ABSTRACT

Motivation: Predicting protein interactions involving peptide recognition domains is essential for understanding the many important biological processes they mediate. It is important to consider the binding strength of these interactions to help us construct more biologically relevant protein interaction networks that consider cellular context and competition between potential binders.

Results: We developed a novel regression framework that considers both positive (quantitative) and negative (qualitative) interaction data available for mouse PDZ domains to quantitatively predict interactions between PDZ domains, a large peptide recognition domain family, and their peptide ligands using primary sequence information. First, we show that it is possible to learn from existing quantitative and negative interaction data to infer the relative binding strength of interactions involving previously unseen PDZ domains and/or peptides given their primary sequence. Performance was measured using cross-validated hold out testing and testing with previously unseen PDZ domain–peptide interactions. Second, we find that incorporating negative data improves quantitative interaction prediction. Third, we show that sequence similarity is an important prediction performance determinant, which suggests that experimentally collecting additional quantitative interaction data for underrepresented PDZ domain subfamilies will improve prediction.

Availability and Implementation: The Matlab code for our SemiSVR predictor and all data used here are available at http://baderlab.org/Data/PDZAffinity.

Contact: gary.bader@utoronto.ca; dengnaiyang@cau.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on July 30, 2010; revised on November 8, 2010; accepted on November 22, 2010

1 INTRODUCTION

Modular domains are the major building blocks of eukaryotic proteins and interaction networks (Pawson and Nash, 2003). These domains usually fold independently and are present in various combinations within a single protein to create a rich repertoire of functionally diverse proteins from a more limited domain set (Vogel et al., 2004). An important subclass of these domains, peptide recognition modules (PRMs), bind to short extended and linear peptide segments in target proteins to mediate protein–protein interactions in eukaryotic cell signaling systems (Pawson and Nash, 2003). Characterizing the interactions of these peptide recognition modules will help us map and understand the many biological processes they mediate.

PRMs generally bind their peptide ligands in the weak (10μM) affinity (binding strength) range (Castagnoli et al., 2004), and sensitive in vitro experimental techniques like phage display (Tonikian et al., 2008, 2009) and peptide/protein microarrays (Jones et al., 2006; Stiffler et al., 2006) have been used to map the binding specificities and protein interactions of large sets of SH3 (Landgraf et al., 2004; Tonikian et al., 2009), SH2 (Huang et al., 2008; Jones et al., 2006), WW (Hu et al., 2004) and PDZ (Stiffler et al., 2006; 2007; Tonikian et al., 2008) domains. However, these experimental techniques are resource intensive, and cannot be readily applied to new members and alleles of PRMs that are increasingly being collected by genome sequencing projects and population-based genetic variation studies (The International HapMap Consortium, 2007). Ideally, a computational model could be developed to predict whether a PRM will bind to a peptide given their primary sequences. Such a model could be used to predict protein interactions from newly sequenced genomes and the effect of mutations on known PRM-mediated protein interactions to guide subsequent experimental characterization. Computational domain–peptide interaction prediction has been studied for multiple PRMs, such as SH3 (Ferraro et al., 2006; Yaffe et al., 2001), SH2 (Sanchez et al., 2008; Wunderlich and Miny, 2009), PDZ (Chen et al., 2008), WW (Schleinkofer et al., 2004) and MHC domains (Jacob and Vert, 2008; Nielsen et al., 2008; Zhang et al., 2009) (which are a special case of peptide binding domains that do not mediate protein–protein interactions). These approaches, except for...
We find that SemiSVR's performance is superior to a previously published method, trained on the set of interaction domains and/or peptides from their primary sequences. Interaction data generated in Stiffler et al. (2007) consists of a positive dataset of PDZ domain–peptide interactions with binding affinity measurements and a negative dataset (non-interacting PDZ domain–peptide pairs, with no binding affinity measurements). Intuitively, the negative interaction data provide qualitative information on the contribution of amino acids to binding affinity that could improve quantitative prediction. Popularity used position weight matrix (PWM) and conventional regression methods like support vector regression (SVR), however, cannot incorporate qualitative negative data. Here, we devised a novel extension of SVR, termed SemiSVR, that considers both quantitative positive and qualitative negative interaction data. We show that SemiSVR, being able to incorporate negative data, is better than SVR and PWM in identifying the stronger interactors among previously unseen peptides. Next, through a feature-encoding framework that considers both the primary sequence of PDZ domains and peptides, we applied SemiSVR to predict relative binding strength of PDZ domain–peptide interactions involving previously unseen PDZ domains. We find that SemiSVR’s performance is superior to a previously published method on the same dataset (Chen et al., 2008) and the naive usage of PWM from sequence-similar PDZ domains.

