Bioinformatics Advance Access published February 9, 2016

# Predicting physiologically relevant SH3 domain mediated protein-protein interactions in yeast

Shobhit Jain [1,2] and Gary D. Bader [1,2*]

[1]Department of Computer Science, University of Toronto, 10 Kings College Road, Toronto, Canada M5S 3G4

[2]The Donnelly Centre, University of Toronto, 160 College Street, Toronto, Canada M5S 3E1

Associate Editor: Prof. Alfonso Valencia

**ABSTRACT**

**Motivation:** Many intracellular signaling processes are mediated by interactions involving peptide recognition modules such as SH3 domains. These domains bind to small, linear protein sequence motifs which can be identified using high-throughput experimental screens such as phage display. Binding motif patterns can then be used to computationally predict protein interactions mediated by these domains. While many protein-protein interaction prediction methods exist, most do not work with peptide recognition module mediated interactions or do not consider many of the known constraints governing physiologically relevant interactions between two proteins.

**Results:** A novel method for predicting physiologically relevant SH3 domain-peptide mediated protein-protein interactions in *S. cerevisae* using phage display data is presented. Like some previous similar methods, this method uses position weight matrix models of protein linear motif preference for individual SH3 domains to scan the proteome for potential hits and then filters these hits using a range of evidence sources related to sequence-based and cellular constraints on protein interactions. The novelty of this approach is the large number of evidence sources used and the method of combination of sequence based and protein pair based evidence sources. By combining different peptide and protein features using multiple Bayesian models we are able to predict high confidence interactions with an overall accuracy of 0.97.

**Availability:** **Do**main-**Mo**tif Mediated Interaction **Pred**iction (DoMo-Pred) command line tool and all relevant datasets are available under GNU LGPL license for download from `http://www.baderlab.org/Software/DoMo-Pred`.

**Contact:** gary.bader@utoronto.ca

## 1 INTRODUCTION

Protein-protein interactions (PPIs) are physical associations between protein pairs in a specific biological context. Their knowledge provide important insights into the functioning of a cell. Previously, experimental detection of PPIs was limited to labor intensive techniques such as co-immunoprecipitation or affinity chromatography (Skrabanek *et al.*, 2008). Though the detected PPIs are largely accurate, these techniques are difficult to apply to whole proteome analysis. This led to the development of various high-throughput PPI detection protocols such as mass-spectrometry combined with affinity-purification, yeast two-hybrid and next-generation sequencing to detect PPIs at whole genome level (Davy *et al.*, 2001; Ito *et al.*, 2001; McCraith *et al.*, 2000; Rain *et al.*, 2001; Uetz *et al.*, 2000; Yu *et al.*, 2011; Braun *et al.*, 2013). However, genome-scale methods are also highly resource intensive and single projects and techniques do not cover all known protein interactions. Further, they only cover interactions in one organism at a time. Computational approaches designed to predict reliable and novel PPIs based on experimental interaction data sets have the advantages that they are inexpensive to apply to genomes, including those that are infeasible to tackle experimentally and this motivates their further development (Skrabanek *et al.*, 2008).

Multiple kinds of protein-protein interactions exist. We focus on interactions involving peptide recognition modules (PRMs), in particular Src homology 3 (SH3), which are important in many cellular signaling processes. These domains bind to small, linear sequence motifs (peptides) within proteins (Pawson and Nash, 2003). SH3 domains are approximately 60 amino acids long with five beta strands organized into two perpendicular beta sheets interrupted by a 3-10 helix (Pawson and Gish, 1992). They often bind to proline-rich regions and multiple classes have been recognized based on their binding motifs. Class I SH3 domains bind to [R/K]xxPxxP and class II bind to PxxPx[R/K] motifs (Mayer, 2001). They can also bind to proline-free regions containing arginine or lysine (Tong *et al.*, 2002). SH3 domains are involved in many regulatory or signaling processes, including endocytosis (Tonikian *et al.*, 2009), actin cytoskeleton regulation (Pawson and Schlessingert, 1993), and tyrosine kinase pathways (Schlessinger, 1994). Experimental methods such as phage display (Tonikian *et al.*, 2008, 2009; Tong *et al.*, 2002) and peptide microarray (MacBeath and Schreiber, 2000; Hu *et al.*, 2004; Stiffler *et al.*, 2007) have been used to identify the peptides binding to PRMs.

The computational problem under focus in this work is to use the SH3 domain binding peptides identified from phage display experiments to predict SH3 domain mediated

*to whom correspondence should be addressed

PPIs in *S. cerevisiae*. A straightforward approach is to construct position weight matrices (PWMs) from phage peptides and scan the whole proteome for potential binding sites in target proteins using some threshold score (Obenauer *et al.*, 2003). The problem with this simple approach is the lack of contextual information, for example, the predicted binding site might not be accessible or it might lie within a structured part of protein (e.g. domain). Tonikian *et al.* (2009) addressed this problem by combining in vitro (phage display, peptide array screening) and in vivo (yeast two-hybrid) data to predict SH3 domain mediated PPIs in yeast. Verifying interactions using multiple experimental techniques improves the PPI confidence but it is both time and resource consuming. Lam *et al.* (Lam *et al.*, 2010) combined comparative and structural genomic features with PWMs to reduce the number of false binding sites. But they did not consider that PPIs are influenced by many cellular constraints including that interacting proteins must be in close proximity and should be part of same process. Peptide-only features are not sufficient for predicting high confidence physiologically relevant PRM mediated PPIs with binding site resolution. Jansen *et al.* (2003), Rhodes *et al.* (2005), Li *et al.* (2008), Zhang *et al.* (2012), and others considered multiple types of cellular constraints and combined different evidence sources for PPI prediction, but their approaches are designed for full length proteins and cannot be used to predict PRM mediated PPIs, including identification of binding sites. More recently, Chen *et al.* (2015) combined limited number of peptide and protein features for predicting PRM mediated PPIs in humans. Their protein features are based on one of the earlier the works in the field ensemble PPI prediction (Jansen *et al.*, 2003). Since then many advances have been made in improving the performance of individual features in PPI prediction (Reimand *et al.*, 2012). Also, their method is not compatible with high-throughput binding peptide data, such as from phage display. Here, we make use of a larger set of evidence sources to predict SH3-mediated PPIs and their binding sites than has been collected previously and combine peptide level and protein level features in a single predictor.

