BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i274–i282
DOI: 10.1093/bioinformatics/bth922

Predicting protein–peptide interactions via a network-based motif sampler

David J. Reiss and Benno Schwikowski*

Institute for Systems Biology, 1441 North 34th street, Seattle, WA 98103-8904, USA

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: Many protein–protein interactions are mediated by peptide recognition modules (PRMs), compact domains that bind to short peptides, and play a critical role in a wide array of biological processes. Recent experimental protein interaction data provide us with an opportunity to examine whether we may explain, or even predict their interactions by computational sequence analysis. Such a question was recently posed by the use of random peptide screens to characterize the ligands of one such PRM, the SH3 domain.

Results: We describe a general computational procedure for identifying the ligand peptides of PRMs by combining protein sequence information and observed physical interactions into a simple probabilistic model and from it derive an interaction-mediated *de novo* motif-finding framework. Using a recent all-versus-all yeast two-hybrid SH3 domain interaction network, we demonstrate that our technique can be used to derive independent predictions of interactions mediated by SH3 domains. We show that only when sequence information is combined with such all versus all protein interaction datasets, are we capable of identifying motifs with sufficient sensitivity and specificity for predicting interactions. The algorithm is general so that it may be applied to other PRM domains (e.g. SH2, WW, PDZ).

Availability: The Netmotsa software and source code, as part of a general Gibbs motif sampling library, are available at http://sf.net/projects/netmotsa

Contact: dreiss@systemsbiology.org

1 INTRODUCTION

Peptide recognition modules (PRMs) are typically found in the context of larger multidomain signaling proteins or complexes. Their specific yet frequent binding events often direct the assembly and targeting of protein complexes involved in a wide range of key cellular processes (Zarrinpar *et al.*, 2003). They have therefore been implicated in a large number of human diseases, from cancer and Alzheimer's to Huntington's disease (Sudol and Hunter, 2000). The SH3 domain is among the most numerous, and most actively studied and widely-understood PRMs to date (Mayer, 2001). Many investigations, using high-resolution structure determination, phage display, and combinatorial chemistry, have revealed the preferred ligands of various specific SH3 domains (Brannetti *et al.*, 2000; Kay *et al.*, 2000, and references therein).

It has been found that the peptide ligands of many PRM domains, including SH3, consist of a proline-rich core. SH3 ligands in particular contain a characteristic **PxxP** consensus (**x** signifies an arbitrary amino acid). Upon further scrutiny, it is observed that the ligands may be classified into two primary consensi, depending upon the orientation of the peptide's binding to the surface of the domain: class-I (**+xΦPxΦP**) and class-II (**ΦPxΦPx+**), where **Φ** is a hydrophobic residue, often leucine or isoleucine; **+** denotes a basic residue, most often arginine or asparagine (Mayer, 2001). Still more detailed studies reveal that the specific affinity for most individual SH3 modules may be ascribed to deviations in their individual ligand peptides from the standard core consensus, or to variations in additional important flanking residues. It has also been found that a few others do not conform to the consensus at all, probably relying upon higher-order structure, or other factors such as cell localization or mediation by other protein interactions or contacts to modulate their affinity (Mayer, 2001).

Tong *et al.* (2002) devised a strategy for examining interactions with SH3 domains on a large scale by combining genome-wide two-hybrid physical interaction tests with the computational prediction of interactions using motifs derived from phage display peptide screens. These two independently derived interaction networks could be compared to each other to derive an 'overlap network', containing only the most significant interactions. Moreover, by identifying the consensus target motifs for each SH3 module, the technique provided a means of identifying the most likely target regions (binding sites) on each SH3 interaction partner.

The work of Tong *et al.* (2002) lends itself naturally to the question of whether the SH3 ligand peptides may also be found using one of the *de novo* motif finding algorithms that have been developed over the past few years, most often for identifying putative transcription factor binding sites in regulatory regions of co-expressed genes (e.g. Bailey and Elkan, 1994; Lawrence *et al.*, 1993). This would provide the clear advantage of allowing us to either specifically target, or perhaps even bypass altogether, some of the difficult and expensive experimental techniques.

*To whom correspondence should be addressed.

The difficulty that arises in any such attempt is that one will not, in general, have more than a few interactions per domain. Restricting the analysis only to the very promiscuous domains would ignore a large fraction of the data. The problem is exacerbated because, as is known in the case of SH3 and several other domains, the consensus motif patterns are rather poorly conserved and would require many examples in order to be detected with any significance. Additionally, the two-hybrid network is known to contain a large number of false positives (Uetz and Hughes, 2000) that will add noise to the training data. The problem is complicated further because many other PRMs (e.g. WW, SH2, WH1) compete with SH3 to bind to proline-rich peptides; proline-rich motifs are therefore the most common sequence motifs in many genomes (Zarrinpar *et al.*, 2003). This is a classic example of trying to find relevant motifs in the 'twilight zone' where the targets are likely to be too subtle, disparate or poorly represented in small numbers to be identified using standard strategies. We argue below and demonstrate later that two such potential strategies, based on current motif-finding technologies, are poorly suited to handle this problem.

Strategy (A) would involve a search for a single motif in all identified SH3 binding partners in the two-hybrid interaction network. Such a method quickly converges to a short ($\sim$11 residue) polyproline pattern with small hints of higher order structure. Clearly, this result lacks the specificity to identify anything but a broad consensus pattern which might represent the ligand consensus. If we were to extend the search to more than one motif across the dataset, we would find it difficult to resolve even the two primary consensus classes for SH3, because they are so similar; individual instances of the motifs deviate more from the consensus models than the two class consensi differ from each other.

An alternate strategy (B) would be to search for a ligand motif pattern for each SH3 domain, in the sequences of the proteins that bind only to it. This will be even more difficult in general because the signal in the small number of binding partners of each domain ($\sim$9 on average, with as few as 1), can be expected to be obscured by a typically large number of false positives in the interaction data (Uetz and Hughes, 2000).