2 MATERIALS AND METHODS

2.1 Data

Our training data is that published in Chen et al. (2008), which is a cleaned subset of interactions with measured affinities originally reported by Stiffler et al. (2007), containing interactions between 82 mouse PDZ domains and 217 mouse genome-derived (genomic) peptides. Briefly, interactions were assessed using a peptide microarray followed by confirmation of positives and measurement of binding affinities by fluorescence polarization (FP), a high-quality affinity measurement method (Stiffler et al., 2007). This resulted in 560 PDZ domain–peptide interactions, involving 82 mouse PDZ domains and 93 peptides, and 1167 negative interactions, involving 82 mouse PDZ domains and 138 peptides, which were confirmed by FP. The 560 positive interactions have measured affinities (each measured as a dissociation constant, Kd) of less than 100 µM (high Kd indicates weak interaction and low Kd indicates strong interaction, see Supplementary Fig. S1 for distribution of the Kds) while the affinities of the non-binding pairs are identified to be greater than the threshold (100 µM), but Kd values are not measured. We call the mixture of both quantitative (positive interactions) and qualitative (negative interactions) ‘semi-quantitative’ data. The number of binding peptides per PDZ domain varies widely. Among the 82 domains, 23 have at least 10 binding peptides, which we use for training (Supplementary Table S1).

2.2 Predictor

We wish to predict quantitative PDZ domain–peptide interactions based on known interactions and affinity data. To do this, we developed a new method called Semi-quantitative SVR (SemiSVR), a novel extension of SVR, to learn how to predict the binding affinity of PDZ domain–peptide interactions from both quantitative binding (positive) data and qualitative non-binding (negative) data (Fig. 1). SVR is an established machine learning method for non-linear regression (Smola and Scholkopf, 2004). We extended this method to take advantage of negative information we have available, which is not considered by other regression-based methods. In non-linear regression, the regression function \( f \) is approximated by a kernel function \( K(x,y) \) as
follows:

\[ f(x) = \sum_{i=1}^{m} a_i K(x_i, x) + b \]  

where \( x_i \) is the known training data, and \( a = [a_1, a_2, \ldots, a_m]^T \) (\( T \) = transpose, \( m \) = training dataset size) is the Lagrange multiplier and \( b \) is the bias threshold. The SemiSVR aims to determine the unknown multiplier \( a \) and the bias \( b \) based on the training data. Given the training dataset \( S = \{ (x_i, y_i) \} : x_i \in \mathbb{R}^n, y_i \in \mathbb{R}^{m+1}, \) where \( x_i, y_i \) are the input features for the positive and negative PDZ domain–peptide pairs, respectively, \( y_i \) is the affinity for positive quantitative data \( x_i \) and the regression value for the training data \( \bar{y} \) is greater than a threshold (i.e. \( \bar{y} > 100 \mu M \)).

For the positive quantitative data \( y_i \), we wish to minimize the \( \varepsilon \)-insensitive loss function, based on the error criterion that leads to \( |f(x_i) - y_i| \leq \varepsilon \), that is the regression values \( f(x) \) on the training data \( x_i \) should have less error than \( \varepsilon \), \( i = 1, 2, \ldots, m \). For the negative quantitative data \( y_i \), we wish to make the regression value \( |f(x_i)| \) on these data satisfy the prior knowledge \( |f(x_i)| \geq \text{known training data} \), and \( \forall y_j \), \( j = 1, 2, \ldots, k \).

The above constraints assume that the final regression function \( f(x) \) can approximate all the data with \( \varepsilon \)-precision. Sometimes, however, we want to allow for some errors. As with standard SVR, a slack variable \( \xi \) and \( \zeta \) can be introduced to cope with otherwise unsatisfiable constraints. Considering all of this and similar to previous work on knowledge-based non-linear kernel approximation (Mangasarian and Wild, 2007), the linear programming form of Semi-SVR is given as:

\[
\begin{align*}
\min_{a, b, y, \xi, \zeta} & \left( \sum_{i=1}^{m} a_i y_i \right) + C \sum_{i=1}^{m} \xi_i + C \sum_{j=1}^{k} \zeta_j \\
\text{s.t.} & \sum_{i=1}^{m} a_i K(x_i, x) + b - y_i \leq \xi_i, \\
& \sum_{i=1}^{m} a_i K(x_i, x) + b - \bar{y} \leq \xi_i, \\
& \sum_{j=1}^{k} K(x_j, x) + b - \bar{y} \leq \zeta_j, \\
& \sum_{j=1}^{k} K(x_j, x) + b - y_j \leq \zeta_j, \\
& \xi_i, \zeta_j \geq 0, \quad i = 1, 2, \ldots, m, j = 1, 2, \ldots, k.
\end{align*}
\]

where \( \varepsilon \) is a user-defined constant that contributes to the \( \varepsilon \)-insensitive loss function, which measures the error of the regression on the training data, and is defined as \( |x| = 0, |x| < \varepsilon \), and equal to \( |x| \) if \( |x| \geq \varepsilon \). \( C_1, C_2, \varepsilon > 0 \) are the penalty parameters determining the trade-off between the regularization term (in order to avoid overfitting) and the empirical error (according to \( \varepsilon \)-insensitive loss function). Here, we drive the error down by minimizing the 1-norm of the errors and together with the 1-norm of \( a \) for complexity reduction or stabilization. Previous work shows the alternative 1-norm for 2-norm regularization achieves equivalent performance (see Mangasarian et al., 2004; 2007). Constraints (3–4) ensure that the positive pairs lie in \( \varepsilon \)-precision with some allowed errors while constraints (5–7) ensure the negative pairs satisfy the prior knowledge within some allowed errors. In practice, all affinities are scaled to the range \([−1, 1]\) after taking log10, which makes the data easier to work with. We select parameters \( C_1, C_2, \varepsilon \) and the kernel parameters (or \( \sigma \)) using grid search (Chang and Lin, 2001).

The input of the semi-quantitative SVR model is the encoded representation of the PDZ domain–peptide pair (see below) and the corresponding binding affinity while the output is the predicted affinity score for each pair. (Higher scores mean weaker interaction while lower scores mean stronger interaction, similar to the scale of biochemical KDs).

All software was developed in Matlab 2008 and source code is available on the web site (http://baderlab.org/Data/PDZAffinity).

As a benchmark, we also developed a nearest neighbor SemiSVR for each test PDZ domain that was trained on the closest PDZ domain with both its binding and non-binding peptides. We only trained a predictor for the closest PDZ domain has \( \geq 10 \) binding peptides (changing this threshold to nearby values does not affect our conclusions).

Generally, there are two strategies to build predictive models for peptide recognition domain-mediated interactions. A single-domain model is trained only on the interactions of an individual PDZ domain (one domain and its binding peptides) while a multi-domain model uses interaction data from multiple PDZ domains.

We tested our models using leave-one-PDZ-domain-out cross-validation, as domain sequence is important for performance. For the single-domain model, we trained on one single PDZ domain associated with all the interaction data, and tested for the held-out PDZ domain. For the multi-domain model, at each run, we trained the SemiSVR model on interaction data involving all PDZ domains but one, and then predicted the relative binding strength of all peptides interacting with the held-out PDZ domain.

### 2.3 Feature encoding

We represented a PDZ domain–peptide as a vector of descriptors including sparse vectors of either the full domain (118AA) or different definitions of the domain binding site (16 or 10AA) and the peptide ligand (10AA) and links between domain and peptide positions. Specifically, a PDZ domain \( \{P_1, P_2, \ldots, P_m\} \) and a peptide \( \{p_1, p_2, \ldots, p_n\} \), where \( P_i, p_j \) represent amino acids at a given position \( \sigma \)-th on the PDZ domain or \( \rho \)-th on the peptide)—in our case, \( n = 118 \) and \( k = 10 \). A PDZ-domain–peptide pair is encoded as a tensor (outer) product between descriptor PDZ and peptide: PDZ*peptide = \( \{p_1, p_2, \ldots, p_n\} \), where the inner product between two tensor product vectors (each one encoding one domain–peptide pair) can be rewritten as a product of two inner products (Jacob and Vert, 2008), we compute the inner product between vectors of any two PDZ domains or vectors of any two peptides. Furthermore, kernels on sequences can replace the inner product between the vectors of any two domains or any two peptides (the kernel trick), such as \( K(PDZ_i)^*p\) (PDZ*peptide) = \( K(PDZ_i, PDZ_j) = \varepsilon(PDZ_i, PDZ_j) \). Thus, when sequences are used as the input for the kernels, one can rewrite the polynomial kernel as follows:

\[
K_{poly}(x, y) = (K_{baseline}(x, y) + 1)^{\phi},
\]

where \( K_{baseline} \) is simply the number of letters the input sequences (domain sequences or peptide sequences, respectively) have in common at the same positions. In practice, all kernels were normalized to 1 on the diagonal by \( K_{normalized}(x, y) = K_{poly}(x, y) / K_{poly}(x, y) \) to make computation easier.