## 2 APPROACH

PRM mediated PPIs do not occur in isolation in the cell. They are influenced by different sequence-based and cellular constraints. For example, SH3 domains can only bind surface accessible regions, interacting proteins must be present in same cellular compartment, and proteins in the same biological process with correlated gene expression profiles are more likely to interact compared to randomly selected protein pairs. Thus, diverse types of information can be used to help predict physiologically relevant protein interactions. In our method, PWMs constructed using peptides from phage display experiments are used to scan the yeast proteome for potential targets. Peptide features: disorder, surface accessibility, peptide conservation, and structural contact are combined using naïve Bayesian integration to score the PWM targets. Another naïve Bayesian model is used to combine protein features: cellular location, biological process, molecular function, gene expression, and sequence signature to score the same targets. Scores from both peptide and protein classifiers are then combined using Bayes theorem to predict physiologically relevant SH3 domain mediated PPIs in yeast. Figure 1 shows the work flow of our PRM mediated PPI prediction pipeline.

## 3 METHODS

### 3.1 Position weight matrix and proteome scanning

Position weight matrices (PWMs) are statistical models for representing sequence motifs. They are real valued $m \times n$ matrices, where $m$ are the amino acids and $n$ is the motif length. They are constructed using peptides from phage display experiments and then used to scan a protein sequences to find motif matches above a certain p-value threshold (Pizzi *et al.*, 2011; Wu *et al.*, 2000). Also, significant positions within the PWMs are identified and used in scoring peptide features: disordered region, surface accessibility, and peptide conservation (see supplementary material for details).

### 3.2 Peptide features

*3.2.1 Disordered region* PRMs bind to small peptide stretches containing a specific motif. Specifically interactions between proteins having SH3 domains and their targets are often mediated by proline rich peptide sequences containing PXXP, [R/K]xxPxxP, PxxPx[R/K] motifs. Proline disrupts the secondary structure of a protein by inhibiting the formation of helices and sheets (Morgan and Rubenstein, 2013). Also, small linear motifs tend to accumulate in disordered regions of protein (Linding *et al.*, 2003; Beltrao and Serrano, 2005; Davey *et al.*, 2010). Beltrao and Serrano showed that the binding sites of SH3 domains in *S. cerevisiae* often lie within the disordered regions of a protein (Beltrao and Serrano, 2005). DISOPRED, a neural network based tool, is used to estimate the probability of the protein region being disordered.

$$DR = \frac{\sum_i p_i = \begin{cases} 1 & \text{if amino acid } i \text{ is disordered} \\ 0 & \text{otherwise} \end{cases}}{N} \quad (1)$$

where $p_i$ is the disorder score of the $i^{th}$ significant amino acid (either 1 for disordered or 0 for ordered) and $N$ is the number of significant amino acids in the binding site.

*3.2.2 Surface accessibility* Sequences present on a protein's surface are more accessible to binding by SH3 domains than those that are buried inside a protein structure. The degree of solvent-accessible surface area of amino acid residues in a sequence indicates its level of exposure and is measured in terms of relative solvent accessibility (RSA) (Lam *et al.*, 2010; Adamczak *et al.*, 2004). We use SABLE (Adamczak *et al.*, 2004) to predict RSA values for target sequences. It uses a neural network based nonlinear regression model for continuous approximation of RSA values. Amino acid residues with RSA value $\geq 25\%$ are considered to be exposed and available for binding (Adamczak *et al.*, 2004).

$$SA = \frac{\sum_i p_i = \begin{cases} 1 & \text{if } RSA >= 25\% \\ 0 & \text{otherwise} \end{cases}}{N} \quad (2)$$

