Computational Prediction Of PDZ Mediated Protein-Protein Interactions

by

Shirley Hui

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Molecular Genetics
University of Toronto

Computational Prediction Of PDZ Mediated Protein-Protein Interactions

Shirley Hui

Doctor of Philosophy

Molecular Genetics
University of Toronto

2013

Abstract

Many protein-protein interactions, especially those involved in eukaryotic signalling, are mediated by PDZ domains through the recognition of hydrophobic C-termini. The availability of experimental PDZ interaction data sets have led to the construction of computational methods to predict PDZ domain-peptide interactions. Such predictors are ideally suited to predict interactions in single organisms or for limited subsets of PDZ domains. As a result, the goal of my thesis has been to build general predictors that can be used to scan the proteomes of multiple organisms for ligands for almost all PDZ domains from select model organisms. A framework consisting of four steps: data collection, feature encoding, predictor training and evaluation was developed and applied for all predictors built in this thesis.

The first predictor utilized PDZ domain-peptide sequence information from two interaction data sets obtained from high throughput protein microarray and phage display experiments in mouse and human, respectively. The second predictor used PDZ domain structure and peptide sequence information. I showed that these predictors are complementary to each other, are capable of predicting unseen interactions and can be used for the purposes of proteome scanning in human, worm and fly. As both positive and negative interactions are required for building a successful predictor, a major obstacle I addressed was the generation of artificial negative interactions for training. In particular, I used position weight matrices to generate such negatives for the positive only phage display data and used a semi-supervised learning approach to overcome the problem of over-prediction (i.e. prediction of too many positives). These predictors are available as a community web resource: http://webservice.baderlab.org/domains/POW. Finally, a Bayesian integration method combining information from different biological evidence sources was used to filter the human proteome scanning predictions from both predictors. This resulted in the construction of a comprehensive physiologically relevant high confidence PDZ mediated protein-protein interaction network in human.

Acknowledgments

I would like to thank my supervisor Gary Bader for his continuous support and patience. This thesis would not be possible without his scientific insight and guidance throughout my Ph.D. studies.

I also thank my supervisory committee members: Charlie Boone, Quaid Morris and Tony Pawson whose encouragement and insightful comments have been most valuable. Thank you also to the many collaborators and contributors who have shared their knowledge and data with me over the years.

I also thank all past and present members of the Bader lab and members of the labs of the CCBR 6th floor for their stimulating discussions (although, it is the fun times we had over the years that I will remember the most!).

Finally, I would like to thank my family and my husband Kevin Leung for their unconditional patience and support and Simon for his constant company.

This thesis is dedicated in loving memory of my brother, Steven Sai Ho Hui.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

AUC: Area under the curve

BiNGO: Biological Network Gene Ontology tool

CO-IP: Co-immunoprecipitation

FDR: False discovery rate

FP: False positive

FPR: False positive rate

GO: Gene ontology

NN: Nearest neighbour

MDSM: Multidomain selectivity model

MYTH: Membrane yeast two-hybrid

PDB: Protein data bank

PDZ: PSD95/DlgA/Zo-1

PPI: protein-protein interaction

PR: Precision/Recall

PRM: Peptide recognition module

PWM: Position weight matrix

ROC: Receiver operating characteristic

SVM: Support vector machine

TCSS: Topological clustering semantic similarity

TN: True negative

TP: True positive

TPR: True positive rate

Chapter 1

Introduction

1 Introduction

1.1 Protein-protein interaction network mapping across the proteome

The human genome contains approximately 23,000 protein-coding genes, which through alternative splicing can direct the synthesis of thousands of different proteins (Flicek et al. 2012). The majority of these proteins interact with other proteins to coordinate a variety of cellular processes including DNA replication, cell cycle control, and signal transduction. The ability to detect these interactions enables the assembly of protein interaction networks, which can be used to better understand how such biological processes are organized. Rewired networks can also be used to study genetic disorders caused by the abolishment of existing interactions or establishment of aberrant ones by mutant proteins (Pawson and Nash 2003). To understand how the genome encodes and how genomic changes may affect the underlying interaction network, correctly detecting protein-protein interactions (PPIs) directly from the proteome is necessary. Such methods must be both accurate and precise to avoid generating too many false positive hits given the large number of possible interactors in a typical eukaryotic proteome.

1.1.1 Experimental and computational methods for studying PPIs

Various experimental methods can be used to detect PPIs *in vivo* and *in vitro* (Phizicky and Fields 1995). The two-hybrid screen is an example of an *in vivo* method that involves detecting reporter gene expression by transcription factor binding onto an upstream activating sequence. In this method, one protein is fused to the transcription factor's DNA binding domain and the other to the transcriptional promoter domain. If the proteins interact, a functional transcription factor is created and results in the expression of the reporter gene. Co-immunoprecipitation (Co-IP), phage display and protein microarrays are examples of *in vitro* methods. Co-IP uses an antibody to immunoprecipitate a target antigen and also co-precipitate any other bound proteins within a sample such as a cell lysate. This is followed by mass spectrometry to identify the bound proteins. Phage display experiments successively select for partners that are displayed on the surface of phage that interact with proteins immobilized on a solid surface. Protein microarrays use labeled proteins to probe for interacting proteins immobilized on a microarray with binding measured semi-quantitatively using a colorimetric assay. The main advantage of *in* *vivo* methods like the two-hybrid screen is that the detection of interactions takes place within the native environment of the cell resulting in more biologically meaningful interactions. However, *in vitro* methods such as phage display or protein microarray can be used in high throughput studies as large random libraries or sets of peptides can be used for selection or screening.

Computational methods to predict PPIs have also been developed and can provide additional evidence for experimental results, provide new insights by combining results from different experiments or prioritize experiments by predicting which proteins are likely binders among a large number of possible binders. Such methods fall into a range of categories from physics to statistics-based methods. Physics-based methods such as protein docking algorithms use geometric and steric considerations to fit two proteins of known structure into a bound complex. Simple sequence-based methods include position weight matrices (PWMs). Given a set of verified ligands, a PWM is matrix of probabilities of observing a particular residue at a given ligand position. PWMs can be used to compute a score indicating the binding preference of a domain for a given peptide. More sophisticated methods employ algorithms from an area of Computer Science called machine learning. Machine learning refers to a family of computational methods that can recognize complex patterns in a given dataset in order to make decisions on new unseen data. For PPI prediction, patterns can be extracted from structure, sequence and other relevant data to train a predictor that will predict if two given proteins will interact. Many different algorithms have been used to predict PPIs including Bayesian methods, neural networks and support vector machines (SVMs) (Bock and Gough 2001; Jansen et al. 2003; Ferraro et al. 2006).

1.1.2 Challenges faced by existing methods

While experimental methods in general can detect PPIs, they face many challenges, such as the proteins may be limited to those that stably fold in a bacteria host, the interaction sites may be blocked by tagging or fusion to another protein, or the binding conditions may be artificially imposed by the investigator. High rates of false positives are also a concern since the detected interactions may be indirect or may never occur *in vivo*. As a result, the authenticity of putative interactions from experiments must be substantiated using additional techniques.

Although computational methods can also correctly predict PPIs, they face several challenges. For physics-based prediction methods, the structures of the proteins are often unavailable or protein flexibility is not considered. Simple methods like PWMs can only represent short binding motifs and often do not account for interdependencies between residues and positions. They may also perform poorly when there are too few experimentally determined binders available for a given protein. Furthermore, such a simple model does not allow for the incorporation of additional biological information such as the gene expression or cellular location of protein pairs to help reduce the number of false positives. Machine learning methods also face certain drawbacks. For example, most machine learning methods require both positive and negative data for training. Therefore, the limited availability of negative interaction data is a common problem, although a recent database has begun to archive such data (Smialowski et al. 2010). Methods which are typically used include shuffled or random ligand sequences. However, the use of such negatives for training results in predictors with lower accuracy when real negatives are used for testing (Lo et al. 2005; Ben-Hur and Noble 2006). Other methods include randomly shuffling the interacting partners or pairing partners which are known to not be in the same cellular compartment. However, these methods create a constraint on the distribution of negatives and make it easier for the predictor to distinguish between positive and negative interactions. This leads to biased estimates of predictor performance when cross validation is used (Ben-Hur and Noble 2006). Due to these challenges, the computational prediction of PPIs is an extremely difficult problem that is not fully addressed by any existing method.

1.2 Peptide recognition modules mediating protein-protein interactions

Many PPIs in eukaryotic signalling systems are mediated by protein recognition modules (PRMs). PRMs are evolutionarily conserved protein domains that fold independently and are organized in different ways to form larger proteins. PRMs have important roles in signal transduction including the assembly of multiprotein complexes, subcellular localization of regulatory proteins and recognition of protein post-translational modifications (Pawson and Nash 2003). These processes are facilitated by PRM-target binding via the recognition of short linear target motifs.

1.2.1 WW and SH3 domains

Many PRMs are known and differ by the set of motifs that they recognize. For example, WW and SH3 domains, bind proline-rich motifs. The WW domain is a short PRM consisting of approximately 40 amino acid residues. These domains fold into a triple stranded beta sheet and contain two tryptophan residues spaced approximately 20 residues apart from each other. WW domains bind short proline motifs (e.g. PPXY, PPLP) and are involved in a variety of processes including receptor signalling and cytoskeleton regulation (Ingham et al. 2005). SH3 domains consist of approximately 60 amino acids and fold into a beta-barrel structure composed of five to six anti-parallel beta strands. In general, these domains also bind to proline-rich motifs (e.g. PXXP) (Tong et al. 2002) which can be accommodated in the SH3 binding site in two possible orientations. Nonconventional motifs that contain arginine or lysine have also been observed and bind in a proline-independent manner (e.g. RKXXYXXY) (Kang et al. 2000). SH3 domains are involved in tyrosine kinase signalling, cytoskeletal organization and cell polarization (Pawson and Nash 2003).

1.2.2 PDZ domains

The **P**SD95/**D**lgA/**Z**o-1 (PDZ) domain is an example of a PRM that is found in increasing abundance in yeast to metazoans with over 250 encoded in the human genome (Ponting 1997). PDZ domains mediate numerous important biological processes, such as ion channel regulation, cell polarity determination and neural development. They are generally found in eukaryotic signalling pathways, often in scaffolding proteins that are responsible for regulating the assembly and localization of intracellular protein complexes to specialized sites in the cell, especially at membranes (Pawson and Nash 2003). Disruption of PDZ domain mediated interactions is associated with diseases such as human papillomavirus, cystic fibrosis and schizophrenia (Moyer et al. 1999; Dev 2004; Doorbar 2006).

The PDZ domain is approximately 80-90 amino acids long and folds into a globular structure consisting of six β strands and two α helices (**Figure 1-1**).

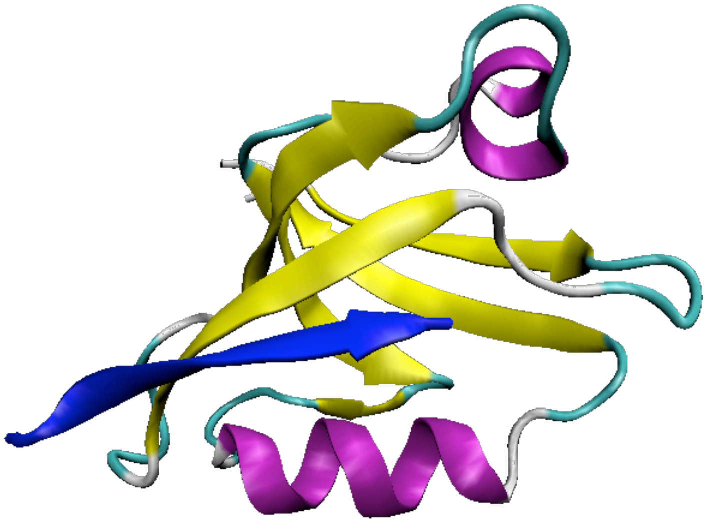


Figure 1-1 3D structure of a bound PDZ domain. A bound peptide is shown in blue, α helices are shown in purple and β strands are shown in yellow. PDB: 1N7F.

Domains bind C-termini with canonical interactions occurring between C-terminal target side chains and a hydrophobic binding pocket formed between domain β2 strand and α2 helix. PDZ domains bind their targets with affinities in the micromolar range through the recognition of short linear motifs. Early peptide library screens grouped their binding specificity into two classes, where class I domains prefer to bind the motif X[T/S]XΦ and class II domains prefer to bind the motif ΦXΦX (where X is any amino acid and Φ is a hydrophobe) (Songyang et al. 1997). More recent studies have found that the PDZ domain can be specific up to seven residues and recognize more than these two classes (Zhang et al. 2006; Tonikian et al. 2008).

1.3 Mapping PDZ domain mediated protein-protein interactions

PDZ domain-peptide interactions have been mapped using a variety of experimental methods. The biological importance of PDZ domains, their simple modes of target and the availability of experimental data sets have also prompted the development of computational methods to predict PDZ domain-peptide interactions by multiple groups.

1.3.1 Experimental methods

In phage display experiments, a large-scale combinatorial peptide library is presented to a given PDZ domain and bound peptides are identified by sequencing the corresponding phage DNA

(Tong et al. 2002; Sidhu et al. 2003; Tonikian et al. 2008). To represent an exhaustive set, extremely large libraries (up to approximately ten billion peptides) can be created containing every possible binding target; 1.3 billion peptides are needed to cover all seven residues using 20 amino acids. In protein microarray experiments, purified domains are immobilized on a solid surface and probed using fluorescently labelled peptides, allowing several hundred domains to be tested for binding against hundreds of peptides simultaneously (MacBeath and Schreiber 2000; Hu et al. 2004; Stiffler et al. 2007). For peptide chip experiments, synthesised peptides are displayed to domains on a protein cellulose membrane chip. These experiments are limited to libraries with sizes in the thousands, so are often designed to use only peptides matching a known binding motif for a given domain type (Landgraf et al. 2004; Wiedemann et al. 2004; Wu et al. 2007; Huang et al. 2008). Either the domains or peptides are displayed on the chip, followed by binding of the interaction partners (Hu et al. 2004; Landgraf et al. 2004; Wiedemann et al. 2004). In both peptide chip and protein microarray experiments, binding is generally measured semi-quantitatively, for example using a colorimetric assay. Yeast two-hybrid assays involve detecting reporter gene expression by transcription factor binding onto an upstream activating sequence. In this method, a ligand of interest is fused to the transcription factor's DNA binding domain (bait) and the domain to the transcriptional promoter domain (prey). High throughput yeast two-hybrid interaction arrays are constructed by taking a collection of yeast strains, each expressing a domain prey, and spotting them on a solid surface. These domains are probed by adding to each strain a vector expressing a ligand of interest (Tonikian et al. 2009; Lenfant et al. 2010).

Quantitative measurements of the strength of PDZ domain-peptide binding can be obtained using different methods such as fluorescence polarization or surface plasmon resonance assays. Fluorescence polarization experiments are performed in solution and involve exciting fluorescent domains with polarized light. Larger bound domains will emit less plane polarized light compared to smaller unbound molecules. Analysis of saturation curves produced by fixing ligand and varying domain concentrations is used to compute the binding constant (Kd) for a given domain-ligand pair (Hu et al. 2004; Stiffler et al. 2007). In surface plasmon resonance experiments, polarized light is used to strike an electrically conducting metal surface between a glass sensor surface and a buffer. The angle of reflected light is detected by a sensorgram and changes as domains, which are immobilized on the sensor surface, interact with ligands. The Kd for a given domain-ligand pair may be computed using the measured association and dissociation rates at different domain concentration (Fournane et al. 2011). These techniques have been applied to detect interactions involving PRMs such as PDZ, SH2 and SH3 domains (Pawson et al. 2001; Tong et al. 2002; Hu et al. 2004; Stiffler et al. 2007; Tonikian et al. 2008).

1.3.2 Computational methods

Computational methods to predict PDZ domain-peptide interactions are based on established bioinformatics, statistical and machine learning techniques. These methods have been used to successfully predict interactions for proteins containing PRMs such as the SH2, SH3 and protein serine–threonine kinase domains (Yaffe et al. 2001; Brinkworth et al. 2003; Lehrach et al. 2006; Chen et al. 2008; Wunderlich and Mirny 2009).

1.3.2.1 Sequence-based methods

The position weight matrix (PWM) is a fast and simple method that captures a domain's binding preferences and can be used to score a list of potential peptide binders. A PWM is constructed based on a set of verified ligands and is a matrix of the probabilities of observing a particular residue at a given ligand position. PWMs are commonly used to compute a score indicating the binding preference of a domain for a given peptide. Tonikian et al. used PWMs to predict human PDZ interactions and to identify viral proteins that mimicked domain specificities (Tonikian et al. 2008). Stiffler et al. developed a variant of the PWM that contained weights describing the relative preference of a PDZ domain for amino acids at positions in the ligand compared to other domains (Stiffler et al. 2007). The inherent limitation of PWMs is their inability to model dependencies between ligand residue positions. PWMs may also perform poorly when there are too few experimentally determined peptide ligands available for a given protein. Furthermore, the PWM model cannot easily consider additional biological information to help reduce the number of false positives.

Other more sophisticated methods have also been used to build predictors of sequence-based PDZ domain-peptide interactions. These predictors learn patterns from the primary amino acid sequences of the domains and peptides of interacting and non-interacting interactions. Eo et al. used a support vector machine (SVM) to predict such interactions, although limited to those involving G-coupled proteins (Eo et al. 2009). Chen et al. used a Bayesian method to predict interactions for the entire PDZ domain family using data from a protein microarray experiment (Chen et al. 2008). The authors demonstrated their model's ability to predict mouse PDZ domain-peptide interactions and, to a lesser extent, interactions in other organisms. Shao et al. developed a regression framework using positive (quantitative) and negative (qualitative) mouse PDZ domain interaction data to predict PDZ domain-peptide binding affinity (Shao et al. 2011). While these methods can predict PDZ domain interactions, their common limitation is that they were trained and validated using limited interaction data for only a subset of PDZ domains. Thus, it is unclear if these can be used to predict interactions for all PDZ domains on a proteome scale.

1.3.2.2 Structure-based methods

Structural features within the domain-binding pocket of the PDZ domain play an important role in determining binding specificity (Skelton et al. 2003; Appleton et al. 2006; Chen et al. 2007). Since domain structure features capture different information about binding compared to sequence features, training with such features should result in a predictor that is complementary to sequence-based predictors. In particular, structure-based predictors would be less dependent on sequence similarity and would predict additional interactions not predicted by sequence-based predictors. Structure-based predictors have been developed to more generally predict PPIs and domain-peptide interactions mediating PPIs. For instance, Hue et al. used a SVM to predict PPIs using a kernel derived from protein structure information (Hue et al. 2010). Other methods using structure information to predict domain-peptide interactions have also been developed. Sanchez et al. used an empirical force field to calculate structure-based energy functions for human SH2 domain interactions (Sanchez et al. 2008). Fernandez-Ballester et al. constructed PWMs of all possible SH3-ligand complexes in yeast using homology modelling (Fernandez-Ballester et al. 2009). Smith et al. used protein backbone sampling to predict binding specificity for 85 human PDZ domains (Smith and Kortemme 2010). Kaufmann et al. developed an optimised energy function to predict the binding specificity of PDZ domain-peptide interactions for 12 PDZ domains (Kaufmann et al. 2011).

1.3.2.3 Integration methods

Bayesian integration is a widely used method for estimating the probability of interaction for a given PPI based on diverse data sources. It is often used due to its simple probabilistic framework and ability to handle missing data. Jansen et al. used Bayesian networks on a feature set of experimental PPI data and genomic features such as, mRNA co-expression, biological function, and essentiality in yeast (Jansen et al. 2003). Rhodes et al. used a semi-naive Bayesian classifier to combine homologous PPI, gene expression, GO Process and domain based sequence datasets in humans (Rhodes et al. 2005). Scott and Barton extended the probabilistic framework for the prediction of human PPIs to include local network topology, co-expression, orthology to known interacting proteins, sub-cellular localization, co-occurrence of domains and post-translational modifications (Scott and Barton 2007). Patil and Nakamura used a naïve Bayes classifier as a means to assign reliability to the PPIs in yeast determined by high-throughput experiments (Patil and Nakamura 2005). Li et al. closely followed the work of Rhodes et al. and used a naïve Bayes classifier to combine different types of indirect biological features (Li et al. 2008).

1.4 Machine learning framework for PDZ domain-peptide interaction prediction

Many existing predictors use machine learning to address the limitations of simple methods such as the PWM. In general, these methods are often binary classifiers that discriminate between objects from two classes using previously available information about those objects. For instance, a predictor may decide if a given PDZ domain and peptide pair will physically interact by analyzing the properties of known interacting and non-interacting PDZ interactions (e.g. primary, secondary or tertiary protein structural features of interactors). Although many types of machine learning methods exist, some of which can also perform quantitative, probabilistic, or multi-class predictions, the construction of a successful predictor follows the same general steps outlined below.

1.4.1 Data Collection

Extensive training data about known domain-peptide interactions (positive examples) and non-interacting domain-peptide pairs (negative examples) are collected from available sources. For PDZ domain-peptide interaction prediction, high throughput interaction data sets from mouse protein microarray and human phage display experiments have frequently been used for training (Stiffler et al. 2007; Tonikian et al. 2008). An important difference between these two data sets is that the protein microarray interactions involve genomic peptides while the phage display interactions may contain non-genomic peptides. Therefore, additional filtering to enrich for genomic peptides in non-genomic phage display data sets is required if the intended application is for example proteome scanning (Hui and Bader 2010). Recently, another data set for fly from a yeast two hybrid study has also become available and may also be suitable for training (Lenfant et al. 2010).

Smaller data sets are useful for blind testing (i.e. testing using examples not used for training) in order to obtain unbiased assessments of predictor performance. Positive interactions for mouse, fly, and worm from protein microarray experiments and negative interactions for human from manual literature curation are available (Chen et al. 2008; Luck et al. 2011). Curated interactions in databases can also be used such as the interactions in Domino and PDZBase databases (Beuming et al. 2005; Ceol et al. 2007). Protein-protein interactions involving various domains and organisms can be obtained from interaction databases such as iRefIndex (Razick et al. 2008), which is a database consolidating PPIs from different databases including BIND (Bader et al. 2001), BioGRID (Stark et al. 2011), CORUM (Ruepp et al. 2010), DIP (Salwinski et al. 2004), HPRD (Mishra et al. 2006), IntAct (Aranda et al. 2010) and MINT (Ceol et al. 2010)

While positive examples of domain-peptide interactions are often described in the literature, reliable evidence of negative interactions is more difficult to compile. For example, phage display data only consists of positive interactions, therefore methods (i.e. random or shuffling) must be used to generate artificial negative interactions if this data set is to be used for training. Although the methods described earlier such as randomly pairing interactions or shuffling peptides have been used for this purpose, methods to generate more biologically meaninful negatives may be more appropriate. I developed such a method (discussed in Chapter 2) which uses PWMs to generate artificial negatives using known binding information (i.e. PWM negatives).

For structure-based predictors, PDZ domain structures can be collected from the Protein Data Bank (Berman et al. 2000) or homology modelled using a variety of tools found at Protein Model Portal, which is a website providing access to structure models generated by different protein structure resources (Arnold et al. 2009). The quality of the homology models is estimated by computing the number of identical residues between the target and template sequence (i.e.

template sequence identity). It has been shown that target-template sequence identity is positively correlated with model quality. In particular, state-of-the-art algorithms can always build high quality models (RMSD < 2 Å) if the target-template sequence identity is higher than 35-40%. Furthermore, there is no significant variation in model quality for targets with sequence similarity between 40-70%. If the similarity is 35%, there is no correlation (Fischer 2006; Zhang 2009). Models with greater than 50% sequence similarity to their template structure are expected to have the correct fold with most inaccuracies arising from structural variation in templates and incorrect reconstruction of loops. The QMEAN score can also be used and is a scoring function measuring multiple geometrical aspects of protein structure including torsion angle potential, secondary structure-specific interaction potentials and solvation exposure potential (Benkert et al. 2008). This score ranges from zero to one with scores closer to one indicating more reliable models. Structure-related information can then be extracted from the structures themselves and include information such as solvent accessibility, hydrogen bonding patterns, electrostatic and hydrophobic potentials.

1.4.2 Feature Encoding Methods

In a pre-processing step, information describing the interactions is systematically represented as vectors of numeric features. This may simply be a sparse binary vector of ones and zeros indicating whether or not a feature is present in a given interaction pair. For example, to encode an amino acid sequence of length five, each residue in the string is represented using a binary vector of length 20 with each bit corresponding to the presence of an amino acid type (1 = present, 0 = not present). The vectors are then concatenated to form a final feature vector of length 100. This can be done for both domain and peptide sequences and the two vector concatenated to form the final feature vector. A more informative representation which also captures which domain and peptide residues are in contact is the ‘contact map’ encoding first described by Chen et al. A contact map contains information about contacting residues in the domain binding site and peptide derived from a protein structure of a PDZ domain complexed with a peptide ligand (Chen et al. 2008). In total, 16 domain binding site positions found to be in contact (< 5.0 Å) with the last five peptide positions were used, based on the three dimensional structure of the mouse α1-syntrophin PDZ domain in complex with a heptapeptide. This corresponded to 38 contacting domain and peptide position pairs. Each amino acid residue pair is numerically encoded as a binary vector of length 400 representing a 20 x 20 binary matrix to capture all possible amino acid pairs. The final encoding consists of a binary vector of size 15200 (38 x 400). Contact maps for other domains are constructed via a multiple sequence alignment. Variations of the contact map can also be created by including fewer or more contacting residue positions. Predictor overfitting may be a concern when such high dimensionality vectors are used for training and therefore it is important select methods such as support vectors machines which have a built in regularization strategy to help avoid this problem.

Non binary information such as structure-based features (i.e. electrostatic potentials, accessible volume) can also be encoding using dense vector encodings. For example, this may involve filling a vector with numeric values describing the feature values in each cell for each domain position considered and concatenating all vectors to form the final feature vector. Peptide information can be encoded using the binary sparse encoding described above. Domain structure and peptide sequence vectors are then concatenated to form the final feature vector.

Finally, feature values may need to be scaled to fall in the range of 0 and 1.0 (or -1 and 1) to avoid feature values with greater numeric ranges from dominating those with smaller numeric ranges and to avoid numerical difficulties during predictor training. This is necessary when using methods such as support vector machines.

1.4.3 Machine Learning

In this thesis, the PDZ domain-peptide interaction prediction problem is treated as a binary classification problem (i.e. binds or does not bind). This is a simplification of the real system because PDZ domains bind their targets at different strengths typically in the micromolar range, where affinity is described by the Kd for example. However, in order to build a quantitative predictor capable of predicting domain-peptide binding strength, training data which includes binding affinity information is required for both positive and negative interactions. Unfortunately, this information is not available in many data sets. Therefore, in order to train with all interaction data available, we focused on the simpler version of the problem by predicting if a given PDZ domain interacts with a given peptide.

1.4.3.1 Support Vector Machine

The support vector machine is a binary machine learning classifier (i.e. yes or no predictions) and was therefore selected as the method of choice to model the problem of PDZ domain-peptide interaction prediction (**Figure 1-2**).

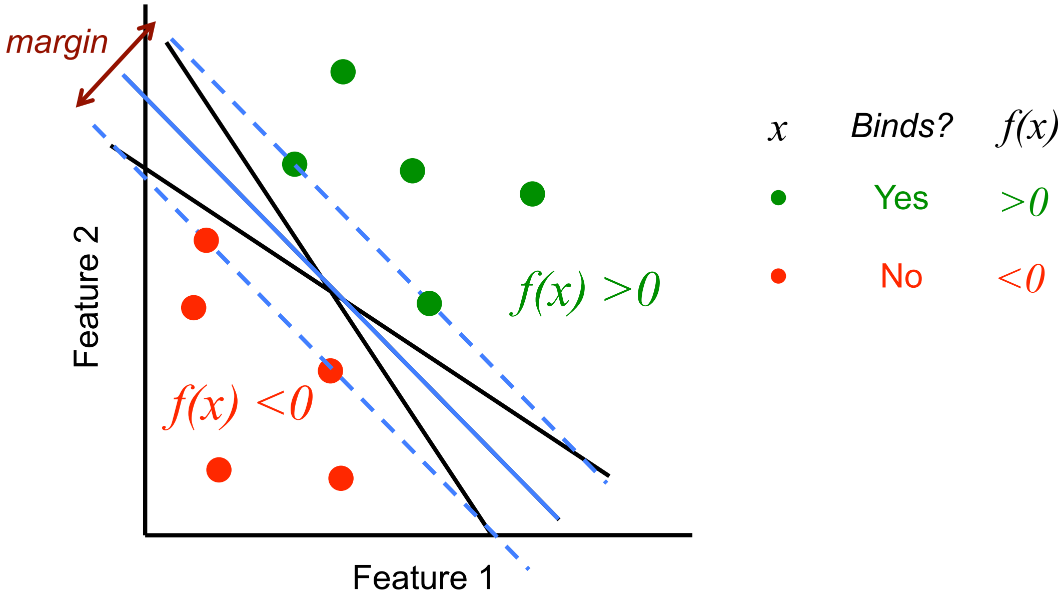


Figure 1-2 Illustration of support vector machine binary classification in 2D. Interaction data is represented in 2D (i.e. two features only) with the red and green points corresponding to positive and negative feature encoded interactions respectively. The SVM tries to optimally separate the two classes. While there are many separating hyperplanes (black lines), the SVM finds the one with the maximum margin (dotted blue lines). The predictions or class membership of any given point is calculated using a decision function *f(x)*. If *f(x)* is greater than 0, the predicted class is positive. If *f(x)* is less than 0, predicted class is negative.

Specifically, given interaction training data consisting of m examples $(x_1,y_1),\ldots,(x_m,y_m)$ where x_i is a feature vector for domain i and peptide i and y is a class label such that $y_i = \{-1, +1\}$, the SVM assigns a class label of +1 if a given interaction feature vector encodes a positive interaction or -1 otherwise (Cristianini and Shawe-Taylor 2000). The decision function is evaluated to assign the binary label:

$$f(\mathrm{x}) = \mathrm{sgn}(\mathrm{w} \bullet \mathrm{x} + b) \quad \textbf{Eq. 1.1}$$

where sgn(0) = +1, otherwise -1. The weight vector w and bias term b describe a maximum margin hyperplane (w,b) that separates positive and negative training examples. For such a hyperplane:

$$\mathrm{w} = \sum_{i=1}^{m} \alpha_i y_i \mathrm{x}_i \quad \textbf{Eq. 1.2}$$

where the α_i's are positive real numbers that maximize the following objective function:

$$\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathrm{x}_i, \mathrm{x}_j) \quad \textbf{Eq. 1.3}$$

subject to the contraints $0 \le \alpha_\mathrm{i} \le C$ for all $i = 1,..,m$ and $\sum_{\mathrm{i=1}}^{\mathrm{m}} \alpha_\mathrm{i} y_i = 0$

where $K(\mathrm{x_i},\mathrm{x_j})$ can be thought of as describing the similarity between two feature vectors, and C is a cost parameter that penalizes training errors. The radial basis function (RBF) kernel was used and is defined as:

$$K(\mathrm{x_i},\mathrm{x_j}) = e^{-\gamma \|\mathrm{x_i} - \mathrm{x_j}\|^2} \quad \textbf{Eq. 1.4}$$

A grid search was used to find locally optimal values for γ and C (Hsu et al. 2010). Instead of explicitly balancing the positive and negative training examples, weighted costs were used according to $C^+ = (n^+/n^-)\, C^-$, where n^+ is the number of positive training interactions and n^- is the number of negative training interactions.

1.4.3.2 Naïve Bayesian Integration

In order to determine if a predicted PDZ protein-protein interaction pair is likely to be physiologically relevant, a Bayesian protein-protein interaction prediction model is used. This method estimates the probability that a given protein pair interacts conditioned on the biological evidence in support of that interaction. A näive Bayesian model simplifies this problem by assuming complete independence between different types of biological evidence. In this thesis, a protein pair is described by a set of features: X_1 = gene expression, X_2 = cellular component, X_3 = molecular function, X_4 = biological process, X_5 = sequence signature, X_6 = binding site conservation. A näive Bayes protein-protein interaction prediction model is then defined as:

$$\underset{Y}{\operatorname{argmax}} P(Y \mid X_1, X_2, ..., X_n) = \underset{Y}{\operatorname{argmax}} \frac{P(X_1, X_2, ..., X_n \mid C) P(Y)}{P(X_1, X_2, ..., X_n)}$$

$$= \underset{Y}{\operatorname{argmax}} P(Y) \prod_i P(X_i \mid Y)$$

$$\underset{Y}{\operatorname{argmax}} P(Y \mid X_1, X_2, ..., X_n) = \underset{Y}{\operatorname{argmax}} \ \log(P(Y)) + \sum_i \log(P(X_i \mid Y)) \qquad \textbf{Eq.4.1}$$

where $P(Y)$ is the class prior probability and $P(X_i|Y)$ is the class-conditional probability. The model is trained on the gold standard training set consisting of positive and negative PPIs.

1.4.4 Predictor Performance Evaluation

Various methods can be used to assess predictor performance. Typically, cross validation methods are used to obtain an estimate of performance with more rigorous blind testing performed using unbiased test examples that were not used for predictor training. A summary of predictor performance can be visualized by plotting receiver operating characteristic (ROC) curves and precision/recall (PR) curves and a single numeric value representing overall performance may be obtained by computing the area under these curves. In the case of domain-peptide binding performance, a domain's predicted and known binding preferences can be visualized and compared using sequence logos and to gain a more general idea of predictor performance.

1.4.4.1 Cross Validation and Blind Testing

Predictor performance can be estimated using various cross validation strategies. A common strategy is ten fold cross validation. This involves partitioning the training data into ten randomly selected interaction sets, independently holding out each set for testing against a predictor trained using the remainder of the data, and computing average performance across all ten runs. Other variations of cross validation designed to estimate predictor performance when specific sets of domains and/or peptides are held out involve holding out 12% of the domains, 8% of the peptides and both 12% of the domains and 8% of the peptides and tested on the rest, again repeating this ten times (Chen et al. 2008). Blind testing should also be performed to obtain an unbiased measure of predictor performance using unseen test data (i.e. data not used for training).

1.4.4.2 ROC and PR Curves

The overall performance from these testing strategies is summarized by computing the area under the ROC and PR curves (**Figure 1-3**). Although ROC curves are insensitive to class distribution changes (i.e. proportion of positives versus negatives changes), PR curves are not (Fawcett 2006). Therefore both curves and AUC scores should be used to provide a more accurate assessment of predictor performance.

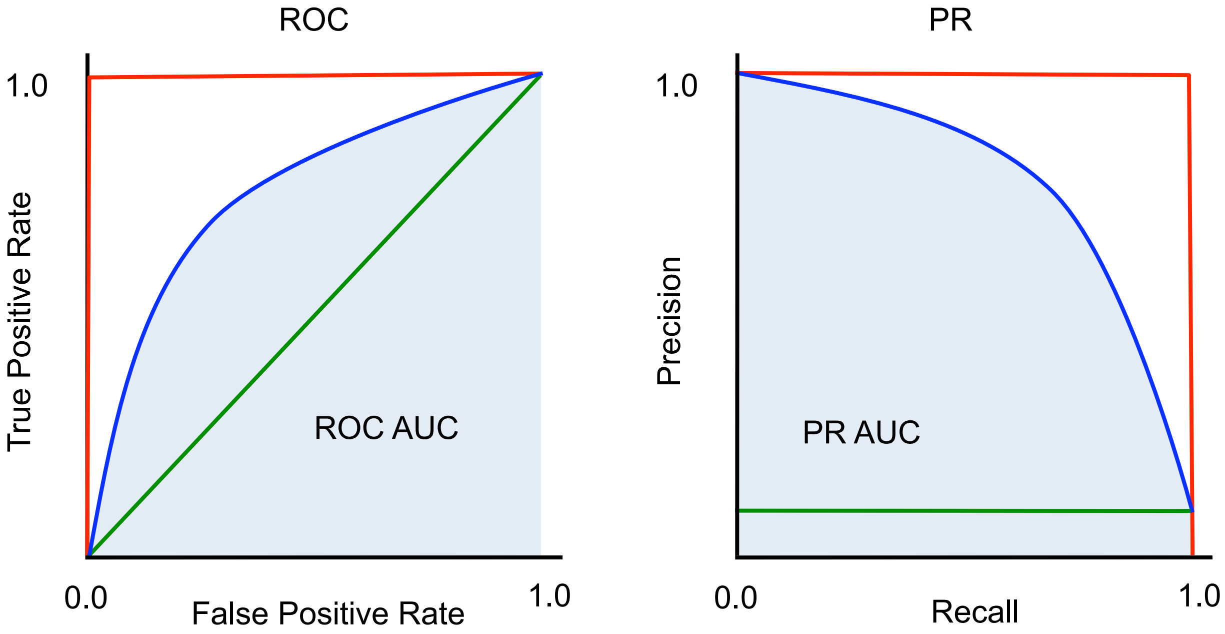


Figure 1-3 ROC and PR curve examples. (Left) Different ROC curves are plotted with the red curve representing perfect predictor performance (i.e. always correctly discriminates between positive and negatives), blue curve representing typical predictor performance (i.e. most of the time correctly discriminates between positives and negatives) and the green curve representing random predictor performance (i.e. randomly discriminates between positives and negatives). (Right) Different PR curves are plotted with the red, blue and green curves representing perfect, typical and random predictor performance. Area under the curves for typical predictor performance is highlighted with light blue shading.

For both curves, predictions are first sorted by decision value (high to low). The ROC curve measures the predictor's true positive rate (TPR) versus false positive rate (FPR) on the set of ordered predictions as the decision value is relaxed (i.e decreased). The curve will be increasing (i.e. TPR increases faster than FPR) with the area under the curve typically ranging between 0.5

and 1.0. The closer to 1.0, the better the performance with a perfect predictor having an ROC AUC of 1.0. Random predictor performance will yield a diagonal curve with AUC of 0.5.

Similarly, precision/recall curves measure the precision versus recall tradeoff as the decision value is relaxed. The shape of this curve will typically be decreasing (i.e. precision falls and recall increases). In general, the area under this curve ranges between a value representing random performance (i.e. number positives/number of examples) to 1.0. The higher the AUC score the better the performance with a perfect predictor having a PR AUC of 1.0.

Individual statistics may also be computed to measure predictor performance including:

- Sensitivity or Recall: TP/(TP+FN)
- Specificity: TN/(TN+FP)
- Precision (PR): TP/(TP + FP)
- F1 Measure: 2 (Precision x Recall) / (Precision + Recall)

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, FP is the number of false positives.

1.4.4.3 Sequence Logos

For a given domain, its binding specificity can be graphically represented using a sequence logo (**Figure 1-4**).

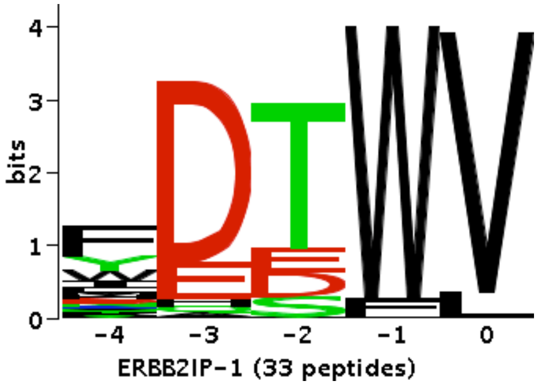


Figure 1-4 Example of a sequence logo for ERBB2IP-1 PDZ domain.

Such a logo helps to visualize the preference of specific residues at different positions of the ligand. These are useful to gain an overall idea of the agreement between predicted and known binding preferences for a given domain. A list of binders is required to create the logo. At each position, the frequency of a residue is depicted by the height of the amino acid character representing it. Residues are stacked on top of each other with the heights of each stack being proportional to the information content (Shannon entropy) measured in bits.

1.5 Thesis Rationale

Protein-protein interactions are involved in almost all biological processes and have been experimentally detected using a variety of methods. Because it is often difficult to determine these interactions experimentally and due to the availability of large high throughput interaction data sets currently available, computational interaction prediction methods have been developed. These methods rely on accurate three dimensional structures of both proteins which are not always available. Often, features of both structures such as surface complementarity, electrostatics and flexibility (i.e. induced fit) must also be solved resulting in a highly complex problem. Therefore, the computational prediction of protein-protein interactions in general is difficult.

However, the ability to accurately predict protein-protein interactions directly from the proteome would enable the assembly of more comprehensive protein interaction networks. These networks could be analyzed to provide insight on the extent to which cellular proteins are connected, how different biological pathways may be linked and how the underlying cellular processes are organized. Fortunately, many protein interactions, such as those mediated by PRMs like PDZ domains, bind their targets through short simple linear motifs. Therefore, the problem of predicting PRM-mediated PPIs is easier solve as factors which make the prediction of PPIs in general (i.e. induced fit, flexibility of large protein molecules) need not be modelled. This would allow a large subset of interactions to be accurately predicted and is a feasible start to solving the more general problem of protein-protein interaction prediction.

My thesis focuses on the computational prediction of PDZ mediated protein-protein interactions. PDZ domains are an ideal model to study this problem because their modes of target recognition are one of the simplest and they are well studied (they have important biological roles and many experimental data sets are available). I will present two predictors (i.e. sequence-based and structure-based) that I built to scan the proteomes of different organisms for interactors of PDZ domains. Using additional biological evidence, I identified a subset of physiologically relevant and high confidence interactions from sequence-based and structure-based predictions. These interactions were used to construct the most comprehensive physiologically relevant PDZ mediated protein-protein interaction network in human to date.

Chapter 2

Predicting PDZ Mediated Protein-Protein Interactions From Sequence

This work was published in *BMC Bioinformatics*, **11**:507: Hui, S., Bader, GD. (2010), Proteome scanning to predict PDZ domain interactions using support vector machines.

Author contributions: I collected the data, developed and implemented the methods and performed the analyses. Gary D. Bader supervised and advised this project.

2 Predicting PDZ mediated protein-protein interactions from sequence

2.1 Introduction

The biological importance of PDZ domains, their simple modes of target recognition and the availability of experimentally determined interactions have prompted the development of PDZ domain interaction prediction methods by multiple groups. Such methods are based on established techniques, which have been used with success to predict interactions for SH2 and SH3 domains, protein serine–threonine kinases and major histocompatibility complex molecules (Yaffe et al. 2001; Donnes and Elofsson 2002; Brinkworth et al. 2003; Lehrach et al. 2006; Wunderlich and Mirny 2009). A practical application of a reliable PDZ domain interaction predictor would be to use it to scan the proteomes of organisms for potential binders of PDZ domains. The results would help direct future experiments to increase the coverage of current PDZ domain interaction networks and expand our knowledge of the roles that PDZ domains play in different biological processes. I developed a primary sequence based predictor of genomic interactions involving PDZ domain family members using a support vector machine. Domain-peptide interaction sequence information was represented using the contact map feature encoding method which captures Unlike published predictors, the predictor is trained using data from two independent high throughput studies using protein microarray and phage display technologies, which makes it more general. Since the phage display data consists of only positive interactions, I have overcome a major issue, which has up to now prevented its straightforward use for predictor training. I addressed this by developing a method for the generation of artificial negative interactions from data consisting of positive interactions only. This method generates more biologically meaningful negatives compared to other commonly used methods that use randomization or shuffling. Through independent testing with published genomic data sets, I demonstrated the predictor's ability to accurately predict interactions in multiple organisms (Chen et al. 2008). I then used the predictor to scan human, worm and fly proteomes to predict binders for different PDZ domains. These predictions were validated using known genomic interactions from PDZBase and protein microarray experiments (Beuming et al. 2005; Chen et al. 2008). Finally a comparison of proteome scanning performance, which depends on minimizing the number of false positives generated, showed the predictor's improved accuracy and precision compared to published predictors. Predicted interactions matched many known protein-protein interactions and were enriched in known and novel biological processes, suggesting that many more predictions are likely to be correct.

