A regularized discriminative model for the prediction of protein–peptide interactions

Wolfgang P. Lehrach\textsuperscript{1,2,*}, Dirk Husmeier\textsuperscript{2} and Christopher K. I. Williams\textsuperscript{1}

\textsuperscript{1}University of Edinburgh, Edinburgh EH1 2QL, UK and \textsuperscript{2}Biomathematics and Statistics Scotland, Edinburgh EH9 3JZ, UK

Received on August 4, 2005; revised on November 23, 2005; accepted on November 25, 2005

ABSTRACT

Motivation: Short well-defined domains known as peptide recognition modules (PRMs) regulate many important protein–protein interactions involved in the formation of macromolecular complexes and biochemical pathways. Since high-throughput experiments like yeast two-hybrid and phage display are expensive and intrinsically noisy, it would be desirable to more specifically target or partially bypass biochemical pathways. Since high-throughput experiments like yeast two-hybrid and phage display are expensive and intrinsically noisy, it would be desirable to more specifically target or partially bypass protein–peptide interactions mediated by SH3 domain proteins in Saccharomyces cerevisiae, where the datasets were obtained with different experimental techniques. The predictions were evaluated with out-of-sample receiver operator characteristic (ROC) curves. In both cases, Laplacian regularization turned out to be crucial for achieving a reasonable generalization performance. The Laplacian-regularized discriminative model outperformed the generative model of Reiss and Schwikowski in terms of the area under the ROC curve on both datasets. The performance was further improved with a hybrid approach, in which our model was initialized with the motifs obtained with the method of Reiss and Schwikowski.

Availability: Software and supplementary material is available from http://lehrach.com/wolfgang/dmf

Contact: wlehrach@ed.ac.uk

1 INTRODUCTION

Peptide recognition modules (PRMs) are specialized compact protein domains that mediate many important protein–protein interactions. They are responsible for the assembly of critical macromolecular complexes and biochemical pathways (Pawson and Scott, 1997), and they have been implicated in carciogenesis and various other human diseases (Sudol and Hunter, 2000). PRMs recognize and bind to peptide ligands that contain a specific structural motif.

One of the most actively studied PRMs is the SH3 domain, which binds to peptide ligands that contain a particular proline-rich core. Tong et al. (2002) carried out two extensive experimental studies to infer the network of SH3-mediated protein–protein interactions in Saccharomyces cerevisiae. They identified 28 SH3 domain proteins in the S.cerevisiae proteome, which were used as baits and screened against conventional and Proline-rich libraries in a yeast two-hybrid experiment (Twyman, 2004). In a second independent study, they screened random peptide libraries by phage display (Twyman, 2004) to identify the consensus sequence for preferred ligands that bind to each PRM. Based on these consensus sequences, they inferred a protein–protein interaction network that links each PRM to proteins containing the preferred ligand. Since both experimental procedures are intrinsically noisy, the two independently inferred interaction networks were found to show only a modest degree of overlap.

Reiss and Schwikowski (2004) addressed the question of whether computational in silico approaches would allow some of the difficult and expensive experimental procedures to be more specifically targeted, or even bypassed altogether. To this end, they developed a probabilistic generative model of the SH3 ligand peptides, based on the widely used Gibbs sampling motif finding algorithm (Lawrence et al., 1993; Liu et al., 1995). Directly applying the standard Gibbs motif sampler to the S.cerevisiae SH3 interaction data faces the difficulty that each SH3 domain is only involved in a small number of interactions (between 1 and 20), which leads to a poor motif conservation and a high susceptibility to random artefacts owing to the small sample size. Conversely, searching for a single motif in all identified SH3 domains lacks the specificity to identify anything but a broad consensus pattern. Reiss and Schwikowski (2004) therefore devised a compromise strategy, where the network information was used as a prior on the structure of individual motifs, which were searched for with a modified version of the Gibbs motif sampler. The prior was adjusted to become discriminative, giving higher probability to those motifs that are distinct from non-binding motifs.