where $p_i$ is the surface accessibility score of $i^{th}$ significant amino

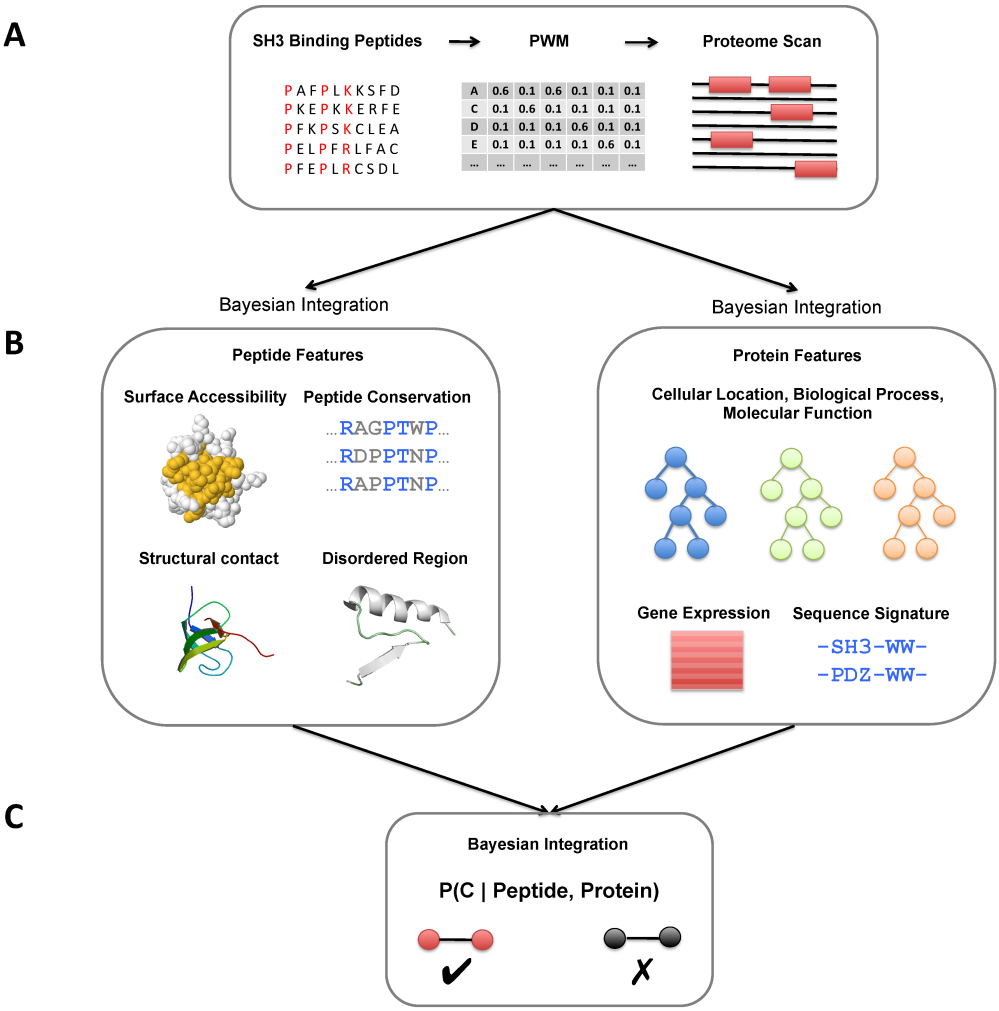



Fig. 1: Work flow of PRM mediated PPI prediction pipeline. (A) Proteome is scanned using a PWM built using experimentally derived binding peptides (e.g. from phage display) of a given SH3 domain for potential interactors. (B) Separate Bayesian classifiers for peptide and protein features. (C) Integration of classifiers for predicting interacting and non-interacting protein pairs.

acid and $N$ is the number of significant amino acids in the binding site.

*3.2.3 Peptide conservation* Biologically relevant peptides binding to yeast SH3 domains are more likely to be conserved in other yeast species (Beltrao and Serrano, 2005; Davey *et al.*, 2010). For measuring the conservation, orthologs of *S. cerevisiae* protein sequences in *C. glabrata, D. hansenii, K. lactis, Y. lipolytica, C. albicans, N. crassa*, and *S. pombe* (an optimal set as selected by (Beltrao and Serrano, 2005)) are identified using INPARANOID (Remm *et al.*, 2001). The orthologous sequences are then aligned with MAFFT (Katoh *et al.*, 2002) and the unweighted sum-of-pairs method from AL2CO (Pei and Grishin, 2001) is used to estimate the conservation score of each position in the multiple sequence alignment (Lam *et al.*, 2010).

$$PC = \frac{\sum_i p_i}{N} \quad (3)$$

where $p_i$ is the conservation score of the $i^{th}$ significant amino acid and $N$ is the number of significant amino acids in the binding site.

*3.2.4 Structural contact* Known 3-D structures of SH3 domains complexed with peptides can be used to assess the binding potential of a query SH3 domain and peptide by reducing residue-residue contacts in 3-D structures to a binary 2-D contact matrix (Chen *et al.*, 2008; Hui and Bader, 2010). Six yeast SH3-peptide co-complex PDB structures (1N5Z, 1SSH, 1ZUK, 2KYM, 2RQW, 2VKN) are used as base models. The Contact Map Analysis (CMA) tool from the SPACE software suite (Sobolev *et al.*, 2005) is used to reduce the 3-D structures to 2-D contact maps with residue level contact area for all base models. Query domain and peptide sequences are aligned with all base models using the Needleman-Wunsch algorithm and BLOSUM 62 substitution matrix to calculate the contact distance between aligned residues.

$$SC = \max_j \frac{\sum_i c_{ij}}{N} \quad (4)$$

where $c_{ij}$ is the normalized contact area of the $i^{th}$ aligned domain and peptide residues of the $j^{th}$ base model. Alignment gaps in contact residues will negatively impact the average contact area as only the aligned residues are used for scoring (a gap at a position associated with a large residue contact area will reduce the SC score more than a gap associated with a smaller residue contact area). $N$ is the number of aligned contact residues.

