Predicting protein-ligand interactions from primary structure*

R. Bandyopadhyay†, X.-X. Tan‡, K. S. Matthews§, D. Subramanian¶
Rice University, 6100 Main Street, Houston TX 77005

January 2002

Abstract

Motivation: One of the key challenges in the post-genomic era is to understand protein-ligand interactions on a large scale. The question is: Given the primary structures of a protein and a ligand, how well can we computationally predict whether the ligand will bind to the protein? Wet laboratory experiments using combinatorial peptide screens and phage display techniques have yielded positive and negative examples of protein-ligand binding [SRH+96, ZPP+00, ACB95]. In this paper, we model the prediction of protein-ligand interactions from primary structure as a classification problem and train naive Bayes classifiers [Mit97] to distinguish between positive and negative examples of protein-ligand interactions Such a predictive model can screen large numbers of potential ligands and save laboratory time and costs.
Results: We demonstrate the power of our approach in predicting interactions between SH3 domains and proline-rich ligands. We use laboratory data gathered from combinatorial peptide library screening [SRH+96] of 8 diverse SH3 domains to construct a body of positive and negative examples. We learn naive Bayes models of ligand binding specificity of these SH3 domains and test them using a cross-validation approach. The models have prediction accuracies of 90% and higher with low false positive and negative rates. In addition, we visualize our classification model to reveal sites on both the ligand and the SH3 domain that contribute to the interaction. We use our classifiers to screen PxxP ligands from Swissprot for given SH3 domains. Over 80% of these ligands are eliminated by our naive Bayes classifiers for 5 of the 8 SH3 domains considered in this paper.

1 Introduction

Protein-protein interactions underlie almost all cellular processes, including DNA replication, transcription, splicing, translation, cell cycle regulation, signal transduction, and cell growth. Identification and analysis of protein-protein interactions is therefore essential for understanding the molecular mechanisms of cellular processes — a prerequisite for most pharmaceutical applications. A number of recognition modules within signaling proteins have been discovered in recent

*Contact author: Devika Subramanian devika@rice.edu. Software and data available upon request.
†Department of Computer Science and W. M. Keck Center for Computational Biology
‡Department of Biochemistry and Cell Biology and W. M. Keck Center for Computational Biology
§Department of Biochemistry and Cell Biology and W. M. Keck Center for Computational Biology
¶Department of Computer Science and W. M. Keck Center for Computational Biology

efforts to explicate intracellular components of signaling pathways [Sti97]. These modules, which include SH2, SH3, WW and EH domains, are believed to mediate protein-protein interactions by binding specific ligands in target proteins. Knowledge of domain-ligand binding can thus be used to predict such protein-protein interactions.

Predicting protein-ligand interactions is an important problem in its own right. Rational drug design research involves panning peptide databases for ligands that bind appropriately to specific proteins. Much of that work [KPO$^+$00] uses knowledge of three dimensional protein structure to find ligands that can dock to known active sites on the protein. Since detailed structural information is available only for a small fraction of known proteins, we focus instead on computational prediction of ligands for proteins for which we have no direct knowledge of three dimensional structure or the binding sites. Our goal, like that of [BG01, Mal99, TDN$^+$02] is to use sequence level information to predict domain-ligand interactions. We wish not only to develop prediction techniques with high accuracy, but also to characterize sequence-level criteria that distinguish interactions from non-interactions for a given target.

We analyze the family of proline-rich ligands that bind to SH3 domains. SH3 domains have now been identified in more than 1100 different proteins in organisms ranging from yeast to humans. SH3 domains are found in kinases, lipases, GTPases, adaptor proteins, structural proteins and others, and these proteins act in diverse processes including signal transduction, cell cycle regulation, and actin organization. Individual SH3 domains display distinct binding specificities. 8 SH3 domains have been extensively studied through combinatorial peptide library screening, yielding rich information on binding specificities of those domains with PxxP (x denotes any amino acid residue) containing ligands [ACB95, SRH$^+$96].

However, many questions remain in the study of SH3 domains and their interaction with PxxP containing ligands. What are the factors that determine the ability of different SH3 domains to recognize particular ligands? Can we predict which of the approximately 5.6×10^5 PxxP containing ligands in Swissprot and other databases bind to a given SH3 domain? Computational answers to these questions promise insights into the mechanisms of interaction between SH3 domains and PxxP ligands in particular, and protein-ligand interactions in general.

We formulate the problem of predicting whether a given SH3 domain drawn from a set S of SH3 domains interacts with a PxxP ligand drawn from a set P of PxxP ligands, as a *classification* problem. Given a labeled training set drawn from $S \times P \times \{+1, -1\}$ consisting of both positive and negative examples of SH3-PxxP interactions, we learn a function that will correctly classify unseen examples.

