BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 13 2009, pages 1632–1639
doi:10.1093/bioinformatics/btp264

Structural bioinformatics

Domain Interaction Footprint: a multi-classification approach to predict domain–peptide interactions

Christian Schillinger[1,2], Prisca Boisguerin[3] and Gerd Krause[1,*]

[1]Leibniz Institute for Molecular Pharmacology, Robert-Roessle-Str. 10, Berlin, [2]FU-Berlin, Department of Biology, Chemistry and Pharmacy, Takustr. 3, 14195 Berlin and [3]Institute of Medical Immunology, Charite-Universitaetsmedizin, Hessischestrasse 3-4, 10115 Berlin, Germany

Received on October 7, 2008; revised on April 1, 2009; accepted on April 15, 2009

Advance Access publication April 17, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: The flow of information within cellular pathways largely relies on specific protein–protein interactions. Discovering such interactions that are mostly mediated by peptide recognition modules (PRM) is therefore a fundamental step towards unravelling the complexity of varying pathways. Since peptides can be recognized by more than one PRM and high-throughput experiments are both time consuming and expensive, it would be preferable to narrow down all potential peptide ligands for one specific PRM by a computational method. We at first present Domain Interaction Footprint (DIF) a new approach to predict binding peptides to PRMs merely based on the sequence of the peptides. Second, we show that our method is able to create a multi-classification model that assesses the binding specificity of a given peptide to all examined PRMs at once.

Results: We first applied our approach to a previously investigated dataset of different SH3 domains and predicted their appropriate peptide ligands with an exceptionally high accuracy. This result outperforms all recent methods trained on the same dataset. Furthermore, we used our technique to build two multi-classification models (SH3 and PDZ domains) to predict the interaction preference between a peptide and every single domain in the corresponding domain family at once. Predicting the domain specificity most reliably, our proposed approach can be seen as a first step towards a complete multi-domain classification model comprised of all domains of one family. Such a comprehensive domain specificity model would benefit the quest for highly specific peptide ligands interacting solely with the domain of choice.

Contact: gkrause@fmp-berlin.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Protein–protein interactions are often mediated through protein domains that function as peptide recognition modules (PRMs), which are small (50–150 amino acids), independently folding domains that bind linear peptides and are found repeatedly in several protein structures (Pawson and Nash, 2003). The PRMs of different families (PDZ, SH3, SH2, WW, PTB, 14-3-3, etc.) are involved in a variety of cellular processes and are therefore an essential key of cell physiology as a whole. Within a family all PRMs share structural commonalities, but differ slightly at specific positions that entails specific binding partners. Identification of interaction partners to different PRMs can immensely benefit in unravelling biochemical pathways. The interaction partners are usually recognized by a short amino acid sequence that contains specific characteristics. In case of the SH3 domains, the binding pocket recognizes poly-prolin motifs of other proteins. PDZ domains always bind to the last four C-terminal residues of their interaction partner. Hence, the prediction of a potential ligands is possible in principle, since the binding mechanisms are very well understood (Feng *et al*., 1994, Songyang *et al*., 1997).

Several methods have been established to elucidate a common pattern of several peptides (Nevill-Manning *et al*., 1998, Schneider and Stephens, 1990, Timothy *et al*., 2006). A good classification method should be able to classify a peptide according to previously learned rules. These rules are subsequently used to classify unknown peptides. Doing this accurately could save expensive and time-consuming experiments. Since PRMs of the same domain family have generally a virtually similar pattern of their peptide ligands, classifying peptides by means of manual inspection is often ambiguously and therefore error-prone. Tong *et al*. (2002) for instance, analysed different SH3 domains and constructed a specific consensus motif of each domain. Those motifs bear such a great resemblance to each other that a manual peptide classification task is all but impossible, since there are peptides that match to all such motifs at once. And as there is usually no classification algorithm coming along with those motifs, one has to do the classification manually. As more interaction data evolved from various experiments, efforts were made to cluster domains in classes and characterize those. PDZ domains, for example, can roughly be divided into four classes based on their ligand specificity (Boisguerin *et al*., 2004, Reina *et al*., 2002, Songyang *et al*., 1997)—class 1 peptides are defined by consensus sequence motif ([S/T]XΦCOOH), class 2 by ([Φ/Ψ]XΦCOOH), class 3 by ([D/E]XVCOOH) and class 4 by (XΨ[D/E]COOH), respectively (Φ: hydrophobic, Ψ: aromatic, COOH: C-terminal carboxylic acid of the protein). But many PDZ domains target in fact a more comprehensive ligand sequence space. That is, to say the set of all recognized ligands by a domain contains also peptides whose sequences does not match the previous stated classification

*To whom correspondence should be addressed.

scheme and cannot categorized by simple rules at all (Vaccaro and Dente, 2002). Hence, we need a computational method that uses information beyond ordinary sequence motifs to circumvent manual classification.

