BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 19 2006, pages 2333–2339
doi:10.1093/bioinformatics/btl403

Structural bioinformatics

A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity

E. Ferraro*, A. Via, G. Ausiello and M. Helmer-Citterich

Centre of Molecular Bioinformatics, Department of Biology, University of Tor Vergata, Rome, Italy

Received on March 30, 2006; revised on July 13, 2006; accepted on July 19, 2006

Advance Access publication July 26, 2006

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Unravelling the rules underlying protein–protein and protein–ligand interactions is a crucial step in understanding cell machinery. Peptide recognition modules (PRMs) are globular protein domains which focus their binding targets on short protein sequences and play a key role in the frame of protein–protein interactions. High-throughput techniques permit the whole proteome scanning of each domain, but they are characterized by a high incidence of false positives. In this context, there is a pressing need for the development of *in silico* experiments to validate experimental results and of computational tools for the inference of domain–peptide interactions.

Results: We focused on the SH3 domain family and developed a machine-learning approach for inferring interaction specificity. SH3 domains are well-studied PRMs which typically bind proline-rich short sequences characterized by the PxxP consensus. The binding information is known to be held in the conformation of the domain surface and in the short sequence of the peptide. Our method relies on interaction data from high-throughput techniques and benefits from the integration of sequence and structure data of the interacting partners. Here, we propose a novel encoding technique aimed at representing binding information on the basis of the domain–peptide contact residues in complexes of known structure. Remarkably, the new encoding requires few variables to represent an interaction, thus avoiding the 'curse of dimension'. Our results display an accuracy >90% in detecting new binders of known SH3 domains, thus outperforming neural models on standard binary encodings, profile methods and recent statistical predictors. The method, moreover, shows a generalization capability, inferring specificity of unknown SH3 domains displaying some degree of similarity with the known data.

Contacts: enrico@cbm.bio.uniroma2.it

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

1 INTRODUCTION

Protein–protein interactions play an essential role in the regulation of cell physiology. Not only can the function of a protein be characterized more precisely through its interactions, but also networks of interacting proteins can shed light on the molecular mechanisms of cell life. The development of a large number of experimental techniques aimed at analysing protein–protein interactions (Ito *et al*., 2000; Uetz *et al*., 2000; Gavin *et al*., 2002; Ho *et al*., 2002; Tong *et al*., 2002; Aebersold and Mann, 2003; Zhu and Snyder, 2003; Landgraf *et al*., 2004), is giving rise to an increasing amount of data. Therefore, the need for validation procedures is becoming ever more pressing.

Several computational methods for the inference of protein–protein interaction have already been developed [see Pazos and Valencia (2002), Bork *et al*. (2004) and Russell *et al*. (2004) as reviews]. Such methods rely on various principles and make use of information on either genes (Gaasterland and Ragan, 1998; Pellegrini *et al*., 1999; Overbeek *et al*., 1999; Marcotte *et al*., 1999; Enright *et al*., 1999) or proteins (Pazos *et al*., 1997; Goh *et al*., 2000; Pazos and Valencia, 2001, 2002). A machine learning approach to the inference of protein–protein interactions is also possible. Milik *et al*. (1998) trained different neural networks to predict MHC-binding peptides using either binary encoding or a biochemical features representation of the amino acids sequence. Bock and Gough (2001) proposed a support vector machine (SVM) learning approach based on the primary structure of the interacting partners and on the physicochemical features of amino acids. A selection of structural and biophysical information has been collected by Zhao *et al*. (2003) to characterize peptide sequences and build a SVM able to identify MHC binders. Martin *et al*. (2005) developed a SVM by combining a sequence-based description of proteins with experimental information, while Nanni and Lumini (2006) proposed an ensemble of machine learning models and a new encoding technique which combines physicochemical indices of amino acids with the occurrence of dipeptides in protein sequences. Reiss and Schwikowski (2004) integrated protein sequence information and observed interactions in a probabilistic model based on the Gibbs sampling motif finding algorithm. This model is aimed at identifying the amino acid sequences of the SH3 domain ligands. Similarly, Lehrach *et al*. (2005) generated a maximum likelihood discriminative model for the direct evaluation of SH3 domain binding motifs in protein sequences. These approaches demonstrate that two main issues have to be faced in order to set up a reliable predictor for protein–protein interactions: the selection of the relevant sequence information and the identification of an effective encoding for the input information. It is known that standard encodings of protein sequences can produce a huge number of input variables (Baldi and Brunak, 1998), thus increasing the input space dimension without enhancing the quantity of available information. This reduces a model's power and may give rise to the 'curse of dimension' (Bishop, 1995). In the above-described works the problem of huge dimensionality has been addressed through the use of models less sensitive to the input space dimension (SVM) or the choice of compressed encoding

*To whom correspondence should be addressed.

of the interaction information. Therefore, a methodology that makes use of both an exhaustive representation of the interaction and a numerical encoding, which avoids the curse of dimensionality without weakening the biological information, is needed.

Over the past few years, it has become increasingly clear that many protein–protein interactions occur within short regions, often <10 amino acids in length within one protein. This is particularly true for protein recognition modules (PRMs), such as Src homology (SH) 2 and 3 domains, WW domains, phosphotyrosine binding domains (PTB), postsynaptic density/disc-large/ZO1 (PDZ) domains, Eps15 homology (EH) domains and 14-3-3 proteins that typically recognize linear regions generally 3–9 amino acids long.