The success of classification learning on the problem of distinguishing SH3-PxxP interactions from non-interactions depends on several factors. First, the training data must be representative of SH3-PxxP interactions and non-interactions. We therefore need to obtain interaction data from a sufficiently diverse set of SH3 domains. Second, the interactions need to be represented in a form that allows the learning algorithm to express a function separating interactions from non-interactions. What makes this step challenging is that the available experimental data does not contain information on interaction sites for the SH3 domains complexed with all the binding ligands. Proper modeling of the interaction between the SH3 domains and the ligands is essential in the representation of the available data. Third, the right balance must be struck between generalization power and accuracy on the training data[Mit97]. No finite set of examples is sufficient to uniquely identify a function that discriminates interactions and non-interactions. It is especially important that we avoid learning patterns that are too specific to the available data. This challenge is the overfitting problem, or the bias variance tradeoff. We use a cross-validation protocol to address this problem for the naive Bayes classification technique.

Here then are the questions we seek to answer in this paper. (1). Given a set of known SH3-PxxP interactions and non-interactions, can we train a statistical machine learning algorithm to distinguish between them based solely on primary structure, and the physico-chemical properties of the residues in the protein and ligand? (2). How well can a learned classifier reproduce the known binding specificities of SH3 domains? How well can it predict the binding to unstudied ligands (e.g., ligands in Swissprot and other databases), and how can we assess its accuracy in the absence of laboratory data? and (3). Can we analyze the classifier to find the primary determiners of the recognition specificity between SH3 domains and their PxxP ligands? Are there a common set of factors, or is there a diversity of underlying mechanisms or recognition rules?

The paper is organized as follows. In Section 2, we derive a parametric model of interaction based on the work of [BVC$^+$00]. We describe a naive Bayes classifier that acquires the parameters of this model from the available data. In Section 3, we perform a series of computational experiments to understand the nature of the learned classifier. We describe differences in the conditional probability distributions learned by the naive Bayes classifier over positive and negative ligand data. The analysis of the differences yields insights into potential interaction sites on the domain and ligand. We then run a cross validation protocol to determine the accuracy of the learned classifier. We observe prediction accuracies of 90% and higher with fairly low false positive and negative rates. Finally, we use the classifiers to screen potential PxxP ligands from Swissprot and other databases. We demonstrate that over 80% of these ligands can be classified as non-binding for 5 of the 8 SH3 domains. This reduces the burden on wet laboratory experiments that map protein-ligand interactions.

2 Methods

2.1 Data

Our goal is to build classification models of protein-ligand interaction based on primary structure alone. The first step is to gather positive and negative examples of protein-ligand interactions and to represent them in a form that will permit induction of an appropriate classifier. Our data on experimentally observed SH3-PxxP interactions, summarized in Table 1, comes from one of the most comprehensive studies of binding preferences of SH3 domains in the literature [SRH$^+$96]. The study finds that a few SH3 domains cross-react with ligands selected by other domains, while most do not. The ligands of domain d_1 that do not cross-react with domain d_2 are chosen as negative examples for d_2 in our paper. It is important to note that our negative examples are true experimental observations. They are not generated by a computational process (e.g., randomized shuffles of sequences) as in [BG01]. The training examples are labeled *pairs* $((s, x), l)$ where s is a protein domain sequence and x is a ligand sequence. $l \in \{+_I, +_{II}, -\}$ indicates whether s binds to x or not. There are two classes of positive examples $+_I$ and $+_{II}$: some ligands bind in an N to C terminal orientation relative to the SH3 domain (class I ligands), and others bind in the opposite orientation (class II ligands). Table 1 shows class I and class II ligands for each of 8 SH3 domains in this paper.

Table 1: The set of SH3 domains and PxxP ligands used in our computational experiments [SRH+96]. There are 126 positive and 106 negative examples spanning 8 SH3 domains.

PDB code	SH3 name (length)	SH3 source	Positives		Negatives
			Class I	Class II	
1abo	Abl (61)	*Homo sapiens*	14	-	18
-	Crk (61)	*Homo sapiens*	-	7	14
-	Cortactin (55)	*Gallus gallus*	11	-	17
1gbq	Grb2 (56)	*Homo sapiens*	19	13	9
1ycs	p53bp2 (58)	*Mus Musculus*	-	13	17
2hsp	PLC-γ (54)	*Homo sapiens*	-	10	17
-	Yes (62)	*Gallus gallus*	11	-	7
1fmk	Src (63)	*Homo sapiens*	19	9	7

2.2 Simple feature vector representations and decision trees

Since most classification algorithms including decision trees and naive Bayes work on feature vectors, we find a feature vector representation of the training examples. We represent each amino acid by a vector $(v_1, \ldots, v_n)$ of n residue-level properties such as hydropathy and charge known to be relevant to protein-ligand interactions [Sti97]. These vectors are computed by the EMBOSS suite of programs [Emb]. A representation of an amino acid sequence is simply the concatenation of vectors that describe each individual amino acid as in [BG01]. Both the protein domains and the binding ligands in our data (see Table 1) are of varying length. The classification techniques we consider in this paper work with fixed length representations of examples.[1] Therefore, we need a technique for normalizing the lengths of the sequences. We multiply align all protein domain sequences so that they all have the same length. Then we align all ligand sequences together so that the core PxxP motif occurs in the same position in all the examples. The aligned sequences are then replaced by the property vector concatenation described above.

