

COMPUTATIONAL MODELS FOR DOMAIN-PEPTIDE MEDIATED PROTEIN-PROTEIN
INTERACTIONS

by

Shobhit Jain

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Department of Computer Science
University of Toronto

© Copyright 2018 by Shobhit Jain

Abstract

Computational Models For Domain-Peptide Mediated Protein-Protein Interactions

Shobhit Jain

Doctor of Philosophy

Department of Computer Science

University of Toronto

2018

Peptide recognition module (PRM) mediated protein-protein interactions (PPIs) are critical for better understanding the relationship between genomes and networks. High-throughput experimental screens, such as phage display, can be used to identify their binding motifs for use in computationally predicting high confidence PPIs with binding site information. Computational approaches for predicting protein interactions are either limited by their inability to predict peptide recognition module mediated interactions or do not consider many known constraints governing these interactions. A novel ensemble method for predicting *in vivo* SH3 domain-peptide mediated PPIs in *S. cerevisiae* and *H. sapiens* using phage display data is presented. As with similar methods, this method uses position weight matrix models of protein linear motif preference in combination with a range of evidence sources related to binding site 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 peptide based and protein pair based evidence sources. A novel semi-supervised training framework is used to train peptide and protein Gaussian naïve Bayes models using both labeled and unlabeled datasets.

Research into the different evidence sources led to the development of state-of-the-art algorithms for predicting PPIs using semantic similarity in the Gene Ontology (GO), gene or protein expression, and network topology. A novel method to compute semantic similarity between GO terms annotated to proteins in interaction datasets which considers the unequal depths of the ontology is developed (TCSS). Most PPI prediction methods rely on observations from a single gene expression profile (GEP) to predict novel interactions. Correlation coefficients from multiple GEPs are combined into a single model to improve PPI prediction. Protein expression is proposed as a new evidence source for predicting PPIs. A machine learning model is developed for predicting high confidence

PPIs using graph descriptors, such as edge density, transitivity, and mutual clustering coefficient of known interaction networks (NTOP). All the algorithms developed during this research are open source and freely available for community use.

Acknowledgements

I am grateful to many people who have helped me throughout my research. Much gratitude is owed to my supervisor, Gary Bader, for mentoring me during my PhD studies and for his continuous support and patience. Gary gave me the freedom to pursue my passions and without his scientific insight and guidance this thesis would not be possible.

I would also like to thank my supervisory committee members: Alan Moses, Michael Brudno and Quaid Morris for taking time out of their busy schedules and helping me out with my research. Their encouragement and insightful comments during numerous committee meetings have been most valuable. I am honored to be a student of the Department of Computer Science at the University of Toronto. I benefited immensely from my interactions in the department with some of the best minds in the field of computer science and all the resources made available to me by the department during my research.

A big thank you to the many colleagues, collaborators and contributors who have shared their knowledge and data with me over the years. A special thank you to Shirley Hui, Brian Law, Juri Reimand, Chris Tan, David Gfeller and Mohamed Helmy for their collaboration on many projects. Chris did the critical assessment of my manuscript which forms the basis of second chapter, David and Mohamed helped with data collection and critical reading of the third chapter of my thesis. Many thanks to Joan Teyra, Haiming Huang and Sachdev S Sidhu for providing me with early access to their datasets and help me in completing the final research chapter of my thesis.

I am indebted most of all to my parents for their extraordinary support throughout my studies. Finally, my wife Anu and my daughter Aaria, who have shared with me all the ups and downs of my PhD journey, receive my utmost gratitude, admiration, and love.

Contents

Acknowledgements	iv
Table of Contents	v
List of Tables	viii
List of Figures	x
1 Introduction	1
1.1 Introduction	1
1.2 Protein-protein interactions	2
1.3 PRM mediated PPIs	3
1.3.1 Src homology 3 (SH3) domains	3
1.3.2 WW domains	3
1.3.3 PSD95/DlgA/Zo-1 (PDZ) domains	4
1.4 Experimental detection	4
1.5 Computational prediction	5
1.5.1 Structure based approaches	6
1.5.2 Sequence based approaches	7
1.5.3 Context based approaches	10
1.5.4 Ensemble approaches	13
1.6 Thesis rationale	18
2 Semantic Similarity	20
2.1 Abstract	20
2.2 Introduction	21

2.3	Approach	25
2.4	Methods	25
2.4.1	Algorithm	25
2.5	Results	31
2.5.1	Data acquisition and processing	31
2.5.2	Model evaluation	32
2.6	Discussion	46
2.7	Conclusions	51
3	DoMo-Pred 1.0	61
3.1	Abstract	61
3.2	Introduction	62
3.3	Approach	64
3.4	Methods	64
3.4.1	Position weight matrix and proteome scanning	64
3.4.2	Peptide features	67
3.4.3	Protein features	69
3.4.4	Bayesian integration	71
3.5	Results	73
3.5.1	Model training	73
3.5.2	Feature selection	75
3.5.3	Model evaluation	78
3.5.4	SH3 domain mediated PPI predictions	81
3.6	Conclusion	86
4	DoMo-Pred 2.0, NTOP, Semi-supervised Training	88
4.1	Abstract	88
4.2	Introduction	89
4.3	Approach	91
4.4	Methods	91
4.4.1	Position weight matrix and proteome scanning	91
4.4.2	Peptide features	93
4.4.3	Protein features	94

4.4.4	Protein expression	96
4.4.5	Network Topology	96
4.4.6	Semi-supervised training of naïve Bayes model	99
4.5	Results	103
4.5.1	Model training	103
4.5.2	Feature selection	105
4.5.3	Model evaluation	106
4.5.4	SH3 domain mediated PPI predictions	109
4.6	Conclusion	112
5	Summary and future directions	115
5.1	Summary of our major contributions	115
5.1.1	Cellular location, biological process, molecular function	115
5.1.2	Gene expression	116
5.1.3	Protein expression	116
5.1.4	Network topology (NTOP)	117
5.1.5	Semi-supervised training	117
5.1.6	Domain-Motif Mediated Interaction Prediction (DoMo-Pred)	118
5.2	Future directions	119
5.2.1	Additional features and other domains	119
5.2.2	PRM-mediated protein-protein interaction networks in human disease	120
	Bibliography	121

List of Tables

1.1	Experimental PPI detection methods based on classification scheme proposed by Phizicky and Fields (1995).	5
2.1	Distribution of positive and negative interactions. Number of interactions in the positive dataset for cellular component (CC), biological process (BP), and molecular function (MF) ontologies.	33
2.2	Area under ROC curves for the <i>S. cerevisiae</i> PPI dataset.	34
2.3	Improvement in F_1 score for the <i>S. cerevisiae</i> PPI dataset.	34
2.4	Improvement in F_1 score for <i>H. sapiens</i> PPI dataset.	41
3.1	Peptide classifier: area under ROC curve (AUROC), area under precision-recall curve (AUPRC), Brier score (BRIER), F_1 -score, Matthews correlation coefficient (MCC) and accuracy (ACC) for different models.	78
3.2	Protein classifier: area under ROC curve (AUROC), area under precision-recall curve (AUPRC), Brier score (BRIER), F_1 -score, Matthews correlation coefficient (MCC) and accuracy (ACC) for different models.	79
3.3	Performance evaluation of peptide and protein classifiers using filtered and unfiltered datasets.	80
3.4	List of yeast SH3 domains from Tonikian <i>et al.</i> (2009) and their binding motifs (trimmed) with significant amino acid positions within those motifs.	82
3.5	Enrichment analysis of predicted high confidence interactors.	85
3.6	Enrichment analysis of predicted MYO3 interactors.	86
4.1	Matthews correlation coefficient, accuracy and F_1 -score for NTOP, mutual clustering coefficient and predicting PPIs by completing defective cliques methods.	99

4.2	Peptide classifier: Matthews correlation coefficient (MCC), accuracy (ACC), and F1-score for different models with increasing number of unlabeled data.	108
4.3	Protein classifier: Matthews correlation coefficient (MCC), accuracy (ACC), and F1-score for different models with increasing number of unlabeled data.	108

List of Figures

2.1	Mutually exclusive sub-graphs.	27
2.2	Graphical illustration of the TCSS algorithm.	29
2.3	ROC curves for <i>S. cerevisiae</i> PPI dataset.	35
2.4	F-score curves for <i>S. cerevisiae</i> PPI dataset.	36
2.5	ROC curves for <i>S. cerevisiae</i> PPI dataset (IEA-).	37
2.6	F ₁ -score curves for <i>S. cerevisiae</i> PPI dataset (IEA-).	38
2.7	ROC curves for <i>S. cerevisiae</i> PPI dataset (IEA+).	39
2.8	F ₁ -score curves for <i>S. cerevisiae</i> PPI dataset (IEA+).	40
2.9	ROC curves for <i>H. sapiens</i> PPI dataset (IEA-).	42
2.10	F ₁ -score curves for <i>H. sapiens</i> PPI dataset (IEA-).	43
2.11	ROC curves for <i>H. sapiens</i> PPI dataset (IEA+).	44
2.12	F ₁ -score curves for <i>H. sapiens</i> PPI dataset (IEA+).	45
2.13	Pearson correlation between gene expression similarity and semantic similarity on <i>S. cerevisiae</i> dataset (BMA approach).	46
2.14	Correlation between semantic similarity and gene expression, sequence similarity, enzyme commission (EC) similarity, protein family (Pfam) similarity (max approach)	47
2.15	Correlation between semantic similarity and sequence, enzyme commission (EC), protein family (Pfam) similarity using online CESSM tool.	48
2.16	Comparison of our topological clustering method and Resnik (MAX) as scoring positive and negative PPIs.	50
2.17	Effect of topology cutoff on (ROC) AUC and F-score for <i>S. cerevisiae</i> PPI dataset (IEA-).	53
2.18	Effect of topology cutoff on (ROC) AUC and F-score for <i>S. cerevisiae</i> PPI dataset (IEA+).	54

