An Integrated Machine Learning System to Computationally Screen Protein Databases for Protein Binding Peptide Ligands*\[S\]

Ling Zhang‡, Chen Shao‡, Dexian Zheng, and Youhe Gao§

A fairly large set of protein interactions is mediated by families of peptide binding domains, such as Src homology 2 (SH2), SH3, PDZ, major histocompatibility complex, etc. To identify their ligands by experimental screening is not only labor-intensive but almost futile in screening low abundance species due to the suppression by high abundance species. An ideal way of studying protein-protein interactions is to use high throughput computational approaches to screen protein sequence databases to direct the validating experiments toward the most promising peptides. Predictors with only good cross-validation were not good enough to screen protein databases. In the current study we built integrated machine learning systems using three novel coding methods and screened the Swiss-Prot and GenBank™ protein databases for potential ligands of 10 SH3 and three PDZ domains. A large fraction of predictions has already been experimentally confirmed by other independent research groups, indicating a satisfying generalization capability for future applications in identifying protein interactions. *Molecular & Cellular Proteomics* 5:1224–1232, 2006.

Experimental screening for protein binding peptides is not only labor-intensive but almost futile in screening for low abundance binding species due to the suppression by high abundance species. A more plausible way of studying protein-protein interactions is by using high throughput computational predictions rather than experimental approaches to screen for interactions from protein sequence databases to direct the validating experiments toward the most promising peptides. A prediction can only be called successful when the overall efficiency and cost of computational prediction plus biological validation are much better than those of experimental screening. A fairly large set of protein interactions are mediated by families of peptide binding domains, such as SH2,1 SH3, PDZ, MHC, etc. that act as receptors to accommodate, in their binding pockets, short peptides in an extended conformation (1, 2). We studied two common domain-ligand interactions mediated by SH3 domain and PDZ domain. SH3 domains selectively bind peptides of 8–11 amino acids on their ligand proteins; PDZ domains bind 4–8 amino acids in the C termini of their partners. Ligands of both PDZ and SH3 domains are of high diversity. Two good reviews provide detailed information on SH3 domain (3) and PDZ domain (4). Two major categories of methods have been developed to predict domain-ligand interactions. The first category is based on sequence information as exemplified by the work on predicting MHC-specific epitopes with protein docking methods (5, 6). This type of method needs intensive computation and the prior knowledge of the three-dimensional structures of the bait proteins. Instead of using exact protein structures, Brannerati et al. (7) and Altuvia and Margalit (8) extracted ligand-contacting residues from the known domain-ligand complex structures to approximately represent domain interface structures of SH3 and MHC, respectively. The interactions of amino acids at each position of the ligand with its contacting residues on the target domain were then scored using a statistical amino acid-amino acid pairwise potential table. By this means, peptide binding score was calculated by simply summing up scores over all the positions of the peptide. This type of method does not reflect the interrelations between different positions on the same peptide.

In contrast to the structure-based methods that include comprehensive and high quality information relevant to interactions in three-dimensional spaces, the second category of prediction methods is based on sequence information from only ligands or from both ligands and domains. One widely used method, Scansite (9) (scansite.mit.edu), calculates the position-specific scoring matrix from the known binding and non-binding peptides of a certain domain to characterize binding profiles. Tong et al. (10) have predicted 20 yeast SH3 binding ligands by the position-specific scoring matrix method. Machine learning approaches like artificial neural network (11, 12) and support vector machine (SVM) (13, 14) have been used in predicting precision; ESP, estimated screening precision; MHC, major histocompatibility complex; MCC, Matthews correlation coefficient; BLU, Boehringer light unit.

From the Proteomics Research Center, National Key Laboratory of Medical Molecular Biology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences/Peking Union Medical College, 5 Dong Dan San Tiao, 100005 Beijing, China

Received, October 24, 2005, and in revised form, January 24, 2006
Published, MCP Papers in Press, March 29, 2006, DOI 10.1074/mcp.M500346-MCP200

1 The abbreviations used are: SH, Src homology; SVM, support vector machine; PNN, probabilistic neural network; SP, screening.
MHC binding epitopes. All these algorithms need to collect a large amount of interaction data for each domain to build one predictor. If a domain can bind ligands belonging to different classes, two or more predictors should be built to characterize different binding profiles (13, 15). Dönnes and Elofsson (13) have reported that at least 20 ligand peptides of a single class for an MHC class I molecule are required for a reliable prediction, and with up to 50 peptides a slight improvement can be achieved. Even more data are required to build two or more predictors for ligands belonging to different classes. Martin et al. (16) have combined the full-length sequence information of both domains and ligands, devised a protein descriptor called signature products to represent interactions between pairs of amino acid sequences, and predicted ligands of SH3 domains using SVM. Because primary sequences are easy to accumulate, the sequence-based methods can be applied to a large variety of domains. However, primary sequences contain insufficient structure information, and these methods do not emphasize enough the quality of data.