The clear path is to choose a middle ground between strategies (A) and (B). Whereas each SH3 module might not bind to a large enough number of proteins to enable its consensus motif to be detected, the network of overlapping sets of interaction partners suggests that there should be a complex pattern of differing levels of similarity between motif models of the different SH3 ligands. This pattern can serve as an additional constraint on the motif detection. In other words, we can choose a compromise strategy between the two methods described above, thereby enabling us to do better than either of the methods alone. We do this by using the network information as a prior on the structure of individual motifs, which we search for using a modified version of the Gibbs sampling algorithm described by Lawrence *et al.* (1993) and Liu *et al.* (1995).

2 THE TRAINING DATA

We use the Tong *et al.* (2002) SH3 yeast two-hybrid interaction network for our training set, although the techniques we develop are designed from the offset to be easily generalizable to networks modulated by any type or number of PRMs and identified by any experimental technique. This network contains 285 interactions between 28 SH3 proteins and 143 SH3 binding partners. Just as important, it is based on all versus all screen in which each SH3 protein was tested against all other proteins in yeast. Each SH3 module interacts with between 1 and 20 partners (average $\sim$ 9) with a roughly flat degree distribution, and each interactor binds to an average of $\sim$2 different SH3 proteins (with a steeply declining power-law degree distribution typical of other observed biological networks).

3 METHODS

3.1 The model

We approach this problem by constructing a probabilistic model describing the likelihood of generating the amino acid sequences of the binding partners of each PRM domain in an interaction network, and then using a Gibbs sampling algorithm to solve for the parameters of the model. We begin with some definitions. Formally, we model the network as a sparse matrix of edges, between a set of PRM domains, $\mathbf{D} = (d_1, d_2, \ldots)$, and a set of proteins with amino acid sequence $\mathbf{S} = (s_1, s_2, \ldots)$. s_j is a vector of residues of length L_j where the k-th residue in s_j is $s_{j,k}$. The edges define the non-zero entries in the matrix, $\mathbf{E} = (\epsilon_{i,j})$; $i = 1, \ldots, |\mathbf{D}|$; $j = 1, \ldots, |\mathbf{S}|$, where each edge $\epsilon_{i,j}$ corresponds to a real probability quantifying our belief in the interaction: $\epsilon_{i,j} = P(\text{interaction}_{i,j} = \text{true})$. Because here we only consider an interaction network derived from a single set of experiments, we use $\epsilon_{i,j} = 1$ if there is an observed interaction between d_i, and d_j; $\epsilon_{i,j} = 0$ otherwise.

Defining for any vector $\vec{v}$, $|\vec{v}| \equiv \sum_i v_i$, each domain d_i is connected via $|\mathbf{E}_i|$ edges to $|\mathbf{E}_i|$ target protein sequences ($\mathbf{E}_i = |\epsilon_{i,1}, \ldots, \epsilon_{i,|\mathbf{E}_i|}|^T$), and likewise each interactor sequence s_j is connected by $|\mathbf{E}_{\cdot j}|$ edges to $|\mathbf{E}_{\cdot j}|$ SH3 domains ($\mathbf{E}_{\cdot j} = |\epsilon_{1,j}, \ldots, \epsilon_{|\mathbf{E}_{\cdot j}|,j}|$). Where there is an interaction $\epsilon_{i,j}$, a binding site $\mathbf{A} = (a_{i,j})$ marks the start of a peptide of length w in s_j (residues $s_{j,a_{i,j}+1}, \ldots, s_{j,a_{i,j}+w}$) that binds to domain d_i. Two sites $a_{i,j}$ and $a_{k,j}$ in s_j that interact with domains d_i and d_k are considered independent. Therefore, s_j may have as many as $|\mathbf{E}_{\cdot j}|$ distinct binding sites, or as few as one. We may, however, add priors into our model if we believe that the two sites should have a higher probability (than random) of being the same.

The consensus binding pattern, or motif, for each domain d_i is modeled as a position-specific scoring matrix (PSSM). The PSSM $\Theta_i \in \mathbf{\Theta}$ is comprised of a w-length vector of independent multinomial distributions, $\theta_{i,j}$, giving the probability

of observing each of the $J = 20$ residues at position j in the motif. Θ_i is therefore a $w \times J$ matrix where $\sum_k \theta_{i,j,k} = 1$ for all j.

The residues in s_j that do not participate in any interactions (background residues) are drawn from a common multinomial distribution, θ_0. We generated θ_0, for this dataset, from the entire translated set of open reading frames (ORFs) in the *Saccharomyces cerevisiae* genome (NCBI, 2002, ftp://ftp.ncbi.nih.gov/refseq). Alternatively, if the dataset were larger and it was expected that the individual motifs were distinct, θ_0 could have been generated from only the SH3 ligand sequences, or even separately for each domain. A higher order Markov process might also be considered to generate the background distribution.

Residues in s_j where there is a binding event with domain d_i at site $a_{i,j}$, i.e. residues $s_{j,a_{i,j}+1}, \ldots, s_{j,a_{i,j}+w}$, are modeled by PSSM Θ_i:

$$P(s_{j,a_{i,j}+1}, \ldots, s_{j,a_{i,j}+w} | \Theta_i) = \prod_{k=1}^{w} \theta_{i,k,s_{j,k+a_{i,j}}}. \quad (1)$$

The likelihood of sequence s_j with binding events $\mathbf{E}_{\cdot j}$ to domains $\mathbf{D}_{\cdot j}$ (with PSSMs $\mathbf{\Theta}_{\cdot j}$) at the corresponding binding sites $\mathbf{A}_{\cdot j}$ may then be written as:

$$P(s_j, \mathbf{E}_{\cdot j}, \mathbf{A}_{\cdot j} | \Theta_{\cdot j}, \theta_0) \propto \prod_{l=1}^{L_j - w} \theta_{0,s_{j,l}} \prod_{i=1}^{|\mathbf{E}_{\cdot j}|} \left(\prod_{k=1}^{w} \frac{\theta_{i,k,s_{j,k+a_{i,j}}}}{\theta_{0,s_{j,k+a_{i,j}}}} \right)^{\epsilon_{i,j}}. \quad (2)$$

We do not exclude the possibility of overlapping binding sites for different domains (in fact, they may be common), and for the case of M such overlapping binding sites, we utilize a mixture of PSSMs, replacing the single motif model $\theta_{i,k,s_{j,k+a_{i,j}}}$ with a mixture of the overlapping motif models, offset by their corresponding binding locations: $\sum_{m=1}^{M} q_m \theta_{m,k,s_{j,k+a_{m,j}}}$. The mixture weights q_m, with $\sum_m q_m = 1$, are determined by the structure of the network, as described in Section 3.2.