Reiss and Schwikowski (2004) encouragingly demonstrate that a probabilistic model trained on protein sequences and observed physical interactions can succeed in independently predicting new protein–protein interactions mediated by SH3 domains. However, a shortcoming of their model is a dependence on tuning parameters that have to be chosen in advance by the user and that are not inferred from the data. Inappropriate values reduce the performance of their algorithm to using standard motif searching algorithms, and

\*To whom correspondence should be addressed.
it is unlikely that universal values applicable to different protein (super-) families exist. Also, the proposed model borrows substantial strength from its heuristic discriminative modification of the prior, which again depends on various tuning parameters.

This paper proposes an alternative in silico method for the prediction of SH3-mediated protein–peptide interactions, which addresses some of the shortcomings of the model introduced by Reiss and Schwikowski (2004). A key feature of our model is that it is discriminative: given a set of protein sequences, the model only attempts to find domains that distinguish between different SH3 binding domains. This is in contrast to the approach of Reiss and Schwikowski (2004), which is based on a generative model of the whole sequence. As discussed in Segal and Sharan (2004), a generative approach can be confounded by repetitive or over-represented motifs that are not related to PRM–peptide interactions, which our discriminative model avoids by formulating the learning problem in terms of a supervised classification problem.

The model we propose is based on a DNA-sequence model applied by Segal et al. (2002) and Segal and Sharan (2004). However, due to the larger size of the alphabet (20 amino acids instead of 4 nt) and the small number of interactions per SH3 domain, their maximum likelihood approach to parameter estimation is bound to lead to serious over-fitting. An essential component of our approach, therefore, is the inclusion of a regularization scheme, resulting in a maximum a posteriori (MAP) or penalized maximum likelihood estimate of the parameters.

2 METHODS

In this section, we first define the problem, followed by an overview of the model of Reiss and Schwikowski (2004). We then derive our discriminative model and describe how to apply it. Let \( D = \{d_i\} \) denote a set of SH3 domains, and \( S = \{s_j\} \) a set of protein sequences. We introduce a binary variable \( e_{ij} \in \{0,1\} \), where \( e_{ij} = 1 \) indicates that sequence \( s_j \) binds to SH3 domain \( d_i \), while \( e_{ij} = 0 \) indicates the absence of an interaction.

We assume that we are given a protein interaction network \( E = \{e_{ij}, d_i \in D, s_j \in S\} \) from a yeast two-hybrid or plasmid expression experiment. The objective is to derive a model that predicts this network from the sequences alone.

2.1 The generative model of Reiss and Schwikowski

Reiss and Schwikowski (2004) model \( P(s|e_{ij} = 1) \), the probability of the sequence \( s_j \) given that it binds to the PRM \( d_i \). The PRM for a domain \( d_i \) is modelled as a position-specific scoring matrix (PSSM) \( \Theta = \{\theta_{ij}\} \), where \( \theta_{ij} \in [0,1] \) is the probability of observing amino acid \( i \) in the \( k \)-th position of the \( i \)-th PSSM (i.e. the PSSM that indicates binding to the PRM \( d_i \)). \( \Theta = \{\theta_{ij}\} \) is the set of all PSSMs. Each position in the PRM is modelled as an independent discrete distribution—in other words, for all \( d_i \), for all positions \( k \), \( \sum_{j=1}^{20} \theta_{i,j,k} = 1 \). They also model the background distribution as a zeroth order Markov model \( \theta_{0,i} \), where again \( \sum_{i=1}^{20} \theta_{0,i} = 1 \).

Given there is an interaction between domain \( d_i \) and sequence \( s_j \), they introduce a hidden variable \( a_{ij} \), where \( a_{ij} = 1 \) indicates the position of the first binding site of the binding motif in sequence \( s_j \). Note that \( a_{ij} \) ranges from 0 to \( n_j - p \), where \( n_j \) is the length of the \( d_i \)-sequence \( s_j \) and \( p \) is the length of the binding motif. \( A = \{a_{ij}\} \) is the set of all hidden location variables. The residues involved in the binding are then modelled as:

\[
P(s_j | a_{ij} = 1, s_j, a_{ij} + 1, \ldots, s_j, a_{ij} + p \mid \Theta, e_{ij} = 1) = \prod_{k=1}^{p} \theta_{i,k, a_{ij} + k}.
\]