In particular SH3 domains belong to a well-known family of 50–70 residue-long PRMs, which are ubiquitous in eukaryotes and bind to short proline-rich peptides characterized by a PxxP core (Sudol, 1998; Mayer, 2001; Musacchio, 2002). Structural studies of peptide–SH3 complexes have demonstrated that peptide ligands preferentially bind in one of two opposite orientations with respect to the SH3 domain (Feng *et al*., 1994; Lim *et al*., 1994), conforming to either class I ([RK]xxPxxP) or class II (PxxPx[RK]) binding consensus, respectively. Individual SH3 domains exhibit specific preferences for variations of their binding consensus, highlighting the promiscuity and the versatility of this kind of PRMs (for an interesting review, see Li, 2005). With the aim of investigating the binding specificity of SH3 domains, a number of experimental strategies have been proposed, some of which are high-throughput (Sparks *et al*., 1996; Kay *et al*., 2000; Cesareni *et al*., 2001; Landgraf *et al*., 2004).

In this work we have developed a methodology which integrates SH3–peptide binding data obtained from high-throughput experiments with information derived from SH3–peptide three-dimensional (3D) complexes. This combined information is numerically encoded *ad hoc*. A neural network model is then trained to infer SH3 domain–peptide binary interactions. The results of this procedure, applied to a set of *Saccharomyces cerevisiae* SH3 domains, are impressive, especially when the experimental data used to set up the method are abundant and of high quality. This is encouraging for the application of our approach to other SH3 domains of the same organism, to SH3 domains of other organisms, and also to other PRMs. In principle, only the availability of the necessary interaction and structural data would limit the possible extension of the methodology to other protein families that recognize more extended protein regions.

2 METHODS

2.1 The dataset

The dataset is composed of 7925 domain–peptide complexes involving 16 SH3 yeast domains and a peptide library of ∼1500 sequences (1379 plus a variable number of domain-specific sequences) extracted from the yeast proteome [The Saccaromyces Genome Database (http://www.yeastgenome.org/), see Table 1] and representing all occurrences of both class I and class II binding *consensi* (Lim *et al*., 1994). True interactions were identified in phage display (Tong *et al*., 2002) and in pep-spot experiments (Landgraf *et al*., 2004), thus allowing us to use both positive and negative information through the identification of binding and non-binding subsets of domain–peptide pairs.

2.2 The SH3 domain–peptide contact space

We processed the sequences of each SH3–peptide interacting pair by selecting only amino acids lying on the interaction surface and directly involved in an inter-molecular contact. Given a domain A and a peptide B in a 3D complex, an amino acid of A is assumed to be in contact with an amino acid of B if the shortest distance between their atoms is smaller than the sum of their van der Waals radii plus a tolerance of 3 Å. To identify the SH3–peptide contact residues, we considered six 3D complexes of known structure [PDB (http://www.pdb.org) ID codes 1ABO, 1EFN, 1OEB, 1N5Z, 1OV3 and 1CKA, corresponding to the SH3 domains of Abl and Fyn kinases, Adapter protein Grb2, Peroxin-13 protein, NCF-1 and C-Crk, respectively] and analyzed them with the software PINQ (Lesk, 1986 and references contained therein).

Table 1. The experimental domain–peptide database. Data are derived from Landgraf *et al*. (2004) and Tong *et al*. (2002)

Domains	Class I peptides	Class II peptides	Class I binders	Class II binders
Boi1	672 (19)	707 (15)	15 (3)	16 (6)
Myo3	(28)	—	(7)	—
Myo5	1139 (34)	—	43 (13)	—
Rvs167	672 (21)	707 (21)	19 (11)	44 (16)
Sho1	672 (30)	—	37 (18)	—
Yfr024	672 (24)	707 (27)	25 (7)	123 (22)
Yhr016	672 (18)	707 (16)	12 (6)	67 (11)
Ygr136	(32)	(24)	(18)	(15)
Ypr154	(34)	(18)	(23)	(8)
Sla1-3	(26)	—	(8)	—
Nbp2	(35)	—	(16)	—
Hr114-1	(31)	—	(14)	—
Hr114-2	(28)	—	(13)	—
Yhl002	(23)	—	(9)	—
Yjl020	(22)	(22)	(4)	(11)
Pex13	(22)	(28)	(10)	(16)
Subtotal	4499 (427)	2828 (171)	151 (180)	250 (105)
Total	4926	2999	331	355

The first number represents the dimension of the peptide library scanned from each domain by pep-spot technology. The number in parenthesis refers to the peptides tested by phage display experiments.

'Virtual' contacts were defined in order to exploit all SH3/peptide interaction data involving peptides whose structure or complex structure is not known. A 'virtual' contact, according to Brannetti *et al*. (2000), is established between a domain and a peptide residue pairs if and only if (1) the SH3 sequence can be aligned to the multiple alignment shown in Figure 1a; (2) the peptide sequence can be aligned to at least one of the peptides of the SH3/peptide complexes of known structure; (3) the SH3 and the peptide have been experimentally tested in a pep-spot or in a phage display experiment. The SH3 and peptide residues contact positions are identified by the SH3 and peptide multiple sequence alignments, respectively.