We build decision trees [Qui93] in a 10-fold cross-validation protocol which predict the ligand-binding specificity of SH3 domains with accuracies close to 100%. However, further inspection reveals that the discrimination between positive and negative examples is made on the basis of artifacts (such as the location of gaps) in the multiple alignment. The decision tree analysis points out the importance of going beyond predictive accuracies in assessing learned models of protein-ligand interactions. It also indicates that concatenation of normalized sequences representing the protein domain and the ligand is not sufficient to capture the significant residues in the interaction. To learn a residue-level interaction model *ab initio* requires significantly more data.

2.3 Features on residue pairs and the naive Bayes classifier

We therefore build a parametric model of protein-ligand interactions based on our current understanding of the binding of SH3 domains to proline-rich ligands using solved structures of SH3-PxxP complexes in the PDB. The parameters of such a model are learned from the available data, and the instantiated model is used for prediction. We learn a separate model for each

[1]Exceptions are hidden Markov models that can accept sequences of varying length.

SH3 domain as in [BVC+00]. Our model is composed of a set of features m_{ij} that characterize pairwise interactions between residues on the SH3 domain and the PxxP ligand [BVC+00]. Here, ij ranges over pairs of positions on the SH3 domain and the PxxP ligand known to be in contact. These contact position pairs are determined from available PDB data on complexes. Modeling the interaction with a common set of residue contact pairs ij for a given SH3 domain is justified by the fact that the structure of the domains appear to be remarkably conserved even in the complexed state [MSW94]. Then, given an SH3 domain and a new ligand, we calculate the features m_{ij} for this pair over all ij, and find the *maximum a posteriori* classification l for it.

$$l = argmax_{v \in \{+_I, +_{II}, -\}} P(v | m_{ij} \forall i, j)$$

We use Bayes Rules to rewrite this expression as

$$l \;=\; argmax_{v \in \{+_I, +_{II}, -\}} \frac{P(m_{ij}, \forall i, j | v) P(v)}{\sum_{w \in \{+_I, +_{II}, -\}} P(m_{ij}, \forall i, j | w) P(w)}$$

$P(+_I)$ is the prior probability of interaction between the SH3 domain s and a class I PxxP ligand p. It can be estimated directly from a large, representative training data set as the fraction of the known class I ligands that bind to the domain. $P(-)$ is the prior probability of non-interaction and it is $1 - P(+_I) - P(+_{II})$. The term $P(m_{ij}, \forall i, j | v)$ is the conditional probability of the conjunction of features m_{ij} over examples with the label $v \in \{+_I, +_{II}, -\}$. Direct estimation of this joint probability requires a combinatorially large data set. Since our training data is sparse, we make the *naive Bayes* assumption to estimate this conditional probability. We assume that each feature m_{ij} is independent of the other features given the classification. The biological import of the naive Bayes assumption is that we treat the overall binding between the domain and ligand as a set of independent interactions between pairs of residues on the domain and the ligand. The quality of predictions made by the model will be used to test the validity of this assumption.

The naive Bayes assumption allows us to replace the joint conditional probability by a product of individual feature conditional probabilities.

$$l \;=\; argmax_{v \in \{+_I, +_{II}, -\}} \frac{\prod_{i,j} P(m_{ij} | v) P(v)}{\sum_{w \in \{+_I, +_{II},, -\}} \prod_{i,j} P(m_{ij} | w) P(w)} \tag{1}$$

The conditional probability $P(m_{ij} = a | +_I)$ is the fraction of class I positive ligands with value a for the feature m_{ij}, while $P(m_{ij} = a | -)$ is the fraction of negative ligands with that property. They can be estimated by a simple counting procedure over the training data.

We now describe the feature m_{ij} over contacting residue position pairs ij where i is on the SH3 domain and j is on the ligand. We only examine the hydropathy and charge of the residues in pair ij. We consider discrete hydropathy values h of -1 (hydrophilic) and 1 (hydrophobic) and discrete charge values c of +1 (positive), 0 (neutral) and -1 (negative). We estimate $P(h_i, h_j | v)$ and $P(c_i, c_j | v)$ for SH3 domain s directly from the training data. Each m_{ij} has four hydropathy pair probabilities and nine charge pair probabilities, for each classification v as shown in Figure 1. However, there are only 11 independent probability parameters for each ij. Thus, if there are N contact pairs in our model, we will estimate $33 * N$ parameters for each of three possible values of v.[2] We make the assumption that the charge and hydropathy distributions for each

[2] In contrast, [BVC+00] estimate $400 * N$ parameters from just the positive ligand data.

contact pair are independent and take their product to determine $P(m_{ij}|v)$, i.e., $P(m_{i,j}|v) = P(h_i h_j|v) * P(c_i c_j|v)$.

A technical difficulty arises in these computations because of the sparsity of the m_{ij} entries. If we never observe a particular combination of hydropathies at a particular contact position pair ij for a class v, the corresponding estimated probability $P(h_i h_j|v)$ will be 0. A 0 probability will dominate the Bayes classifier described in Equation 1. To avoid this difficulty, we adopt a Bayesian approach to estimating probability. Let n be the total number of training examples, and let n_v be the number that belongs to class v. The probability of membership in class v is n_v/n. However, if $n_v = 0$, we use the m-estimate $\frac{n_v+mp}{n+m}$, where p is a prior estimate of the probability we wish to determine, and m is the equivalent sample size.[3]

In summary, our classification algorithm involves two steps. First, for each of the 8 SH3 domains, we use the labeled training data to learn $P(v)$ (priors) and individual feature conditional probabilities $P(m_{ij}|v)$ for the 3 classes: class I ligand, class II ligand, and non-binding ligand. Then, the model for an SH3 domain s is used to classify a new ligand with equation 1 which chooses the most likely label given the learned model.