The likelihood of sequence \( s_j \) with binding events \( E_j \) to domains \( D = \{d_i\} \) (with PSSMs \( \Theta_{ij} \)) at the corresponding binding sites \( A_{ij} \) may then be written as (see Reiss and Schwikowski, 2004):

\[
P(s_j, E_j, A_{ij} \mid \Theta_{ij}, \theta_0) \\
\propto \prod_{q=1}^{n_j} \theta_{0,i_{j+1}} \prod_{j=1}^{p} \prod_{k=1}^{p} \theta_{i,k, a_{ij} + k} e_{ij}.
\]

The Gibbs motif sampler works by sampling the location parameters \( a_{ij} \) and the PSSM parameters \( \Theta_{ij} \) from the posterior distribution with Gibbs sampling, iterating between sampling \( \Theta_{ij} \) given \( a_{ij} \) and then \( a_{ij} \) given \( \Theta_{ij} \). The posterior distributions depend on the data via sufficient statistics that are summarized in the matrices \( C_i \) whose elements are defined as \( C_{i,j,k,l} = \delta(s_j, a_{ij} + l) \), where \( \delta \) is the indicator function. In words \( C_{i,j,k} \) is 1 if the \( k \)-th position of the binding motif in sequence \( s_j \) that binds to PRM domain \( d_i \) is amino acid \( l \). Otherwise, it is zero. As opposed to the standard Gibbs sampler, Reiss and Schwikowski (2004) made use of the protein–protein interaction information \( E = \{e_{ij}\} \) in computing the modified sufficient statistics \( \tilde{C}_{i,j} \), which they define as follows:

\[
\tilde{C}_{i,j} = \sum_{a_{ij}} e_{a_{ij}, C_{a_{ij}}} + p_0 \sum_{a_{ij}} e_{a_{ij}, \bar{C}_{a_{ij}}}.
\]

The third term encourages similarity of the binding motifs in sequences that bind to the same PRM domain. The second term encourages different binding motifs of all SH3 domains to be similar. The first term increases the similarity between the binding motifs for SH3 domains linked by ’promiscuous’ sequences, i.e. those sequences binding to more than one SH3 domain. While this approach is intuitively appealing, the scheme depends on two tunable parameters \( p_0 \) and \( p_1 \), which have to be set by the user in advance. Furthermore, the counting matrices are adjusted in a discriminatory way: when two sites equally match a given PSSM \( \Theta_{ij} \), then the one that is most dissimilar to a third, highly conserved but non-binding PSSM \( \Theta_{ij} \), should preferably be chosen. The implementation of this scheme depends on some further user-tunable parameters, with the same resulting difficulties.

2.2 A discriminative model

In our discriminative approach, we do not directly model \( \theta_0 \), the background distribution and \( \Theta_{ij} \), the motif distribution. Instead, we directly model the probability of the occurrence of a binding motif for the \( i \)-th PRM in sequence \( s_j \): \( P(e_{ij} | s_j) \). We start with a very similar model to Reiss and Schwikowski (2004). The probability of a sequence given a binding site position \( m \) is shown in Equation (4). Equation (5) shows the probability of a sequence without a motif.

\[
P(s_j | e_{ij} = 1, a_{ij} = m) = \prod_{q=1}^{n_j} \theta_{0,i_{j+1}} \prod_{k=1}^{p} \theta_{i,k, a_{ij} + k} e_{ij}.
\]

\[
P(s_j | e_{ij} = 0) = \prod_{q=1}^{n_j} \theta_{0,i_{j+1}}
\]

We now sum over \( a_{ij} \) the possible binding positions for the PRM with a uniform position prior \( P(a_{ij} = m) = 1/(n_j - p + 1) \).

\[
P(s_j | e_{ij} = 1) = \prod_{q=1}^{n_j} 1/(n_j - p + 1) \sum_{m=0}^{n_j - p} \prod_{k=1}^{p} \theta_{i,k, a_{ij} + k} e_{ij}.
\]