The definition of virtual contact is supported by the theoretical basis of homology modelling. An SH3 sequence, which can be aligned to a multiple alignment of SH3 domains of known structure, can be homology-modelled with high reliability, depending on its sequence similarity to the template structure (Sali, 1995; Srinivasan *et al*., 1996). A poly-proline sequence, which has been shown to interact with an SH3 domain is very likely to assume a poly-proline II structure. It is therefore reasonable to assume that the pattern of contacts identified on the crystal structures of a number of complexes is conserved in the homologous sequences and can be extended to the phage display or pep-spot interaction data.

Virtual contacts can lead to a binding or to a non-binding interaction, depending on the results of the experimental tests. In case of non-binding interaction, two main possibilities can be envisaged:

(1) The domain or the peptide does not assume the expected 3D conformation;

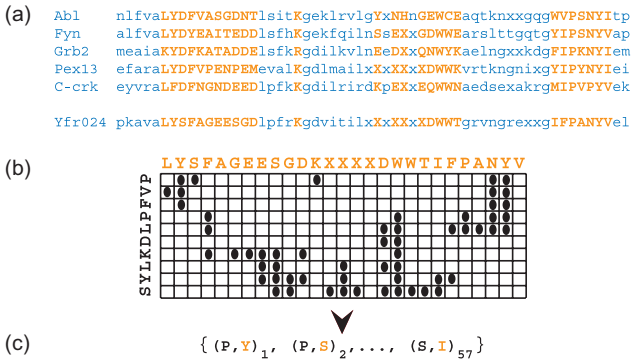

Fig. 1. The identification of PAIRs of a query domain–peptide pair follows three steps: (**a**) the sequence of the query domain (Yfr024 in the reported example), for which no structural data are available, is aligned to the sequences of the domains of known structure previously used to extract contact information. From the alignment, the contact positions of the query domain are inferred (uppercase residues). (**b**) The inferred contact positions of the query domain (27 for an SH3 domain) and the positions of the query peptide (10 for an SH3 ligand) are aligned respectively to the columns and to the rows of the contact matrix (the matrix shown represents class I complexes). (**c**) PAIRs involved in the interaction are identified from the cells filled with a black dot in the matrices (57 in the class I matrix). The subscript index in a PAIR indicates the position of the contact along the matrix in a lexicographic order, namely reading the filled cells from the first top left to the last bottom right.

(2) The domain and the peptide fold in the expected SH3 and poly-proline II conformations, but no complex is formed between them.

We foresee that most of the non-binding cases represented in our dataset fall in this second ensemble and we assume that no binding is detected since the residue pairs that are responsible for the binding in the complexes of known structure are changed into residue pairs that are not compatible with the complex formation.

The relative frequency of virtual contacts leading or not leading to the formation of complexes can be organized in a contact matrix, as described below.

2.3 Contact matrix

A contact matrix for an SH3–ligand complex is characterized by 27 columns (the number of the domain positions involved in the interaction) and 10 rows (the number of peptide positions involved in the interaction) (Brannetti *et al*., 2000). The matrix contains 27×10 elements, each corresponding to a pair of residues, one belonging to the SH3 domain and one to the peptide (Fig. 1b). Out of the 270 potential pairs of interacting residues, those belonging to class I and class II complexes are 57 and 53, respectively.

The contact residues of a new SH3 domain–peptide pair are identified as follows (Fig. 1). The sequence of the domain is aligned to the multiple sequence alignment of the SH3 domains used to build the contact matrix (Fig. 1a); 27 contact positions are derived from the alignment and assigned to the columns of the 27×10 contact matrix. A total of 10 peptide positions are assigned to the rows of the contact matrix, by aligning the last proline of the class I motif to the first row from the top (in Fig. 1b, the class I matrix). The contact residues are then inferred from the contact matrix filled cells (Fig. 1c).

2.4 Pair of interacting residues

A pair of interacting residues (PAIR) is thus defined as the object $(p_i, d_j)_k$, $i \in \{1, \ldots, 10\}$, $j \in \{1, \ldots, 27\}$, $k \in \{1, \ldots, N\}$, where p_i and d_j are the i-th and j-th interacting residues of the peptide and the domain, respectively, while k is the order index of the contact position (Fig. 1c). This procedure is used to transform the dataset of domain–peptide sequence partners into a dataset of PAIR arrays. Class I and class II contact matrices identify 57 and 53 PAIRs, respectively.

The relevance of a PAIR in the formation of a complex can be assessed by examining its frequency within the binding and the non-binding peptide subsets.

Given the PAIR $(p, d)_k$, in the position k, where $(p, d) \in \Sigma \times \Sigma$, Σ represents the set of all amino acids plus the insertions, the relative frequencies $f_k^{(+)}(p, d)$ and $f_k^{(-)}(p, d)$ were defined as follows:

$$f_k^{(+)}(p,d) = \frac{n_k^{(+)}(p,d)}{\sum_{r',\, r'' \in \Sigma} n_k^{(+)}(r', r'')}$$

$$f_k^{(-)}(p,d) = \frac{n_k^{(-)}(p,d)}{\sum_{r',\, r'' \in \Sigma} n_k^{(-)}(r', r'')},$$

where $n_k^{(+)}(p, d)$ and $n_k^{(-)}(p, d)$ indicate the number of occurrences of $(p, d)_k$ in the binding and non-binding subsets respectively, and k represents the generic contact position. Hence, the PAIRs distributions for any given position k are $\Phi_k^{(+)} = \{f_k^{(+)}(p, d)\}_{(p,\, d) \in \Sigma \times \Sigma}$ and $\Phi_k^{(-)} = \{f_k^{(-)}(p, d)\}_{(p,\, d) \in \Sigma \times \Sigma}$.