2.19	Topology cutoff for <i>S. cerevisiae</i> PPI dataset.	55
2.20	Effect of topology cutoff on (ROC) AUC and F-score for <i>H. Sapiens</i> PPI dataset (IEA-).	56
2.21	Effect of topology cutoff on (ROC) AUC and F-score for <i>H. Sapiens</i> PPI dataset (IEA+).	57
2.22	Topology cutoff for <i>H. Sapiens</i> PPI dataset.	58
3.1	Work flow of PRM mediated PPI prediction pipeline.	65
3.2	Change in average area under the curve (AUC) with the number of yeast gene expression datasets used for predicting PPIs.	70
3.3	SH3 domain binding motifs in MINT database	74
3.4	Negative peptide set motif	74
3.5	Prediction efficacy of individual peptide and protein features.	75
3.6	Distribution of positive and negative dataset score for peptide and protein features.	76
3.7	Maximal information coefficients for peptide and protein feature sets.	77
3.8	Performance of peptide, protein, and combined classifiers on the curated SH3 domain mediated PPI set. (Note: small size of curated validation dataset prevents the variance from being estimated.)	81
4.1	Work flow of PRM mediated PPI prediction pipeline.	92
4.2	Prediction efficacy of individual peptide and protein features.	105
4.3	Maximal information coefficients for peptide and protein feature sets.	107
4.4	Recursive feature elimination plots for peptide and protein classifiers.	107
4.5	Performance of peptide, protein and combined classifiers on the curated SH3 domain mediated PPI set. (Note: small size of curated validation dataset prevents the variance from being estimated.)	109
4.6	Human SH3 domain specificity map.	110
4.7	Human SH3 domain specificity map.	111
4.8	Enrichment map of proteins involved in SH3 domain mediated PPIs predicted by DoMo-Pred.	113

Chapter 1

Introduction

Some sections in this chapter were published in FEBS Letters, 14;586(17):2751-63: Reimand J., Hui S., Jain S., Law B., Bader GD. (2012), Domain-mediated protein interaction prediction: From genome to network.

Author contributions: I contributed to the introduction and physiologically relevant protein-protein interactions section. Other authors contributed to different sections of this paper. Gary D. Bader supervised and advised this work.

1.1 Background¹

Almost all cellular processes are controlled by specific protein-protein interactions (PPIs) that are ultimately encoded in the genome. Understanding this relationship between interaction networks and genome will not only help us in predicting biologically relevant protein interactions directly from the genome but will also help us in understanding how genomic changes impact interaction networks both over evolution and within a population or an individual organism. Advances in experimental technologies have led to the availability of large datasets of genomes and protein interaction networks, but still it is difficult to accurately relate the two. Computational methods can play an important role in bridging this gap. Ideally, methods should be developed to predict PPIs along with their binding sites directly from the genome. Once binding sites are known, we can identify how changes at the DNA level affect those sites and thus the interactions. However, accurately predicting PPIs along with their binding sites directly from genome is difficult. But

¹This section is derived from our published work (Reimand *et al.*, 2012).

methods can be developed to predict important subclasses of interactions more accurately. One such subclass is that of peptide recognition module (PRM) mediated interactions. PRMs are protein domains which recognize short, linear amino acid sequences in other proteins. Interactions mediated by them are essential for a normal cellular life and any deviations often result in abnormal cellular behavior and disease (Pawson and Nash, 2003). They are widespread in eukaryotic genomes and their binding preferences can be determined using high-throughput experimental techniques. Once the binding preferences of PRMs are known, computational methods can be developed to predict physiologically relevant (true positives) PRM mediated PPIs, that is, interactions which are more likely to take place *in vivo* (Tong *et al.*, 2002; Reimand *et al.*, 2012).

1.2 Protein-protein interactions

Proteins are essential macromolecules which are involved in almost all cellular processes such as transport, signaling, regulation, respiration, metabolism, development, repair and control of genes. Proteins do not usually work alone but interact with other proteins, forming complexes and networks. PPIs are physical associations between protein pairs in a specific biological context. Their knowledge provide important insights into the functioning of a cell. PPIs can be divided into two major groups: permanent and transient. Permanent interactions, as the suggests, are strong and irreversible. On the other hand, transient interactions are usually reversible and take place in a cellular context (Perkins *et al.*, 2010). Transient interactions are involved in many biological processes such as regulation of biochemical pathways and signaling cascades. Because of their crucial role in many disease related pathways transient PPIs are also important drug targets (Ozbabacan *et al.*, 2011). Previously, experimental detection of PPIs was limited to labor intensive techniques such as co-immunoprecipitation or affinity chromatography (Skrabaneck *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).

1.3 PRM mediated PPIs

Multiple kinds of transient PPIs exist. Our focus is on those involving PRMs, such as SH3, PDZ, and WW domains. These domains bind to small, linear sequence motifs (peptides) within proteins. They are involved in important biological processes including signaling systems and human diseases (Reimand *et al.*, 2012). Their binding preferences can be identified using high-throughput experimental techniques such as phage display, peptide chips, and yeast two-hybrid (Tonikian *et al.*, 2008, 2009; Tong *et al.*, 2002; Landgraf *et al.*, 2004; Hu *et al.*, 2004).

1.3.1 Src homology 3 (SH3) domains

SH3 domains are approximately 60 amino acids long and fold into a beta-barrel structure composed of five to six anti-parallel beta strands. SH3 domain has a flat, hydrophobic surface which consists of three shallow pockets with conserved aromatic residues. They often bind to proline-rich regions containing a core PxxP motif flanked by positively charged residues (where x is any amino acid). Class I domains bind to ligands conforming to the consensus sequence [R/K]xxPxxP and class II domains recognize PxxPx[R/K] sequence (Mayer, 2001; Teyra *et al.*, 2012). Though, more recently it has been found that SH3 domains have much wider binding specificity. In some cases they can even bind to proline-free regions (Tong *et al.*, 2002; Tian *et al.*, 2006; Kim *et al.*, 2008; Pires *et al.*, 2003). SH3 domains are involved in many regulatory or signaling processes, including endocytosis, actin cytoskeleton regulation, and tyrosine kinase pathways (Tonikian *et al.*, 2009; Schlessinger, 1994).

1.3.2 WW domains

WW domains are 30–40 amino acid in length and fold into a triple stranded beta sheet and contain two tryptophan residues spaced approximately 20 residues apart from each other. They have a flat and hydrophobic binding surface (Wintjens *et al.*, 2001). Like SH3 domains, they also recognize proline-rich motifs such as xPPxY (Dalby *et al.*, 2000). Proteins with WW domains are involved in many regulatory and signaling processes, including growth control, ubiquitin-mediated proteolysis, transcription, and control of cytoskeleton (Reimand *et al.*, 2012; Salah *et al.*, 2012).

1.3.3 PSD95/DlgA/Zo-1 (PDZ) domains

PDZ domains are one of the simplest PRMs, since they mostly bind to C-terminal tails of other proteins. PDZ domains are 80 – 90 amino acids in length and folds into a globular structure consisting of six β strands and two α helices. PDZ domains prefer hydrophobic residues. Their binding specificities can be divided into two classes, where class I domains prefer to bind $x[T/S]x\Phi$ motif and class II domains prefer motif $x\Phi x\Phi$ (where Φ is a hydrophobic amino acid) (Songyang *et al.*, 1997). Though, more recently it has been found that PDZ domains can recognize more than these two classes (Tonikian *et al.*, 2008). PDZ domains regulate many signaling and regulatory processes including ion channels, localize signaling components to the membrane, participate in cell polarity, and are involved in neural development (Tonikian *et al.*, 2008; Lee and Zheng, 2010; Reimand *et al.*, 2012).

1.4 Experimental detection

Experimental PPI detection techniques can be broadly classified into three major categories: physical methods, library-based methods, and genetic methods (Phizicky and Fields, 1995). Some of the widely used experimental techniques are summarized in Table 1.1. In yeast-two-hybrid (Y2H) assays, pairs of proteins to be tested for interaction are expressed as fusion proteins (bait & prey) in yeast. The bait protein is fused to a transcription factor DNA binding domain, the prey protein, is fused to a transcription factor activation domain. When expressed in a yeast cell containing the appropriate reporter gene, interaction of the bait with the prey brings the DNA binding domain and the activation domain into close proximity thus creating a functional transcription factor. This triggers transcription of the reporter gene. The interaction can then be detected by expression of the linked reporter genes (Phizicky and Fields, 1995; Chien *et al.*, 1991; Fields and Song, 1989; von Mering *et al.*, 2002).

Another widely used PPI detection technique is affinity-purification mass spectrometry (AP-MS). The principle behind AP-MS based protein interaction detection experiments is using a protein as an affinity reagent to isolate its binding partners and indentifying them using MS. AP-MS methods have three essential components: bait presentation, affinity purification of the complex, and analysis of the bound proteins (Aebersold and Mann, 2003). In a generic strategy based on tandem affinity purification (TAP) the protein of interest is tagged by a TAP tag containing sequence recognizable by an antibody. The tagged proteins are introduced into the host cells or

Type	Method	Description
Physical	Affinity chromatography	Separation of protein mixtures based on specific interactions
	Affinity blotting	Fractionating protein mixtures using PAGE
	Immunoprecipitation	Precipitating a protein antigen with antibody
	Cross-linking	Detect proteins that interact with a given test protein ligand by probing
Library-based	Protein probing	Labeled protein as a probe to screen an expression library
	Phage display	Sequences expressed in phage library bind to target protein
	Two-hybrid system	Uses transcriptional activity as a measure of PPI
Genetic	Synthetic lethal effects	Mutations in two genes can cause death while mutation in either alone does not
	Overproduction phenotypes	Overproduction of mutant or wild-type proteins

Table 1.1: Experimental PPI detection methods based on classification scheme proposed by Phizicky and Fields (1995).

organism and expressed to optimal levels. Cell extracts are prepared and the fusion protein as well as associated partners are recovered by passing the extract through a tag specific antibody column and a calmodulin binding column. The elution consisting of the protein of interest and its interacting protein partners are analyzed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and identified by MS using electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI) techniques (Puig *et al.*, 2001; Aebersold and Mann, 2003)

As discussed earlier, many intracellular signalling processes are mediated by interactions between peptide recognition modules (or domains) and small, continuous sequence motifs (peptides) within proteins. Phage display technology has been especially useful in studying such PRM interactions. Phage technology permits the display of extremely diverse libraries of peptides or proteins on the surface of phage particles. Once created these libraries can then be selected by binding to immobilized proteins (Sidhu *et al.*, 2003). In the past few years, there have been numerous developments in this technology to make it applicable to a variety of protein-protein, protein-peptide, and domain-peptide interactions (Phizicky and Fields, 1995; Smith, 1985; Tonikian *et al.*, 2007). Peptide chip technology is another way of detecting peptides binding to PRMs. Potential binding peptides are immobilized on a glass chip and selected by a domain (Carducci *et al.*, 2012). Peptide chip technology is limited in its ability to identify novel motifs as only a small number of peptides can be immobilized.