2.2 Results

2.2.1 The predictor achieves high cross validation results

The predictor achieved high AUC scores from multiple cross validation testing. The highest ROC and PR AUCs of 0.939 and 0.896 respectively were obtained when 10% of interactions were held out for testing. For tests that involved holding out all interactions for a given domain, the AUC scores were lower. In particular, the leave-12%-of-domains-out test yielded ROC and PR AUC scores of 0.851 and 0.764 and the leave-12%-domain-and-8%-peptides-out test yielded ROC and PR AUC scores of 0.87 and 0.794. Finally, the leave-8%-peptides-out test yielded higher ROC and PR AUCs of 0.893 and 0.838 (**Figure 2-1**).

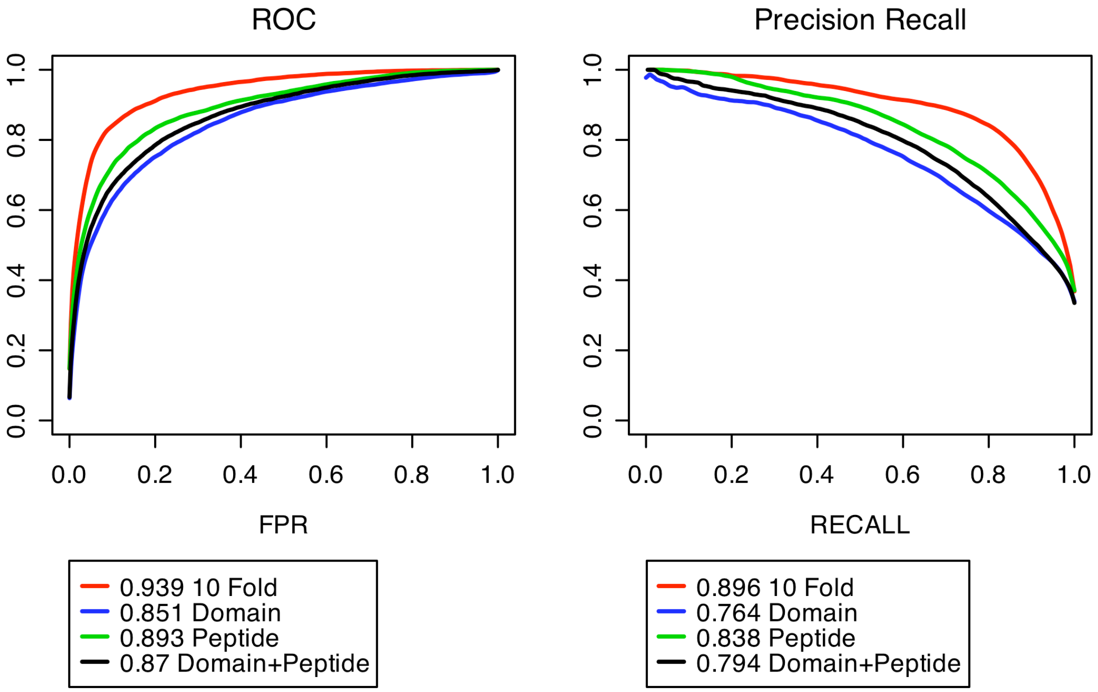


Figure 2-1 Predictor performance estimation using cross validation. Predictor performance measured using ten fold (red), leave-12%-of-domains-out (blue), leave-8%-of-peptides-out (green), leave-12%-of-domains-and-8%-of-peptides-out (black) cross validation.

2.2.2 Predictor performance depends on nearest neighbour training domain sequence similarity

The lower AUC results for leave-domain-out cross validation strategies suggests that the predictor's performance for a given test domain depends on its level of similarity to the training domains. To determine the degree of this dependency I performed leave-one-domain-out cross validation and divided the AUC scores according to the binding site similarity of the held out domain to that of its nearest training neighbour. The similarity between two binding site sequences *a* and *b* was computed as:

$$Similarity(a,b) = \frac{\sum_{i=1}^{n} match(a_i b_i)}{n}$$

Eq. 2.1

where match is 1 if a_i=b_i and n is the length of the sequences. This was repeated using a simple nearest neighbour predictor (NN) and the results were compared. The nearest neighbour predictor determines whether a given interaction is positive or negative using a nearest neighbour criterion. The nearest neighbour criterion is evaluated by computing the similarity between a test interaction and all other training interactions (where interactions are represented as a domain binding site–peptide sequence pair). The training interaction with the lowest distance is then set to be the test interaction's nearest neighbour. Thus if the nearest neighbour is a positive interaction, the test interaction is predicted to be positive, otherwise it is predicted to be negative. In total, interactions for 82 mouse domains from protein microarray and 20 human domains from phage display were used to build the NN predictor.

The results showed that, indeed, the predictor achieves higher performance for domains that are more similar to the training set. In particular, the predictor was on average better than the nearest neighbour method for testing domains with over 60% sequence similarity to their nearest training neighbour (**Figure 2-2** Top Row). Presumably, this means the predictor learned non-trivial patterns in the data features instead of simply indentifying similarities in the sequences as the NN predictor did. For tested peptides, this dependence was not as apparent, which indicates that the predictor's performance is more dependent on domain sequence similarity than peptide sequence similarity (**Figure 2-2** Bottom Row).

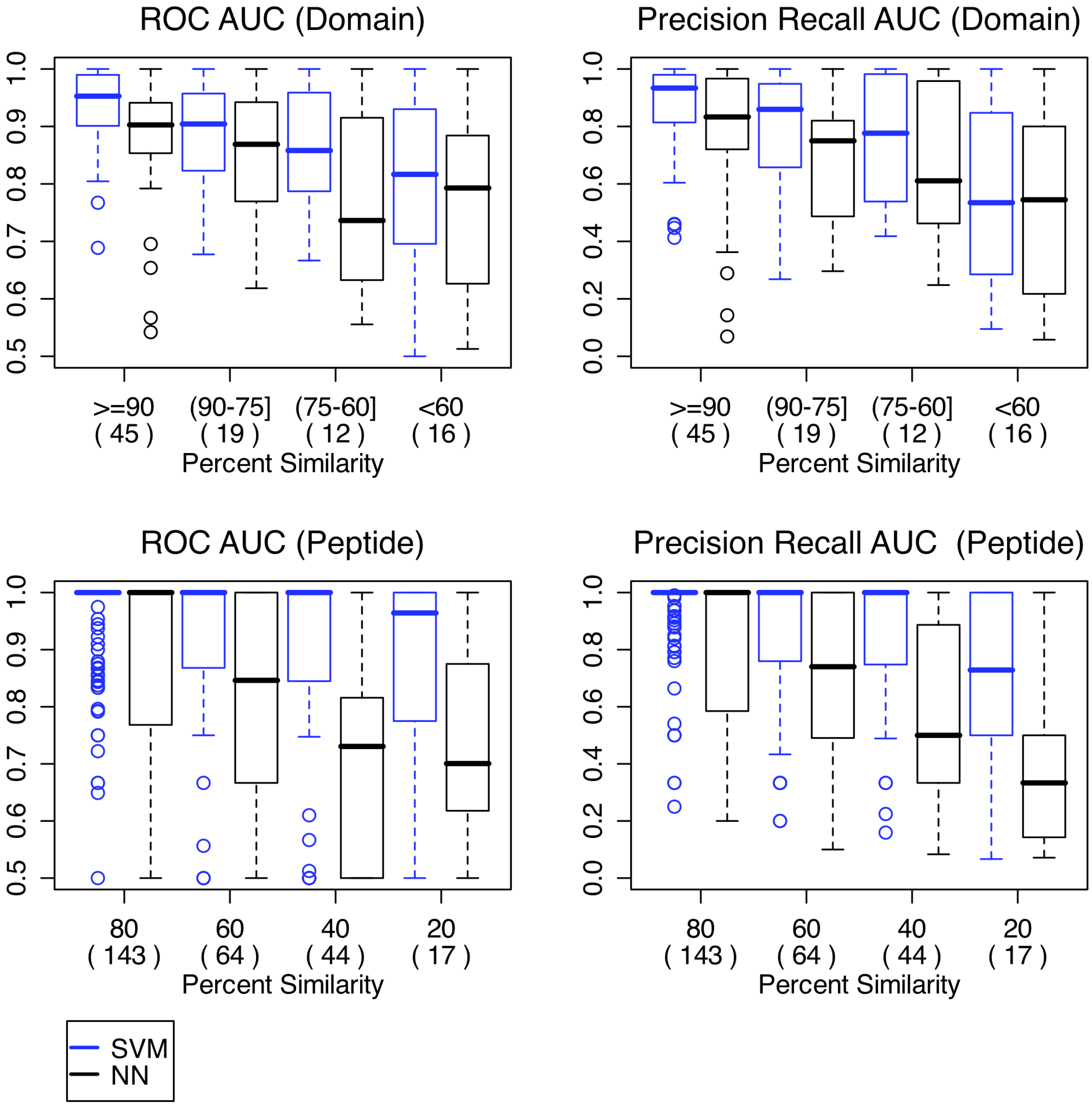


Figure 2-2 Predictor performance dependence on testing and nearest training neighbour sequence similarity. (Top Row) Using leave-one-domain-out cross validation, domain specific ROC and PR AUC scores for predictor (blue) and nearest neighbour predictor (black) were grouped according to a given testing domain's similarity to its nearest training neighbour. (Bottom Row) The same was done for peptides using leave-one-peptide-out cross validation. The similarity between two domains was calculated as the percentage of matched residues between their binding site sequences. The similarity between two peptides was calculated as the percentage of matched residues. Numbers in parentheses indicate the number of domains or peptides in each boxplot.

2.2.3 Evaluating predictor performance using a series of independent tests across organisms

I next validated the choice of data and methods for three major parameters affecting predictor performance: training data, feature encoding and artificial negatives. Each parameter was examined independently by comparing the predictor to other SVMs built using different values for the parameter of interest while holding the other two parameter values fixed. Predictor performance was assessed using data for mouse, worm and fly from independent protein microarray experiments, which all contain positive and negative interactions (Chen et al. 2008) (**Table 2-1**).

		Domains		**Interactions**	
Organism	**Source**	**# Pos**	**# Neg**	**# Pos**	**# Neg**
Fly	Protein microarray	7	7	34	106
Worm	Protein microarray	6	6	59	88
Mouse	Protein microarray	11	19	52	74
Human	PDZBase	13	-	38	-

Table 2-1 Summary of data for independent genomic testing and prediction validation.

2.2.3.1 Training with both mouse protein microarray and human genomic-like phage display data improves predictor performance

I first validated the use of mouse protein microarray and human genomic-like phage display data for training. The predictor was compared to other SVMs built using data from single experimental data types (mouse/protein microarray or human/phage display), both experimental data types (mouse/protein microarray and human/phage display) and both experimental data types but with human phage display data enriched in genomic-like or non genomic-like interactions. For all predictors, contact map features were used to encode the data and PWMs were used to generate artificial negatives. A comparison of predictor performance showed that the SVM built using mouse and human genomic-like data for training was better than the other SVMs for the worm and fly tests (**Figure 2-3** Top Row). All predictors had lower scores for the mouse test. To explain the latter observation, for each test I computed the binding site similarity of each testing domain to its nearest training neighbour. I found that the mouse domains were on average 65% similar to their nearest training neighbours, while the worm and fly testing domains were on average 80% and 87% similar to their nearest training neighbours respectively. Therefore the observed pattern of performance was consistent with the earlier observation that predictor performance decreased as the similarity between testing domains to their nearest training neighbours decreased. These results validate the use of both mouse protein microarray and human genomic-like phage display interactions for predictor training.

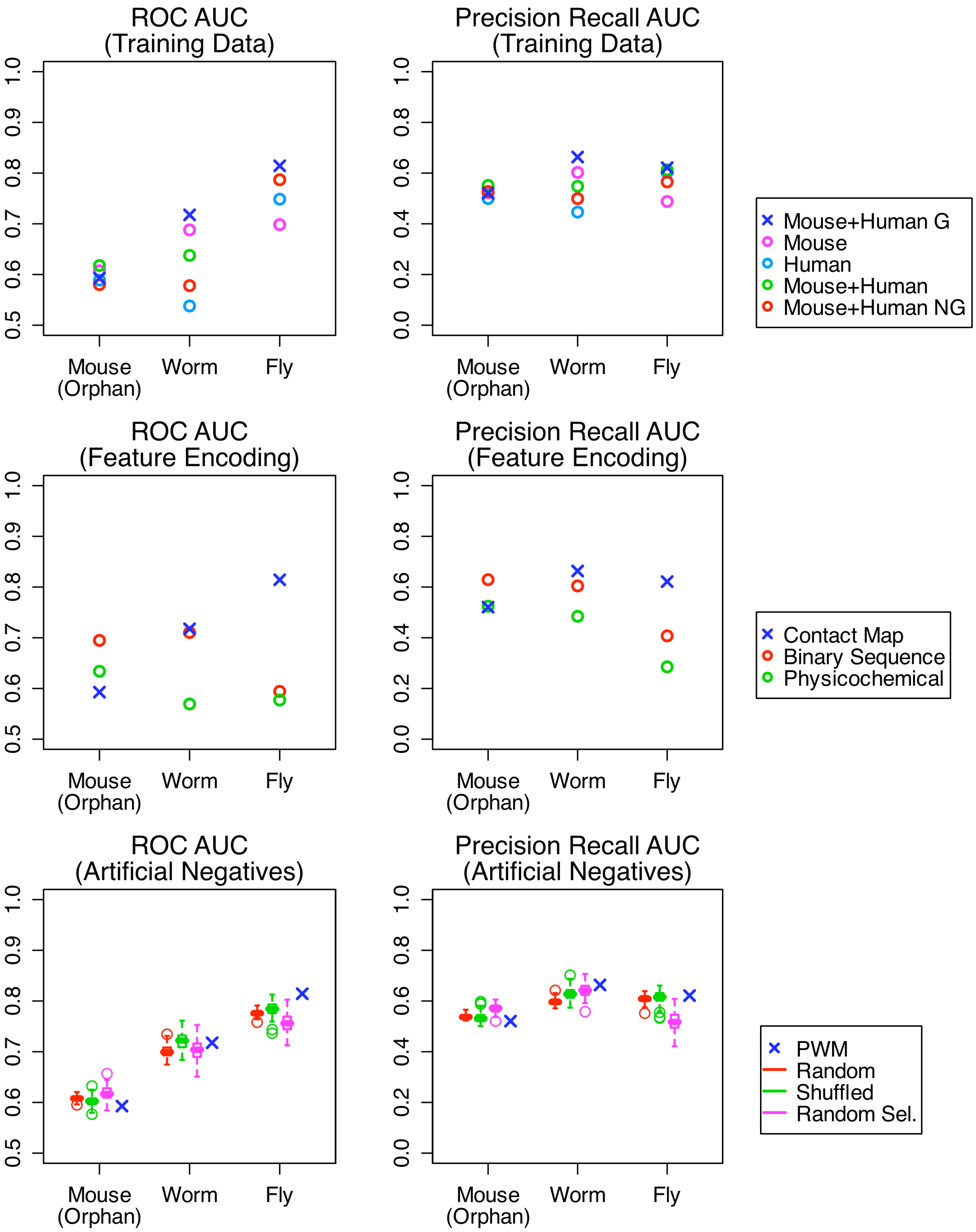


Figure 2-3 Comparison of independent genomic test performance of different SVMs. Blue x denotes data or method used by the predictor in all panels. (Top Row) A comparison of predictors trained using data from one experiment: mouse from Chen et al. (magenta) or human from Tonikian et al. (light blue), from two experiments: mouse and human (green) and from two experiments with data enriched in genomic-like or non genomic-like human data: mouse and genomic-like human (blue) and mouse and non genomic-like human (red). (Middle Row) A comparison of predictors trained using data encoded using different feature encodings: binary sequences (red), physicochemical properties (green), contact map (blue). (Bottom Row) A comparison of predictors trained using different methods for generating artificial negatives for phage display: random peptides (red), shuffled peptides (green), randomly selected peptides (magenta), PWM selected peptides (blue). One hundred different predictors trained using different random, shuffled and randomly selected peptides were built.

2.2.3.2 Contact map feature encoding is better compared to other sequence based feature encoding strategies

I next validated the choice of using the contact map feature encoding. The predictor was compared to SVMs built using binary sequence or physicochemical property-based encodings. All predictors used mouse protein microarray and human genomic-like training data and PWMs to generate artificial negatives. For the binary sequence encoding, binary vectors were created using a vector of length 20 with each element representing an amino acid and initially set to zero. A single residue was represented by placing a one in the position representing that residue. A binary vector was created for each residue in a domain-peptide interaction pair, with the final vector of length 20 amino acids x (length of domain binding site sequence + length of the peptide sequence). For physicochemical features, a vector of five real numbers describing over 500 different physicochemical properties for each amino acid residue was created for a domain-peptide interaction sequence (Atchley et al. 2005). Thus, final vectors were of length 5 x (length of the domain binding site sequence + length of the peptide sequence). The predictor performance comparison showed that except for the mouse test, the SVM trained using contact map feature encoded data had the highest scores (**Figure 2-3** Middle Row). I again attributed the low performance on the mouse test to the dissimilarity of the test domains to the training domains. Predictors with better mouse test performance did not generalize to the worm or fly tests, supporting the conclusion that the mouse test is not ideal. These results indicate that the contact map feature encoding for predictor training is better compared to binary and physicochemical property based encodings.

2.2.3.3 The use of PWMs is a valid method for generating artificial negatives

Finally, I validated the use of PWMs for generating artificial negatives for the phage display training data (i.e. PWM negatives). The predictor was compared to other SVMs built using random, shuffled, and randomly selected artificial negatives. All predictors used mouse protein microarray and human genomic-enriched phage display training data encoded using contact map features. Random negatives were created using random residues concatenated into peptides of length five. Shuffled negatives were created by shuffling residues in the positive peptides. Randomly selected negatives were created by randomly selecting peptides from the same set of peptides used to select negatives in the PWM method. I created 100 different artificial negative data sets from the phage display data and measured the mean predictor performance. Over all the tests, the average predictor ROC and PR AUC scores were 0.71 and 0.60, respectively, which were slightly higher than the overall average ROC and PR scores for the other predictors (**Figure 2-3** Bottom Row). Specifically, the average ROC and PR scores were 0.70 and 0.58 for random negatives, 0.70 and 0.59 for shuffled negatives and 0.69 and 0.58 for randomly selected negatives. Although the scores were similar for all predictors within each test, the average ROC and PR scores for mouse, worm and fly tests showed that all predictors performed poorly for the mouse test but were better for the worm test. For the fly test however the predictor using PWM negatives was in general better. This suggests that the PWM negatives are a reasonable choice for artificial training negatives with its importance for improving predictor performance more evident in cases where the testing domain is highly similar to the training domains.

2.2.4 Proteome scanning predictions are validated by known PDZ domain peptide interactions in human, worm and fly

The predictor was used to scan the human proteome (defined by genome assembly Ensembl:GRCh37.56) (Hubbard et al. 2009) to predict binders for 13 human PDZ domains with available validation data in PDZBase (Beuming et al. 2005). In total, 41,193 unique transcript tails of length five, out of 77,748 transcripts corresponding to 23,675 genes from the human proteome, were scanned. I also scanned the worm and fly proteomes (defined respectively by genome assemblies Ensembl:WS200.56 and Ensembl:BDGP5.13.56) (Hubbard et al. 2009) for binders for six and seven PDZ domains respectively, with known genomic interactions (Chen et al. 2008). For worm, 19,864 unique transcript tails of length five, out of 27,533 transcripts corresponding to 20,158 genes, were scanned. For fly, 14,691 unique transcript tails of length five, out of 21,309 transcripts corresponding to 20,158 genes, were scanned. In all cases, very few known genomic interactions per domain (on average 2.2 human, 4.2 worm and 9.8 fly) were available for validation of the domains tested making accurate assessment of predictor performance difficult. Furthermore, a blind negative interactions were unavailable and therefeore false positive rates could not be computed. Nonetheless, the results reported here serve as a reasonable performance estimate.

For human, over 65% (19 out of 29) of PDZBase interactions 13 human domains were correctly predicted. Of the three remaining domains, MAGI2-2 and MAGI3-1 had no PDZBase interactions correctly predicted, but these domains had only one known interaction each. Two other domains (PDZK1-1 and SNTG1-1) also had only one known interaction each however the predictor correctly identified the single interaction for these domains. Further experimental validation and more detailed literature searches should be carried out to obtain a more reliable assessment of predictor performance for these domains. For the last domain (MLLT-4), only one out of six known interactions was predicted, however compared to the other domains tested, this domain was the most dissimilar to its nearest training neighbour with a similarity of 0.68. It also had no homologs in the training data making it a challenging test case. Please see **Appendix A, Table A-1** for detailed results.

For worm and fly, 25% (15 out of 60) and 37% (11 out of 30) of protein microarray interactions respectively were correctly predicted. Although this is much lower than the human proteome scanning result, the false positive rates are both quite low at approximately 4%. In particular, in worm and fly, none of the known interactions were predicted for DSH-1 despite it having a reasonable number of known interactions (11 and three respectively) and being very similar to its nearest training neighbour (over 0.8). In fly, there were no predictions for PAR6-1 even though it too was very similar to its nearest training neighbour (1.0). Through further analysis, I found that in each case, the nearest training neighbours DSH-1 and PAR6B-1 in mouse had only three and two training interactions respectively. This suggests the possibility that predictor performance might also depend on the abundance of nearest neighbour training data. However, a single exception to this is that the predictor did not predict any known interactions for PATJ-2, which had a reasonable amount of validation data (seven interactions) and was very similar to its nearest training neighbour (over 0.81), which also had adequate data (16 interactions). Please see **Appendix A, Table A-2, Table A-3** for detailed results. Thus, in general, the predictor is more likely to correctly predict interactions for domains that are well represented in the training data in terms of sequence similarity and interaction abundance.

2.2.5 Predicted binding specificities are consistent with experimentally determined binding specificities

Since known interactions are limited, I compared the predicted and experimental binding specificities to determine if the set of predictions was consistent with their corresponding set of experimental binders, at a high level. Four of the human domains had adequate genomic-like binders from phage display experiments (ten or more), which were used to create PWMs to summarize their binding specificities. These were then graphically represented as sequence logos. For worm and fly, PWMs were created for five and three domains, respectively, that had five or more binders determined from protein microarray experiments. I then created PWMs using the corresponding predicted binders and computed the similarity between the predicted and experimentally determined binding specificities. The similarity between two PWMs *a* and *b* (i.e. binding specificities) was determined using the following:

$$\text{Distance}_{PWM}(a,b) = \frac{1}{\sqrt{2}} \sum_{i=1}^{n} \sqrt{\sum_{L \in (20aa's)} (a_{i,L}, b_{i,L})^2} \qquad \textbf{Eq. 2.2}$$

$$\text{Similarity}_{\text{PWM}}(a,b) = 1.0 - \text{Distance}_{\text{PWM}}(a,b) \qquad \textbf{Eq. 2.3}$$

where n is the number of columns in the PWM. This metric is normalized such that 0 represents perfectly similar PWMs and 1 represents perfectly dissimilar PWMs. The similarity between two PWMs is therefore 1 minus the distance.

The average PWM similarity was 67% and the predicted binding specificities corresponded to known PDZ domain binding classes I and II (**Figure 2-4**). Two domains (DSH-1 from worm, PATJ-2 from fly) had binding specificity similarities much lower than the average (less than

60%), however these results were not unexpected, given the poor results for these two domains shown above.

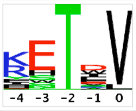
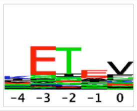
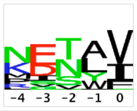
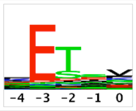
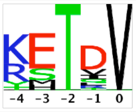
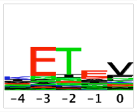
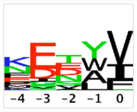
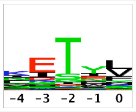
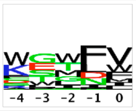
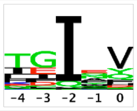
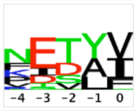
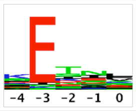
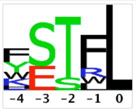
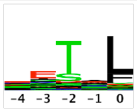
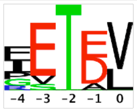
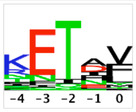
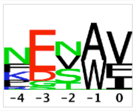
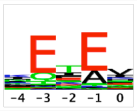
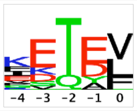
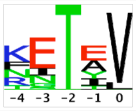
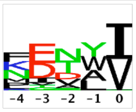
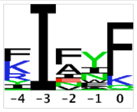
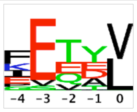
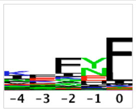
Domain Name	NN Sim	Experiment	SVM Predicted	Profile Sim	Domain Name	NN Sim	Experiment	SVM Predicted	Profile Sim
DLG1-2 Human	1			0.751	LIN7-1 Worm	1			0.688
DLG3-2 Human	1			0.682	MPZ1-6 Worm	0.69			0.729
MLLT4-1 Human	0.69			0.62	STN2-1 Worm	0.81			0.688
PDZK1-1 Human	0.81			0.691	LAP4-2 Fly	0.88			0.725
DLG1-3 Worm	0.94			0.671	LAP4-3 Fly	0.75			0.735
DSH-1 Worm	0.81			0.507	PATJ-2 Fly	0.81			0.565

Figure 2-4 Comparison of predicted and experimental binding specificities. A comparison of phage display determined and predicted PDZ domain binding specificities for the last five terminal binding positions visualized as sequence logos. For human, only domains with ten or more peptides from phage display experiments (Tonikian et al. 2008) were compared. For worm and fly, domains with an adequate (five or more) number of peptides from protein microarray experiments (Chen et al. 2008) were compared.

Although the experimental and predicted binding specificities were generally consistent, there were some discrepancies. For example the human phage display sequence logos show a clear preference for T at p-2 and V at p0 while this preference is not as strong for the predicted sequence logos. This is because phage display experiments only find optimal binders. However, such binders may not exist in the proteome, leading to the domain preferring a less optimal binder. This may be biologically advantageous as weak binders may allow for easier interaction regulation. To determine whether this was the case in the data, I scanned the human proteome with the optimal phage display PWMs and created genomic sequence logos with the top 1% of binders. The predicted sequence logos were all more similar to the genomic phage display sequences logos than they were to the optimal phage display sequence logos (**Figure 2-5**). Therefore, some discrepancies between experiment and predicted logos are not unexpected. Overall, these results show that the predicted binding specificities are generally consistent with those that are experimentally determined.

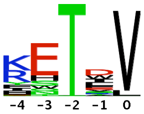
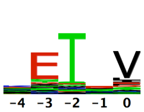
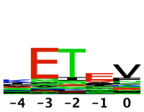
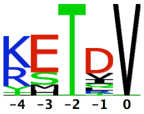
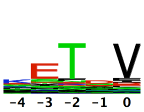
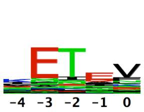
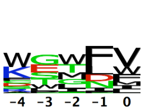
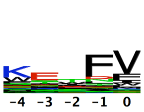
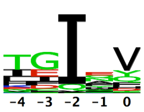
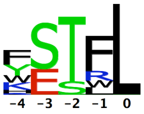
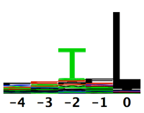
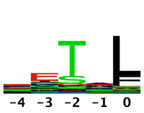
Domain Name	NN Sim	Optimal	Genomic	SVM Predicted	Optimal Profile Sim	Genomic Profile Sim
DLG1-2 Human	1				0.751	0.886
DLG3-2 Human	1				0.682	0.86
MLLT4-1 Human	0.69				0.62	0.624
PDZK1-1 Human	0.81				0.691	0.851

Figure 2-5 Comparison of optimal and genomic phage display binding specificities. Optimal and genomic phage display sequence logos were compared to the corresponding predicted SVM sequence logos for the last five terminal binding positions. Only the four human PDZ domains from **Figure 2-4** were compared.

2.2.6 Many predictions correspond to known PPIs involving PDZ domain containing proteins

To provide additional support for the predictions, I calculated how many corresponded to known PPIs. Specifically, I scanned the human proteome for potential binders for 213 human PDZ domains with known PPIs in the iRefIndex database (Razick et al. 2008). If the protein containing the given domain was found to interact with another protein whose C-terminal tail matched the predicted binder, the prediction was considered to correspond to a known PPI. The predictor successfully predicted interactions corresponding to known PPIs for 75 of the 213 PDZ domains with an average of 19% of known PPIs successfully predicted per domain (see **Appendix A, Table A-4, Table A-5** for detailed results). The number of PPIs successfully predicted per domain was significant ($p < 0.05$) for all but 19 domains. Significance testing was performed using Fisher's exact test, which asked whether the observed number of PPIs predicted for a given domain could be achieved at random. Since many PDZ domain containing proteins may contain multiple PDZ domains, it is not possible to uniquely assign a PPI to a PDZ domain. This could result in erroneous false negative or true positive statistics for the above tests, thus they should be regarded as a rough estimate of predictor performance. There were not enough PPI data in iRefIndex to carry out the same analysis for worm and fly domains.

2.2.7 The predictor is better at proteome scanning compared to other sequence based predictors

Cross validation and a series of independent tests show that the predictor can accurately predict PDZ domain-peptide interactions, however, a major issue with most predictors used to scan a proteome is the generation of too many false positives. I thus compared the proteome scanning performance of the predictor and other published prediction methods - the multidomain selectivity model (MDSM) by Stiffler et al. and the additive model by Chen et al. which are both state-of-the art and trained using mouse protein microarray data in their original publications (Stiffler et al. 2007; Chen et al. 2008). For the MDSM model, the binding preference of a given peptide was computed using the model parameters corresponding to its nearest model domain as determined by the Hamming distance between the binding site sequences. A given peptide is predicted to be positive if the binding preference score is greater than a predetermined threshold (i.e. parameter $m = 5$ according to the original publication). In total 74 mouse PDZ domains were modelled. For the additive model, I used the model parameters as specified in the tutorial provided in the supplemental material of the original publication (Chen et al. 2008). The value of tau used was -0.3978. In total, 82 mouse domains from the Stiffler et al. protein microarray experiment were used for training in the original publication. PWMs representing the baseline for comparison were built per domain using their known binders and represented their binding preferences. Thus the cells of the position weight matrices contain the log probability of each residue at each of the positions in the binding peptide. The peptdies in the phage display library were constructed using a NNK codon set (where N represents a 25% mix each of adenine, thymine, guanine, and cytosine nucleotides; and K represents a 50% mix each of thymine and guanine nucleotides). As a result, some amino acids occur more frequently than others. This bias was corrected for by dividing the PWM residue frequencies by their expected frequencies using the NNK codon set (Skelton et al. 2003). To avoid negative infinity values in the PWM, any residues with a frequency of zero were assigned the pseudocount of 0.01. The binding preference of a domain for a given peptide sequence was then computed by summing the weights in the matrix corresponding to each residue and position in the given sequence. If the score was above a specified cut off, the peptide is predicted to bind otherwise it is predicted to not bind. Using the nearest neighbour PWM of a given test domain (as determined by binding site sequence similarity), a list of peptides was evaluated and ordered in descending order by PWM score. The top 1% of this ordered list was then predicted to be binders. In total, interactions for 82 mouse from protein microarray and 20 human domains from phage display experiments as described in the paper were used to build the PWMs.

I used the F1 measure to compare predictor performance since it summarizes the precision/recall performance of a predictor and is used in document retrieval where the recovery of relevant documents from a large number of possibilities is critical. For all predictors, the majority of F1 measures are low (less than 0.1). This is likely due to the high level of incompleteness in the benchmark used to validate the predictions. However, the results show that the predictor achieves a higher average F1 measure (0.037) than the other predictors demonstrating its improved accuracy and precision. In comparison, the average F1 measures were 0.02, 0.005 and 0.016 for the MDSM, additive model and PWM predictor respectively. For fly and worm domains, the false positive rate (FPR) was approximately 4% and substantially (over four times) lower than the FPRs of the other predictors (**Figure 2-6**).

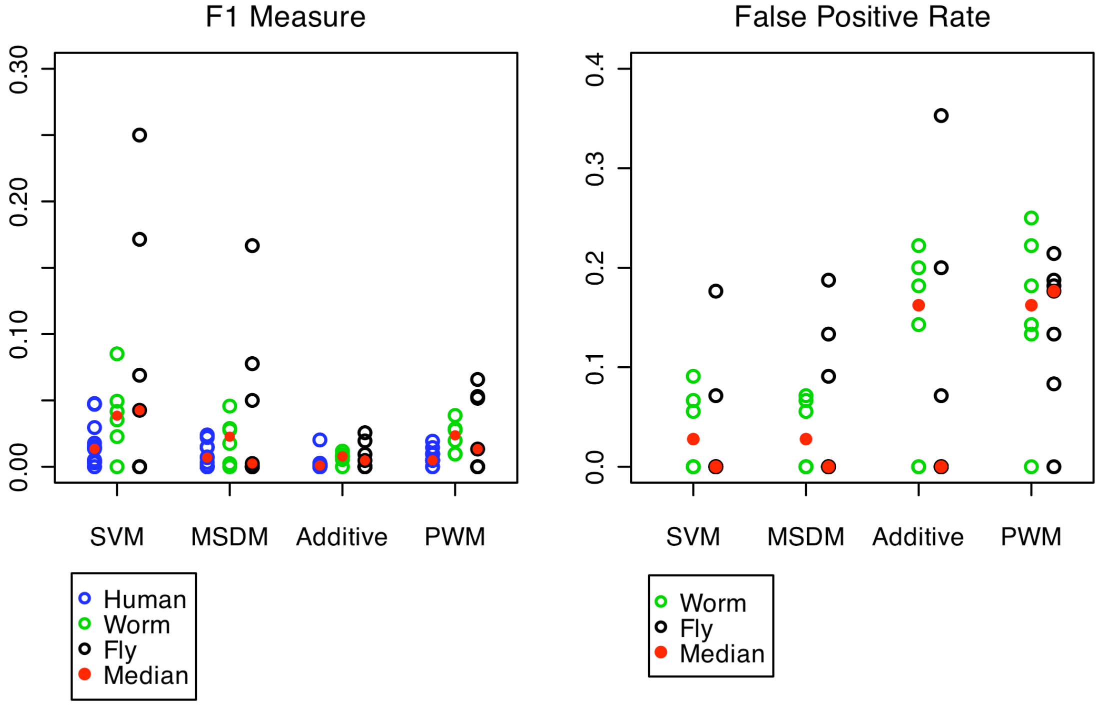


Figure 2-6 Comparison of proteome scanning performances for SVM and other published predictors. A comparison of predictor performance evaluated using F1 measures and FPRs for 13 human (blue), six worm (green) and seven fly (black) PDZ domains. Three different predictors were compared: MDSM, additive model and a PWM predictor. PDZBase interactions were used to validate human predictions. Protein microarray interactions from Chen et al. were used to validate fly and worm predictions. The median is denoted by the red circle. No FPRs were calculated for human predictions since there are no negative human validation interaction data. MDSM and the additive model were trained in their original publications using mouse protein microarray data only. The PWM predictor was trained using the same mouse and human data as the predictor.

The performance of the MDSM and the predictor was close and the predictor’s improved performance may be due to its use of a larger training data set (both phage display and microarray). To more directly compare these two predictors, I trained an SVM with only mouse microarray data and compared the performance. The results show that no predictor method is clearly better than the other. The MDSM's performance is not consistent as shown by the fly test results, which has similar testing and training data sets, and is expected to be an easy test (**Figure 2-7**).

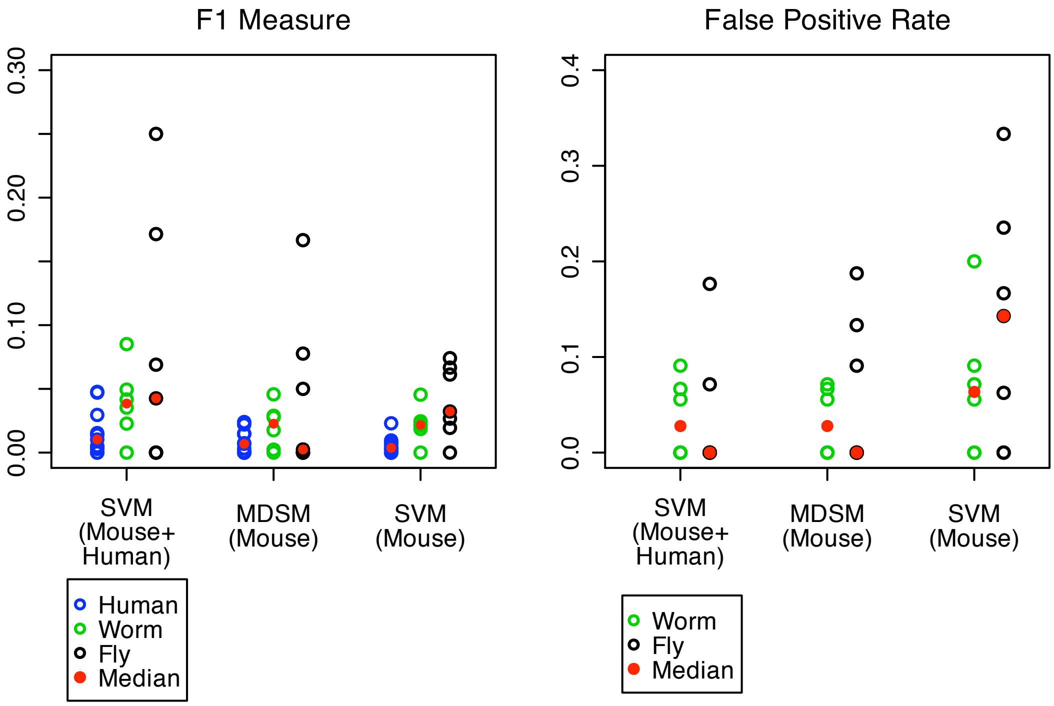


Figure 2-7 Comparison of MDSM and SVM predictor performance. MDSM and SVM performance were evaluated using F1 measures and FPRs for 13 human (blue), six worm (green) and seven fly (black) PDZ domains. The median is denoted by the red circle. No FPRs were calculated for human predictions since there are no negative human validation interaction data. Both predictors were trained using microarray data only.

On the other hand, the performance of the SVM trained only using microarray data is more consistent, but has a higher FPR compared to the MDSM. These results suggest that the predictor performance improvement is likely due to the use of more training data. It may be possible to modify the MDSM method to accept phage display data as training, though the SVM method naturally accepts this data without method modification – a clear advantage in terms of flexibility. Overall, these results demonstrate the predictor's improved performance over other published predictors for proteome scanning of PDZ domain interactions.

2.2.8 Predicted interactions highlight PDZ domain involvement in different biological processes

To demonstrate how the predictions can be used to further our understanding of PDZ domains and the biological processes they mediate, I performed GO biological process term enrichment analysis of the predicted binders in human using the BiNGO (Biological Network Gene Ontology tool) software library (Maere et al. 2005). The hypergeometric test was used to compute a p-value assessing the GO term enrichment for a given set of predicted target genes. Multiple testing correction was performed using the Benjamini and Hochberg False Discovery Rate (FDR) correction. Almost all PDZ domain target lists were statistically enriched ($p < 0.05$) for known PDZ domain processes such as ion transport and localization (see **Appendix A, Table A-6** for detailed results). Interestingly, the biological process 'photoreceptor cell maintenance' was found enriched only among the predicted genes for the PDZ domain containing protein PDZK1-1. These genes include those that encode proteins associated with Usher (USH1G, USH2A) and Bardet-Biedl syndromes (BBS10); both are genetic human diseases of the cilia with wide ranging symptoms including retinal degeneration (Eley et al. 2005). Although disruption of PDZ mediated interactions are known for Usher syndrome, such a disruption involving PDZK1-1 has not been reported for either. Since the validity of the predicted binders is supported by the successful prediction of known interactions in PDZBase and iRefIndex (1 out of 1 and 4 out of 24 respectively), with experimental validation, these potential PDZ domain mediated interactions may provide further insight into the molecular mechanisms underlying Usher and Bardet-Biedl syndromes. There was not enough information to perform the analysis with worm and fly targets.

2.3 Discussion

I have presented a predictor, which can be used to more accurately and precisely scan proteomes of organisms for potential binders of PDZ domains. The results of this predictor can help prioritize biological experiments. In addition, since the predictions are predicted in vitro interactions, they can also be used as input to computational methods aiming to predict likely in vivo interactions by including multiple lines of evidence, such as co-expression and binding site conservation (Jansen et al. 2003; Li et al. 2008). In both cases the predictions will be useful for substantially reducing the number of candidates that need to be considered for more focused analyses. Given the success of the proteome scanning results I also expect the predictor to perform well in organisms which are closely related to human, worm and fly.

An interesting result from my work is that binding site sequence information at contacting positions in the domain was the most effective feature encoding method among the ones I tried.

The poor performance obtained by the other encoding methods (flatly representing binary sequence or physicochemical properties) suggest that by explicitly encoding contacting domain and peptide position pairs, sequence information need only be used to obtain good predictor performance. While I showed that this results in a predictor that relies to some degree on binding site sequence similarity, I also showed that this dependence only exists for the domain and not the peptide. I established a sequence similarity threshold of 60% for testing domains, which may act as a rough indicator of the limits of the predictor and can be used identify poorly characterized PDZ domains in current data sets.

The use of PWMs to generate artificial negatives was motivated by previous work that showed the importance of training with artificial negatives, which resemble real negative interactions. In one study, predictors were trained using random and shuffled negatives to show that this resulted in predictors with lower accuracy when real sequences were used for testing (Lo et al. 2005; Ben-Hur and Noble 2006). In other work, artificial negatives were generated by pairing proteins with different co localizations or randomly pairing proteins known to not interact. It was shown that this created a constraint on the distribution of the negatives making it easier for the predictor to distinguish between positive and negative interactions. This led to biased estimates of predictor performance when cross validation was used (Ben-Hur and Noble 2006). Since the PWM negatives were selected from peptides involved in real positive interactions, they are biological sequences and their distribution is expected to be closer to biologically meaningful interactions. This may result in a more realistic learning problem for the predictor and may reduce the bias in predictor accuracy estimation and benefit predictor performance in practice. However, PWMs may have high false positive rates due to limitations such as their inability to model dependencies between ligand positions. These shortcomings may be responsible for the modest improvement in independent testing performance between predictors trained using PWM generated and other negatives.

Although many of the proteome scanning predictions were validated using known interactions, the lack of a complete benchmark of genomic PDZ domain interactions contributes to the low F1 measures (most are less than 0.1). This may be addressed to some degree by using more validation data from experiments or literature searches, which I expect to help improve the accuracy of the F1 and FPR measurements. In the case of two fly domains LAP4-2 and LAP4-3, the SVM did achieve higher F1 measures of 0.17 and 0.25 respectively. The predictor predicted many known interactions but also predicted a very small number of fly proteins as potential binders (34 and 8 respectively). In general, the predictor made far less positive predictions than the other predictors, which raises the question of whether the predictor is simply more conservative (by making fewer predictions) or actually more precise (by making fewer and more accurate predictions) compared to other predictors. Again, this cannot be fully answered without more validation data, however the predictor's higher F1 and lower FPR scores are strong evidence supporting the latter case.

In genomic tests, predictor performance was consistently poor for the mouse test, which consisted of domains that were highly dissimilar to the training domains. Based on the finding that predictor performance depends on the similarity between testing and training domains, this result was not unexpected. However, even if the similarity between testing and training domains is similar, predictor performance can still be poor. This was discovered while scanning the fly proteome for binders of PATJ-2. I found that the nearest training neighbour for this domain according to binding site sequence similarity did not correspond to its known human homolog, which was present in the training data. This highlighted a limitation generally faced by sequence based predictors: if the training domains best representing a given testing domain do not share similar sequence features, the correct binding specificity may not be properly learned. This may occur for two domains with structurally or physicochemically similar binding sites encoded with very different amino acid sequences. This may be the reason for the predictor's inability to predict any known interactions for PATJ-2. Exploring structural domain features useful for predictor training may determine if this is the case.