2.5 PAIRs binding significance

The comparison of the relative frequencies makes it possible to gauge which PAIRs are specific for binding and which for non-binding, and which PAIRs have no role in the SH3 domain–peptide interaction.

We introduced the following rule of significance: if $f_k^{(+)} > 0$ and $f_k^{(-)} = 0$, the corresponding PAIR makes a positive contribution to the formation of the complex. If $f_k^{(+)} = 0$ and $f_k^{(-)} > 0$, the PAIR provides a negative contribution to the formation of the complex. If $f_k^{(+)} > 0$ and $f_k^{(-)} > 0$, the relevance of the PAIR in the formation of the complex depends on the difference between $f_k^{(+)}$ and $f_k^{(-)}$: If $f_k^{(+)}$ is greater than $f_k^{(-)}$, the PAIR is more relevant for binding whereas, if $f_k^{(+)}$ is lower than $f_k^{(-)}$, the PAIR is more relevant for non-binding. This rule of significance clearly depends on whether the distributions $\Phi_k^{(+)}$ and $\Phi_k^{(-)}$ display a meaningful difference.

A numerical code $C_k(p, d)$ can be proposed for a PAIR $(p, d)_k$, according to its binding significance:

$$(p,d)_k \rightarrow C_k(p,d) = \frac{f_k^{(+)}(p,d) - f_k^{(-)}(p,d)}{\max\left(f_k^{(+)}, f_k^{(-)}\right)}$$

$$C_k(p,d) = \begin{cases} 1 \text{ if } f_k^{(+)} > 0,\ f_k^{(-)} = 0 \\ c \in (-1, 1) \text{ if } f_k^{(+)},\ f_k^{(-)} > 0 \\ -1 \text{ if } f_k^{(-)} > 0,\ f_k^{(+)} = 0 \end{cases}$$

For each pair of domain–peptide sequences, the set of N PAIRs is thus transformed into an N-tuple of PAIR numerical codes.

We stress that the set Σ includes an auxiliary character 'X' that represents insertions in domain contacts alignment (Fig. 1 and Brannetti *et al*., 2000) and is also used to complete peptide sequences shorter than 10 amino acids. Therefore a PAIR that combines a domain residue and an unknown peptide residue, or an insertion in the domain sequence and a peptide residue, or an insertion and an unknown residue, is considered to be meaningless and a null numerical code is assigned to it.

2.6 The input space

Each class I interaction between an SH3 domain and a peptide in the dataset is thus described by 57 real variables, varying between -1 and $+1$ and representing the contacts existing between the domain and the peptide. As described above, each variable encodes for a single contact between a domain and a peptide, independently of the other contacts and of the relative PAIR frequencies. This corresponds to the assumption of an independent contacts representation, which is clearly an approximation simplifying the analysis of binding information. This assumption implies a noisy input

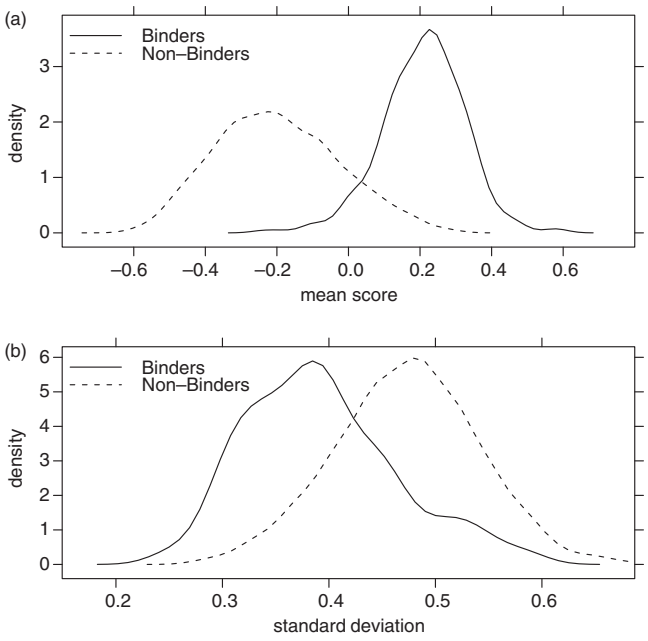

Fig. 2. The average and the standard deviation of the 57 numerical codes that represent each interaction, enhance the difference between positive (binders) and negative interactions (non-binders). We added these indicators as auxiliary input variables for the neural network. (**a** and **b**) Figures show the distributions of the mean numerical code (mean score) and of the standard deviation evaluated for all the interacting (solid line) and non-interacting (dashed line) domain–peptide pairs included in our dataset.

space. In fact, some contact positions might be non-significant for discriminating binders from non-binders: In this case the corresponding distributions $\Phi_k^{(+)}$ and $\Phi_k^{(-)}$ do not differ meaningfully. This implies that the numerical codes associated to the PAIRs in those non-significant positions will fluctuate randomly around zero as a sort of white noise.