There exist several methods to predict domain–peptide interactions whose predictive capabilities have all been tested with the phage display dataset of Tong *et al*. (2002) and are thus comparable. Brannetti *et al*. (2000) first incorporated structural information of a domain by SH3-specific matrices. This matrices can be derived from SH3/peptide complexes of the Protein Data Bank (Bernstein *et al*., 1997) and were used to establish a residue–residue contact database for predicting domain–peptide interactions. The machine learning method of Ferraro *et al*. (2006) relies also on the domain–peptide contact residues in complexes of known structures and enhances the contact matrix approach by a neural network and an appropriate encoding of the interacting residues. With it the accuracy of prediction of domain–peptide interactions could further be improved. Nevertheless, a necessary prerequisite of both methods is the necessity of structural information. Even though both of them show a generalization ability to a certain degree, those techniques are not applicable to cases where not very reliable structure is available. Contrary to this dependency, Reiss and Schwikowski (2004) developed a probabilistic generative model of the SH3 ligand peptides which is based on a modified Gibbs motif sampler (Lawrence, 1993). They were able to identify ligand peptides of PRMs by combining protein sequences and physical interaction data what makes their approach suitable for PRMs without any known structure. Lehrach *et al*. (2005) replaced the generative approach of Reiss and Schwikowski (2004) by a Laplacian-regularized discriminative model and reached together with Ferraro *et al*. (2006) the best predictive power aimed at inferring the domain recognition specificity of the SH3 phage display dataset so far. Another sequence-only method is the related, Hidden Markov Model (HMM)-based work of McLaughlin *et al*. (2006). For a more comprehensive overview of computational methods, see Shoemaker and Panchenko (2007) and Lee *et al*. (2007).

Aside from the prediction of domain–peptide interactions, it is of great interest to infer specific ligand peptides that preferably interact solely with one domain despite the similar consensus patterns of other PRMs. This demands a multi-classification approach that can reliably classify a peptide with respect to several PRMs of the same domain family.

In this article, we introduce Domain Interaction Footprint (DIF) as a new sequence-based approach to predict domain–peptide interactions and show that its performance to separate binding and non-binding peptides is better than those of any other method. Additionally, we show that our technique can be applied to multi-classification tasks where it assigns a given peptide to its appropriate PRM out of a set of domains that all share a very similar consensus pattern of their ligand peptides.

2 METHODS

2.1 Datasets

The SH3 dataset is composed of 25 experimental result sets whereas each one of those contains the corresponding binding and non-binding peptides of one SH3 domain that were identified in phage display by Tong *et al*. (2002).

The examined peptides can be divided into classes 1 and 2, whereas class 1 is defined by the consensus sequence motif (+xΦPxΦP) and class 2 by (ΦPxΦPx+), where Φ is a hydrophobic and + is a basic residue (Mayer, 2001). Some domains interact not only with class 1 peptides, but also with class 2 ones and are therefore listed twice in the dataset as two different experiments have been carried out. An overview of the used domains as well as the number of binding and non-binding peptides is assembled in Table 1. Further information about this dataset can be obtained in the supplementary material of Tong *et al*. (2002). In order to use the provided sequence information with our approach, we prolonged shorter sequences with a non-defined value *X* so that all sequences are of the same length. This value has no impact in further steps as it is simply being ignored.

Table 1. Experimentally verified binding and non-binding peptides determined for different SH3 and PDZ domains

Domain	Binding peptides	Non-binding peptides
SH3		
Boi1 Class 1	3	16
Boi1 Class 2	6	11
Boi2 Class 1	8	9
Bzz1-1 Class 1	14	17
Bzz1-2 Class 1	13	17
Myo3 Class 1	7	21
Myo5 Class 1	13	21
Nbp2 Class 1	16	19
Pex13 Class 1	10	12
Pex13 Class 2	16	12
Rvs167 Class 1	11	10
Rvs167 Class 2	16	5
Sho1 Class 1	18	12
Sla1-3 Class 1	8	18
Yfr024 Class 1	7	17
Yfr024 Class 2	22	6
Ygr136 Class 1	18	14
Ygr136 Class 2	15	9
Yhl002 Class 1	9	14
Yhr016 Class 1	6	16
Yhr016 Class 2	11	6
Yjl020 Class 1	4	18
Yjl020 Class 2	11	11
Ypr154 Class 1	23	11
PDZ		
AF6	16	
ERBIN	19	
SNA1	15	
N1P1	9	

Additionally, we used also a PDZ dataset that consists of binding and non-binding peptides of four different PDZ domains [AF6 (Protein AF-6, UniProt P55196), SNA1 (Protein Syntrophin-1, UniProt Q13424), ERBIN (Protein LAP2, UniProt Q96RT1), N1P1 (NHERF1 PDZ 1, UniProt O14745)] to create a multi-classification model as all involved peptides have been experimentally tested against every PDZ domain (Supplementary Material, Table S1). The experimental verification of single-binding peptides makes it possible to build a multi-dimensional classifier with disjunctive sets of peptides.

2.1.1 Peptide array synthesis and incubation The peptides tested with the PDZ domains were synthesized on N-modified CAPE-membrane (Bhargava *et al*., 2002) and prepared with a MultiPep SPOT-robot (INTAVIS Bioanalytical Instruments AG, Germany). Array design was performed using the inhouse software LISA 1.71. Peptide arrays for the AF6-, ERBIN- and SNA1-PDZ domain arrays were generated and incubated as described in

Boisguerin *et al*. (2004) whereas peptide arrays for N1P1 were synthesized using the improved method of inverted peptides as described in Boisguerin *et al*. (2007). N1P1 incubations were performed analogous to the other PDZ domains with a concentration of 10 g/ml and a detection system consisting in anti-His(mouse) (1:3000, Sigma)/anti-mouse-HRP (1:2600, Calbiochem).

