystal structure of the hCASK PDZ domain reveals the structural basis of class II PDZ domain target recognition. *Nature structural biology* 5(4):317-325.

27. Harris BZ & Lim WA (2001) Mechanism and role of PDZ domains in signaling complex assembly. *Journal of cell science* 114(Pt 18):3219-3231.

28. Chi CN, Engstrom A, Gianni S, Larsson M, & Jemth P (2006) Two conserved residues govern the salt and pH dependencies of the binding reaction of a PDZ domain. *The Journal of biological chemistry* 281(48):36811-36818.

29. Harris BZ, Lau FW, Fujii N, Guy RK, & Lim WA (2003) Role of electrostatic interactions in PDZ domain ligand recognition. *Biochemistry* 42(10):2797-2805.

30. Smock RG & Gierasch LM (2009) Sending signals dynamically. *Science (New York, N.Y.)* 324(5924):198-203.

31. Akiva E, Friedlander G, Itzhaki Z, & Margalit H (2012) A dynamic view of domain-motif interactions. *PLoS computational biology* 8(1):e1002341.

32. Hillier BJ, Christopherson KS, Prehoda KE, Bredt DS, & Lim WA (1999) Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex. *Science (New York, N.Y.)* 284(5415):812-815.

33. Hurd TW, Gao L, Roh MH, Macara IG, & Margolis B (2003) Direct interaction of two polarity complexes implicated in epithelial tight junction assembly. *Nature cell biology* 5(2):137-142.

34. London TB, Lee HJ, Shao Y, & Zheng J (2004) Interaction between the internal motif KTXXXI of Idax and mDvl PDZ domain. *Biochemical and biophysical research communications* 322(1):326-332.

35. Wong HC*, et al.* (2003) Direct binding of the PDZ domain of Dishevelled to a conserved internal sequence in the C-terminal region of Frizzled. *Molecular cell* 12(5):1251-1260.

36. Uemura T, Mori H, & Mishina M (2004) Direct interaction of GluRdelta2 with Shank scaffold proteins in cerebellar Purkinje cells. *Molecular and cellular neurosciences* 26(2):330-341.

37. Zhang Y, Appleton BA, Wu P, Wiesmann C, & Sidhu SS (2007) Structural and functional analysis of the ligand specificity of the HtrA2/Omi PDZ domain. *Protein science : a publication of the Protein Society* 16(8):1738-1750.

38. Ellencrona K, Syed A, & Johansson M (2009) Flavivirus NS5 associates with host-cell proteins zonula occludens-1 (ZO-1) and regulating synaptic membrane exocytosis-2 (RIMS2) via an internal PDZ binding mechanism. *Biological chemistry* 390(4):319-323.

39. Sengupta D & Linstedt AD (2010) Mitotic inhibition of GRASP65 organelle tethering involves Polo-like kinase 1 (PLK1) phosphorylation proximate to an internal PDZ ligand. *The Journal of biological chemistry* 285(51):39994-40003.

40. Zimmermann P*, et al.* (2002) PIP(2)-PDZ domain binding controls the association of syntenin with the plasma membrane. *Molecular cell* 9(6):1215-1225.

41. Zimmermann P (2006) PDZ domain-phosphoinositide interactions in cell-signaling. *Verhandelingen - Koninklijke Academie voor Geneeskunde van Belgie* 68(4):271-286.

42. Sugi T, Oyama T, Morikawa K, & Jingami H (2008) Structural insights into the PIP2 recognition by syntenin-1 PDZ domain. *Biochemical and biophysical research communications* 366(2):373-378.

43. Gallardo R, Ivarsson Y, Schymkowitz J, Rousseau F, & Zimmermann P (2010) Structural diversity of PDZ-lipid interactions. *Chembiochem : a European journal of chemical biology* 11(4):456-467.

44. Bezprozvanny I & Maximov A (2001) Classification of PDZ domains. *FEBS letters* 509(3):457-462.

45. Gerek ZN & Ozkan SB (2010) A flexible docking scheme to explore the binding selectivity of PDZ domains. *Protein science : a publication of the Protein Society* 19(5):914-928.

46. Tonikian R*, et al.* (2008) A Specificity Map for the PDZ Domain Family. *PLoS Biology* 6(9).

47. Reina J*, et al.* (2002) Computer-aided design of a PDZ domain to recognize new target sequences. *Nature structural biology* 9(8):621-627.