1.5 Computational prediction

Computational methods provide a complementary approach to detecting PPIs experimentally. In general, all computational approaches make use of accurate experimental PPIs to predict novel PPIs or assess PPIs reported by high-through experiments (Pitre *et al.*, 2008). The computational PPI

prediction can be formulated both as a prediction and validation problem with similar solutions:

1. **Prediction:** Given two proteins predict whether they will interact with each other.
2. **Validation:** Given a protein-protein interaction detected by an experiment (high-throughput) assign a confidence score to it.

Different PPI detection methods can be broadly classified into four different categories based on the type of information used by them either structure based, sequence based, contextual, or ensemble of these information.

1.5.1 Structure based approaches

Structural approaches for predicting PPIs in general make use of three-dimensional protein structures or protein complex structures available through protein structure database. The prediction process generally starts with identifying homologous protein or protein complex structures for the query sequences and then using interface information to model the interaction. Hue *et al.* used a support vector machine (SVM) to predict domain-domain interactions using a kernel derived from protein structure information at a large scale (Hue *et al.*, 2010). MULTIPROSPECTOR uses threading and protein-protein interfacial energy for PPI predictions (Lu *et al.*, 2002). InterPreTS assess the fit of any possible interacting pair on the homologous three-dimensional complex by using empirical potentials (Aloy and Russell, 2003). PRISM uses structure and evolutionary conserved residue similarity of query sequences to structurally known protein interfaces formed between dimers, trimers, or higher protein complexes for PPI prediction (Aytuna *et al.*, 2005). HOMCOS predicts interacting protein pairs and their interaction sites by homology modeling of complex structures (Fukuhara and Kawabata, 2008).

Structural features within the binding pocket of a PRM play an important role in determining its binding specificity (Reimand *et al.*, 2012). Therefore, structural information play an important role in accurately predicting PRM mediated PPIs. Sanchez *et al.* used an empirical force field to calculate structure-based energy functions for human SH2 domain mediated interactions (Sanchez *et al.*, 2008). Fernandez-Ballester *et al.* used SH3 structural features to homology model most SH3 domains in yeast and then constructed positional matrices of all possible SH3-ligand complexes to predict SH3 domain mediated interactions (Fernandez-Ballester *et al.*, 2009). Hui *et al.* used a SVM trained on PDZ domain structure and peptide sequences for predicting PPIs (Hui and Bader, 2010). Smith *et al.* used protein backbone sampling using independent Monte Carlo simulations

to predict binding specificity for human PDZ domains (Smith and Kortemme, 2010). Kaufmann *et al.* developed an optimized energy function using PDZ domain-peptide interfaces to improve the binding specificity of PDZ domains (Kaufmann *et al.*, 2011). Structural approaches are usually limited by the small number of protein sequences with accurate structural information. However, they allow for more accurate prediction along with the identification of binding residues (Skrabanek *et al.*, 2008).

1.5.2 Sequence based approaches

Sequence based methods for interaction prediction utilizes the information about genes or protein sequences to predict PPIs. Completely sequenced genomes provide genomic information such as gene order, evolutionary conservation, co-localization, and fusion to be used in PPI prediction (Pitre *et al.*, 2008; Skrabanek *et al.*, 2008). Co-localization based PPI prediction methods make use of the notion that the gene which interact with each other are kept in close physical proximity to each other on the genomes (Dandekar *et al.*, 1998; Overbeek *et al.*, 1999; Skrabanek *et al.*, 2008; Tamames *et al.*, 1997). Phylogenetic evolution based methods takes into account co-occurrence of gene pairs across different genomes (Pellegrini *et al.*, 1999). Gene fusion is complementary to both co-localization and phylogenetic analysis. Gene fusion represents the physical fusion of two separate parent genes into a single multi-functional gene (Skrabanek *et al.*, 2008).

Sequence signature ²

Protein interactions can also be predicted based on correlated sequence motifs. These motifs are learned from existing PPIs using only sequence data and characterize direct binding, but also could be related to protein function, which is in turn predictive of PPIs (Shen *et al.*, 2007). Methods based on information content analyze co-occurring subsequences of proteins with experimentally verified interactions, and use these patterns for predicting new interactions. Pitre *et al.* (2006) developed Protein-protein Interaction Prediction Engine (PIPE), which looks for the co-occurrences of subsequences of a protein pair in known interactions. Najafabadi and Salavati (2008) introduced a codon usage based method as a predictor for PPIs. Sprinzak and Margalit (2001) attempted to identify over-represented sequence signatures in known PPIs and then used this information for predicting novel interactions.

Machine learning methods use sequence information regarding a gold standard set of positive

²This section is reproduced from our published work (Reimand *et al.*, 2012)

and negative PPIs to classify new pairs of potentially interacting proteins. Various approaches mainly differ in their encoding of sequence features and choice of learning functions. For instance, Martin and co-workers (Martin *et al.*, 2005) encoded the sequence information for a protein pair by a product of signatures, which is then used by a support vector classifier (SVC). Shen *et al.* (2008) proposed a SVC based classifier in which protein sequences are encoded by conjoint triads i.e. frequencies of 3 continuous amino acid long subsequences. Guo *et al.* (2008) used a feature vector comprising of auto-correlation values of 7 different physicochemical scales for protein sequences. Nanni and Lumini (2006) proposed a new method based on an ensemble of K-Local Hyperplane Distance Nearest Neighbor (HKNN) classifiers, where each HKNN is trained using a different physicochemical property of the amino-acids. Roy *et al.* (2009) explored the contribution of pure amino acid composition for protein interaction prediction using naïve Bayes (NB), SVM, and maximum entropy classifier.

A major limitation of sequence signatures for predicting protein interactions is the generally weak correlation between sequence and functional similarity. Limitations of these machine learning methods are the lack of well-defined true negative examples. For instance, Yu *et al.* evaluated the impact of positive-to-negative ratio in training and test sets for SVM based methods and found that it had considerable effect on classifier accuracy (Yu *et al.*, 2010). Lastly, use of sequence signatures to refine high-resolution PRM-mediated interaction networks must avoid duplicate counting of the domain-motif interaction knowledge already used to generate the original network.

Position weight matrix

As discussed earlier, experimental methods such as phage display and peptide microarray have been used to identify the peptides binding to PRMs. A straightforward computational approach for predicting PRM mediated PPIs is to construct a position weight matrix (PWM) using these peptides and scan the whole proteome for potential binding sites in target proteins using some threshold score (Obenauer *et al.*, 2003). Position weight matrices (PWMs) are statistical models for representing sequence motifs. They are real valued $m \times n$ matrices, where m is the size of alphabet (20 amino acids for protein sequences) and n is the motif length. PWMs contain a weight for each alphabet symbol i at each position j in the motif. They are used to scan the proteome and compute a score using alphabet weights indicating the binding preference of a domain containing protein for a peptide in another protein. Stiffler *et al.* (2007) constructed a single PWM model for many PDZ domains and used weights describing the preference of individual domains for amino

acids at different positions in the peptide. Tonikian *et al.* (2008) used experimental phage display data to derive PWMs for PDZ domain binding peptides and predicted PPIs using PWM scanning. Tonikian and co-workers again used PWMs from phage display experiment, peptide array screening, and Y2H assays for generating a yeast SH3 domain specificity map. A major issue with the PWM 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). Also, the assumption of independence between residue positions might affect its performance. PWMs may also perform poorly when too few experimentally determined peptides are available for a given protein (Reimand *et al.*, 2012; Teyra *et al.*, 2012).

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).

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). Surface accessibility can be predicted computationally from protein structures using tools such as the Eukaryotic Linear Motif (ELM) structure filter [101]. Amino acid sequence-based predictors such as PHDacc or SABLE are useful when no known protein structure is available [102].

1.5.3 Context based approaches ³

Proteins will only interact if they recognize each other, and are temporally and spatially co-located in the cell. Domain-peptide interaction predictors described above allow us to discover protein pairs that recognize each other. Additional sources of evidence must be considered to accurately score domain-peptide interactions by their physiological relevance, such as the correlation of the expression profiles of the corresponding genes, their involvement in related biological processes, and their presence in the same cellular compartment. Gene expression profiles, cellular location of proteins, functional annotation (molecular function and biological process), sequence signatures, literature references, and known experimental interactions can be obtained from diverse biological data sources and combined for predicting physiologically relevant protein interactions. Consequently, a number of computational methods have been developed for evaluating protein interactions using single sources of evidence, while others combine multiple types of knowledge in ensemble approaches. As domain-mediated networks are only now emerging, few methods have been developed specifically for these data. However, the collection of methods developed for analysing traditional protein-protein interaction networks can be combined with sequence- and structure-based domain-peptide interaction prediction methods discussed above to define high-resolution PRM-mediated interaction networks.

Cellular location, biological process, molecular function

Proteins are more likely to interact with each other when they are co-localised in the same cellular compartment or part of the same biological processes. Gene Ontology (GO) is a useful and popular taxonomy that contains a hierarchy of controlled terms regarding cellular location, biological process and molecular function (The Gene Ontology Consortium, 2000). GO terms are used to annotate genes and proteins based on experimental and computational evidence as well as literature curation. This resource can be used to quantify the functional relationship between different proteins using a straightforward comparison of associated annotations, that is, two proteins are related if they have many annotations in common. More elaborate semantic similarity measures consider the entire GO hierarchy in comparing two interacting proteins, that is, two proteins are related if they have many similar annotations in common.

Semantic similarity provides a quantitative measurement of the likeness of concepts belonging to an ontology. In the context of PPIs, higher semantic similarity scores between GO terms annotated

³This section is reproduced from our published work (Reimand *et al.*, 2012)

to a protein pair indicate a higher likelihood of these proteins interacting *in vivo*. Guo *et al.* compared a number of graph-based and information content-based semantic similarity methods in distinguishing true and false human PPIs, and concluded that the average (AVG) method by Resnik performed best in AUROC analysis (Resnik, 1995; Lin, 1998; Jiang and Conrath, 1997; Guo *et al.*, 2006). Xu *et al.* compared the AVG and maximum (MAX) methods by Resnik to a number of semantic similarity methods specifically developed for GO, and concluded that the MAX method by Resnik outperforms others when considering the three ontologies of GO either individually or together (Resnik, 1995; Tao *et al.*, 2007; Schlicker *et al.*, 2006; Wang *et al.*, 2007; Xu *et al.*, 2008).