We now describe how contact pairs ij on the SH3 domain and PxxP ligand are chosen. We use the method outlined in [BVC$^+$00], which is summarized in Figure 1. First, the 8 SH3 domains are multiply aligned to normalize their lengths [Emb]. Then the contact residues on each domain are identified as in [BVC$^+$00]. This process yields 19 contact residues in 6 non-contiguous regions for the 8 SH3 domains. These are the "relevant" SH3 residues to consider in the model. On the ligand side, we use the scheme in [BVC$^+$00] based on the knowledge of consensus residues gathered from the experimental literature [ACB95, SRH$^+$96]. We align class I ligands as xxxxxx**P**xx**P** so that their PxxP cores occupy positions 7 through 10. For class II ligands, we align the ligands as xxx**P**xx**P**x so that their PxxP cores occupy positions 4 through 7. Negative examples for SH3 domains in the class I or II conformation are aligned as per the corresponding positive examples. For SH3 domains such as Src which bind to ligands in both class I and II conformations, negative examples are aligned in both modes.

The structural data makes clear that not all 19×10 or 19×8 pairwise contacts between residues in the SH3 domain and a PxxP ligand are feasible. The matrix in Table 1, taken from [BVC$^+$00] identifies the observed contact pairs in the PDB and labels them according to whether class I or class II ligands are involved in the interaction. These are the pairs ij over which the probability distributions of hydropathy and charge are estimated. Note that only 25% of the entries in the matrix are filled. Our empirical finding is that classification models which take all pairwise interactions into account, rather than the selected quarter of the interactions are swamped by noise added by the biologically implausible contact pairs.

3 Results and Discussion

We test our model construction and prediction algorithm on the SH3-PxxP data in Table 1. First, we determine and visualize (see Figure 2) differences, if any, in the conditional probability distributions $P(m_{ij}|v)$ for $v \in \{+_I, +_{II}, -\}$ which characterize ligands that bind to SH3 domains in a class I or class II conformation, and those that do not bind. The detection of significant statistical differences between the learned distributions for the 3 classes provides evidence for the

[3]We assume uniform priors and set $p = 1/k$, where k is the number of possible values of the feature. m determines how heavily to weight p relative to the observed data. We choose $m = 0.1n$ in our experiments.

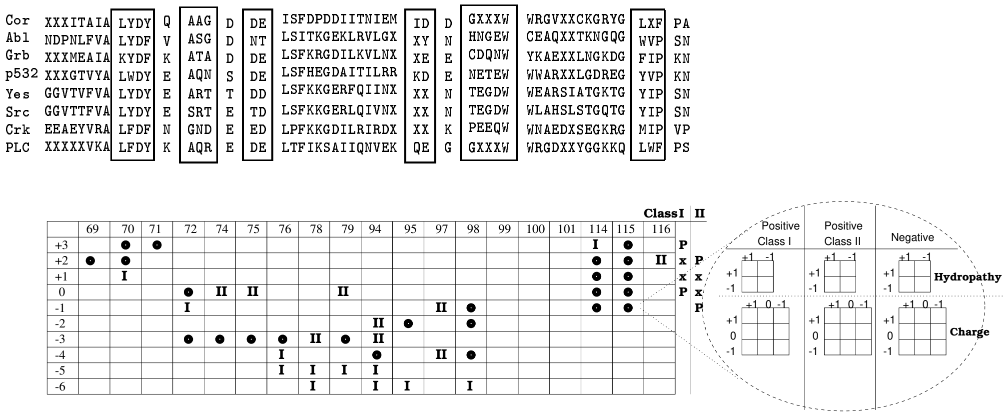

Figure 1: Multiple alignment of the 8 SH3 domains:(Table 1). The 6 boxed regions of contiguous residues are in contact with a PxxP ligand in some SH3-PxxP complex in the PDB. The interaction matrix has columns corresponding to the 19 SH3 contact residues from the boxed regions above. Each entry indicates the type of contact between an SH3 residue and a ligand residue. This table is adapted from [BVC+00]. The conditional probability distributions associated with each filled cell in the matrix on hydropathy and charge for the 3 classes I, II and - are schematically indicated.

appropriateness of the chosen features m_{ij}. It also ensures that classification methods such as the naive Bayes algorithm will generate good predictions. Second, we test the accuracy of predictions (see Table 3) produced by the naive Bayes algorithm using a standard n-fold cross-validation protocol [Mit97]. The ultimate test of a prediction algorithm is its performance on unseen data. We use an ensemble of learned classifiers (see Table 3) to classify the approximately 5.6×10^5 PxxP ligands extracted from Swissprot. The benefit of a classification model such as naive Bayes is that we can directly predict both binding and non-binding ligands without setting arbitrary thresholds on probability of binding to identify them.