While the predictor performs better than published methods on proteome scanning, it can clearly be improved. One way to do this is to consider additional relevant features, such as information related to protein structure. For example, it has been shown that entropic and thermodynamic features of PDZ domain binding can vary considerably across PDZ domains and even for the same PDZ domain bound to different ligands (Fuentes et al. 2004; Basdevant et al. 2006). Therefore, including dynamic features such as electrostatic or non-polar contributions between contacting residues may be used to help improve SVM performance. Another approach would be to use an SVM with a structure based kernel for PDZ domains. Indeed, recent work showed that an SVM using a structure based kernel was successful in the more general problem of predicting protein-protein interactions (Hue et al. 2010). The main challenge for both these approaches is that 3D structures are not available for the majority of PDZ domains and homology modelling would be needed to increase the number of domains available for training and testing. A structure-based approach may also be used to generate more accurate biologically meaningful artificial negatives for training. Thus, until larger training datasets are available, a combination of strategies may be required to predict PDZ domain interactions, involving both sequence and structure-based methods, to maximize coverage and prediction performance. Nonetheless, here I have shown that sequence similarity is an important feature for accurately predicting PDZ domain interactions and it will be interesting to see how general this feature is for other domains.

2.4 Methods

2.4.1 PDZ domain-peptide interaction training data

The predictor was trained using data from mouse protein microarray and human phage display experiments (Stiffler et al. 2007; Tonikian et al. 2008). Interactions were collected in the form of domain-peptide sequence pairs, where domains were represented by their binding site and peptides were five residues in length. For both mouse and human PDZ domains, those whose binding site did not align well with other PDZ domains were omitted. Human domains that lacked adequate data (less than ten interactions), or were difficult to generate artificial negative interactions for, were also not used. This left 82 out of 85 mouse and 31 out of 54 human PDZ domains. Since phage display data may contain non-genomic interactions, I filtered the human phage display data to create a data set enriched in genomic-like interactions. First, an interaction was considered to be genomic-like if the last four residues of the interacting peptide matched a human protein tail (defined by genome assembly Ensembl:GRCh37.56), otherwise it was defined as non genomic-like. Then, domains were categorized as genomic-like, non genomic-like, dual or non-specific, depending on the number of unique genomic-like or non genomic-like interacting peptides they bound to (**Table 2-2**).

Category	**# Unique genomic-like peptides**	**# Unique non genomic-like peptides**
Genomic-like	≥ 10	< 10
Non genomic-like	< 10	≥ 10
Dual	≥ 10	≥ 10
Non specific	< 10	< 10

Table 2-2 Domain category definitions based on the number of unique genomic-like and non genomic-like peptides.

To enrich for genomic-like interactions I did not use any data from non genomic-like domains and removed all non genomic-like interactions from the dual domains. Domains with less than ten unique genomic-like peptides after this filtering were not used. Finally, data from genomic-like and non specific domains (that had a combined total of ten or more peptides) were used without any filtering. This resulted in a small number of non genomic-like interactions being included, but allowed us to increase the amount of phage display data usable for training. In total, data for 20 human and 82 mouse domains were used for training (**Table 2-3**).

		Domains		**Interactions**	
Organism	**Source**	**# Pos**	**# Neg**	**# Pos**	**# Neg**
Mouse	Protein microarray	82	72	643	1324
Human	Phage display	20	-	363	-
Human	Artificial negatives	-	20	-	745
	Total	102	92	1006	2069

Table 2-3 Summary of domain-peptide interaction data used for training.

2.4.2 Artificial negative interactions for phage display

In order to train an effective binary predictor both positive and negative interaction data were required. Therefore, I generated artificial negative interactions for the human phage display data since they only contained positive interactions. Based on previous research the proper selection of artificial negatives is important for successful predictor training and evaluation (Lo et al. 2005; Ben-Hur and Noble 2006). Random and shuffled peptide sequences have been commonly used, but since these negatives do not resemble real sequences, they have been shown to produce predictors with lower accuracy when predicting real negative interactions (Lo et al. 2005). I generated artificial negative interactions for training based on positive interactors (peptide ligands) modelled using PWMs. Therefore a PWM for a given PDZ domain was used to select likely negative interactors for that domain from a set of unique real interactors for all domains. Specifically, all unique peptides from the positive training interactions were put into a list to create a pool of peptides. Given a domain with a corresponding set of positive peptide sequences (representing positive interactions determined from phage display), the following steps were taken to select artificial negatives:

Step 1: A PWM was built using the positive peptide sequences and the minimum PWM score amongst the positive peptides was set to be the cutoff.

Step 2: All unique peptides in the pool were scored with the PWM from Step 1 and sorted in descending order according to PWM score. Walking down the sorted list, peptides were selected based on two criteria:

1. Low scoring: the PWM score must be lower than the cutoff
2. Low redundancy: The similarity of the peptide to peptides already selected must be below the redundancy threshold (must have less than three residues in common with negative peptides already selected).

For the 20 human phage display domains, a total of 745 artificial negative interactions were generated.

When selecting negative peptides in Step 2, only peptides with less than three residues in common with those already selected were used. I optimized this redundancy threshold by building different SVMs trained using artificial negatives selected using different redundancy thresholds (i.e. 1,2,3,4,5). For example, using a low threshold (less than one residue in common) would allow fewer but a more diverse set of negatives to be selected than using a higher threshold (less than five in common) which would allow a greater number but an overall less diverse set of negatives to be selected. The predictor with the highest ROC and PR AUCs was used and corresponded to a redundancy threshold of three (**Figure 2-8**).

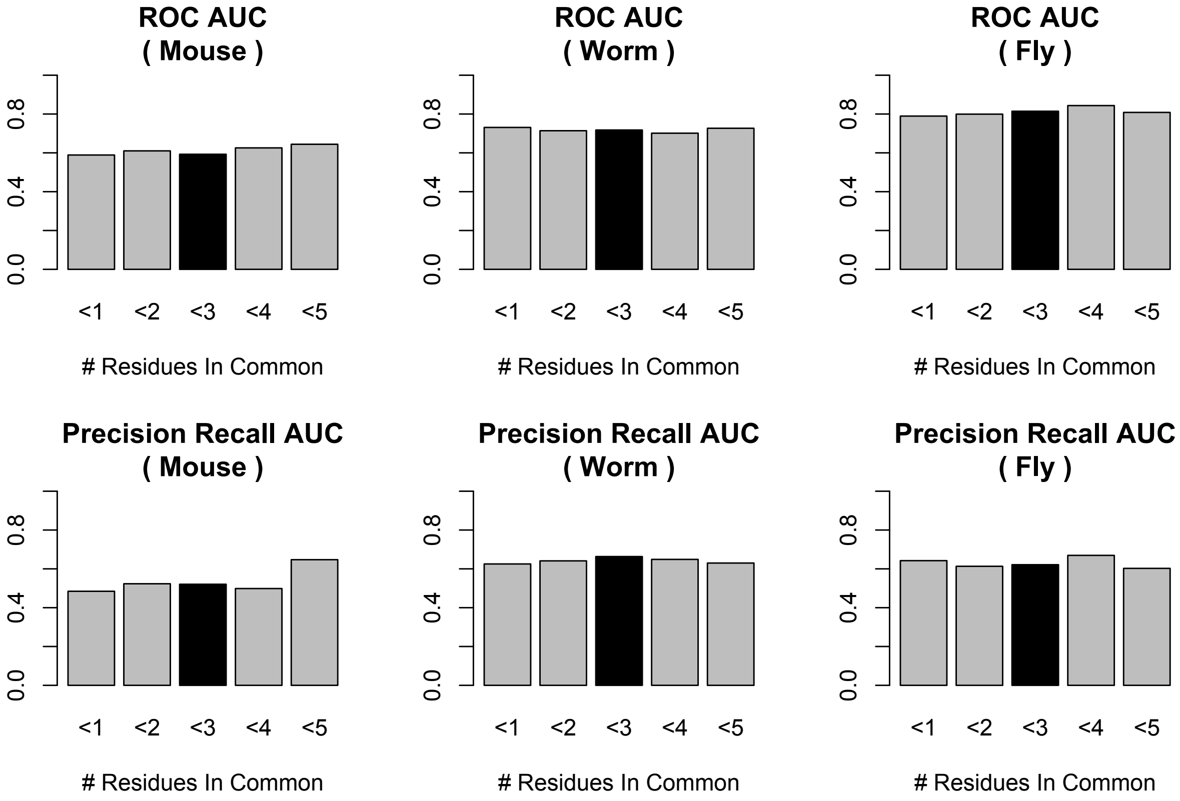


Figure 2-8 Selecting peptide data redundancy threshold. (Top Row) ROC AUC comparison for predictors trained using data with different levels of peptide redundancy. (Bottom Row) PR AUC comparison for predictors trained using data with different levels of peptide redundancy. Black coloured bars indicate the number used for the final predictor.

2.4.3 Primary sequence based feature encoding

Domain-peptide interactions were encoded as a vectors of numeric values representing features of a positive or negative interactions. Interacting domain-peptide residue pairs were encoded using the 'contact map' encoding method (Chen et al. 2008). Contact maps for other domains were constructed via a multiple sequence alignment.

2.4.4 Optimization of SVM parameters

The RBF kernel parameter γ and the SVM cost parameter C were optimized by performing a coarse two dimensional grid search over combinations of C ={2,4,6,8,10} and γ = {2,4,6,8,10}, with a finer grid search over combinations of C = {2,3,4,5,6} and γ = {3,4,5}. A ten fold cross validation using the training data was performed to evaluate the average ROC AUC score for each combination of γ and C. The parameters values yielding the predictor with the highest ROC AUC score were used. LibSVM was used to build the SVMs (Chang and Lin 2011).

Chapter 3

Predicting PDZ Mediated Protein-Protein Interactions From Structure

This work was published in *BMC Bioinformatics*, **14**:12: Hui, S., Xing, X., Bader, GD. (2013), Predicting PDZ mediated protein-protein interactions from structure.

Author contributions: I collected the data, developed and implemented the methods and performed the analyses. Xiang Xing developed the POW! Website under my supervision. Gary D. Bader supervised and advised this project.

3 Predicting PDZ mediated protein-protein interactions from structure

3.1 Introduction

In the previous chapter, I presented a sequence-based predictor to scan proteomes of multiple organisms for binders of PDZ domains. Although this predictor is more accurate and precise at proteome scanning compared to previous sequence-based predictors, like others, it performs better on sequences similar to those in the training set. It is known that structure features within the domain binding pocket play important roles in determining binding specificity (Skelton et al. 2003; Appleton et al. 2006; Chen et al. 2007). Since domain structure features capture different information about binding compared to sequence features, I hypothesized that training with such features would result in a predictor that is complementary to the sequence-based predictor. In particular, such a predictor would be less dependent on sequence similarity and would predict additional interactions not predicted by the sequence-based predictor. This would expand the coverage of PDZ domain C-terminal peptide interactions that can currently be predicted by sequence-based predictors alone.

In this chapter, I present a structure-based predictor for PDZ domain-peptide interactions that can be used for proteome scanning. This predictor uses a variety of structure features that are known to play roles in protein structure stability and facilitating PPIs. Through leave-12%-of-domains-out cross validation, I show that the structure-based predictor depends less on training-testing domain sequence similarity compared to the previous sequence-based predictor. Based on human proteome scanning results, I also show that the structure-based predictions correspond to known experimentally determined PDZ domain-peptide interactions and known PPIs involving PDZ domain containing proteins. A substantial number of the structure-based predictions correspond to known PPIs not previously predicted by the sequence-based predictor (48% increase), confirming that the structure-based predictor finds different interactions than the sequence-based predictor. Using predictions from both methods, I created a functional map of all predicted human PDZ mediated PPIs and identify xenobiotic metabolism as a novel biological process enriched in PDZ interactors.

Finally, a website was created called POW! PDZ domain-peptide interaction prediction (http://webservice.baderlab.org/domains/POW), which enables users to run the sequence-based and structure-based predictors online for human, mouse, fly and worm.

3.2 Results

3.2.1 The structure-based predictor achieves high cross validation results

To estimate the generality of the predictor, I ran multiple cross validation tests and plotted the ROC and PR curves to summarize the performance. The predictor achieves high ROC and PR AUC scores compared to random performance AUCs over all cross validation strategies. In particular the ten fold cross validation ROC and PR AUCs were 0.96 and 0.936, respectively (random ROC AUC 0.5, PR AUC 0.253). The leave-8%-of-peptides-out cross validation ROC and PR AUCs were 0.935 and 0.909 respectively (random ROC AUC 0.5, PR AUC 0.358). The leave-12%-of-domains-and-8%-of-peptides-out cross validation out ROC and PR AUCs were 0.927 and 0.886 respectively (random ROC AUC 0.5, PR AUC 0.347). Finally, slightly lower AUCs were obtained for the leave-12%-of-domains-out cross validations, which achieved 0.872 and 0.785 respectively (random ROC AUC 0.5, PR AUC 0.33) (**Figure 3-1**).

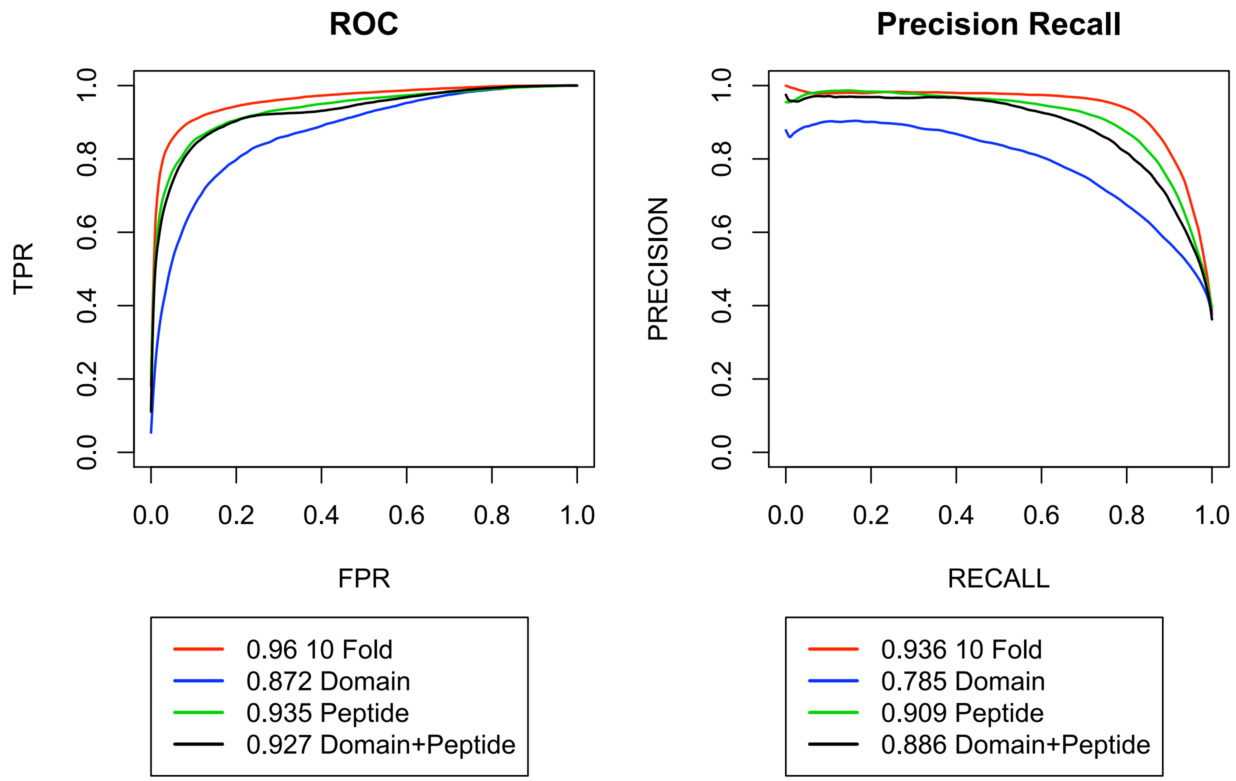


Figure 3-1 Predictor performance estimation using cross validation. Predictor performance measured using ten fold (red), leave-12%-of-domains-out (blue), leave-8%-of-peptides-out (green), leave-12%-of-domains-and-8%-of-peptides-out (black) cross validation.

Like the previous sequence-based predictor, the cross validation results were lower for strategies that involved leaving sets of domains out. A one-tailed t-test showed that the mean AUC scores were significantly higher for the structure-based predictor compared to those of the sequence-based predictor (p-value < 0.025) (**Table 3-1**).

	ROC		**PR**	
	Structure	**Sequence**	**Structure**	**Sequence**
10 Fold (95% CI)	**0.96** **(0.957~ 0.962)**	0.939 (0.936~0.941)	**0.936** **(0.932~0.940)**	0.896 (0.890~0.900)
Domain (95% CI)	**0.872** **(0.860 ~0.882)**	0.851 (0.839~0.862	**0.785** **(0.765~0.805)**	0.764 (0.747~0.779)
Peptide (95% CI)	**0.935** **(0.929~ 0.941)**	0.893 (0.883~0.902)	**0.909** **(0.898~0.918)**	0.838 (0.825~0.850)
Domain+Peptide	**0.927**	0.87	**0.886**	0.794

(95% CI)	**(0.919~ 0.934)**	(0.862~0.877)	**(0.875~0.896)**	(0.783~0.804)

Table 3-1 Comparison of structure-based and sequence-based predictor cross validation results. Structure-based predictor achieves significantly better cross validation results than the sequence-based predictor.

3.2.2 The structure-based predictor successfully predicts a limited number of interactions in different organisms

Blind testing was performed to obtain an unbiased measure of predictor performance and to determine if the predictor could correctly predict interactions in other organisms not represented in the training set (such as fly and worm). I used interaction data for 13 mouse, seven worm and six fly PDZ domains with interactions from previous protein microarray experiments which were not previously used for training (Chen et al. 2008) (**Table 3-2**).

		Domain		**Interactions**	
Organism	**Source**	**# Pos**	**# Neg**	**# Pos**	**# Neg**
Mouse	Protein microarray	8	13	32	36
Worm	Protein microarray	6	7	59	88
Fly	Protein microarray	6	6	34	106

Table 3-2 Summary of domain-peptide interaction data used for blind testing.

Homology models were generated by SWISS-MODEL and have at least 40% sequence identity to their template structures and no binding site gaps. The average template sequence similarity was 92%, 61% and 61% for mouse, worm and fly domains, respectively. An NMR structure was available for one fly domain (PAR6-1) and the first model was used (1RY4 A). One mouse domain (CHAPSYN-110-1) was removed from the test set because its performance was consistently poor for both sequence-based and structure-based predictors (see **Appendix B, Table B-2** for details on blind testing domains).

The blind test results show that the structure-based predictor is able to correctly predict many unseen interactions in fly, worm and mouse (**Figure 3-2**).

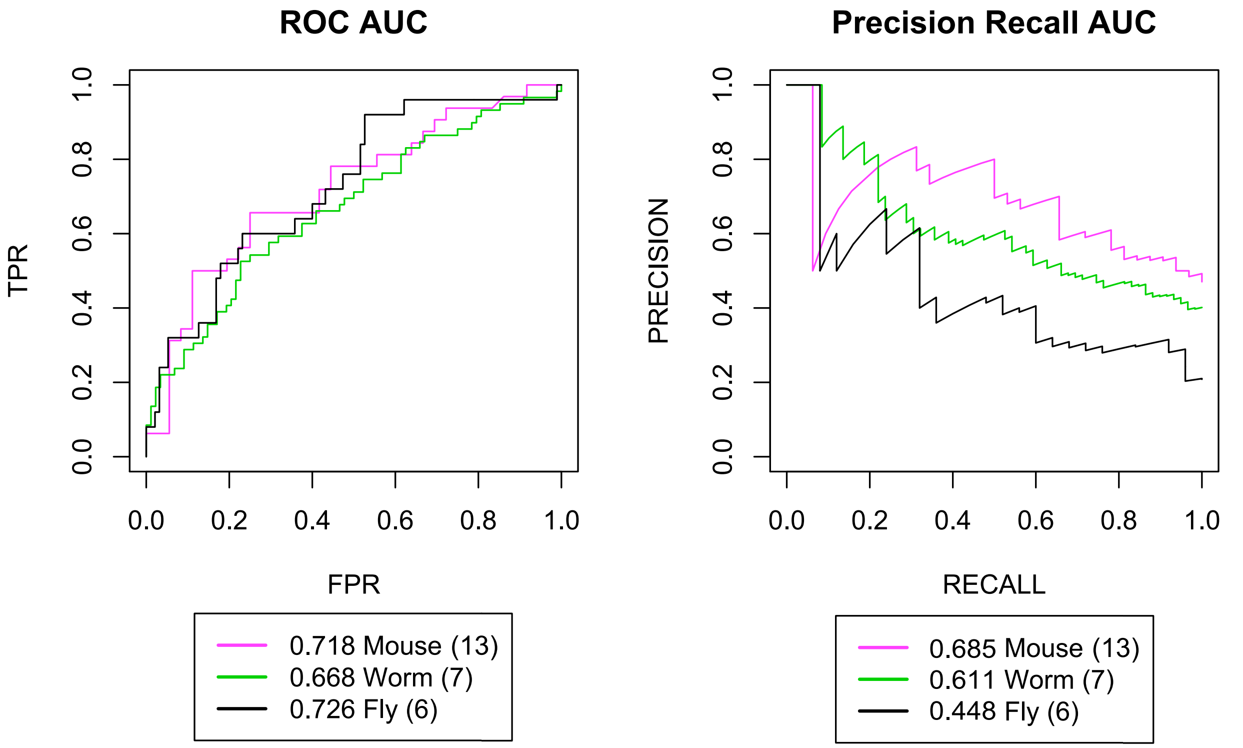


Figure 3-2 Blind testing performance results for mouse, worm and fly. ROC and Precision/Recall curves were computed for mouse (magenta), worm (green) and fly (black) blind tests. Test data was obtained from published protein microarray experiments (Chen et al. 2008). Number of PDZ domains tested is noted in parentheses.

However, compared to the sequence-based predictor, the structure-based predictor performance is similar for mouse (**Table 3-3**), but somewhat worse for worm and fly blind tests. Since these data sets are small, additional data is required to accurately compare predictor performance.

	ROC AUC		**PR AUC**	
	Structure	**Sequence**	**Structure**	**Sequence**
Mouse	0.718	0.709	0.685	0.723
Worm	0.668	0.718	0.611	0.663
Fly	0.726	0.799	0.448	0.591

Table 3-3 Comparison of structure-based and sequence-based predictor blind testing performance.

3.2.3 The structure-based predictor is less dependent on training-testing domain sequence similarity

The performance of the previous sequence-based predictor depends on how similar in binding site sequence a given testing domain is to its nearest training domain. In particular, as the domain binding site sequence similarity decreases so does the predictor's average performance until it is comparable to that of a naïve nearest neighbour sequence predictor. To more rigorously compare structure-based and sequence-based predictor performance as training-testing domain sequence similarity varies, I performed a leave-12%-of-domains-out cross validation with domain sequence similarity-based training set filtering for each predictor. For each fold, 12% of domains and their interactions were held out, and of the remaining domains, only those and their corresponding interactions were retained for training if the domain sequence similarity was less than a given threshold for all testing domains. All training sets had no more than 500 interactions. Ten folds were executed and repeated ten times for a total of 100 runs. For each run, the ROC and PR AUCs were computed and plotted as box plots according to the similarity threshold (**Figure 3-3**).

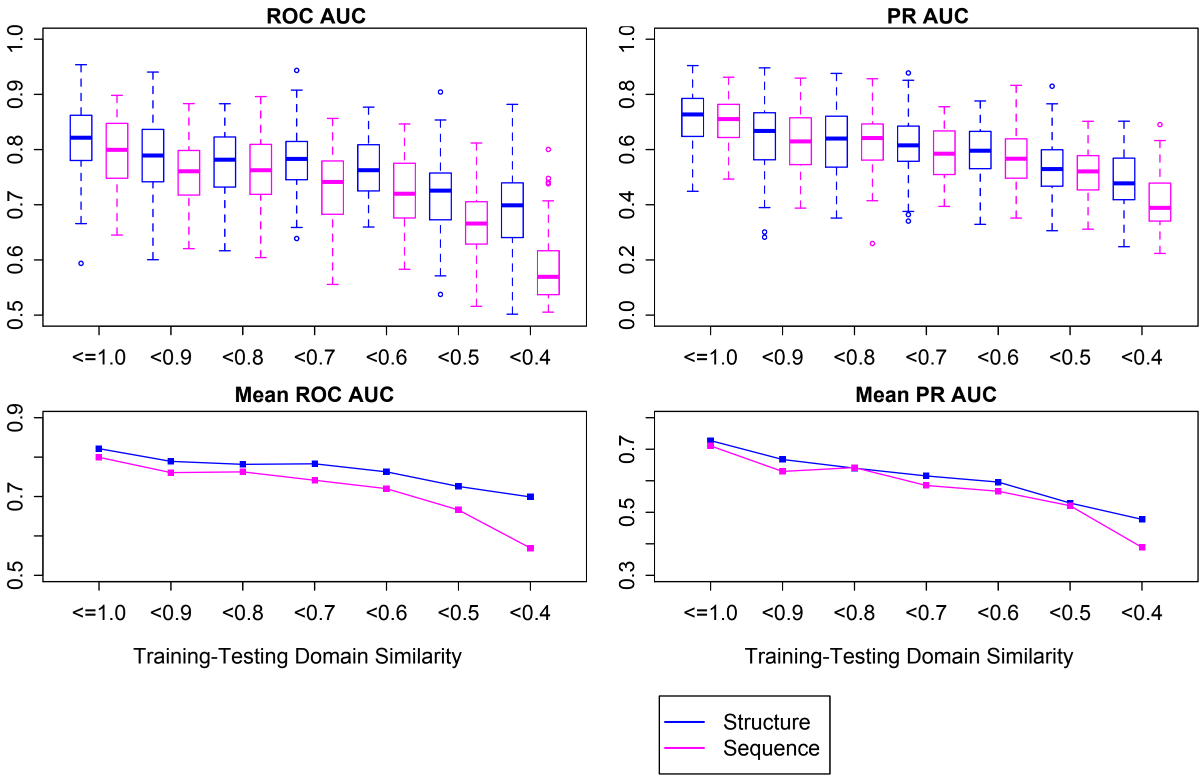


Figure 3-3 Predictor performance dependence on training-testing domain sequence similarity. Leave-12%-of-domains-out cross validation was performed with domains retained for training in each fold if their sequence similarity to all testing domains was less than a given threshold. This was performed for structure-based (blue) and sequence-based predictors (magenta). ROC and PR AUC scores were computed for each run and displayed in box plots according to training-testing domain sequence similarity threshold (top left and right). Based on significance testing using a one-tailed t-test, the mean structure-based predictor ROC and PR AUC scores are significantly higher than the sequence-based predictors scores when training-testing domain sequence similarity is < 0.7 (*p*-value < 0.029). The mean AUC scores for structure-based (blue) and sequence-based (magenta) predictors are plotted against sequence similarity threshold (bottom left and right).

A one-tailed t-test showed that the mean ROC and PR AUC scores were significantly higher for the structure-based predictor when training-testing domain sequence similarity is < 0.7 (*p*-value < 0.029). These results show that on average, the structure-based predictor is less dependent on training-testing domain sequence similarity compared to the sequence-based predictor at lower similarity thresholds.

3.2.4 Structure-based predictions are validated by known PDZ domain-peptide interactions

The predictor was used to scan the human C-terminal proteome (defined by genome assembly Ensembl:GRCh37.64) for binders of 45 PDZ domains with known interactions in PDZBase that I could obtain structures and compute features for. For each domain, this involved scanning 43827 unique C-termini of length five (including splice variants). Structures for these domains were obtained from the PDB or were homology modelled and are at least 35% sequence similar (average over 80%) to their template structures. The minimum QMEAN score for these models is 0.36 (average 0.78). Please see **Appendix B, Table B-3** for details about domains used for scanning.

The structure-based predictor has a true positive rate (TPR) of 0.36 and precision of 0.0033 and correctly predicted interactions for 22 of the 45 domains. For these domains approximately 73% of known PDZ domain-peptide interactions in PDZBase, an independent data source not used for training, were predicted (see **Appendix B, Table B-4** for detailed results). The sequence-based predictor had a higher TPR of 0.46 and correctly predicted interactions for 28 out of 45 domains. For these domains, 65% of known PDZ interactions were predicted and the precision was 0.0024. Although the sequence-based predictor has a higher TPR than the structure-based predictor, its precision and coverage of known PDZ domains is lower. This is likely because the sequence-based predictor predicts on average more interactions per domains than the structure-based predictor (average 426.89 and 239.71 per domain respectively). The low precision for both predictors is due to the few known interactions per domain that are available from PDZBase (average 2.2 interactions per domain).

I also tested the false positive rate of the predictor using two real negative data sets for human, which were used in a recent study (Luck et al. 2011) to benchmark another recent sequence-based predictor (Chen et al. 2008). The first data set consists of 466 experimentally validated negative interactions involving peptides that contain a PDZ binding motif found from the literature. The second data set consists of 133 negative literature-described interactions involving peptides with a non-binding PDZ motif caused by a mutation. The structure-based predictor made predictions for 410 negative interactions from the first data set and 126 negative interactions from the second data set, which resulted in an FPR of 0.145 and 0.0, respectively. The sequence-based predictor had a FPR of 0.09 and 0.0, and made predictions for 421 and 128 negative interactions for the first and second data sets, respectively. Compared to the structure-based and sequence-based predictors, the Chen et al. sequence-based predictor has a much higher FPR of 0.482 and 0.256 for the first and second data sets, respectively (see **Appendix B, Table B-8, Table B-9** for detailed results).

3.2.5 Many structure-based predictions correspond to known PDZ domain containing protein-protein interactions

To determine how many structure-based predicted interactions correspond to known PPIs, I scanned the human proteome to predict interactions for 218 human PDZ domains with known PPIs (that I could obtain structures and compute structure features for). Known PPIs were retrieved from iRefIndex (Razick et al. 2008). In total, 61 XRAY and nine NMR structures (only the first models used) were obtained from the PDB and 148 homology models were created. All models had a template sequence similarity of at least 22% (average 72%) and QMEAN score of at least 0.36 (average 0.78) Please see **Appendix B, Table B-3** for details about domains used for scanning.

In total, 88 domains had predicted interactions that corresponded to known PPIs, with an average of greater than 21% of known PPIs being correctly predicted per domain. The number of PPIs successfully predicted per domain was significant (p-value < 0.05, Fisher's exact test) for all but ten domains. A caveat of this result is that PDZ domain containing proteins may contain multiple PDZ domains and other domains, so it is not possible to uniquely assign a PPI to a PDZ domain. This could result in erroneous false negative or true positive statistics for the above tests. However, the results still serve as an estimate of predictor performance and show that the predictor is able to predict many known human PPIs.

3.2.6 The structure-based predictor is complementary to the sequence-based predictor

I also compared the structure-based predictor's proteome scanning predictions to the ones obtained using the sequence-based predictor. In total, the results for 221 domains where both predictors were able to make predictions were compared. A total of 172 out of 925 known PPIs were predicted using both methods, 116 were unique to the sequence predictor and 56 were unique to the structure-based predictor (**Figure 3-4**). Thus the sequence and structure-based predictors both predict unique known PPIs and are complementary.

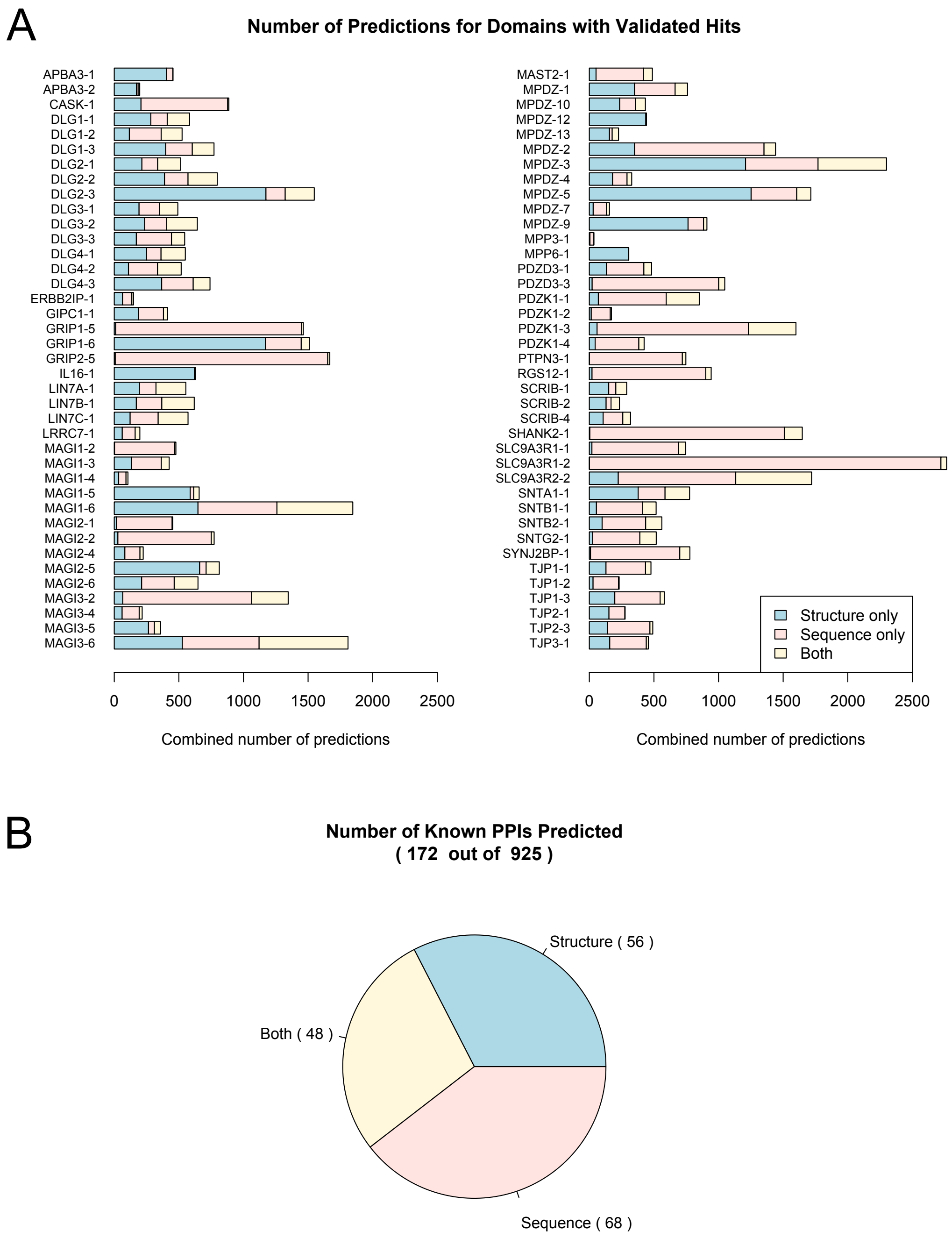


Figure 3-4 Summary of predictions for domains with hits validated by known PPIs. (A) Breakdown of the number of proteome scanning predictions per domain made by the structure-based predictor only (blue), sequence-based predictor only (pink), and both predictors (yellow). Only domains with hits matching known PPIs (physical and experimental interactions) in

iRefIndex are shown. (B) Pie chart of the number of validated hits predicted by the structure-based predictor only (blue), sequence-based predictor only (pink), both predictors (yellow).

To better understand how unique predictions are made, I compared the results in more detail. The unique structure based predictions arise for different reasons. Some domains (43 domains) are more challenging for the sequence-based predictor, which returns a low number of hits per domain (ten or less) with none corresponding to known PPIs (e.g. APBA1-1, CNKSR2-1, IL16-1, IL16-3) (**Table 3-4**).

Domain Name	#P	#TP	#Pred. Struct.	#Pred. Seq.	#Pred. Both	#TP Struct.	#TP Seq.	#TP Both
APBA1-1	7	0	9	1	0	0	0	0
ARHGEF11-1	11	1	273	0	0	1	0	0
CNKSR2-1	8	0	16	8	0	0	0	0
DLG1-1	23	11	283	127	173	2	0	9
DLG1-2	23	12	117	246	162	3	1	8
DLG5-3	2	0	2	239	0	0	0	0
IL16-1	5	1	621	0	6	1	0	0
IL16-3	5	0	80	0	5	0	0	0
MLLT4-1	19	0	8	47	0	0	0	0
MPDZ-6	13	0	1	339	0	0	0	0
MPDZ-8	13	0	4	75	1	0	0	0
MPDZ-12	13	4	437	3	2	4	0	0
MPP3-1	7	1	5	30	0	0	1	0
MPP6-1	16	1	302	3	0	1	0	0
PDZD2-3	1	0	671	0	0	0	0	0
PDZD2-5	1	0	316	0	2	0	0	0
RAPGEF6-1	4	0	1529	0	5	0	0	0
SCRIB-3	14	3	344	0	0	3	0	0

Table 3-4 Subset of validation results for human PDZ domain proteome scanning predictions against known interactions in iRefIndex. Details for all domains scanned are found in **Appendix B, Table B-9**.

The structure predictor fares better for nine of these domains (ARHGEF11-1, IL16-1, IL16-3, MPDZ-12, MPP6-1, PDZD2-3, PDZD2-5, RAPGEF6-1, SCRIB-3) and is able to predict many more hits per domain (on average approximately 510 hits) with on average approximately three known hits per domain. On the other hand, the structure-based predictor has difficulty predicting hits for 19 domains (e.g. DLG5-3, MPDZ-6, MPDZ-8), of which four are better predicted by the sequence-based predictor (MLLT4-1, MPDZ-8, MPP3-1, PDZD2-2; average 383 hits) with on average one known PPI hit per domain. In another scenario, two domains may have identical binding sites at the sequence level (e.g. DLG1-1 and DLG2-1), but be different at the structure level. The sequence-based predictor cannot distinguish between the two domains in this case, even though the domains may actually bind different proteins. While the structure-based predictor uses features corresponding to ten core positions, these features are computed by considering the entire domain structure. Therefore, even if two domains have the same binding site residues, the resulting features will be different if their whole domain structures are different. The structure-based predictor’s ability to distinguish between domains with highly similar binding site sequences helps explain why it is able to predict more unique interactions than the sequence-based predictor. Overall, these results demonstrate situations where the structure-based predictor can be used to make predictions for domains that otherwise could not be easily predicted by the sequence-based predictor and thus shows that both methods are complementary.

3.2.7 Structure-based predicted binding specificities recapitulate experimental binding specificities

Since validation data is limited, I more generally assessed the results of proteome scanning by comparing predicted binding specificities to those known from phage display. I constructed position weight matrices to summarize the domain’s amino acid binding preference at each position in the ligand, using all predicted interacting peptides from C-terminal proteome scanning. Sequence logos were then used to visually represent the binding specificities. In total, 26 domains could be compared (had > 4 genomic peptides from phage display experiments), covering known PDZ domain binding classes I and II (see **Appendix B, Figure B-1** for all logos).

For several domains, the structure-based predicted binding specificity is more similar to the phage display determined binding specificity than the sequence-based predicted binding specificity, and better recapitulates the preference of residues at specific positions (**Figure 3-5**).

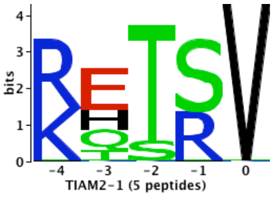
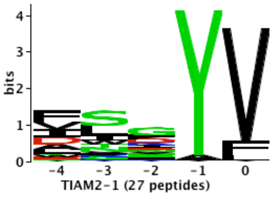
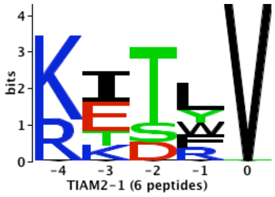
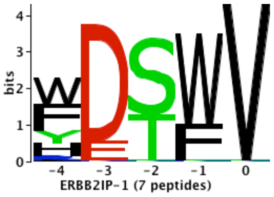
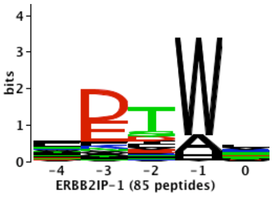
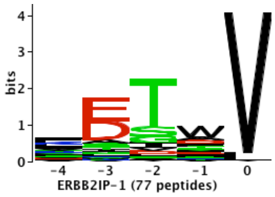
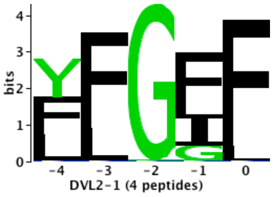
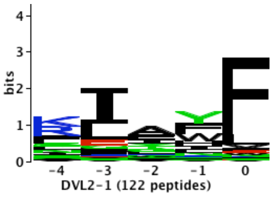
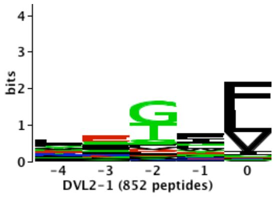
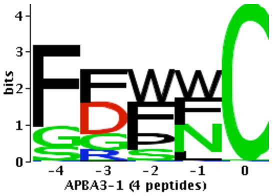
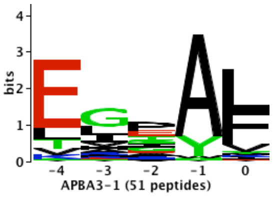
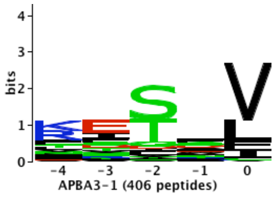
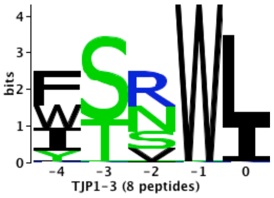
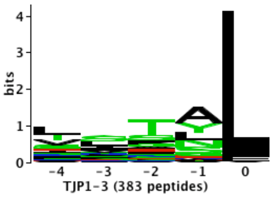
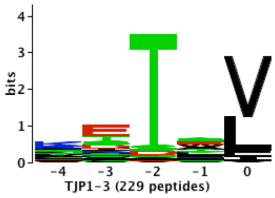
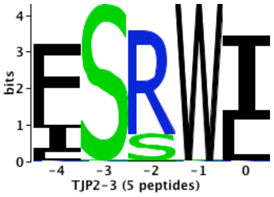
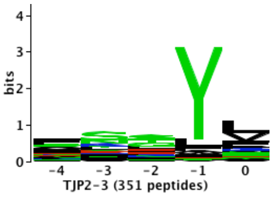
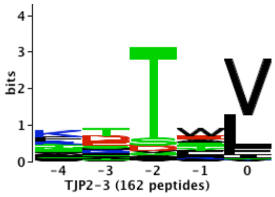
Phage Display	Predicted Logo (Sequence)	Predicted Logo (Structure)	Sim (Sequence)	Sim (Structure)
			0.47	**0.736**
			0.619	**0.692**
			0.561	**0.584**
			0.4	**0.5**
			**0.537**	0.454
			**0.421**	0.378

Figure 3-5 Comparison of a subset of predicted and phage display determined PDZ domain binding specificities. Phage display determined and predicted PDZ domain binding specificities for the last five terminal binding positions were visualized as sequence logos. The binding specificity similarity between two domains was computed using the normalized Euclidean distance between their corresponding position weight matrices (**Eq. 2.3**). Non-genomic phage display peptides were removed from the set of binders for each domain. Only domains with four

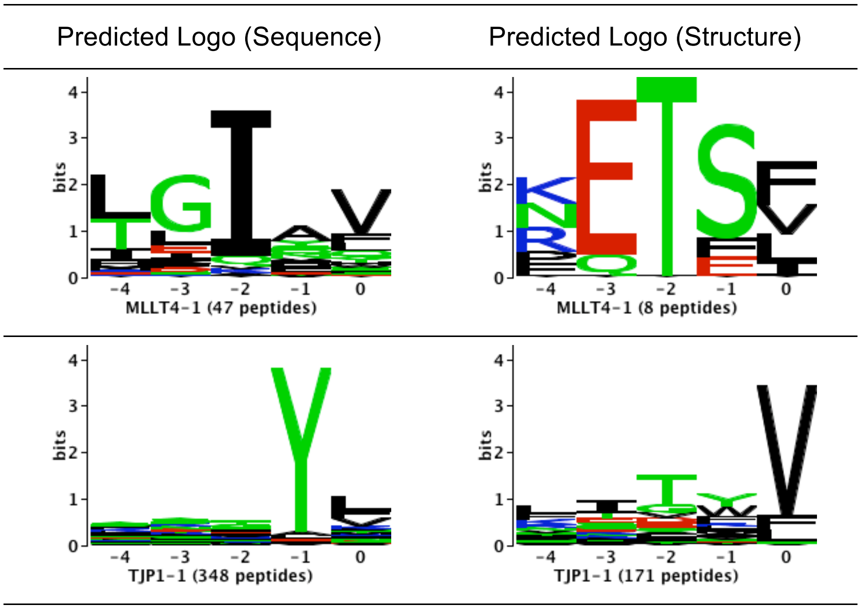
or more peptides after this filter were used to create sequence logos describing the domain's binding specificity. Based on a previously established protocol, a peptide was considered to be genomic if the last four residues could be found in a proteomic tail, otherwise it was considered to be non genomic. Numbers in bold indicate which similarity (sequence or structure) is higher (i.e. which predicted logo is closer to the experimental logo).