The likelihood of the complete data, given the parameters, is

$$P(\mathbf{S}, \mathbf{A}, \mathbf{E} | \mathbf{\Theta}, \theta_0) = \prod_{j=1}^{|\mathbf{S}|} P(s_j, \mathbf{E}_{\cdot j}, \mathbf{A}_{\cdot j} | \Theta_{\cdot j}, \theta_0). \quad (3)$$

The main distinctions between our model [Equation (3)] and that for the common site sampler (Lawrence *et al.*, 1993), which assumes one motif instance per sequence [e.g. Equation (1) in Liu *et al.* (1995)] are that here we are counting over interactions (through their likelihood) rather than over sequences, and utilizing mixtures of motif models for cases of multiple overlapping motifs. Other than these details, the resulting conditional distributions which we use during the Gibbs sampling are identical [see Liu *et al.* (1995) for their derivation].

3.2 The motif prior

The Gibbs sampling algorithm enables us to sample over individual conditional probabilities, updating prior expectations to posterior distributions and thereby sampling the joint likelihood. These conditional probability distributions are derived by Liu *et al.* (1995):

$$P(a_{i,j} | \mathbf{A}_{i,\hat{j}}, \mathbf{S}, \mathbf{E}) \propto \prod_{k=1}^{W} \frac{\bar{\theta}_{i,k,s_{j,k+a_{i,j}}}}{\bar{\theta}_{0,s_{j,k+a_{i,j}}}}, \quad (4)$$

where the $\bar{\theta}$ are the posterior means of θ, i.e. $\bar{\theta} \propto \int \theta P(\theta) d\theta$. $P(a_{i,j} | \mathbf{A}_{i,\hat{j}}, \mathbf{S}, \mathbf{E})$ are the predictive update distributions and $\mathbf{A}_{i,\hat{j}}$ denotes the set of all sites in all sequences other than $a_{i,j}$ (Liu *et al.*, 1995).

We define a $w \times J$ counting matrix $C_{i,j}$ for a chosen location $a_{i,j}$ in sequence s_j, as $C_{i,j,k,l} = \delta(s_{j,a_{i,j}+k} = l)$, and an alignment matrix over all sites that bind to d_i as $\mathbf{C}_i = \sum_j \epsilon_{i,j} C_{i,j}$, then we may use a mixture of Dirichlet distributions $\left[\sum_k q_k \mathcal{D}(\Theta_i | \mathbf{C}_i, \vec{\alpha}_k)\right]$ as a conjugate prior on the Θ_i. Then, we find that $\bar{\Theta}_i \propto \sum_k q'_k \mathcal{D}(\Theta_i | \mathbf{C}_i + \vec{\alpha}_k)$, where the $\vec{\alpha}_k$ are 'pseudocounts', which may be thought of as additional observations, added to the observed counts (Durbin *et al.*, 1998). Now, in addition to the Dirichlet mixture components of Sjolander *et al.* (1996) that capture chemical similarities between the residues, we can include further prior information by adding additional pseudocounts to the observed alignment counts.

Previous work on SH3 (Section 1) and other PRMs suggests that the binding peptides for most SH3 domains are similar. We capture this prior information by adding a global pseudocount component, $C_g = \sum_{i,j} \epsilon_{i,j} C_{i,j}$.

We may specify a further prior that captures the local pattern of binding that we see in the observed interaction network $\mathbf{E}$. By adding the prior assumption that binding sites on 'promiscuous' proteins are likely to bind to many different domains [which is hinted at by the enzyme-linked immunosorbent assay (ELISA) experiments on Las17 by Tong *et al.* (2002)], we would expect that models Θ_i for domains ($D_i \in \mathbf{D}_{\cdot j}$) that bind to protein s_j should be similar. This means that the Θ_i of those domains with a high degree of overlap in their binding partner sets would be more similar than those of two domains with distinct sets of partners. We incorporate this prior information into $P(\Theta_i)$ as an appropriately weighted set of pseudocounts that describes all alignment counts in $\mathbf{C}_{\cdot j}$: $\mathbf{C}_{\cdot j} = \sum_i \epsilon_{i,j} C_{i,j}$.

3.3 The discriminative prior

This model does not take full advantage of the fact that our rather unique training data, having been generated from an all-versus-all two-hybrid screen, contains explicit information on a large number of interactions that do not occur. This negative interaction information tells us that a putative binding site $a_{i,j}$ in sequence s_j, that binds to SH3 domain

d_i, must not only be (1) similar to the motifs $\Theta_{\cdot j}$, but it should also be (2) distinct from all $\Theta_{\cdot \hat{j}}$, where $D_{\cdot \hat{j}}$ are s_j's non-binding domains. Point (1) above is already included in the model as described in Section 3.2. Point (2) may be incorporated into the model through judicious use of a non-uniform site-based prior $P(a_{i,j})$. This type of prior distribution may, in general, be incorporated into our model [Equation (2)], in the exponent, with Equation (4) then becoming

$$P(a_{i,j}|\mathbf{A}_{i,\hat{j}}, \mathbf{S}, \mathbf{E}) \propto P'(a_{i,j}) \prod_{k=1}^{W} \frac{\bar{\theta}_{i,k,s_{j,k+a_{i,j}}}}{\bar{\theta}_{0,s_{j,k+a_{i,j}}}}, \qquad (5)$$

where $P'(a_{i,j})$ is the posterior mean distribution of $P(a_{i,j})$. Typical Gibbs samplers utilize a uniform prior, and thus $P'(a_{i,j}) = 1/(L_i - w)$. We instead use a non-uniform prior $P_d(a_{i,j})$, described below.