2.2 Domain Interaction Footprint

A DIF is related to one specific protein recognition module and can be seen as an advanced description of the most important properties of all peptides with whom the PRM interacts with. Each given peptide contributes with its properties partly to the model of the investigated domain interaction pattern.

Our approach of building a predictive domain interaction model used for sequence-function analysis consists of three major steps (Fig. 1) . First, we encode our data to be utilizable by our algorithm. This is done by using specific property-based numerical values for amino acids, so-called amino acid indices (Kawashima and Kanchisa, 2000). To keep our approach as general as possible, we do not assume any previous knowledge of the examined domain. On account of this, we do not make a preliminary selection of amino acid's properties, but let our algorithm choose the most meaningful of these indices by correlation-based feature selection. Thus, we capture the specific and appropriate properties for each domain separately. Eventually, we create a DIF by combining descriptors that were built with the empirically observed range of the selected properties.

2.2.1 Sequence encoding The structural information and therefore the function of a protein is due to the combination of 20 different amino acids. Even though no explicit rule has yet been derived to conclude a protein's 3D structure merely from its sequence, the shape is undoubtedly based on the underlying characteristics of the amino acids in terms of volume, charge and hydrophobicity among others. The approach relies purely on the sequence analysis of the peptides in relation to the biological effect. The sequence of the PRM is not considered.

In our approach, we try to capture the functional aspect of a peptide by its amino acid sequence. For this purpose, we encode each amino acid by specific physiochemical and biochemical properties. There are several parameters such as logP, Verloop parameters for volume, parameters for hydrophobicity, for polarization, for frequency of occurrence in secondary structure elements, for flexibility, for surface description, etc. The current database of Kawashima and Kanchisa (2000) contains 544 such amino acid indices that we all take into consideration. A formal depiction follows.

Let $S=(s_1,\cdots,s_{n_s})$ be a list of sequences, with n_s as the total number of sequences. Each sequence $s_i=(a_{i,1},\cdots,a_{i,n_a})$ is in turn composed of n_a amino acids. As amino acids can be described by several specific properties and characteristics, we define a list of n_p different properties by $P=(p_1,\cdots,p_{n_p})$. Encoding the amino acid sequences in S with properties P leads to an encoded sequence set $p(S)$ containing accordingly encoded sequences $p(s_i)=(p_1(a_{i,1}),\ldots,p_{n_p}(a_{i,n_a}))$, with a total of $n_e(=n_a\cdot n_p)$ features.

2.2.2 Feature adaptation As each PRM interacts with specific peptides, it is reasonable that all of these peptides share a common consensus pattern which in turn typifies an implicit interaction rule. To tackle the domain specificity, we capture the characteristic features of each PRM separately by scaling down the order of magnitude for the descriptor set to significant descriptors. The selection of the most relevant properties strongly reduces the number of relevant descriptors and provide a basis to model the mentioned implicit interaction rule.

There are $2^{n_e}-1$ possible subsets consisting of at least one feature that can be built with P. Taking an assumed peptide length of 10 into account as well as the 544 amino acid indices, it would result in about 2^{5440} different subsets. To find a preferably good descriptive model out of the vast number of subsets, we need to select the most meaningful features. For this purpose, we use correlation-based feature selection (CFS) to evaluate the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. The evaluation method

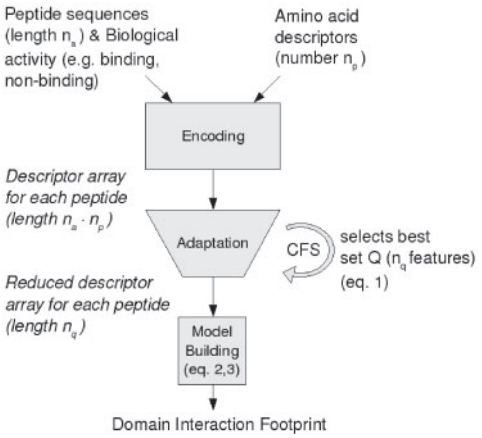

Fig. 1. Schematic overflow of the algorithm.

tries to select those features, whose values differ systematically between classes (i.e. binding and non-binding). Redundant features are eliminated since they are highly correlated with at least one of the remaining features (Kohavi and Sommerfield, 1995). The term correlation is not intended to refer specifically to classical linear correlation; rather it is used to refer to a degree of dependence or predictability of one variable (position-specific amino acid index) with another. We start with no feature and search through the search space by adding single features until five consecutive subsets show no improvement over the current best subset. The acceptance of a new feature depends on the extent to which it predicts classes in areas of the sequence space not already predicted by other features. To avoid exhaustive enumeration of all possible subsets, we use a confined search strategy as described by Langely (1994). This presumably best subset is labelled Q. The size of $Q(=n_q)$ is strongly reduced with regard to the initial number of features $(n_a\cdot n_p)$.

The core of CFS is given through the following subset evaluation function

$$M_A=\frac{d\overline{r_{cf}}}{\sqrt{d+d(d-1)\overline{r_{ff}}}} \tag{1}$$

where M_A is the heuristic merit of subset A containing d features, $\overline{r_{cf}}$ is the mean feature–class correlation $(f\in A)$ and $\overline{r_{ff}}$ is the average feature–feature intercorrelation. The CFS approach given in Equation (1) is, in fact, Pearson's correlation coefficient, where all variables have been standardized. A more detailed survey can be found in Hall (1999).