Prediction of protein-protein interactions using GO has several limitations. Notably, GO annotations are often noisy, as more than one third of all annotations and 75% of human gene annotations are assigned using automated methods (Reimand *et al.*, 2007). Such low-confidence annotations, labelled as 'Inferred from Electronic Annotation' (IEA), should be excluded from predictions when higher quality annotations are available. Additionally, the structure of GO is often unbalanced since some biological processes are studied more extensively than others, leading to ascertainment biases in predicting protein interactions. As semantic similarity measures use knowledge structured in the form of ontologies, other ontologies could be substituted for GO. Some describe highly structured biological pathway mechanisms, such as the BioPAX pathway representation standard (Demir *et al.*, 2010). Large amounts of curated pathway data are available in this format, such as from the Reactome pathway database (Matthews *et al.*, 2009). Further development of semantic similarity methods that consider such ontologies could improve PPI prediction.

Gene and protein expression

Gene expression is a popular measure for assessing the confidence and biological relevance of predictions from high-throughput PPI experiments. As proteins must be expressed in order to interact, interacting proteins should be co-expressed at the same time and have similar gene expression profiles. The association between protein interactions and correlated gene expression profiles has been demonstrated in several studies. Co-expressed genes in yeast and bacteriophage T7 were shown to be enriched in protein interactions, and clusters of gene expression profiles frequently contained interacting proteins in yeast (Ge *et al.*, 2001). Jansen *et al.* demonstrated a strong correlation between the gene expression profiles of yeast proteins involved in the same complex (Jansen *et al.*, 2002). Bhardwaj *et al.* compared the gene expression profiles of interacting and random gene pairs in *E. coli*, and concluded that genes encoding for interacting proteins have a stronger expression

pattern correlation that is also more conserved than for random protein pairs (Bhardwaj and Lu, 2005). Consequently, PPI prediction methods frequently use strong co-expression of genes as an evidence source for protein interactions (Li *et al.*, 2008; Rhodes *et al.*, 2005).

While gene expression data is a useful source of evidence, it has a number of inherent limitations. Adler *et al.* studied curated Reactome pathways in the context of the human gene expression atlas, and concluded that co-expression is sufficient for reconstructing pathways such as metabolism and translation, while dynamic signalling processes are captured to a lesser extent (Adler *et al.*, 2009a). Liu *et al.* noted that six large protein complexes, including the ribosome, provided the majority of the signal between expression correlation and protein interactions in several gene expression datasets in yeast, while many other protein complexes did not show the association (Adler *et al.*, 2009a). Further, complex tissue-specific and developmental programs regulate gene expression in multicellular organisms, meaning that the global co-expression of potentially interacting proteins is not necessarily informative of their co-expression in a given cellular state. Future work to improve the confidence of co-expression data for high-resolution PRM-mediated interaction networks will involve novel methods for determining global co-expression of genes using multiple expression datasets (Adler *et al.*, 2009b).

Emerging experimental technologies have now made it possible to move from the human genome map to the proteome map with direct measurements of proteins and peptides. Kim and co-workers used high-resolution Fourier-transform mass spectrometry to produce a draft map of human proteome. They did in-depth proteomic profiling of 30 histologically normal human samples, including 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells, which resulted in identification of proteins encoded by 17,294 genes. As described earlier, gene expression profiles across various experimental conditions or tissues have been utilized to investigate the likelihood of co-expressed genes to physically interact at the protein level. With the availability of protein expression data, we showed that protein expression patterns should be a better predictor of PPIs than gene expression measured at the mRNA level (Kim *et al.*, 2014).

Network topology

Much work has been done in defining the relationship between PPI network topology and biological function, with the conclusion that two proteins that have many shared neighbours in a PPI network are more likely to interact (Sharan *et al.*, 2007). The property of highly connected components, i.e., network cohesiveness, in small-world networks is often used to assess the confidence of predicted

protein-protein interactions. Goldberg and Roth showed that true interactions have higher neighbourhood cohesiveness as compared to false interactions (Goldberg and Roth, 2003). Conversely, a predicted PPI is more likely to be true if it shows a higher degree of neighbourhood cohesiveness. Bader *et al.* proposed that interacting proteins with shared interactors are more likely to be biologically relevant (Bader *et al.*, 2004). Yu *et al.* predicted interactions in protein networks by completing their partially connected components, applying the assumption that proteins within the same protein cluster are likely to interact with each other (Yu *et al.*, 2006).

An important challenge for network-based protein interaction predictors is the identification of appropriate topological clusters in networks. Larger cluster sizes lead to an increased rate of false positives, while overly small clusters have few positive predictions. Clustering and cohesiveness analysis of PRM-mediated protein interaction networks may require additional research, as their topological properties may differ from traditional PPI networks. Finally, prediction of PRM-mediated protein interactions based on known interactions will require careful filtering of data to avoid duplicate counting of evidence.

1.5.4 Ensemble approaches

Each of the structure, sequence or context based approaches have the ability to classify protein pairs as "interacting" or "non-interacting" independently. Ensemble approaches for PPI prediction goes a step further and combine these individual approaches into a single model for improved predictions. By transforming multiple direct (protein-protein interaction databases) and indirect biological data sources (or evidence) into a feature vector representing every pair of proteins, the task of predicting pairwise protein interactions can be formalized as a binary classification problem. Many different research groups have independently suggested using supervised learning methods for predicting protein interactions (Patil and Nakamura, 2005; Rhodes *et al.*, 2005; Stelzl *et al.*, 2005; Li *et al.*, 2008; Chen and Liu, 2005; Qi *et al.*, 2005; Mohamed *et al.*, 2010; Scott and Barton, 2007; Eom and Zhang, 2006; Bader *et al.*, 2004; Gilchrist *et al.*, 2004; Jansen *et al.*, 2003; Lee *et al.*, 2004; Jaimovich *et al.*, 2006; Zhang *et al.*, 2004; Ben-Hur and Noble, 2005; Yamanishi *et al.*, 2004; Lin *et al.*, 2004). However, choice of biological evidence and strategy to combine them into a single model varies greatly.

Bayesian integration

Bayesian integration is the most widely used ensemble technique for PPI prediction. Although other machine learning approaches have been used for this task, such as logistic regression, random forests, decision trees, and support vector machines, Bayesian integration remains the method of choice due to its simple probabilistic framework and ability to handle missing data (Reimand *et al.*, 2012). Jansen *et al.* proposed the use of Bayesian networks on a feature set of experimental PPI data (direct evidence) and genomic features such as, mRNA co-expression, biological function, and essentiality (indirect evidence) in *Saccharomyces cerevisiae*. They fed the results of naïve Bayes model of indirect evidence and fully connected Bayesian network of direct evidence to another naïve Bayes classifier for predicting PPIs (Jansen *et al.*, 2003). Rhodes *et al.* employed a similar strategy using a semi-naïve (partially connected) Bayes classifier to combine homologous PPI, gene expression, GO Process and domain based sequence evidence (Rhodes *et al.*, 2005). Scott and Barton extended the probabilistic framework for the prediction of human PPIs with more features, which include local network topology, co-expression, orthology to known interacting proteins and the full-Bayesian combination of subcellular localization, co-occurrence of domains and post-translational modifications (Scott and Barton, 2007). Patil and Nakamura used a naïve Bayes classifier as a means to assign reliability to the PPIs in *Saccharomyces cerevisiae* determined by high-throughput experiments (Patil and Nakamura, 2005). Li *et al.* closely followed the work of Rhodes *et al.* (Rhodes *et al.*, 2005) and used a naïve Bayes classifier to combine different types of indirect biological features (Li *et al.*, 2008). More recently, Zhang *et al.* combined structural, functional, evolutionary and expression information using Bayesian framework for predicting PPIs (Zhang *et al.*, 2012a,b).

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 Bayes model simplifies this problem by assuming complete independence between different types of biological evidence. For a protein pair described by a set of features (X_1, X_2, \dots, X_n) a naïve Bayes PPI prediction model is defined as,

$$\begin{aligned}
 \operatorname{argmax}_Y P(Y|X_1, X_2, \dots, X_n) &= \operatorname{argmax}_Y \frac{P(X_1, X_2, \dots, X_n|C) P(Y)}{P(X_1, X_2, \dots, X_n)} \\
 &= \operatorname{argmax}_Y P(Y) \prod_i P(X_i|Y) \tag{1.1} \\
 \operatorname{argmax}_Y \log(P(Y|X_1, X_2, \dots, X_n)) &= \operatorname{argmax}_Y \log(P(Y)) + \sum_i \log(P(X_i|Y))
 \end{aligned}$$

where $P(Y)$ is the class prior probability and $P(X_i|Y)$ is the class-conditional probability. If the number of classes Y are small ("interacting" or "non-interacting" for PPI prediction problem) then usually class priors are estimated by treating $P(Y)$ as a multinomial distribution (or categorical distribution) $P(Y) = \Pi_Y$. Biological features (X_1, X_2, \dots, X_n) could be continuous or discrete. If the features are continuous then they are usually discretized and modeled using a multinomial probability distribution $P(X_i|Y) = \text{Mult}(X_i; \theta_{iY}) \propto \Theta_{iY}^{X_i}$. Putting it altogether, the naïve Bayes model for PPI predictions is defined as,

$$\operatorname{argmax}_Y \log(P(Y|X_1, X_2, \dots, X_n)) = \operatorname{argmax}_Y \log(\Pi_Y) + \sum_i \log(\Theta_{iY}^{X_i}) \quad (1.2)$$

the parameters Π_Y and $\Theta_{iY}^{X_i}$ are learned from the training dataset. More complex Bayesian models (semi-naïve, fully connected) though may be more accurate but are computationally expensive.

Logistic regression

Bader *et al.* used a logistic regression approach with statistical and topological descriptors to predict the biological relevance of PPIs obtained from high-throughput screens for yeast (Bader *et al.*, 2004). A logistic regression model for PPI prediction learns the function of the form $f : X \rightarrow P(Y|X)$ where $X = (X_1, X_2, \dots, X_n)$ is a vector containing discrete or continuous variables (features) and Y is "interacting" or "non-interacting" class (Mitchell, 1997). Logistic regression uses a sigmoid function to parameterize the probability distribution $P(Y|X)$. The parameterized form used by logistic regression classifier is,

$$P(Y|X) = \frac{1}{1 + e^{-(\Theta_o + \sum_{i=1}^n \Theta_i X_i)}} \quad (1.3)$$

where the model parameters Θ_i are learned from training set.