To encode the PDZ domain, we used the same alignment as published by Chen et al. (2008) since we want to compare our method to Chen’s method using the same encoding (e.g. the same 16 binding sites). This alignment represents the conserved part of the domain containing all conserved secondary structure elements and the canonical binding site. See Supplementary Material for details and Supplementary Fig. S2 for the pairwise identity distribution of all pairs of 82 PDZ domain sequences based on this alignment. For the peptide, we use the entire length of 10 amino acids in all experiments.

There are many different feature encodings and kernels that could be used for our prediction task. We tried encoding PDZ domains using PreFAR features (Li et al., 2006), which includes amino acid physicochemical properties, sequence pattern frequency and correlations, conventional sparse encoding where each position is represented as a vector of length 20 (one element for every amino acid type) has a one in the element corresponding to the amino acid at that position and the rest of the 19 elements are set to zeros, and all vectors are concatenated; encoding peptides using 5 factors Atchley et al. (2005) and 11 factors (Liu et al., 2006), which are also based on amino acid physicochemical properties. We encoded the peptide using all above encodings, except for PreFAR, for which the peptide sequence is too short. However, none of these encodings (Gaussian kernel) resulted in
better performance than using the above described sequence-based encoding with a polynomial kernel (Supplementary Table S2).

3 RESULTS

Our goal is to predict binding strength of a previously unseen PDZ domain–peptide pair based on the primary sequence of the domain and peptide and quantitative interaction data. To address this, we applied regression analysis to published PDZ domain–peptide binding affinity data obtained using a combination of protein microarray and fluorescence polarization experiments (Chen et al., 2008; Stiffler et al., 2007).

3.1 Incorporating negative data for quantitative interaction prediction

3.1.1 Single domain models

Often, peptides that bind a PDZ domain will be modeled using a PWM, one per domain. The PWM method has been shown to capture binding energy (Stormo, 2000) and is often used for predicting PRM domain–peptide interactions (Tong et al., 2002; Tomkian et al., 2008). As a basic test of the modeling capability of the SemiSVR method and to compare it to the established PWM method (see Supplementary Materials), which was trained on quantitative data only, we trained it using positive quantitative and negative peptide data of an individual PDZ domain. We then tested the ability of each method to distinguish the stronger binding peptide among a pair of peptides randomly held out from the training peptide set (run for all possible peptide pairs, either two binders or a binder and a non-binder) and generated a percentage success rate for each of the 23 PDZ domains that bound at least 10 peptides. We found that the SemiSVR method performs better than the PWM method at the same task for the vast majority of PDZ domains (21/23, average performance of 0.79 versus 0.72, P-value = 0.0023, ranksum testing, Table 1). We had similar results when comparing to SVR (Table 1). Hence, incorporating negative data in regression analysis through SemiSVR improves quantitative prediction of interacting peptides, and even a simple application of the SemiSVR method given a set of peptides per domain is useful.

3.1.2 Multi-domain model

We next trained the SemiSVR on interaction data of multiple PDZ domains to predict quantitative domain–peptide interactions involving previously unseen PDZ domains. We tested this using leave-one-PDZ-domain-out cross-validation, where we trained the SemiSVR model on interaction data involving all PDZ domains but one, and then predicted the relative binding strength of all peptides interacting with the held-out PDZ domain. To measure performance, we correlated the SemiSVR score with actual binding affinities using Pearson and Spearman’s correlation coefficients. Since too few data points lead to inconclusive correlation results, we assessed the performance only for the 23 PDZ domains that bound to 10 or more peptides.

To enable the SemiSVR to learn from interaction data of multiple PDZ domains, the primary sequence of each PDZ domain and peptide in our training set was encoded as a feature vector (as compared with single domain testing where only peptides were encoded). We evaluated various ways of encoding these features (see Section 2 and Supplementary Table S2). For every PDZ domain–peptide interaction, we combined the feature vectors with the interaction binding affinity for regression analysis. We used a pairwise encoding with a polynomial kernel, which captures all pairs of amino acids between all domain and peptide positions, as this predictor performed best in initial experiments (see Section 2 and Supplementary Fig. S3).