In order to reduce this type of noise we added, as auxiliary input variables, the average and the standard deviation (SD) of the 57 numerical codes of each interaction (Fig. 2).

2.7 The neural network

All the encoded information was used to build and test a neural network model. The architecture of the network is feed-forward, with a single hidden layer. A single output unit is a linear combination of sigmoid hidden units.

In the training phase, the output of the network is forced to assume the value 1 if the input refers to true binding partners, 0 otherwise.

For the simulation we used the Stuttgart University software JavaNNS, i.e. the Java version of the SNNS package (http://www-ra.informatik.uni-tuebingen.de/SNNS/).

2.8 Control procedures

We formulated and carried out four control procedures in order to investigate the efficacy of the method. In the first procedure, the neural network results were compared with the results of a position specific scoring matrix (PSSM) (Henikoff and Henikoff, 1997). Using the Emboss (Rice *et al*., 2000) routines 'prophecy' and 'profit' (EMBOSS, http://emboss.sourceforge.net/), a PSSM was built for each domain displaying a sufficient number of binders. The matrices were built on data extracted from the neural network training sets and tested on the corresponding test sets. To compare the PSSM and neural network results, we obtained an overall PSSM performance by assembling the results of each domain-specific matrix.

To evaluate how the neural network approach improves ligand identification, we compared our results to the results of the SH3-SPOT methodology (Brannetti and Helmer-Citterich, 2003), which also relies on information on the frequencies of contact residues in domain–peptide pairs. SH3-SPOT does not make use of non-binding peptides data and is not based on a learning method. The SH3-SPOT score of a domain–peptide pair is derived from the sum of the residue–residue pair frequencies in the domain–peptide contact positions (Brannetti *et al*., 2000).

The PAIR representation involves two essential aspects: a structure-based approach to the SH3 domain–peptide interaction and a novel compressed encoding that strongly reduces the input space dimension. In order to verify the improvement carried out by each aspect, we defined as supplementary control procedures both a sequence-based and a structure-based approach, using a standard orthogonal encoding (Wu, 1997; Baldi and Brunak, 1998). The sequence-based approach (BinSeq in the following) considers the entire sequence of both the domain (58 amino acids) and the peptide (14 amino acids), while the structure-based approach (Bin3D in the following) deals only with the contact residues of both domain (27 non-contiguous residues, see above and Fig. 1) and peptide (10 contiguous residues, see above) sequences. In both cases, the information is encoded using 20 binary variables for each amino acid. The corresponding input spaces have 1440 and 740 dimensions respectively.

2.9 Models evaluation

Each model (NN-PAIRs, PSSM, SH3-SPOT, Bin3D NN and BinSeq NN) was evaluated by a 5-fold cross-validation. The dataset is divided into five subsets of equal size. The training and testing of every model was carried out five times, using on each occasion one distinct subset for testing and the remaining four subsets for training.

In the PAIRs approach, each training set was used to define the PAIRs, to assess their binding significance, and to build the models. The remaining test set was expressed as PAIRs, which were translated into numerical *N*-ples using the code previously built on the training set.

Since binding and non-binding experimental data in the dataset were unbalanced, the binders in the training set were replicated until an equal proportion was established. The test sets were left unbalanced. A validation set is randomly extracted from each training set and is used for the stopping criterion.

An average performance and its related error were evaluated from the results of the five test sets.

The output of each model is normalized in the range [0,1] and a receiver operating characteristic (ROC) curve is plotted as a representation of the true positive rate (TP/(TP + FN), or sensitivity versus the false positive rate (FP/(FP + TN) or 1 − specificity for different value of a decision threshold. In order to obtain a single measure of the model performance that is independent of the decision threshold, the accuracy was evaluated as the area under the curve (AUC) (Bradley, 1997).

2.10 Class II peptides

The methodology was also applied to the interactions between SH3 domain and class II peptides. The structure of the interaction matrix is the same, with 27 columns for domain positions and 10 rows for peptide positions. In this case only 53 pairs of interacting positions emerged from the analysis of class II complexes. This means that 53 PAIRs represent a domain–peptide pair in case of class II peptides. As for class I data, we added the average and the SD of the numerical codes for each domain–peptide pair. Therefore, the corresponding neural network is characterized by 55 inputs.

3 RESULTS

Integrating both structural and sequence information we have developed a neural network-based methodology for inferring binary domain–peptide interactions. The interaction interface is encoded in a set of numerical variables representing spatial contact positions.

Table 2. Averaged AUC values for the neural network model and control models applied to class I and class II datasets

Method	Class I peptides	Class II peptides
NN-PAIRs	0.922 ± 0.005	0.921 ± 0.008
NN Bin3D	0.89 ± 0.02	0.90 ± 0.01
NN BinSeq	0.887 ± 0.005	0.88 ± 0.02
SH3-SPOT	0.78 ± 0.02	0.74 ± 0.02
PSSM[a]	0.60 ± 0.03	0.69 ± 0.08

[a]PSSM performances are evaluated on 13 domains for class I peptides and on 4 domains on class II peptides. For the remaining domains, the matrices did not resolve binders from non-binders.

We focused on yeast SH3 domains as an ideal domain–peptide interaction template.