Random forest and decision trees

Lin *et al.* repeated the experiments in Jansen *et al.* (2003) with random forest and logistic regression classifiers and concluded that random forest approach gives highly accurate classifications on complete datasets. They also discussed the importance of different features and concluded that

the biological function category was the most informative (Lin *et al.*, 2004). Zhang *et al.* used a decision tree to integrate high-throughput protein interaction datasets and other gene and protein pair characteristics to predict co-complexed pairs of proteins (Zhang *et al.*, 2004). Qi *et al.* used direct and indirect information about interaction pairs to constructs a random forest (collection of decision trees) to determine the similarity between protein pairs and then using k-nearest neighbor approach to classify protein pairs (Qi *et al.*, 2005).

Random forest is an ensemble classifier that consists of many decision trees. Random forest as defined by Breiman and Schapire is "a classifier consisting of a collection of tree-structured classifiers $\{h(X, \Theta_k), k = 1, \dots\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input X " (Breiman and Schapire, 2001). The construction of tree-structured classifiers usually work top-down by choosing a feature X_i from a randomly selected set of features (X_1, X_2, \dots, X_m) (where $m < n$ and $X = (X_1, X_2, \dots, X_n)$) at every step which best splits the training data.

Support vector classifier

Yamanishi *et al.* presented a method to infer protein interaction networks using a variant of kernel canonical correlation analysis. They transformed each genomic dataset into a symmetric positive definite kernel function and then summed all the genomic kernels (Yamanishi *et al.*, 2004). Ben-Hur and Noble extended the kernel approach of Yamanishi *et al.* (2004) by integrating pairwise, sequence, non-sequence genomic kernels to build a support vector classifier (Ben-Hur and Noble, 2005). A symmetric positive definite kernel function is a real-valued function $K(p_i, p_j)$ for a set of proteins (p_1, p_2, \dots, p_n) satisfies the following properties.

$$K(p_i, p_j) = K(p_j, p_i) \tag{1.4}$$

$$\sum_{i=1}^n r_i r_j K(p_i, p_j) \geq 0 \text{ where } r_i, r_j \in \mathfrak{R}^n \tag{1.5}$$

Commonly used kernel functions $K(p_i, p_j)$ are Gaussian RBF (1.6) and linear (1.7).

$$K(p_i, p_j) = \exp\left(-\|p_i - p_j\|^2 / 2\sigma^2\right) \quad (1.6)$$

$$K(p_i, p_j) = p_i \cdot p_j \quad (1.7)$$

Different biological datasets are represented by suitable kernel functions (K_1, K_2, \dots, K_n). These kernels can be integrated by linear combination (1.8) and fed to a support vector machine for making predictions.

$$K = \sum_{i=1}^n K_i \quad (1.8)$$

Other approaches

All the above mentioned approaches consider protein pairs independently when inferring the presence of PPIs. Jaimovich *et al.* considered the neighborhood interaction pairs together and integrated genomic features using a relational Markov network for simultaneous prediction of PPIs in yeast (Jaimovich *et al.*, 2006). Apart from the above mentioned machine learning approaches, Stelzl *et al.* proposed a "voting" based approach for predicting PPIs in human. They classified high confidence PPIs based on the votes cast by experimental, topological, and GO information in their favor (Stelzl *et al.*, 2005). Brown and Jurisica presented a web-based database of predicted interactions between human proteins. It combines the literature-derived human PPI from BIND, HPRD and MINT, with predictions made from model organisms. They also evaluated their predictions using protein domains, gene co-expression and Gene Ontology terms (Brown and Jurisica, 2005).

Domain-peptide interaction prediction approaches

Previously, discussed ensemble approaches are designed for full length proteins and cannot be used to predict PRM mediated PPIs, including identification of binding sites. Tong *et al.* (2002) combined *in vitro* phage-display ligand consensus sequences with *in vivo* large-scale two-hybrid physical interaction experiments to predict SH3 domain mediated PPIs. Tonikian *et al.* (2009) combined phage display, peptide array screening and yeast two-hybrid data using Bayesian integration to predict SH3 domain mediated PPIs in yeast. Lam *et al.* (2010) combined comparative and structural genomic features with PWMs to reduce the number of false 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).

1.6 Thesis rationale

The proposed research is focused on developing computational methods for predicting physiologically relevant PRM mediated PPIs using peptides identified from phage display or peptide microarray experiments in *S. cerevisiae* and *H. sapiens*. As discussed earlier, the straightforward approach of constructing PWMs from peptides and scanning the whole proteome for potential binding sites in target proteins using some threshold score leads to too many false positives and is not sufficient to predict high confidence interactions because of missing sequence, structure, or contextual information. 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.* (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.* (2012b), 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 works in the field of 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.

In this thesis, we develop novel algorithms which make use of a larger set of evidence sources to predict PRM-mediated PPIs and their binding sites by combining peptide level and protein level features in a single predictor. PRM mediated PPIs do not occur in isolation in the cell. They

are influenced by different 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. Our proposed research has two major goals:

- Developing methods for processing of biological information that can help predict physiologically relevant PRM mediated PPIs.
- Developing methods for integrating processed biological information to predict high confidence PRM mediated PPIs.

We have identified peptide features: disorder, surface accessibility, peptide conservation, and structural contact as evidence sources for predicting high confidence binding sites. We have also identified protein features: cellular location, biological process, molecular function, gene expression, protein expression, sequence signature, and network topology as evidence sources for predicting PPIs. As discussed earlier, available methods for processing features like cellular location, biological process, molecular function, gene expression, and network topology have certain limitations and we will be focusing on improving them. We will also focus on using new experimental datasets like protein expression for PPI prediction. Next, we will explore machine learning models for integrating peptide and protein features into a single model. We will quantify the performance of our proposed models using different statistical measures and perform any potential comparisons to previous methods. Finally, we aim to construct a high confidence PRM mediated PPI network for yeast and humans with binding site resolution.

Chapter 2

An improved method for scoring protein-protein interactions using semantic similarity within the Gene Ontology

This work was published in BMC Bioinformatics, 11:562: Jain, S. and Bader, GD. (2010), An improved method for scoring protein-protein interactions using semantic similarity within the Gene Ontology.

Author contributions: I collected the data, developed and implemented the method and performed the analyses. Gary D. Bader supervised and advised this project.

2.1 Abstract

Semantic similarity measures are useful to assess the physiological relevance of protein-protein interactions (PPIs). Semantic similarity measures quantify the gene function similarity between two or more proteins. Proteins that interact in the cell are likely to be in similar locations or involved in similar biological processes compared to proteins that do not interact. Thus the more semantically similar the gene function annotations are among the interacting proteins in ontologies capturing protein location and process information like the Gene Ontology (GO), the more likely the inter-

action is to be physiologically relevant. However, most semantic similarity measures used for PPI confidence assessment do not consider the unequal depth of term hierarchies in different biological knowledge areas in gene function annotation systems, like GO. We describe an improved algorithm, Topological Clustering Semantic Similarity (TCSS), to compute semantic similarity between GO terms annotated to proteins in interaction datasets, that considers the different levels of biological knowledge representation depth in different branches of the GO graph. The central idea is to divide the GO graph into sub-graphs and score PPIs higher if participating proteins belong to the same sub-graph as compared to if they belong to different sub-graphs. The TCSS algorithm performs better than other semantic similarity measurement techniques that we evaluated on tests to distinguish true from false protein interactions, correlation with gene expression or protein families. We show an average improvement of 4.6 times in F_1 scores over Resnik, the next best method, on our *Saccharomyces cerevisiae* PPI test and 2 times on our *Homo sapiens* PPI test using cellular component, biological process and molecular function GO ontologies.

2.2 Introduction

Gene Ontology (GO)(The Gene Ontology Consortium, 2000) is a useful and popular taxonomy of controlled biological terms that can be used to assess the functional relationship between different gene products. GO organizes knowledge about gene function in a directed acyclic graph (DAG) of terms and their relationships. It is organized in three orthogonal ontologies capturing knowledge about cellular location, biological process and molecular function (The Gene Ontology Consortium, 2000). Experts annotate GO terms to genes in different organisms based on diverse evidence sources. GO has become the most used ontology and annotation system for assessing the confidence and biological relevance of high-throughput experiments based on the notion that if two or more genes are related by an experiment, they should also be related by known gene function. For instance, GO is often used as a benchmark for protein-protein interaction (PPI) experimental mapping and prediction (Li *et al.*, 2008; Patil and Nakamura, 2005; Rhodes *et al.*, 2005; Stelzl *et al.*, 2005), protein function prediction (Jensen *et al.*, 2003; Chen and Xu, 2005; Nariai *et al.*, 2007), and pathway analysis (Shen *et al.*, 2008). In this paper, we are specifically interested in the use of GO as a metric for scoring protein-protein interactions (PPIs).

The relationship between gene products annotated to GO is quantified either simply from the annotated terms (for instance, by finding a set of common GO terms annotated to gene products)

or more globally by using semantic similarity measures that consider the entire GO DAG. The GO DAG is a complex network of over 31,000 terms and 46,900 relations (GO release March, 2010). The cellular component ontology of GO describes gene product locations at the levels of sub-cellular structure and macromolecular complexes through over 2,650 terms and 5,000 relations. The molecular function ontology of GO is described using over 8,650 terms and 10,150 relations. The complexity of biological process ontology is even greater with over 18,500 terms and 38,700 relations. The large number of terms and relations describing the cellular knowledge covered by GO makes it difficult to naively quantify relationships between gene products. For example, *Saccharomyces cerevisiae* proteins RPL10 (annotated to GO cellular component term 'large ribosomal subunit') and SQT1 (annotated to GO cellular component term 'ribosome') physically interact with each other but do not share a GO term. Often, a sub-set of GO terms or a reduced version of GO, like *GO slim* (The Gene Ontology Consortium, 2000), is used for relating genes. This makes GO terms and annotations easier to work with and compare, but valuable information is lost in the simplification.