The existing computational models have only been assessed by cross-validation and have failed to provide any evidence of their performance on screening of the whole protein database. To computationally screen protein databases and focus the experimental efforts on the most likely interactions, we set up an integrated prediction system, taking advantage of both types of methods. We extracted high quality structural data and a large quantity of aligned sequences of both interacting partners and processed the data with machine learning approaches. (a) Quality of data. We extracted only information relevant to interaction by taking potential interface residues rather than using the full-length sequence. (b) Quantity of data. We collected and aligned a family of domains and their ligands and combined the interacting partners into a single prediction system. Because there are structural and sequential similarities among the domains in each family, one pair of known interaction can complementarily provide interaction information for other similar ones. Therefore, less data were needed for each domain to train the predictors. (c) Presentation of data. We developed three novel coding methods to represent different aspects of interactions between interface residues, namely orthogonal dot product, physicochemical product, and structural matrix. (d) Processing of data. We used two machine learning approaches, SVM and probabilistic neural network (PNN), to set up three independent predictors. (e) Neural network ensemble. We assembled the result of the three independent predictors to achieve better generalization capability. (f) Biological filtering. We filtered the candidate ligands with biological information, i.e. the protein subcellular localization, to make further improvement of the system. The flow chart of our method is shown in Fig. 1.

EXPERIMENTAL PROCEDURES

SH3 and PDZ Structural Alignments and Interface Residue Extraction

539 SH3 domains and 613 PDZ domains in Swiss-Prot protein database (a few from TrEMBL database) (www.ebi.ac.uk/swissprot/) (17) (release 45.1) were identified by SMART (Simple Modular Architecture Research Tool) (18) (smart.embl-heidelberg.de/). 15 SH3-ligand and 12 PDZ-ligand complex structures were collected from Protein Data Bank (19) (www.rcsb.org/pdb). The 15 SH3 sequences were multiple-aligned to make a structural profile by ClustalW 3.0 (20) (available at www.ebi.ac.uk/clustalw/). The profile was manually adjusted by putting together regions of the same secondary structure and inserting gaps into loop regions. Other SH3 sequences with no resolved structures were multiple-aligned according to the structural profile with some manual adjustment. An SH3 residue was defined as an interface residue when, in any of the 15 complexes of resolved crystallographic structures, the shortest distance between the residue and the SH3 binding ligands is less than 3 Å. The interface residues were considered most relevant to protein interaction and were used to represent full-length sequences of the SH3 domain. A similar procedure was used in extracting PDZ interface residues. Ultimately 27 positions on SH3 and 23 positions on PDZ sequences were extracted as potential interface residues. (Structural alignments of 15 SH3 domains and 12 PDZ domains and interface residue extraction can be found in Supplemental Fig. S1; the full-length sequences and extracted interface residues of 539 SH3 and 613 PDZ domains can be found in Supplemental Table S1.)

Pretreatment of Ligand Peptides

For proteins interacting with PDZ domains, the C-terminal 8 amino acids were treated as ligand peptides and arrayed according to their positions. SH3 ligands in the training set were aligned by fitting into a very loose consensus and trimmed to 10 amino acids. Class I ligands were fit to motif XXXXXX(X/P)(X/P)(X/P), and class II were fit to X(X/P)XX(P/Q)XXX. Atypical peptides were set with Pro at position 5 or 7 of the 10 amino acids (X, any amino acid; P, hydrophobic amino acid). Peptides less than 10 amino acids were complemented with random amino acids. After alignment of the training set, Pro residues of the peptides were aligned at position 2, 5, 7, or 10. Therefore, each peptide in the test set was pretreated with Pro at any one position of 2, 5, 7, or 10. Pretreatment of peptides in the SH3 test set is not mandatory; nonetheless it saves computational screening time by reducing the number of peptide segments in the protein database.

Interaction Data Representation

Dot Product of Orthogonal Coding—Each interface residue on domains and each residue on ligands were coded in traditional orthogonal method by 20 dimensions of 0-1 vectors. Then a domain and a ligand could be described using vectors \( a = (a_1, \ldots, a_20)^T \in \mathbb{R}^{20} \) and \( b = (b_1, \ldots, b_20)^T \in \mathbb{R}^{20} \), respectively, where \( n \) indicates the number of interface residues on the domain and \( m \) indicates the number of residues on the ligand. To describe interactions between pairs of amino acids, a tensor product between vectors \( a \) and \( b \) was defined to be \( t(a_1, b_1, a_2, b_2, \ldots, a_m, b_m) \in \mathbb{R}^{20m} \). Using this definition, the coding for pairs of amino acids, called orthogonal dot product, was established.