3.1 The encoding procedure

First, we want to emphasize the efficacy of the encoding procedure that transforms both sequence and structural information in a moderate number of numerical variables, thus avoiding the problem of huge dimensionality typical of machine learning approaches (Wu, 1997, Bock and Gough, 2001, Martin *et al*., 2005). The encoding requires a two-step procedure in which relevant pairs of interacting residues (PAIRs, see Methods) are identified from sequence and 3D structure, in order to represent the domain–peptide interaction. Next, a numerical encoding is proposed, taking into account the role of each PAIR in participating in the formation of the complex. The PAIRs' distributions $\Phi_k^{(+)}$ and $\Phi_k^{(-)}$ (see Methods), by which we define such a role, do not always display a sharp difference, thus producing a noisy representation of the interactions. We reduced such noise by adding the average and the SD of the numerical codes to the input variables used to represent the contact positions. These supplementary variables help to discriminate binders from non-binders (Fig. 2).

3.2 The inference of new binders

Performances of the neural network and control models are reported in Table 2 in the form of the area under the ROC curve (AUC) (Bradley, 1997). These results are obtained considering the entire dataset of pep-spot and phage display interactions (see Table 1 and Methods). The three neural approaches clearly outperform the other models (see Table 2 and Fig. 3). For class I data, the PAIRs neural network achieves 0.92 of averaged global accuracy, while the Bin3D and the BinSeq networks attains a lower but considerable performance of 0.89 and 0.88, respectively. SH3-SPOT and PSSM attain only 0.77 and 0.60, respectively. Analogously, for class II interactions, the NN-PAIRs accuracy is 0.92, against 0.90 and 0.88 of Bin3d and BinSeq neural networks respectively. Again, the SH3-SPOT and PSSM performances reach only 0.74 and 0.69 respectively. The NN-PAIRs method gives two pivotal advantages over the BinSeq and the Bin3D procedures: (1) A higher AUC value is reached, which guarantees a better overall accuracy of the model; (2) NN-PAIRs involves a small number of parameters with respect to BinSeq and Bin3D. This ensures that the network generalization error (Baum and Haussler, 1990; Barron, 1994) is lower for NN-PAIRs and that such model can also be applied in cases with small amounts of available training data.

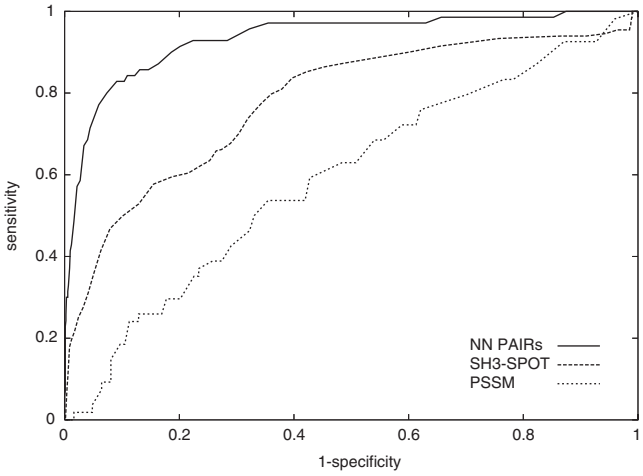

Fig. 3. ROC curves for the NN-PAIRs, PSSM and SH3-SPOT performance in the case of class I specificity inference. The NN curve is the closest to the optimal point (0;1) which corresponds to the maximum sensitivity and the minimum false positive rate. The NN curve is always above the control curves for acceptable values of false positive rate.

In order to compare the performance of our model with other recent statistical predictors (Reiss and Schwikowski, 2004, Lehrach *et al*., 2006) we applied our method to the phage display dataset (Tong *et al*., 2002) largely used in those applications. In this case the data consist of 426 interactions between 16 SH3 domains of baker's yeast and peptides selected by phage display experiments on a random peptide library (Tong *et al*., 2002). The lower number of interactions required a reduction of the complexity of the neural network. Nevertheless, the performance, with an average AUC of 0.83 and a standard error of 0.04, is higher than the one of the Reiss and Schwikowski model (AUC = 0.79) (Reiss and Schwikowski, 2004) and comparable with the results of Lehrach *et al*. (AUC = 0.83) (Lehrach *et al*., 2006). Note that both the BinSeq and the Bin3D encoding schemes generate a redundant neural model with a number of parameters higher than the available data and with a performance (AUC) <0.78.

3.3 The generalization to unknown SH3 domains

In order to assess its generalization ability, our method was applied for inferring the interaction specificity of SH3 domains, which were not present in the training set. We selected the pep-spot interaction data (available for the six SH3 domains: Boi1, Myo5, Rvs167, Sho1, Yfr024 and Yhr016) as the ideal test data since they cover the entire yeast proteome. Then we designed a 'cross-validation sampling' where the test set comprises the interaction data of one of the six domains listed above and the training set consists of the interaction data of all the remaining 15 domains of our dataset (Table 1).

The model reaches an optimal generalization capability in predicting the specificity of Yhr016 and Yfr024 (AUC = 0.91 and AUC = 0.89, respectively) and displays a good inference of the specificities of Myo5 and Rvs167 (respectively, AUC = 0.74, AUC = 0.70), while the global accuracy for Sho1 and Boi1 domains is lower (AUC = 0.52 and AUC = 0.41, respectively) (Fig. 4).