For example, the structure-based method better predicts the preference for polar residues at position -4 and a Thr or Ser at position -1 for TIAM2-1, for hydrophobic Val residue at position 0 for ERBB2IP-1 and for hydrophilic residues such as Gly or Thr at position -2 for DVL2-1 (position numbering counted backwards from the zero C-terminal position) (**Figure 3-5** Rows 1-3).

Three domains, APBA3-1, TJP1-3 and TJP2-3 had both structure-based and sequence-based predicted binding specificity similarities much lower than the average (0.5 or less) (**Figure 3-5** Rows 4-5). This seems to be caused by poor representation of these domains in the training set. More validation data should be used to more reliably compare the binding specificities for these domains in the future. Furthermore, since phage display experiments select optimal binders and cellular interactions may not be optimal (e.g. to aid interaction regulation), some differences between phage display and proteome scanning-based profiles were expected. In general, the similarity between the structure-based predicted and experimentally determined binding specificities is high (0.636).

3.2.8 Predicted binding specificities are supported by known structural determinants of PDZ domain binding specificity

As noted above, there are many cases where the structure-based predicted binding specificity is closer to the experimental binding specificity than the sequence-based predicted binding specificity. For some examples, the structure-based predicted binding specificity better predicts the experimental binding specificity at certain positions. To examine if this is caused by specific structural features used by the structure-based predictor, I searched the literature to find known structure determinants influencing these specific amino acid preferences and compared them to the results. For MLLT4-1, the structure-based predictions indicate a preference for a hydrophilic Thr residue at position -2 (**Figure 3-6** Row 1). The preference for a hydrophilic Thr residue at position -2 is explained by the findings of Chen et al. (Chen et al. 2007). Their work showed that the Thr preference at position -2 is due to its interaction with Gln at position α2-1 of the domain, which forms a hydrophilic binding site pocket at position -2. This preference is reflected in the structure-based predicted binding specificity, whereas a completely different preference for a hydrophobic Ile residue at this position is predicted by the sequence-based predictor. The domain TJP1-1 is another example where the predicted structure and sequence-based binding specificities are very different (**Figure 3-6** Row 2). Appleton et al., showed that this domain has a bi-specific preference for Trp or Tyr at position -1 (Appleton et al. 2006). The Trp preference is accommodated through main chain interactions with β2 and β3 strands, while the Tyr preference is accomplished through hydrogen bonding with Asp at position β3-5 of the domain. The bi-specific preference for a Trp or Tyr at position -1 is reflected in the structure-based binding specificity, while only a preference for Tyr is indicated in the sequence-based binding specificity. Finally, the predicted binding specificities for domain DVL2-1 are very different (**Figure 3-6** Row 3). Zhang et al. found that the -2 binding site of the domain actually accommodates a Gly-Tyr pair (Zhang et al. 2006). The preference for a Gly at position -2 is reflected in the predicted structure-based binding specificity whereas there is no obvious preference in the predicted sequence-based binding specificity. Since the binding specificities for these examples are determined by specific domain structure features, this helps explain why the structure-based predictor can better predict their binding preferences than the sequence-based predictor.



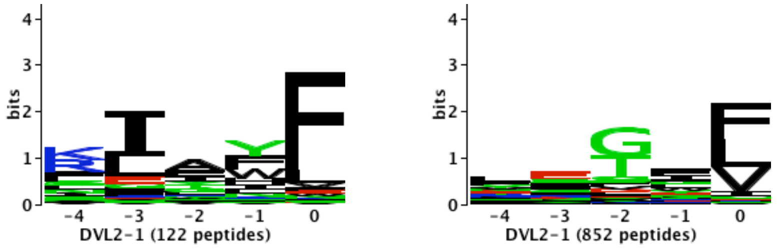


Figure 3-6 Comparison of a subset of sequence-based and structure-based predicted PDZ domain binding specificities.

3.2.9 A functional map of PDZ domain biology highlights PDZ involvement in a variety of biological processes

To identify gene functions better predicted by sequence or structure-based methods, I performed GO-based gene function enrichment analysis on all predicted PDZ targets. The results were visualized using an enrichment map, which groups related gene function terms to ease identification of functional themes (**Figure 3-7**).

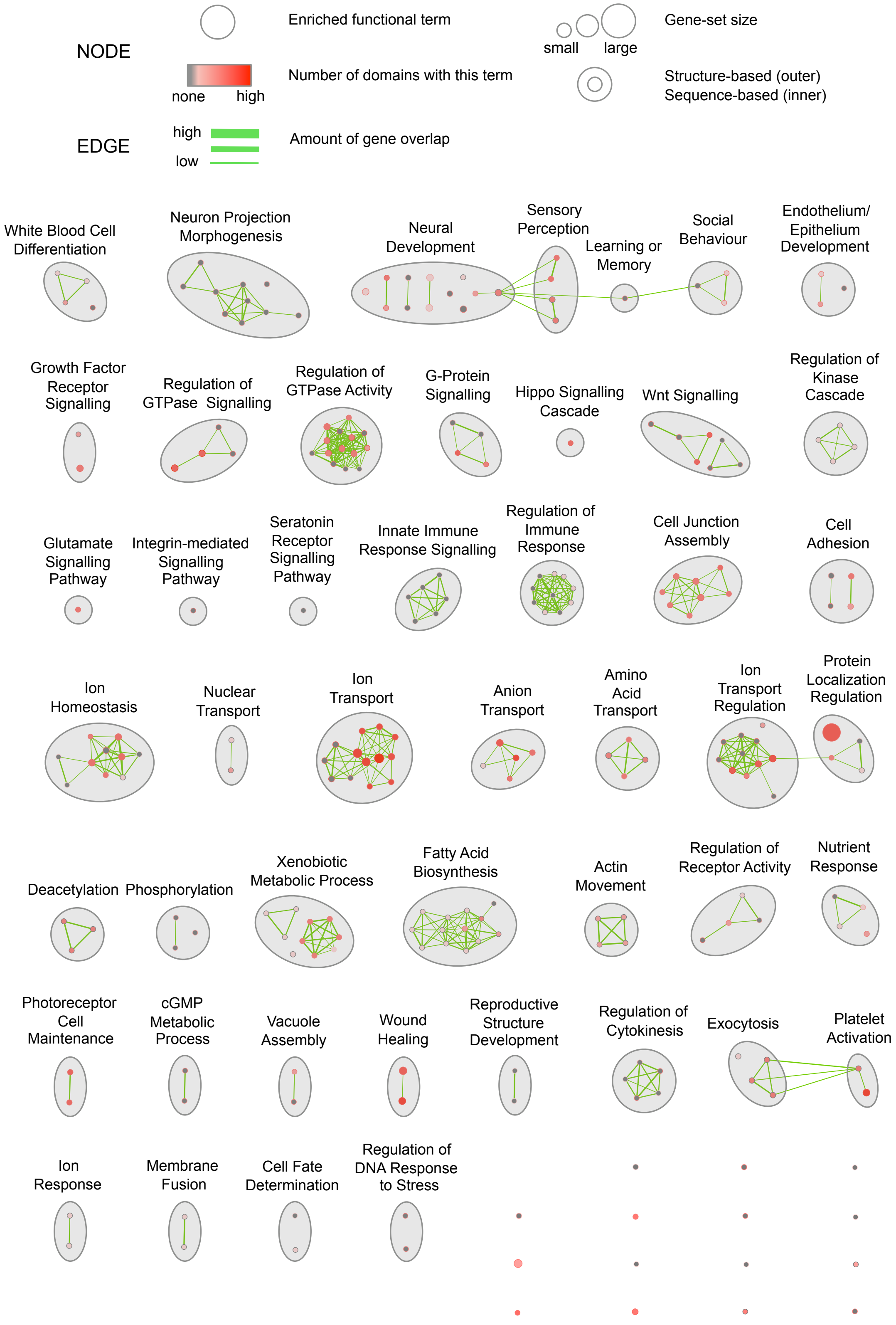

NODE
Enriched functional term
Gene-set size
small
large
Number of domains with this term
none
high
Structure-based (outer)
Sequence-based (inner)
EDGE
high
low
Amount of gene overlap
White Blood Cell Differentiation
Neuron Projection Morphogenesis
Neural Development
Sensory Perception
Learning or Memory
Social Behaviour
Endothelium/ Epithelium Development
Growth Factor Receptor Signalling
Regulation of GTPase Signalling
Regulation of GTPase Activity
G-Protein Signalling
Hippo Signalling Cascade
Wnt Signalling
Regulation of Kinase Cascade
Glutamate Signalling Pathway
Integrin-mediated Signalling Pathway
Seratonin Receptor Signalling Pathway
Innate Immune Response Signalling
Regulation of Immune Response
Cell Junction Assembly
Cell Adhesion
Ion Homeostasis
Nuclear Transport
Ion Transport
Anion Transport
Amino Acid Transport
Ion Transport Regulation
Protein Localization Regulation
Deacetylation
Phosphorylation
Xenobiotic Metabolic Process
Fatty Acid Biosynthesis
Actin Movement
Regulation of Receptor Activity
Nutrient Response
Photoreceptor Cell Maintenance
cGMP Metabolic Process
Vacuole Assembly
Wound Healing
Reproductive Structure Development
Regulation of Cytokinesis
Exocytosis
Platelet Activation
Ion Response
Membrane Fusion
Cell Fate Determination
Regulation of DNA Response to Stress

Figure 3-7 A functional map of predicted PDZ domain biology. An enrichment analysis of the GO biological process terms associated with the predicted PDZ targets for each of the domains from structure-based and sequence-based human proteome scanning was performed. The results were visualized as a network where the nodes represent gene-sets. The colour of the node border represents the number of domains that the gene-set was seen enriched for, among the structure-based predictions. The colour of the node centre represents number of domains that the gene-set was seen enriched for, among the sequence-based predictions. Edges represent the overlap between two connected gene-sets with the thickness corresponding to the number of genes overlapping.

Enrichment results from both sequence and structure-based predictions were plotted on the same map to ease identification of overlapping or unique themes, with sequence-based enrichment scores corresponding to node centre colour and structure-based scores corresponding to node border colour. For example, a number of themes are enriched in targets from both methods, such as 'photoreceptor cell maintenance, 'hippo signalling' and 'cell junction assembly' (i.e. node centre and border are red). Other themes are only enriched in sequence-based (i.e. border is grey, node centre is red) or structure-based targets (i.e. border is red, node centre is grey). For example, 'neuron projection morphogenesis', 'regulation of cytokinesis', and 'innate immune response signalling' themes contain terms only enriched in structure-based predictions, while 'actin movement', 'membrane fusion' and 'nuclear transport' are enriched only in sequence-based targets.

I also compared the themes from the predictions to those from 1249 known PDZ mediated PPIs in the iRefIndex database. Some themes were enriched only in known targets (e.g. 'DNA damage checkpoint', 'negative regulation of angiogenesis'), however many known themes were covered by the predictors (e.g. 'cell junction assembly', 'ion homeostasis', 'neural development'). I identified the theme 'xenobiotic metabolic process' (enriched in both sequence-based and structure-based predictions) to be novel as it did not correspond to any themes seen in the known interaction network and did not have any PDZ interactions reported in the literature (based on a manual search). For this theme, both predictors predicted PDZ domain interactions with enzymes that are important for catalyzing foreign compounds in the xenobiotic metabolism pathway. For example the sequence-based predictor predicted the domain DVL1L1-1 to interact with cytochrome P450 (HGNC:CYP19A1) and dimethylaniline monooxygenase (HGNC:FMO1)

(Eling and Curtis 1992; Omiecinski et al. 2011), FRMPD4-1 to interact with various glutathione S-transferases (e.g. HGNC:GSTA1, GSTA2, GSTA3), MAST4-1 to interact with prostaglandin G/H synthase (HGNC:PTGS1). The domains SDCBP-1, SDCBP2-1 were predicted by the structure-based predictor to interact with bisphosphate nucleotidase (HGNC:BPNT1). The domains CAR14-1, CNKRS2-1, CNKRS3-1, SNX27-1, WHRN2-1 and the domains DLG4-2, GRIP1-1, MAGI2-6, MPDZ-1, TJP2-3 and TJP3-3 were predicted by the sequence-based and structure-based predictors respectively to interact with various sulfotransferases (e.g. HGNC:SULT1C2, SULT4A1, SULT1B1, SULT1E1, SULT1A1, SULT1A2, SULT1A4) (**Figure 3-8**).

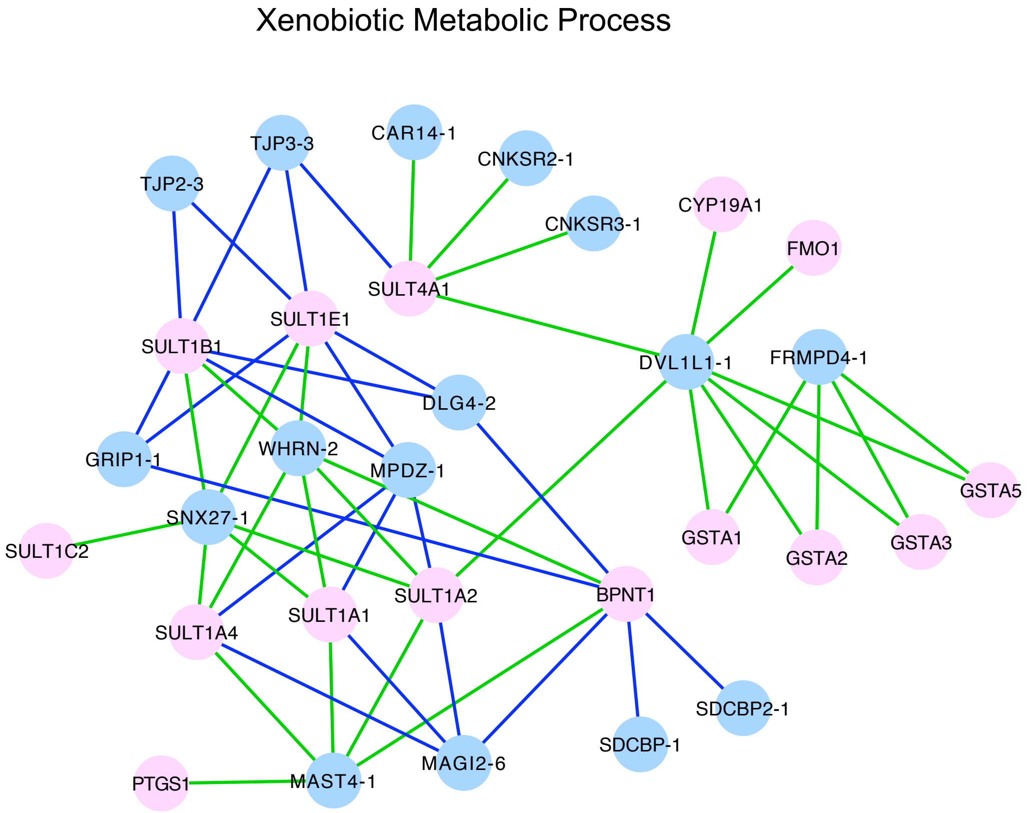


Figure 3-8 A network view of predicted novel PDZ interactions in xenobiotic metabolism. PDZ domains are shown as blue nodes and labelled using their gene names. Protein targets are shown as pink nodes and labelled using their HGNC gene symbols. Blue edges represent structure-

based only predicted interactions. Green edges represent sequence-based only predicted interactions. Only interactions involving proteins with GO annotations are presented.

In some cases, although the themes were also enriched in the iRefIndex map, only limited information about PDZ domain involvement in the associated process was found in the literature. These themes represent opportunities for the predictions to shed light on the role of PDZ domains where little is currently known. One example is ‘wound healing’, where both predictors predicted PDZ domains to interact with proteins involved in different stages of wound healing. These included platelet activators and aggregators (e.g. HGNC:CD9 (Zhang et al. 2012), P2RY12 (Klepeis et al. 2004)), growth factor receptors (e.g. HGNC:PDGFRA (Lynch et al. 1987), TGFBR1 (Liu et al. 2011), HGF (Bevan et al. 2004)), plasma membrane calcium-transporting ATPases (e.g. HGNC:ATP2B1, ATP2B2, ATP2B3, ATP2B4 (Talarico 2010)), calcium-activated potassium channels (e.g. HGNC:KCNMA1, KCNMB2 (Becchetti and Arcangeli 2010)), fibrinogen (HGNC:FGG) (Laurens et al. 2006), coagulation factors (e.g. HGNC:F8, F11 (Inbal and Dardik 2006)), immune system proteins such as chemokines (e.g. HGNC:CXCR1, CXCR2, CCL19 (Gillitzer and Goebeler 2001)), tumour necrosis factors (e.g. HGNC:TNFAIP6, TNF (Barrientos et al. 2008)) and inhibitor of nuclear factor kappa-β kinase (HGNC:IKBKB) (Barrientos et al. 2008)).

Finally, the predictions also suggested additional targets for well studied processes that are known to involve PDZ domains. For ‘Wnt signalling’, both predictors predicted known interactions between the domain MAGI3-2 and Frizzled-4 and 7 as well as domains DLG4-1,2 and frizzled-1,2,4 and 7 (Wawrzak et al. 2009). However, several other PDZ domains were also predicted to interact with Frizzled family members. Some examples include AHNAK2-1, CAR14-1, CNKSR2-1 (structure-based) and MPDZ-13, PDZRN4-1, SYNJ2BP-1 (sequence-based) which are all predicted to interact with one or more Frizzled family members (HGNC:FZD1, FZD2, FZD4, FZD7, FZD10). Interactions which may negatively regulate Wnt signalling were also predicted and involve F-box-like proteins (HGNC:TBL1X, TBL1XR1 (Lagna et al. 1999)) and human colorectal mutant cancer protein (HGNC:MCC) (Fukuyama et al. 2008).

Many functional themes identified consist of multiple different enriched terms containing multiple proteins, predicted to interact with several PDZ domains. These patterns involve many protein targets and are unlikely to occur by chance. Thus, this functional analysis provides additional validation of the prediction methods and highlights novel PDZ interactors involved in a variety of biological processes.

3.3 Discussion

I have presented a structure-based predictor of PDZ domain-peptide interactions that can be used to scan C-terminal proteomes to predict PDZ domain mediated PPIs. This predictor utilizes domain structure features derived from the whole domain, focusing on a core peptide-binding site defined by ten highly conserved amino acid positions. Combined with the use of experimentally determined and computationally generated training negative interactions, the predictor achieves high cross validation results and is expected to generalize well to unseen interactions in practice. Compared to the previous sequence-based predictor, the structure-based predictor is less dependent on training-testing domain sequence similarity and predicts many new validated interactions in human. As a result, the structure-based predictor is complementary to the sequence-based predictor and both should be used to identify candidates for further biological experiments and to expand our knowledge of PDZ domain mediated PPIs.

An important technical result of this work is the use of computationally generated negatives to supplement training and reduce over-prediction. I showed that the negative interactions in current experimental data sets do not adequately cover the negative proteome space resulting in a predictor that returns many hits that are likely false positives. While this problem is more apparent for the structure-based predictor, it also affects the sequence-based predictor, as there are several domains where sequence-based proteome scanning predicts thousands of hits, and likely affects other sequence-based predictors. Since additional experimentally determined negatives for training are limited, using computationally generated negatives is required. While PWMs can be used to computationally generate such negatives as discussed in the previous chapter, such methods do not model dependencies between ligand positions and depend on a user or naively defined cutoff to discriminate between positives and negatives. Here, I use a semi supervised learning approach utilizing an SVM to generate additional negatives, since SVMs can better address the limitations faced by PWMs. As a result, the proteome scanning performance was improved by reducing the number of false positive hits that would otherwise be returned. As this problem is not unique to the structure-based predictor, training with additional negatives is likely to benefit other predictors as well.

Comparing proteome scanning hits to known PPIs, there is only a moderate overlap in hits predicted by both the structure-based and sequence-based predictor. While this suggests that the predictors are complementary and thus should both be used, there are cases when using either the structure-based or sequence-based predictor to find interactors may be more appropriate. For example, when the training-testing domain sequence similarity is < 0.7, the structure-based predictor may be more useful, since its performance is less dependent on sequence similarity at lower similarity levels. In fact, when the sequence similarity is very low the sequence-based predictor may fail to return any predictions. For other domains, a reliable structure may not be obtained or modelled, or the required structure features cannot be successfully generated. In this case, the sequence-based predictor may be the only predictor that can be used. However, for the majority of cases, both predictors should be used to find as many hits as possible for a given domain.

Although PDZ domains can recognize motifs internal to a protein, most data is available for domain-C-terminal binding, thus both the structure-based and sequence-based predictors have been trained using this data and are best suited for the prediction of such interactions. Although other similar methods exist on the web, they can only predict that a protein containing a PDZ domain interacts with another protein (Szklarczyk et al. 2011) or are best suited for interactions between PDZ domains and specific types of proteins (e.g. membrane proteins) (Bhardwaj et al. 2007). Thus, I expect the website will be useful to biologists in helping to further map the many processes mediated by PDZ domains.

While the current structure-based predictor performs well, other domain structure related features should be considered in the future. For example, it is known that the structural flexibility of the PDZ domain binding pocket can contribute to the domain’s ability to bind specific ligands (Zhang et al. 2006; Chen et al. 2007). Recently, a model of PDZ domain backbone flexibility was used to successfully predict domain binding specificity, but for a subset of human PDZ domains (Smith and Kortemme 2010). Thus, domain backbone flexibility features should be considered as they may help to improve predictor performance. Another structure related feature, which should also be considered, is binding pocket geometry and shape. Although I explored the use of 3D-Zernike descriptors (La et al. 2009), their use did not benefit the predictor. However, there are other shape descriptors such as real spherical harmonic coefficients that could be investigated that may improve predictor performance (Morris et al. 2005). Although I have built an entirely structure-based predictor, additional features including sequence features can be combined to build a single predictor that utilizes all available types of information. Finally, since the predictor predicts in vitro interactions, incorporating contextual information such as co-expression and protein location will help to build a more physiologically relevant map of PDZ domain mediated protein-protein interactions.

3.4 Methods

3.4.1 Domain binding site definition

A number of positions in the PDZ domain that are in close contact with the peptide are important for binding (Chen et al. 2008; Tonikian et al. 2008). For this work, I defined the binding site using ten domain binding site positions (core positions) that are in close contact with the peptide ligand (< 4.5 Å) across nine PDZ domain structures. In total, 218 out of 267 human PDZ domains could be used because they don't have gaps in their binding sites based on a PDZ family multiple sequence alignment (eight structures), and I could obtain structures and compute features for them (41 structures). For mouse, fly and worm, respectively, 178 of 237, 85 of 117 and 64 of 81 known PDZ domains are supported with 11, 14 and seven of the remaining domains containing gaps. All PDZ domains were defined by HMMER 3.0 (Eddy 2011) against UniProt defined PDZ proteins as of Apr 2012. Overall, the structure-based predictor supports the majority of PDZ domains (i.e. 82%, 74%, 73% and 79% of known PDZ domains) for human, mouse, fly and worm, respectively.

Although previous studies used a binding site definition of 16 domain positions (a superset of the ten used here), these positions were identified from only a single PDZ domain-peptide complex structure (Chen et al. 2008) and many domains contain gaps using this larger 16-position binding site definition (based on a multiple sequence alignment with other PDZ domains). To justify the use of using the smaller ten-position binding site definition, I used the results of different cross validation strategies to compare two predictors built using training sets defined using the different binding site definitions. Using the 16-position binding site definition, 556 positive and 1167 negative interactions corresponding to 58 domains were used for training. The cross validation AUC scores for the ten-position domain binding site definition was higher across all strategies. This indicates that the information in the smaller binding site definition is adequate to achieve good predictor results and it was not necessary to train with additional features from the 16 binding site positions. Since the ten positions are also based on multiple PDZ domain structures, these positions likely capture more general features about PDZ domain binding compared to the 16 positions which were derived from a single structure and may contain noise when applied to other PDZ domains. Finally, using the minimum number of features for training helps to prevent the predictor from becoming overfit and further justifies the use of the ten-position binding site definition (**Table 3-5**).

	ROC AUC		**PR AUC**	
	10 positions	16 positions	10 positions	16 positions
10 Fold	**0.96**	0.936	**0.936**	0.894
Domain	**0.872**	0.840	**0.785**	0.708
Peptide	**0.935**	0.907	**0.909**	0.844
Domain+Peptide	**0.927**	0.925	**0.886**	0.878

Table 3-5 Cross validation results for predictors trained using a ten-position vs. a 16-position domain binding site definition.

3.4.2 Domain structure data

The initial set of PDZ domain structures consists of one NMR and 17 X-ray structures for human collected from the Protein Data Bank (PDB) (Berman et al. 2000) with corresponding interaction data from phage display or protein microarray experiments (Stiffler et al. 2007; Tonikian et al. 2008). Five NMR structures were collected from the PDB for mouse. For NMR structures, only the first model was used. Homology models were used to increase the number of structures available for domain structure feature encoding. In total, 11 human and 54 mouse PDZ domain models were modelled by SWISS-MODEL (Arnold et al. 2006) (downloaded Feb-Sep 2011) through the Protein Model Portal, which is a website providing access to structure models generated by different protein structure resources (Arnold et al. 2009). All training models have greater than 50% sequence similarity to their template structure (average 90%). The minimum QMEAN score for the training models is 0.520 (average 0.836). Please see **Appendix B, Table B-1** for details on all training domains.

3.4.3 Domain-peptide interaction data

PDZ domain-peptide interactions were collected from published high throughput phage display and protein microarray experiments for human and mouse, respectively (Stiffler et al. 2007; Tonikian et al. 2008). Since the phage display data consisted of only positive interactions (of which many could be non-genomic, meaning not similar to any genomic peptide), I used the protocol described in the previous chapter to filter the interactions for genomic interactions and to generate artificial negative interactions. A minor modification of this procedure was adopted to allow for the inclusion of additional class II type PDZ domains to increase coverage of the PDZ family – the minimum number of genomic peptides required for inclusion was relaxed from ten to four. Only domains with both positive and negative interaction data were used for predictor training.

3.4.4 Domain structure feature encoding

Structure features across the entire PDZ domain structure were computed and values corresponding to the ten core binding site positions were extracted from the larger list of features computed for all domain positions. Four types of structure features (detailed below) involved in protein folding and stability were computed to describe the PDZ domain structure (**Figure 3-9**).

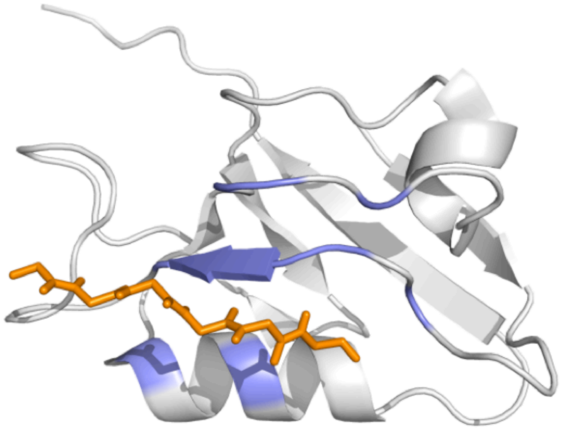


Figure 3-9 3D structure of a complexed PDZ domain in complex with a peptide. The ten core domain binding sites are highlighted in blue and the bound peptide is in orange. PDB:2OQS (NMR first model).

In total, the PDZ domain structure as defined by the core positions was represented by a vector of length 240 features. Each value in the feature vector was scaled to lie between 0 and 1. Details regarding software parameters used to compute the following structure features are available in **Appendix B**.

3.4.4.1 Solvent accessibility, hydrogen bonding and positive phi angle properties

The first feature type consists of five values describing protein structure and were computed using the JOY web server (Mizuguchi et al. 1998). Solvent accessibility indicates whether the protein surface in the area at the given core residue position is available to interact with ligands. Therefore, the first value indicates whether a given residue is solvent accessible or inaccessible. Patterns of hydrogen bonding are important in forming protein secondary and tertiary structure and are known to be important for canonical C-terminal peptide binding to the PDZ domain. The next three values indicate if there is a residue side chain hydrogen bonded to a main chain amide, carbonyl or another side chain. Finally, since positive main chain phi angles may restrict what types of residues may be accommodated at the given position, the last value indicates if the residue has a positive phi angle. These binary features (i.e. absence is 0, presence is 1) were computed for each core residue position resulting in a binary vector of length 50 (5 features x 10 core positions).

3.4.4.2 Solvent accessible area

The second feature type consists of a single value indicating how much surface (i.e. area) for a core residue is available for binding to a ligand residue. This feature was computed using the SURFV software (Sridharan et al. 1992) for each residue resulting in a numeric vector of length 10 (1 feature x 10 core positions).

3.4.4.3 Electrostatic potential and hydrophobicity

Protein-protein interactions are facilitated by the electrostatic and hydrophobic complementarity of molecular surfaces. Therefore, the third and fourth feature types describe the electrostatic potential and hydrophobicity along the surface of the domain. At each core residue position, nine values were sampled from the surface resulting in a total of 90 electrostatic and 90 hydrophobicity values (9 features x 10 core positions). These features were generated by the VASCo software (Steinkellner et al. 2009).

Three-dimensional geometric descriptors were investigated but were not included because they resulted in inferior cross validation performance (**Figure 3-10**).

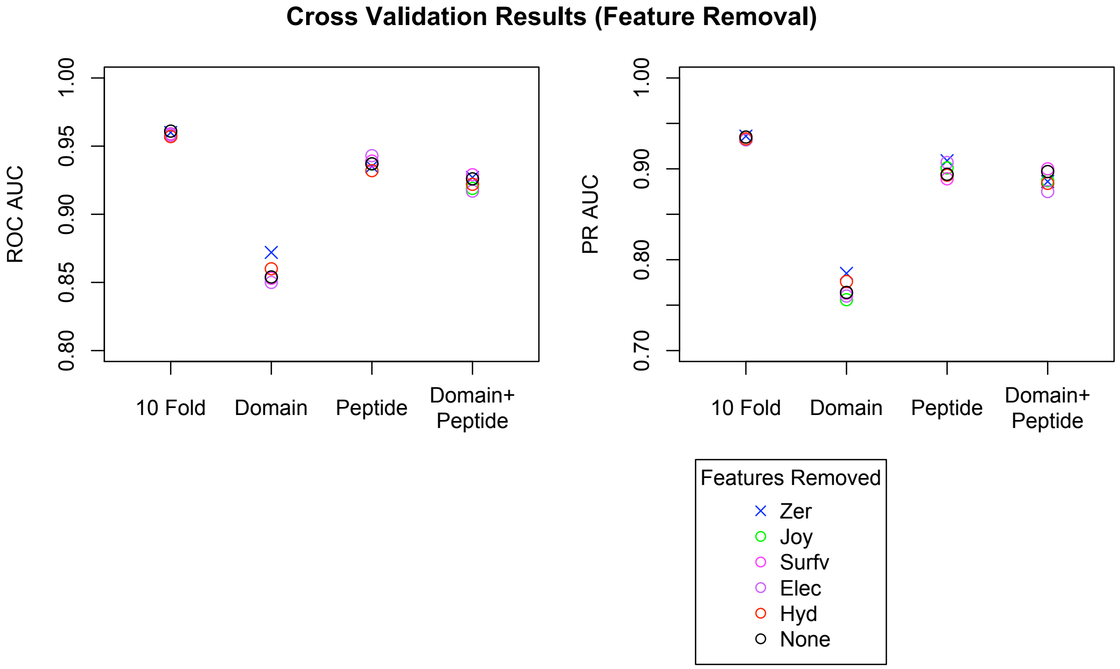


Figure 3-10 Cross validation results for structure-based predictors trained using different combinations of structure features. Initially, five types of structure features were considered for feature encoding: Joy (solvent accessibility, hydrogen bonding), Surfv (solvent accessible area), VASCo (electrostatics), VASCo (hydrophobicity) and 3D Zernike descriptors (structure shape). Five predictors were trained with all but one of the feature sets and the performance for multiple cross validation strategies was measured. For all strategies except for the leave-12%-of-domains-out, the performance across all predictors is comparable. For the strategy that involved leaving sets of domains out, the performance improves only if the 3D Zernike descriptors are not used. Therefore, the final domain structure feature encoding did not include these features.

3.4.5 Peptide sequence feature encoding

Peptides were encoded using a sparse binary vector encoding with each residue in the peptide of length five represented using a binary vector of length 20. The vectors were concatenated to form the final feature vector of length 100.

3.4.6 Semi supervised negative training set expansion

An initial predictor was built using the data for 88 PDZ domains described above. A preliminary assessment of the predictor’s proteome scanning performance was performed by scanning the human proteome (defined by genome assembly Ensembl:37.64) for each domain in the training set. This initial predictor returned a large number of hits (1000 or more) for over half of the domains with an average number of predictions returned per domain of over 2000. (**Figure 3-11** Left boxplot).

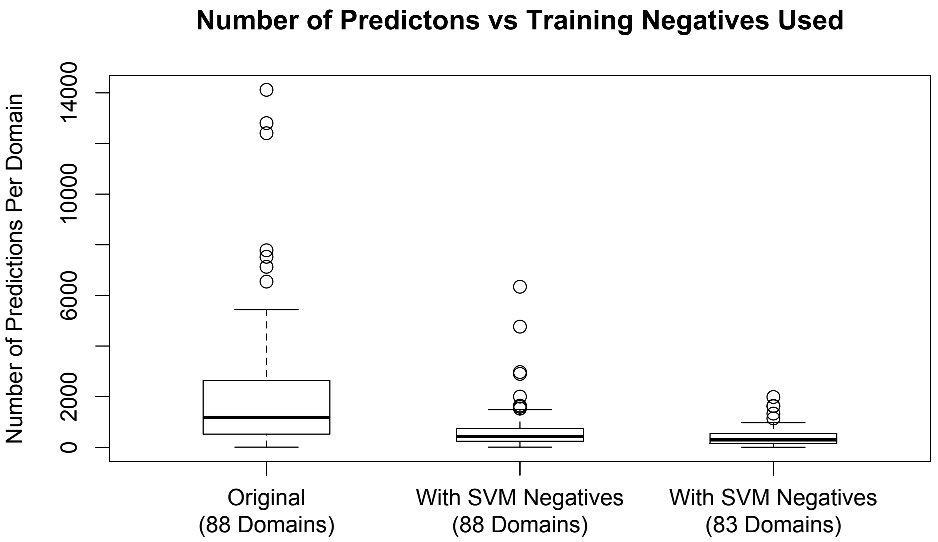


Figure 3-11 Number of hits returned by different structure-based predictors during negative training set expansion. (Left Boxplot) An initial predictor was built using all available training data corresponding to 88 PDZ domains. When proteome scanning was performed for only the training domains, the predictor returned a large number of hits (1000 or more) for over half of the domains. In general, the mean number of predictions returned per domain was over 2000. (Middle Boxplot) Additional negative training data was generated by using an SVM to scan a pool of proteomic human or mouse peptides. The resulting predictor predicted 1000 or more hits for 18% of training domains with a mean number of predictions returned per domain of 685.

(Right Boxplot) For five domains, the predictor still predicted over 2000 interactions and I considered these to be outliers and removed these domains from the training set. The final predictor uses training data for 83 PDZ domains. The average number of predictions per training domain returned by the final predictor was 406

Since previous phage display experiments detected fewer than a hundred binders per domain among billions of random peptides, the majority of these initial predictions are likely false positives. I surmised that the initial negative training data did not adequately cover the negative proteomic interaction space. Therefore, I used a semi supervised learning approach similar to a method previously used to expand negative training data sets when there are no negatives initially available (Wang et al. 2006). This predictor was used to scan the human proteome for interactors of training domains as I did for the initial predictor. I found that adding negatives reduced the number of hits returned per domain. Specifically, when I used this predictor to scan the human proteome for interactors of training domains, fewer domains (16 out of 88 domain or 18%) still had 1000 or more predicted hits. For all but one domain, which had no change, the number of predictions returned per domain was lower than before, with an average number of predictions returned per domain of approximately 685 (**Figure 3-11** Middle Boxplot). However, for five domains, the predictor still predicted over 2000 interactions. These were considered to be outliers and removed these domains from the training set. The above steps were repeated to train the final predictor using a total of 942 positive and 1843 negative interactions involving 83 PDZ domains and 872 peptides. A final scan for only training domains, revealed that the final predictor predicted 1000 or more hits for only five out of 83 domains (approximately 6% of training domains). The average number of predictions per domain returned by the final predictor was approximately 400 (**Figure 3-11** Right Boxplot). I did not remove any more domains from the training set to avoid removing too many positive interactions from the data set. The final predictor was trained using a total of 942 positive and 1843 negative interactions involving 83 PDZ domains and 872 peptides (**Table 3-6**).

		Domain		**Interactions**	
Organism	**Source**	**# Pos**	**# Neg**	**# Pos**	**# Neg**
Mouse	Protein microarray	58	53	527	1026
Mouse	SVM Negatives	-	24	-	210
Human	Phage Display	25	-	415	-

Human	PWM Negatives	-	25	-	407
Human	SVM Negatives	-	20	-	200
	Totals	83	-	942	1843

Table 3-6 Summary of domain-peptide interaction data used for training. PWM negatives are artificial negative interactions generated using PWMs are described in Chapter 2.

3.4.7 Functional enrichment analysis

A gene function enrichment analysis was performed on the predicted sequence-based and structure-based gene interactors using GO biological process terms (Ashburner et al. 2000). The BiNGO (Biological Network Gene Ontology tool) software library (Maere et al. 2005) was used to determine the enriched terms. The hypergeometric test was used to compute a p-value assessing the GO term enrichment for a given set of predicted genes. Multiple testing correction was performed using the Benjamini and Hochberg False Discovery Rate correction. GO v1.2 (downloaded Dec 7, 2011) and human GO annotations (downloaded Dec 7, 2011) were used. Only gene-sets with between five and 300 genes were used from the GO ontology (defined by the GMT file dated Dec 6, 2011). A list of enriched terms (p-value < 0.05 and FDR < 0.1) with more than one gene interactor and associated with more than two domains were retained. To better interpret the structure-based and sequence-based enrichment results, I created an enrichment map, a network-based visual representation of enriched terms that groups similar terms and eases identification of functional themes. The Enrichment Map Cytoscape plugin software to create the enrichment map (Shannon et al. 2003; Merico et al. 2010), using the parameters p-value < 0.05, FDR Q value < 0.1 and "Jaccard + overlap similarity" cutoff = 0.517.

Chapter 4

Predicting Physiologically Revelant PDZ Mediated Protein-Protein Interactions in Human

This chapter is the basis of a manuscript which will be submitted for publication: Hui, S., Jain, S., Yao, Z., Stagljar, I., Bader, GD. (2013). Predicting physiologically relevant PDZ mediated protein-protein interactions in human.

Author contributions: I collected the data, performed the analyses and PDZ domain cloning. Shobhit Jain developed and implemented the Bayesian integration system. Zhong Yao performed the membrane yeast two-hybrid experiments. Igor Stagjlar supervised the experiments. Gary D. Bader supervised and advised this project.

4 Predicting physiologically relevant PDZ mediated protein-protein interactions in human

4.1 Introduction

Although computational predictors (including the ones presented in the previous chapters) can be used to predict PDZ domain-peptide interactions, these interactions may not be physiologically relevant (i.e. occur in the cell) resulting in a potentially large number of false positives. It is well known that protein-protein interactions are influenced by different cellular constraints. For example, for an interaction to take place both the proteins should have correlated gene expression profiles, be part of same biological process and be present in the same cellular compartment. Therefore, information about proteins obtained from diverse biological data sources such as gene expression profiles, cellular location of proteins, functional annotation (molecular function and biological process), sequence signatures, literature, known experimental interactions can be used to identify physiologically relevant interactions among a given set of predicted interactions. These different biological data sources can then be combined using machine learning approaches to classify protein pairs as interacting or non-interacting.

In this chapter, a Bayesian integration system was used to combine gene expression profiles, gene function similarity (molecular function, biological process), cellular location information, sequence signatures and binding site conservation to score predicted PDZ domain-peptide interactions from sequence-based and structure-based SVM predictors. The result is a set of high confidence and predicted physiologically relevant interactions resulting in a 97% reduction in the number of initial predictions. Using the reduced set of interactions, I created a high confidence and physiologically relevant PDZ interaction map in human. I also characterized the PDZ domain targets, by performing a gene function enrichment analysis and showed that the interactors are enriched in known and novel PDZ mediated biological processes. Finally, several novel interactions involving the Frizzled-7 G protein-coupled receptor protein were verified using membrane yeast two-hybrid assay (MYTH).

4.2 Results

4.2.1 Bayesian integration system achieves high cross validation results

We used ten fold cross validation to estimate the performance of the Bayesian integration system. The system achieves high ROC and PR AUC scores of 0.83 and 0.82 respectively and is estimated to perform well in practice (**Figure 4-1**).

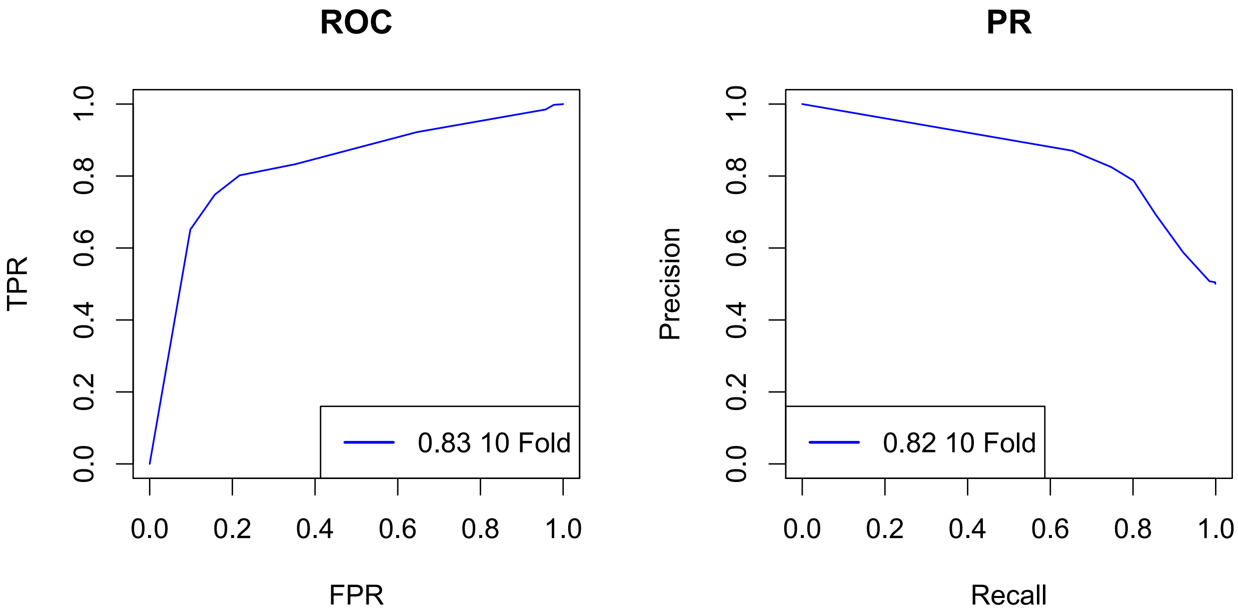


Figure 4-1 Bayesian integration performance estimation using ten fold cross validation.