Such a prior should give higher probability to these sites that are distinct from the non-binding motifs $\Theta_{\cdot \hat{j}}$. This is a particularly difficult requirement, especially in the case of SH3 where all motifs (the $\Theta_{\cdot \hat{j}}$ as well as the $\Theta_{\cdot j}$) are known to be similar in most cases. The ideal preference may be stated like this: if two sites equally match the $\Theta_{\cdot j}$, then the one that is most dissimilar to the $\Theta_{\cdot \hat{j}}$ should preferentially be chosen.

We implement this simple expectation as follows: when a new site $a_{i,j}$ is to be sampled from $P(a_{i,j}|\mathbf{A}_{i,\hat{j}}, \mathbf{S}, \mathbf{E})$ [Equation (4)], we compute for that site $p_k = P(a_{i,j}|\Theta_k)$ via Equation (1) for each $\Theta_k \in \Theta_{\cdot j}$, and also for each $\Theta_{k'} \in \Theta_{\cdot \hat{j}}$. A comparison of these two sets of scores ($\mathbf{p}_k$ and $\mathbf{p}_{k'}$) against each other in which most of the $\mathbf{p}_k$ are greater than the $\mathbf{p}_{k'}$ should produce a favorable probability $P_d(a_{i,j})$. The significance of a Student's t-test or Wilcoxon rank test may be used to do this comparison (Siegel, 2003). We find that the rank test works best in our case where the number of elements in $\mathbf{p}_k$ is often small. In either case, the significance of the difference in distributions $P_d(a_{i,j})$ equals 0 for no distinction (when $\bar{\mathbf{p}}_k \leq \bar{\mathbf{p}}_{k'}$) or 1 when $\bar{\mathbf{p}}_k$ is significantly greater than $\bar{\mathbf{p}}_{k'}$.

The strength of our discriminative prior (i.e. the amount by which this discrimination influences the choices of binding sites) may be adjusted by adding a pseudocount q_d to the posterior distribution in Equation (5), i.e. $P'_d(a_{i,j}) = q_d + P_d(a_{i,j})$, and then renormalizing. A choice of 0 for q_d means that the discrimination (i.e. the result of the rank or t-test) will strongly influence our choice of a given site. Otherwise, a choice of, e.g. 10 for q_d means that this prior should account for ∼10% of the overall decision to choose the site.

3.4 The algorithm

The Gibbs sampling approach allows us to sample the joint distribution of our interaction model [Equation (3)] by iterating over each interaction (all $\epsilon_{i,j} \neq 0$) to choose the most probable value for $a_{i,j}$ by sampling from its conditional probability distribution when all remaining binding sites are left fixed [Equation (5)]. We start by choosing an initial (random) site in sequence s_j for each of the binding events (edges in the two-hybrid network; $\epsilon_{i,j} \neq 0$), and proceed to iterate over the edges, choosing a new binding site $a_{i,j}$ in s_j by sampling from Equation (5), after removing the previous $a_{i,j}$ from $C_{i,j}$. The PSSM, $\bar{\theta}_{i,j}$, used to compute this distribution is calculated, temporarily for each edge, from the alignment counts $\mathbf{C}$, using various pseudocounts derived in Section 3.2. In particular, $\bar{\theta}_{i,j}$ is computed from a dirichlet mixture of the individual pseudocount components described above ($C_{i,j}$), added and appropriately weighted:

$$C_{i,j} = \sum_k \epsilon_{k,j}\mathbf{C}_{k,j} + p_0 \sum_k \sum_l \epsilon_{k,l}\mathbf{C}_{k,l} + p_1 \sum_k \epsilon_{i,k}\mathbf{C}_{i,k}, \qquad (6)$$

Once a new $a_{i,j}$ is chosen, the corresponding counting matrix $C_{i,j}$ is updated and the procedure repeated on a new interaction.

The influence of the network-based components of the model, C_g and $\mathbf{C}_{\cdot j}$, on the overall procedure are adjusted simply by scaling their mixture coefficients, which we call q_g and q_1, respectively. These tunable parameters represent the user's degree of belief in the expectations, respectively, that all motif models should be similar on a global scale (global similarity of binding sites), and that the motif models for all SH3 domains that bind to a particular sequence should be similar (local network-informed similarity of binding sites). They may be seen as parameters which influence the degree of over- or under-fitting of the model to the data. In practice on the SH3 network, with a sufficiently high choice for p_1 (e.g. 10%), p_g is not required, and we choose $p_g = 0$. Once a new $a_{i,j}$ is chosen, the corresponding counting matrix $C_{i,j}$ is updated and the procedure repeated on a new interaction.

Following Lawrence *et al.* (1993), we compute the maximum a posteriori probability (MAP) estimate of the model given each sampled set of variables (including the priors), and use the highest scoring set that is obtained during a repeated number of iterations of the sampling procedure. A simplified summary of the algorithm is described in Figure 1.