Physicochemical Product Coding—Each amino acid was represented by five physicochemical properties: hydrophobicity (21), charge (22), van der Waals volume (23), flexibility (24), and bulkiness (25). Every property was classified into five groups by their value (The property classification of each of the 20 amino acids can be found in Supplemental Fig. S2). For one property, one amino acid was as-
signed to class 1, 2, 3, 4, or 5; the combination of two amino acids could be 11, 22, 33, 44, 55, 12, 13, 14, 15, 23, 24, 25, 34, 35, or 45 (in total 15 combinations) coded in 15 binaries. A pair of amino acids was represented by five properties, each with 15 combinations, so that a pair of amino acids was coded in $15 \times 5 = 75$ binaries, including 70 dimensions of zero and 5 dimensions of one. Each interface residue on domains and each residue on ligands were paired and coded in a set of 75 binaries. Then the interaction between a domain ($m$ contacting residues) and a ligand ($n$ residues) was coded by an element in the matrix.

Previous studies have used the measurements of the physicochemical properties of amino acids directly to code peptide sequences. But for the same property of the same amino acid, different literature reported slightly different measurements. We found that when using the previously published physicochemical coding method even a slight difference between measurements led to a different prediction performance. The differences did not interfere with the present physicochemical product coding method.

**Matrix of Structurally Interacting Potentials**—Betancourt and Thirumalai (26) have given a scale for interaction energies between the naturally occurring amino acid residues called potential matrix. The affinity of each amino acid pair composed of any of the 20 amino acids is expressed as a constant in the $20 \times 20$ matrix. Interaction between an interface residue on a domain and a residue on a ligand is coded by an element in the matrix.

### Machine Learning Approaches

**SVM** is a machine learning approach based on statistical learning theory. A full coverage of SVM is given by Vapnik (27). The decision rules developed by the system generate a discrete decision ($>0$, interaction; $<0$, no interaction) upon introduction of a new set of putative interaction pairs. SVM learning was implemented using SVM-light (28) (svmlight.joachims.org).

**PNN** is a kind of artificial neural network that is suitable for classification problems. We used the neural network toolbox from commercial software, MATLAB (Version 6.5), to design PNNs. Introductions of this toolbox are available at www.mathworks.com.

PNN shows good classification and generalization capabilities but has difficulty in dealing with high dimensional input vectors; SVM is suitable for the classification of sparse, high dimensional inputs. We used SVM to process orthogonal dot product and physicochemical product-encoded datasets and PNN to process matrix-encoded datasets.

**5-Fold Cross-validation**

The training set was randomly divided into five equally sized subsets. Each subset was used in turn as a test set, whereas the remaining four subsets were used to train the predictors. TP (true positive), FP (false positive), TN (true negative), and FN (false negative) in the five test sets were counted. The performance of each predictor was

---

**Fig. 1. Flow chart of experiment design.** Step 1, extracting potential interface residues on PDZ and SH3 domains. Step 2, setting up three independent machine learning predictors with encoded interaction and non-interaction PDZ and SH3 data and then assembling three predictors for PDZ and SH3. Step 3, computational screening of 11,146 proteins from Swiss-Prot human protein database for three PDZ domains and 13,000 peptides total from Swiss-Prot and GenBank™ for 10 SH3 domains to predict their potential ligands. The prediction results were compared with the peptide SPOT array experiment results in which SH3 domains were tested against the same 13,000 peptides as those tested in computational prediction, but PDZ domains were tested against only a subset of Swiss-Prot (6,223 proteins). DB, Database.
evaluated by calculating the mean value of three indices in the five tests with accompanying errors: precision, TP/(TP + FP); sensitivity, TP/(TP + FN); and Matthews correlation coefficient (MCC) (29).

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}
\]  

(Eq. 1)

In a totally correct prediction, MCC = 1; in a totally incorrect prediction, MCC = -1.

Test Set Obtained from Peptide SPOT Arrays

Three PDZ domains (Erbin, Lap2_Human; Af-6, Afad_Human; and Sna1, Sna1_Human) were tested against the C-terminal peptide of 11 amino acids from 6,223 human proteins in peptide SPOT array experiments by Boisguerin et al. (30) and Wiedemann et al. (31). The strongest 100 interactions were selected as positive ligands by the authors for each PDZ. In SH3 experiments by Landgraf et al. (32), 10 SH3 domains (Aqp1, Abp1_yeast; Boi1, Bob1_yeast; Boi2, Boi2_yeast; Myo5, Mys5_yeast; Rvs167, R167_yeast; Sho1, Ss81_yeast; Yhr016c, Yhh6_yeast; Yfr024c, Yfj4_yeast; Amphiphysin, Amph_human; and Endophilin, Sh32_human) were chosen to screen peptide arrays containing 672–2,032 peptides using Boehringer light unit (BLU) to represent the affinity between each SH3 and each 13-amino acid peptide sequence. The positive ligands were those with BLU larger than mean BLU plus 2 times standard deviation among total peptides tested. The rest of the peptides were regarded as non-binders.