3.1 Characterizing differences between positive and negative ligands

Our model of interaction captures the differences between the positive (class I and class II) and negative ligands in a set of probability distributions of the form $P(m_{ij}|v)$ where $v \in \{+_I, +_{II}, -\}$. We use the *Kullback-Leibler (KL)* divergence defined as follows

$$KL(p,q) = \sum_{x \in S} p(x) log \left(\frac{p(x)}{q(x)} \right)$$

to measure the difference between distributions p and q defined over a common basis set S.[4] It is always non-negative and is zero only when $p = q$. Information-theoretically, KL-divergence represents the cost of representing a distribution p using another distribution q. Since KL-divergence is non-symmetric, we adopt the symmetric form $sKL(p,q) = KL(P,q) + KL(q,p)$ which can be loosely treated as a measure of the "distance" between distributions.

Recall that we have distributions of the form $P(m_{ij}|v)$ for each of contact position pairs ij shown in the matrix in Figure 1. For every ij, we compute the $sKL(P(m_{ij}|+_I), P(m_{ij}|-))$ (resp.

[4]Note that $0 log \frac{0}{0} = 0$ and $p log \frac{p}{0} = \infty$.

$sKL(P(m_{ij}|+_{II}, P(m_{ij}|-)))$ for SH3 domains to which ligands bind in a class I conformation (resp. class II conformation). Since we have 19 potential contact residues on the SH3 domain and up to 10 residues on the ligand, we aggregate contiguous residues on the SH3 domain into six regions to aid in the visualization of the distributional differences between positive and negative ligands. The sKL divergence for an SH3 region r and a ligand position j pair is computing by averaging the divergences over all the contact position pairs ij where position i is in the region r. We then have a 6×10 matrix with each entry containing a number that characterizes the distance between the charge and hydropathy distributions of positive and negative ligands for that region/position pair. Figure 2 shows the aggregated sKL-divergence over the 6 regions on the Src SH3 domain (resp. p53bp2 SH3 domain) and 10 (resp. 8) residues on positive and negative class I (resp. II) ligands. The annotated peaks in Figure 2 have sKL-divergence values higher than two standard deviations over the mean.

The peaks in Figure 2 identify the residue pairs on the SH3 domain and PxxP ligand where there is the most difference in charge and hydropathy distributions between positive and negative ligands. We interpret them to be the most likely contact regions between ligands and the SH3 domain. It confirms conventional wisdom [BVC+00, KWS00] that the interactions between the SH3 domain and the ligand is limited to a few non-contiguous residues on the SH3 domain. Since positive and negative ligands are aligned the same way, there are no significant differences between them with respect to their PxxP core. All differences are constrained to positions to the left of the PxxP core for class I ligands. In particular, ligand positions 4 (P_{-3}) and 6 (P_{-5}) for the Src SH3 domain are likely to be involved in binding, which is confirmed by [BVC+00]. For class II ligands, one position to the right of the PxxP core and two to the left appear to be significant. For p53bp2, ligand residues in position 2 (P_{+3}) and 8 (P_{-3}) appear to be most likely participants in the interaction. These are consistent with current knowledge of the binding specificities of the two domains [SRH+96, BVC+00].

A summary of the peaks in sKL-divergence for all 8 SH3 domains is reported in Table 2. We were able to confirm all ligand position interactions in the table from the literature [ACB95, KWS00, BVC+00, SRH+96]. Our peak matrix also indicates that the most common mechanism for class I ligand binding is for ligand position 4 to interact with regions 2 and 3 (residues 74-79) of the SH3 domain, and ligand position 6 to interact with regions 5 and 6 (residues 97-101 and 114-116) of the SH3 domain. SH3 domains of Abl and Cortactin show individual differences.[5] For class II ligand binding, the most common mechanism is for ligand position 2 to interact with region 3 (residues 78-79) of the SH3 domain. PLC and p53bp2 have individual differences. What is striking is that our analysis uses very coarse information about structure and all available data from positive and negative ligands for 8 SH3 domains, and is able to arrive at a fairly detailed picture of the sites on both the SH3 domain and the ligand that participate in the binding.

3.2 Testing prediction quality using cross-validation

We test the performance of our naive Bayes classifier for each SH3 domain using a 4-fold[6] cross validation protocol. Cross-validation is a standard technique used when the amount of training data is not large enough to have an independent test data set. For each SH3 domain we merge

[5] The interaction between ligand position 1 and residues 94-95 in region 4 on the Abl SH3 domain is possibly a spurious peak caused by a preponderance of proline in the first position in the small training sample.

[6] We also ran 5-fold and 10-fold cross validation studies and the results are within 2% of the performance numbers shown in Table 3.