Semantic similarity is a technique used to measure the likeness of concepts belonging to an ontology. Most early semantic similarity measures (Resnik, 1995; Lin, 1998; Jiang and Conrath, 1997) were developed for linguistic studies in natural language processing. Recently, semantic similarity measurement methods have been applied to and further developed and tailored for biological uses (Schlicker *et al.*, 2006; Wang *et al.*, 2007; Tao *et al.*, 2007; Pesquita *et al.*, 2007). A semantic similarity function returns a numerical value describing the closeness between two (or sometimes more) concepts or terms of a given ontology (Pesquita *et al.*, 2009). In the context of PPI datasets, semantic similarity can be used as an indicator for the plausibility of an interaction because proteins that interact in the cell (*in vivo*) are expected to participate in similar cellular locations and processes. For example, a high semantic similarity value between GO cellular component terms annotated to proteins indicates that proteins are in close proximity and thus have a higher probability of interaction compared to proteins randomly selected from the proteome (Li *et al.*, 2008; Rhodes *et al.*, 2005; Xia *et al.*, 2006). Thus, semantic similarity measures are useful for scoring the confidence of a predicted protein-protein interaction and have the advantage of using the full information stored in the ontology, compared to methods using slim versions of the ontology.

Semantic similarity measures can be broadly classified into two groups: edge based and node based. Edge based methods (Yu *et al.*, 2005; Cheng *et al.*, 2004; Wu *et al.*, 2005; del Pozo *et al.*, 2008) determine semantic similarity based on the shared paths between two terms in a given ontology,

whereas node based methods (Resnik, 1995; Lin, 1998; Jiang and Conrath, 1997) rely on comparing the properties of the input terms (nodes), their ancestors, or descendants. One commonly used property is the specificity, or the information content (entropy), of the common ancestors between a pair of terms, which captures the notion of closeness in the DAG - the more specific the common ancestors of the terms, the closer the terms. The information content of a term c can be defined as the negative log likelihood of the term (eq. 2.1),

$$IC(c) = -\ln p(c) \quad (2.1)$$

where $p(c)$ is the probability of occurrence of the term c in a specific corpus (e.g. GO annotations) (Pesquita *et al.*, 2009). While calculating $p(c)$ in GO the descendants of term c are also considered. For example, the probability of occurrence of the term 'cytosol' in the cellular component hierarchy of GO for *S. cerevisiae* defined by the number of genes assigned to it is 0.104 and its information content is 0.98. Comparative studies to determine the best semantic similarity measurement method have shown that performance on a variety of tests varies greatly depending upon the type of biological datasets used (Lei and Dai, 2006; Guo *et al.*, 2006; Pesquita *et al.*, 2008; Xu *et al.*, 2008). For example, for function prediction Resnik's (Resnik, 1995) and Graph Information Content (simGIC) (Pesquita *et al.*, 2008) work best and for cellular location prediction, the support vector machine (SVM) based method Lei and Dai (2006) is preferable. Guo *et al.* (2006) compared a number of semantic similarity methods (Resnik (Resnik, 1995) (avg), Lin (Lin, 1998), Jiang (Jiang and Conrath, 1997), and graph similarity-based methods (Gentleman *et al.*, 2005)) on a test to distinguish true from false human PPIs. They used proteins within a complex or neighboring each other in Kyoto Encyclopedia of Genes and Genomes (KEGG) regulatory pathways as a positive PPI dataset and randomly chosen protein pairs as a negative interaction dataset. From receiver operating characteristic (ROC) curve performance analysis they concluded that Resnik (avg) is better than other measures at distinguishing positive from negative PPIs. Xu *et al.* (2008) compared the Resnik (Resnik, 1995) (MAX, avg), Tao (Tao *et al.*, 2007), Schlicker (Schlicker *et al.*, 2006; Schlicker and Albrecht, 2008), and Wang (Wang *et al.*, 2007) semantic similarity measurement methods using a similar test with a *S. cerevisiae* PPI dataset from the Database of Interacting Proteins (DIP). They used Schlicker's rfumSimAll method which considers all three ontologies. From ROC analysis they found that the Resnik (MAX) method is best for all the three GO ontologies. Thus, recent independent studies show Resnik's method for calculating semantic similarity is best for measuring

the likelihood of true PPIs.

Resnik’s method defines semantic similarity between two ontology terms s and t for a given set C of common ancestors of s and t as,

$$r(s, t) = \max_{c \in C} [-\ln(p(c))] \quad (2.2)$$

where $p(c)$ is the frequency of proteins annotated to term c and its descendants in the ontology. However, in most cases, proteins are assigned to more than one term in the same GO ontology. Suppose, proteins A and B are annotated to sets of GO terms S and T respectively. the semantic similarity between A and B is defined as the maximum information content (Resnik (MAX)) of the set $S \times T$ (2.3).

$$sim(A, B) = \max_{s_i, t_j \in S, T} r(s_i, t_j) \quad (2.3)$$

or as the average information content (Resnik (avg)) of the set $S \times T$ (2.4).

$$sim(A, B) = \frac{\sum_{s_i, t_j \in S, T} r(s_i, t_j)}{n \times m} \quad (2.4)$$

where s_i and t_j are the GO terms in sets S and T respectively, $r(s_i, t_j)$ is the information content of the lowest common ancestor of terms s_i and t_j , and n and m are the set sizes. Resnik (MAX) has been found to be a better measure of likelihood for PPIs. The use of the MAX function with Resnick’s method to score PPIs, instead of an ‘average’ function, makes sense because proteins in PPIs only need to be in close proximity (similar cellular component terms) or in a similar biological process once, among all possible combinations annotation terms, to be biologically relevant.

Resnik’s measure calculates semantic similarity based only on the information content of a common ancestor. Therefore, it cannot differentiate between any two term pairs with same common ancestor even if they are in different parts of the GO DAG. For example, proteins A and B annotated to the same cellular component term, e.g. ‘cytoplasm’, will have the same semantic similarity value as proteins C and D annotated to different terms, e.g. ‘nucleus’ and ‘mitochondria’, which have ‘cytoplasm’ as a common ancestor. Thus, Resnik’s measure does not consider some of the information contained in the taxonomy by focusing only on the information content of a single ancestor term (Sevilla *et al.*, 2005). Lin’s and Jiang’s measures consider the information content of two terms along with that of a common ancestor but tend to overestimate similarity if the

terms are higher up in the ontology (Sevilla *et al.*, 2005). For example, Lin’s method will assign a score of 1 if two proteins are present in a same general compartment, e.g. ‘cytoplasm’. Similar arguments also hold for molecular function and biological process GO ontologies. Further, the structure of GO is unbalanced with some paths having more details (depth) than others. This could be due to a particular path describing a more complex biological structure or to a particular focus of GO curators as they work to complete the ontology. For example, the ‘intracellular’ term of GO component has more depth than the ‘extracellular’ term (for *S. cerevisiae* GO DAG ‘extracellular’ term has a depth of 0 and ‘intracellular’ has a depth of 7), because there are many more biological terms associated with cell internals versus immediate cell externals. The ideal solution to these problems is to use a balanced GO DAG and annotation, but this is difficult to construct automatically (Alterovitz *et al.*, 2010). Alternatively, we can develop semantic similarity scoring methods that consider the unbalanced nature of GO. In this paper, we have used the successful idea of information content from Resnik (MAX) and introduced clustering of similar GO terms into sub-graphs in a new semantic similarity algorithm, Topological Clustering Semantic Similarity (TCSS), which outperforms Resnik’s method for distinguishing positive from negative protein interactions and other tests.

2.3 Approach

Topological Clustering Semantic Similarity (TCSS) algorithm computes semantic similarity between GO terms annotated to proteins in interaction datasets. TCSS considers the different levels of biological knowledge representation depth in different branches of the GO graph. TCSS clusters similar GO terms in a given ontology and creates a hierarchical graph structure with proteins belonging to the same sub-graph scored higher as compared to proteins belonging to different sub-graphs.

2.4 Methods

2.4.1 Algorithm

The goal of TCSS is to find subsets of GO terms defining similar concepts (e.g. nucleus related terms vs. mitochondrion related terms) and score gene products belonging to a similar subset higher than if they belong to different sets. In an effort to normalize the depth of the GO DAG across

the ontology, the algorithm first defines mutually exclusive (non-overlapping) sub-graphs (sets of connected GO terms) rooted at major nodes. These sub-graphs are collapsed as single nodes to form a meta-graph and a two-level semantic similarity calculation is performed, as described below.

Topology based clustering

To normalize the depth of terms across the GO DAG, semantic similarity between terms, s and t , is calculated within a sub-graph instead of the complete GO graph. Sub-graphs consist of terms defining related concepts (e.g. all terms relating to the 'nucleus') and are defined based on a threshold on the information content of all terms present in a given ontology. The topological information content (ICT) of a term depends upon its specificity in the graph and is defined as shown in equation (2.5)

$$ICT(t) = -\ln\left(\frac{|\text{descendants of } t|}{|\text{total terms in ontology}|}\right) \quad (2.5)$$

where t is a term in the ontology (Zhang *et al.*, 2006). The terms which are more specific (i.e. terms which are present in the lower levels, closer to the leaves in the ontology graph) will have high information content as compared to less specific ones (i.e. terms which are present in the upper levels of the ontology graph closer to the root). An ICT cutoff (referred to as the 'topology cutoff') is defined in a pre-calculation step (see Implementation details). All terms with ICT values below the topology cutoff are treated as nodes of the meta-graph. Terms are removed from meta-graph if their ICT values are within 20% of their parent ICT values. This is done to increase the dissimilarity between meta-graph nodes. For each node in the meta-graph a sub-graph is created from all descendants terms of that node.