Applying Bayes’ rule,

\[
P(e_{ij} = 1 | s_j) = \frac{P(s_j | e_{ij} = 1)P(e_{ij} = 1)}{P(s_j)}.
\]
where \( P(s_j) = \sum_{i=0}^{T} P(s_j | e_{ij})P(e_{ij}) \) and combining Equations (5) and (6), we get

\[
P(e_{ij} = 1 | s_j) = \log \left( \log \left( \frac{1}{N_p - 1 + \sum_{m=0}^{n-1} \exp \left\{ \sum_{l=1}^{L} W_{i,k,l,s_{m-1}} \right\} } + T_i \right) \right),
\]

where we have defined \( W_{i,j,k,l} = \log(\theta^{(l)} / \theta^{(k)}), T_i = \log(P(e_{ij} = 1)P(e_{ij} = 0)) \) and \( \log(x) = (1 + e^{-x})^{-1} \). We can now apply this discriminative model, which corresponds to Equation (2) in Segal et al. (2002), and the equation in Section 2.1 of Segal and Sharan (2004), to infer the presence of an interaction between a peptide sequence and an SH3 domain.

### 2.3 Parameter estimation
Having specified the model, we next need to estimate the parameters, which are the set of weights \( W = \{ W_{i,k,l} \} \) and thresholds \( T = \{ T_i \} \). A standard way to optimize these parameters, adopted for instance in Segal et al. (2002) and Segal and Sharan (2004), is to follow a maximum likelihood approach. Given the training data \( D \), which is the set of all training sequences \( s_j \) and binding interaction indicator variables \( e_{ij} \), we want to maximize the log likelihood \(-E_D\) :

\[
-E_D = \log P(D | W, T) = \sum_{i,j} e_{ij} \log P(e_{ij} = 1 | s_j, W, T) + (1 - e_{ij}) \log (1 - P(e_{ij} = 1 | s_j, W, T))
\]

Note that \( P(e_{ij} = 1 | s_j, W, T) \) has been defined in Equation (8). It is straightforward to derive the partial derivatives \( \partial E_D / \partial (W_{i,k,l}) \) and \( \partial E_D / \partial (T_i) \), which allow us to apply an iterative gradient-based optimization scheme.

### 2.4 Regularization
A shortcoming of the maximum likelihood approach discussed in the previous section is its susceptibility to over-fitting as for each SH3 domain, we have \( p + 1 \) parameters to estimate. This exceeds the number of peptide sequences an SH3 domain binds to and hence calls for the implementation of an effective regularization scheme. A standard approach widely applied in machine learning is to impose a prior probability on the weights \( W \) such that large weight values are discouraged and a priori a value of zero is assumed. Define \( E_W \) to denote a function of \( W \) that is monotonically increasing with the magnitude of the weights \( W \). We define the prior

\[
P(W | \alpha) = \frac{1}{Z} \prod \exp (-\alpha E_W(W)),
\]

where \( Z \) is a normalization constant, and \( \alpha \) represents a scale factor. This choice of prior is particularly meaningful in our application. From the definition of the weights in the text below Equation (8) it is seen that \( W_{i,k,l}=0 \) corresponds to the assumption that the amino acid distribution at the \( k \)-th position of the \( i \)-th motif, \( \theta_{i,k} \), is equal to the background distribution \( \theta_0 \). Consequently, the \( k \)-th amino acid occurring in the \( k \)-th motif position provides no information about whether the amino acid is part of the background or part of the motif, which considering the larger number of parameters compared with sequences will be a common occurrence.

A priori commonly used in machine learning is the Gaussian distribution for \( P(W | \alpha) \) [see, for instance, MacKay (1992)], where

\[
E_W(W) = \frac{1}{2} \sum_{i,k,l} W_{i,k,l}^2.
\]

Less widely applied, but for our application particularly appropriate, is the Laplacian prior (Williams, 1995):

\[
E_W(W) = \sum_{i,k,l} | W_{i,k,l} |.
\]

The difference between these priors will be explained shortly. Note that we have left the thresholds \( T \) unregularized, as suggested in Williams (1995). This corresponds to a uniform prior

\[
P(T) = \text{Constant}.
\]

From our definition of \( T \), from below Equation (8), this is seen to correspond to a lack of prior knowledge about the global connectivities of the different SH3 domains. If this knowledge is available, it is straightforward to replace \( P(T) \) by a more informative prior. Now, the ideal approach would be to follow a fully Bayesian approach and sample the parameters \( [W, T] \) and the so-called hyperparameter \( \alpha \) from the posterior distribution \( P(W, T | \alpha) \).