2.2.3 Model building Once an appropriate subset Q has been selected, we use it to create a classification model. Therefore, we define a function $f(l,z)$ that returns a value at position z from a list l and a Boolean function $b(y)$ that returns 1, if y is true and 0 otherwise. In addition, we define a position-based list $X(S,k)=(f(s_1,k),\ldots,f(s_{n_s},k)), k\in\mathbb{N}\wedge k\leq n_s$, that contains all list entries at a given position k. Let then

$$D_k=(\min((X(q(S),k)),\max((X(q(S),k))) \tag{2}$$

be a descriptor of the lowest and highest value contained in list $X(q(S),k)$, respectively. Hence, we can define a model $\mathrm{DIF}^{q(S)}$ that describes the range of all properties within a defined set of sequences as follows:

$$\mathrm{DIF}^{q(S)}=(D_1,\ldots,D_{|Q|}) \tag{3}$$

Such a model captures the observed scope of the position-specific amino acid properties of the provided data and can therefore be seen as a descriptive pattern of S. Based on such a model, it is rather simple to estimate how well an unknown sequence s_0 fits into those generalized DIF. For this purpose, we introduce a fitness function F [Equation (4)] that takes two arguments: a

Table 2. Averaged AUC and AUC01 values of our proposed DIF and other recently published methods applied to the SH3 dataset

Method	AUC	AUC01
Neural network PAIRs[a]	0.83	n.a.
Informatively initialized discriminative[b]	0.83	0.44
Network-based motif sampler[c]	0.79	0.17
SH3-SPOT[d]	0.76	n.a.
PSSM[e]	0.64	n.a.
DIF	0.95	0.81

The AUC (AUC01) score was averaged twice: once over all cross-validation experiments and again over all SH3 domains. An AUC score of 0.5 reflects utter randomness, while a value of 1 implies perfect prediction; AUC01 represents the calculated AUC up to a false positive rate of 0.1 (the calculated AUC01 value is stated as a fraction of the maximum area 0.1).
n.a., not available.
[a]Ferraro *et al*. (2006).
[b]Lehrach *et al*. (2005).
[c]Reiss and Schwikowski (2004).
[d]Brannetti *et al*. (2000).
[e]Position-specific scoring matrix built by Ferraro *et al*. (2006).

peptide and a DIF the peptide will be evaluated to. The possible outcome of F is between 0 and 1, whereas 0 stands for no commonness at all and 1 for a perfect match.

$$F(s^0, \mathrm{DIF}^{q(S)}) = \frac{\sum_k b(cond_1) \cdot b(cond_2)}{k} \quad (4)$$

$$cond_1 = f(q(s^0), k) \geq f(f(\mathrm{DIF}^{q(S)}, k), 1) \quad (5)$$

$$cond_2 = f(q(s^0), k) \leq f(f(\mathrm{DIF}^{q(S)}, k), 2) \quad (6)$$

We used this methodology to build a series of classifiers. For each investigated SH3 domain a binary classifier was created to discriminate between binding and non-binding peptides. In that cases, the classifier was composed of two DIFs, one for the binding and one for the non-binding peptides. A peptide s^0 is then assigned to the DIF with the highest score.

As there is no limitation in creating DIFs of different datasets, a multi-classification approach can comprise as many DIFs as wanted. But notice that Step 2, feature adaption, will select different properties with regard to a different input. We combined binding data of all PDZ domains to build one multi-classifier and selected appropriate amino acid properties by presenting all datasets at once. Hence, this multi-domain model can assess to which PRM a given peptide binds most probably. We also created a multi-domain model for all SH3 domains.

2.3 Method evaluation

In order to compare the results, our analysis was performed in the same way as the other methods. Like the other approaches, we used a 10-fold cross-validation to evaluate each DIF. Therefore, the dataset is divided into 10 subsets of equal size, whereas the ratio of class sizes is kept the same. Then 10 experiments were performed in which nine distinct subsets were used to build a model and the remaining one to evaluate it. Subsequently, the average performance was calculated out of the 10 results. Furthermore, we randomized the input data in different runs to attain unbiased results. In case of multi-classification, we created a DIF for each domain and assigned an examined peptide to the domain with the highest fitness score [Equation (4)].

In order to compare the performance of the introduced approach to the recently published methods NN-PAIRs, PSSM (Ferraro *et al*., 2006), Network-based Motif Sampler (Reiss and Schwikowski, 2004), Informatively Initialized Discriminative (Lehrach *et al*., 2005) and SH3-SPOT (Brannetti *et al*., 2000), we used like the other authors the area under the receiver operating characteristic (ROC) curve (AUC) (Bradley, 1997) to evaluate the accuracy with one single measure (Table 2). An AUC of 0.5 reflects random prediction, while AUC = 1 implies perfect prediction. The final AUC score is averaged over all 10 runs of the cross-validation. The AUC value of the multi-classification model refers to the average score of all the possible pairs of class combinations. Additionally, we calculated the AUC up to 0.1, where the number of false positives is low, since the upper range of a ROC curve is largely irrelevant for useful predictions. Note that the maximum AUC up to a false positive rate of 0.1 is 0.1 and that the calculated areas are stated as a fraction of this. We refer to this score as AUC01. Even so AUC01 is more appropriate to assess the prediction performance, we use both scores since AUC01 is not available for all of the other approaches.