### Table 1. Performance comparison of single domain SemiSVR, SVR and PWM on 23 PDZ domains in leave two domain-peptide interactions out cross-validation testing

<table>
<thead>
<tr>
<th>PDZ domain</th>
<th>SemiSVR</th>
<th>SVR</th>
<th>PWM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAPS SYN-110_2/3</td>
<td>0.75</td>
<td>0.57</td>
<td>0.71</td>
</tr>
<tr>
<td>CHAPS SYN-110_3/3</td>
<td>0.86</td>
<td>0.60</td>
<td>0.79</td>
</tr>
<tr>
<td>GM1582_2/3</td>
<td>0.74</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>HTRAX_1/1</td>
<td>0.73</td>
<td>0.66</td>
<td>0.70</td>
</tr>
<tr>
<td>LIN7C_1/1</td>
<td>0.89</td>
<td>0.59</td>
<td>0.76</td>
</tr>
<tr>
<td>MAGI-2_2/6</td>
<td>0.85</td>
<td>0.55</td>
<td>0.73</td>
</tr>
<tr>
<td>MAGI-2_6/6</td>
<td>0.71</td>
<td>0.67</td>
<td>0.69</td>
</tr>
<tr>
<td>MAGI-3_1/5</td>
<td>0.71</td>
<td>0.49</td>
<td>0.64</td>
</tr>
<tr>
<td>MAL52_1/1</td>
<td>0.55</td>
<td>0.40</td>
<td>0.60</td>
</tr>
<tr>
<td>OMP52_1/1</td>
<td>0.77</td>
<td>0.63</td>
<td>0.65</td>
</tr>
<tr>
<td>PDZK3_1/1</td>
<td>0.78</td>
<td>0.64</td>
<td>0.70</td>
</tr>
<tr>
<td>PDZ-KG31_1/1</td>
<td>0.82</td>
<td>0.80</td>
<td>0.68</td>
</tr>
<tr>
<td>SCR5_2/3</td>
<td>0.69</td>
<td>0.37</td>
<td>0.65</td>
</tr>
<tr>
<td>SCR95_3/3</td>
<td>0.82</td>
<td>0.70</td>
<td>0.80</td>
</tr>
<tr>
<td>TFP-B1_2/5</td>
<td>0.83</td>
<td>0.60</td>
<td>0.77</td>
</tr>
<tr>
<td>SAP102_2/3</td>
<td>0.81</td>
<td>0.63</td>
<td>0.66</td>
</tr>
<tr>
<td>SAP97_1/3</td>
<td>0.74</td>
<td>0.57</td>
<td>0.69</td>
</tr>
<tr>
<td>SAP97_2/3</td>
<td>0.74</td>
<td>0.50</td>
<td>0.71</td>
</tr>
<tr>
<td>SCR81_3/4</td>
<td>0.84</td>
<td>0.59</td>
<td>0.75</td>
</tr>
<tr>
<td>SHANK1_1/1</td>
<td>0.91</td>
<td>0.88</td>
<td>0.81</td>
</tr>
<tr>
<td>SHANK1_1/2</td>
<td>0.88</td>
<td>0.82</td>
<td>0.80</td>
</tr>
<tr>
<td>G1-SYNTROPHIN_1/1</td>
<td>0.87</td>
<td>0.58</td>
<td>0.79</td>
</tr>
<tr>
<td>Z01_1/1</td>
<td>0.75</td>
<td>0.51</td>
<td>0.75</td>
</tr>
<tr>
<td>Average Performance</td>
<td>0.79</td>
<td>0.61</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Numbers indicate the average percentage of correct predictions. Bold numbers indicate the best performance.

### Table 2. Performance comparison of SemiSVR and SVR on 23 PDZ domains with associated peptides for Multi-domain model testing

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>SemiSVR</th>
<th>SVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman’s correlation</td>
<td>0.605</td>
<td>0.501</td>
</tr>
<tr>
<td>Pearson correlation</td>
<td>0.653</td>
<td>0.574</td>
</tr>
</tbody>
</table>

Performance comparison based on leave-one-PDZ-domain out cross-validation. A pairwise polynomial kernel (\(P = 2\)) using the whole PDZ (118AA) and whole peptide (10AA) as feature input was used for both predictors. Bold numbers indicate the best performance.
3.2 Performance determinants of quantitative prediction