To determine the contribution of each biological evidence source to the integration system, we repeated the ten fold cross validation using a predictor built with all evidence sources except one. The results show that only by using all sources may the highest ROC and PR AUCs (0.83 and 0.82 respectively) be achieved (**Figure 4-2**).

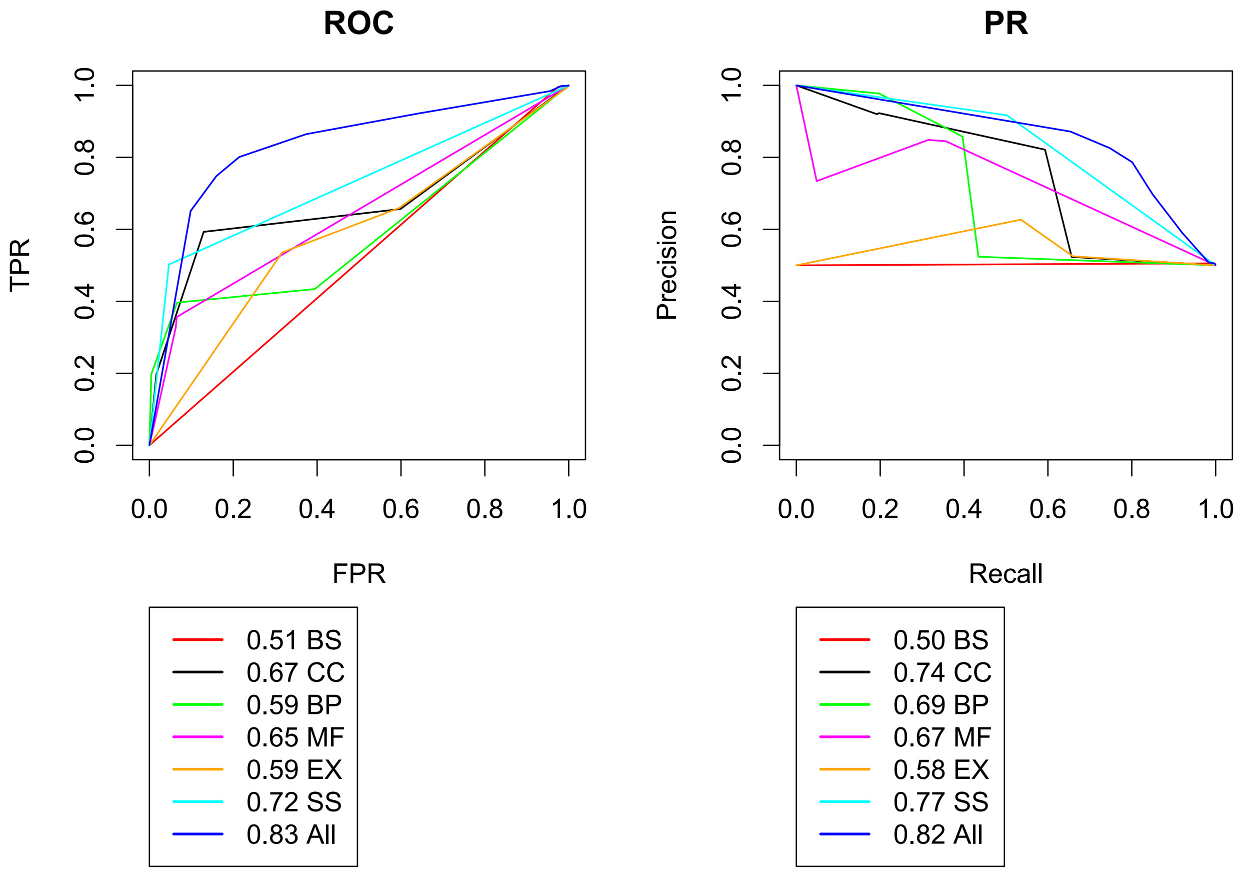


Figure 4-2 Ten fold cross validation results for predictors built using all evidence sources except one. BS = binding site conservation, CC = cellular component, BP = biological process, MF = molecular function, EX = expression correlation, SS = sequence signature.

4.2.2 Bayesian integration system correctly predicts blind interactions in human

Blind tests for human were carried out to obtain an unbiased measure of the performance of the integration system. Positive interactions were obtained from PDZBase (Beuming et al. 2005) and negative interactions were obtained from Luck et al. and were manually curated from the literature (Luck et al. 2011). In total, 59 domains and 51 positive and 68 negative interactions in human were used and the predictor achieved ROC and PR AUC scores of 0.695 and 0.656 respectively (**Figure 4-3**).

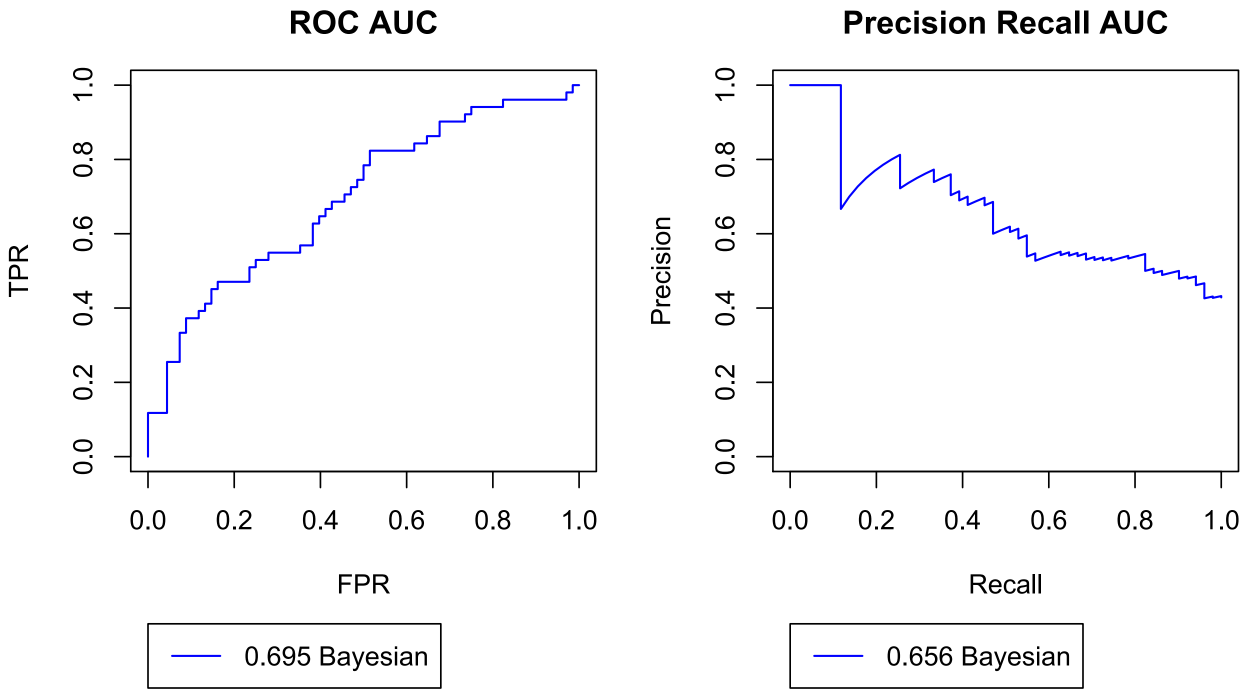


Figure 4-3 Blind testing performance for the Bayesian integration system in human.

These results show that the integration system is able to correctly predict many blind positive and negative interactions in human.

4.2.3 Bayesian integration system substantially reduces the number of initial predictions

The sequence-based and structure-based predictors were used to scan the human proteome for interactions for 222 and 215 PDZ domains, respectively. In total, there were 106,792 unique interactions made by either the sequence-based or structure-based predictor. This initial set was used as input into the Bayesian integration system. To obtain a final set of high confidence predictions, only interactions with scores 0.9 or above and predicted by both predictors were selected from this set. This cutoff was chosen based on cross validation results which showed that at this threshold, the estimated true positive rate is 0.65 while the false positive rate is 0.097 (**Figure 4-1**). The number of unique interactions in the final high confidence set was 3,380 involving 127 PDZ domains and is a 97% reduction in the number of initial interactions.

A subset of predictions that could be verified as true or false positives according to PDZBase (Beuming et al. 2005) and Luck et al. (Luck et al. 2011) were used to estimate if more true positive or more false positive interactions were filtered out by the integration system. Ideally, the latter is desired as this indicates that the system produces a better quality set of predictions by filtering out more incorrect false positive predictions. The number of true and false positives present in the set of interactions before Bayesian integration was 51 and 68 respectively. After integration, the number of true and false positives was 28 and 14 respectively. Therefore, of the interactions that were filtered out, 79% or 54 interactions were false positives while 45% or 23 were true positives (**Figure 4-4**).

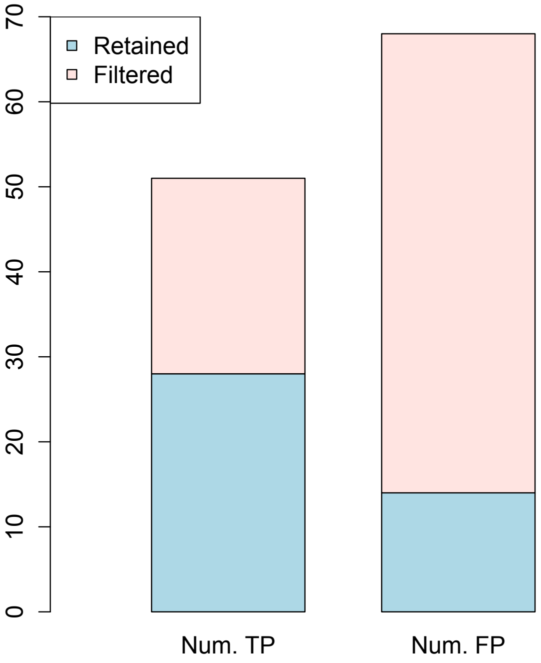


Figure 4-4 Number of true and false positives filtered out using Bayesian integration. The number of initial predictions was 106,792. The number of retained interactions (i.e. not filtered) was 3,380 (blue). The number of filtered interactions was 103,412 (pink).

Although a larger set of known interactions should be used for more rigorously verification, this result suggests that the integration system removes more false positives than true positives.

4.2.4 Construction of a physiologically relevant PDZ mediated protein-protein interaction network in human

A high confidence physiologically relevant PDZ mediated protein-protein interaction network was constructed using the Bayesian integrated interactions (**Figure 4-5**).

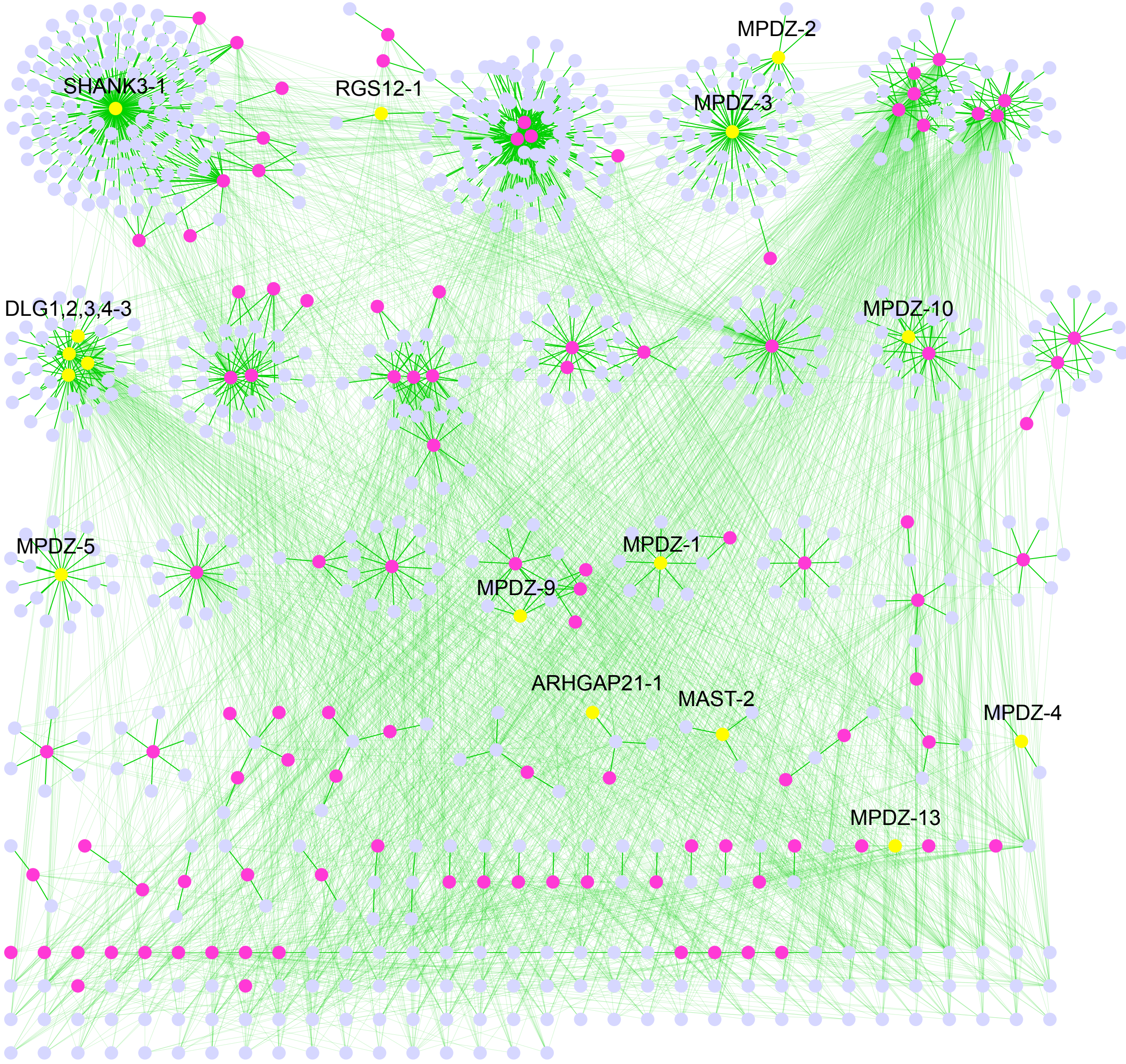


Figure 4-5 Physiologically relevant high confidence protein-protein interaction network in human. The network consists of 127 domains 773 proteins and 3,380 edges. To ease identification of structure in the network, the clusterMaker Cytoscape plugin (Morris et al. 2011) was used to layout the network. The MCL algorithm was used to organize domains (i.e. pink nodes) into clusters (i.e. nodes connected by darker green edges) connected to interacting proteins (i.e. purple nodes) and to visualize inter-connected clusters (i.e. lighter green edges). Yellow nodes highlight domains discussed in the main text. All nodes and edges are depicted in the network.

Several large clusters are clearly visible and suggest examples of promiscuous PDZ domains (i.e. domains which are connected to over 50 binding partners) and may be associated with different functional roles. For example, one of the largest clusters is centered around SHANK3-1 which is connected to 189 partners with functions enriched in a variety of processes such as 'ion transport', 'neurogenesis', 'photoreceptor cell maintenance' and 'actin filament organization' (p-value < 0.05). In contrast, the network is dominated by many small clusters (i.e. domains which are connected to 10 or less binding partners) which may indicate more selective domains with more specific biological functions. For example, ARHGAP21-1 is connected to six binding partners with the majority of functions enriched in the regulation of GTPase activity (p-value $<$ 0.05). In general, the node degree distribution follows a power law distribution ($p > 0.272$, K-S test) and the average node degree is 26.7.

The network also shows a large number of inter-cluster edges (i.e. edges which connect many domains to the same binding partners). This suggests a potentially high degree of cross-selectivity which is a known characteristic of PDZ domains. Over 990 domain pairs have high binding overlap scores of over 0.25, where overlap is computed as the intersection of interactions / union of interactions for domain pairs with a union of 10 or more interactions. These include related domains such as the DLG1,2,3,4-3 domains which have a high average binding site sequence similarity (> 0.921) and high average overlap score of 0.614. This type of group appears as multiple pink nodes within the same cluster. On the other hand, the domain pair RGS12-1 and MAST2-1 which are unrelated domains, have a low binding site sequence similarity of 0.438, but have an overlap score of 0.307. These domains are depicted as nodes in different clusters. The binding preferences of these examples were visualized as sequence logos and show that these domains bind similar class I targets (similarity is > 0.83) (**Figure 4-6** First and Second Columns).

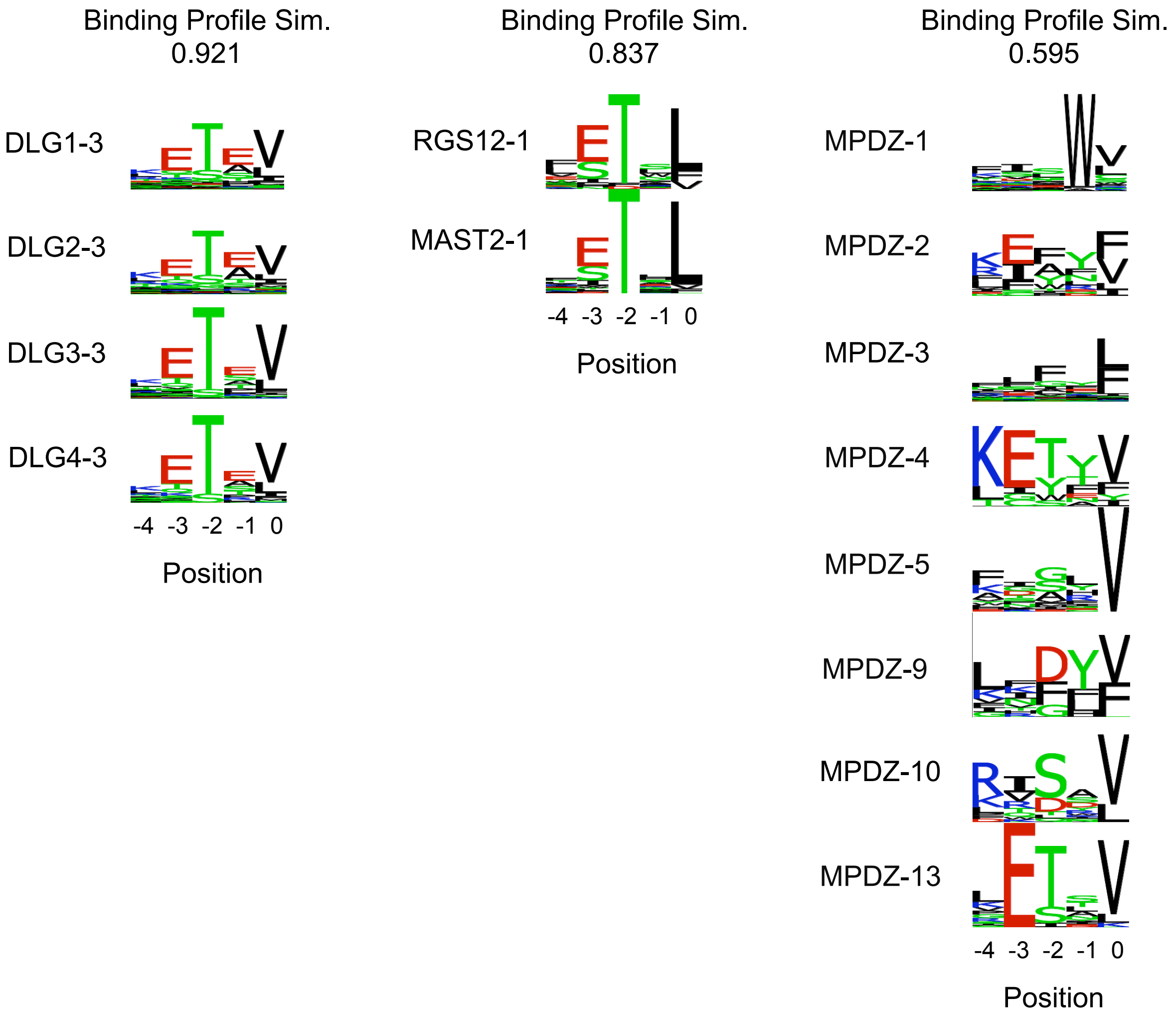


Figure 4-6 Binding specificities of PDZ domains with high and low degrees of target overlap.

In contrast, some clusters have less overlap and suggest cases where cross-selectivity may be minimized. For example, MPDZ-1,2,3,4,5,9,10,13 domains have an average overlap of 0.03 and bind different targets (**Figure 4-6** Third Column). As these are domains on the same multiple PDZ containing protein (MPDZ), this may enable multiple binding partners to interact simultaneously and efficiently during protein complex assembly.

As I have shown here, network visualization can be used to highlight cases for more detailed study of the network interactions themselves to further our understanding of different properties of PDZ domains (i.e. promiscuity vs. selectivity, cross-selectivity). Although the network is constructed using high confidence predicted interactions, experimental validation to verify and support any findings is required.

4.2.5 Predicted PDZ targets are enriched in known and novel biological functions

To determine biological processes which are enriched among the predicted PDZ binding targets, an enrichment map was created using the high confidence physiologically relevant predictions (**Figure 4-7**).

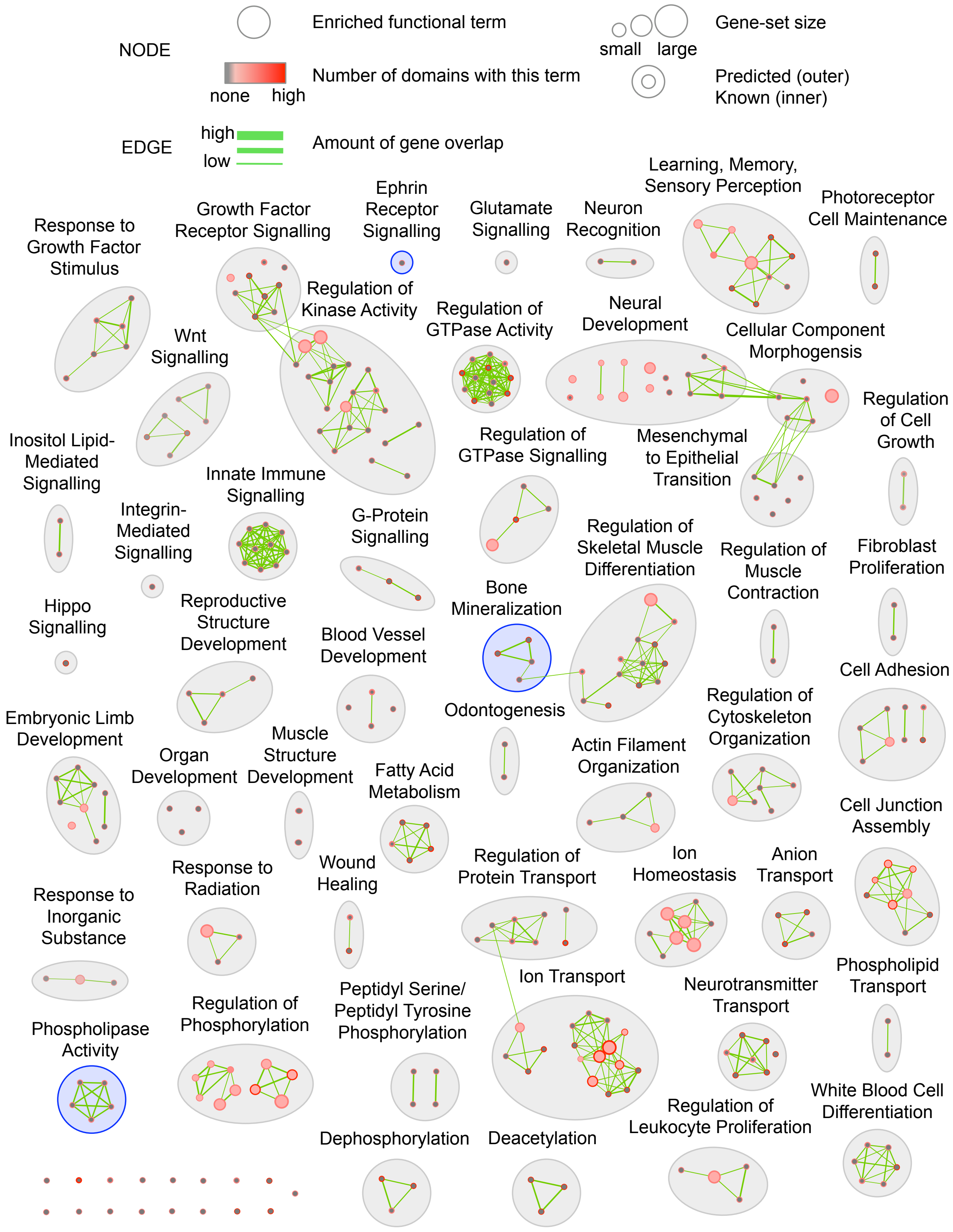

NODE
Enriched functional term
Number of domains with this term
none
high
Gene-set size
small
large
Predicted (outer)
Known (inner)
EDGE
high
low
Amount of gene overlap
Response to Growth Factor Stimulus
Growth Factor Receptor Signalling
Ephrin Receptor Signalling
Glutamate Signalling
Neuron Recognition
Learning, Memory, Sensory Perception
Photoreceptor Cell Maintenance
Regulation of Kinase Activity
Regulation of GTPase Activity
Neural Development
Cellular Component Morphogensis
Wnt Signalling
Regulation of Cell Growth
Inositol Lipid-Mediated Signalling
Innate Immune Signalling
Regulation of GTPase Signalling
Mesenchymal to Epithelial Transition
Integrin-Mediated Signalling
G-Protein Signalling
Regulation of Skeletal Muscle Differentiation
Regulation of Muscle Contraction
Fibroblast Proliferation
Hippo Signalling
Reproductive Structure Development
Bone Mineralization
Blood Vessel Development
Cell Adhesion
Regulation of Cytoskeleton Organization
Embryonic Limb Development
Odontogenesis
Muscle Structure Development
Organ Development
Actin Filament Organization
Fatty Acid Metabolism
Cell Junction Assembly
Regulation of Protein Transport
Ion Homeostasis
Anion Transport
Response to Radiation
Wound Healing
Response to Inorganic Substance
Phospholipid Transport
Ion Transport
Neurotransmitter Transport
Peptidyl Serine/ Peptidyl Tyrosine Phosphorylation
Regulation of Phosphorylation
Phospholipase Activity
White Blood Cell Differentiation
Regulation of Leukocyte Proliferation
Dephosphorylation
Deacetylation

Figure 4-7 A functional map of physiologically relevant PDZ domain biology. An enrichment analysis of the GO biological process terms associated with the predicted targets for each of the

domains after Bayesian integration was performed. The results were visualized as a network where the nodes represent gene-sets. The colour of the node border represents the number of domains that the gene-set was seen enriched for, among predicted targets. The colour of the node centre represents the number of domains that the gene-set was seen enriched for among known targets. Edges represent the overlap between two connected gene-sets with the thickness corresponding to the number of genes overlapping. Blue circled clusters represent new themes not seen in the previous enrichment map (**Figure 3-7**) or in iRefIndex.

In this map, I also compared the enrichments to those obtained from known PDZ domain interactors found in the iRefIndex database (Razick et al. 2008). Therefore, nodes represent enriched terms and edges connect terms with overlapping genes. Node borders are coloured according to whether or not the term was seen enriched among predicted interactors (grey means term is not enriched, red means term is enriched), while node center colourings correspond to enrichment seen among known interactors. Functional themes are clusters of related terms and are circled for easy identification. In order to visualize enrichment with respect to predicted interactors, nodes with enrichment seen only for known interactors were not included.

Several biological processes known to involve PDZ domains (i.e. neural development, cell junction assembly, ion transport) are enriched among predicted and known targets (nodes with red borders and centers) supporting the validity of many of the predictions. This map was compared to the previous enrichment map (**Figure 3-7**) which was created using the full set of interactions before Bayesian integration. In general, many of the terms in the previous map are still present after filtering including those related to ‘white blood cell differentiation’, ‘innate immune response signalling’ and ‘fatty acid metabolism. However, some themes are no longer present such as ‘xenobiotic metabolism’, ‘vacuole assembly’ and ‘regulation of cytokinesis’. Missing themes are those that may not be physiologically relevant because they consist of targets that may be considered to bind PDZ domains only *in vitro*. Finally, several new themes (i.e. not seen in the previous map) emerge as a result of Bayesian integration. This is due to the large reduction in the number of predicted interactors per domain (i.e. sampling size) as a result of filtering which increases the statistical power of enrichments that were previously too weak to pass significance testing. Some examples of these themes are highlighted as blue coloured clusters and include ‘regulation of phospholipase activity’, ‘ephrin receptor signalling’, ‘bone mineralization’ and are also not enriched among known interactors in iRefIndex. Many novel (i.e. not found in the literature or in iRefIndex) interactions are predicted. For example, for 'ephrin receptor signalling', PDZ domains INADL-5, 9 and MPDZ-11 are predicted to interact with ephrin type-A receptor 7 (HGNC:EPHA7) (Pasquale 1997). For 'bone mineralization', several domains including DLG1-1, LIN7A-1, MAGI1-6 are predicted to interact with activin receptor type-2A (HGNC: ACVR2A) (Ebisawa et al. 1999). For 'regulation of phospholipase activity', domains DLG1,2,4-3 are predicted to interact with endothelin-1 receptor (HGNC: EDNRA) (Ambar and Sokolovsky 1993).

4.2.6 Novel PDZ mediated interactions are experimentally validated

Frizzled receptors are a subset of the G protein-coupled receptor family and are involved in mediating Wnt signalling pathways which are responsible for establishing basic developmental processes in the embryo and tissue homeostasis of organs in the adult (Schulte and Bryja 2007). In mammals, there are ten different Frizzled proteins. Eight contain C-terminal PDZ binding motifs and have been shown to interact with PDZ domains of other proteins (Wawrzak et al. 2009). For example, interactions between Frizzled-1,2,4,7 proteins and Disheveled PDZ domains are important for proper Wnt signalling function in Xenopus (Umbhauer et al. 2000; Wong et al. 2003). Frizzled-4,5,7,8 have also been shown to interact with MAGI3 in mouse to mediate ciliogenesis and non canonical Wnt signalling (Yao et al. 2004). Finally Frizzled-1,2,4,7 are also known to interact with DLG4-1 and 2 to facilitate clustering of adenomatous polyposis coli proteins (Hering and Sheng 2002).

Many interactions involving Frizzled-7 (FZD7) were predicted of which several are known including those between Frizzled-7 and domains in the PSD-95 family of proteins and the MAGI3-2 domain (Yao et al. 2004). We selected five novel PDZ interactions involving Frizzled-7 and experimentally validated them using membrane yeast two-hybrid (**Figure 4-8**).

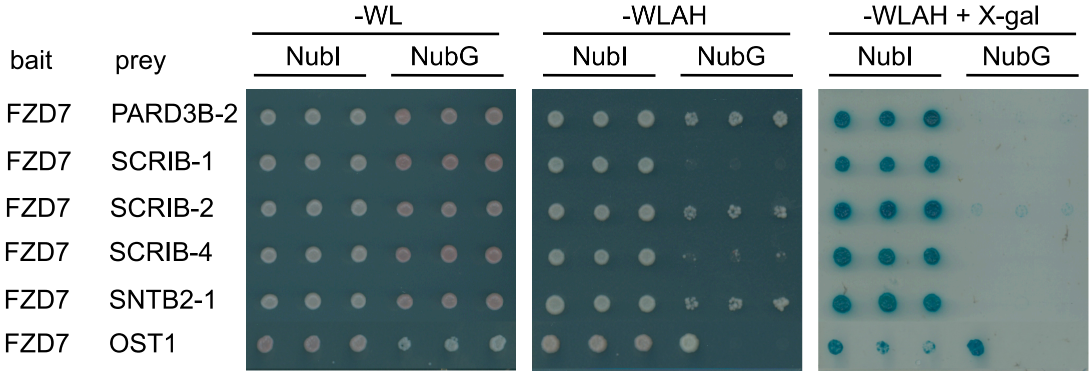


Figure 4-8 Membrane yeast two-hybrid validation of novel PDZ interactions involving Frizzled-7 (FZD7). The bait (FZD7) construct and prey construct (NubI or NubG version) were cotransformed into yeast cells (S. cerevisiae NMY51). NubI version of prey is used as expression control. Prey OST1 is used as negative control.

Two confirmed interactions involve Scribble (SCRIB-2, SCRIB-4) which, along with Frizzled, is a core component functioning within the planar cell polarity pathway (McNeill 2010). Two other confirmed interactions involve PAR3B and SNTB2 proteins which may play a role in the proper formation of renal epithelial tight junctions and neuromuscular junctions respectively (Lumeng et al. 1999; Gao et al. 2002). Interestingly, Frizzled expression has been reported in renal and skeletal muscle cells (Janssens et al. 2004; Korkut and Budnik 2009). Since PDZ proteins often function as scaffolding proteins, these interactions suggest novel PDZ involvement possibly in the organization of protein complexes to facilitate proper cell polarity and formation of cell junctions.

4.3 Discussion

In this chapter, I have presented a Bayesian integration system which assigns confidence scores to interactions in a given set of PDZ mediated PPIs based on different lines of biological evidence. Using this system, a set of previously predicted PDZ domain-peptide interactions in human was substantially reduced by 97%. Blind testing using a limited number of interactions suggests that the interactions removed are mostly false positives resulting in a higher quality set of predictions.

Other Bayesian methods exist that integrate biological evidence from different sources to score protein interactions (Li et al. 2008), however, our method consists of several improvements. For example, compared to existing methods, our system uses more gene expression data sets and this has been shown to improve predictor performance (Jain 2011). We also use gene function similarity to determine whether two proteins are annotated to similar GO terms. This measurement has also been shown to be better than the ones used by other methods which merely assess similarity using the intersection of terms between proteins (Jain and Bader 2010).

Although the method works well, the evidence sources used have limitations and therefore, false positives, although reduced, may still exist. For example, two proteins may be annotated to the same GO cellular component (or have a high semantic similarity score for cellular component), but this does not necessarily mean that they are interacting. Similarly, even though two proteins are found to be co-expressed they may not interact *in vivo* since other factors such as protein concentration also affect whether two proteins will interact. Despite these limitations, using multiple lines of evidence in combination helps to improve predictor performance and addresses the limitations faced when using only single sources of evidences.

A large number of predictions were filtered out by the Bayesian integration system and our use of a strict cutoff score. Many of these predictions may actually be filtered out because the Bayesian system does not compute a score for them due to a lack of information for one or both proteins involved in the interactions. As more information about proteins is available and used for training, the more accurate and comprehensive the system will be.

Since the Bayesian system was trained using randomly paired negative human interactions, the the predictor's performance may improve if real negatives were used for training. Such negatives are available in a limited number for human (Luck et al. 2011) however, mouse negatives from protein microarray experiments (Stiffler et al. 2007) could be considered for training the system in the future.

Finally, since our system uses information at the protein level, it can be used to filter other domain mediated protein-protein interaction predictions as well (e.g. WW or SH3 mediated interactions).

4.4 Methods

4.4.1 SVM prediction of PDZ mediated protein-protein interactions

Prediction of PDZ domain-peptide interactions were performed using the sequence-based and structure-based predictors in Chapters 1 and 2. These predictors were trained using experimentally determined PDZ domain-peptide interactions from high throughput protein microarray and phage display experiments for mouse and human, respectively (Stiffler et al. 2007; Tonikian et al. 2008). For the sequence-based predictor, residue information at contacting positions in the domain binding site and peptide were obtained from a PDZ domain structure complexed with a peptide ligand. For the structure-based predictor, domain structure features were mined from the binding sites (consensus site determined from nine PDZ complex structures) of PDZ domain structures (i.e. experimental structures or homology models). Features used include factors known to facilitate protein folding and stability such as electrostatics, hydrophobicity, solvent accessibility, patterns of hydrogen bonding and phi torsion angles. Amino acid sequence features were used for the peptide. The predictors were used to scan the human proteome (defined by genome assembly Ensembl:GRCh37.64) for targets of hundreds of PDZ domains.

4.4.2 Gold standard training set

A gold standard training set was created using 1322 known interactions involving at least one PDZ containing protein from iRefIndex (Razick et al. 2008). A negative interaction set of equal size was created by randomly pairing proteins from the known interaction set and ensuring that they did not correspond to known positive interactions.

4.4.3 Bayesian integration of biological evidences

4.4.3.1 Cellular location, biological process, molecular function

The GO is a popular taxonomy of controlled biological terms that can be used to assess the functional relationship between different proteins (Ashburner et al. 2000). GO organizes knowledge of cellular location, biological process, and molecular function of different proteins in three orthogonal ontologies. The strength of the relationship between proteins annotated to these ontologies can be quantified using semantic similarity. A high semantic similarity value between two proteins indicates that they participate in similar pathways or cellular components and are thus more likely to physically interact in the cell than randomly selected proteins. The Topological Clustering Semantic Similarity (TCSS) metric was used to compute the semantic similarity between GO terms annotated to proteins in the predicted protein interaction dataset (Jain and Bader 2010).

4.4.3.2 Gene Expression

If two or more genes are similarly expressed over multiple conditions in a gene expression experiment, they are more likely to be related in function. Multiple studies have shown that a strong correlation exists between gene expression profiles of interacting protein pairs when compared to random pairs (Ge et al. 2001; Grigoriev 2001; Jansen et al. 2002; Bhardwaj and Lu 2005). Therefore, high correlation between gene expression profiles of interacting proteins provides evidence in support of that interaction. Gene expression profiles from 117 studies were downloaded from the GeneMANIA gene function prediction website (www.genemania.org) (Mostafavi et al. 2008) and an average Pearson correlation was calculated using Fisher's *z* transformation (Faller 1981).

4.4.3.3 Sequence signature

Regions or sites of interest in a protein sequence (i.e. sequence signatures) can be used to predict novel interactions between two proteins (Shen et al. 2008). Such regions may correspond to short sequence motifs, binding sites, enzyme active sites or other local secondary structure. In particular, protein interactions have been predicted in the past by identifying pairs of domains enriched among a set of known interacting proteins (Ng et al. 2003; Betel et al. 2004; Rhodes et al. 2005).

We use the information content (IC) score as defined below to determine co-occurring domains within proteins with experimentally verified interactions.

$$IC(A,B) = \sum_{i,j} -\log_2\left(\frac{p_{ij}}{p_i p_j}\right) \qquad \textbf{Eq. 4.1}$$

where in the verified protein-protein interaction set, p_{ij} is the probability of seeing domain *i* in one protein and domain *j* in the other protein, p_i is the probability of seeing domain *i*, p_i is the probability of seeing domain *j*. Protein A is predicted to interact with protein B if IC(A,B) is above a given threshold. Information about domains for a given protein was obtained from the Protein Domains section, Domains subsection in Ensembl (defined by genome assembly Ensembl:GRCh37.62) using BioMart. Domains were indentified by their InterPro PFAM IDs.

4.4.3.4 Binding site conservation

The more conserved a binding site is, the more functionally relevant it is. Therefore, the conservation score for a given PDZ target was computed by finding a given protein's ortholog among mouse, worm and fly proteomes. Orthology information was downloaded from Ensembl's BioMart (June 2012). If the ortholog existed, the score was determined by computing the Hamming distance between the ortholog's last five residues with the given protein's last five residues (score between 0 and 1.0). This was done for all mouse, worm and fly orthologs and the maximum score was reported. If there were no orthologs, the score was set to be -1.

4.4.4 Protein-protein interaction network

The clusterMaker Cytoscape plugin was used to layout and determine highly connected nodes in the protein-protein interaction network (Morris et al. 2011). The network interactions indicated with edge attribute values of one were input into the MCL clustering algorithm. Default clustering parameters were used (weak edge pruning = 10E-15, number of iterations = 16, max residual value = 0.001, max number of threads = 0). Inter-cluster edges were restored after the network was automatically laid out. Nodes representing protein interactors were coloured purple, nodes representing domains were coloured pink. Edges connecting intra-cluster nodes (i.e. edges forming a cluster) were coloured a more opaque shade of green than edges connecting inter-cluster nodes (i.e. edges connecting clusters).

4.4.5 Membrane Yeast Two-Hybrid Assay

MYTH assay was performed as described previously (Snider et al. 2010). Briefly, FZD-7 and Ryk cDNAs were cloned to bait vector pTMBVa by gap repair. Prey cDNAs were cloned to pGPR3N by Gateway LR cloning. A pair of bait and prey vectors were cotransformed into NMY51 yeast cells, and the yeast cells were grown on SD-WL plates. After colony formation, three independent colonies from each assay were picked and grown on SD - WLAH or SD – WLAH + X-gal plates. Positive interactions were counted as those that could grow on SD - WLAH or SD – WLAH + X-gal plate. NubI version of preys were used as expression control.

Chapter 5

Summary And Future Directions

5 Summary and future directions

5.1 Thesis Summary

My thesis focuses on building computational predictors of PDZ domain-peptide interactions for the purposes of proteome scanning. These predictors were trained using high throughput interaction data from mouse and human with additional biological evidence sources added in a second stage to identify physiologically relevant interactions. A machine learning framework to build the predictors was established and followed throughout the thesis. The main components of the framework were data collection, feature encoding, predictor construction and performance evaluation. By following this framework, the predictors could be systematically and efficiently built and this enabled easier comparison.

In Chapter 2, I presented a sequence-based PDZ domain-peptide interaction predictor which was built using a support vector machine. This predictor was trained using protein microarray data in mouse and phage display data in human. In order to use the phage display data for training, which only contained positive interactions, I developed a novel method to generate artificial negative interactions using positive weight matrices. Using cross-validation and a series of independent tests, I showed that the predictor successfully predicted interactions in different organisms (i.e. mouse, worm and fly). I then used the predictor to scan the proteomes of human, worm and fly to predict binders for several PDZ domains. Predictions were validated using known genomic interactions and published protein microarray experiments. Based on the results, novel PDZ interactions potentially associated with Usher and Bardet-Biedl syndromes were predicted. A comparison of performance measures for the predictor and other existing sequence-based predictors demonstrated the predictor's improved accuracy and precision at proteome scanning.

In Chapter 3, I presented a structure-based predictor of PDZ domain-peptide interactions. Since domain structure is known to influence binding specificity I hypothesized that structural information could be used to predict new interactions not predicted by the sequence-based predictor presented in Chapter 2. A technical result of this work was the use of a semi supervised predictor to computationally generate artificial negatives to supplement training and reduce the problem of over-prediction. This predictor was also used to scan the human proteome for ligands of hundreds of PDZ domains. By comparing the structure-based predictions to the sequence-based proteome scanning predictions, I showed that indeed the structure-based predictor is complementary to the sequence-based predictor, finding unique known and novel protein-protein interactions. Furthermore, I showed that the structure-based predictor is also less dependent on training-testing domain sequence similarity. A functional enrichment analysis of the sequence and structure-based predicted PDZ targets was used to create a map of PDZ domain biology. This map highlighted PDZ domain involvement in diverse biological processes, some only found by the structure-based predictor. Based on this analysis, novel PDZ domain involvement in xenobiotic metabolism was identified and new interactions for other processes including wound healing and Wnt signalling were suggested. An online resource (http://webservice.baderlab.org/domains/POW) was made to enable users to access the two predictors.

In Chapter 4, I presented a Bayesian integration system to combine gene expression profiles, gene function similarity (molecular function, biological process, cellular component), sequence signatures and binding site conservation to score the PDZ domain-peptide interactions predicted by the sequence-based and structure-based predictors. The result was a set of high confidence and physiologically relevant interactions representing a substantial reduction in the number of initial predictions. A comparison of predictor performance measures showed that the integration system mainly filtered out false positives resulting in a set of higher quality predictions. Using this reduced set of interactions I created a high confidence and physiologically relevant PDZ interaction map for human. The PDZ domain targets were analyzed, by performing an enrichment analysis which showed that the targets were enriched in known and novel PDZ mediated biological processes.