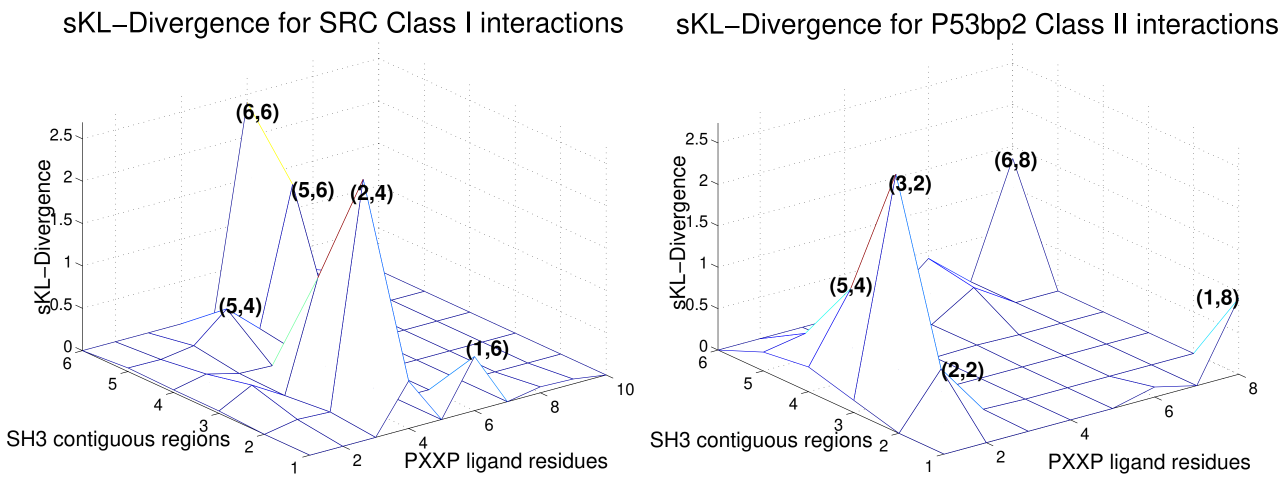

Figure 2: The sKL-divergences between distributions for positive and negative ligands for the Src SH3 domain and the p533bp2 domain. The peaks identify the SH3 region-PxxP residue pairs where there is the most difference in charge and hydropathy distributions between positive and negative ligands.

	1	2	3	4	5	6	7...10
1							
2				A,S, C,Y			
3		A		A,S C,Y			
4	A,G C						
5	G				C	G,S Y	
6						G,S Y	
							PXXP

	1	2	3	4	5	6	7	8
1								
2								
3		K,G S,B						
4		K,G	S,P					
5					S,P			
6								B
				P	X	X	**P**	

Table 2: SH3 region/ligand residue pairs displaying statistically significant differences between the positive and negative ligands for Class I (left) and Class II (right). SH3 domains: Abl (A), Grb2N (G), Cortactin (C), Src (S), Yes (Y), CrkN (K), P53bp2(B) and PLC (P)

the available positive and negative ligand data and shuffle them randomly. The shuffled data is split into 4 chunks. A classifier is learned on 3 of the chunks and the set aside chunk is used as a test set for assessing the accuracy, precision (100−false positive percentage) and recall (100−false negative percentage) of that classifier. Each chunk is used as a test set in turn and performance numbers are averaged over the four runs.

Our naive Bayes classifier estimates the conditional probabilities $P(m_{ij}|v)$ for each of the 3 choices ($+_I$, $+_{II}$, and $-$) for class v, as explained in the previous section. However, we cannot use the proportion of the 3 classes in the training data for the priors $P(v)$ because the data set is not representative of the class probabilities. About half the ligands for each SH3 domain are positive, yet the common belief about the prior probability of a PxxP ligand binding with an SH3 domain is not 0.5! Since no objective measure of the prior probability on the classes is available, we experiment with choices of the priors ranging from 0.1 to 0.8 in steps of 0.1. The results are displayed in Figure 3. This figure shows that accuracy, precision and recall of the naive Bayes classifier are remarkably robust to the choice of the priors in this range. That is, the conditional probability distributions $P(m_{ij}|v)$ are sufficient to determine the class labels for this range of priors.

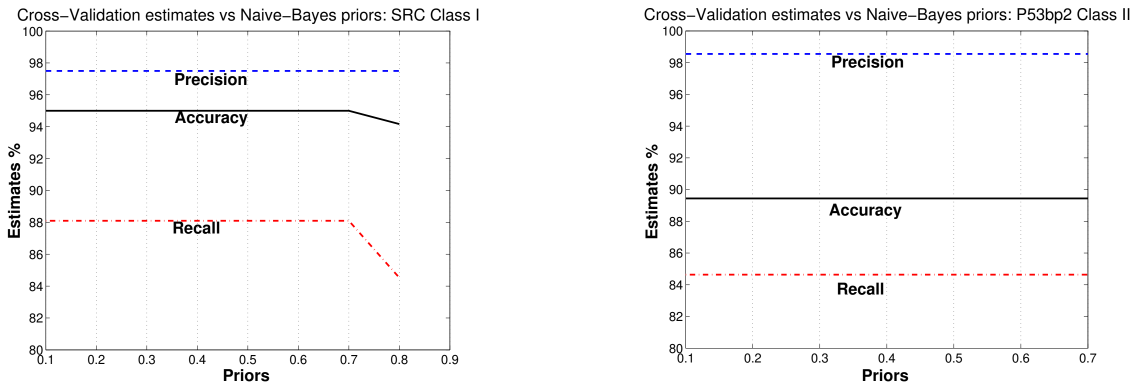

Figure 3: Prediction accuracy, precision and recall of naive Bayes classifier are remarkably insensitive to choice of prior probabilities on class labels.