### 3.3 Protein features

*3.3.1 Cellular location, biological process, molecular function* Physical PPIs require proteins to be in close proximity to each other i.e. they should co-localize in the same cellular compartment. Also, interacting proteins are more likely to be part of same biological process or have the same function. The Gene Ontology (GO) contains a hierarchy of controlled terms describing cellular location, biological process, and molecular function of proteins (The Gene Ontology Consortium, 2000). The functional relationship between two proteins can be quantified using GO. Semantic similarity can be used to quantify relationships between different GO terms in an ontology. The higher the semantic similarity score between GO terms annotated to two proteins, more likely that they will interact with each other (Jain and Bader, 2010). Topological Clustering Semantic Similarity (TCSS) (Jain and Bader, 2010) is an accurate semantic similarity measure for PPI prediction. It normalizes the GO hierarchy before computing semantic similarity, according to cutoffs defined in the original TCSS paper.

$$CC = TCSS(a, b, ontology = C, cutoff = 2.4) \quad (5)$$

$$BP = TCSS(a, b, ontology = P, cutoff = 3.5) \quad (6)$$

$$MF = TCSS(a, b, ontology = F, cutoff = 3.3) \quad (7)$$

where $a$ and $b$ are the query proteins and $C, P, F$ are the cellular component, biological process, and molecular function ontologies.

*3.3.2 Gene expression* Gene expression as a measure for assessing the confidence and biological relevance of high-throughput PPIs is based on the notion that the cell is optimized to co-express genes if they function together and if they function together, they are more likely to physically interact than by chance (Bhardwaj and Lu, 2005; Grigoriev, 2001; Ge *et al.*, 2001; Jansen *et al.*, 2002). Most PPI prediction methods that make use of gene expression profile (GEP) correlation with PPIs to predict novel interactions (Li *et al.*, 2008; Rhodes *et al.*, 2005) rely on observations from a single expression dataset which can lead to many false positives and true negatives, as not all genes are expressed under a particular set of experimental conditions. Using multiple GEPs clearly improves the performance of a predictor as shown in Figure S1. Correlation coefficients from 86 gene expression profiles from GeneMANIA (Warde-Farley *et al.*, 2010) for a given pair of genes are combined using Fisher's z transformation (Faller, 1981; Jain and Bader, 2010)

$$EX = 1 - \frac{e^{2\bar{z}} + 1}{e^{2\bar{z}} - 1} \quad (8)$$

$$\bar{z} = N^{-1} \sum_{i=1}^{N} \frac{1}{2} \ln\left(\frac{1 + r_i}{1 - r_i}\right) \quad (9)$$

where $N$ is the number of profiles and $r_i$ is the Pearson correlation of the $i^{th}$ profile.

*3.3.3 Sequence signature* Sequence signature based PPI prediction methods are based on the notion that protein domains are correlated with specific functions. For instance, it has been shown that functionally related proteins have similar domain composition or they belong to the same "domain club" (Jin *et al.*, 2009). Information content of co-occurring InterPro (Apweiler *et al.*, 2001) signatures extracted from sequences of an experimentally verified set of 22,707 PPIs from DIP (Salwinski *et al.*, 2004) is used to score novel interactions, as described by Sprinzak and Margalit (Sprinzak and Margalit, 2001).

$$SS = \sum_{ij} -\log_2\left(\frac{p_{ij}}{p_i p_j}\right) \quad (10)$$

where $p_{ij}$ is the probability of seeing motif $i$ on one protein and motif $j$ on other protein in the experimentally verified PPI set, $p_i$ is the probability of seeing motif $i$ and $p_j$ is the probability of seeing motif $j$ in the same set.

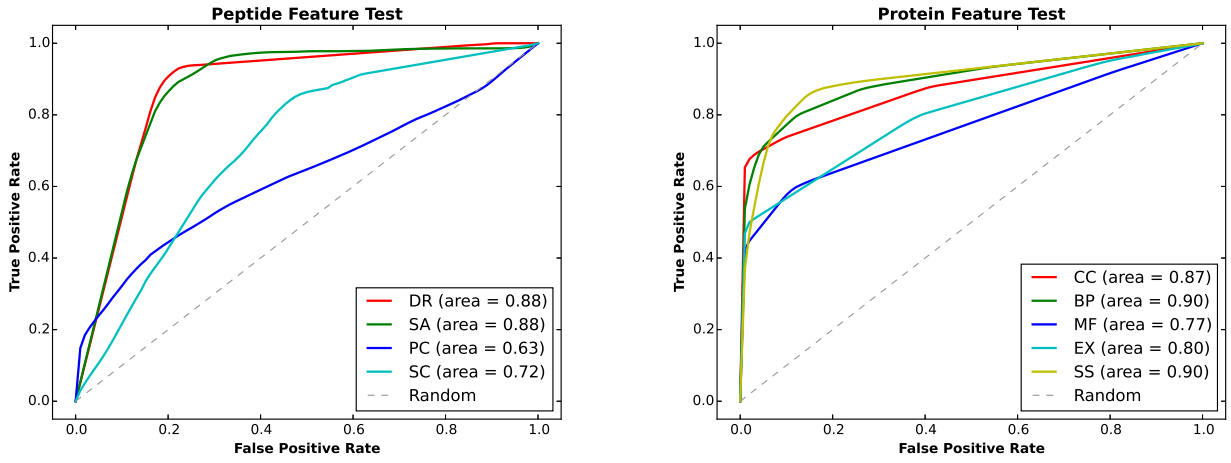
### 3.4 Bayesian integration

The objective of a Bayesian PPI prediction model is to estimate the probability that a given protein pair interacts, conditioned on the biological evidence in support of that interaction. A naïve Bayesian model simplifies this problem by assuming independence between different types of biological evidence. While modeling the PRM mediated PPI prediction problem a set of observations are made on domain-peptides while others are made on full-length proteins. Assuming that peptide and protein features are independent of each other, two separate naïve Bayes models $M_{pep}$ for peptide features and $M_{pro}$ for protein features are built to independently assess the class probability $Y$. The posterior probabilities $P(Y|M_{pep})$ and $P(Y|M_{pro})$ are combined using Bayes' theorem (Mitchell, 1997) (see supplementary material for further details).