To explore which aspects of our input features are most important for prediction performance, we trained a multi-domain SemiSVR model using subsets of PDZ domain sequences. We used 16 binding positions from the a1-syntrophin PDZ (a1synPDZ) structure described in Chen et al. (2008), and also 10 core binding positions derived from the intersection of all binding sites in nine available PDZ domain–peptide structures described in Tonikian et al. (2008). Using the whole PDZ sequence gave better overall performance, although the binding site encoding gives comparable performance (Supplementary Table S3), achieving Spearman’s correlation of 0.605, 0.594, 0.594 and Pearson correlation of 0.653, 0.636, 0.649 for whole PDZ sequence, 16 binding positions and 10 core binding positions, respectively. This suggests that additional information is present in non-binding site positions that improves performance. Next, we assessed the relationship between the predictor performance and percent sequence identity of the test PDZ to its nearest domain in the training set. We observed a positive correlation between performance and sequence identity (Fig. 2, Spearman’s correlation, 0.498; P = 0.0157). To further study this trend, we progressively removed all training PDZ domain interactions that are above a sequence similarity threshold to the test PDZ domain and retrained a SemiSVR model for each test domain (whole PDZ, pairwise polynomial kernel). We observed that the average SemiSVR performance decreased as the level of similarity of the test PDZ domain to the closest PDZ domain in the training set decreased (Supplementary Fig. S4). Hence, sequence similarity between a test
Table 4. Performance of our SemiSVR versus local information-based models using different PDZ domain similarity definitions

<table>
<thead>
<tr>
<th>Performance measurement</th>
<th>Spearman’s correlation</th>
<th>Pearson correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SemiSVR</td>
<td>Naïve PWM transfer (Identity)</td>
<td>0.303</td>
</tr>
<tr>
<td></td>
<td>Naïve PWM transfer (BLOSUM62)</td>
<td>0.305</td>
</tr>
<tr>
<td></td>
<td>Nearest neighbor SemiSVR</td>
<td>0.471</td>
</tr>
<tr>
<td></td>
<td>118AA</td>
<td>0.605</td>
</tr>
<tr>
<td></td>
<td>10BS</td>
<td>0.326</td>
</tr>
<tr>
<td></td>
<td>0.354</td>
<td>0.286</td>
</tr>
</tbody>
</table>

Fig. 3. SemiSVR can predict changes in affinity resulting from point mutations introduced into known binding peptides of the a1syn PDZ domain. The three wild-type peptides are denoted by asterisks (*). Each mutant within a set is labeled by a different shape. Residue mutations are highlighted in red. One KIF1B mutant had no measurable binding, so it was excluded from our analysis. Performance of the SemiSVR on peptide mutation of a1synPDZ is very high (Spearman’s correlation, 0.921; $P < 1e-16$ and Pearson correlation, 0.922; $P = 1.414e-07$). All affinities are scaled to the range [0, 1] after taking log10.

3.3 A global approach improves the prediction performance

One potential advantage of our encoding framework approach is that we can incorporate interaction data of multiple PDZ domains (global) rather than just close neighbors (local) to improve prediction. To investigate this, we trained a set of ‘nearest neighbor’ SemiSVR predictors using only interaction data of the single domain with the highest sequence similarity to each test PDZ domain, ensuring that enough interaction data are used to create a viable predictor, and compared their performance to our multi-domain SemiSVR. In addition, since the SemiSVR’s performance is correlated with the sequence similarity of a test PDZ domain to those in training data, we also assessed how the naïve usage of PWM on the only interaction data of the nearest PDZ neighbor performs for quantitative interaction prediction.

We tested different ways to identify the nearest neighbor using whole PDZ, 16 and 10 position binding sites, and based on amino acid identity and scoring matrix BLOSUM62. Nearest neighbor SemiSVR performed better than the naïve PWM transfer method, presumably because negative interactions help with prediction. However, our multi-domain SemiSVR gave the best performance overall (Table 4). Thus, while sequence similarity is an important factor and nearest neighbors are important contributors to performance, our multi-domain SemiSVR uses additional information from across the PDZ domain family to improve performance.

3.4 Validation of the method using blind PDZ domain–peptide affinity measurements

We next tested the SemiSVR model on newly measured PDZ domain–peptide interactions that were not used for training. The third PDZ domain of the human Scribble protein PDZ was cloned, expressed and purified and binding affinities to 57 peptides from natural human proteins were measured using fluorescence polarization (see Supplementary Material and Supplementary Table 4). Only the third PDZ domain was used because the other three had less than 45% sequence identity to the training set. This resulted in 36 binding peptides, enough for a confident performance assessment. The result shows that our SemiSVR method can accurately predict PDZ domain–peptide interactions (Spearman’s correlation, 0.74; $P = 8.85e-7$). We found similar results testing our model on interactions involving domains in Fly and Worm data (Supplementary Table S5).