RESULTS

System Setup—Based on secondary structure conformation, 539 SH3 and 615 PDZ domains were structurally aligned. 27 positions on SH3 sequences and 23 positions on PDZ sequences were extracted as potential interface residues. SH3 ligands were trimmed to 10 amino acids of protein internal sequence; PDZ ligands were the protein C-terminal 8-residue peptide.

598 interactions between 42 SH3 domains and their ligands along with 770 non-interaction pairs of 19 SH3 domains and the corresponding peptides (7, 10) were collected as the training set for SH3 interaction prediction. 338 interactions between 105 PDZ domains and 210 proteins (33) were collected as the training set for PDZ interaction prediction (details are in Supplemental Table S2.). To create negative data, the ligand sequence from each of these 338 interacting pairs was randomly rearranged. The protein domain and the shuffled sequence were regarded as negative and entered into the training set. SH3 positive data were duplicated to make the positive and negative ratio in the training set 1:1 approximately.

Each coding method emphasizes each of the three different aspects that affect protein and ligand interactions. Orthogonal dot product was a product modification of orthogonal coding (34), which sets each amino acid pair furthest apart in a distinct dimension. Physicochemical product represented the physical and chemical features of amino acid combinations. Structural matrix indicated the likelihood of two amino acids to interact in three-dimensional structures. All three predictors are oriented around interaction information between protein domains and their binding partners with no other irrelevant information.

Three predictors were trained independently with the three encoded training sets by machine learning approaches. The performance of each predictor was tested using 5-fold cross-validation (Fig. 2) by adjusting several parameters. The performance of each predictor was evaluated by calculating the mean value of three indices in the five tests with accompanying errors: precision, TP/(TP + FP); sensitivity, TP/(TP + FN); and Matthews correlation coefficient (MCC) (29).
three indices: precision, sensitivity, and MCC. To computationally screen protein database with higher reliability and to reduce the false positives, we set the predictors to high precision with compromised sensitivity (Table I) rather than choosing the predictor with the best MCC as usually done in other studies. More detailed rationales can be found under “Discussion.”

The prediction results of the three independent component predictors are combined by majority voting to build an assembled system. A ligand would finally be determined positive or negative if it was predicted to be binding or non-binding by two or more predictors. The 5-fold cross-validation results of the assembled system are also shown in Table I. The assembled system performed no better than the best performed individual predictor on 5-fold cross-validation. However, neural network ensemble did improve the accuracy of prediction in database screening as shown in Tables II and III. Ensemble of the three independent predictors would dramatically improve the generalization capability (35), which is the key for database screening but nonetheless cannot be shown by the cross-validation result generated from very limited data.

Prediction of Ligands for 13 Domains and Comparison with Peptide SPOT Array Screening Experiments—The three PDZ predictors computationally screened 11,146 human proteins from Swiss-Prot protein database with redundant C-terminal 8-amino acid sequences pre-excluded to predict potential ligands for three PDZ domains. The three SH3 predictors screened 13,000 peptides from 3,657 proteins (subsets of Swiss-Prot human protein database and GenBank™ yeast protein database (36)) for potential ligands for 10 SH3 domains (32). The prediction results were compared with the experimental screening results in peptide SPOT arrays, excluding any examples learned in training set. After ensemble, the average overlap of SH3 and PDZ prediction with experiments was 28.74 and 17.07%, respectively (Fig. 3). The prediction results of each predictor and the overlap between predictions and experiments for the 13 domains are shown in Tables II and III (Details of the predicted sequences by the assembled system and the predictions validated by experiments can be found in Supplemental Table S3.).

A measurement called screening precision (SP) was used to evaluate a predictor’s performance, whereas screening total ligand database was used to predict binding ligands of a domain. SP represents the generalization capability of predictors.

\[
SP = \frac{\text{confirmed predicted ligands}}{\text{totally predicted ligands}} \times 100\% \quad \text{(Eq. 2)}
\]

The last row in Tables II and III shows SP of each predictor and the assembled system. Despite that each domain learned less than 10 positive ligands on average, the prediction and experimental results overlapped well. Because PDZ peptide SPOT array experiments screened only a part of the Swiss-
Prot database (6,223 proteins out of a total of 11,146), experi-
mentally unconfirmed prediction results may not be false
positives but real positives not included in the experiments.