4 RESULTS

4.1 Interaction prediction

Following Tong *et al.* (2002), we may estimate how well our computationally identified SH3 ligand motifs can be used to predict, or confirm, physical interactions. We may write the likelihood of an arbitrary sequence s_j conditioned on the fact that it binds to SH3 domain D_i with motif model Θ_i by integrating Equation (2) over all potential binding sites and

input: interaction network $\mathbf{E}$, protein sequences $\mathbf{S}$
for each $\epsilon_{i,j} \in \mathbf{E}$: randomly choose binding site $a_{i,j} \in \mathbf{A}$ in sequence $s_j \in \mathbf{S}$
repeat:
  for each $\epsilon_{i,j} \in \mathbf{E}$:
    add up counting matrices $\mathbf{C}_i$, $\mathbf{C}_{\cdot j}$, $\mathbf{C}_g$ (§3.2), ignoring the current site $a_{i,j}$ (*i.e.* using only sites $a_{i,\hat{j}}$)
    compute $C_{i,j}$ (*Eq.* (6)); add additional pseudocounts (Sjolander *et al.*, 1996)
    compute PSSM $\bar{\Theta}_{i,j}$ from $C_{i,j}$ (§3.2)
    compute the "discriminative posterior" $P'_d(a_{i,j})$ (§3.3)
    compute $P(a_{i,j}|\mathbf{A}_{i,\hat{j}}, \mathbf{S}, \mathbf{E})$ (*Eq.* (5)) using $\Theta_{i,j}$, s_j, and $P'_d(a_{i,j})$
    sample a new site $a_{i,j}$ from $P(a_{i,j}|\mathbf{A}_{i,\hat{j}}, \mathbf{S}, \mathbf{E})$
  Compute a new MAP score given the newly sampled sites
  Store the current parameters ($\mathbf{A}'$, Θ') if the MAP score is the largest yet seen
until a fixed number of iterations passes, or the best MAP score does not change for a fixed number of iterations
output: $\mathbf{A}'$, Θ'

Fig. 1. The network-based Gibbs sampling procedure.

applying Bayes' rule:

$$P(\epsilon_{i,j} = 1|s_j, \theta_0, \Theta_i) = \text{logit}\Bigg[\log\Bigg(\frac{P(\epsilon_{i,j}=1)}{P(\epsilon_{i,j}=0)}\prod_{k=1}^{L_j}\theta_{0,s_{j,k}} \times \sum_{l=1}^{L_j-W} P(a_{i,j})\prod_{m=1}^{W}\frac{\theta_{i,m,s_{j,l+m}}}{\theta_{0,s_{j,l+m}}}\Bigg)\Bigg], \quad (7)$$

where $\text{logit}(x) \equiv (1 + \mathrm{e}^{-x})^{-1}$. $P(\epsilon_{i,j} = 1)/P(\epsilon_{i,j} = 0)$ quantifies our prior expectation that there is indeed an edge between s_j and d_i. We use the observed ratio of edges to non-edges in the SH3 two-hybrid interaction network for this prior.

We can apply Equation (7) for an arbitrary protein sequence $s_{j'}$, using our derived SH3 ligand models Θ_i to compute $\epsilon_{i,j'} = P(\text{interaction}_{i,j'} = \text{true})$ for that sequence. We can then compute a predicted interaction network as Tong *et al.* (2002) did. For each predicted edge in our network, we ensure that the models were not learned using the sequence(s) and their corresponding interaction(s) being tested. Prior to computing $\epsilon_{i,j'}$, we therefore cull $s_{j'}$ and its interactions $\mathbf{E}_{\cdot j'}$ from the dataset, and re-learn the model parameters from this subset of the data. Repeating this procedure for all proteins in the two hybrid dataset allows us to construct a prediction network that is independent of the two-hybrid network. To directly compare our results to those of Tong *et al.* (2002), we choose a P-value cut-off for selecting interactions so that our network has the same number (394) of edges as their predicted network (Fig. 2).

The predicted network reveals a highly connected core complex centered on Las17, similar to the complex identified in Tong *et al.* (2002). The predictions of Tong *et al.* (2002), computed with ligand motifs obtained via phage display screens, resulted in a network of 394 interactions among 206 proteins, of which 59 also existed in the two-hybrid network (expected overlap of <1). We find consistently that our algorithm, with a P-value cut-off chosen to result in ∼400 interactions, identifies ∼50 interactions that overlap the two-hybrid network, for a range of the various user-tunable parameters (e.g. q_g, q_1 and q_d). This number is only slightly smaller than the overlap of Tong *et al.* (2002), a fact, which might be surprising considering that our training (and comparison) interaction dataset is based solely upon considerably noisy two-hybrid measurements (Uetz and Hughes, 2000). An example of such an 'overlap network' (between our computationally predicted network and the two-hybrid network) is shown in Figure 2b. It is apparent that the overlap network is also dominated by the core complex of SH3 domain proteins.

Interestingly, our predicted network does not overlap the (Tong *et al.*, 2002) predicted network by significantly more than it does with the observed (two-hybrid) network. The typical amount of intersection among the three networks is shown in Figure 3; whereas only ∼17% of each of the three networks intersect each other independently, about two-thirds of each overlap network agrees. This network of ∼35 interactions may perhaps be considered a truly high-confidence 'core network', which agrees among the three independent techniques.

To assess how well our algorithm performs relative to the two simpler algorithms (A) and (B) described near the end of Section 2, we ran the same procedure described above, using technique (A), with (1) a standard Gibbs site sampler and (2) MEME (Bailey and Elkan, 1994), and strategy (B) also with (1) the Gibbs sampler and (2) MEME. We found that the size of the 'overlap network' was consistently larger using our algorithm than it was for any of these four experiments ($p \lesssim 10^{-10}$). These results are summarized in column 2 of Table 1.