The calculation of AUC as well as the feature selection were carried out by adapting and using the Weka framework (Holmes *et al*., 1994).

3 RESULTS

3.1 Model building

The prerequisite of our approach is the binding information of an examined PRM in terms of the sequences of its peptide ligands. To build up a classifier we need at least two sets of binding information that can be derived either from another PRM or from non-binding peptides.

Building the actual DIF of a PRM consists of three major steps. First, our algorithm encodes the sequences of the binding peptides with all available amino acid indices which results in a large multi-dimensional feature space that numerically describes all known properties of the amino acids of the sequences. Next, the algorithm adapts this feature space specifically to one or more given PRM. Therefore, it selects the most relevant features by eliminating redundancy within one set of binding information and high intercorrelation between itself and all other sets. Eventually, the remaining position-dependent properties are used to build one DIF for each examined set that are subsequently combined to one classifier.

3.2 Single-domain model

To compare the performance of our algorithm with recent methods, we referred for the evaluation of our single-domain approach to a well-known dataset. We used the SH3 phage display dataset of Tong *et al*. (2002) and validated our method as previously done by others (see Section 2 for more details). All prerequisites of our algorithm as aforementioned were fulfilled since sets of binding and non-binding peptides are available. We used this information to create 25 binary classifiers one for each dataset and calculated the average AUC value (Supplementary Material, Table S2).

Our result of AUC = 0.95 (AUC01 = 0.81) outperforms all other methods (Table 2). Using the identical dataset the performance of the approach of Ferraro *et al*. (2006) and the method of Lehrach *et al*. (2005) (AUC01 = 0.44) can both be stated with an average AUC of 0.83. The approach of Reiss and Schwikowski (2004) attains AUC = 0.79 (AUC01 = 0.17), which is by and large comparable with the result of Brannetti *et al*. (2000) (AUC = 0.76). A position-specific scoring matrix reaches only an average of 0.64.

3.3 Multi-domain model

Since interaction partners of PRMs can be described by a sequence consensus motif and those are virtually similar between PRMs of the same domain family, it is even more challenging to discriminate between binding peptides of two similar PRMs than to discriminate

Table 3. Performance of the two multiple domain models DIF-SH3 and DIF-PDZ

Model	AUC
DIF-SH3	**0.88**
M1DIF-Boi1c2	0.95
M1DIF-Boi2c1	0.97
M1DIF-Bzz1-1c1	0.96
M1DIF-Bzz1-2c1	0.78
M1DIF-Myo3c1	0.87
M1DIF-Myo5c1	0.80
M1DIF-Nbp2c1	0.98
M1DIF-Pex13c2	0.97
M1DIF-Rvs167c1	0.93
M1DIF-Rvs167c2	0.93
M1DIF-Sho1c1	0.77
M1DIF-Sla1-3c1	0.96
M1DIF-Yfr024c1	0.97
M1DIF-Yfr024c2	0.94
M1DIF-Ygr136c1	0.73
M1DIF-Ygr136c2	0.72
M1DIF-Yhl002c1	0.76
M1DIF-Yhr016c1	0.88
M1DIF-Yhr016c2	0.92
M1DIF-Yjl020c2	0.97
M1DIF-Ypr154c1	0.92
M1DIF-Ypr154c2	0.84
DIF-PDZ	**0.89**
M2DIF-AF6	0.91
M2DIF-ERBIN	0.93
M2DIF-SNA1	0.84
M2DIF-N1P1	0.87

They comprise the binding information of the SH3 and PDZ datasets, respectively. The AUC value of DIF-SH3 (DIF-PDZ) (in bold) is the mean of all M1DIFs (M2DIFs) . M1DIFs and M2DIFs are part of the DIF-SH3 and DIF-PDZ multi-classification model, respectively. Their corresponding AUC score is attained by a one-against-all measurement.

between binding peptides and non-binding peptides at large as the latter not necessarily share a common motif.

We attained remarkably results with both a PDZ and a SH3 multi-domain model. The average performance of our PDZ multi-domain model is AUC = 0.89 (AUC01 = 0.81), whereas the AUC values of the single classifiers range from 0.84 to 0.93 (Table 3). As even the lowest AUC score of the AF6 model can be seen as a very good result of overall accuracy, the classifier on the whole assures a strong prediction capability. The AUC values of the individual classifiers of the SH3 multi-domain model range from 0.72 to 0.98 with an overall average of 0.88.

4 DISCUSSION AND CONCLUSION

We developed a sequence-based approach to predict the interaction of PRMs with peptides. The approach extracts for this purpose the major amino acid properties responsible for the domain–peptide interaction and generates a DIF. Used with knowledge of non-binding peptides, such a DIF can be used as a general binary classification method to distinguish between binding and non-binding peptides. Additionally, we derived a multi-classification model that is able to discriminate between peptides binding to different domains of the same family by incorporating several datasets at once into our algorithm. We want to emphasize that no additional structural information of the domain itself is needed and no adjustment of any parameter need to be made at any time. The algorithm is therefore easy to use and can be applied to all sort of domains very quickly.

We applied our single-domain method to a well-known SH3 domain dataset as well as to a dataset of PDZ domains. The former one was chosen to evaluate our algorithm, because it has previously been used by other approaches (Brannetti *et al*., 2000, Ferraro *et al*., 2006, Lehrach *et al*., 2005, Reiss and Schwikowski, 2004) what makes our result comparable. We employed both datasets to build two different multi-domain models, one for the SH3 and one for PDZ domains, which allowed us to predict the interaction of a given peptide to all investigated domains, either PDZ or SH3, at once.