The current SP of PDZ undervalues the performance of the
system. The assembled prediction system dramatically improved
the screening precision. The ensemble has many superior
features. (a) Each predictor was trained to focus on one prop-
erty rather than trained with a universal code containing all
kinds of information. The complexity in machine learning is
therefore reduced (37). (b) Because the error overlapping the
three independent predictors was greatly decreased, the as-
sembled system acquired higher accuracy (38). (c) The as-
sembled system integrated information from many aspects. A
new predictor handling new aspects of information could
easily be incorporated into the system to improve the gener-
alization capability.

Generalization of Ligands Belonging to Different Classes—
There are two major classes of common motifs that SH3
binds: (XH11001PXH9272) and (PH9272PXH11001X
(H11001, basic amino acid; H9272, hydrophobic amino acid) and
other uncommon variants (3). There are two major classes of
PDZ binding ligands, characterized as
X(S/T)XH9272-COOH and
XH9272-XH9272-COOH, and a large variety of atypical ligands (4). We
integrated the ligands of different classes into one system to

### Table II
**Prediction results of three PDZ domains and verification results by peptide SPOT array experiments**

| Positive samples learned | Positive/whole set | Dot product (SVM) | Physicochemical product (SVM) | Structural matrix (PNN) | Ensemble |
|-------------------------|-------------------|------------------|--------------------------|---------------------|----------|
|                         |                   | Pred | Conf | Pred | Conf | Pred | Conf | Pred | Conf | Pred | Conf |
| Erbin                   | 4                 | 99/11,143 | 252  | 41  | 134  | 37  | 453  | 14  | 134  | 29  |
| Af-6                    | 10                | 96/11,138 | 277  | 28  | 107  | 16  | 450  | 25  | 96   | 14  |
| Sna1                    | 5                 | 96/11,141 | 431  | 41  | 473  | 44  | 257  | 16  | 227  | 35  |
| Total                   | 19                | 291/33,422 | 960  | 110 | 714  | 97  | 1,160 | 55  | 457  | 78  |
| SP (%)                  |                   | 11.46 | 13.59 | 4.74 | 17.07 |

**TABLE III**
**Prediction results of 10 SH3 domains and verification results by peptide SPOT array experiments**

| Positive samples learned | Positive/whole set | Dot product (SVM) | Physicochemical product (SVM) | Structural matrix (PNN) | Ensemble |
|-------------------------|-------------------|------------------|--------------------------|---------------------|----------|
|                         |                   | Pred | Conf | Pred | Conf | Pred | Conf | Pred | Conf | Pred | Conf |
| Abp1                    | 13                | 14/1,327 | 59   | 6   | 8    | 1   | 3    | 1   | 10   | 1   |
| Boi1                    | 9                 | 27/1,379 | 13   | 4   | 4    | 0   | 15   | 3   | 3    | 3   |
| Boi2                    | 8                 | 10/1,343 | 12   | 3   | 5    | 0   | 1    | 0   | 1    | 0   |
| Myo5                    | 15                | 28/1,141 | 58   | 20  | 51   | 9   | 135  | 17  | 49   | 17  |
| Rvs167                  | 13                | 15/672 | 28   | 7   | 3    | 1   | 30   | 5   | 7    | 4   |
| Sho1                    | 8                 | 34/1,023 | 45   | 20  | 25   | 14  | 8    | 3   | 11   | 10  |
| Yhr016c                 | 11                | 11/672 | 35   | 8   | 35   | 6   | 29   | 4   | 23   | 5   |
| Yfr024c                 | 24                | 80/1,379 | 249  | 72  | 321  | 56  | 87   | 20  | 174  | 55  |
| Amphiphysin             | 12                | 98/2,032 | 497  | 89  | 418  | 77  | 47   | 12  | 310  | 76  |
| Endophilin              | 3                 | 95/2,032 | 395  | 72  | 109  | 24  | 47   | 9   | 101  | 27  |
| Total                   | 116               | 412/13,000 | 1,391 | 301 | 979  | 188 | 407  | 74  | 689  | 198 |
| SP (%)                  |                   | 21.64 | 19.20 | 18.18 | 28.74 |

The assembled prediction system dramatically improved
the screening precision. The ensemble has many superior
features. (a) Each predictor was trained to focus on one prop-
erty rather than trained with a universal code containing all
kinds of information. The complexity in machine learning is
therefore reduced (37). (b) Because the error overlapping the
three independent predictors was greatly decreased, the as-
sembled system acquired higher accuracy (38). (c) The as-
sembled system integrated information from many aspects. A
new predictor handling new aspects of information could
easily be incorporated into the system to improve the gener-
alization capability.