48. Poulat F*, et al.* (1997) The human testis determining factor SRY binds a nuclear factor containing PDZ protein interaction domains. *The Journal of biological chemistry* 272(11):7167-7172.

49. Barritt DS*, et al.* (2000) The multi-PDZ domain protein MUPP1 is a cytoplasmic ligand for the membrane-spanning proteoglycan NG2. *Journal of cellular biochemistry* 79(2):213-224.

50. Nielsen PA, Baruch A, Giepmans BN, & Kumar NM (2001) Characterization of the association of connexins and ZO-1 in the lens. *Cell Communication & Adhesion.* 8(4-6):213-217.

51. Marfatia SM*, et al.* (1996) Modular organization of the PDZ domains in the human discs-large protein suggests a mechanism for coupling PDZ domain-binding proteins to ATP and the membrane cytoskeleton. *The Journal of cell biology* 135(3):753-766.

52. Kornau HC, Schenker LT, Kennedy MB, & Seeburg PH (1995) Domain interaction between NMDA receptor subunits and the postsynaptic density protein PSD-95. *Science (New York, N.Y.)* 269(5231):1737-1740.

53. Skrabanek L, Saini HK, Bader GD, & Enright AJ (2008) Computational prediction of protein-protein interactions. *Molecular biotechnology* 38(1):1-17.

54. Berman HM*, et al.* (2002) The Protein Data Bank. *Acta crystallographica. Section D, Biological crystallography* 58(Pt 6 No 1):899-907.

55. Letunic I, Doerks T, & Bork P (2009) SMART 6: recent updates and new developments. *Nucleic acids research* 37(Database issue):D229-232.

56. Encinar JA*, et al.* (2009) ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics*, England), Vol 25, pp 2418-2424.

57. Schymkowitz J*, et al.* (2005) The FoldX web server: an online force field. *Nucleic acids research* 33(Web Server issue):W382-388.

58. Smith CA & Kortemme T (2010) Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. *Journal of molecular biology* 402(2):460-474.

59. Stiffler MA*, et al.* (2007) PDZ domain binding selectivity is optimized across the mouse proteome. *Science (New York, N.Y.)* 317(5836):364-369.

60. Chen JR, Chang BH, Allen JE, Stiffler MA, & MacBeath G (2008) Predicting PDZ domain-peptide interactions from primary sequences. *Nature biotechnology* 26(9):1041-1045.

61. Schillinger C, Boisguerin P, & Krause G (2009) Domain Interaction Footprint: a multi-classification approach to predict domain-peptide interactions. *Bioinformatics*, England), Vol 25, pp 1632-1639.

62. Shao X*, et al.* (2011) A regression framework incorporating quantitative and negative interaction data improves quantitative prediction of PDZ domain-peptide interaction from primary sequence. *Bioinformatics* 27(3):383-390.

63. Li L*, et al.* (2011) DomPep--a general method for predicting modular domain-mediated protein-protein interactions. *PloS one* 6(10):e25528.

64. Kundu K & Backofen R (2014) Cluster based prediction of PDZ-peptide interactions. *BMC genomics* 15 Suppl 1:S5.

65. Kundu K, Mann M, Costa F, & Backofen R (2014) MoDPepInt: an interactive web server for prediction of modular domain-peptide interactions. *Bioinformatics* 30(18):2668-2669.

66. Brinkworth RI, Breinl RA, & Kobe B (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proceedings of the National Academy of Sciences of the United States of America* 100(1):74-79.

67. Saunders NF, Brinkworth RI, Huber T, Kemp BE, & Kobe B (2008) Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC bioinformatics* 9:245.

68. Ellis JJ & Kobe B (2011) Predicting protein kinase specificity: Predikin update and performance in the DREAM4 challenge. *PloS one* 6(7):e21169.

69. Oyarzún P, Ellis JJ, Bodén M, & Kobe B (2013) PREDIVAC: CD4+ T-cell epitope prediction for vaccine design that covers 95% of HLA class II DR protein diversity. *BMC Bioinformatics* 14:52.

70. Oyarzun P*, et al.* (2015) A bioinformatics tool for epitope-based vaccine design that accounts for human ethnic diversity: Application to emerging infectious diseases. *Vaccine* 33(10):1267-1273.

71. Lenfant N*, et al.* (2010) A genome-wide study of PDZ-domain interactions in C. elegans reveals a high frequency of non-canonical binding. *BMC genomics* 11:671.

72. Ceol A*, et al.* (2007) DOMINO: a database of domain-peptide interactions. *Nucleic acids research* 35(Database issue):D557-560.