Since this distribution is not available in closed form and cannot directly be sampled from, we have to resort to a numerical approximation with Markov chain Monte Carlo. Neal (1996) has applied this scheme in the context of neural networks, where he observed a significant improvement in the generalization performance over maximum likelihood. However, the computational costs are quite excessive, and we here follow a computationally less demanding approach. Rather than sample from the posterior distribution, we find the parameters that maximize this distribution, the so-called MAP estimate

\[
[W, T]_{\text{MAP}} = \arg \max_{W,T} P(W, T | D, \alpha).
\]

Now

\[
P(W, T | D, \alpha) \propto P(D | W, T)P(W | \alpha)P(T).
\]

From Equations (9), (10) and (13) we see that the optimization of the thresholds \( T \) remains unaffected by the proposed regularization scheme.

For the weights \( W \), however, we now have to minimize the modified cost function

\[
E^*(W) = E_D(W) + \alpha E_W(W),
\]

where we have two alternative functions for \( E_W \); see Equations (11) and (12). These two priors can be justified in a maximum entropy sense under different inference assumptions, as discussed in Williams (1995). The practical difference between the two priors can be understood from the derivatives of the regularization term \( E_W \). In the Gaussian case, this derivative is proportional to the size of the weights \( | W_{i,k,l} | \). This implies that large weights are more heavily penalized than small weights, and the model tends to end up with a large number of small weights. For the Laplacian prior, the derivative is constant: \( \partial E_W / \partial (W_{i,k,l}) \propto 1 \). This imposes less of a penalty on large weights, while driving small weights more strongly down to zero. In fact, the discontinuity of the derivative at the origin \( W_{i,k,l}=0 \) can be used for a pruning scheme, as discussed in Williams (1995).

The proposed regularization scheme seems to depend on the hyperparameter \( \alpha \). In fact, this hyperparameter can be integrated out

\[
P(W) = \int P(W | \alpha)P(\alpha)\,d\alpha.
\]

Since \( \alpha \) is a scale parameter, it is reasonable to use the improper \((1/\alpha)\) ignorance prior (Williams, 1995). It is then straightforward to show that \( E_W(W) \) as in Equations (10) and (12),

\[
-\log P(W) = |W| \log E_W,
\]

where \( |W| \) is the number of weight parameters. Replacing \( P(W | \alpha) \) by \( P(W) \) in Equation (15):

\[
P(W, T | D) \propto P(D | W, T)P(W)P(T),
\]

and noting that the threshold parameters \( T \) are unregularized (corresponding to a uniform prior), this leads to the following modification of the objective function \( E^* \) [compare with Equation (16)]:

\[
E^*(W) = E_D(W) + |W| \log E_W.
\]
Now, taking derivatives we get
\[
\frac{\partial E^*}{\partial W_{i,k,l}} = \frac{\partial E_D}{\partial W_{i,k,l}} + \frac{\partial E_W}{\partial W_{i,k,l}},
\]
where the effective hyperparameter \( \alpha = (|W|/E_D) \) is determined adaptively during training. Hence, as opposed to Reiss and Schwikowski (2004), we have no arbitrary parameters that would need to be hand-tuned by the user. As an aside, we notice that the integration over the hyperparameters has been criticized by MacKay (1999) on the grounds that in conjunction with the MAP approximation it may lead to over-regularization. An alternative proposed by MacKay (1992) is a computationally more expensive maximum likelihood type II optimization of \( \alpha \). Interestingly, these approaches lead to identical results when using the Laplacian prior (Williams, 1995), hence rendering the MAP approach more valid than in the Gaussian case.

The regularization method proposed in this section can easily be generalized to allow for more than one hyperparameter \( \alpha \). In fact, we divided the weights \( W_{i,k,l} \) into separate weight groups, one for each SH3 domain protein, where each weight group was associated with a separate hyperparameter. Such weight groups have been found in previous studies to improve the generalization performance of neural networks (MacKay, 1992). To reduce the opacity of the notation, we have not made this modification explicit in the text.