3.5 Predicting effect of peptide mutations

As another test of our SemiSVR method, we predicted the change in binding affinity of PDZ domain–peptide interactions resulting from amino acid changes in the peptide. We used a previously published dataset of PDZ–peptide affinities measured with fluorescence polarization (Chen et al., 2008) in which five single point mutations were introduced into each of three wild-type binding peptides (from proteins: Kv1.5, Nav1.5 and KIF1B) that bind the a1syn PDZ domain. The SemiSVR model successfully predicted the relative affinity change (increase or decrease versus wild type) for all mutants (i.e. $14/14 = 100\%$; one mutated KIF1B ligand had no measurable binding affinity). The correlations between the predicted and actual affinities of the mutated peptides for the SemiSVR are very high (Spearman’s correlation, 0.921; $P < 1e-16$ and Pearson correlation 0.922; $P = 1.414e-07$) (Fig. 3). Therefore, our method can correctly predict the direction and relative magnitude of affinity changes in the mutant ligand compared with the wild type.

3.6 Binary classification of PDZ domain–peptide interactions

Next, we assessed the performance of the multi-domain SemiSVR method on the presumably easier binary classification task—to predict whether a PDZ domain will bind a peptide or not. We performed leave-one-PDZ-domain-out cross-validation on the 23 PDZ domains with sufficient (>10) positive and negative peptides for the SemiSVR model and computed the average area under the
receiver operating characteristic curve (ROC AUC). The SemiSVR model was trained as before with all 81 non-test PDZ domains. The average ROC AUC score was 0.88 (Supplementary Fig. S5A).

To compare this result with that of a previous method for binary prediction published in Chen et al. (2008), we used their bootstrap testing approach. (i) PDZ bootstrap: leave 12% out for testing; (ii) peptide bootstrap: leave 8% out for testing; and (iii) both PDZ and peptide bootstrap. The SemiSVR performed well in this test (AUC of 0.862 ± 0.016, 0.853 ± 0.021 and 0.848 ± 0.017, respectively, Supplementary Fig. S5B), which is comparable to the published performance of Chen’s model [AUC 0.9] (confidence interval (CI) 0.84–0.96), 0.84 (CI 0.76–0.89) and 0.87 (CI 0.67–0.98), respectively.

4 DISCUSSION

Inferring the relative strength of protein–peptide interactions mediated by peptide recognition modules (PRMs) will lead to better understanding of cellular processes. Here, we show that it is possible to predict affinity of PDZ domain–peptide interactions based on primary sequence information. We also show that incorporating both positive and negative interaction data using a novel SemiSVR approach improves prediction. This approach is also successful at predicting which PDZ domain–peptide pairs are likely to interact (binary prediction).

Based on the experimental data, a threshold of 100 μM separates quantitative ‘positive’ data from qualitative ‘negative’ data. Changing this threshold to more stringent values (i.e. 20 and 10 μM) did not change our results (Supplementary Table S6).

Although our method is mainly based on sequence similarity, it is interesting to analyze how much physicochemical factors contribute to our prediction performance. To investigate this, we assessed how well each of 11 properties from the ‘11-factor’ encoding (Liu et al., 2006) can be used individually for quantitative prediction of PDZ–peptide interactions using SemiSVR. We found that isoelectric point, hydrophilicity scale, polarity, average accessible surface area, van der Waals parameter epsilon and steric parameter are most important for performance, in decreasing order, suggesting they are the physicochemical factors that mostly modulate the binding strength of PDZ–peptide interactions (Supplementary Fig. S6).