73. UniProt C (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic acids research* 42(Database issue):D191-198.

74. Kanehisa M*, et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research* 42(Database issue):D199-205.

75. Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28(1):27-30.

76. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, & Tatusova T (2007) FLAN: a web server for influenza virus genome annotation. *Nucleic acids research* 35(Web Server issue):W280-284.

77. Pei J, Kim BH, & Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research* 36(7):2295-2300.

78. Xing Y, Xu Q, & Lee C (2003) Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. *FEBS letters* 555(3):572-578.

79. Saunders NF & Kobe B (2008) The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information. *Nucleic acids research* 36(Web Server issue):W286-290.

80. Beuming T, Skrabanek L, Niv MY, Mukherjee P, & Weinstein H (2005) PDZBase: a protein-protein interaction database for PDZ-domains. *Bioinformatics* 21(6):827-828.

81. Kato Y, Yoshida K, Watanabe C, Sai Y, & Tsuji A (2004) Screening of the interaction between xenobiotic transporters and PDZ proteins. *Pharmaceutical research* 21(10):1886-1894.

82. Tsunoda S*, et al.* (1997) A multivalent PDZ-domain protein assembles signalling complexes in a G-protein-coupled cascade. *Nature* 388(6639):243-249.

83. Yang N*, et al.* (1998) Cofilin phosphorylation by LIM-kinase 1 and its role in Rac-mediated actin reorganization. *Nature* 393(6687):809-812.

84. Ikemoto M*, et al.* (2000) Identification of a PDZ-domain-containing protein that interacts with the scavenger receptor class B type I. *Proceedings of the National Academy of Sciences of the United States of America* 97(12):6538-6543.

85. Silver DL & Tall AR (2001) The cellular biology of scavenger receptor class B type I. *Current opinion in lipidology* 12(5):497-504.

86. Walsh NP, Alba BM, Bose B, Gross CA, & Sauer RT (2003) OMP peptide signals initiate the envelope-stress response by activating DegS protease via relief of inhibition mediated by its PDZ domain. *Cell* 113(1):61-71.

87. Young JC & Hartl FU (2003) A stress sensor for the bacterial periplasm. *Cell* 113(1):1-2.

88. Golebiewski L, Liu H, Javier RT, & Rice AP (2011) The avian influenza virus NS1 ESEV PDZ binding motif associates with Dlg1 and Scribble to disrupt cellular tight junctions. *Journal of virology* 85(20):10639-10648.

89. Soubies SM*, et al.* (2013) Deletion of the C-terminal ESEV domain of NS1 does not affect the replication of a low-pathogenic avian influenza virus H7N1 in ducks and chickens. *The Journal of general virology* 94(Pt 1):50-58.

90. Wang Y*, et al.* (2013) Towards a better understanding of the novel avian-origin H7N9 influenza A virus in China. *Scientific Reports* 3:2318.

91. Ernst A*, et al.* (2014) A structural portrait of the PDZ domain family. *Journal of molecular biology* 426(21):3509-3519.

92. Tiwari G & Mohanty D (2014) Structure-based multiscale approach for identification of interaction partners of PDZ domains. *Journal of chemical information and modeling* 54(4):1143-1156.

93. Beigel JH*, et al.* (2005) Avian influenza A (H5N1) infection in humans. *The New England journal of medicine* 353(13):1374-1385.

94. Gao HN*, et al.* (2013) Clinical findings in 111 cases of influenza A (H7N9) virus infection. *The New England journal of medicine* 368(24):2277-2285.

95. Zielecki F*, et al.* (2010) Virulence determinants of avian H5N1 influenza A virus in mammalian and avian hosts: role of the C-terminal ESEV motif in the viral NS1 protein. *Journal of virology* 84(20):10708-10718.

96. Yu J*, et al.* (2011) PDlim2 selectively interacts with the PDZ binding motif of highly pathogenic avian H5N1 influenza A virus NS1. *PloS one* 6(5):e19511.

97. Jackson D, Hossain MJ, Hickman D, Perez DR, & Lamb RA (2008) A new influenza virus virulence determinant: the NS1 protein four C-terminal residues modulate pathogenicity. *Proceedings of the National Academy of Sciences of the United States of America* 105(11):4381-4386.

98. Fan S*, et al.* (2013) Synergistic Effect of the PDZ and p85beta-Binding Domains of the NS1 Protein on Virulence of an Avian H5N1 Influenza A Virus. *Journal of virology* 87(9):4861-4871.