2.5 The algorithm
We adapted the parameters with conjugate gradients, using the MATLAB implementation in the NETLAB library (Nabney, 2002). We rescaled the objective function of Equation (9) by assuming that there is a small \( \zeta = 10^{-8} \) chance of measurements being incorrect, the effect of this being to constrain the objective function to remain finite within floating point accuracy, leading to a significantly faster rate of convergence. This corresponds to the model of uncertainty in measurements discussed in Deng et al. (2002). The weights \( W_{i,k,l} \) were regularized according to Equation (21) and we updated the effective hyperparameter \( \alpha \) every 10 iterations, as described below Equation (21). After each update, the search direction of the conjugate gradients method was reset.

3 SIMULATIONS
We evaluated the performance of the proposed discriminative model with the different regularization schemes on the phage display and yeast two-hybrid protein interaction data of Tong et al. (2002). We removed SH3 domain proteins that only bind to a single peptide, as there would be no way to validate these interactions on an independent test set. With this modification, the phage display dataset contains 17 SH3 domains, 207 binding partners and 381 interactions, while the yeast two-hybrid dataset (displayed in Fig. 1) has 28 SH3 domains, 143 binding partners and 285 interactions. Further details can be found in the Supplementary material of Tong et al. (2002).

We evaluated the generalization performance with a 10-fold cross-validation scheme where the data were randomly partitioned into 10-folds. The generalization performance was then evaluated on the current fold, and the other 9-folds were used for training. We obtained an average out-of-sample performance by repeating this for all 10-folds.

For comparison with Reiss and Schwikowski (2004), we measured the performance in terms of ROC curves, which are obtained by subjecting the predicted posterior probabilities \( P(e_{ij} | x) \) to various threshold parameters \( \theta \in [0, 1] \). By numerically integrating over the whole parameter range \( \theta \in [0, 1] \) we obtain the area under the ROC curve. This so-called AUROC score ranges from 0.5 for a random predictor to 1.0 for a perfect predictor, with larger values generally indicating a better performance; see Section 4.3 for a more specific discussion. Since the left part of the ROC curves, where the number of false positives is low, is often of particular interest, we also restrict the integral to false positive values of less than 0.1. We refer to the resulting score as AUROC1.

To evaluate the performance of the generative model of Reiss and Schwikowski (2004), we used the software provided by the authors, which is available from http://sf.net/projects/netmotsa. Recall that the generative model depends on various tuning parameters, which are not inferred from the data but rather have to be set by the user in advance. For our comparative study, we used the default values defined in the software of Reiss and Schwikowski (2004). These parameters had been optimized by the authors on the same dataset as used in our study; hence they should reflect a quasi-ideal performance.

4 RESULTS
In earlier simulations, we have found that the Laplacian regularization scheme achieved a significant improvement on both the unregularized as well as the Gaussian regularized models. Due to space restrictions imposed by the journal, we have relegated a detailed discussion of these findings to the supplementary online material. In the simulations reported in the following sections, we have compared three approaches: (1) the generative model of Reiss and Schwikowski (2004); (2) the proposed discriminative model, where the weights were initialized from the PSSMs learned with the generative model and (3) an ensemble of discriminative models; this ensemble was created by training 10 models from different initializations, and keeping the five models with the highest training set scores. In what follows, we will refer to these methods as (1) GEN, (2) DIS-I and (3) DIS-E, respectively; see Table 1 for a summary. For training the discriminative models, the Laplacian regularization scheme was applied throughout.

4.1 Assessing the prediction performance
The top left panel of Figure 2 shows the ROC curves obtained for the yeast two-hybrid network. Both discriminative methods, DIS-I and DIS-E, clearly outperform GEN in the right part of the graph, for false positive rates (FPR) greater than 0.3. This is reflected in higher overall AUROC scores, as seen from Table 2. In the left part, for FPR < 0.3, the three methods perform more or less equally well. Plotting the ROC curves for values of FPR < 0.1 at a higher resolution, as done in the bottom left panel of Figure 2, reveals that DIS-I and GEN achieve the same performance (AUROC01 = 0.17), which is slightly better than that of DIS-E.