5.2 Future Directions

5.2.1 Proteome scanning for other domains

The PDZ domain is one of many known PRMs and has been the focus of this thesis because of its simple mode of target recognition and availability of high throughput experimental interaction data. Data sets are also now available for other PRMs including WW, SH3 and SH2 domains and have resulted the development of predictors for these domains as well. As discussed in this thesis, such predictors should be built using all available data for training and should utilize a variety of different features (not just sequence information) in order to build predictors which have the greatest coverage, accuracy and can be used for the purposes of proteome scanning in multiple organisms. As more data sets are published, this will enable the construction of multiple predictors each capable of predicting interactions for a family of domains. Not only does this depend on the availability of experimentally determined interactions but the availability of other types of data which can be used for training, including solved protein structures and cellular contextual information about protein interactions. As more predictors are built, we can start to obtain a more complete picture of the cellular interactome mediated by PRMs.

5.2.2 Predictor training using additional features

Sequence and structure-based information have been the main source of features for predictor training in this thesis. However, the use of different and complementary features for training will result in a predictor that is capable of identifying new interactions compared to the ones predicted by existing methods. This would help to further expand the current coverage of interactions and to strengthen confidence in current predictions. For instance, protein backbone flexibility has been shown to produce a predictor that is capable of predicting interactions for a subset of PDZ domains (Smith and Kortemme 2010) and should be considered as an additional structure-related feature that can be explored for predictor training. Binding pocket geometry and shape is another feature which should also be used. Although this was explored in the form of 3D-Zernike descriptors (La et al. 2009) in Chapter 3, it was shown not to benefit the structure-based predictor. However, other shape descriptors can be investigated such as real spherical harmonic coefficients (Morris et al. 2005) to see if they improve predictor performance. Extending these types of features to other domains or incorporating them as additional features for PDZ interaction prediction will improve predictor performance and coverage.

The addition of other biological sources of evidence would benefit the Bayesian integration system discussed in Chapter 4. These would include information from network topology analysis (based on the principle that two proteins that have many shared neighbors in a protein-protein interaction network are more likely to interact) and text mining (interaction information automatically extracted from the literature). Ideally information about the abundance or concentrations of proteins within the cell would also be available.

5.2.3 Mapping changes in PDZ mediated interaction networks

Another useful future direction for this work is to study PDZ mediated network rewiring caused by evolutionary or disease-related mutations. Kim et al., studied the changes in C-terminal binding sequence of PDZ targets from PDZBase and phage display experiments and suggested rewiring of PDZ domain-peptide interactions as a mechanism for the development of new protein functions in human (Kim et al. 2012). Ideally, PDZ interaction networks across different organisms could be aligned to provide insight into the function and evolution of different parts of the proteome involving domain containing proteins. Since the predictors presented here can perform proteome scanning for PDZ domains in multiple organisms (human, mouse, worm and fly), it is now possible to construct such networks for PDZ proteins and further enable the study of network evolution with methods such as network alignment algorithms.

Studying disease-related mutations and their rewiring effects on underlying domain mediated interaction networks can help to study the functional impact of such changes. For example, point mutations in phosphorylation sites on cancer genes and their effects on the post translational modifications by protein kinase domains were recently studied (Reimand and Bader 2012). Information about disease genes, disease mutations including cancer mutations are available in various databases (Hamosh et al. 2005; Stenson et al. 2009; Forbes et al. 2011) and can be used to identify disease-related mutations in PDZ proteins in specific-disease pathways. Predictors can then be used to determine changes in binding targets thus highlighting candidates for further study. Since the predictors discussed here are suitable for proteome scanning for unmutated PDZ domains, additional work needs to be performed to assess their ability to accurately predict interactions for mutated PDZ domains. However, if possible this type of analysis can shed more light on known and new roles of PDZ domains in disease.

References

Ambar, I. and Sokolovsky, M. (1993). "Endothelin receptors stimulate both phospholipase C and phospholipase D activities in different cell lines." Eur J Pharmacol **245**(1): 31-41.

Appleton, B. A., Zhang, Y., Wu, P., Yin, J. P., Hunziker, W., Skelton, N. J., Sidhu, S. S. and Wiesmann, C. (2006). "Comparative structural analysis of the Erbin PDZ domain and the first PDZ domain of ZO-1. Insights into determinants of PDZ domain specificity." J Biol Chem **281**(31): 22312-22320.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K. and Hermjakob, H. (2010). "The IntAct molecular interaction database in 2010." Nucleic Acids Res **38**(Database issue): D525-531.

Arnold, K., Bordoli, L., Kopp, J. and Schwede, T. (2006). "The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling." Bioinformatics **22**(2): 195-201.

Arnold, K., Kiefer, F., Kopp, J., Battey, J. N., Podvinec, M., Westbrook, J. D., Berman, H. M., Bordoli, L. and Schwede, T. (2009). "The Protein Model Portal." J Struct Funct Genomics **10**(1): 1-8.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Atchley, W. R., Zhao, J., Fernandes, A. D. and Druke, T. (2005). "Solving the protein sequence metric problem." Proc Natl Acad Sci U S A **102**(18): 6395-6400.

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T. and Hogue, C. W. (2001). "BIND--The Biomolecular Interaction Network Database." Nucleic Acids Res **29**(1): 242-245.

Barrientos, S., Stojadinovic, O., Golinko, M. S., Brem, H. and Tomic-Canic, M. (2008). "Growth factors and cytokines in wound healing." Wound Repair Regen **16**(5): 585-601.

Basdevant, N., Weinstein, H. and Ceruso, M. (2006). "Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study." J Am Chem Soc **128**(39): 12766-12777.

Becchetti, A. and Arcangeli, A. (2010). "Integrins and ion channels in cell migration: implications for neuronal development, wound healing and metastatic spread." Adv Exp Med Biol **674**: 107-123.

Ben-Hur, A. and Noble, W. S. (2006). "Choosing negative examples for the prediction of protein-protein interactions." BMC Bioinformatics **7 Suppl 1**: S2.

Benkert, P., Tosatto, S. C. and Schomburg, D. (2008). "QMEAN: A comprehensive scoring function for model quality assessment." Proteins **71**(1): 261-277.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-242.

Betel, D., Isserlin, R. and Hogue, C. W. (2004). "Analysis of domain correlations in yeast protein complexes." Bioinformatics **20 Suppl 1**: i55-62.

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

Bevan, D., Gherardi, E., Fan, T. P., Edwards, D. and Warn, R. (2004). "Diverse and potent activities of HGF/SF in skin wound repair." J Pathol **203**(3): 831-838.

Bhardwaj, N. and Lu, H. (2005). "Correlation between gene expression profiles and protein-protein interactions within and across genomes." Bioinformatics **21**(11): 2730-2738.

Bhardwaj, N., Stahelin, R. V., Zhao, G., Cho, W. and Lu, H. (2007). "MeTaDoR: a comprehensive resource for membrane targeting domains and their host proteins." Bioinformatics **23**(22): 3110-3112.

Bock, J. R. and Gough, D. A. (2001). "Predicting protein--protein interactions from primary structure." Bioinformatics **17**(5): 455-460.

Brinkworth, R. I., Breinl, R. A. and Kobe, B. (2003). "Structural basis and prediction of substrate specificity in protein serine/threonine kinases." Proc Natl Acad Sci U S A **100**(1): 74-79.

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L. and Cesareni, G. (2010). "MINT, the molecular interaction database: 2009 update." Nucleic Acids Res **38**(Database issue): D532-539.

Ceol, A., Chatr-aryamontri, A., Santonico, E., Sacco, R., Castagnoli, L. and Cesareni, G. (2007). "DOMINO: a database of domain-peptide interactions." Nucleic Acids Res. **35**: D557-560.

Chang, C.-C. and Lin, C.-J. (2011). "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology **2**(3): 27:21--27:27.

Chen, J. R., Chang, B. H., Allen, J. E., Stiffler, M. A. and MacBeath, G. (2008). "Predicting PDZ domain-peptide interactions from primary sequences." Nat Biotechnol **26**(9): 1041-1045.

Chen, Q., Niu, X., Xu, Y., Wu, J. and Shi, Y. (2007). "Solution structure and backbone dynamics of the AF-6 PDZ domain/Bcr peptide complex." Protein Sci **16**(6): 1053-1062.

Cristianini, N. and Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge; New York, Cambridge University Press.

Dev, K. K. (2004). "Making protein interactions druggable: targeting PDZ domains." Nat Rev Drug Discov **3**(12): 1047-1056.

Donnes, P. and Elofsson, A. (2002). "Prediction of MHC class I binding peptides, using SVMHC." BMC Bioinformatics **3**: 25.

Doorbar, J. (2006). "Molecular biology of human papillomavirus infection and cervical cancer." Clin Sci (Lond) **110**(5): 525-541.

Ebisawa, T., Tada, K., Kitajima, I., Tojo, K., Sampath, T. K., Kawabata, M., Miyazono, K. and Imamura, T. (1999). "Characterization of bone morphogenetic protein-6 signaling pathways in osteoblast differentiation." J Cell Sci **112 (Pt 20)**: 3519-3527.

Eddy, S. R. (2011). "Accelerated Profile HMM Searches." PLoS Comput Biol **7**(10): e1002195.

Eley, L., Yates, L. M. and Goodship, J. A. (2005). "Cilia and disease." Curr Opin Genet Dev **15**(3): 308-314.

Eling, T. E. and Curtis, J. F. (1992). "Xenobiotic metabolism by prostaglandin H synthase." Pharmacol Ther **53**(2): 261-273.

Eo, H. S., Kim, S., Koo, H. and Kim, W. (2009). "A machine learning based method for the prediction of G protein-coupled receptor-binding PDZ domain proteins." Mol. Cells **27**: 629-634.

Faller, A. J. (1981). "An Average Correlation Coefficient." Journal of Applied Metereology **20**(2): 203-205.

Fawcett, T. (2006). "An introduction to ROC analysis." Pattern Recogn Lett **27**: 861-874.

Fernandez-Ballester, G., Beltrao, P., Gonzalez, J. M., Song, Y. H., Wilmanns, M., Valencia, A. and Serrano, L. (2009). "Structure-based prediction of the Saccharomyces cerevisiae SH3-ligand interactions." J. Mol. Biol. **388**: 902-916.

Ferraro, E., Via, A., Ausiello, G. and Helmer-Citterich, M. (2006). "A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity." Bioinformatics **22**(19): 2333-2339.

Fischer, D. (2006). "Servers for protein structure prediction." Curr Opin Struct Biol **16**(2): 178-182.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A. K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H. S., Ritchie, G. R., Ruffier, M., Schuster, M., Sobral, D., Tang, Y. A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X. M., Harrow, J., Herrero, J., Hubbard, T. J., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A. and Searle, S. M. (2012). "Ensembl 2012." Nucleic Acids Res **40**(Database issue): D84-90.

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J. W., Campbell, P. J., Stratton, M. R. and Futreal, P. A. (2011). "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer." Nucleic Acids Res **39**(Database issue): D945-950.

Fournane, S., Charbonnier, S., Chapelle, A., Kieffer, B., Orfanoudakis, G., Trave, G., Masson, M. and Nomine, Y. (2011). "Surface plasmon resonance analysis of the binding of high-risk mucosal HPV E6 oncoproteins to the PDZ1 domain of the tight junction protein MAGI-1." Journal of molecular recognition : JMR **24**(4): 511-523.

Fuentes, E. J., Der, C. J. and Lee, A. L. (2004). "Ligand-dependent dynamics and intramolecular signaling in a PDZ domain." J Mol Biol **335**(4): 1105-1115.

Fukuyama, R., Niculaita, R., Ng, K. P., Obusez, E., Sanchez, J., Kalady, M., Aung, P. P., Casey, G. and Sizemore, N. (2008). "Mutated in colorectal cancer, a putative tumor suppressor for serrated colorectal cancer, selectively represses beta-catenin-dependent transcription." Oncogene **27**(46): 6044-6055.

Gao, L., Macara, I. G. and Joberty, G. (2002). "Multiple splice variants of Par3 and of a novel related gene, Par3L, produce proteins with different binding properties." Gene **294**(1-2): 99-107.

Ge, H., Liu, Z., Church, G. M. and Vidal, M. (2001). "Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae." Nat Genet **29**(4): 482-486.

Gillitzer, R. and Goebeler, M. (2001). "Chemokines in cutaneous wound healing." J Leukoc Biol **69**(4): 513-521.

Grigoriev, A. (2001). "A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae." Nucleic Acids Res **29**(17): 3513-3519.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. and McKusick, V. A. (2005). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." Nucleic Acids Res **33**(Database issue): D514-517.

Hering, H. and Sheng, M. (2002). "Direct interaction of Frizzled-1, -2, -4, and -7 with PDZ domains of PSD-95." FEBS Lett **521**(1-3): 185-189.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J. (2010). A practical guide to support vector classifications. D. o. C. S. National Taiwan University.

Hu, H., Columbus, J., Zhang, Y., Wu, D., Lian, L., Yang, S., Goodwin, J., Luczak, C., Carter, M., Chen, L., James, M., Davis, R., Sudol, M., Rodwell, J. and Herrero, J. J. (2004). "A map of WW domain family interactions." Proteomics **4**(3): 643-655.

Huang, H., Li, L., Wu, C., Schibli, D., Colwill, K., Ma, S., Li, C., Roy, P., Ho, K., Songyang, Z., Pawson, T., Gao, Y. and Li, S. S. (2008). "Defining the specificity space of the human SRC homology 2 domain." Molecular & cellular proteomics : MCP **7**(4): 768-784.

Hubbard, T. J., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S. and Flicek, P. (2009). "Ensembl 2009." Nucleic Acids Res **37**(Database issue): D690-697.

Hue, M., Riffle, M., Vert, J. P. and Noble, W. S. (2010). "Large-scale prediction of protein-protein interactions from structures." BMC bioinformatics **11**: 144.

Hui, S. and Bader, G. D. (2010). "Proteome scanning to predict PDZ domain interactions using support vector machines." BMC Bioinformatics **11**: 507.

Inbal, A. and Dardik, R. (2006). "Role of coagulation factor XIII (FXIII) in angiogenesis and tissue repair." Pathophysiol Haemost Thromb **35**(1-2): 162-165.

Ingham, R. J., Colwill, K., Howard, C., Dettwiler, S., Lim, C. S., Yu, J., Hersi, K., Raaijmakers, J., Gish, G., Mbamalu, G., Taylor, L., Yeung, B., Vassilovski, G., Amin, M., Chen, F., Matskova, L., Winberg, G., Ernberg, I., Linding, R., O'Donnell, P., Starostine, A., Keller, W., Metalnikov, P., Stark, C. and Pawson, T. (2005). "WW domains provide a platform for the assembly of multiprotein networks." Mol Cell Biol **25**(16): 7092-7106.

Jain, S. (2011). Literature Review: Computational Methods For (In Vivo) Protein-Protein Interaction Prediction. Toronto, University of Toronto**:** 45.

Jain, S. and Bader, G. D. (2010). "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology." BMC Bioinformatics **11**: 562.

Jansen, R., Greenbaum, D. and Gerstein, M. (2002). "Relating whole-genome expression data with protein-protein interactions." Genome Res **12**(1): 37-46.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. and Gerstein, M. (2003). "A Bayesian networks approach for predicting protein-protein interactions from genomic data." Science **302**(5644): 449-453.

Janssens, N., Andries, L., Janicot, M., Perera, T. and Bakker, A. (2004). "Alteration of frizzled expression in renal cell carcinoma." Tumour Biol **25**(4): 161-171.

Kang, H., Freund, C., Duke-Cohan, J. S., Musacchio, A., Wagner, G. and Rudd, C. E. (2000). "SH3 domain recognition of a proline-independent tyrosine-based RKxxYxxY motif in immune cell adaptor SKAP55." EMBO J **19**(12): 2889-2899.

Kaufmann, K., Shen, N., Mizoue, L. and Meiler, J. (2011). "A physical model for PDZ-domain/peptide interactions." J Mol Model **17**: 315-324.

Kim, J., Kim, I., Yang, J. S., Shin, Y. E., Hwang, J., Park, S., Choi, Y. S. and Kim, S. (2012). "Rewiring of PDZ domain-ligand interaction network contributed to eukaryotic evolution." PLoS Genet **8**(2): e1002510.

Klepeis, V. E., Weinger, I., Kaczmarek, E. and Trinkaus-Randall, V. (2004). "P2Y receptors play a critical role in epithelial cell communication and migration." J Cell Biochem **93**(6): 1115-1133.

Korkut, C. and Budnik, V. (2009). "WNTs tune up the neuromuscular junction." Nat Rev Neurosci **10**(9): 627-634.

La, D., Esquivel-Rodriguez, J., Venkatraman, V., Li, B., Sael, L., Ueng, S., Ahrendt, S. and Kihara, D. (2009). "3D-SURFER: software for high-throughput protein surface comparison and analysis." Bioinformatics **25**(21): 2843-2844.

Lagna, G., Carnevali, F., Marchioni, M. and Hemmati-Brivanlou, A. (1999). "Negative regulation of axis formation and Wnt signaling in Xenopus embryos by the F-box/WD40 protein beta TrCP." Mech Dev **80**(1): 101-106.

Landgraf, C., Panni, S., Montecchi-Palazzi, L., Castagnoli, L., Schneider-Mergener, J., Volkmer-Engert, R. and Cesareni, G. (2004). "Protein interaction networks by proteome peptide scanning." PLoS biology **2**(1): E14.

Laurens, N., Koolwijk, P. and de Maat, M. P. (2006). "Fibrin structure and wound healing." J Thromb Haemost **4**(5): 932-939.

Lehrach, W. P., Husmeier, D. and Williams, C. K. (2006). "A regularized discriminative model for the prediction of protein-peptide interactions." Bioinformatics **22**(5): 532-540.

Lenfant, N., Polanowska, J., Bamps, S., Omi, S., Borg, J. P. and Reboul, J. (2010). "A genome-wide study of PDZ-domain interactions in C. elegans reveals a high frequency of non-canonical binding." BMC genomics **11**: 671.

Li, D., Liu, W., Liu, Z., Wang, J., Liu, Q., Zhu, Y. and He, F. (2008). "PRINCESS, a protein interaction confidence evaluation system with multiple data sources." Mol Cell Proteomics **7**(6): 1043-1052.

Liu, J., Johnson, K., Li, J., Piamonte, V., Steffy, B. M., Hsieh, M. H., Ng, N., Zhang, J., Walker, J. R., Ding, S., Muneoka, K., Wu, X., Glynne, R. and Schultz, P. G. (2011). "Regenerative phenotype in mice with a point mutation in transforming growth factor beta type I receptor (TGFBR1)." Proc Natl Acad Sci U S A **108**(35): 14560-14565.

Lo, S. L., Cai, C. Z., Chen, Y. Z. and Chung, M. C. (2005). "Effect of training datasets on support vector machine prediction of protein-protein interactions." Proteomics **5**(4): 876-884.

Luck, K., Fournane, S., Kieffer, B., Masson, M., Nomine, Y. and Trave, G. (2011). "Putting into practice domain-linear motif interaction predictions for exploration of protein networks." PLoS One **6**(11): e25376.

Lumeng, C., Phelps, S., Crawford, G. E., Walden, P. D., Barald, K. and Chamberlain, J. S. (1999). "Interactions between beta 2-syntrophin and a family of microtubule-associated serine/threonine kinases." Nat Neurosci **2**(7): 611-617.

Lynch, S. E., Nixon, J. C., Colvin, R. B. and Antoniades, H. N. (1987). "Role of platelet-derived growth factor in wound healing: synergistic effects with other growth factors." Proc Natl Acad Sci U S A **84**(21): 7696-7700.

MacBeath, G. and Schreiber, S. L. (2000). "Printing proteins as microarrays for high-throughput function determination." Science **289**(5485): 1760-1763.

Maere, S., Heymans, K. and Kuiper, M. (2005). "BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks." Bioinformatics **21**(16): 3448-3449.

McNeill, H. (2010). "Planar cell polarity: keeping hairs straight is not so simple." Cold Spring Harb Perspect Biol **2**(2): a003376.

Merico, D., Isserlin, R., Stueker, O., Emili, A. and Bader, G. D. (2010). "Enrichment map: a network-based method for gene-set enrichment visualization and interpretation." PLoS One **5**(11): e13984.

Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T. M., Menon, S., Hanumanthu, G., Gupta, M., Upendran, S., Gupta, S., Mahesh, M., Jacob, B., Mathew, P., Chatterjee, P., Arun, K. S., Sharma, S., Chandrika, K. N., Deshpande, N., Palvankar, K., Raghavnath, R., Krishnakanth, R., Karathia, H., Rekha, B., Nayak, R., Vishnupriya, G., Kumar, H. G., Nagini, M., Kumar, G. S., Jose, R., Deepthi, P., Mohan, S. S., Gandhi, T. K., Harsha, H. C., Deshpande, K. S., Sarker, M., Prasad, T. S. and Pandey, A. (2006). "Human protein reference database--2006 update." Nucleic Acids Res **34**(Database issue): D411-414.

Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S. and Overington, J. P. (1998). "JOY: protein sequence-structure representation and analysis." Bioinformatics **14**(7): 617-623.

Morris, J. H., Apeltsin, L., Newman, A. M., Baumbach, J., Wittkop, T., Su, G., Bader, G. D. and Ferrin, T. E. (2011). "clusterMaker: a multi-algorithm clustering plugin for Cytoscape." BMC Bioinformatics **12**: 436.

Morris, R. J., Najmanovich, R. J., Kahraman, A. and Thornton, J. M. (2005). "Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons." Bioinformatics **21**(10): 2347-2355.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. and Morris, Q. (2008). "GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function." Genome Biol **9 Suppl 1**: S4.

Moyer, B. D., Denton, J., Karlson, K. H., Reynolds, D., Wang, S., Mickle, J. E., Milewski, M., Cutting, G. R., Guggino, W. B., Li, M. and Stanton, B. A. (1999). "A PDZ-interacting domain in CFTR is an apical membrane polarization signal." J Clin Invest **104**(10): 1353-1361.

Ng, S. K., Zhang, Z. and Tan, S. H. (2003). "Integrative approach for computationally inferring protein domain interactions." Bioinformatics **19**(8): 923-929.

Omiecinski, C. J., Vanden Heuvel, J. P., Perdew, G. H. and Peters, J. M. (2011). "Xenobiotic metabolism, disposition, and regulation by receptors: from biochemical phenomenon to predictors of major toxicities." Toxicol Sci **120 Suppl 1**: S49-75.

Pasquale, E. B. (1997). "The Eph family of receptors." Curr Opin Cell Biol **9**(5): 608-615.

Patil, A. and Nakamura, H. (2005). "Filtering high-throughput protein-protein interaction data using a combination of genomic features." BMC Bioinformatics **6**: 100.

Pawson, T., Gish, G. D. and Nash, P. (2001). "SH2 domains, interaction modules and cellular wiring." Trends Cell Biol **11**(12): 504-511.

Pawson, T. and Nash, P. (2003). "Assembly of cell regulatory systems through protein interaction domains." Science **300**(5618): 445-452.

Phizicky, E. M. and Fields, S. (1995). "Protein-protein interactions: methods for detection and analysis." Microbiol Rev **59**(1): 94-123.

Ponting, C. P. (1997). "Evidence for PDZ domains in bacteria, yeast, and plants." Protein Sci **6**(2): 464-468.

Razick, S., Magklaras, G. and Donaldson, I. M. (2008). "iRefIndex: a consolidated protein interaction database with provenance." BMC bioinformatics **9**: 405.

Reimand, J. and Bader, G. D. (2012). "Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers." Molecular Systems Biology **Accepted**.

Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A. and Chinnaiyan, A. M. (2005). "Probabilistic model of the human protein-protein interaction network." Nat Biotechnol **23**(8): 951-959.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C. and Mewes, H. W. (2010). "CORUM: the comprehensive resource of mammalian protein complexes--2009." Nucleic Acids Res **38**(Database issue): D497-501.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. and Eisenberg, D. (2004). "The Database of Interacting Proteins: 2004 update." Nucleic Acids Res **32**(Database issue): D449-451.

Sanchez, I. E., Beltrao, P., Stricher, F., Schymkowitz, J., Ferkinghoff-Borg, J., Rousseau, F. and Serrano, L. (2008). "Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm." PLoS Comput. Biol. **4**: e1000052.

Sanner, M. F., Olson, A. J. and Spehner, J. C. (1996). "Reduced surface: an efficient way to compute molecular surfaces." Biopolymers **38**(3): 305-320.

Schulte, G. and Bryja, V. (2007). "The Frizzled family of unconventional G-protein-coupled receptors." Trends Pharmacol Sci **28**(10): 518-525.

Scott, M. S. and Barton, G. J. (2007). "Probabilistic prediction and ranking of human protein-protein interactions." BMC Bioinformatics **8**: 239.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res **13**(11): 2498-2504.

Shao, X., Tan, C. S., Voss, C., Li, S. S., Deng, N. and Bader, G. D. (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.

Shen, R., Chinnaiyan, A. M. and Ghosh, D. (2008). "Pathway analysis reveals functional convergence of gene expression profiles in breast cancer." BMC Med Genomics **1**: 28.

Sidhu, S. S., Bader, G. D. and Boone, C. (2003). "Functional genomics of intracellular peptide recognition domains with combinatorial biology methods." Curr Opin Chem Biol **7**(1): 97-102.

Skelton, N. J., Koehler, M. F., Zobel, K., Wong, W. L., Yeh, S., Pisabarro, M. T., Yin, J. P., Lasky, L. A. and Sidhu, S. S. (2003). "Origins of PDZ domain ligand specificity. Structure determination and mutagenesis of the Erbin PDZ domain." J Biol Chem **278**(9): 7645-7654.

Smialowski, P., Pagel, P., Wong, P., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., Rattei, T., Frishman, D. and Ruepp, A. (2010). "The Negatome database: a reference set of non-interacting protein pairs." Nucleic Acids Res **38**(Database issue): D540-544.

Smith, C. A. and Kortemme, T. (2010). "Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains." J Mol Biol **402**(2): 460-474.

Snider, J., Kittanakom, S., Damjanovic, D., Curak, J., Wong, V. and Stagljar, I. (2010). "Detecting interactions with membrane proteins using a membrane two-hybrid assay in yeast." Nat Protoc **5**(7): 1281-1293.

Songyang, Z., Fanning, A. S., Fu, C., Xu, J., Marfatia, S. M., Chishti, A. H., Crompton, A., Chan, A. C., Anderson, J. M. and Cantley, L. C. (1997). "Recognition of unique carboxyl-terminal motifs by distinct PDZ domains." Science **275**(5296): 73-77.

Sridharan, S., Nicholls, A. and Honig, B. (1992). "A new vertex algorithm to calculate solvent accessible surface area." J. Biophys. **61**(A174).

Stark, C., Breitkreutz, B. J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J. M., Winter, A., Dolinski, K. and Tyers, M. (2011). "The BioGRID Interaction Database: 2011 update." Nucleic Acids Res **39**(Database issue): D698-704.

Steinkellner, G., Rader, R., Thallinger, G. G., Kratky, C. and Gruber, K. (2009). "VASCo: computation and visualization of annotated protein surface contacts." BMC Bioinformatics **10**: 32.

Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Thomas, N. S. and Cooper, D. N. (2009). "The Human Gene Mutation Database: 2008 update." Genome Med **1**(1): 13.

Stiffler, M. A., Chen, J. R., Grantcharova, V. P., Lei, Y., Fuchs, D., Allen, J. E., Zaslavskaia, L. A. and MacBeath, G. (2007). "PDZ domain binding selectivity is optimized across the mouse proteome." Science **317**(5836): 364-369.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J. and von Mering, C. (2011). "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored." Nucleic Acids Res **39**(Database issue): D561-568.

Talarico, E. F., Jr. (2010). "Plasma membrane calcium-ATPase isoform four distribution changes during corneal epithelial wound healing." Mol Vis **16**: 2259-2272.

Tong, A. H., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C. W., Fields, S., Boone, C. and Cesareni, G. (2002). "A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules." Science **295**(5553): 321-324.

Tonikian, R., Xin, X., Toret, C. P., Gfeller, D., Landgraf, C., Panni, S., Paoluzi, S., Castagnoli, L., Currell, B., Seshagiri, S., Yu, H., Winsor, B., Vidal, M., Gerstein, M. B., Bader, G. D., Volkmer, R., Cesareni, G., Drubin, D. G., Kim, P. M., Sidhu, S. S. and Boone, C. (2009). "Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins." PLoS Biol. **7**: e1000218.

Tonikian, R., Zhang, Y., Sazinsky, S. L., Currell, B., Yeh, J. H., Reva, B., Held, H. A., Appleton, B. A., Evangelista, M., Wu, Y., Xin, X., Chan, A. C., Seshagiri, S., Lasky, L. A., Sander, C., Boone, C., Bader, G. D. and Sidhu, S. S. (2008). "A specificity map for the PDZ domain family." PLoS Biol **6**(9): e239.

Umbhauer, M., Djiane, A., Goisset, C., Penzo-Mendez, A., Riou, J. F., Boucaut, J. C. and Shi, D. L. (2000). "The C-terminal cytoplasmic Lys-thr-X-X-X-Trp motif in frizzled receptors mediates Wnt/beta-catenin signalling." EMBO J **19**(18): 4944-4954.

Wang, C., Ding, C., Meraz, R. F. and Holbrook, S. R. (2006). "PSoL: a positive sample only learning algorithm for finding non-coding RNA genes." Bioinformatics **22**(21): 2590-2596.

Wawrzak, D., Luyten, A., Lambaerts, K. and Zimmermann, P. (2009). "Frizzled-PDZ scaffold interactions in the control of Wnt signaling." Adv Enzyme Regul **49**(1): 98-106.

Wiedemann, U., Boisguerin, P., Leben, R., Leitner, D., Krause, G., Moelling, K., Volkmer-Engert, R. and Oschkinat, H. (2004). "Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides." Journal of molecular biology **343**(3): 703-718.

Wong, H. C., Bourdelas, A., Krauss, A., Lee, H. J., Shao, Y., Wu, D., Mlodzik, M., Shi, D. L. and Zheng, J. (2003). "Direct binding of the PDZ domain of Dishevelled to a conserved internal sequence in the C-terminal region of Frizzled." Mol Cell **12**(5): 1251-1260.

Wu, C., Ma, M. H., Brown, K. R., Geisler, M., Li, L., Tzeng, E., Jia, C. Y., Jurisica, I. and Li, S. S. (2007). "Systematic identification of SH3 domain-mediated human protein-protein interactions by peptide array target screening." Proteomics **7**(11): 1775-1785.

Wunderlich, Z. and Mirny, L. A. (2009). "Using genome-wide measurements for computational prediction of SH2-peptide interactions." Nucleic Acids Res **37**(14): 4629-4641.

Yaffe, M. B., Leparc, G. G., Lai, J., Obata, T., Volinia, S. and Cantley, L. C. (2001). "A motif-based profile scanning approach for genome-wide prediction of signaling pathways." Nat Biotechnol **19**(4): 348-353.

Yao, R., Natsume, Y. and Noda, T. (2004). "MAGI-3 is involved in the regulation of the JNK signaling pathway as a scaffold protein for frizzled and Ltap." Oncogene **23**(36): 6023-6030.

Zhang, J., Dong, J., Gu, H., Yu, S., Zhang, X., Gou, Y., Xu, W., Burd, A., Huang, L., Miyado, K., Huang, Y. and Chan, H. C. (2012). "CD9 is critical for cutaneous wound healing through JNK signaling." J Invest Dermatol **132**(1): 226-236.

Zhang, Y. (2009). "Protein structure prediction: when is it useful?" Curr Opin Struct Biol **19**(2): 145-155.

Zhang, Y., Yeh, S., Appleton, B. A., Held, H. A., Kausalya, P. J., Phua, D. C., Wong, W. L., Lasky, L. A., Wiesmann, C., Hunziker, W. and Sidhu, S. S. (2006). "Convergent and divergent ligand specificity among PDZ domains of the LAP and zonula occludens (ZO) families." J Biol Chem **281**(31): 22299-22311.

Appendix A

Predicting PDZ Protein-Protein Interactions From Sequence

This work was published in *BMC Bioinformatics*, **11**:507: Hui, S., Bader, GD. (2010), Proteome scanning to predict PDZ domain interactions using support vector machines..

Author contributions:

I collected the data, developed and implemented the methods and performed the analyses. Gary D. Bader supervised and advised this project.

A. Detailed summary of proteome scanning results

The following is a summary of the results of proteome scanning in different organisms using the SVM, MDSM, additive model and PWM predictor. Method is the name of the predictor used, Domain is the name of the domain that the proteome is being scanned for, NN Sim is the similarity of the scanning domain to its nearest training neighbour, Num predicted is the number of positive predictions made by the predictor, #TP is the number of positive predictions validated to be positive, #FP is the number of positive predictions that were validated to be negative, #Valid Positives is the number of positive validation interactions, #Valid Negatives is the number of negative validation interactions. Only validation interactions involving genomic peptides (as defined by the Ensembl genome assemblies) were used.

Human

The human proteome was scanned to predict interactions for 13 human PDZ domains with available interactions from PDZBase (Beuming et al. 2005). In total, 41,193 unique transcript tails of length five out of 77,748 transcripts corresponding to 23,675 genes from the human proteome were scanned (defined by Ensembl:GRCh37.56 genome assembly).

Table A-1 Summary of human proteome scanning results for SVM and other predictors.

Method	Domain	NN Sim	Num Predicted	#TP	#FP	#Valid Positives	#Valid Negatives
SVM	DLG1-1	1.0	283	2	0	2	0
SVM	DLG1-2	1.0	389	3	0	3	0
SVM	MPDZ-10	1.0	199	3	0	4	0
SVM	ERBB2IP-1	1.0	83	2	0	2	0
SVM	DLG3-2	1.0	389	1	0	2	0
SVM	LIN7B-1	1.0	422	1	0	2	0
SVM	DLG4-1	0.9375	223	2	0	2	0
SVM	DLG4-2	0.9375	294	2	0	2	0
SVM	PDZK1-1	0.8125	551	1	0	1	0
SVM	MLLT4-1	0.6875	36	1	0	6	0
SVM	MAGI3-1	1.0	1185	0	0	1	0
SVM	MAGI2-2	1.0	694	0	0	1	0
SVM	SNTG1-1	1.0	680	1	0	1	0
Method	**Domain**	**NN Sim**	**Num**	**#TP**	**#FP**	**#Valid**	**#Valid**

			Predicted			Positives	Negatives
MDSM	DLG1-1	1.0	269	2	0	2	0
MDSM	DLG1-2	0.875	269	3	0	3	0
MDSM	MPDZ-10	1.0	2534	1	0	4	0
MDSM	ERBB2IP-1	1.0	825	0	0	2	0
MDSM	DLG3-2	0.875	269	1	0	2	0
MDSM	LIN7B-1	1.0	165	2	0	2	0
MDSM	DLG4-1	0.9375	269	2	0	2	0
MDSM	DLG4-2	0.8125	269	2	0	2	0
MDSM	PDZK1-1	0.9375	11	0	0	1	0
MDSM	MLLT4-1	0.6875	285	1	0	6	0
MDSM	MAGI3-1	0.6875	1070	0	0	1	0
MDSM	MAGI2-2	0.75	1070	0	0	1	0
MDSM	SNTG1-1	0.875	613	1	0	1	0
Method	**Domain**	**NN Sim**	**Num Predicted**	**#TP**	**#FP**	**#Valid Positives**	**#Valid Negatives**
Additive	DLG1-1	1.0	2094	2	0	2	0
Additive	DLG1-2	1.0	2241	3	0	3	0
Additive	MPDZ-10	1.0	52	0	0	4	0
Additive	ERBB2IP-1	1.0	395	0	0	2	0
Additive	DLG3-2	1.0	2241	1	0	2	0
Additive	LIN7B-1	1.0	2734	1	0	2	0
Additive	DLG4-1	0.9375	1960	2	0	2	0
Additive	DLG4-2	0.9375	2041	2	0	2	0
Additive	PDZK1-1	0.8125	0	0	0	1	0
Additive	MLLT4-1	0.6875	93	1	0	6	0
Additive	MAGI3-1	1.0	1846	0	0	1	0
Additive	MAGI2-2	1.0	2406	1	0	1	0
Additive	SNTG1-1	1.0	1723	1	0	1	0
Method	**Domain**	**NN Sim**	**Num Predicted**	**#TP**	**#FP**	**#Valid Positives**	**#Valid Negatives**
PWM	DLG1-1	1.0	412	1	0	2	0
PWM	DLG1-2	1.0	412	3	0	3	0
PWM	MPDZ-10	1.0	412	4	0	4	0
PWM	ERBB2IP-1	1.0	412	2	0	2	0
PWM	DLG3-2	1.0	412	1	0	2	0
PWM	LIN7B-1	1.0	412	2	0	2	0

PWM	DLG4-1	0.9375	412	1	0	2	0
PWM	DLG4-2	0.9375	412	2	0	2	0
PWM	PDZK1-1	0.8125	412	1	0	1	0
PWM	MLLT4-1	0.6875	412	2	0	6	0
PWM	MAGI3-1	1.0	412	0	0	1	0
PWM	MAGI2-2	1.0	412	0	0	1	0
PWM	SNTG1-1	1.0	412	1	0	1	0

Worm

The worm proteome was scanned to predict interactions for 6 worm PDZ domains with positive and negative interactions from protein microarray experiments (Chen et al. 2008). In total, 19,864 unique transcript tails of length five out of 27,533 transcripts corresponding to 20,158 genes in the worm proteome were scanned (defined by genome assembly Ensembl:WS200.56).

Table A-2 Summary of worm proteome scanning results for SVM and other predictors.

Method	Domain	NN Sim	Num Predicted	#TP	#FP	#Valid Positives	#Valid Negatives
SVM	DLG1-1	0.8125	44	1	1	4	18
SVM	DLG1-3	0.9375	87	4	1	7	15
SVM	DSH-1	0.8125	14	0	0	11	4
SVM	LIN7-1	1.0	159	3	1	11	11
SVM	MPZ1-6	0.6875	144	4	0	18	4
SVM	STN2-1	0.8125	256	3	0	8	14
Method	**Domain**	**NN Sim**	**Num Predicted**	**#TP**	**#FP**	**#Valid Positives**	**#Valid Negatives**
MDSM	DLG1-1	0.75	110	1	1	4	18
MDSM	DLG1-3	0.9375	168	4	1	7	15
MDSM	DSH-1	0.8125	2598	3	0	11	4
MDSM	LIN7-1	1.0	61	1	0	11	11
MDSM	MPZ1-6	0.6875	85	0	0	18	4
MDSM	STN2-1	0.8125	200	3	1	8	14
Method	**Domain**	**NN Sim**	**Num Predicted**	**#TP**	**#FP**	**#Valid Positives**	**#Valid Negatives**
Additive	DLG1-1	0.8125	730	2	4	4	18
Additive	DLG1-3	0.9375	864	4	3	7	15
Additive	DSH-1	0.8125	79	0	0	11	4
Additive	LIN7-1	1.0	1177	7	2	11	11

Additive	MPZ1-6	0.6875	713	3	0	18	4
Additive	STN2-1	0.8125	1086	4	2	8	14
Method	**Domain**	**NN Sim**	**Num Predicted**	**#TP**	**#FP**	**#Valid Positives**	**#Valid Negatives**
PWM	DLG1-1	0.8125	199	2	4	4	18
PWM	DLG1-3	0.9375	199	1	2	7	15
PWM	DSH-1	0.8125	199	1	0	11	4
PWM	LIN7-1	1.0	199	3	2	11	11
PWM	MPZ1-6	0.6875	199	3	1	18	4
PWM	STN2-1	0.8125	199	4	2	8	14

Fly

The fly proteome was scanned to predict interactions for 7 fly PDZ domains with positive and negative interactions from protein microarray experiments (Chen et al. 2008). In total, 14,691 unique transcript tails of length five out of 21,309 transcripts corresponding to 20,158 genes were scanned (defined by genome assembly Ensembl:BDGP5.13.56).

Table A-3 Summary of fly proteome scanning results for SVM and other predictors.

Method	**Domain**	**NN Sim**	**Num Predicted**	**#TP**	**#FP**	**#Valid Positives**	**#Valid Negatives**
SVM	MAGI-4	0.8125	92	2	3	2	17
SVM	DLG1-1	0.9375	112	4	0	4	15
SVM	DSH-1	0.9375	49	0	0	3	16
SVM	LAP4-2	0.875	30	3	1	5	14
SVM	LAP4-3	0.75	8	2	0	8	11
SVM	PAR6-1	1.0	0	0	0	1	18
SVM	PATJ-2	0.8125	184	0	0	7	12
Method	**Domain**	**NN Sim**	**Num Predicted**	**#TP**	**#FP**	**#Valid Positives**	**#Valid Negatives**
MDSM	MAGI-4	0.8125	192	0	0	2	17
MDSM	DLG1-1	0.9375	76	2	2	4	15
MDSM	DSH-1	0.9375	1641	2	3	3	16
MDSM	LAP4-2	0.875	8	0	0	5	14
MDSM	LAP4-3	0.75	95	4	1	8	11
MDSM	PAR6-1	1.0	3	0	0	1	18
MDSM	PATJ-2	0.625	5	1	0	7	12
Method	**Domain**	**NN Sim**	**Num**	**#TP**	**#FP**	**#Valid**	**#Valid**

			Predicted			**Positives**	**Negatives**
Additive	MAGI-4	0.8125	843	2	6	2	17
Additive	DLG1-1	0.9375	849	4	3	4	15
Additive	DSH-1	0.9375	98	0	0	3	16
Additive	LAP4-2	0.875	307	4	1	5	14
Additive	LAP4-3	0.75	300	3	0	8	11
Additive	PAR6-1	1.0	18	0	0	1	18
Additive	PATJ-2	0.625	30	0	0	7	12
Method	**Domain**	**NN Sim**	**Num Predicted**	**#TP**	**#FP**	**#Valid Positives**	**#Valid Negatives**
PWM	MAGI-4	0.8125	147	0	3	2	17
PWM	DLG1-1	0.9375	147	4	2	4	15
PWM	DSH-1	0.9375	147	1	3	3	16
PWM	LAP4-2	0.875	147	5	3	5	14
PWM	LAP4-3	0.75	147	4	2	8	11
PWM	PAR6-1	1.0	147	0	0	1	18
PWM	PATJ-2	0.8125	147	0	1	7	12

B. Protein-protein interaction evidence to support PDZ domain peptide predictions

Physical human protein-protein interactions (PPIs) were collected from the iRefIndex database (Razick et al. 2008). Only interactions annotated with UniProt ids from UniProtKB/Swiss-Prot were used (since the corresponding sequences were manually annotated and reviewed). A PPI was counted as corresponding to a domain peptide interaction prediction if the protein containing the domain was found in iRefIndex to interact with the protein containing the peptide. To test the significance of the number of predictions found to be in iRefIndex for a given domain, a Fisher's exact test was performed and asked whether the observed number predictions could be achieved at random. In total, 213 human PDZ domains with PPIs in iRefIndex were analyzed. The SVM predicted interactions for 192 domains with 75 domains having predictions corresponding to at least one iRefIndex interaction. The SVM did not make predictions for the remaining 21 domains.

Table A-4 Identities of human PDZ domains in iRefIndex. The identities of the 75 human PDZ domains whose proteome predictions correspond to at least one protein-protein interaction from iRefIndex are listed. UniProt Domain Name is the name of the domain using the UniProt protein name. UniProt Domain Sequence Positions are the start and end positions of the domain sequence along the UniProt protein sequence. UniProt ID is the identifier of the UniProt protein. Tonikian Domain Name is the name of the domain used in Tonikian et al.