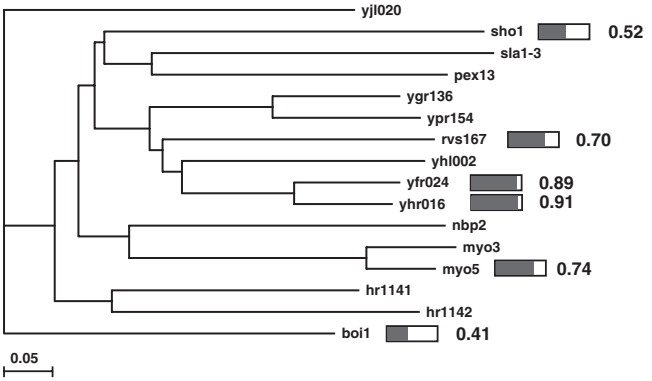

Fig. 4. The generalization capability of the model depends on the similarity of the unknown SH3 domain with the domains belonging to the training set. Closest similarities correspond to higher level of generalization. The phylogenetic tree in the figure is obtained from the multiple alignment of the 16 SH3 domains considered in this work. The figure also displays the level of accuracy that the model reaches for the six domains on which we tested the generalization capability. Yfr024 and Yhr016 show considerable accuracy owing to their strong sequence similarity (89 with 79% of sequence identity) and to a broad sharing of interacting peptides. Analogously, the good accuracies evaluated for Myo5 and Rvs167 are related to the similarity with Myo3 and the tern Yhl002, Yfr024 and Yhr016, respectively.

4 DISCUSSION

The problem of SH3 domain specificity in baker's yeast was investigated with the aim of developing a computational methodology focused on protein–peptide interaction inference and characterized by some important innovations. It is worth noting that both positive (interacting) and negative (non-interacting) experimentally confirmed domain–peptide pairs were used. The use of reliable negative information rather than the artificial randomization of the pairs to obtain negative cases (Bock and Gough, 2001, Martin *et al*., 2005), enhanced the discrimination capability of the method, thus making the neural network specificity particularly significant.

Since two interacting partners establish a network of contacts between residues, protein–peptide interaction can be analyzed in the light of residue–residue contacts across the crystal structure interaction interface. In this context, we identified the 'PAIRs' as crucial elements of our approach to the inference of domain interaction specificity.

The PAIR's representation can be used to characterize domain–peptide and, more generally, protein–protein interacting pairs. The transformation of the partner sequences into a set of encoded PAIRs projects the interaction information in a numerical space that comprises the input space of the neural network classifier. The PAIR's input space is characterized by a manageable dimension, an important property that makes it possible to circumvent the curse of dimension (Bishop, 1995). Moreover, the numerical variables thus introduced [<60 in this application corresponding to the orthogonal encoding of three residues (Baldi and Brunak, 1998)] have a precise meaning, since they represent the contact positions involved in an SH3–ligand 3D crystal complex. Notice that a small number of variables describing putative interacting partners would make it possible in future to add further variables representing other biophysical properties of proteins.

The integration of the encoded PAIRs with a complex model like neural network makes the inference of new interactors of SH3 domains extremely accurate.

Specifically, the overall performance of the NN-PAIRs is 92% for both class I peptides and class II peptides. It outperforms SH3-SPOT and PSSM results for both classes of peptides. Our method also performs well when compared with recently developed probabilistic tools.

Our methodology is based on two innovative features: a structure-based model which allowed the selection of a relatively low number of residue–residue pairs involved in the binding, and a knowledge-based encoding which allowed us to summarize the interaction information in a very compact and comprehensive form.

In order to dissect the different contributions, two control procedures were designed: BinSeq, where the orthogonal encoding was applied to the whole domain and peptide sequences ($\sim$70 residues), and Bin3D, where the orthogonal encoding was applied to the residues selected for belonging to the domain–peptide complex interface ($\sim$40 residues).

The structure-based selection of residues involved in the binding results in a better performance of the Bin3D over the BinSeq control procedure. The knowledge-based encoding introduced in this work allows a further improvement of the results, as shown by the better statistical parameters attained and shown in Table 2. Moreover, the reduced input space of NN-PAIRs enlarges the application capability of the model to small sized training data.

It should be noticed that our methodology, based on a representative encoding of the domain–peptide pairs, is strongly dependent on the quality and dimension of the training set. Nevertheless, it reveals a capability of generalization that permits the inference of the specificity of unknown SH3 domains. Obviously, such capability is influenced by the similarity of the new domain to the domains used to train the model (Fig. 4).

It is well known that interaction data obtained through high-throughput *in vitro* experiments often exhibit cases of false negatives and false positives. Even if the procedure described in this work benefits from such experiments, it must be emphasized that the problem of false negatives and positives has been neglected, our main interest being the methodological approach of protein–peptide interaction inference.

Nevertheless, we have shown that, by merging a suitable representation of the interacting partners, an appropriate encoding of the variables, and a complex model such as a neural network, it is possible to identify new SH3–domain binding peptides, based on the information enclosed in the residue–residue contacts of known interactors. This suggests that structural complexes contain more information on specificity than a mere sequence consensus derived from known binders: Facing residues of two binding partners define affinity and specificity of the interaction and make it possible to characterize protein–protein interactions more accurately.

In the near future we intend to apply the method to other families of protein domains and, subsequently, to further develop this approach in order to provide experimentalists with a powerful computational support for the construction of peptide libraries and for the investigation of domain–peptide interactions.