The right panel of Figure 2 shows the ROC curves obtained for the phage display network. The discriminative methods outperform the generative model, both in terms of overall (top panel) and left side (bottom panel) performance. This improvement is considerably improved when starting the training simulations from an informative initialization (DIS-I). Also, we found that the performance of all methods is consistently better for the phage display network than for the yeast two-hybrid network (see Section 5).
4.2 Locating binding regions

To test whether the proposed model is actually able to locate the binding sites, we focused on Las17, which can form protein complexes containing multiple SH3 domains. Tong et al. (2002) have applied an enzyme-linked immunosorbent assay (ELISA) to identify the region of the target protein that binds the SH3 domain. They focused on five proline-rich peptide fragments of Las17, whose locations are indicated in the caption of Figure 3. In our study, we removed Las17 from the training set, and repeated the training simulations for both the yeast two-hybrid as well as the phage display data. We then tested whether the binding locations of SH3 domain proteins interacting with Las17 could be correctly predicted. We evaluated the models in three different ways. In the first evaluation, we took the interactions of the ELISA experiment, reported in Tong et al. (2002), as true interactions. We applied the models to these segments separately, ranked the SH3 domains for each segment according to their segment-specific binding scores, and obtained the ROC curves from these rankings. The results are shown in Figure 4. In the second evaluation, we selected the threshold that resulted in a total of
14 predicted SH3 domains. This number is equal to the total number of interactions detected with the ELISA experiments when omitting Yfr024 (Yfr024 binds only to a single sequence and was therefore omitted from our study for the reason discussed above). We then compared the predictions between the different models and with the ELISA experiment. The results are shown in Figure 3. In the third evaluation, we tested whether the predictions obtained with the different models were significantly better than obtained by chance. For each proline-rich segment, we ranked the SH3 domains according to the binding score predicted by the model. We divided the SH3 domains into two classes: those detected as binding with the ELISA experiment, and those not detected as binding. We then applied the Wilcoxon rank sum (or Mann–Whitney) test to obtain the P-value under the null hypothesis that the in silico predicted score is independent of the ELISA experiment.

The top panel of Figure 3 shows the thresholded predictions obtained for the yeast two-hybrid data. Both GEN and DIS-I predict 5, while DIS-E predicts 4 true positive interactions. The slightly worse performance of DIS-E is in accordance with the ROC curves in the left panel of Figure 4, where DIS-E shows a poorer performance in the very left part of the graphs. Lowering the
The threshold turns out to be beneficial only for the discriminative models: DIS-I predicts 1, and DIS-E predicts 4 additional true interactions. Again, this finding is consistent with the ROC curves in Figure 4, where both discriminative models outperform GEN.

The three models show a modest degree of complementarity. GEN fails to predict any SH3 domain protein binding to the second segment of Las17, and this failure persists even as the threshold is lowered. To the contrary, both discriminative models predict at
least one SH3 protein-binding to this segment, and this number increases as the threshold is lowered.

The bottom panel of Figure 3 shows the predictions obtained for the phage display data. Among the 14 highest scoring interactions, GEN predicts 6 true positives, while DIS-I and DIS-E predict only 5 and 4, respectively. However, among the next 14 highest scoring interactions, GEN only gains 2 extra true positives. DIS-E gains 5 extra true positives, and thus performs slightly better than GEN. Both methods are noticeably outperformed by DIS-I, which predicts all but two interactions. This improved performance is consistent with the ROC curves, shown in the right panel of Figure 4. While GEN obtains higher true positive rates in the leftmost region of the graph, both discriminative models experience a considerable performance boost at a false positive rate of 0.18, and DIS-E shows the best performance overall.

The results discussed in this section are summarized in Table 3. For the yeast two-hybrid data, both discriminative models slightly outperform the generative model: DIS-E (AUROC = 0.68) > DIS-I (AUROC = 0.65) > GEN (AUROC = 0.60). For the phage display data, the discriminative model with the informative initialization outperforms the generative model: DIS-I (AUROC = 0.84) > GEN (AUROC = 0.77) > DIS-E (AUROC = 0.71). The performance on the phage display data is, overall, better than that on the yeast two-hybrid data, with all $p$-values significant. For the yeast two-hybrid data, only the $p$-values obtained for the discriminative methods would be regarded as significant.