UniProt Domain Name	UniProt Domain Sequence Positions	UniProt ID	Tonikian Domain Name	UniProt Domain Name	UniProt Domain Sequence Positions	UniProt ID	Tonikian Domain Name
ARHGC-1	72-151	Q9NZN5		NHRF2-1	11-90	Q15599	
GIPC1-1	133-213	O14908		PARD3-3	590-680	Q8TEW0	PARD3-3
LIN7B-1	93-175	Q9HAP6		MPDZ-4	565-630	O75970	MPDZ-4
MAGI2-1	17-101	Q86UL8		MPDZ-7	1151-1239	O75970	MPDZ-7
MAGI2-2	426-510	Q86UL8		MPDZ-10	1629-1708	O75970	MPDZ-10
MAGI2-4	778-860	Q86UL8		MPDZ-13	1959-2038	O75970	MPDZ-13
MAGI2-5	920-1010	Q86UL8		NHRF2-2	151-227	Q15599	SLC9A3R2-2
MAGI2-3	605-683	Q86UL8		DLG4-3	313-390	P78352	DLG4-3
MAGI2-6	1147-1229	Q86UL8		MPDZ-2	257-333	O75970	MPDZ-2
MAST2-1	967-1055	Q9Y2H9		SCRIB-4	1110-1194	Q14160	
MPP3-1	137-212	Q13368		ZO2-1	33-120	Q9UDY2	
NHRF1-1	14-94	O14745		DLG1-1	224-307	Q12959	DLG1-1
NHRF1-2	154-234	O14745		DLG1-2	319-402	Q12959	DLG1-2
NHRF3-2	134-215	Q5T2W1		DLG1-3	466-543	Q12959	DLG1-3
NHRF3-4	378-458	Q5T2W1		DLG3-2	226-309	Q92796	DLG3-2
NHRF3-3	243-323	Q5T2W1		MAGI1-2	472-554	Q96QZ7	
NHRF4-1	115-196	Q86UT5		MAGI1-3	634-719	Q96QZ7	MAGI1-2
NHRF4-3	329-412	Q86UT5		MAGI1-4	813-895	Q96QZ7	
PDLI1-1	3-85	O00151		MAGI1-6	1124-1206	Q96QZ7	
PDZ11-1	47-129	Q5EBL8		MAGI3-4	751-831	Q5TCQ9	MAGI3-3
PDZD2-2	334-419	O15018		MAGI3-5	876-963	Q5TCQ9	
PTN3-1	510-582	P26045		MAGI3-5	1046-1128	Q5TCQ9	
RGS12-1	22-98	O14924		PTN13-2	368-1449	Q12923	PTPN13-2
RGS3-1	299-376	P49796		SCRIB-1	728-811	Q14160	SCRIB-1
SHAN1-1	663-757	Q9Y566		SCRIB-2	862-947	Q14160	SCRIB-2
SHAN2-1	247-341	Q9UPX8		SCRIB-3	1004-1093	Q14160	
SNTB1-1	112-195	Q13884		DLG2-2	193-279	Q15700	
SNTB2-1	115-198	Q13425		DLG2-1	98-184	Q15700	

SNTG1-1	57-140	Q9NSN8		DLG2-3	421-501	Q15700	
SNTG2-1	73-156	Q9NY99		LAP2-1	1323-1406	Q96RT1	ERBB2IP-1
SYJ2B-1	13-100	P57105		LRRC7-1	1448-1531	Q96NW7	LRRC7-1
APBA3-2	485-560	O96018		CSKP-1	490-566	O14936	CASK-1
DLG3-1	130-217	Q92796		AFAD-1	1009-1087	P55196	MLLT4-1
DLG3-3	379-465	Q92796		SNTA1-1	87-166	Q13424	SNTA1-1
DLG4-2	160-246	P78352		MAGI3-2	435-517	Q5TCQ9	
DLG4-1	65-151	P78352		MAGI3-3	603-679	Q5TCQ9	
INADL-8	1437-1520	Q8NI35		NHRF3-1	9-86	Q5T2W1	PDZK1-1
MPDZ-8	1350-1433	O75970		NHRF2-1	11-90	Q15599	

Table A-5 Number of predicted interactions that correspond to protein-protein interactions in iRefIndex. UniProt Domain Name is the name of the domain using the UniProt protein name.

UniProt Domain Name	**# iRefIndex PPIs predicted**	**# iRefIndex PPIs**	***p*-value**	**UniProt Domain Name**	**# iRefIndex PPIs predicted**	**# iRefIndex PPIs**	***p*-value**
ARHGC-1	1	14	0.566	NHRF2-1	12	44	2.48e-12
GIPC1-1	4	42	7.76e-06	PARD3-3	1	26	0.0311
LIN7B-1	1	11	0.107	MPDZ-4	2	9	0.0081
MAGI2-1	1	10	0.124	MPDZ-7	1	9	0.027
MAGI2-2	2	10	0.0117	MPDZ-10	4	9	6.53e-08
MAGI2-4	1	10	0.0325	MPDZ-13	1	9	0.0137
MAGI2-5	1	10	0.0344	NHRF2-2	15	44	2.33e-11
MAGI2-3	1	10	0.122	DLG4-3	13	130	1.41e-10
MAGI2-6	1	10	0.0952	MPDZ-2	1	9	0.0591
MAST2-1	2	6	0.0017	SCRIB-4	1	11	0.0534
MPP3-1	1	1	0.000631	ZO2-1	1	11	0.0844
NHRF1-1	15	57	7.45e-15	DLG1-1	13	83	1.98e-14
NHRF1-2	24	57	1.74e-14	DLG1-2	14	83	5.21e-14
NHRF3-2	1	24	0.0763	DLG1-3	10	83	3.09e-09
NHRF3-4	3	24	0.00141	DLG3-2	9	48	6.6e-10
NHRF3-3	8	24	3.08e-06	MAGI1-2	4	24	0.00014
NHRF4-1	1	5	0.0408	MAGI1-3	3	24	0.000176
NHRF4-3	3	5	0.000206	MAGI1-4	2	24	0.000724
PDLI1-1	1	14	0.0748	MAGI1-6	6	24	4.54e-05
PDZ11-1	1	4	0.0307	MAGI3-4	1	12	0.0426

PDZD2-2	1	5	0.123	MAGI3-5	1	12	0.0256
PTN3-1	1	5	0.0861	MAGI3-6	1	12	0.31
RGS12-1	4	19	0.000715	PTN13-2	1	23	0.111
RGS3-1	3	11	0.026	SCRIB-1	1	11	0.0357
SHAN1-1	2	21	0.0913	SCRIB-2	1	11	0.0292
SHAN2-1	1	13	0.364	SCRIB-3	1	11	0.161
SNTB1-1	4	14	9.74e-06	DLG2-2	8	41	4.28e-09
SNTB2-1	3	20	0.00105	DLG2-1	8	41	3.53e-10
SNTG1-1	1	12	0.181	DLG2-3	6	41	1.46e-06
SNTG2-1	1	1	0.0114	LAP2-1	2	33	0.00203
SYJ2B-1	3	5	5.71e-05	LRRC7-1	2	13	0.000731
APBA3-2	1	7	0.00289	CSKP-1	3	53	0.0396
DLG3-1	9	48	3.98e-11	AFAD-1	1	58	0.0495
DLG3-3	7	48	1.95e-07	SNTA1-1	4	28	9.53e-05
DLG4-2	14	130	2.88e-11	MAGI3-2	5	12	1.31e-05
DLG4-1	13	130	7.37e-12	MAGI3-3	1	12	0.0199
INADL-8	1	15	0.0653	NHRF3-1	4	24	0.000272
MPDZ-8	1	9	0.0141				

C. GO biological process term enrichment

GO biological process term enrichment analysis was performed to determine statistically overrepresented annotations in the genes of predicted binders for the PDZ domains used in proteome scanning tests. The hypergeometric test was used to compute a *p*-value to assess GO term enrichment for a set of predicted genes. Since this results in testing the significance of all GO terms in the given set of genes in a single analysis, multiple testing correction was performed using the Benjamini and Hochberg False Discovery Rate (FDR) correction with a significance level of 0.05. The BiNGO (Biological Network Gene Ontology tool) (Maere et al. 2005) software library was used. Only manually annotated GO terms were used.

Table A-6 Enriched GO biological process terms in genes of predicted binders. GO ID is the GO process term identifier, *p*-value is the hypergeometric test statistic corrected for multiple testing, Description is the GO term description. GO terms are ordered by increasing *p*-value.

Only GO terms with $p < 0.05$ are displayed. Domains with no terms satisfying this cutoff are indicated by an asterisk and only the top 10 GO terms are displayed.

DLG1-2

GO ID	*p*-value	Description
6813	2.658E-3	potassium ion transport
30001	2.658E-3	metal ion transport
6811	3.062E-3	ion transport
6812	3.481E-3	cation transport
15672	8.531E-3	monovalent inorganic cation transport

DLG1-2

GO ID	*p*-value	Description
6811	2.774E-4	ion transport
6813	2.167E-3	potassium ion transport
6812	5.264E-3	cation transport
30001	5.264E-3	metal ion transport
6810	1.151E-2	transport
15672	1.685E-2	monovalent inorganic cation transport
51234	2.034E-2	establishment of localization

DLG3-2

GO ID	*p*-value	Description
6811	2.774E-4	ion transport
6813	2.167E-3	potassium ion transport
6812	5.264E-3	cation transport
30001	5.264E-3	metal ion transport
6810	1.151E-2	transport
15672	1.685E-2	monovalent inorganic cation transport
51234	2.034E-2	establishment of localization

DLG4-1

GO ID	*p*-value	Description
6813	2.658E-3	potassium ion transport
30001	2.658E-3	metal ion transport
6811	3.062E-3	ion transport
6812	3.481E-3	cation transport

DLG4-2

GO ID	*p*-value	Description
6811	2.774E-4	ion transport
6813	2.167E-3	potassium ion transport
6812	5.264E-3	cation transport
30001	5.264E-3	metal ion transport
6810	1.151E-2	transport
15672	1.685E-2	monovalent inorganic cation transport
51234	2.034E-2	establishment of localization

ERBB2IP-1 *

GO ID	*p*-value	Description
32581	2.557E-1	ER-dependent peroxisome biogenesis
16557	2.557E-1	peroxisome membrane biogenesis
45046	2.557E-1	protein import into peroxisome membrane
55114	2.557E-1	oxidation reduction
6338	2.557E-1	chromatin remodeling
7155	2.557E-1	cell adhesion
22610	2.557E-1	biological adhesion
51016	2.557E-1	barbed-end actin filament

					capping
			51693	2.557E-1	actin filament capping
			15917	2.557E-1	aminophospholipid transport
LIN7B-1*			**MAGI2-2***		
GO ID	***p*-value**	**Description**	**GO ID**	***p*-value**	**Description**
6811	1.414E-1	ion transport	7389	3.909E-1	pattern specification process
35176	1.414E-1	social behavior	35176	3.909E-1	social behavior
6813	1.414E-1	potassium ion transport	6812	3.909E-1	cation transport
6812	1.414E-1	cation transport	6810	3.909E-1	transport
30001	1.414E-1	metal ion transport	7264	3.909E-1	small GTPase mediated signal transduction
30516	1.414E-1	regulation of axon extension	6813	3.909E-1	potassium ion transport
32927	1.414E-1	positive regulation of activin receptor signaling pathway	51234	3.909E-1	establishment of localization
51705	1.414E-1	behavioral interaction between organisms	51179	3.909E-1	localization
1935	1.414E-1	endothelial cell proliferation	32927	3.909E-1	positive regulation of activin receptor signaling pathway
50808	1.414E-1	synapse organization and biogenesis	51705	3.909E-1	behavioral interaction between organisms
MAGI3-1			**MLLT4-1***		
GO ID	***p*-value**	**Description**	**GO ID**	***p*-value**	**Description**
6813	1.458E-2	potassium ion transport	33081	5.388E-2	regulation of T cell differentiation in the thymus
51234	1.768E-2	establishment of localization	46620	5.388E-2	regulation of organ growth
6810	1.768E-2	transport	303	5.388E-2	response to superoxide
6811	1.768E-2	ion transport	45541	5.388E-2	negative regulation of cholesterol biosynthetic process
51179	1.768E-2	localization	48538	5.388E-2	thymus development
			45939	5.388E-2	negative regulation of steroid metabolic process
			45540	5.388E-2	regulation of cholesterol biosynthetic process
			1890	5.388E-2	placenta development
			305	5.388E-2	response to oxygen radical

			50810	7.339E-2	regulation of steroid biosynthetic process
MPDZ-10*			**PDZK1-1**		
GO ID	***p*-value**	**Description**	**GO ID**	***p*-value**	**Description**
6813	7.25E-2	potassium ion transport	6811	2.389E-4	ion transport
1508	1.822E-1	regulation of action potential	45494	5.702E-3	photoreceptor cell maintenance
30001	1.822E-1	metal ion transport			
15672	1.822E-1	monovalent inorganic cation transport	**SNTG1-1**		
6342	1.822E-1	chromatin silencing	**GO ID**	***p*-value**	**Description**
31507	1.822E-1	heterochromatin formation	6810	2.251E-2	transport
42391	1.822E-1	regulation of membrane potential	51234	2.251E-2	establishment of localization
45814	1.822E-1	negative regulation of gene expression, epigenetic	46942	3.625E-2	carboxylic acid transport
6812	1.822E-1	cation transport	6813	3.625E-2	potassium ion transport
19226	1.822E-1	transmission of nerve impulse	15849	3.625E-2	organic acid transport

Appendix B

Predicting PDZ Protein-Protein Interactions From Structure

This work was published in *BMC Bioinformatics*, **14**:12: Hui, S., Xing, X., Bader, GD. (2013), Predicting PDZ mediated protein-protein interactions from structure.

Author contributions: I collected the data, developed and implemented the methods and performed the analyses. Xiang Xing developed the POW! Website under my supervision. Gary D. Bader supervised and advised this project.

A. Parameters for structure feature generation software

1. Solvent accessibility and hydrogen bonding properties

- Joy website (Mizuguchi et al. 1998): http://tardis.nibio.go.jp/cgi-bin/joy/joy.cgi
- PDB files were uploaded and the resulting LaTEX output file was downloaded and parsed.

2. Solvent accessible area

- Surfv sotfware (Sridharan et al. 1992): http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:SURFace_Algorithms
- The software was run using the parameters: single format flag = on, resolution = 2, probe size = 1.4, last 3 parameters = 1, 0 and 0.

3. Electrostatic and hydrophobicity

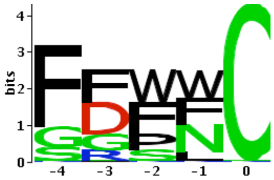
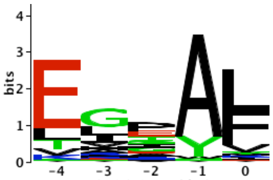
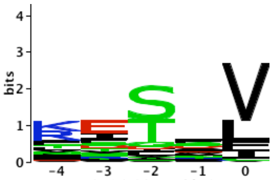
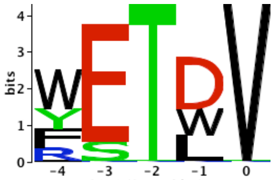
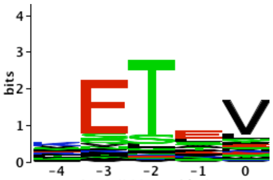
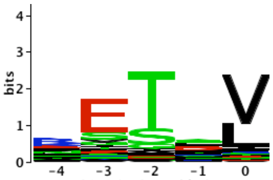
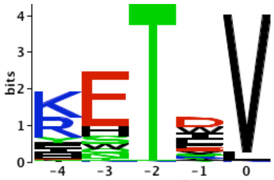
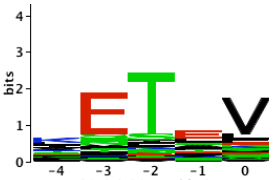
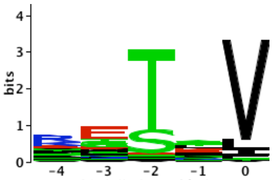
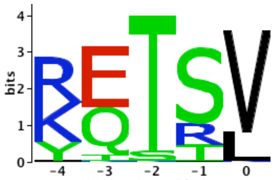
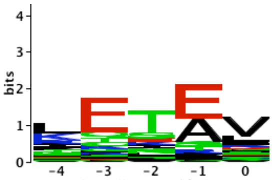
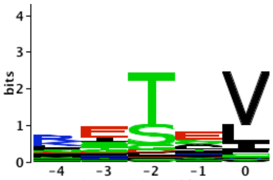
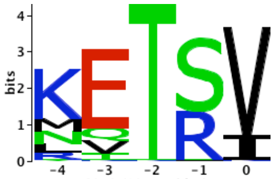
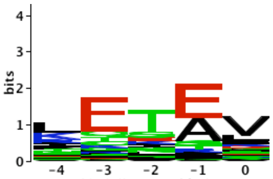
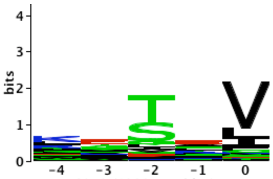
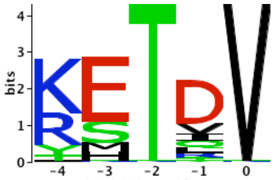
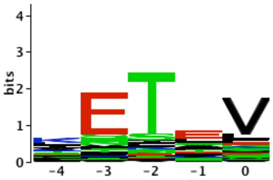
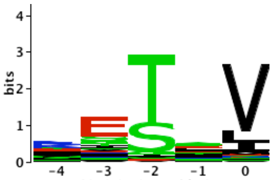
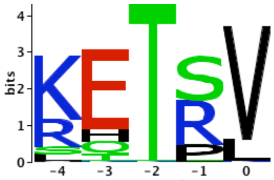
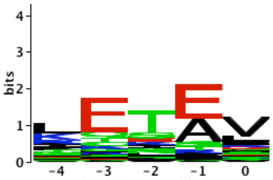
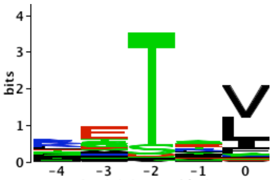
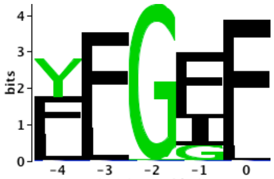
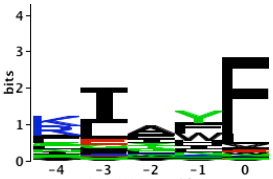
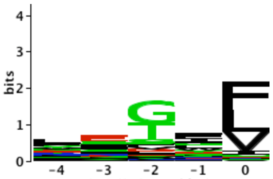
- VASCo sotfware (Steinkellner et al. 2009): http://genome.tugraz.at/VASCo
- The software uses the program DelPhi (Rocchia et al. 2001; Rocchia et al. 2002) to compute the electrostatic potentials and HydroCalc to compute the hydrophobicity values. The default parameters, as distributed in the VASCo package for these programs, were used. Both programs require the calculation of surface points which as performed by the MSMS software (Sanner et al. 1996). For all programs the default probe size of 1.4 was used.

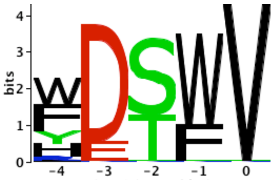
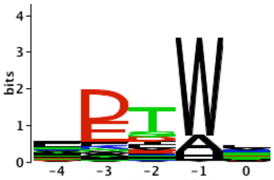
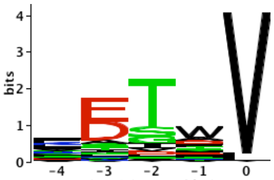
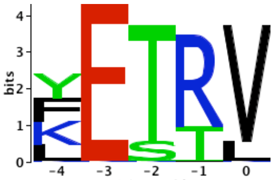
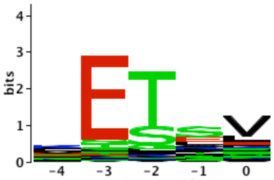
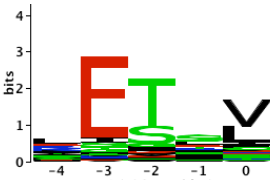
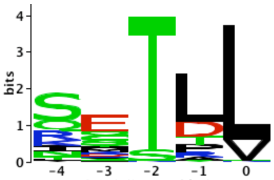
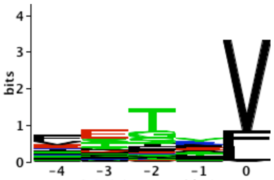
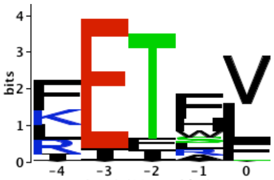
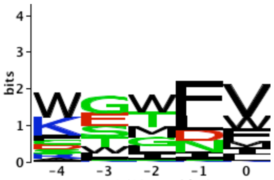
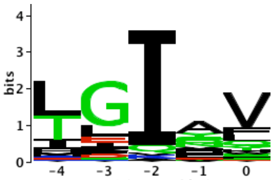
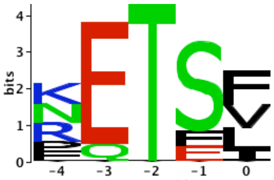
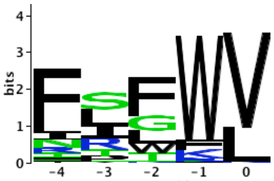
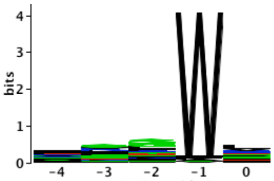
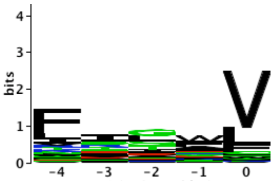
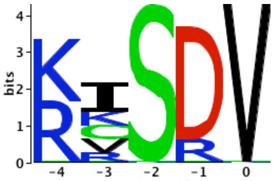
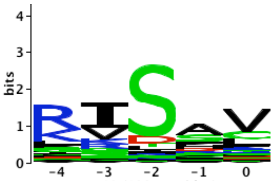
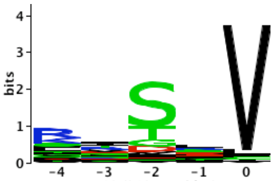
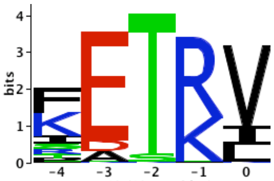
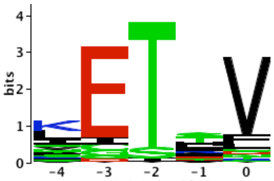
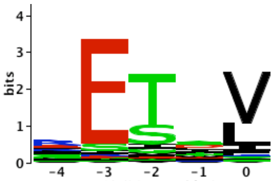
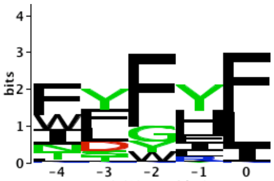
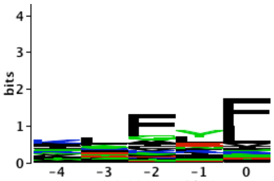
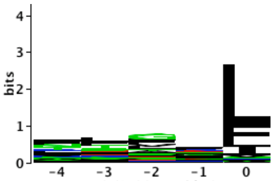
4. 3D binding pocket shape

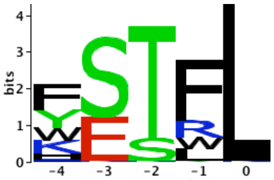
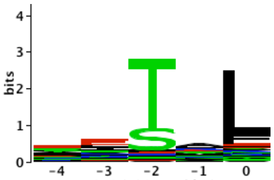
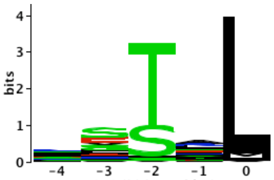
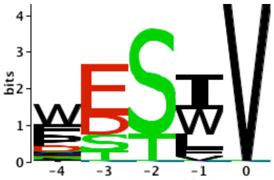
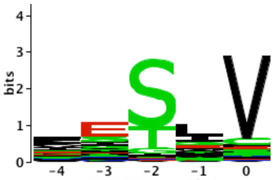
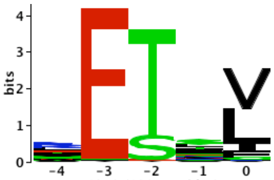
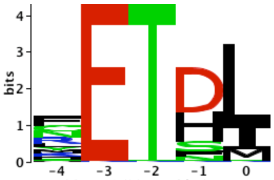
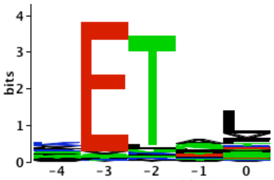
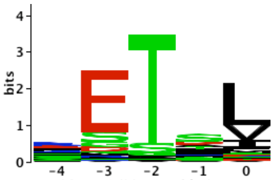
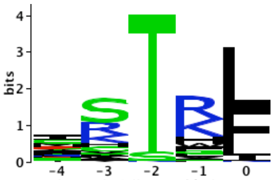
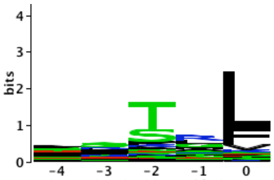
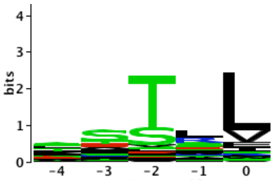
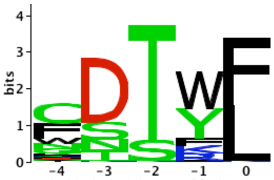
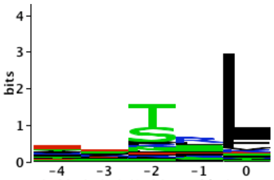
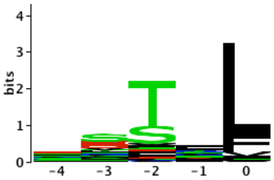
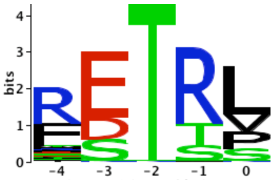
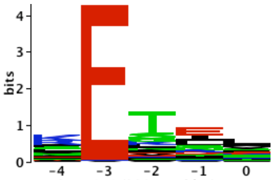
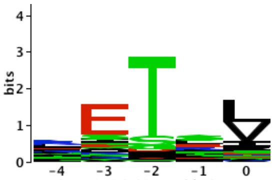
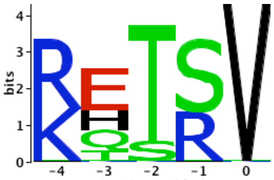
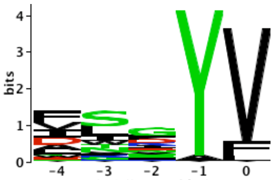
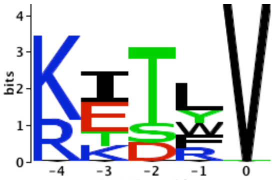
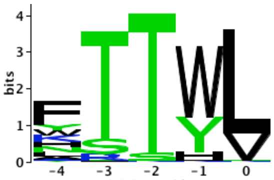
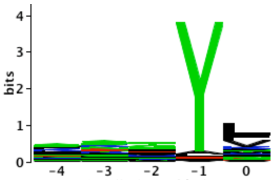
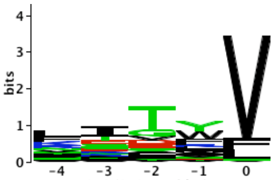
- 3D-Surfer website (La et al. 2009): http://dragon.bio.purdue.edu/3d-surfer
- PDB coordinates corresponding to the binding pocket (defined by 10 core positions) were uploaded and the resulting Zernike descriptors were collected.

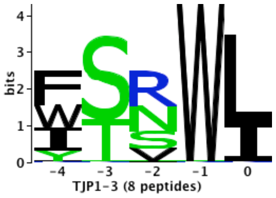
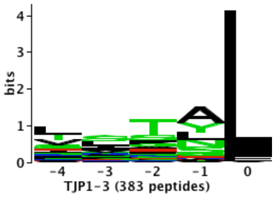
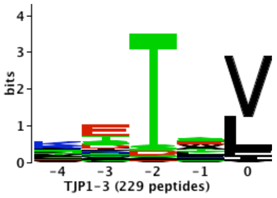
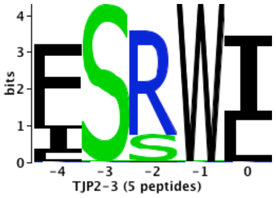
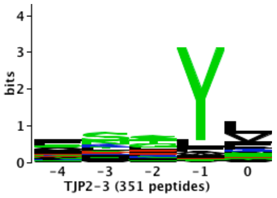
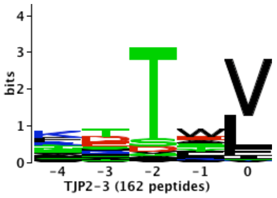
B. Comparison of predicted and experimentally determined genomic binding specificities for human PDZ domains

Figure B-1 Comparison of predicted and experimentally determined binding specificities for human PDZ domains. The predicted and phage display determined binding specificities for 26 domains with four or more genomic peptides were visualized as sequence logos and compared. The binding specificity similarity between two domains was computed using the normalized Euclidean distance between their corresponding position weight matrices (See Section D). Non-genomic phage display peptides were removed from the set of binders for each domain. Based on a previously established protocol, a peptide was considered to be genomic if the last four residues can be found in a proteomic tail, otherwise it was considered to be non genomic. Numbers in bold indicate which similarity (sequence or structure) is higher (i.e. which predicted logo is closer to the experimental logo).

Phage Display	Predicted Logo (Sequence)	Predicted Logo (Structure)	Sim (Sequence)	Sim (Structure)
 APBA3-1 (4 peptides)	 APBA3-1 (51 peptides)	 APBA3-1 (406 peptides)	0.4	**0.5**
 DLG1-1 (6 peptides)	 DLG1-1 (300 peptides)	 DLG1-1 (456 peptides)	**0.698**	0.673
 DLG1-2 (22 peptides)	 DLG1-2 (408 peptides)	 DLG1-2 (279 peptides)	0.781	**0.812**
 DLG1-3 (10 peptides)	 DLG1-3 (374 peptides)	 DLG1-3 (566 peptides)	0.634	**0.711**
 DLG2-3 (10 peptides)	 DLG2-3 (374 peptides)	 DLG2-3 (1399 peptides)	**0.642**	0.642
 DLG3-2 (16 peptides)	 DLG3-2 (408 peptides)	 DLG3-2 (472 peptides)	0.709	**0.709**
 DLG4-3 (9 peptides)	 DLG4-3 (374 peptides)	 DLG4-3 (497 peptides)	0.647	**0.718**
 DVL2-1 (4 peptides)	 DVL2-1 (122 peptides)	 DVL2-1 (852 peptides)	0.561	**0.584**

 ERBB2IP-1 (7 peptides)	 ERBB2IP-1 (85 peptides)	 ERBB2IP-1 (77 peptides)	0.619	**0.692**
 LIN7A-1 (5 peptides)	 LIN7A-1 (359 peptides)	 LIN7A-1 (426 peptides)	**0.701**	0.696
 MAGI1-2 (27 peptides)	 MAGI1-2 (475 peptides)	 MAGI1-2 (11 peptides)	0.612	**0.627**
 MLLT4-1 (19 peptides)	 MLLT4-1 (47 peptides)	 MLLT4-1 (8 peptides)	**0.604**	0.523
 MPDZ-1 (14 peptides)	 MPDZ-1 (410 peptides)	 MPDZ-1 (446 peptides)	0.69	**0.746**
 MPDZ-10 (8 peptides)	 MPDZ-10 (199 peptides)	 MPDZ-10 (313 peptides)	0.678	**0.691**
 MPDZ-13 (13 peptides)	 MPDZ-13 (70 peptides)	 MPDZ-13 (205 peptides)	**0.777**	0.728
 MPDZ-3 (12 peptides)	 MPDZ-3 (1091 peptides)	 MPDZ-3 (1739 peptides)	**0.781**	0.713

 PDZK1-1 (13 peptides)	 PDZK1-1 (781 peptides)	 PDZK1-1 (326 peptides)	0.693	**0.757**
 PTPN13-2 (14 peptides)	 PTPN13-2 (222 peptides)	 PTPN13-2 (108 peptides)	**0.793**	0.562
 SCRIB-1 (23 peptides)	 SCRIB-1 (139 peptides)	 SCRIB-1 (234 peptides)	**0.796**	0.771
 SHANK3-1 (35 peptides)	 SHANK3-1 (1523 peptides)	 SHANK3-1 (864 peptides)	**0.771**	0.765
 SLC9A3R2-2 (14 peptides)	 SLC9A3R2-2 (1496 peptides)	 SLC9A3R2-2 (811 peptides)	0.621	**0.675**
 SNTA1-1 (7 peptides)	 SNTA1-1 (398 peptides)	 SNTA1-1 (568 peptides)	0.629	**0.741**
 TIAM2-1 (5 peptides)	 TIAM2-1 (27 peptides)	 TIAM2-1 (6 peptides)	0.47	**0.736**
 TJP1-1 (13 peptides)	 TJP1-1 (348 peptides)	 TJP1-1 (171 peptides)	0.516	**0.569**

			**0.537**	0.454
			**0.421**	0.378

C. Structure information for PDZ domains used for training and testing

Table B-1 Structure information for PDZ domains used for predictor training. In total, 83 PDZ domains were used for training. Domain structures were obtained from the PDB or homology modelled through the Protein Model Portal. For NMR structures, only the first model was used. All homology models were generated by SWISS-MODEL and have greater than 50% sequence similarity to their template structure (average 90%). Model quality is estimated using template sequence ID (percentage of residues between target and template sequences that are identical) and QMEAN score (a scoring function that measures multiple geometrical aspects of protein structure, ranging from 0 to 1 with higher values indicating more reliable models).

Domain Name	Organism	Experiment	PDB	Template PDB	Template Seq ID	QMEAN Score
CASK-1	Human	XRAY	1KWA			
DLG1-1	Human	SWISS-MODEL		1ZOK A	0.99	0.603
DLG1-2	Human	XRAY	2G2L			
DLG1-3	Human	SWISS-MODEL		1PDR A	1.00	0.938
DLG2-3	Human	XRAY	2HE2			
DLG3-2	Human	XRAY	2FE5			
DLG4-3	Human	XRAY	1TP3			
DVL2-1	Human	XRAY	2REY			
ERBB2IP-1	Human	XRAY	1MFL			
INADL-2	Human	NMR	2DLU			
LRRC7-1	Human	SWISS-MODEL		2H3L B	0.75	0.856
MAGI1-4	Human	SWISS-MODEL		1UEW A	0.68	0.949
MAGI3-3	Human	SWISS-MODEL		1UEW A	0.63	0.924

MPDZ-1	Human	SWISS-MODEL		2O2T A	0.98	0.855
MPDZ-3	Human	SWISS-MODEL		2IWN A	0.96	0.955
MPDZ-10	Human	XRAY	2OPG			
MPDZ-12	Human	SWISS-MODEL		2IWP B	1.00	0.887
MPP6-1	Human	SWISS-MODEL		2E7K A	0.75	0.771
PDLIM4-1	Human	XRAY	2V1W			
PDZK1-1	Human	SWISS-MODEL		2EDZ A	0.89	0.813
PSCDBP-1	Human	XRAY	2Z17			
SCRIB-1	Human	XRAY	2W4F			
SCRIB-2	Human	XRAY	1WHA			
SLC9A3R2-2	Human	XRAY	2HE4			
SNTA1-1	Human	SWISS-MODEL		1QAV A	0.99	0.804
A1-SYNTROPHIN-1	Mouse	NMR	1Z86			
B1-SYNTROPHIN-1	Mouse	SWISS-MODEL		1QAV A	0.82	0.742
CHAPSYN-110-2	Mouse	SWISS-MODEL		1BYG A	0.98	0.968
CHAPSYN-110-3	Mouse	SWISS-MODEL		2HE2 B	1.00	0.998
CIPP-3	Mouse	SWISS-MODEL		2DMZ A	0.93	0.749
CIPP-5	Mouse	SWISS-MODEL		2D92 A	0.91	0.835
CIPP-8	Mouse	SWISS-MODEL		2DM8 A	0.97	0.689
CIPP-9	Mouse	SWISS-MODEL		2QG1 A	0.70	0.990
CIPP-10	Mouse	SWISS-MODEL		2IWO A	0.68	0.802
DVL1-1	Mouse	NMR	1MC7			
DVL3-1	Mouse	SWISS-MODEL		1L6O A	0.97	0.911
ERBIN-1	Mouse	SWISS-MODEL		2H3L A	0.98	0.836
GRIP1-6	Mouse	SWISS-MODEL		1N7F A	1.00	0.784
HARMONIN2-1	Mouse	SWISS-MODEL		1X5N A	0.95	0.696
LARG-1	Mouse	SWISS-MODEL		2OMJ A	0.99	0.842
LIN-7C-1	Mouse	SWISS-MODEL		2DKR A	0.93	0.799
LRRC7-1	Mouse	SWISS-MODEL		2H3L B	0.75	0.969
MAGI-1-6	Mouse	SWISS-MODEL		2R4H C	0.99	0.738
MAGI-2-2	Mouse	SWISS-MODEL		1UEQ A	1.00	0.904
MAGI-2-5	Mouse	SWISS-MODEL		1UEW A	0.99	0.755
MAGI-2-6	Mouse	SWISS-MODEL		1WFV A	1.00	0.835
MAGI-3-1	Mouse	SWISS-MODEL		1UEQ A	0.77	0.818
MAGI-3-2	Mouse	SWISS-MODEL		1UJV A	0.61	0.728
MAGI-3-5	Mouse	SWISS-MODEL		2R4H C	0.66	0.748
MALS2-1	Mouse	SWISS-MODEL		2DKR A	0.99	0.808

MPP7-1	Mouse	SWISS-MODEL		3O46 A	0.93	0.895
MUPP1-1	Mouse	SWISS-MODEL		2O2T A	0.95	0.849
MUPP1-11	Mouse	SWISS-MODEL		2QG1 A	0.95	1.000
MUPP1-10	Mouse	SWISS-MODEL		2OPG B	0.99	0.917
MUPP1-12	Mouse	SWISS-MODEL		2IWP B	0.92	0.871
MUPP1-13	Mouse	SWISS-MODEL		2FNE B	0.96	0.892
MUPP1-5	Mouse	SWISS-MODEL		2D92 A	0.62	0.859
NHERF2-2	Mouse	SWISS-MODEL		2HE4 A	0.93	0.821
NNOS-1	Mouse	SWISS-MODEL		1QAV B	1.00	0.825
OMP25-1	Mouse	SWISS-MODEL		1JIK A	0.96	0.943
PAR6B-1	Mouse	SWISS-MODEL		1NF3 D	1.00	0.667
PDZK11-1	Mouse	SWISS-MODEL		1WI2 A	1.00	0.781
PDZK1-1	Mouse	NMR	2EDZ			
PDZK1-3	Mouse	SWISS-MODEL		2D90 A	1.00	0.867
PSD95-2	Mouse	SWISS-MODEL		3GSL A	1.00	1.000
PSD95-3	Mouse	SWISS-MODEL		1TP5 A	1.00	0.894
PTP-BL-2	Mouse	NMR	1VJ6			
SAP102-2	Mouse	SWISS-MODEL		1FE5 A	1.00	0.992
SAP102-3	Mouse	SWISS-MODEL		3JXT B	1.00	0.940
SAP97-1	Mouse	SWISS-MODEL		1ZOK A	1.00	0.664
SAP97-2	Mouse	SWISS-MODEL		2I0L B	1.00	1.000
SAP97-3	Mouse	SWISS-MODEL		1PDR A	1.00	0.936
SCRB1-1	Mouse	SWISS-MODEL		2W4F A	0.96	0.825
SCRB1-2	Mouse	SWISS-MODEL		1WHA A	0.94	0.855
SCRB1-3	Mouse	SWISS-MODEL		3GSL A	0.54	0.520
SEMCAP3-1	Mouse	SWISS-MODEL		1UHP A	0.96	0.816
SHANK1-1	Mouse	SWISS-MODEL		1Q3O A	1.00	0.700
SHANK3-1	Mouse	SWISS-MODEL		1Q3O A	0.86	0.636
SHROOM-1	Mouse	SWISS-MODEL		2EDP A	0.53	0.731
SLIM-1	Mouse	SWISS-MODEL		1VB7 A	1.00	0.777
ZO-1-1	Mouse	SWISS-MODEL		2H2C A	1.00	0.960
ZO-2-1	Mouse	SWISS-MODEL		2CSJ A	1.00	0.879
ZO-3-1	Mouse	SWISS-MODEL		2CSJ A	0.51	0.879
HTRA2-1	Human	XRAY	2PZD			
MLLT4-1	Human	NMR	1XZ9			
APBA3-1	Human	SWISS-MODEL		2YT7 A	1.00	0.806
SHANK3-1	Human	SWISS-MODEL		1Q3O A	0.86	0.719

PDZ-RGS3-1	Mouse	NMR	1WHD

Table B-2 Structure information for PDZ domains used for blind testing. Blind testing was performed using interaction data from mouse, worm and fly protein microarray experiments. In total, 13 mouse, 7 worm and 6 fly PDZ domains were used. Homology models were generated by SWISS-MODEL. All models have at least 40% sequence identity to their template structures. An NMR structure was available for one fly domain and the first model was used. The average template sequence similarity was 0.92, 0.61 and 0.61 for mouse, worm and fly domains, respectively. One mouse domain (CHAPSYN-110-1) was removed from the test set because its performance was consistently poor for both predictors. Model quality is estimated using template sequence ID (percentage of residues between target and template sequences that are identical) and QMEAN score (a scoring function that measures multiple geometrical aspects of protein structure, ranging from 0 to 1 with higher values indicating more reliable models).

	Domain				**Template**	**Template**	**QMEAN**
	Name	**Organism**	**Experiment**	**PDB**	**PDB**	**Seq ID**	**Score**
1	CIPP-7	Mouse	SWISS-MODEL		2DAZ A	0.91	0.888
2	GOPC1-1	Mouse	SWISS-MODEL		2DCD2 A	1.00	0.833
3	GRIP1-4	Mouse	SWISS-MODEL		1P1D A	0.99	0.549
4	GRIP1-5	Mouse	SWISS-MODEL		1P1D A	0.99	0.78
5	IL-16-3	Mouse	SWISS-MODEL		1X6D A	0.88	0.904
6	MAGI2-3	Mouse	SWISS-MODEL		1UJV A	0.99	0.757
7	MAGI2-4	Mouse	SWISS-MODEL		1UEP A	1.00	0.729
8	MUPP1-2	Mouse	SWISS-MODEL		2DLU A	0.69	
9	NHERF1-2	Mouse	SWISS-MODEL		2OZF A	0.94	0.884
10	PAR-3B-3	Mouse	SWISS-MODEL		1WG6 A	1.00	0.572
11	PAR-6G-1	Mouse	SWISS-MODEL		1NF3 A	0.89	0.764
12	PDZK1-4	Mouse	SWISS-MODEL		2VSP A	0.75	0.757
13	SCRIB-4	Mouse	SWISS-MODEL		1UJU A	0.95	0.823
1	DLG1-1	Worm	SWISS-MODEL		3GSL A	0.51	0.886
2	DLG1-3	Worm	SWISS-MODEL		3JXT A	0.62	0.999
3	DSH-1	Worm	SWISS-MODEL		2F0A A	0.67	0.46
4	LIN7-1	Worm	SWISS-MODEL		2DKR A	0.83	0.684
5	MPZ1-6	Worm	SWISS-MODEL		2IWQ A	0.52	0.869
6	NAB-1-1	Worm	SWISS-MODEL		2FN5 A	0.70	0.739

7	STN-2-1	Worm	SWISS-MODEL		2DKR A	0.40	0.597
1	DLG1-1	Fly	SWISS-MODEL		1ZOK A	0.70	0.579
2	DSH-1	Fly	SWISS-MODEL		3CBZ A	0.81	0.702
3	LAP4-2	Fly	SWISS-MODEL		1WHA A	0.51	0.865
4	MAGI-4	Fly	SWISS-MODEL		1UEW A	0.50	0.729
5	PATJ-2	Fly	SWISS-MODEL		2IWN A	0.53	0.842
6	PAR6-1	Fly	NMR	1RY4 A			

Table B-3 Structure information for PDZ domains used for proteome scanning in human. Proteome scanning was performed for 218 human PDZ domains, which have known interactions in iRefIndex. In total, 61 X-ray and nine NMR structures (only the first models used) were obtained from the PDB and 148 homology models were created (template sequence similarity minimum 22%, average 72%). Model quality is estimated using template sequence ID (percentage of residues between target and template sequences that are identical) and QMEAN score (a scoring function that measures multiple geometrical aspects of protein structure, ranging from 0 to 1 with higher values indicating more reliable models).