The achievement of our single-domain method applied to the SH3 dataset is very promising as it outperforms all other methods compared to. While reaching an average AUC score of 0.95 and an AUC01 score of 0.81, both the method of Ferraro *et al*. (2006) and Lehrach *et al*. (2005) (AUC01 = 0.44) reached an AUC score of 0.83 with the identical dataset, whereas the proposed algorithms of Brannetti *et al*. (2000) and Reiss and Schwikowski (2004) achieved only 0.76 and 0.79 (AUC01 = 0.17), respectively. Such a good result is remarkable, since we do not use any inferred information of the domain's structure directly as the methods of Ferraro *et al*. (2006) and Brannetti *et al*. (2000) do. Without such structural information, the method is applicable to cases where no hints of the 3D structure are known. Nevertheless, we capture important aspects of the domain's structure indirectly by its propensity toward certain properties of its peptide ligands. The identified properties' range of observed amino acids at a fixed peptide position can be seen as a mirror image representation of the counterpart properties of the surrounding PRM. In other words, each residue participates with its own value in creating the domain's mould, or footprint. Seeing only small volume residues at a particular position of the peptide, for example, can be interpreted as a spatial constraint due to a narrow cleft of the bulky surrounding of the PRM. We take not only volume constraints into consideration but also any other properties represented in the amino acid index database of Kawashima and Kanchisa (2000) like electrostatics, charges, hydrophobicity, etc. The achieved good performance and improvement compared with other approaches is likely to be attributed to the incorporation of those encoded amino acids. Defining a scope of favourable values for specific features is disparate to methods using frequencies of observed amino acids. Whereas the latter need many samples to infer a deductive scheme, the former are better suitable to interpolate from a low sample size.

One natural concern of machine learning techniques is always overfitting. Therefore, we took the same line as the other approaches that we compared our method to and used cross-validation as it was done before. Those approaches have shown successfully that overfitting is not a problem, even though they utilize up to 57 parameters. As a result, we conclude that our technique is legitimate as it uses comparable or less parameters in 23 out of 25 cases (Supplementary Material, Table S4).

Additionally, we merged the binding peptides of both domains with the same length into one dataset and measured the discriminative ability of our method on that combined dataset. The classifier is based upon three features and yielded as expected

Table 4. Sequence motifs of peptide ligands for different PDZ domains derived

	Position			
Domain	−3	−2	−1	0
AF6	∘, Φ, Ψ, +	∘, Φ, Ψ	∘, Φ, Ψ	∘, Φ, Ψ
ERBIN	∘, Φ, +	∘, Ψ	∘, Φ, Ψ, +	∘, Φ, Ψ, −
SNA1	∘, Φ, +, −	∘, Φ	∘, Φ	∘, Φ, +, −
N1P1	∘, Φ, +, −	∘, Φ, Ψ, −	∘, Φ, Ψ, −	∘, Φ, Ψ, +, −

Amino acids of the peptide ligands are encoded as follows: Ψ (aromatic): F, Y, W; Φ (hydrophobic): V, L, I, M, C, A; + (positive): D, E; − (negative): R, K, H; ∘ (others): N, Q, P, G, S, T. Ligand positions are numbered in reverse from the very last C-terminal ligand residue, which is denoted as 0.

Table 5. Advanced sequence motif of used amino acid indices to build descriptors of the PDZ multi-classifier

Peptide position	Accession No.	Description
0	ZASB820101	Hydrophobicity scale
0	QIAN880113	Alpha helix propensity
0	RACS820114	Value of theta (i−1)
0	AURR980112	Helix capping
−1	LEVM760103	Side chain angle theta (AAR)
−1	AURR980110	Helix capping
−1	TANS770109	frequency of coil
−1	ZASB820101	Hydrophobicity scale
−1	RICJ880108	Alpha helices preference
−1	PONP800105	Hydrophobic packing
−1	LEVM760101	Hydrophobic parameter
−1	FAUJ880103	van der Waals volume
−1	BIGC670101	Residue volume
−2	PONP800104	Surrounding hydrophobicity
−2	CHOP780205	Helix propensity
−3	MAXF760106	Frequency of alpha region
−3	HUTJ700103	Entropy of formation
−3	GEIM800103	Alpha-helix indices
−3	CHOP780208	Beta-sheet propensity

Accession number can be used to obtain more details of the corresponding publication at http://www.genome.jp/aaindex.

a very good result (AUC = 0.95), since the peptides represent completely different binding mechanisms, which is quite similar to a binding/non-binding classification. Details can be found in the Supplementary Material (Table S3).

The structure-based method of Ferraro *et al*. (2006) was also applied to a larger pep-spot dataset of 7327 tested peptides (Landgraf *et al*., 2004). Using that bigger dataset improves the neural network approach to an accuracy of AUC = 0.92. This is due to more training instances where the net can learn from as the method is based on interacting residues of the peptide and the domain. The relevance of such an interacting pair in the peptide–domain complex is weighted by its frequency. More pairs in terms of a bigger dataset leads therefore to a more accurate model. The smaller SH3 phage display set entails a worse performance. Hence, our single-domain approach is more favourable on small and medium datasets.