**Generalization of Ligands Belonging to Different Classes—**
There are two major classes of common motifs that SH3
binds: (+XH11001PXH9272) and (PH9272PXH11001X
(+, basic amino acid;$, hydrophobic amino acid) and
other uncommon variants (3). There are two major classes of
PDZ binding ligands, characterized as X(S/T)XH9272-COOH and
XH9272-XH9272-COOH, and a large variety of atypical ligands (4). We
integrated the ligands of different classes into one system to

---

**Fig. 3.** The overlap between our prediction and the Peptide SPOT (Pep-spot) Array experiment results. **a,** prediction and experiment results for the three PDZ domains. **b,** prediction and experiment results for the 10 SH3 domains. **DB,** database.
were included in both training and test sets. However, in SH3 peptide array experiments, there were a number of peptides having the motif, which satisfies the consensus of both class I and class II motifs (X, any of the 20 amino acids: +, basic amino acid; P, hydrophobic amino acid). The peptides with the +X₈PX₆PX₊ motif in SH3 tests were not classified as either class I or class II for this analysis.

Biological Filters Complemented to Predictions—To reduce the number of experiments necessary for validating and increase the experimental success rate, it is recommended to select the prediction results by biological filters, e.g. similar gene expression profiles, subcellular co-localization, shared function, etc. Using a protein subcellular localization database, DBSubLoc (39), (www.bioinfo.tsinghua.edu.cn/SubLoc/), we excluded the proteins that were not co-localized in the same cellular compartment as proteins containing the target domain. As the existing database (August 2005) has not annotated localization information for every protein in Swiss-Prot, we only provide several examples to show the effect of biological filtering in Table V (Detailed localization information of each predicted interactor can be found in Supplemental Table S4.). The screening precision of our system was increased by more than 5 percentage points from 24.52% (76 of 310) to 31.25% (14 of 43) for Amphiphysin SH3 and from 14.58% (14 of 96) to 19.64% (11 of 53) for Af-6 PDZ. Biological filters will play more important roles as more information in databases are available. Other biological information could help further improve the prediction.

### DISCUSSION

Previous prediction methods have shown their performances by cross-validation on the training set but provided no evidence of their practical performances on database screening. However, there is a broad gap between the performance on database screening and cross-validation. In our prediction, screening precision of each predictor was much lower than the precision from cross-validation on the training set. There are two main explanations. First, unlike in the training dataset, the percentage of interactors in the database is very low. Because most peptides in the database are negative for binding to a particular domain, a small proportion of falsely predicted ligands would appear to be a relatively large number compared with that of the true positives in the prediction result. Second, good performance on cross-validation cannot guarantee the predictor a high generalization capability because the cross-validation result is based on very limited data.

To train a balanced learning machine, we set the ratio of positive/negative examples in the training set to 1:1 (34);
however, the odds of a domain binding ligands in the protein database (defined as \( r \)) is usually one in hundreds if not more. The precision in database screening of a predictor can be estimated from cross-validation indices by the formula below. (For derivation of Eq. 3, refer to the supplemental materials.)

\[
\text{Estimated screening precision (ESP)} = \frac{\text{precision} \times r}{\text{precision} \times r + (1 - \text{precision}) \times (1 - r)} \times 100\%
\] (Eq. 3)

In the formula above, precision was adopted from the cross-validation result on the training set. The formula here is based on the hypothesis that the predictor generalizes well on the whole protein database. In other words, taking any independent part of the protein database for the test set, in which the numbers of positive and negative data are equal, precision and sensitivity of the predictor are constant. In our prediction, \( r_{PDZ} \) and \( f_{SH3} \) were estimated to be 1 in 115 (291 of 33,442) and 1 in 32 (412 of 13,000), respectively. ESP of PDZ dot product predictor, for example, was calculated to be only 17.49\% by the above formula; this is much lower than the precision from cross-validation (95.80\%).

However, ESP calculated by the formula still overestimated the practical precision in database screening. Note that the premise of this formula that the predictor has good generalization capability is rarely achieved in reality. That is because of the following. (a) When the training set is very small, performance obtained from cross-validation cannot guarantee the actual performance on a large sample space (27). Patterns learned from very limited known interactions can rarely be generalized to the whole protein database. (b) It was difficult to obtain experimental negative data; hence shuffled ligands were created as negative data for training. It is intrinsically easier to separate real protein sequences from shuffled ones than to distinguish the real sequences of binding proteins from real sequences of non-binding proteins. The results generated by predictors fed with shuffled sequences might not adequately reflect the true precision (40); therefore cross-validation overrates the actual classification capability of the predictor, and hence ESP overrates SP. (c) The generalization capabilities of machine learning approaches, such as SVM and PNN, are not perfect daynows.