### 4.3 Biological validation and application

An important practical application of the proposed method would be the cleaning and filtering of high-throughput interaction data. Our conjecture is that protein interactions that are assigned a higher posterior probability score in silico are more reliable than those with a lower score. We would therefore assume that interactions found with both the yeast two-hybrid as well as the phage display experiment have, on average, higher posterior probability scores than those found with only one experiment. Phrased differently, we would assume that the intersection of the sets of interactions found with yeast two-hybrid and phage display shows an enrichment for higher scoring in silico interactions. To test this conjecture, we extracted for both experiments the 400 highest scoring interactions; this is the number of interactions detected experimentally with phage display. When training our model on the yeast two-hybrid data, we recovered 25% of the interactions in the intersection set, but only 9% of the interactions in the non-intersection set. Hence, in both training simulations, we found that the subset of more reliable interactions (i.e. those interactions found with both experimental methods) was noticeably enriched for high-scoring in silico predictions. This finding corroborates our hypothesis that the predicted interaction scores are biologically consistent and suggests that our method could offer a useful tool for filtering noisy high-throughput protein interaction data.

### 5 DISCUSSION

The model we propose is based on the assumption that protein interactions are mediated by short peptide segments that bind to PRM domains. This assumption is valid for the phage display data, which explains why all the models achieve a better performance here than on the yeast two-hybrid interaction network.

Our simulations suggest that the randomly initialized discriminative model achieves a performance at least as good as the generative model of Reiss and Schwikowski (2004). While the AUROC01 score for the yeast two-hybrid network is slightly worse, the overall AUROC scores have noticeably improved. The discriminative model also shows some complementarity to the generative model with respect to locating the binding regions.

When initializing the discriminative model with the PSSMs predicted by the generative model, its performance further improves. With this informative initialization, the discriminative model outperforms the generative model of Reiss and Schwikowski (2004) both in terms of predicting protein interactions as well as locating binding regions. The improvement is particularly noticeable for the phage display network, where the data are more in line with the model assumptions.

Note that the model we have proposed has only been trained on proteins in the SH3 interaction networks. Hence, the objective of our approach is to predict the probability of a particular protein interaction, given that the protein is in the interaction network. In principle, it is straightforward to generalize this approach to not only distinguish between the different protein interactions, but also to predict whether the protein is in the interaction network. All that is required is to include an extra output node representing non-binding background sequences. However, the inclusion of background sequences, which substantially outnumber the binding sequences, would substantially increase the computational costs of the training scheme, and has therefore not been attempted.

We therefore suggest a hybrid approach in which, at the first stage, the generative model of Reiss and Schwikowski (2004) is applied to predict if a protein sequence is binding, and our discriminative model is applied at the second stage to predict which protein the sequence binds to. Our simulations suggest that this combined scheme should outperform the generative model of Reiss and Schwikowski (2004) owing to the better performance of the proposed discriminative model at the second stage.

This improved performance is presumably the consequence of two important modifications. First, the hyperparameters in our approach, which are in some way akin to the tuning parameters in the model of Reiss and Schwikowski (2004), have been integrated out. As a consequence of this integration, our scheme depends on some effective hyperparameters that are automatically updated during training (see Subsection 2.4). This renders our approach independent of any tweaking parameters that would otherwise
have to be hand-tuned by the user. The second improvement is related to the discriminative nature of our model. Note that Reiss and Schwikowski (2004) also tried to include a discriminative feature into their generative model by penalizing the detection of over-represented but non-discriminative motifs. However, this approach is rather heuristic and introduces another user-defined tuning parameter. The model applied in our study is a proper discriminative model per se, which has been consistently derived within the probabilistic context and dispenses with the need for hand-tuning another parameter.

ACKNOWLEDGEMENTS

Wolfgang Lehrach is supported by an EPSRC postgraduate student grant and Dirk Husmeier is supported by the Scottish Executive Environmental and Rural Affairs Department (SEERAD). Computing time was generously provided by Tony Travis on the Scottish Agricultural Research Institute (SARI) Linux Beowulf cluster.

Conflict of Interest: none declared.

REFERENCES