GO terms often have multiple parents, which could result in overlapping sub-graphs (a term is present in two sub-graphs). Each GO term in the cellular component ontology has on average 1.9 edges, whereas the ratio is 2.1 for the biological process ontology, and 1.2 for the molecular function ontology. All relationships (or edges) are considered and treated equally. Sub-graph overlap is removed in two steps (Figure 2.1):

- *Edge removal by transitive reduction.* The GO DAG gives rise to partial orders \leq on its vertices, where $u \leq v$ when there exists a directed edge from u to v . However, u and v could connect via many different GO DAG paths. For example, the GO graph with paths

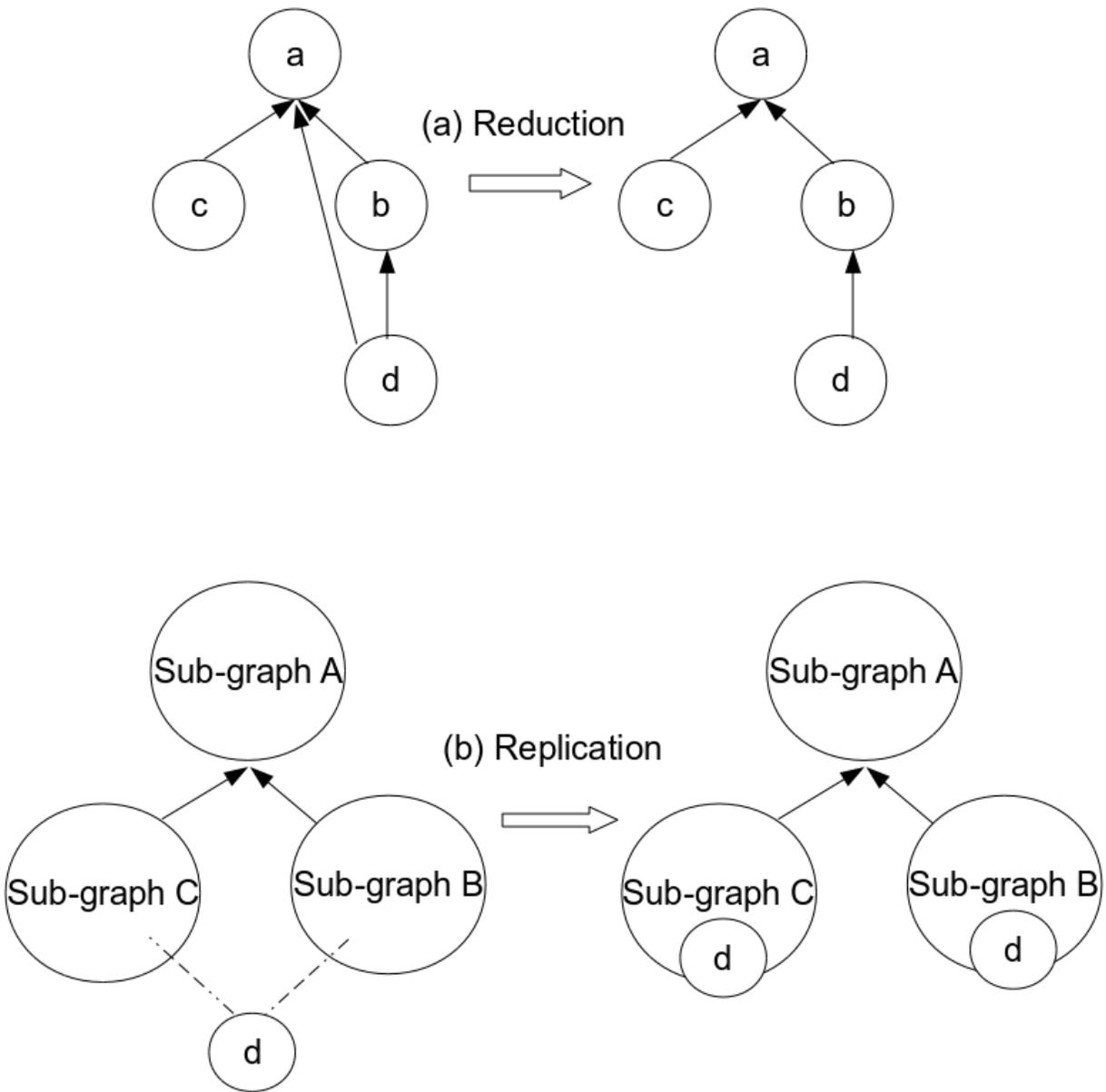


Figure 2.1: Mutually exclusive sub-graphs. (a) Transitive reduction - suppose a , b , c , and d are the nodes in graph G with directed edges as shown in gure (a). Let the number of genes annotated to each node is 1. Then the total annotation of node a in G (annotation of a and its descendants) is 4. Transitive reduction of G will result in G' without edge $d \rightarrow a$ and with same total annotation of 4 as a can still be reached from d . (b) Replication - suppose term d is common to both the sub-graphs B and C then term d will be copied to both the sub-graphs.

$a \rightarrow b \rightarrow c$ and $a \rightarrow c$ has the same reachability as the GO graph with relationships $a \rightarrow b \rightarrow c$. Thus, the transitive reduction of GO graph G results in the smallest graph $R(G)$ such that, the transitive closure of G is same as the transitive closure of $R(G)$. This results in 14% and 6% fewer edges in cellular component and biological process ontologies respectively, reducing the likelihood of sub-graph overlap. There was no significant reduction in the molecular function ontology.

- *Term duplication.* After the reduction step, if a term still belongs to more than one sub-graph then it and its descendants are replicated in each sub-graph. Such a situation arises with a term having disjunctive ancestors (having independent paths from the ancestors to the term) belonging to different sub-graphs (Couto *et al.*, 2005).

Finally, all sub-graphs are connected into a hierarchy based upon the position of their root terms in original graph to construct a meta-graph (Figure 2.2). Meta nodes representing sub-graphs are labeled using the GO term of their sub-graph root.

Normalized scoring

Notation: G^m and G^s denote the meta and sub graphs. G_i^s is the i^{th} sub-graph. t_i is the i^{th} term belonging to either the meta or sub-graph. ICS and ICM are the normalized information content values of a term t (denoted by ICA) in the sub and meta graphs, respectively. LCA is the lowest common ancestor (or the ancestor with maximum information content) of any two given terms.

We developed a semantic similarity scoring system on the constructed meta-graph that results in more balanced semantic similarity scores compared to scoring the GO DAG directly. The system scores protein pairs in the same sub-graph higher than if they belong to different sub-graphs. The annotation information content (ICA) of all the terms present in an ontology is calculated based on the frequency of gene products annotated to a term and its children is shown in equation (2.6).

$$ICA(t) = -\ln \left(\frac{annotation(t)}{|\text{total gene products annotated to the ontology}|} \right) \quad (2.6)$$

$$annotation(t) = |(\text{gene products} \in t) \cup \bigcup_{c \in \text{descendants}(t)} (\text{gene products} \in c)|$$

where t is a term in the ontology. The annotation information content values lie in the range $[0, \infty)$

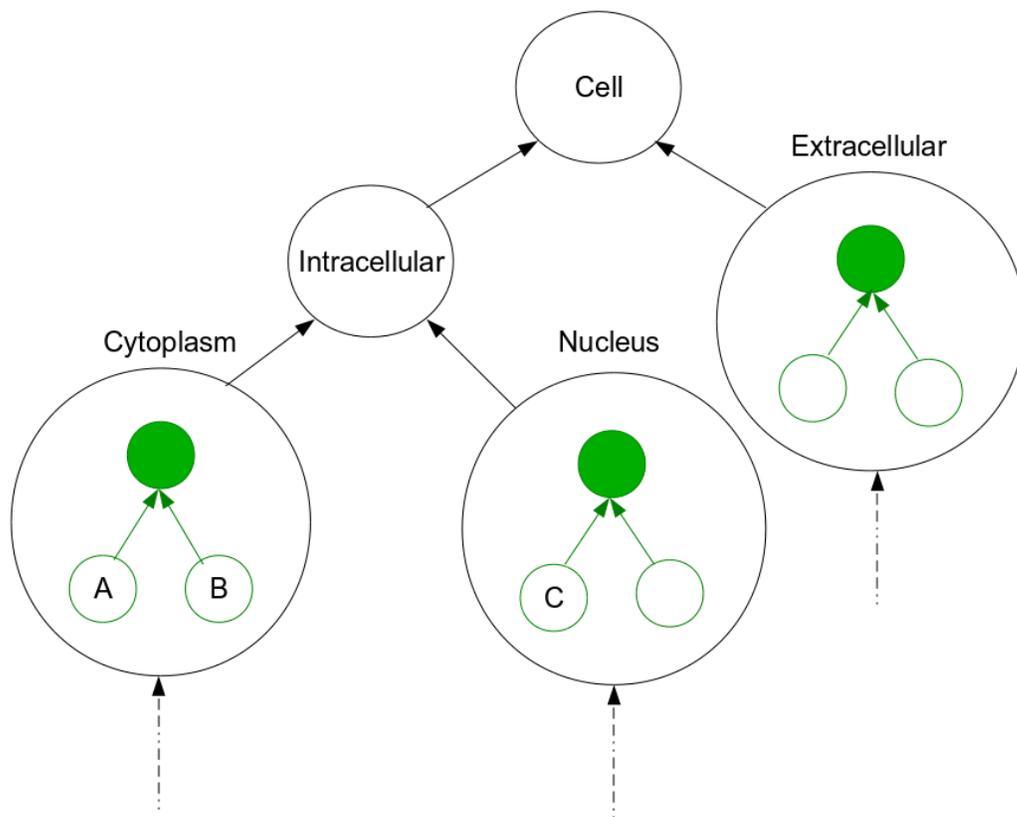


Figure 2.2: Graphical illustration of the algorithm. Nodes in the higher level graph and sub-graphs are shown by black and green circles, respectively. Root nodes of sub-graphs are shown by solid green circles and are equivalent to the corresponding higher level node. Terms A and B belong to the same sub-graph, therefore the semantic similarity score between them will be computed based on their common ancestor term 'Cytoplasm' (solid green). Terms B and C belong to different sub-graphs, therefore their semantic similarity score will be computed based on the common ancestor term 'Intracellular'.

and are normalized to $[0, 1]$ by dividing with the maximum information content in a sub-graph or meta-graph. For a term t_i belonging to the i^{th} sub-graph G_i^s , the sub-graph information content (ICS) of t_i is defined as shown in equation (2.7).

$$ICS(t_i) = \frac{ICA(t_i)}{\max_{t_i \in G_i^s} ICA(t_i)} \quad (2.7)$$

For a term t_i in meta-graph G^m , the information content (ICM) of t_i is calculated as shown in equation (2.8).

$$ICM(t_i) = \frac{ICA(t_i)}{\max_{t_i \in G^m} ICA(t_i)} \quad (2.8)$$

It is possible that gene products A and B are annotated to more than one GO term. Let, S and T be the sets of GO terms annotated to gene products A and B , respectively. Then the semantic similarity between gene products A and B is defined by the maximum approach, as shown in equation (2.9).

$$Sim_{max}(A, B) = \max_{s_i, t_j \in S, T} \begin{cases} ICM(LCA(s_i, t_j)), & \text{if } s_i \in G_i^s \text{ and } t_j \in G_j^s \\ ICS(LCA(s_i, t_j)), & \text{if } s_i, t_j \in G_i^s \end{cases} \quad (2.9)$$

where $LCA(s_i, t_j)$ is the lowest common ancestor (or the common ancestor with maximum information content) of the terms s_i and t_j . If both the terms s_i and t_j belong to the same sub-graph then their lowest common ancestor will be in that sub-graph otherwise it will belong to the meta-graph.