From the discussion above, it could be concluded that cross-validation precision is higher than ESP, and ESP overrates SP. The performance on cross-validation could not represent the practical performance in database screening. For a new prediction model, the best testing strategy is experimental validation (41).

In the cases when no experimental validation is available, how to choose a predictor that might have higher database screening capability? It can be estimated according to the ESP formula. Taking the dot product PDZ predictor for example, when the predictor was optimized to the best MCC (MCC = 0.8106, precision = 90.53\%), ESP was calculated to be only 8.25\%. ESP decreased a lot, even though precision of this predictor was not too much lower than the predictor we used to screen the database (precision = 97.01\%). Classifiers with lower precision would suffer from much more false positives and not be effective in helping the biological experimental designs. (A prediction will be called successful only when the total efficiency of the computational screen and the consequent experimental validation are higher than that of de novo experimental screening.) To reduce false positive predictions in database screening, we chose a predictor with high precision and moderate sensitivity on cross-validation rather than the one with the highest MCC.

SH3 prediction results seemed to be more successful than PDZ prediction results. There are three possible explanations. (a) The SP indices might underestimate the PDZ predictors in reality because the peptide SPOT array experiments of PDZ domains based on which SP was calculated screened only a part of the Swiss-Prot database. (b) The positive odds of SH3 binding ligands in the database was higher than that of PDZ binding ligands. Calculated by the formula, ESP of SH3 predictors was higher than PDZ predictors, and thus SH3 predictors would exhibit better performance if generalized well. (c) In the SH3 training set, experimental negatives in addition to created shuffled sequences were included. The predictors were better trained to discriminate the binders and non-binders, resulting in a better successful prediction rate. Using experimentally tested negative data in the training set could improve the practical performance of a predictor.

CONCLUSION

We predicted peptide ligands of 10 SH3 and three PDZ domains. 20–30\% of the predictions have already been experimentally confirmed by other independent research groups. Compared with the previous sequence-based prediction method, which requires a large number of interaction data for each domain (41), our system learned less than 10 ligands per domain on average. The system could also generalize to predict ligands belonging to different classes not included in the training set; it was impossible to predict different classes of ligands using methods solely based on one class of ligands. Predictions could be further improved by filters based on supplemental biological information.

This system can potentially be used to predict ligands of other peptide binding domains (SH2, PTB, etc.). Domain-domain interactions, such as between G-proteins and G-protein-coupled receptors, could also be predicted in similar procedures by extracting interface residues from each domain.

Acknowledgments—We thank Sucan Ma, Rui Tian, Shijuan Gao, and Ali Song for generous help in biological experiments; Xiaolin Yang and Fuxin Li for valuable suggestions in SVM and artificial neural network; and Weizhi Chen for critical reading and revision of the manuscript.
Protein Binding Ligand Prediction

This work was supported in part by The National Basic Research Program Grant 2004CB520804 and National Natural Science Foundation Grants 30270657, 30230150, and 3037030. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