Table 3 shows the performance of our naive Bayes classifier for each of the 8 SH3 domains. The prior probability of a positive ligand is taken to be 0.1 and a 4-fold cross-validation protocol is used to generate the results. Note that the accuracy of prediction is over 80% for all but one SH3 domain: Src SH3 binding to ligands of class II. The reason for the lower mean and higher variance in prediction accuracy for this class is the low sample size both for positive and negative examples. Further, the existing examples are too diverse for the naive Bayes classifier to learn feature distributions conditioned on the class that are sufficiently "far apart". The ligand binding specificity of 4 of the 8 SH3 domains can be predicted with accuracies greater than 90%. As the sample sizes of positive and negative examples for each of the SH3 domains grows, the variances reported in the table for prediction accuracy will get smaller. The false positive rate is under 10% for 5 of the 8 SH3 domains. The outliers are Cortactin with class I ligands, Src with class II ligands, and PLC with class II ligands. For all of these cases there is imbalance in the available set of positive and negative examples The false negative rates are higher than the false positive rates. However, this result is an artifact of the simple classification rule (equation 1) that we use. By adjusting the decision rule to vote for a class only if the others are separated from it by a given probability margin, we can get the two rates to be equal. The high variances in both false positive and negative rates are the result of low sample sizes (see Table 1). As more data becomes available, we expect the variance to decrease considerably. To test this conjecture, we used all the available data to construct a single SH3-PxxP interaction model, instead of building a model for each SH3 domain. The performance of this model is also tested in a cross-validation protocol, and the results are in the row labeled `All` in Table 3. Note that the variance in performance with more data is much much lower. The mean prediction accuracy is lower, however, indicating that a one-size-fits-all model for capturing SH3-PxxP interactions is inherently less accurate than an SH3-specific scheme. The overall results of Table 3 do vindicate the naive Bayes assumption of treating the individual residue pair interactions as being conditionally independent.

3.3 Screening ligands in Swissprot

The real utility of a learned classifier is a function of how effective it is in screening ligands for SH3 domains for testing by wet laboratory experiments. We predict the class labels for approximately 110,000 PxxP ligands sampled from Swissprot. To check whether a PxxP ligand can bind in a class I (resp. class II) conformation with a specific SH3 domain, we align it as xxxxxxPxxP (resp.

Table 3: Left: Performance of 8 SH3 domains under 4-fold cross-validation (prior=0.1). Right: Percentages of PxxP ligands in Swissprot predicted to bind in the class I and class II mode Column A is for $P(+_c|\prod_{ij} m_{ij}) = 1$ and column B is for $P(+_c|\prod_{ij} m_{ij}) > 0.999$.

SH3	Class	Accuracy	Precision	Recall
Src	I	87.5 ± 13.5	90.7 ± 11.4	81.7 ± 29.8
Grb2N	I	93.2 ± 9.5	98.8 ± 5.3	82.9 ± 30
Yes	I	92.9 ± 11.3	100 ± 0	76.9 ± 37.3
Abl	I	85.4 ± 6.9	94.9 ± 9	80.3 ± 10.4
Cortactin	I	80.3 ± 10.6	71.4 ± 31.8	83.6 ± 16.2
Src	II	77.1 ± 23.9	75.7 ± 34.3	83.3 ± 25.8
Grb2N	II	97.6 ± 6.4	94.1 ± 10.1	100 ± 0
CrkN	II	100 ± 0	100 ± 0	100 ± 0
PLC	II	81 ± 14.8	62.3 ± 41.3	91.7 ± 16.1
p53bp2	II	85.1 ± 14.8	89.5 ± 20.1	81.7 ± 23.4
All	I	74.8 ± 3.5	80.8 ± 1.5	66.7 ± 9.1
All	II	75.9 ± 5.6	51.6 ± 13.2	95.8 ± 2.9

SH3	% $+_I$	% A	% B
Src	39.77	8.02	18.09
Yes	33.4	9.65	26.39
Grb2N	40.59	8.94	30.33
Abl	17.21	7.5	12.37
Cortactin	10.58	0.4	4.61
SH3	% $+_{II}$	% A	% B
CrkN	21.92	9.7	15.21
p53bp2	3.54	0	1.15
Grb2N	20.46	0	10.32
PLC	9.63	0	7.5
Src	38.69	0.8	19.06

xxxPxxPx) and run a cross-validated committee of naive Bayes classifiers learned from positive and negative ligands for that SH3 domain. A cross-validated committee of naive Bayes classifiers is a set of n classifiers learned during n-fold cross-validation. The final classification for each ligand is the majority vote among the n classifiers. In Table 3, we report the fraction of ligands for each SH3 domain that are judged to be in $\{+_I, +_{II}\}$ by a cross-validated committee of 3 naive Bayes classifiers trained on the data in Table 1. The classifiers for Abl, CrkN, p53bp2, Cortactin and PLC screen 80% or more of the ligands in Swissprot and other databases. It appears that Src, Grb2N and Yes are more promiscuous in ligand binding, since the classifiers are only able to screen out 60% or more of these ligands. As more data accumulates, we expect our estimates on screening power to be more reliable. In columns 2 and 3 of Table 3 we show the effect of changing the decision-making criterion on our classifiers. Instead of using equation 1, we let the classifier label a ligand as positive only if its posterior probability $P(+|m_{ij}, \forall ij)$ is identically one (resp. > 0.999 for column B). These stringent criteria which may be more appropriate at this time because of the paucity of data allow the classifier to screen out 70% to 99% of the ligands.