ACKNOWLEDGEMENTS

The authors are grateful to Gianni Cesareni for helpful discussion and to Barbara Brannetti for important suggestions. The authors thankfully acknowledge the support of Telethon (GGP04273), GENEFUN, AIRC, a PNR 2001-2003 (FIRB art.8) and a PNR 2003-2007 (FIRB art.8).

Conflict of Interest: none declared.

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Baldi,P. and Brunak,S. (1998) *Bioinformatics: The Machine Learning Approach,* MIT Press, Cambridge, MA, 1998.

Barron,A.R. (1994) Approximation and estimation bounds for artificial neural networks. *Mach. Learn.*, **14**, 115–133.

Baum,E.B. and Haussler,D. (1990) What size net gives valid generalization? *Neural comput.*, **1**, 151–160.

Bishop,C.M. (1995) *Neural networks for Pattern Recognition* (Oxford University Press).

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

Bork,P. *et al.* (2004) Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.*, **14**, 292–299.

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

Brannetti,B. and Helmer-Citterich,M. (2003) iSPOT: A web tool to infer the interaction specificity of families of protein modules. *Nucleic Acids Res.*, **31**, 3709–3711.

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.

Cesareni,G. *et al.* (2001) Can we infer peptide recognition specificity mediated by SH3 domains? *FEBS Lett.*, **513**, 38–44.

Enright,A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

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

Gaasterland,T. and Ragan,M.A. (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genomics*, **3**, 199–217.

Gavin,A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Goh,C.-S. *et al.* (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**, 283–293.

Henikoff,S. and Henikoff,J.G. (1997) Embedding strategies for effective use of information ifrom multiple sequence alignments. *Protein Sci.*, **6**, 698–705.

Ho,Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Ito,T. *et al.* (2000) Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA*, **97**, 1143–1147.

Kay,B.K. *et al.* (2000) The importance of being proline: The interaction of proline-rich motifs in signaling proteins with their cognate domains. *FASEB J.*, **14**, 231–241.

Landgraf,C. *et al.* (2004) Protein interaction networks by proteome peptide scanning. *PLOS Biol.*, **2**, 94–103.

Lehrach,W.P. *et al.* (2006) A regularized discriminative model for the prediction of protein–protein interactions. *Bioinformatics*, **22**, 532–540.

Lesk,A.M. (1986) Integrated access to sequence and structural data. In *Biosequences: Perspectives and User Services in Europe*. Saccone,C. (ed). EEC, Bruxelles, pp. 23–28.

Li,S.S. (2005) Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction. *Biochem. J.*, **390**, 641–53.

Lim,W.A. *et al.* (1994) Structural determinants of peptide-binding orientation and of sequence specificity in SH3 domains. *Nature*, **372**, 375–379.

Marcotte,E.M. *et al.* (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

Martin,S. *et al.* (2005) Predicting protein–protein interactions using signature products. *Bioinformatics*, **21**, 218–226.

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

Milik,M. *et al.* (1998) Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nature*, **16**, 753–756.

Musacchio,A. (2002) How SH3 domains recognize proline. *Adv. Protein Chem.*, **61**, 211–68.

Nanni,L. and Lumini,A. (2006) An ensemble of K-local hyperplanes for predicting protein–protein interactions. *Bioinformatics*, **22**, 1207–1210.

Overbeek,R. *et al.* (1999) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol.*, **1**, 93–108.

Pazos,F. and Valencia,A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.*, **14**, 609–614.

Pazos,F. and Valencia,A. (2002) *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, **47**, 219–227.

Pazos,F. *et al.* (1997) Correlated mutations contain information about protein–protein interaction. *J. Mol. Biol.*, **271**, 511–523.

Pellegrini,M. *et al.* (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.

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

Rice,P. *et al.* (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in genetics*, **16**, 276–277.

Russell,R.B. *et al.* (2004) A structural perspective of protein–protein interactions. *Curr. Opin. Struct. Biol.*, **14**, 313–324.

Sali,A. (1995) Modeling mutations and homologous proteins. *Curr. Opin. Biotechnol.*, **4**, 437–451.

Sparks,A.B. *et al.* (1996) Distinct ligand preferences of Src homology 3 domains form Src, Yes, Abl, Cortactin, p53bp2, PLCγ, Crk, and Grb2. *Proc. Natl Acad. Sci. USA*, **93**, 1540–1544.

Srinivasan,N., Guruprasad,K. and Blundell,T.L. (1996) Comparative modelling of proteins. In Sternberg,M.J.E., Rickwood,D. and Hames,B.D. (eds), *Protein Structure Prediction, A practical approach*, Oxford University press, Oxford.

Sudol,M. (1998) From Src homology domains to other signalling modules: proposal of the 'protein recognition code'. *Oncogene*, **17**, 1469–1474.

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

Uetz,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–631.

Valencia,A. and Pazos,F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.*, **12**, 368–373.

Wu,C.H. (1997) Artificial neural networks for molecular sequence analysis. *Comp. Chem.*, **21**, 237–256.

Zhao,Y. *et al.* (2003) Application of support vector machines for T-cell epitopes prediction. *Bioinformatics*, **19**, 1978–1984.

Zhu,H. and Snyder,M. (2003) Protein chip technology. *Curr. Opin Chem. Biol.*, **7**, 55–63.