## 4 RESULTS

### 4.1 Model training

The goal is to construct a generalized model which can predict high confidence, in vivo yeast SH3 domain - peptide physical interactions. To achieve this, both peptide and protein classifiers are trained on their respective positive and negative datasets. The peptide classifier is trained on a high confidence set of 628 SH3 domain-peptide interactions in yeast from the MINT database **(P1)** and an equal number of random selected negative interactions **(N1)**. The protein classifier is trained on a high confidence set of 5,215 pairwise yeast PPIs from the iRefIndex database **(P2)** and an equal number of randomly selected negative interactions **(N2)** (see supplementary material for details).




(a) (b)

Fig. 2: Prediction efficacy of individual (a) peptide features: disordered region (DR), surface accessibility (SA), peptide conservation (PC), structural contact (SC); and (b) protein features: cellular component (CC), biological process (BP), molecular function (MF), gene expression (EX), sequence signature (SS).

### 4.2 Feature selection

Figure 2 shows the discriminatory power of individual features for peptide and protein classifiers. Disordered region (DR) and surface accessibility (SA) perform much better in separating positives from negatives as compared to structural contact (SC) and peptide conservation (PC). Prediction efficacy of PC is least among the peptide features. This is due to the difficulty distinguishing positive and negative interactions because both of these sets have high conservation scores caused by the high similarity of protein sequences (and peptides they contain) in general across different yeast species (Figure S2). Biological process (BP), cellular component (CC), and sequence signature (SS) outperform molecular function (MF) and gene expression (EX) in the protein feature set. Proteins could have the same molecular function but still belong to different processes and this could be one of the reasons behind molecular function feature's weak performance. Gene expression data alone is not as powerful as others in discriminating positives from negatives (Kim *et al.*, 2014), which may be due to its moderate correlation with protein expression (i.e. gene expression may not imply that a functioning protein will be available for interaction) (Vogel and Marcotte, 2012).

Highly correlated features can negatively effect the performance of a naïve Bayesian classifier. Maximal information coefficient (MIC) is used to quantify the correlation between different features. DR and SA in the peptide feature set and CC and BP in the protein feature set are correlated with MICs of 0.72 and 0.5 respectively. The effect of correlation on classifier performance is measured by comparing different models without one of the correlated features. Further, to identify the feature subset which maximizes the performance of both classifiers, all possible combinations of features are compared using different statistical measures, such as area under ROC curve (AUROC), area under precision-recall curve (AUPRC), Brier score (BRIER), $F_1$-score, Matthews correlation coefficient (MCC) and accuracy (ACC). Peptide and protein classifiers with all features outperformed other models on at least one of statistical measure (see supplementary material for details).

### 4.3 Model evaluation

Blind validation protocols is used to assess the predictive power of peptide $M_{pep}$ and protein $M_{pro}$ naïve Bayesian classifiers. The majority of interactions in the P1 dataset are from two peptide array experiments (Tonikian *et al.*, 2009; Landgraf *et al.*, 2004). This could lead to an experimental bias therefore, for blind testing, the peptide classifier is trained using interactions only from peptide array experiments and tested using interactions from all other experiments (no overlap between training and test data sets). Similarly, to make an unbiased assessment, the protein classifier was trained using P2 dataset but tested using the 2,304 interactions (with no missing information) from the core subset of Database of Interacting Proteins (DIP) that do not overlap the P2 training set and are based on different filtering criteria compared to the MINT-inspired score used to select the iRefIndex P2 training set Salwinski *et al.* (2004). The DIP core database includes PPIs derived from both small-scale and large-scale experiments that have been scored by quality of experimental methods, occurrence of interaction between paralogs (PVM), probable domain-domain interactions between protein pairs (DPV), and comparison with expression profiles (EPR) (Salwinski

| Test | Classifier | MCC | ACC | $F_1$-score | AUROC |
|---|---|---|---|---|---|
| Filtered | Peptide | 0.74 | 0.87 | 0.87 | 0.92 |
| | Protein | 0.68 | 0.83 | 0.83 | 0.94 |
| Unfiltered | Peptide | 0.72 | 0.86 | 0.86 | 0.92 |
| | Protein | 0.63 | 0.80 | 0.80 | 0.92 |

**Table 1.** The filtered set has no missing values for any of the features, whereas unfiltered includes all feature data (as would be the case in a real world prediction scenario). Matthews correlation coefficient (MCC) threshold score $\geq 0.9$, accuracy (ACC), $F_1$-score and area under ROC curve (AUROC) of protein and peptide classifiers for blind and 10-fold cross-validation tests are shown. MCC, ACC, and $F_1$-score are reported at threshold score $\geq 0.9$.

*et al.*, 2004). In a real world prediction scenario, both classifiers are expected to encounter cases with missing information. Therefore, the performance of both classifiers is also tested using an unfiltered blind set. The results are summarized in Table 1. The AUROC for peptide clasifier is 0.92 and ACC lies within the range $[0.86, 0.87]$. The protein classifier has an AUROC within the range $[0.92, 0.94]$ and ACC is between $[0.80, 0.83]$.