3.4 Conclusions

The problem of predicting protein-ligand interactions from primary structure has been addressed by [BVC+00] and [Mal99]. [Mal99] performs a discriminant analysis of a very large database of ligands that bind to MHC molecules. The interaction model chosen by [Mal99] is of a peptide sliding over the MHC molecule, while our interaction model, derived using the methodology of [BVC+00], involves non-contiguous residues on the SH3 domain. In [BVC+00], prediction of binding specificities is done on the basis of positive ligands alone, since it is not a classification model. We believe that the use of negative ligands in the construction of a discriminant model allows for more effective screening of unseen ligands, and more detailed characterizations of the interaction between SH3 domain and PxxP ligand.

We now revisit the three questions raised at the start of the paper. We have shown that the naive Bayes classification algorithm can be used to learn a high-performance classifier that successfully predicts binding specificities of individual SH3 domains. The classifier uses knowledge of key residues on the SH3 domains and PxxP motif-containing ligands. Charge and hydropathy distributions on residue pairs are sufficient to discriminate between positive and negative ligands for an SH3 domain. The learned classifier is also effective at screening 70-90% of unseen ligands in the Swissprot database for most of the SH3 domains. The sKL-divergence analysis of distributional differences between positive and negative ligands reveals commonalities in binding mechanisms for the SH3-PxxP interaction as well as individual differences for particular SH3 domains. In future, we shall incorporate more experimental interaction data and other information regarding species, tissue, and function of proteins to improve prediction. We have demonstrated that the naive Bayes classifier packs enough punch to eliminate a vast number of costly experimental verifications in spite of the limited experimental data.

Acknowledgements

This work is supported by NSF IRI-9796046 (to D. Subramanian), and NIH GM22441 (to K. S. Matthews). X.-X. Tan is supported by a training fellowship from the Keck Center for Computational Biology (NLM Grant 5T15LM07093). R. Bandyopadhyay is supported by N00014-96-1-0538 from the ONR (to D. Subramanian).

References

[ACB95] K. Alexandropoulos, G. Cheng, and D. Baltimore. Proline-rich sequences that bind to Src homology 3 domains with individual specificities. *Proceedings of the National Academy of Sciences*, 92:3110–3114, 1995.

[BG01] J. R. Bock and D. A. Gough. Predicting protein-protein interactions from primary structure. *Bioinformatics*, 17(5):455–460, 2001.

[BVC+00] B. Brannetti, A. Via, G. Cestra, G. Cesareni, and M. H. Citterich. SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family. *Journal of Molecular Biology*, 298:313–328, 2000.

[Emb] EMBOSS: The European Molecular Biology Open Software Suite.

[KPO+00] T. Klabunde, H. M. Petrassi, V. B. Oza, P. Raman, J. W. Kelly, and J. C. Sacchettini. Rational design of potent human transthyretin amyloid disease inhibitors. *Nature Structural Biology*, 7:312 – 321, 2000.

[KWS00] B. K. Kay, M. P. Williamson, and M. Sudol. The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. *The FASEB Journal*, 14:231–241, 2000.

[Mal99] R. R. Mallios. Class II MHC quantitative binding motifs derived from a large molecular database with a versatile stepwise discriminant analysis meta-algorithm. *Bioinformatics*, 15:432–439, 1999.

[Mit97] T. M. Mitchell. *Machine Learning.* McGraw Hill, 1997.

[MSW94] A. Musacchio, M. Saraste, and M. Wilmanns. High resolution crystal structures of tyrosine kinase SH3 domains complexed with proline-rich peptides. *Nature Structure Biology*, 1:546–551, 1994.

[Qui93] J. R. Quinlan. *C4.5:programs for machine learning.* Morgan Kaufmann, 1993.

[SRH+96] A. B. Sparks, J. E. Rider, N. G. Hoffman, D. M . Fowlkes, L. A. Quiliam, and B. K. Kay. Distinct ligand preferences of Src homology 3 domains from Src, Yes, Abl, Cortactin, p53bp2, PLCγ, Crk, and Grb2. *Proceedings of the National Academy of Sciences*, 93:1540–1544, 1996.

[Sti97] W. E. Stites. Protein-protein interactions: interface structure, binding thermodynamics, and mutational analysis. *Chem. Rev.*, 97:1233–1250, 1997.

[TDN+02] A. H. Y. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. V. Hogue, S. Fields, C. Boone, and G. Cesarani. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295:321–324, 2002.

[ZPP+00] A. Zucconi, S. Panni, S. Paoluzi, L. Castagnoli, L. Dente, and G. Cesareni. Domain repertories as a tool to derive protein recognition rules. *FEBS Letters*, 480:49–54, 2000.