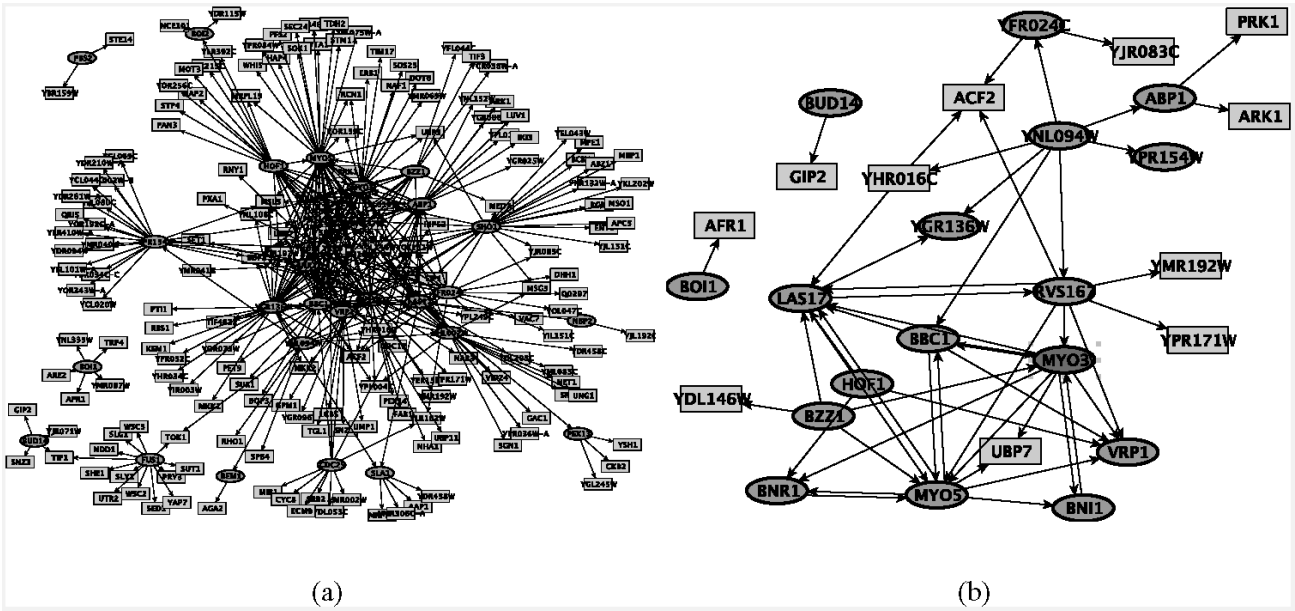

Fig. 2. Predicted (**a**) and overlap (**b**) SH3 interaction networks. Proteins containing SH3 domains are drawn as dark ovals; other interactors are light rectangles.

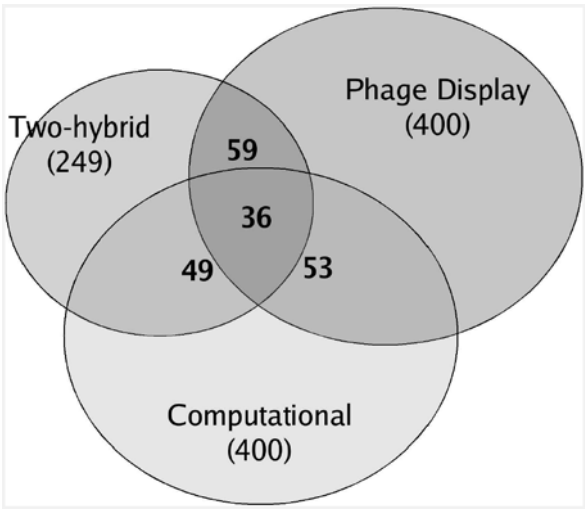

Fig. 3. Size of intersection between our computationally predicted network and the two networks (two-hybrid and predicted based upon phage display) of Tong *et al.* (2002).

Furthermore, we computed the rate of true positives against the rate of false positives (as measured against the two-hybrid network) over a wide range of predicted network sizes, to derive an receiver operating characteristic (ROC) curve. The area under the curve, for which a value of 0.5 represents no correlation at all, and 1.0 implies full correlation, was preferentially higher using our algorithm than it was for the four test cases described above ($p \lesssim 10^{-4}$; Table 1, column 3).

Table 1. Comparison of the results of our algorithm (row 5) against the two algorithms (A) and (B) described in Section 1 using a Gibbs sampler (1) and MEME (2).

Algorithm	Overlap interactions	ROC integral	Motif match scores
A, 1	40 ± 3.1	0.76 ± 0.010	0.47 ± 0.06
A, 2	27 ± 2.7	0.66 ± 0.011	0.41 ± 0.04
B, 1	41 ± 2.9	0.77 ± 0.008	0.41 ± 0.05
B, 2	35 ± 1.6	0.72 ± 0.013	0.42 ± 0.03
Our algorithm	49 ± 2.7	0.79 ± 0.008	0.55 ± 0.08

See text (Sections 4.1 and 4.2) for an explanation of the three columns.

4.2 Binding peptide consensus identification

We display a sample of the ligand motifs for each SH3 domain, identified by this algorithm, as motif logos (Schneider and Stephens, 1990, http://www.lecb.ncifcrf.gov/~toms/paper/logopaper/), in Figure 4. Clearly, the algorithm converges on proline-rich peptides (many even seemingly PxxP-like), even for domains in which there are very few interactions. We also see common SH3-binding residues, such as leucine, arginine and others (Section 1), often in their expected flanking positions surrounding or within the proline-rich core. However, polyproline strings clearly dominate the signal, and reveal a clear detriment of our technique. The same feature of the algorithm that directs the sampler to converge on proline-rich peptides also serves to weakens the effect of any higher order signal in the individual motifs. This is a classic example