REFERENCES

1. Pawson, T., and Nash, P. (2003) Assembly of cell regulatory systems through protein interaction domains. Science 300, 445–452
2. Pawson, T., and Scott, J. D. (1997) Signaling through scaffold, anchoring, and adaptor proteins. Science 278, 2075–2080
3. Mayer, B. J. (2001) SH3 domains: complexity in moderation. J. Cell Sci. 114, 1253–1263
4. Noury, C., Grant, S. G., and Borg, J. P. (2003) PDZ domain proteins: plug and play! Sci. STKE 2003, RE7
5. Tong, J. C., Tan, T. W., and Ranganathan, S. (2004) Modeling the structure of bound peptide ligands to major histocompatibility complex. Protein Sci. 13, 2523–2532
6. Michelin, O., and Karplus, M. (2002) Binding free energy differences in a TCR-peptide-MHC complex induced by a peptide mutation: a simulation analysis. J. Mol. Biol. 324, 547–569
7. Brannetti, B., Via, A., Cestra, G., Cesareni, G., and Helmer-Citterich, M. (2000) SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family. J. Mol. Biol. 296, 313–328
8. Altuvia, Y., and Margalit, H. (2004) A structure-based approach for prediction of MHC-binding peptides. Methods 34, 454–459
9. Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res. 31, 3635–3641
10. Tong, A. H., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Volkmer-Engert, R., and Cesareni, G. (2004) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295, 321–324
11. Honeyman, M. C., Brusic, V., Stone, N. L., and Harrison, L. C. (1998) Neural network-based prediction of candidate T-cell epitopes. Nat. Biotechnol. 16, 966–969
12. Brusic, V., Rudy, G., Honeyman, G., Hammar, J., and Harrison, L. (1998) Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. Bioinformatics 14, 121–130
13. Dönnes, P., and Elofsson, A. (2002) Prediction of MHC class I binding peptides, using SVMHC. BMC Bioinformatics 3, 25–38
14. Bhaisin, M., and Raghava, G. P. (2004) SVM based method for predicting HLA-DRβ1*0401 binding peptides in an antigen sequence. Bioinformatics 20, 421–423
15. Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A., and Stevanovic, S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50, 213–219
16. Martin, S., Roe, D., and Faulon, J. L. (2005) Predicting protein–protein interactions using signature products. Bioinformatics 21, 218–226
17. Boeckmann, B., Barocho, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O’Donovan, C., Phan, I., Pibout, S., and Schneider, M. (2003) The SWISS-PROT protein knowledge-base and its supplement Trembl in 2003. Nucleic Acids Res. 31, 365–370
18. Schultz, J., Milpetic, F., Bork, P., and Ponting, C. P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. Proc. Natl. Acad. Sci. U. S. A. 95, 5857–5864
19. Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., and Westbrook, J. (2000) The Protein Data Bank and the challenge of structural genomics. Nat. Struct. Biol. 7, (suppl.) 957–959
20. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680
21. Kyte, J., and Doolittle, R. F. (1982) A simple method for displaying the hydroptic character of a protein. J. Mol. Biol. 157, 105–132
22. Bull, H. B., and Breeze, K. (1974) Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues. Arch. Biochem. Biophys. 161, 665–670
23. Chothia, C. (1975) Structural invariants in protein folding. Nature 254, 304–308
24. Bhaskaran, R., and Ponnuswamy, P. K. (1994) Dynamics of amino acid residues in globular proteins. Int. J. Pept. Protein Res. 43, 180–191
25. Zimmerman, J. M., Eliezer, N., and Simha, R. (1968) The characterization of amino acid sequences in proteins by statistical methods. J. Theor. Biol. 21, 170–201
26. Betancourt, M. R., and Thirumalai, D. (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci. 8, 361–369
27. Vapnik, V. N. (2000) The Nature of Statistical Learning Theory, 2nd Ed., Springer, New York
28. Joachims, T. (1999) Making large-scale SVM learning practical, in Advances in Kernel Methods—Support Vector Learning (Ikofl, B., Burges, C., Smola, A., eds.) pp. 169–184, MIT Press, Cambridge, MA
29. Matthews, B. W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442–451
30. Boisguerin, P., Leben, R., Ay, B., Radziwill, G., Moelling, K., Dong, L., and Volkmer-Engert, R. (2004) An improved method for the synthesis of cellulosic membrane-bound peptides with free C termini useful for PDZ domain binding studies. Chem. Biol. 11, 449–459
31. Wiedemann, U., Boisguerin, P., Leben, R., Leitner, D., Krause, G., Moelling, K., Volkmer-Engert, R., and Oschkinat, H. (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. J. Mol. Biol. 343, 703–718
32. Landgraf, C., Panni, S., Montecchi-Palazzi, L., Castagnoli, L., Schneider-Mergener, J., Volkmer-Engert, R., and Cesareni, G. (2004) Protein interaction networks by proteome peptide scanning, PILOs Biol. 2, 94–103
33. Beuming, T., Skrabaneck, L., Niv, M. Y., Mukherjee, P., and Weinstein, H. (2005) PDZBase: a protein–protein interaction database for PDZ-domain proteins. Bioinformatics 21, 827–828
34. Baldi, P., and Brunak, S. (2001) Bioinformatics: the Machine Learning Approach, 2nd Ed., pp. 97, 115, and 126, MIT Press, Cambridge, MA
35. Hansen, L. K., and Salamon, P. (1999) Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. 21, 993–1001
36. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2005) GenBank. Nucleic Acids Res. 33, D34–D38
37. Baum, E., and Haussler, D. (1989) What size net gives valid generalization? Neural Comput. 1, 151–160
38. Perrone, M., and Cooper, L. (1993) When Networks Disagree: Ensemble Method for Neural Networks, Chapman-Hall, London
39. Guo, T., Hua, S., Ji, X., and Sun, Z. (2004) DBSubLocc: database of protein subcellular localization. Nucleic Acids Res. 32, D122–D124
40. Lo, S. L., Cai, C. Z., Chen, Y. Z., and Chung, M. C. (2005) Effect of training datasets on support vector machine prediction of protein-protein interactions. Proteomics 5, 876–884
41. Brusic, V., Bajic, V. B., and Petrovsky, N. (2004) Computational methods for prediction of T-cell epitopes—a framework for modelling, testing, and applications. Methods 34, 436–443