The efficacy of the combined peptide and protein model was tested on the manually curated SH3 domain mediated PPI set from Tonikian *et al.* (2009). Tonikian and co-workers curated interactions supported by multiple experiments through an exhaustive literature search. Not all interactions (especially those identified using two hybrid and overlay assays) in this set are mapped to the peptide sequence within the interacting partner (Tonikian *et al.*, 2009). Therefore, these sequences are scanned using the three P1 training set PWMs to identify binding sites and significant amino acid positions within those sites. Peptide and protein classifiers are trained on P1 & N1 (no overlap with curated set) and P2 & N2 datasets, respectively. A randomized negative test set is created in the same way as N1. Results from different statistical measures are summarized in Figure 3. The combined classifier outperforms both the peptide and protein classifiers on the curated set.

### 4.4 SH3 domain mediated PPI predictions

30 PWMs representing multiple binding specificities of 25 SH3 domains in yeast are constructed using phage display data from Tonikian *et al.* (2009) as described in section 3.1 (Table S1 & S2). These PWMs are then used to predict SH3 domain-peptide interactions using the combined classifier. 534 unique PPIs (1,481 binding sites) are predicted as positives for the stringent p-value PWM threshold of $1e-05$ with no missing features (Table S3). Approximately 55% (295 PPIs, 1,139 binding sites) of these interactions are known at the PPI level (iRefIndex & MINT) and at least 172 (464 binding sites) out of 295 PPIs are known SH3 domain mediated interactions at the peptide level (with $\geq$ 60% overlapping binding site). For example, the FUS1p SH3 domain is known to bind the STE5p protein (verified by two-hybrid assay and phage display) via an R(S/T)(S/T)SL motif, supported by two separate studies (Nelson *et al.*, 2004; Kim *et al.*, 2008). This interaction is part of the predicted set. 143 (203 binding sites) out of 239 (342 binding sites) novel interactions are of high confidence with the combined classifier scores $\geq 0.9$. Biological pathway enrichment (KEGG (Kanehisa, 2002) and Reactome (Croft *et al.*, 2014)) of the interactors reveal that a number of over-represented processes or pathways are associated with known SH3 domain biology such as endocytosis (Tonikian *et al.*, 2009; Xin *et al.*, 2013), MAPK signaling (Lyons *et al.*, 1996), and Rho GTPase signaling (Bishop and Hall, 2000) (Table S4). For example, some interacting partners of the MYO3 SH3 domain are found to be enriched in PI3K/AKT signaling. AKT is known to regulate actin organization and cell motility during endocytosis (Koral *et al.*, 2014; Enomoto *et al.*, 2005). MYO3 is also implicated in actin organization for the internalization step in endocytosis (Toret and Drubin, 2006) (Table S5). These examples support our results and suggest that our predicted interactions are biologically relevant.

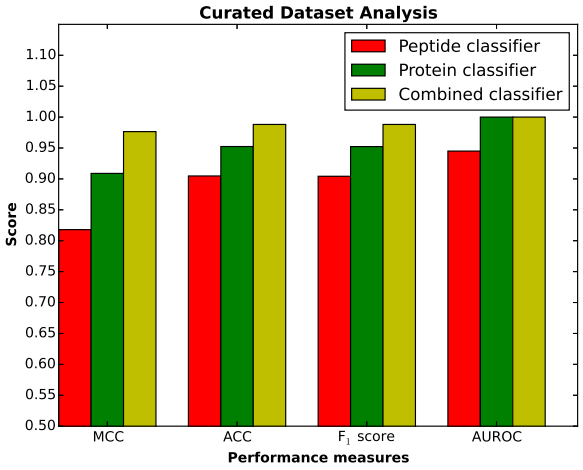



Fig. 3: Performance of peptide, protein, and combined classifiers on the curated SH3 domain mediated PPI set.

## 5 CONCLUSION

We developed a novel method for predicting physiologically relevant PPIs in yeast. This method combines diverse binding site (peptide) features, including presence in a disordered region of the protein, surface accessibility, conservation across different yeast species, and structural contact with the SH3 domain, as well as protein features such as cellular proximity, shared biological process, similar molecular function, correlated gene expression and sequence signature. Two separate Bayesian models are used to combine peptide and protein features. Their respective posterior probabilities are further combined using Bayes rule for predicting high confidence interactions. The combination of peptide and protein models achieved a higher accuracy of 0.97 compared

to individual models on a curated benchmark dataset from Tonikian *et al.* (2009). Disordered region and surface accessibility data from the peptide feature set and biological process, cellular location and sequence signature information from the protein feature set are able to separate positive from negative interactions significantly better than other features. The method presented is generic and modular in nature. Given binding peptide and feature data, we expect it can be used to predict other PRM mediated PPIs in yeast and other organisms. Additional features such as network topology, protein expression, and text mining derived protein relationships can be added to our framework. Future development includes testing this method on other PRMs in different organisms, especially human.

## IMPLEMENTATION

The DoMo-Pred command line tool is implemented using Python 2.7 and C++. It is available for download under the GNU LGPL license from `http://www.baderlab.org/Software/DoMo-Pred`

## ACKNOWLEDGEMENTS

We thank David Gfeller for help collecting binding peptide data and Mohamed Helmy for critical reading of the manuscript.

*Funding*: This work was supported by the Canadian Institutes of Health Research grant to GDB (MOP-84324).