Given that physical forces between the domain and the peptide 3D structures determine affinity, and our observation that SemiSVR performance correlates with sequence similarity between PDZ domains in testing and training sets, we postulate that natural PDZ domains with similar sequences have similar 3D structures that determine affinity in similar ways. This is supported by the observation that PDZ specificities are conserved from worm to human (Tonikian et al., 2008). It has been shown that it is easy to mutate PDZ domains to bind non-natural ligands; however, we only see a limited set of PDZ domain specificities in nature (Ernst et al., 2009; Tonikian et al., 2008). These observations are consistent with a constrained model of PDZ specificity evolution where a set of initial PDZ domain specificities evolved, and that these were then expanded to form a finite number of subfamilies, each functionally similar down to the level of affinity determination. This model predicts that each subfamily has a characteristic structure and mode of determining binding affinity with a ligand. Regardless, we find that information useful for prediction is taken from the entire PDZ domain family and this improves prediction performance compared with using a naive nearest neighbor-based predictor. As our method is trained on interaction data of natural PDZ domains, it may not do well at quantitative interaction prediction involving synthetic PDZ domains that have multiple mutations not found in our training data. We have noticed, in other work, that synthetic mutations may cause large changes in specificity, and presumably affinity (Ernst et al., 2009; Tonikian et al., 2008). This may occur by drastically changing the binding mode, for instance, by causing the peptide to rotate. We do not notice these types of large specificity changes arising from small sequence differences in natural PDZ domains, possibly because they disrupt normal PDZ function. The reduced predictive ability on synthetic PDZ domains, at least for specificity, has also been recently noticed (Smith and Kortemme, 2010). However, we were not able to test this due to lack of sufficient affinity data on synthetic PDZ domains.

We observed that some PDZ domains share identical subsequences in the 10 and 16 binding positions but bind the same peptides with different affinity. For example, both Dv11 (1/1) and Dv33 (1/1) share identical subsequences in their 16 binding positions yet bind to peptide Caspr4 with 79.298 μM and 30.756 μM Kd, respectively. Assuming the affinities are measured accurately suggests that additional sequence positions are modulating the binding strength of PDZ–peptide interactions. This is supported by previous work showing that sets of non-binding positions coupled with a binding site contribute to the binding energy (Lockless and Ranganathan, 1999). It has also been found that mutations in these sites may affect the structure of the binding site and thus alter binding affinity (Lockless and Ranganathan, 1999). Although our best predictor was obtained using the full PDZ sequence, the performance was only somewhat improved on average compared with using either 10 or 16 binding positions. This may be due to the limitation of our sequence-based approach that fails to capture structural features of PDZ domains and their ligands that are important for binding. Additional experimental data about how structural variation in the binding site combined with affinity data would be useful in the future to further address the importance of non-binding site positions on affinity.

The binding strength of domain–peptide interactions may also be affected by the presence of other partners bound. It will be interesting to examine the potential competition of PDZ binding sites bound by multiple PDZ domains expressed at different concentrations using our method. It will also be important to extend our method in the future to consider co-operativity (Gibson, 2009).

The performance of the SemiSVR depends on sequence similarity of test PDZ domains to those in the training set with sufficient binding peptides. The human Scribble PDZ domain we tested is fairly close to domains in the training set (94% similar), and thus is a good test of our approach. Because of this, we expect our method is immediately applicable to PDZ domains in multiple species that are close to the domains in our training set. We thus used our method to predict relative affinities for a set of reasonably close mouse and human PDZ domains (<60% domain sequence identity) to putative mouse and human PDZ ligands and included it as a convenient starting set (see Supplementary Tables S7 and S8), which is useful for prioritizing future experiments.

Our results highlight the need to collect experimental domain–peptide binding data covering PDZ sequence space to improve prediction methods. This means we need to measure
affinities for domains that are less sequence related to those with known peptide affinities. We also need more affinity data from other species to make a more general conclusion about cross-species generality. Our future work will include incorporating more quantitative and qualitative interaction data from multiple sources into a prediction model to improve performance. For example, it might be possible to use phase display data to improve coverage and performance, which only includes qualitative positive PDZ domain–peptide pairs (Tonkian et al., 2008). Furthermore, quantitative prediction can potentially be improved by considering additional information about the domain and peptide, such as the 3D structure features of PDZ domains (Hue et al., 2010; Stein and Aloy, 2010; Beuming et al., 2009) and co-evolving residues (Halabi et al., 2009) in PDZ domain–peptide pairs. We plan to further develop our method along these lines and hope to increase its utility and accuracy in predicting quantitative interactions involving PDZ domains and apply it to other peptide recognition modules.

ACKNOWLEDGEMENTS

We thank D. Gfeller, S. Hui and L. Li for useful comments on the manuscript, R. Ammar and other Bader lab members for helpful suggestions and discussions and J.R. Chen for providing the backfitting algorithm code and the PDZ domain–peptide data.

Funding: X.S. is supported by a scholarship from the Chinese Scholarship Council. N.D. is supported by the key project of the National Nature Science Foundation of China (no. 10631070) and the National Nature Science Foundation of China (no. 10971223, 11071252). This project was supported by the Canadian Institutes of Health Research (MOP-84324).

Conflict of Interest: none declared.

REFERENCES