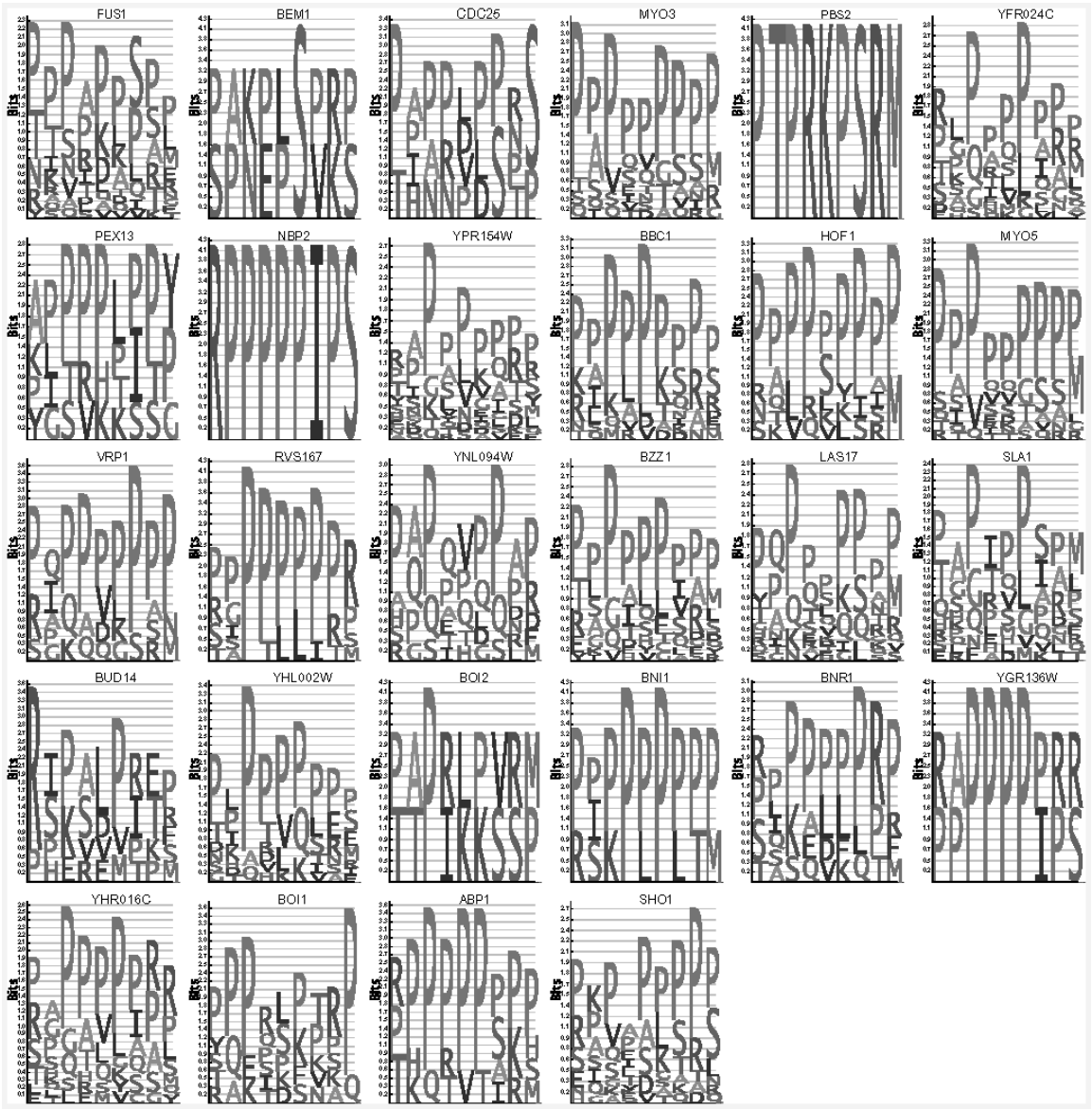

Fig. 4. Sample logos of the individual SH3 ligand motifs for each SH3 domain in the Tong *et al.* (2002) two-hybrid SH3 network, resulting from a single run of our algorithm.

of trying to find the best compromise between over- and under-fitting the model to the available (noisy) data.

We computed how similar our computationally derived consensus binding motifs (such as those displayed in Fig. 4) are to those computed from the phage display experiments of Tong *et al.* (2002) (their table 2). This was done by generating a set of PSSMs from their consensi, and computing the Pearson correlation coefficients of those PSSMs with our computationally derived motifs (Pietrokovski, 1996). This measurement, which lies in the range $[-1,1]$, was often higher for the consensi derived from our algorithm than it was using any of the four methods identified above ($p \lesssim 10^{-3}$; Table 1, column 3).

4.3 Binding site prediction

The likelihood of an interaction with an arbitrary binding site $a_{i,j}$, in sequence s_j (i.e. the binding occurs with residues $s_{j,a_{i,j}+1}, \ldots, s_{j,a_{i,j}+w}$) can be derived analogously to Equation (7), and results in

$$P(\epsilon_{i,j} = 1 \mid a_{i,j}, s_{j,a_{i,j}+1}, \ldots, s_{j,a_{i,j}+w}, \theta_0, \Theta_i) = \operatorname{logit}\left[\log\left(\frac{P(\epsilon_{i,j}=1)}{P(\epsilon_{i,j}=0)} \prod_{k=1}^{w} \frac{\theta_{i,k,s_{j,k+a_{i,j}}}}{\theta_{0,k,s_{j,k+a_{i,j}}}}\right)\right]. \quad (8)$$

The ratio $p_{\text{site}} \equiv P(\epsilon_{i,j}=1)/P(\epsilon_{i,j}=0)$ is a site-based prior, describing our expectation that any site in s_j is indeed a

binding site, given that an interaction does occur somewhere in the sequence. Where we have a prior expectation of n_{site} binding sites per interaction, we use a uniform prior distribution $p_{\text{site}} = n_{\text{site}}/(L_j - w)$, with $n_{\text{site}} \equiv 1$. This allows us, for a given domain D_i and interactor sequence s_j in which we predict $P(\epsilon_{i,j}) = 1$ via Equation (7), to identify putative binding sites up to a certain probability cut-off. We have chosen to perform this computational analysis on the protein Las17, whose binding sites with various SH3 domains were also determined experimentally (via ELISA experiments) by Tong *et al.* (2002).