## REFERENCES

Adamczak, R., Porollo, A., and Meller, J. (2004). Accurate prediction of solvent accessibility using neural networks–based regression. *Proteins: Structure, Function, and Bioinformatics*, **56**(4), 753–767.

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J., and Zdobnov, E. M. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*, **29**(1), 37–40.

Beltrao, P. and Serrano, L. (2005). Comparative genomics and disorder prediction identify biologically relevant SH3 protein interactions. *PLoS Comput Biol*, **1**(3), e26.

Bhardwaj, N. and Lu, H. (2005). Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics*, **21**(11), 2730–2738.

Bishop, A. and Hall, A. (2000). Rho gtpases and their effector proteins. *Biochem J*, **348**, 241–255.

Braun, P., Aubourg, S., Van Leene, J., De Jaeger, G., and Lurin, C. (2013). Plant protein interactomes. *Annual review of plant biology*, **64**, 161–187.

Chen, J. R., Chang, B. H., Allen, J. E., Stiffler, M. A., and MacBeath, G. (2008). Predicting PDZ domain–peptide interactions from primary sequences. *Nature biotechnology*, **26**(9), 1041–1045.

Chen, T. S., Petrey, D., Garzon, J. I., and Honig, B. (2015). Predicting peptide-mediated interactions on a genome-wide scale.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., *et al.* (2014). The reactome pathway knowledgebase. *Nucleic acids research*, **42**(D1), D472–D477.

Davey, N. E., Edwards, R. J., and Shields, D. C. (2010). Computational identification and analysis of protein short linear motifs. *Frontiers in Bioscience*, **15**, 801–825.

Davy, A., Bello, P., Thierry-Mieg, N., Vaglio, P., Hitti, J., Doucette-Stamm, L., Thierry-Mieg, D., Reboul, J., Boulton, S., Walhout, A. J., Coux, O., and Vidal, M. (2001). A protein-protein interaction map of the Caenorhabditis elegans 26S proteasome. *EMBO Rep*, **2**(9), 821–828.

Enomoto, A., Murakami, H., Asai, N., Morone, N., Watanabe, T., Kawai, K., Murakumo, Y., Usukura, J., Kaibuchi, K., and Takahashi, M. (2005). Akt/PKB regulates actin organization and cell motility via girdin/ape. *Developmental cell*, **9**(3), 389–402.

Faller, A. (1981). An average correlation coefficient. *Journal of Applied Metereology*, **203-205**, 20.

Ge, H., Liu, Z., Church, G. M., and Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. *Nat Genet*, **29**(4), 482–486.

Grigoriev, A. (2001). A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast saccharomyces cerevisiae. *Nucleic Acids Res*, **29**(17), 3513–3519.

Hu, H., Columbus, J., Zhang, Y., Wu, D., Lian, L., Yang, S., Goodwin, J., Luczak, C., Carter, M., Chen, L., *et al.* (2004). A map of WW domain family interactions. *Proteomics*, **4**(3), 643–655.

Hui, S. and Bader, G. D. (2010). Proteome scanning to predict PDZ domain interactions using support vector machines. *BMC Bioinformatics*, **11**, 507.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, **98**(8), 4569–4574.

Jain, S. and Bader, G. D. (2010). An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, **11**, 562.

Jansen, R., Greenbaum, D., and Gerstein, M. (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Res*, **12**(1), 37–46.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**(5644), 449–453.

Jin, J., Xie, X., Chen, C., Park, J. G., Stark, C., James, D. A., Olhovsky, M., Linding, R., Mao, Y., and Pawson, T. (2009). Eukaryotic protein domains as functional units of cellular evolution. *Science signaling*, **2**(98), ra76–ra76.

Kanehisa, M. (2002). The kegg database. *silico simulation of biological processes*, **247**, 91–103.

Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, **30**(14), 3059–3066.

Kim, J., Lee, C. D., Rath, A., and Davidson, A. R. (2008). Recognition of non-canonical peptides by the yeast fus1p SH3 domain: elucidation of a common mechanism for diverse SH3 domain specificities. *Journal of molecular biology*, **377**(3), 889–901.

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., *et al.* (2014). A draft map of the human proteome. *Nature*, **509**(7502), 575–581.

Koral, K., Li, H., Ganesh, N., Birnbaum, M. J., Hallows, K. R., and Erkan, E. (2014). Akt recruits dab2 to albumin endocytosis in the proximal tubule. *American Journal of Physiology-Renal Physiology*, **307**(12), F1380–F1389.

Lam, H. Y. K., Kim, P. M., Mok, J., Tonikian, R., Sidhu, S. S., Turk, B. E., Snyder, M., and Gerstein, M. B. (2010). MOTIPS: automated motif analysis for predicting targets of modular protein domains. *BMC Bioinformatics*, **11**, 243.

Landgraf, C., Panni, S., Montecchi-Palazzi, L., Castagnoli, L., Schneider-Mergener, J., Volkmer-Engert, R., and Cesareni, G.

(2004). Protein interaction networks by proteome peptide scanning. *PLoS biology*, **2**(1), e14.

Li, D., Liu, W., Liu, Z., Wang, J., Liu, Q., Zhu, Y., and He, F. (2008). PRINCESS, a protein interaction confidence evaluation system with multiple data sources. *Mol Cell Proteomics*, **7**(6), 1043–1052.

Linding, R., Russell, R. B., Neduva, V., and Gibson, T. J. (2003). GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*, **31**(13), 3701–3708.