UniProt	UniProt Id	Start Index	End Index	Experiment	PDB ID	Template PDB ID	Template Seq ID	QMEAN Score
AHNAK2-1	Q8IVF2	122	195	SWISS-MODEL		3SHW A	0.39	0.603
APBA1-1	Q02410	660	741	SWISS-MODEL		1U3B A	1.00	0.810
APBA1-2	Q02410	755	821	SWISS-MODEL		1U3B A	1.00	0.581
APBA2-1	Q99767	571	653	SWISS-MODEL		1U3B A	0.85	0.792
APBA2-2	Q99767	666	733	SWISS-MODEL		1U3B A	0.93	0.514
APBA3-1	O96018	396	479	SWISS-MODEL		2YT7 A	1.00	0.806
APBA3-2	O96018	491	559	XRAY	2YT8			
ARHGAP21-1	Q5T5U3	49	159	NMR	2YUY			
ARHGAP23-1	Q9P227	52	156	SWISS-MODEL		2YUY A	0.81	0.609
ARHGEF11-1	O15085	51	118	SWISS-MODEL		2DLS A	1.00	0.921
ARHGEF12-1	Q9NZN5	77	147	SWISS-MODEL		2OMJ A	1.00	0.847
CAR14-1	Q9BXL6	570	659	SWISS-MODEL		1Z87 A	0.30	0.442
CASK-1	O14936	489	573	XRAY	1KWA			
CNKR1-1	Q969H4	198	286	SWISS-MODEL		2DKR A	0.22	0.458
CNKSR2-1	Q8WXI2	225	293	SWISS-MODEL		2E7K A	0.29	0.643
CNKSR3-1	Q6P9H4	219	288	SWISS-MODEL		2E7K A	0.29	0.688
CYTIP-1	O60759	76	163	XRAY	2Z17			
DEPTOR-1	Q8TB45	330	408	SWISS-MODEL		2D90 A	0.31	0.879
DLG1-1	Q12959	221	312	SWISS-MODEL		1ZOK A	0.99	0.603
DLG1-2	Q12959	316	406	XRAY	2G2L			

DLG1-3	Q12959	463	544	SWISS-MODEL		1PDR A	1.00	0.938
DLG2-1	Q15700	100	183	SWISS-MODEL		2WL7 A	0.98	0.997
DLG2-2	Q15700	196	278	SWISS-MODEL		2BYG A	1.00	0.953
DLG2-3	Q15700	418	518	XRAY	2HE2			
DLG3-1	Q92796	134	216	XRAY	2I1N			
DLG3-2	Q92796	223	315	XRAY	2FE5			
DLG3-3	Q92796	389	464	XRAY	1UM7			
DLG4-1	P78352	67	151	SWISS-MODEL		3GSL B	1.00	0.966
DLG4-2	P78352	163	245	SWISS-MODEL		3GSL A	1.00	0.991
DLG4-3	P78352	301	416	XRAY	1TP3			
DLG5-3	Q8TDM6	1353	1426	XRAY	1UIT			
DLG5-4	Q8TDM6	1509	1580	SWISS-MODEL		2QG1 A	0.32	0.594
DVL1-1	O14640	254	337	SWISS-MODEL		1MC7 A	1.00	0.624
DVL1L1-1	P54792	260	340	SWISS-MODEL		2KAW A	0.99	0.629
DVL2-1	O14641	270	353	XRAY	2REY			
DVL3-1	Q92997	252	335	SWISS-MODEL		1L6O A	0.96	0.977
ERBB2IP-1	Q96RT1	1321	1413	XRAY	1MFL			
FRMPD1-1	Q5SYB0	67	133	SWISS-MODEL		2FNE C	0.33	0.791
FRMPD2-2	Q68DX3	950	1036	SWISS-MODEL		1VJ6 A	0.61	0.910
FRMPD2-3	Q68DX3	1080	1168	SWISS-MODEL		1B8Q A	0.38	0.518
FRMPD3-1	Q5JV73	62	132	SWISS-MODEL		1WHD A	0.34	0.819
FRMPD4-1	Q14CM0	79	156	SWISS-MODEL		2EDV A	0.36	0.837
GIPC1-1	O14908	136	211	SWISS-MODEL		3GGE A	0.65	0.852
GIPC2-1	Q8TF65	125	200	XRAY	3GGE			
GIPC3-1	Q8TF64	120	195	SWISS-MODEL		3GGE A	0.63	0.698
GOPC-1	Q9HD26	293	369	XRAY	2DC2			
GORASP2-1	Q9H8Y8	5	76	SWISS-MODEL		3RLE A	1.00	0.857
GRD2I-1	A4D2P6	10	85	SWISS-MODEL		2EDV A	0.36	0.509
GRD2I-2	A4D2P6	279	357	SWISS-MODEL		2KV8 A	0.36	0.686
GRIP1-1	Q9Y3R0	56	135	SWISS-MODEL		2QT5 A	1.00	0.866
GRIP1-2	Q9Y3R0	154	237	XRAY	2JIL			
GRIP1-3	Q9Y3R0	261	335	SWISS-MODEL		1V62 A	0.64	0.804
GRIP1-4	Q9Y3R0	472	562	SWISS-MODEL		1P1D A	0.99	0.624
GRIP1-5	Q9Y3R0	577	657	SWISS-MODEL		1P1D A	0.98	0.804
GRIP1-6	Q9Y3R0	676	753	SWISS-MODEL		1N7E A	1.00	0.953
GRIP1-7	Q9Y3R0	1008	1084	SWISS-MODEL		1M5Z A	0.99	0.902
GRIP2-1	Q9C0E4	52	130	SWISS-MODEL		2QT5 A	0.79	0.889
GRIP2-2	Q9C0E4	151	227	SWISS-MODEL		2QT5 A	0.70	0.962
GRIP2-3	Q9C0E4	254	331	XRAY	1V62			
GRIP2-4	Q9C0E4	466	543	SWISS-MODEL		1X5R A	1.00	0.766
GRIP2-5	Q9C0E4	561	640	SWISS-MODEL		1P1D A	0.88	0.704
GRIP2-6	Q9C0E4	659	736	SWISS-MODEL		1N7E A	0.90	1.000

GRIP2-7	Q9C0E4	944	1021	SWISS-MODEL		1M5Z A	0.70	
HTRA1-1	Q92743	370	468	NMR	2YTW			
HTRA2-1	O43464	359	442	SWISS-MODEL		2PZD B	1.00	0.838
HTRA3-1	P83110	350	441	XRAY	2P3W			
IL16-1	Q14005	221	301	SWISS-MODEL		2ENO A	0.49	0.778
IL16-3	Q14005	1117	1189	SWISS-MODEL		1X6D A	1.00	0.866
INADL-1	Q8NI35	138	219	SWISS-MODEL		2DB5 A	1.00	0.792
INADL-2	Q8NI35	231	342	NMR	2DLU			
INADL-3	Q8NI35	369	451	SWISS-MODEL		2DMZ A	0.99	0.845
INADL-5	Q8NI35	692	769	XRAY	2D92			
INADL-6	Q8NI35	1073	1155	XRAY	2EHR			
INADL-7	Q8NI35	1243	1319	XRAY	2DAZ			
INADL-8	Q8NI35	1441	1518	XRAY	2DM8			
INADL-9	Q8NI35	1537	1613	SWISS-MODEL		2QG1 A	0.75	1.000
INADL-10	Q8NI35	1684	1760	SWISS-MODEL		2IWP B	0.66	0.973
LDB3-1	O75112	11	83	XRAY	1RGW			
LIMK1-1	P53667	168	256	SWISS-MODEL		2YUB A	0.40	0.583
LIMK2-1	P53671	155	238	SWISS-MODEL		2YUB A	0.95	0.622
LIN7A-1	O14910	111	188	SWISS-MODEL		2DKR A	0.92	0.793
LIN7B-1	Q9HAP6	96	172	XRAY	2DKR			
LIN7C-1	Q9NUP9	96	173	SWISS-MODEL		2DKR A	0.94	0.827
LMO7-1	Q8WWI1	1042	1129	XRAY	2EAQ			
LRRC7-1	Q96NW7	1454	1535	SWISS-MODEL		2H3L B	0.75	0.856
MAGI1-2	Q96QZ7	478	544	XRAY	2KPK			
MAGI1-3	Q96QZ7	646	722	SWISS-MODEL		3BPU A	0.95	0.756
MAGI1-4	Q96QZ7	849	923	SWISS-MODEL		2Q9V A	0.97	0.949
MAGI1-5	Q96QZ7	1001	1092	SWISS-MODEL		1UEW A	0.67	0.752
MAGI1-6	Q96QZ7	1155	1230	SWISS-MODEL		2R4H A	0.96	0.901
MAGI2-1	Q86UL8	26	98	SWISS-MODEL		2HE4 A	0.35	0.535
MAGI2-2	Q86UL8	429	497	XRAY	1UEQ			
MAGI2-3	Q86UL8	605	684	SWISS-MODEL		1UJV A	1.00	0.824
MAGI2-4	Q86UL8	783	860	XRAY	1UEW			
MAGI2-5	Q86UL8	923	1008	SWISS-MODEL		1UEW A	1.00	0.708
MAGI2-6	Q86UL8	1150	1227	XRAY	1WFV			
MAGI3-2	Q5TCQ9	440	504	SWISS-MODEL		1UEQ A	0.74	0.725
MAGI3-3	Q5TCQ9	603	680	SWISS-MODEL		3SOE A	1.00	0.924
MAGI3-4	Q5TCQ9	757	833	SWISS-MODEL		1UEP A	0.63	0.750
MAGI3-5	Q5TCQ9	879	961	SWISS-MODEL		1UEW A	0.63	0.754
MAGI3-6	Q5TCQ9	1049	1126	SWISS-MODEL		1WFV A	0.65	1.000
MAST1-1	Q9Y2H9	974	1052	XRAY	3PS4			
MAST2-1	Q6P0Q8	1104	1193	SWISS-MODEL		2KQF A	1.00	0.710
MAST3-1	O60307	950	1039	SWISS-MODEL		3KHF B	1.00	0.875

MLLT4-1	P55196	1014	1091	SWISS-MODEL		1XZ9 A	1.00	0.530
MPDZ-1	O75970	136	228	SWISS-MODEL		2O2T A	0.98	0.855
MPDZ-2	O75970	258	334	SWISS-MODEL		2DLU A	0.68	0.758
MPDZ-3	O75970	373	464	SWISS-MODEL		2IWN A	0.95	0.955
MPDZ-4	O75970	562	630	SWISS-MODEL		2DAZ A	0.38	0.850
MPDZ-5	O75970	703	784	SWISS-MODEL		2D92 A	0.63	0.836
MPDZ-6	O75970	1011	1077	SWISS-MODEL		3B76 B	0.39	0.601
MPDZ-7	O75970	1151	1240	XRAY	2IWQ			
MPDZ-8	O75970	1353	1429	SWISS-MODEL		2DAZ A	0.76	0.785
MPDZ-9	O75970	1487	1562	SWISS-MODEL		2DKR A	0.39	0.732
MPDZ-10	O75970	1623	1717	XRAY	2OPG			
MPDZ-11	O75970	1728	1805	XRAY	2QG1			
MPDZ-12	O75970	1862	1945	SWISS-MODEL		2IWP B	1.00	0.887
MPDZ-13	O75970	1990	2070	XRAY	2FNE			
MPP1-1	Q00013	74	150	XRAY	2EV8			
MPP3-1	Q13368	141	216	SWISS-MODEL		3O46 A	0.80	0.853
MPP4-1	Q96JB8	161	233	SWISS-MODEL		3O46 A	0.59	0.912
MPP5-1	Q8N3R9	260	333	XRAY	1VA8			
MPP6-1	Q9NZW5	129	207	SWISS-MODEL		2E7K A	0.75	0.771
MYO18A-1	Q92614	225	310	SWISS-MODEL		1G9O A	0.29	0.631
NOS1-1	P29475	20	96	XRAY	1QAV			
PARD3-1	Q8TEW0	282	349	SWISS-MODEL		2DB5 A	0.29	0.587
PARD3-3	Q8TEW0	597	667	SWISS-MODEL		2K1Z A	0.97	0.630
PARD3B-1	Q8TEW8	211	292	SWISS-MODEL		2O2T A	0.30	0.716
PARD3B-2	Q8TEW8	391	471	SWISS-MODEL		2KOJ A	0.61	0.842
PARD3B-3	Q8TEW8	507	593	SWISS-MODEL		1WG6 A	0.99	0.533
PARD6A-1	Q9NPB6	160	248	SWISS-MODEL		1RZX A	0.84	0.779
PARD6B-1	Q9BYG5	162	240	XRAY	1NF3			
PARD6G-1	Q9BYG4	163	241	SWISS-MODEL		1NF3 D	0.90	0.682
PDLIM1-1	O00151	7	83	XRAY	2PKT			
PDLIM2-1	Q96JY6	1	85	XRAY	2PA1			
PDLIM3-1	Q53GG5	11	85	SWISS-MODEL		1V5L A	0.96	0.706
PDLIM4-1	P50479	1	90	XRAY	2V1W			
PDLIM5-1	Q96HC4	12	84	XRAY	2UZC			
PDLIM7-1	Q9NR12	10	82	XRAY	2Q3G			
PDZD11-1	Q5EBL8	50	126	SWISS-MODEL		1WI2 A	1.00	0.835
PDZD2-2	O15018	342	417	SWISS-MODEL		2DM8 A	0.47	0.812
PDZD2-3	O15018	592	665	SWISS-MODEL		2ENO A	0.44	0.968
PDZD2-4	O15018	728	814	SWISS-MODEL		2JRE A	0.35	0.533
PDZD2-5	O15018	2626	2694	SWISS-MODEL		1X6D A	0.52	0.846
PDZD3-1	Q86UT5	121	194	SWISS-MODEL		1G9O A	0.41	0.821
PDZD3-2	Q86UT5	231	298	SWISS-MODEL		2OCS A	0.39	0.866

PDZD3-3	Q86UT5	333	410	SWISS-MODEL		2V9O E	1.00	1.000
PDZD4-1	Q76G19	130	215	SWISS-MODEL		1WH1 A	0.75	0.710
PDZD7-1	Q9H5P4	86	169	NMR	2EEH			
PDZK1-1	Q5T2W1	1	108	SWISS-MODEL		2EDZ A	0.89	0.813
PDZK1-2	Q5T2W1	142	210	NMR	2EEI			
PDZK1-3	Q5T2W1	247	321	SWISS-MODEL		2D90 A	0.88	0.855
PDZK1-4	Q5T2W1	384	456	SWISS-MODEL		2VSP D	1.00	0.932
PDZRN3-1	Q9UPQ7	257	340	NMR	1UHP			
PDZRN3-2	Q9UPQ7	429	505	SWISS-MODEL		1WH1 A	1.00	0.730
PDZRN4-1	Q6ZMN7	232	315	SWISS-MODEL		1UHP A	0.70	0.839
PDZRN4-2	Q6ZMN7	412	488	SWISS-MODEL		1WH1 A	0.79	0.782
PICK1-1	Q9NRD5	25	101	XRAY	2GZV			
PPP1R9A-1	Q9ULJ8	509	590	SWISS-MODEL		3HVQ C	1.00	0.933
PPP1R9B-1	Q96SB3	498	578	XRAY	3EGG			
PTPN13-1	Q12923	1096	1176	SWISS-MODEL		2DKR A	0.49	0.788
PTPN13-2	Q12923	1371	1445	SWISS-MODEL		1Q7X A	0.99	0.355
PTPN13-3	Q12923	1504	1584	SWISS-MODEL		2OGP A	0.39	0.688
PTPN13-4	Q12923	1793	1866	SWISS-MODEL		2DKR A	0.35	0.616
PTPN13-5	Q12923	1891	1955	SWISS-MODEL		1UEZ A	0.39	0.772
PTPN3-1	P26045	513	596	SWISS-MODEL		2VPH A	0.71	0.830
PTPN4-1	P29074	520	603	XRAY	2CS5			
RADIL-1	Q96JH8	976	1062	NMR	1UM1			
RAPGEF6-1	Q8TEU7	538	610	SWISS-MODEL		1UF1 A	0.51	0.821
RGS12-1	O14924	26	97	XRAY	2KV8			
RGS3-1	P49796	302	374	XRAY	2F5Y			
RHPN1-1	Q8TCX5	542	611	SWISS-MODEL		1VAE A	0.43	0.662
RHPN2-1	Q8IUC4	524	592	XRAY	2VSV			
RIMS1-1	Q86UR5	608	689	SWISS-MODEL		2CSS A	1.00	0.734
SCRIB-1	Q14160	725	816	XRAY	2W4F			
SCRIB-2	Q14160	853	958	NMR	1WHA			
SCRIB-3	Q14160	1007	1091	SWISS-MODEL		3GSL A	0.44	0.745
SCRIB-4	Q14160	1106	1190	XRAY	1UJU			
SDCBP-1	O00560	117	193	XRAY	1YBO			
SDCBP-2	O00560	198	274	SWISS-MODEL		1NFE A	1.00	0.865
SDCBP2-1	Q9H190	111	186	SWISS-MODEL		1W9E B	0.69	0.872
SDCBP2-2	Q9H190	195	265	SWISS-MODEL		1NFE A	0.70	0.869
SHANK2-1	Q9UPX8	250	339	SWISS-MODEL		1Q3O A	0.90	0.832
SHANK3-1	Q9BYB0	565	668	SWISS-MODEL		1Q3O A	0.86	0.719
SHROOM3-1	Q8TF72	36	111	SWISS-MODEL		2EDP A	0.63	0.745
SHROOM4-1	Q9ULL8	19	93	NMR	2EDP			
SIPA1-1	Q96FS4	690	759	SWISS-MODEL		2EEH A	0.32	0.723
SIPA1L1-1	O43166	957	1026	SWISS-MODEL		2YT8 A	0.31	0.684

SIPA1L2-1	Q9P2F8	959	1026	SWISS-MODEL		1G9O A	0.33	0.804
SIPA1L3-1	O60292	975	1042	SWISS-MODEL		2YT8 A	0.34	0.671
SLC9A3R1-1	O14745	20	92	XRAY	1G9O			
SLC9A3R1-2	O14745	159	232	SWISS-MODEL		2KRG A	1.00	0.824
SLC9A3R2-2	Q15599	147	229	XRAY	2HE4			
SNTA1-1	Q13424	83	171	SWISS-MODEL		1QAV A	0.99	0.804
SNTB1-1	Q13884	117	193	SWISS-MODEL		2VRF A	0.86	0.993
SNTB2-1	Q13425	118	196	XRAY	2VRF			
SNTG1-1	Q9NSN8	60	137	SWISS-MODEL		1Z87 A	0.45	0.769
SNTG2-1	Q9NY99	76	155	SWISS-MODEL		1Z87 A	0.54	0.639
SYNJ2BP-1	P57105	16	99	SWISS-MODEL		2JIK A	1.00	0.977
SYNPO2-1	Q9UMS6	7	89	SWISS-MODEL		1WF7 A	0.38	0.696
SYNPO2L-1	Q9H987	7	89	SWISS-MODEL		2EDP A	0.43	0.644
TIAM1-1	Q13009	856	920	XRAY	2D8I			
TIAM2-1	Q8IVF5	891	977	SWISS-MODEL		1KY9 B	0.36	0.588
TJP1-1	Q07157	26	108	XRAY	2H2C			
TJP1-2	Q07157	189	263	XRAY	2RCZ			
TJP1-3	Q07157	429	499	SWISS-MODEL		1UF1 A	0.40	0.783
TJP2-1	Q9UDY2	35	118	XRAY	1CSJ			
TJP2-2	Q9UDY2	307	386	XRAY	3E17			
TJP2-3	Q9UDY2	518	581	SWISS-MODEL		1UF1 A	0.38	0.783
TJP3-1	O95049	16	90	SWISS-MODEL		2H2B A	0.56	0.966
TJP3-2	O95049	195	272	SWISS-MODEL		2OSG A	0.58	0.594
TJP3-3	O95049	394	461	SWISS-MODEL		1UM7 A	0.43	0.792
USH1C-1	Q9Y6N9	91	166	SWISS-MODEL		3K1R A	0.99	1.000
USH1C-2	Q9Y6N9	220	289	SWISS-MODEL		2KBS A	1.00	0.932
USH1C-3	Q9Y6N9	460	527	SWISS-MODEL		1V6B A	0.96	0.837
WHRN-1	Q9P202	144	217	XRAY	1UEZ			
WHRN-2	Q9P202	288	358	XRAY	1UF1			
WHRN-3	Q9P202	820	888	SWISS-MODEL		1UFX A	1.00	1.000

D. Validation results of structure-based predictions against known PDZ domain-peptide interactions

Table B-4 Validation results for human PDZ domain proteome scanning predictions against known interactions in PDZBase. Proteome scanning predictions for 45 human PDZ domains were validated against known PDZ domain-peptide interactions in PDZBase. Several statistics were calculated including: #P (number of positives), #TP (total number of true positives), #Pred. Struct. (number of predictions predicted only by the structure-based predictor), #Pred. Seq. (number of predictions predicted only by the sequence-based predictor), #Pred. Both (number of predictions predicted by both), #TP Struct. (number of true positives predicted by the structure-based predictor only), #TP Seq. (number of true positives predicted by the sequence-based predictor only), #TP Both (number of true positives predicted by both).

Domain Name	#P	#TP	#Pred. Struct.	#Pred. Seq.	#Pred. Both	#TP Struct.	#TP Seq.	#TP Both	Template Seq ID	QMEAN Score
ARHGEF11-1	2	0	273	0	0	0	0	0	1.00	0.921
CASK-1	6	2	207	671	9	0	2	0		
DLG1-1	2	2	283	127	173	0	0	2	0.99	0.603
DLG1-2	3	3	117	246	162	0	1	2		
DLG2-1	2	2	214	122	178	0	0	2	0.98	0.997
DLG2-2	2	2	389	182	226	0	0	2	1.00	0.953
DLG3-1	2	1	192	159	141	0	0	1		
DLG3-2	2	2	235	171	237	1	0	1		
DLG4-1	2	2	250	112	188	0	0	2	1.00	0.966
DLG4-2	2	2	110	225	183	0	0	2	1.00	0.991
ERBB2IP-1	2	2	64	72	13	0	0	2		
GIPC1-1	6	2	188	193	32	1	1	0	0.65	0.852
GOPC-1	2	0	21	1	0	0	0	0		
GRD2I-1	1	0	18	354	5	0	0	0	0.36	0.509
INADL-6	1	0	243	19	1	0	0	0		
INADL-8	1	0	597	73	116	0	0	0		
LIN7A-1	1	1	195	128	231	0	0	1	0.92	0.793
LIN7B-1	2	2	171	197	251	1	0	1		
LIN7C-1	1	1	123	217	231	0	0	1	0.94	0.827
MAGI2-2	1	0	28	723	22	0	0	0		
MAGI2-6	2	1	212	252	184	0	0	1		
MAGI3-2	1	0	66	997	283	0	0	0	0.74	0.725
MLLT4-1	6	1	8	47	0	0	1	0	1.00	0.530
MPDZ-10	4	3	235	121	78	0	0	3		
MPDZ-13	2	1	156	21	49	1	0	0		
MPP1-1	1	0	109	309	25	0	0	0		
MPP5-1	1	0	44	4	1	0	0	0		
PDZD3-3	1	1	22	979	47	0	1	0	1.00	1.000
PDZK1-1	1	1	70	525	256	0	0	1	0.89	0.813
PICK1-1	5	0	65	0	0	0	0	0		
PTPN13-2	2	1	80	194	28	0	1	0	0.99	0.355
PTPN13-3	2	0	184	0	0	0	0	0	0.39	0.688
PTPN13-4	2	0	53	0	0	0	0	0	0.35	0.616

PTPN13-5	1	0	3	6	0	0	0	0	0.39	0.772
PTPN3-1	1	1	1	719	27	0	1	0	0.71	0.830
PTPN4-1	2	1	73	147	61	1	0	0		
SLC9A3R1-1	7	4	20	670	56	0	4	0		
SLC9A3R1-2	2	2	1	2720	44	0	1	1	1.00	0.824
SLC9A3R2-2	3	3	224	909	587	0	0	3		
SNTA1-1	3	2	378	208	190	0	0	2	0.99	0.804
SNTB2-1	2	1	99	337	125	0	1	0		
SNTG1-1	1	1	7	681	39	0	1	0	0.45	0.769
SNTG2-1	1	1	26	365	127	0	0	1	0.54	0.639
TJP1-2	2	0	29	197	5	0	0	0		
USH1C-1	2	0	80	186	13	0	0	0	0.99	1.000

Table B-5 Validation results for human PDZ domain proteome scanning predictions against known negative interactions (with PDZ binding motifs) in the literature. Proteome scanning predictions for 74 human PDZ domains were validated against experimentally determined negative interactions involving peptides with PDZ binding motifs (found from the literature) for a total of 410 interactions (Luck et al. 2011).

Domain Name	**#FP**	**#N**	**Domain Name**	**#FP**	**#N**	**Domain Name**	**#FP**	**#N**
CASK-1	1	2	INADL-6	1	4	MPDZ-12	3	5
DLG1-1	1	3	INADL-7	2	5	MPDZ-13	0	3
DLG1-2	1	3	INADL-8	1	4	MPDZ-2	0	2
DLG1-3	1	2	INADL-9	0	5	MPDZ-3	0	2
DLG3-1	0	2	LIN7A-1	0	1	MPDZ-4	0	2
DLG3-2	0	2	LIN7B-1	0	6	MPDZ-5	1	2
DLG3-3	0	2	MAGI1-2	0	19	MPDZ-6	0	2
DLG4-1	1	5	MAGI1-3	2	23	MPDZ-7	0	2
DLG4-2	0	4	MAGI1-4	0	18	MPDZ-8	0	3
DLG4-3	1	5	MAGI1-5	7	26	MPDZ-9	1	3
ERBB2IP-1	0	13	MAGI1-6	5	20	PICK1-1	0	3
GIPC1-1	0	5	MAGI2-1	1	7	PTPN13-1	0	3
GOPC-1	0	8	MAGI2-2	1	8	PTPN13-2	0	5
GRIP2-1	0	5	MAGI2-3	1	11	PTPN13-3	0	2
GRIP2-2	2	5	MAGI2-4	4	12	PTPN13-4	0	2
GRIP2-3	0	5	MAGI2-5	8	13	PTPN13-5	0	3
GRIP2-4	0	2	MAGI2-6	1	6	PTPN3-1	0	2

GRIP2-5	0	2	MAGI3-2	2	8	SHANK2-1	0	1
GRIP2-6	2	2	MAGI3-3	0	10	SLC9A3R1-1	0	1
GRIP2-7	0	1	MAGI3-4	1	11	SLC9A3R1-2	0	1
INADL-1	0	5	MAGI3-5	3	13	SNTA1-1	0	2
INADL-10	0	5	MAGI3-6	2	10	TJP1-1	1	3
INADL-2	0	5	MPDZ-1	1	2	TJP1-2	0	3
INADL-3	0	5	MPDZ-10	1	2	TJP1-3	1	3
INADL-5	2	5	MPDZ-11	0	5			

Table B-6 Validation results for human PDZ domain proteome scanning predictions against known negative interactions (with no PDZ binding motifs) in the literature. Proteome scanning predictions for 47 human PDZ domains were validated against known negative interactions involving mutated peptides with non-binding PDZ motifs (found from the literature) for a total of 126 interactions (Luck et al. 2011).

Domain Name	#FP	#N	Domain Name	#FP	#N	Domain Name	#FP	#N
ERBB2IP-1	0	1	MAGI2-1	0	3	MPDZ-2	0	3
GIPC1-1	0	8	MAGI2-2	0	3	MPDZ-3	0	3
INADL-1	0	2	MAGI2-3	0	3	MPDZ-4	0	3
INADL-10	0	2	MAGI2-4	0	3	MPDZ-5	0	3
INADL-2	0	2	MAGI2-5	0	3	MPDZ-6	0	3
INADL-3	0	2	MAGI2-6	0	3	MPDZ-7	0	3
INADL-5	0	2	MAGI3-2	0	1	MPDZ-8	0	3
INADL-6	0	2	MAGI3-3	0	1	MPDZ-9	0	3
INADL-7	0	2	MAGI3-4	0	1	PICK1-1	0	4
INADL-8	0	2	MAGI3-5	0	1	PTPN13-1	0	2
INADL-9	0	2	MAGI3-6	0	1	PTPN13-2	0	3
MAGI1-2	0	4	MPDZ-1	0	3	PTPN13-3	0	2
MAGI1-3	0	5	MPDZ-10	0	3	PTPN13-4	0	1
MAGI1-4	0	4	MPDZ-11	0	3	PTPN13-5	0	2
MAGI1-5	0	5	MPDZ-12	0	3	MPDZ-2	0	3
MAGI1-6	0	5	MPDZ-13	0	3			

Table B-7 Validation results for worm PDZ domain proteome scanning predictions against experimentally determined interactions. Proteome scanning was performed for six worm PDZ domains with interactions from protein microarray experiments. Several statistics were calculated including the ones from **Table B-6** as well as the following: #N (number of negatives), #FP Struct. (number of false positives predicted by the structure-based predictor only), #FP Seq. (number of false positives predicted by the sequence-based predictor only), #FP Both (number of false positives predicted by both).

Domain Name	#P	#TP	#Pred. Struct.	#Pred. Seq.	#Pred. Both	#TP Struct.	#TP Seq.	#TP Both	#N	#FP Struct.	#FP Seq.	#FP Both
DLG1-1	4	4	251	17	28	3	0	1	18	6	0	1
DLG1-3	7	4	36	65	22	0	3	1	15	3	1	0
DSH-1	11	6	523	5	8	6	0	0	4	1	0	0
LIN7-1	11	4	33	91	72	1	2	1	11	0	0	1
MPZ1-6	18	12	194	78	68	8	1	3	4	1	0	0
STN2-1	8	4	19	212	47	1	1	2	14	0	0	0

Table B-8 Validation results for fly PDZ domain proteome scanning predictions against experimentally determined interactions. Proteome scanning was performed for six worm PDZ domains with interactions from protein microarray experiments. Several statistics were calculated including the ones from **Table B-6** as well as the following: #N (number of negatives), #FP Struct. (number of false positives predicted by the structure-based predictor only), #FP Seq. (number of false positives predicted by the sequence-based predictor only), #FP Both (number of false positives predicted by both).

Domain Name	#P	#TP	#Pred. Struct.	#Pred. Seq.	#Pred. Both	#TP Struct.	#TP Seq.	#TP Both	#N	#FP Struct.	#FP Seq.	#FP Both
DLG1-1	4	4	115	43	66	0	0	4	16	4	0	0
DSH-1	4	0	77	37	6	0	0	0	16	3	0	0
LAP4-2	5	4	15	15	13	1	0	3	15	2	1	0
LAP4-3	9	5	81	0	8	3	0	2	11	1	0	0
MAGI-4	3	2	54	66	26	0	1	1	17	2	1	2
PAR6-1	2	0	1	0	0	0	0	0	18	0	0	0

Table B-9 Validation results for human PDZ domain proteome scanning predictions against known interactions in iRefIndex. Proteome scanning results for 221 human PDZ domains with both structure-based and sequence-based predictions were validated against known human PPIs in iRefIndex. A prediction is considered to be a true positive if the domain involved is found in a known PPI where one of the proteins contains the domain. See **Table B-6** caption for details about the statistics calculated.

Domain Name	#P	#TP	#Pred. Struct.	#Pred. Seq.	#Pred. Both	#TP Struct.	#TP Seq.	#TP Both
APBA1-1	7	0	9	1	0	0	0	0

APBA1-2	7	0	56	23	19	0	0	0
APBA3-1	2	1	404	49	2	1	0	0
APBA3-2	2	1	172	13	12	0	0	1
ARHGAP21-1	4	0	7	1760	17	0	0	0
ARHGEF11-1	11	1	273	0	0	1	0	0
ARHGEF12-1	3	0	451	0	0	0	0	0
CAR11-1	6	0	0	878	0	0	0	0
CASK-1	13	2	207	671	9	1	1	0
CNKSR1-1	9	0	2	1	0	0	0	0
CNKSR2-1	8	0	16	8	0	0	0	0
DEPTOR-1	4	0	215	9	13	0	0	0
DLG1-1	23	11	283	127	173	2	0	9
DLG1-2	23	12	117	246	162	3	1	8
DLG1-3	23	12	398	206	168	6	0	6
DLG2-1	9	2	214	122	178	1	0	1
DLG2-2	9	4	389	182	226	3	0	1
DLG2-3	9	4	1174	149	225	2	0	2
DLG3-1	22	4	192	159	141	1	0	3
DLG3-2	22	3	235	171	237	0	0	3
DLG3-3	22	4	171	272	102	1	0	3
DLG4-1	28	8	250	112	188	4	1	3
DLG4-2	28	8	110	225	183	3	2	3
DLG4-3	28	9	367	244	130	3	1	5
DLG5-1	2	0	0	25	0	0	0	0
DLG5-2	2	0	0	20	0	0	0	0
DLG5-3	2	0	2	239	0	0	0	0
DLG5-4	2	0	2	97	2	0	0	0
DVL1-1	11	0	717	68	34	0	0	0
DVL1L1-1	1	0	81	173	7	0	0	0
DVL2-1	18	0	793	63	59	0	0	0
DVL3-1	9	0	92	89	15	0	0	0
ERBB2IP-1	8	1	64	72	13	1	0	0
GIPC1-1	25	3	188	193	32	3	0	0
GOPC-1	7	0	21	1	0	0	0	0
GORASP1-1	2	0	0	69	0	0	0	0
GORASP2-1	7	0	54	30	2	0	0	0
GRD2I-1	1	0	18	354	5	0	0	0

GRD2I-2	1	0	41	1127	44	0	0	0
GRIP1-1	32	0	166	104	7	0	0	0
GRIP1-2	32	0	20	6	4	0	0	0
GRIP1-3	32	0	69	14	1	0	0	0
GRIP1-4	32	0	1	746	0	0	0	0
GRIP1-5	32	1	11	1439	13	0	1	0
GRIP1-6	32	1	1170	277	63	1	0	0
GRIP1-7	32	0	789	5	5	0	0	0
GRIP2-1	9	0	8	35	0	0	0	0
GRIP2-2	9	0	12	253	19	0	0	0
GRIP2-3	9	0	9	215	6	0	0	0
GRIP2-4	9	0	30	155	2	0	0	0
GRIP2-5	9	1	8	1644	16	0	1	0
GRIP2-6	9	0	1672	214	40	0	0	0
GRIP2-7	9	0	473	2	12	0	0	0
HTRA1-1	2	0	1	0	0	0	0	0
HTRA2-1	7	0	90	0	0	0	0	0
IL16-1	5	1	621	0	6	1	0	0
IL16-2	5	0	0	194	0	0	0	0
IL16-3	5	0	80	0	5	0	0	0
IL16-4	5	0	0	0	0	0	0	0
INADL-1	7	0	16	294	20	0	0	0
INADL-10	7	0	28	43	8	0	0	0
INADL-2	7	0	269	191	97	0	0	0
INADL-3	7	0	230	26	0	0	0	0
INADL-4	7	0	0	0	0	0	0	0
INADL-5	7	0	1231	39	51	0	0	0
INADL-6	7	0	243	19	1	0	0	0
INADL-7	7	0	137	8	13	0	0	0
INADL-8	7	0	597	73	116	0	0	0
INADL-9	7	0	45	23	2	0	0	0
LDB3-1	1	0	152	71	12	0	0	0
LIMK1-1	10	0	100	3	0	0	0	0
LIMK2-1	3	0	3	2	0	0	0	0
LIN7A-1	5	1	195	128	231	1	0	0
LIN7B-1	6	3	171	197	251	1	0	2
LIN7C-1	8	2	123	217	231	1	0	1

LMO7-1	3	0	71	87	7	0	0	0
LNX1-1	11	0	0	32	0	0	0	0
LNX1-2	11	0	0	197	0	0	0	0
LNX1-3	11	0	0	31	0	0	0	0
LNX1-4	11	0	0	0	0	0	0	0
LNX2-1	3	0	0	6	0	0	0	0
LNX2-2	3	0	0	236	0	0	0	0
LNX2-3	3	0	0	57	0	0	0	0
LNX2-4	3	0	0	12	0	0	0	0
LRRC7-1	4	1	62	100	36	0	0	1
MAGI1-1	13	0	0	21	0	0	0	0
MAGI1-2	13	3	2	466	9	0	2	1
MAGI1-3	13	2	135	229	61	0	1	1
MAGI1-4	13	2	34	57	15	0	0	2
MAGI1-5	13	2	588	27	42	2	0	0
MAGI1-6	13	5	648	611	587	2	0	3
MAGI2-1	8	1	17	431	5	0	1	0
MAGI2-2	8	2	28	723	22	0	2	0
MAGI2-3	8	0	38	0	0	0	0	0
MAGI2-4	8	1	82	117	25	0	1	0
MAGI2-5	8	2	661	50	102	1	0	1
MAGI2-6	8	1	212	252	184	0	1	0
MAGI3-1	10	0	0	7	0	0	0	0
MAGI3-2	10	6	66	997	283	0	2	4
MAGI3-3	10	0	21	0	0	0	0	0
MAGI3-4	10	1	61	132	23	0	1	0
MAGI3-5	10	1	265	46	48	0	0	1
MAGI3-6	10	4	527	594	688	0	0	4
MAST1-1	33	0	7	387	43	0	0	0
MAST2-1	6	2	52	367	69	0	2	0
MAST3-1	4	0	21	516	44	0	0	0
MLLT4-1	19	0	8	47	0	0	0	0
MPDZ-1	13	2	350	314	96	2	0	0
MPDZ-10	13	7	235	121	78	2	1	4
MPDZ-11	13	0	112	25	1	0	0	0
MPDZ-12	13	4	437	3	2	4	0	0
MPDZ-13	13	3	156	21	49	2	0	1

MPDZ-2	13	3	350	1002	89	3	0	0
MPDZ-3	13	1	1209	561	530	1	0	0
MPDZ-4	13	2	180	112	36	1	0	1
MPDZ-5	13	6	1252	353	109	6	0	0
MPDZ-6	13	0	1	339	0	0	0	0
MPDZ-7	13	2	30	103	23	1	1	0
MPDZ-8	13	0	4	75	1	0	0	0
MPDZ-9	13	4	764	120	26	4	0	0
MPP3-1	7	1	5	30	0	0	1	0
MPP4-1	1	0	9	135	1	0	0	0
MPP5-1	5	0	44	4	1	0	0	0
MPP6-1	16	1	302	3	0	1	0	0
MPP7-1	1	0	0	187	0	0	0	0
MYO18A-1	6	0	0	363	1	0	0	0
NOS1-1	8	0	114	8	7	0	0	0
PARD3-1	24	0	18	1	3	0	0	0
PARD3-2	24	0	0	131	0	0	0	0
PARD3-3	24	0	27	1	0	0	0	0
PARD3B-1	6	0	58	104	7	0	0	0
PARD3B-2	6	0	145	431	51	0	0	0
PARD3B-3	6	0	104	23	3	0	0	0
PARD6A-1	12	0	108	0	0	0	0	0
PARD6B-1	7	0	26	0	0	0	0	0
PARD6G-1	5	0	115	0	0	0	0	0
PDLIM1-1	5	0	58	234	5	0	0	0
PDLIM4-1	1	0	173	92	41	0	0	0
PDLIM5-1	1	0	111	108	6	0	0	0
PDLIM7-1	12	0	40	27	0	0	0	0
PDZD11-1	2	0	30	308	43	0	0	0
PDZD2-1	1	0	0	1	0	0	0	0
PDZD2-2	1	0	1	1113	5	0	0	0
PDZD2-3	1	0	671	0	0	0	0	0
PDZD2-4	1	0	111	18	3	0	0	0
PDZD2-5	1	0	316	0	2	0	0	0
PDZD2-6	1	0	0	20	0	0	0	0
PDZD3-1	3	1	132	290	60	0	1	0
PDZD3-2	3	0	4	11	3	0	0	0

PDZD3-3	3	3	22	979	47	0	3	0
PDZD3-4	3	0	0	17	0	0	0	0
PDZD4-1	1	0	118	0	0	0	0	0
PDZK1-1	9	3	70	525	256	0	3	0
PDZK1-2	9	1	15	147	9	0	1	0
PDZK1-3	9	5	60	1171	368	1	2	2
PDZK1-4	9	2	45	338	41	1	1	0
PDZRN3-1	2	0	158	37	3	0	0	0
PDZRN3-2	2	0	185	0	0	0	0	0
PICK1-1	31	0	65	0	0	0	0	0
PPP1R9A-1	2	0	87	6	3	0	0	0
PPP1R9B-1	14	0	11	11	0	0	0	0
PREX1-1	3	0	0	11	0	0	0	0
PREX2-1	2	0	0	16	0	0	0	0
PRX-1	6	0	0	2	0	0	0	0
PSMD9-1	3	0	0	0	0	0	0	0
PTPN13-1	7	0	391	2	2	0	0	0
PTPN13-2	7	0	80	194	28	0	0	0
PTPN13-3	7	0	184	0	0	0	0	0
PTPN13-4	7	0	53	0	0	0	0	0
PTPN13-5	7	0	3	6	0	0	0	0
PTPN3-1	9	1	1	719	27	0	1	0
PTPN4-1	10	0	73	147	61	0	0	0
RADIL-1	14	0	73	947	21	0	0	0
RAPGEF2-1	13	0	0	0	0	0	0	0
RAPGEF6-1	4	0	1529	0	5	0	0	0
RGS12-1	1	1	21	880	42	0	0	1
RGS3-1	2	0	94	2684	31	0	0	0
RHPN2-1	10	0	60	200	23	0	0	0
RIMS1-1	17	0	108	93	7	0	0	0
RIMS2-1	3	0	0	108	0	0	0	0
SCRIB-1	14	3	150	55	84	1	0	2
SCRIB-2	14	3	130	38	65	1	0	2
SCRIB-3	14	3	344	0	0	3	0	0
SCRIB-4	14	2	107	152	60	0	1	1
SDCBP-1	10	0	8	123	4	0	0	0
SDCBP-2	10	0	17	1	0	0	0	0

SHANK1-1	3	1	0	1105	0	0	1	0
SHANK2-1	13	2	4	1505	139	0	1	1
SHANK3-1	8	0	224	883	640	0	0	0
SHROOM2-1	4	0	0	1586	0	0	0	0
SHROOM3-1	1	0	20	508	27	0	0	0
SIPA1-1	3	0	252	0	0	0	0	0
SIPA1L1-1	2	0	2	0	0	0	0	0
SIPA1L3-1	2	0	0	0	0	0	0	0
SLC9A3R1-1	14	6	20	670	56	0	6	0
SLC9A3R1-2	14	8	1	2720	44	0	8	0
SLC9A3R2-1	16	9	0	693	0	0	9	0
SLC9A3R2-2	16	11	224	909	587	0	2	9
SNTA1-1	8	4	378	208	190	1	0	3
SNTB1-1	3	2	55	358	104	0	0	2
SNTB2-1	5	1	99	337	125	0	0	1
SNTG1-1	3	0	7	681	39	0	0	0
SNTG2-1	1	1	26	365	127	0	0	1
SNX27-1	2	2	0	728	0	0	2	0
STXBP4-1	1	0	0	20	0	0	0	0
SYNJ2BP-1	8	4	9	691	78	0	3	1
SYNPO2-1	3	0	77	252	6	0	0	0
TAX1BP3-1	2	0	0	262	0	0	0	0
TIAM1-1	5	0	12	0	0	0	0	0
TJP1-1	26	2	129	306	42	0	1	1
TJP1-2	26	1	29	197	5	0	1	0
TJP1-3	26	1	197	351	32	0	1	0
TJP2-1	14	1	152	123	2	1	0	0
TJP2-2	14	0	0	1	0	0	0	0
TJP2-3	14	1	140	329	22	0	0	1
TJP3-1	4	1	158	283	16	0	1	0
TJP3-2	4	0	0	1	0	0	0	0
TJP3-3	4	0	128	214	55	0	0	0
USH1C-1	2	0	80	186	13	0	0	0
USH1C-2	2	0	89	1	4	0	0	0
USH1C-3	2	0	43	0	0	0	0	0