Although our approach is very robust some prerequisites should be considered prior an application. The method is designated when several verified peptide interaction partners exist as training set and is therefore not suitable in cases where only one domain–peptide complex is available.

Furthermore, we created sequence motifs of the peptide ligands of the PDZ domains and show that there is no general differentiation between them (Table 4). No domain can be described by an unique pattern. There are only very few features that in some cases can separate the motif of one PRM from the others. A peptide, for instance, with a negative charged amino acid (Arg, His, Lys) at position −2 (positions are counted backwards, beginning at 0) interacts only with the N1P1 domain according to the current dataset. Peptides with a positive charged amino acid (Asp, Glu) at position −1 are unique of peptide ligands of SNA1.

These examples show the rationale of highlighting position-based properties that can be used for classification, albeit no trivial pattern exists to classify all peptides. We take this idea further and describe an amino acid in greatest detail by all recently known properties. After selecting the most distinguishable properties into an advanced sequence motif (Table 5), we have built a classifier that can reliably predict the interaction partner of the different PDZ domains.

To build a multi-domain model, binding information of each participating domain for a given peptide ought to be available, otherwise the prediction cannot be evaluated unequivocally. That is why we chose the PDZ dataset as an example for the advanced sequence motif, since the data used for the SH3 multi-domain model cannot be stated as collectively exhaustive. In other words, there is in the case of the SH3 dataset at least one peptide *P* that binds to domain *A*, but we have no information whether *P* binds to domain *B* as well (nor have we information whether *P* not binds to *B*). That poses the question how we are supposed to handle any prediction concerning *P* and *B*. If our approach predicts that peptide *P* binds to domain *B*, we have no way to verify or falsify that prediction. In case of a simple binary class decision problem of one domain, we are not facing this sort of problem as we possess mutually exclusive information of all available peptides. This changes, of course, if we incorporate peptides from another domain, where we have no information regarding the current examined domain. Nevertheless, we combined the SH3 datasets into one multi-domain model and received good results (Table 3). We, therefore, conclude that the binding information of Tong's dataset is mutually exclusive albeit this is not explicitly stated.

The PRMs of the PDZ domain family show a virtually similar consensus pattern of their binding peptides (Table 4), however, our multi-domain method achieves a very good result (PDZ model: AUC = 0.89). We suggested an advanced sequence motif made up of varying position-based amino acid characteristics (Table 5) as a possible discriminator to separate the peptide ligands of one PRM from all the others. Note that in our approach the essential properties like hydrophobicity is not only described by one parameter (Φ in Table 4), but by four different hydrophobic parameters such as hydrophobicity scale, hydrophobic parameter, surrounding hydrophobicity and hydrophobic packing (Table 5). The more detailed description of volume and steric properties by the essential side-chain parameters value of theta (i−1), side-chain angle theta (AAR), van der Waals volume and residue volume (Table 5) also contribute to the improved performance. In contrast to binary classifications where consensus patterns are extracted to accentuate

differences between classes or common major characteristics of one class, such a multi-class discriminator can not only focus on one class but has to be composed of features whose presence or absence assign the class belonging. Subsequently, such an advanced motif emphasizes the characteristics' variety which is comparatively referred to. Each of the examined PDZ domains can therefore be defined by its preference for a peculiar specificity of that motif.

Such a multi-classification model can eventually improve the search for interaction partners preferably binding to the domain of choice. The complex pattern used for classification might be used to search for new potential interaction candidates in sequence databases. There are just few steps needed. First, the whole sequences are splitted according to the size of peptides the DIF was trained with. After that the descriptor of the DIF is used to transform the encoded sequence in an evaluable form. Finally, the evaluation takes place. Besides, it can be used to narrow down the complete space of possible peptides for synthesis. Such an approach might also be useful in suggesting sequences for a specific peptide library that can be used by experimentalists. Although regulation of PDZ domain-mediated interactions has been a major focus over the last 10 years, only a handful of reports describe negative regulation involving phosphorylation of ligand residues in position −2 (Chung *et al*., 2004, Cohen *et al*., 1996, Chetkovich *et al*., 2002) position −3 (Matsuda *et al*., 1999, Chung *et al*., 2000) or position −5 (Tian *et al*., 2006). The phosphorylated amino acids at positions −2 and −3 could be simulated with aspartic acid such as reported by Cohen *et al*. (1996). Due to the fact that we used the whole amino acid set during the experiment, we could also take into consideration potential phosphorylation and therefore incorporate the result in the search for new binding proteins *in vivo*. Generally, our proposed method can be used in the analysis of any other PRM. Predicting potential peptide ligands for peptide recognition modules would benefit the understanding of biological networks as new interaction partners would enhance our knowledge of existing pathways and system biology of the cell. Considering the good performance of our proposed method, costs of experiments can be reduced by proposing a selection of preferably good candidates at which the examination can focus on.

ACKNOWLEDGEMENTS

We thank D.R. Madden (Department of Biochemistry, Dartmouth Medical School, Hanover, NH03755) for providing the N1P1 PDZ domain.

Funding: Research Group 806 of the Deutsche Forschungsgesellschaft (DFG); Deutsche Forschungsgemeinschaft (DFG Grant VO885/3-1 to P.B.).

Conflict of Interest: none declared.