A comparison of the most likely predicted binding sites on Las17 with its various predicted interactors based upon Equation (8) shows that we do not have the sufficient specificity to accurately predict binding sites using our probabilistic model ($p = 0.69$). When we performed the identical analysis using the phage display-derived motifs of Tong *et al.* (2002) we see a somewhat more significant ability to predict binding sites ($p = 0.28$). Even this result seems to be at odds with the analysis performed in the paper in which only one of 15 binding sites were incorrectly predicted. Such a disparity reveals one of the weaknesses of our model with regard to SH3: we use only one consensus for each domain (as opposed to two, which the phage display experiments are capable of resolving). Further, our combined model results in the blurring of the individually specific motifs, which seems to diminish the specific resolving power of the motifs in predicting individual binding sites. It should also be pointed out, however, that the particular ELISA experiments performed by Tong *et al.* (2002) on Las17 are subjected to some of the same potential systematic effects the phage display experiments are, which could be artificially enhancing the agreement between the two experimental methods in their work.

5 DISCUSSION

There have been several attempts to predict the ligands of SH3 and other PRM domains in the recent past, using methods, such as profile scanning (Obenauer *et al.*, 2003), neural networks (Chang and Page, 2002) and structural models (Brannetti *et al.*, 2000), with varying degree of success. All these techniques are specific to the particular system that is being investigated. We have described a method for identifying such ligands using only sequence and high-throughput interaction data, without requiring any additional prior assumptions on the system, or any type of structural information. We have shown that our technique is capable of characterizing the peptides that bind to sets of SH3 domains and thereby predicting which proteins these domains will interact with, nearly as accurately as the motifs derived from phage display experiments. However, our technique is not able to correctly identify the individual binding sites that the domains bind to.

There is clearly information that plays an important role in this system, that we are not including in our statistical model. As a simple example, additional prior information on the selection of interaction sites $a_{i,j}$, e.g. based on modeled or observed three-dimensional structures (such as residue burial predictions), or other sequence-based prior knowledge, may be incorporated into the model as a non-uniform prior $P(a_{i,j})$ (Section 3.3). More intelligent choices of motif priors (Section 3.2), such as inclusion of the PxxP signature, or some sort of discrimination between class-I and class-II motifs, would also probably help, although it would result in a loss of generality of the technique to other systems. We have also ignored the structure or sequence of the SH3 domains completely, and perhaps this is the ultimate limitation of our technique. Such information could potentially be included into the model, in a variety of ways (many of which, again, could result in loss of generality).

While we have tried to keep the algorithm as general as possible, there remain three user-tunable parameters (q_g, q_1, and q_d) that must be chosen for each training interaction data set. As with many algorithms, choosing the appropriate combination of parameters would be more of an art than a science, and if one were to apply this algorithm to a different dataset, choosing inappropriate parameters will result only in the effectiveness of the algorithm falling back to those of the standard Gibbs sampler or MEME (Table 1). In such a case, a good place to start would be to compare the target interaction network with that which we have used in this paper. For example, for a more loosely connected graph or one in which the motifs are expected to be more similar, one would increase q_g and decrease q_1 and q_d; and vice versa for a more tightly connected graph or one in which the motifs are expected to be dissimilar.

We have only tested our algorithm on the system of interacting SH3 domains in yeast, but it could potentially be used in the analysis of other PRM domains, such as WW, SH2, PDZ and Vasp, or in other species, once the results of any similar all versus all interaction screens become available. We believe that a major limitation on the performance of the algorithm lies in the quality of the interaction data that it is trained on. Reducing the false positive rate of the training data by incorporating positive (and negative, where available) interactions derived from lower throughput techniques can be expected to increase the predictive power of our method.

Finally, we believe we have developed a framework that is general and flexible enough that it could, with few modifications, be applied to completely new systems of interactions between various biomolecules. Such potential targets of this analysis could include other domain–peptide interaction systems (e.g. immune response interactions) and protein–DNA interaction sets (Lee *et al.*, 2002).

ACKNOWLEDGEMENTS

We gratefully acknowledge the generous input and suggestions of Stanley Fields, Becky Drees, Andrew Siegel and

Gary Bader. The interaction networks were rendered using Cytoscape (Shannon *et al.*, 2003).

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, June 1998. AAAI Press, Menlo Park, CA, pp. 28–36.

Brannetti,B. Via,A., Cestra,G., Cesareni,G. and Helmer-Citterich,M. (2000) SH3-spot: an algorithm to predict preferred ligands of different members of the SH3 gene family. *J. Mol. Biol.*, **298**, 313–328.

Durbin,R. Eddy,S., Krough,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

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

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwoald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lee,T., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae. Science*, **298**, 799–804.

Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156–1170.

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

NCBI (2002) Reference genomes curated by ncbi staff members. ftp://ftp.ncbi.nih.gov/refseq.

Obenauer,J.C., Cantky,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.

Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.

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

Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Siegel,A. (2003) *Practical Business Statistics, 5th edn.* McGraw-Hill/Irwin, New York.

Sjolander,K., Karplus,K., Brown,M., Hughey,R., Krough,A., Mian,I.S. and Haussler,D. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Bio sci.*, **12**, 327–345.

Chang,R. and Page,D. (2002) A neural network approach for studying the relationship between protein sequences and protein–protein interactions. In *ISMB2002*, Edmonton, Canada.

Sudol,M. and Hunter,T. (2000) New wrinkles for an old domain. *Cell*, **103**, 1001–1004.

Tong,A.H.Y., Dress,B., Nandelli,G., Bader,G.D., Brannetti,B., Castagnoli,L., Evagrgelista,M., Ferracuti,S., Nelson,B. and Paoluzi,S. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **295**, 321–324.

Uetz,P. and Hughes,R. (2000) Systematic and large-scale two-hybrid screens. *Curr. Opin. Microbiol.*, **3**, 303–308.

Zarrinpar,A., Bhattacharyya,R.P. and Lim,W.A. (2003) The structure and function of proline recognition domains. *Sci. STKE*, **179**, RE8.