Lyons, D. M., Mahanty, S. K., Choi, K.-Y., Manandhar, M., and Elion, E. A. (1996). The SH3-domain protein bem1 coordinates mitogen-activated protein kinase cascade activation with cell cycle control in saccharomyces cerevisiae. *Molecular and Cellular Biology*, **16**(8), 4095–4106.

MacBeath, G. and Schreiber, S. L. (2000). Printing proteins as microarrays for high-throughput function determination. *Science*, **289**(5485), 1760–1763.

Mayer, B. J. (2001). SH3 domains: complexity in moderation. *Journal of Cell Science*, **114**(7), 1253–1263.

McCraith, S., Holtzman, T., Moss, B., and Fields, S. (2000). Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc Natl Acad Sci U S A*, **97**(9), 4879–4884.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.

Morgan, A. A. and Rubenstein, E. (2013). Proline: the distribution, frequency, positioning, and common functional roles of proline and polyproline sequences in the human proteome. *PLoS One*, **8**(1), e53785.

Nelson, B., Parsons, A. B., Evangelista, M., Schaefer, K., Kennedy, K., Ritchie, S., Petryshen, T. L., and Boone, C. (2004). Fus1p interacts with components of the hog1p mitogen-activated protein kinase and cdc42p morphogenesis signaling pathways to control cell fusion during yeast mating. *Genetics*, **166**(1), 67–77.

Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Research*, **31**(13), 3635–3641.

Pawson, T. and Gish, G. D. (1992). SH2 and SH3 domains: from structure to function. *Cell*, **71**(3), 359–362.

Pawson, T. and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**(5618), 445–452.

Pawson, T. and Schlessingert, J. (1993). SH2 and SH3 domains. *Current Biology*, **3**(7), 434–442.

Pei, J. and Grishin, N. V. (2001). AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**(8), 700–712.

Pizzi, C., Rastas, P., and Ukkonen, E. (2011). Finding significant matches of position weight matrices in linear time. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, **8**(1), 69–79.

Rain, J. C., Selig, L., Reuse, H. D., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A., and Legrain, P. (2001). The protein-protein interaction map of Helicobacter pylori. *Nature*, **409**(6817), 211–215.

Reimand, J., Hui, S., Jain, S., Law, B., and Bader, G. D. (2012). Domain-mediated protein interaction prediction: From genome to network. *FEBS Lett*, **586**(17), 2751–2763.

Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, **314**(5), 1041–1052.

Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, **23**(8), 951–959.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res*, **32**(Database issue), D449–D451.

Schlessinger, J. (1994). SH2/SH3 signaling proteins. *Current Opinion in Genetics & Development*, **4**(1), 25–30.

Skrabanek, L., Saini, H. K., Bader, G. D., and Enright, A. J. (2008). Computational prediction of protein-protein interactions. *Mol Biotechnol*, **38**(1), 1–17.

Sobolev, V., Eyal, E., Gerzon, S., Potapov, V., Babor, M., Prilusky, J., and Edelman, M. (2005). SPACE: a suite of tools for protein structure prediction and analysis based on complementarity and environment. *Nucleic Acids Research*, **33**(suppl 2), W39–W43.

Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, **311**(4), 681–692.

Stiffler, M. A., Chen, J. R., Grantcharova, V. P., Lei, Y., Fuchs, D., Allen, J. E., Zaslavskaia, L. A., and MacBeath, G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. *Science*, **317**(5836), 364–369.

The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nat Genet*, **25**(1), 25–29.

Tong, A. H. Y., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., *et al.* (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **295**(5553), 321–324.

Tonikian, R., Zhang, Y., Sazinsky, S. L., Currell, B., Yeh, J.-H., Reva, B., Held, H. A., Appleton, B. A., Evangelista, M., Wu, Y., Xin, X., Chan, A. C., Seshagiri, S., Lasky, L. A., Sander, C., Boone, C., Bader, G. D., and Sidhu, S. S. (2008). A specificity map for the PDZ domain family. *PLoS Biol*, **6**(9), e239.

Tonikian, R., Xin, X., Toret, C. P., Gfeller, D., Landgraf, C., Panni, S., Paoluzi, S., Castagnoli, L., Currell, B., Seshagiri, S., Yu, H., Winsor, B., Vidal, M., Gerstein, M. B., Bader, G. D., Volkmer, R., Cesareni, G., Drubin, D. G., Kim, P. M., Sidhu, S. S., and Boone, C. (2009). Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. *PLoS Biol*, **7**(10), e1000218.

Toret, C. P. and Drubin, D. G. (2006). The budding yeast endocytic pathway. *Journal of Cell Science*, **119**(22), 4585–4587.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, **403**(6770), 623–627.

Vogel, C. and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, **13**(4), 227–232.

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., and Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*, **38**(Web Server issue), W214–W220.

Wu, T. D., Nevill-Manning, C. G., and Brutlag, D. L. (2000). Fast probabilistic analysis of sequence function using scoring matrices. *Bioinformatics*, **16**(3), 233–244.

Xin, X., Gfeller, D., Cheng, J., Tonikian, R., Sun, L., Guo, A., Lopez, L., Pavlenco, A., Akintobi, A., Zhang, Y., *et al.* (2013). SH3 interactome conserves general function over specific form. *Molecular Systems Biology*, **9**(1).

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X., *et al.* (2011). Next-generation sequencing to generate interactome datasets. *Nature Methods*, **8**(6), 478–480.

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., *et al.* (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, **490**(7421), 556–560.