REFERENCES

Bernstein,F.C. *et al.* (1997) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Bhargava,S. *et al.* (2002) A complete substitutional analysis of VIP for better tumor imaging properties. *J. Mol. Recognit.*, **15**, 145–153.

Boisguerin,P. *et al.* (2004) An improved method for the synthesis of cellulose membrane-bound peptides with free C termini is useful for PDZ domain binding studies. *Chem. Biol.*, **11**, 449–459.

Boisguerin,P. *et al.* (2007) Characterization of a Putative Phosphorylation Switch: Adaptation of SPOT Synthesis to Analyze PDZ Domain Regulation Mechanisms. *Chembiochem.*, **8**, 2302–2307.

Bradley,A.P. (1997) The use of the area under ROC curve in the evaluation of the machine learning algorithms. *Pattern Recognit.*, **30**, 1145–1159.

Brannetti,B. *et al.* (2000) SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family. *J. Mol. Biol.*, **298**, 313–328.

Chetkovich,D.M. *et al.* (2002) Phosphorylation of the postsynaptic density-95 (PSD-95)/Discs Large/Zona Occludens-1 binding site of stargazin regulates binding to PSD-95 and synaptic targeting of AMPA receptors. *J. Neurosci.*, **22**, 5791–5796.

Chung,H.J. *et al.* (2000) Phosphorylation of the AMPA receptor subunit GluR2 differentially regulates its interaction with PDZ domain-containing proteins. *J. Neurosci.*, **20**, 7258–7267.

Chung,H.J. (2004) Regulation of the NMDA receptor complex and trafficking by activity-dependent phosphorylation of the NR2B subunit PDZ ligand. *J. Neurosci.*, **24**, 10248–10259.

Cohen,N.A. *et al.* (1996) Binding of the inward rectifier K channel Kir 2.3 to PSD-95 is regulated by protein kinase A phosphorylation. *Neuron*, **17**, 759–767.

Feng,S. *et al.* (1994) Two binding orientations for peptides to the Src SH3 domain: development of a general model for SH3-ligand interactions. *Science*, **266**, 1241–1247.

Ferraro,E. *et al.* (2006) A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity. *Bioinformatics*, **22**, 2333–2339.

Hall,M.A. (1999) Correlation-based feature subset selection for machine learning. PhD thesis. Department of Computer Science, University of Waikato. Available at http://www.cs.waikato.ac.nz/~mhall/thesis.pdf.

Holmes,G. *et al.* (1994) Weka: a machine learning workbench. In *Proceedings of the 1994 Second Australia and New Zealand Conference on Intelligent Information Systems*, pp. 357–361. Available at http://hdl.handle.net/10289/1138

Kawashima,S. and Kanchisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374.

Kohavi,R. and Sommerfield, D. (1995) Feature subset selection using the wrapper method: overfitting and dynamic search space topology. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*. Available at https://www.aaai.org/Papers/KDD/1995/KDD95-049.pdf

Landgraf,C. *et al.* (2004) Protein interaction networks by proteome peptide scanning. *PLoS Biol.*, **2**, e14.

Langely,P. (1994) Selection of relevant features in machine learning. In *AAAI Fall Symposium on Relevance*, AAAI Press, CA, USA, pp. 140–144.

Lawrence,C.E. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lehrach,W.P. *et al.* (2005) A regularized discriminative model for the prediction of protein-peptide interactions. *Bioinformatics*, **22**, 532–540.

Lee,D. *et al.* (2007) Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.*, **8**, 995–1005.

Matsuda,S. *et al.* (1999) Phosphorylation of Serine-880 in GluR2 by protein kinase C prevents its C terminus from binding with glutamate receptor-interacting protein. *J. Neurochem.*, **73**, 1765–1768.

Mayer,B.J. (2001) SH3 domains: complexity in moderation. *J. Cell Sci.*, **114**, 1253–1263.

McLaughlin,M.A. *et al.* (2006) Prediction of binding sites of peptide recognition domains: an application on Grb2 and SAP SH2 domains. *J. Mol. Biol.*, **357**, 1322–1334.

Nevill-Manning,C.G. *et al.* (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl Acad. Sci. USA*, **95**, 5865–5871.

Pawson,T. and Nash,P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **18**, 445–452.

Reina,J. *et al.* (2002) Computer-aided design of a PDZ domain to recognize new target sequences. *Nat. Struct. Biol.*, **9**, 621–627.

Reiss,D.J. and Schwikowski,B. (2004) Predicting protein-peptide interactions via a network-based motif sampler. *Bioinformatics*, **20**, 274–282.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Shoemaker,B.A. and Panchenko,A.R. (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PloS Comput. Biol.*, **3**, 595–601.

Songyang,Z. *et al.* (1997) Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science*, **275**, 73–77.

Tian,Q.B. *et al.* (2006) Interaction of LDL receptor-related protein 4 (LRP4) with postsynaptic scaffold proteins via its C-terminal PDZ domain-binding motif, and its regulation by Ca2+/calmodulin-dependent protein kinase II. *Eur. J. Neurosci.*, **23**, 2864–2876.

Timothy,L. *et al.* (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, 369–373.

Tong,A.H.Y. *et al.* (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **295**, 321–324.

Vaccaro, P.and Dente, L. (2002) PDZ domains: troubles in classification. *FEBS Lett.*, **512**, 345–349.

Wiedemann,U. *et al.* (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. *J. Mol. Biol.*, **343**, 703–718.