Best-match average approach

Let, S and T be the sets of GO terms annotated to gene products A and B respectively. Then semantic similarity between gene products A and B based upon the best-match average approach (Azuafe *et al.*, 2006; Wang *et al.*, 2007) is defined by the equation (2.10).

$$Sim_{bma}(A, B) = \frac{\sum_{s_i \in S} Sim(s_i, T) + \sum_{t_j \in T} Sim(t_j, S)}{|S| \times |T|} \quad (2.10)$$

where $Sim(u_i, V)$ is defined as (2.11),

$$Sim(u_i, V) = \max_{v_j \in V} \begin{cases} ICM(LCA(u_i, v_j)), & \text{if } u_i \in G_i^s \text{ and } v_j \in G_j^s \\ ICS(LCA(u_i, v_j)), & \text{if } u_i, v_j \in G_i^s \end{cases} \quad (2.11)$$

where u_i is a term annotated to a gene product and V is the set of terms annotated to the other gene product.

2.5 Results

2.5.1 Data acquisition and processing

- **Ontology data:** Ontology data was downloaded from the Gene Ontology database (The Gene Ontology Consortium, 2000) (dated March 2010) containing 31,382 ontology terms subdivided into 2,689 cellular component, 18,545 biological process and 8,688 molecular function terms.
- **GO Annotation data:** Gene annotations for GO terms were downloaded from the Gene Ontology database for *S. cerevisiae* (dated February 2010) (Christie *et al.*, 2004) and *H. Sapiens* (dated August 2010) (Consortium *et al.*, 2010). Electronically inferred annotations (IEA) lack manual review therefore, we designed two sets of tests one with IEA annotations and one without. In our implementation, we only consider the most specific GO gene annotations. For example, if gene A is annotated to terms X and Y (and X is an ancestor of Y), then we only consider annotation to Y. This is because in ontologies a term is an aggregate of its descendants. This pre-filtering of GO could impact the results of some methods used in our analysis. For instance, in CESSM tests, correlation between SimGIC semantic similarity and EC similarity for the molecular function ontology increases by 25% and correlation with sequence similarity decreases by 15% if all the annotations are considered, however all other changes we noticed were minor and didn't change our results.
- **Interaction dataset:** To evaluate the performance of TCSS against other semantic similarity measures on the problem of scoring PPI confidence we created positive and negative interaction datasets for *S. cerevisiae* and *H. sapiens*.
 - *S. cerevisiae*: We retrieved 4,598 unique pairwise *S. cerevisiae* PPIs from the core set of Database of Interacting Proteins (DIP) (dated December 2009) (Salwinski *et al.*, 2004). The DIP core database records data derived from both small-scale and large-scale exper-

iments that have been validated by the occurrence of the interaction between paralogous proteins in different species (Salwinski *et al.*, 2004). The positive dataset for CC, BP, and MF ontologies comprised interactions with both proteins annotated to terms (other than root) in their respective ontologies (Table 2.1). The negative dataset with the same number of PPIs as the positive set was generated by randomly selecting proteins from genes in the GO annotation files that are not known to be positive in a set of all known (45,448) yeast PPIs from iRefWeb (September 2010), a metadatabase containing the ten largest primary PPI databases (Razick *et al.*, 2008).

- *H. sapiens*: We retrieved 2077 unique pairwise PPIs (with three or more publications) for *H. sapiens* from DIP (dated June 2010). The positive dataset for CC, BP, and MF ontologies comprised interactions with both proteins annotated to terms (other than root) in their respective ontologies (Table 2.1). The negative interaction dataset contained an equal number of randomly selected interactions from a pool of all possible interactions in human minus all known (43,935) iRefWeb (Razick *et al.*, 2008) known PPIs.
- **Gene expression datasets:** The gene expression dataset for *S. cerevisiae* was downloaded from GeneMANIA (Warde-Farley *et al.*, 2010) (dated August 2010) and contained data from 39 different microarray experiments. Test datasets were prepared from 5000 randomly picked *S. cerevisiae* gene pairs randomly selected from a list of all possible pairs of proteins in our gene expression data set, including an equal number of random and known PPIs (PPIs in the DIP core set have higher than average expression correlation). This was done independently for CC, BP, and MF annotations of GO (including IEA annotations).
- **CESSM dataset:** Collaborative Evaluation of GO-based Semantic Similarity Measures (CESSM) is an online tool for the automated evaluation of GO-based semantic similarity measures in terms of performance against sequence, Pfam (protein family) and EC (enzyme commission number) similarity (Pesquita *et al.*, 2008). Protein pair (from multiple species), GO (dated August 2010), and UniProt GO annotations (dated August 2008) were downloaded from CESSM.

2.5.2 Model evaluation

In the previous section we presented a new algorithm, Topological Clustering Semantic Similarity (TCSS), to compute semantic similarity between GO terms annotated to proteins that normalizes

	<i>S. cerevisiae</i>		<i>H. sapiens</i>	
	DIP (core)		DIP (core)	
	IEA+	IEA-	IEA+	IEA-
CC	4469	4425	1431	1054
BP	4385	4326	1435	1204
MF	3858	3583	1441	1288

Table 2.1: Distribution of positive and negative interactions. Number of interactions in the positive dataset for cellular component (CC), biological process (BP), and molecular function (MF) ontologies.

GO DAG branch depth. We compared the performance of TCSS with other semantic similarity measures given by Resnik (Resnik, 1995), Lin (Lin, 1998), Wang (Wang *et al.*, 2007), Schlicker (simRel method) (Schlicker *et al.*, 2006), Jiang (Jiang and Conrath, 1997), Pesquita (SimGIC) (Pesquita *et al.*, 2007) on the problem of scoring PPIs. Performance analysis of TCSS was done using receiver operating characteristic (ROC) and F_1 measures. ROC grades the performance of classifiers as a trade-off between true positive rate (TPR) and false positive rate (FPR). We also used the F_1 measure, which is the harmonic mean of precision (the proportion of retrieved information that is actually relevant) and recall (the proportion of relevant information that is retrieved) and indicates the classifier’s ability to retrieve relevant information. The evaluation was done separately for cellular component (CC), biological process (BP), and molecular function (MF) ontologies.

***Saccharomyces cerevisiae* PPI test**

S. cerevisiae positive and negative protein interaction sets were used to evaluate the above mentioned semantic similarity measures for their ability to distinguish positives from negatives. TCSS, Resnik, Lin, Jiang and Schlicker were tested using both the maximum (MAX) and best-match average (BMA) approach of combining multiple GO gene annotations and Wang was tested using only the BMA approach, as only BMA was used in the original Wang publication and is the only option available in the author’s implementation. BMA averages scores when multiple combinations of GO terms are possible (for gene products annotated with multiple terms). SimGIC considers multiple GO annotations while calculating semantic similarity scores, thus MAX and BMA methods are not relevant for it. We focused initial tests on manually annotated GO annotations (“without” annotations with IEA evidence codes (IEA-)), but also tested with all annotations, including electronic annotations (“with” annotations with IEA evidence codes (IEA+)).

TCSS and Resnik consistently showed the best performance for all three ontologies in ROC

		IEA-			IEA+		
		CC	BP	MF	CC	BP	MF
TCSS	max	0.83	0.89	0.73	0.83	0.89	0.75
	bma	0.82	0.88	0.72	0.83	0.88	0.74
Resnik	max	0.83	0.89	0.73	0.83	0.89	0.75
	bma	0.81	0.87	0.72	0.83	0.88	0.74
Lin	max	0.80	0.87	0.70	0.79	0.87	0.72
	bma	0.79	0.85	0.68	0.80	0.86	0.72
Jiang	max	0.75	0.85	0.72	0.73	0.85	0.73
	bma	0.73	0.84	0.70	0.72	0.84	0.73
Schlicker	max	0.70	0.81	0.65	0.70	0.81	0.67
	bma	0.69	0.82	0.64	0.71	0.82	0.68
SimGIC		0.73	0.75	0.64	0.73	0.76	0.68
Wang		0.74	0.83	0.72	0.76	0.82	0.73

Table 2.2: Area under ROC curves for the *S. cerevisiae* PPI dataset. Tests were performed separately for cellular component (CC), biological process (BP), and molecular function (MF) ontologies. *Best-match average* and *maximum* approaches were used for datasets with (IEA+) and without (IEA-) electronic annotations. The best ROC scores are in bold.

Best-match average						Maximum					
IEA-			IEA+			IEA-			IEA+		
CC	BP	MF	CC	BP	MF	CC	BP	MF	CC	BP	MF
7.36	6.66	1.36	3.0	6.0	2.66	8.53	5.54	1.53	5.74	5.51	1.83

Table 2.3: Improvement in F_1 score for the *S. cerevisiae* PPI dataset. Average improvement in F_1 scores achieved by TCSS over Resnik for best-match average and maximum approaches. TCSS does 6 times better than Resnik for cellular component (CC), 5.9 times for biological process (BP), and 1.9 times for molecular function (MF) ontologies on average.

analysis under different conditions (Table 2.2, Figures 2.3 (MAX, IEA-), 2.5 (BMA, IEA-), 2.7 (MAX, BMA, IEA+)). Since it is not clear from ROC analysis which of TCSS and Resnik performs better, we compared their F_1 scores at different semantic similarity cutoffs for all the three ontologies (Figures 2.4 (MAX, IEA-), 2.6 (BMA, IEA+), 2.8 (MAX, BMA, IEA+)). TCSS showed average improvements of 6 times for CC, 5.9 times for BP, and 1.9 times for MF in retrieving relevant information over Resnik (Table 2.3) mainly due to the faster increase in true positive rate for TCSS at a given score threshold.

Homo sapiens PPI test

To test the generality of the method for PPI scoring, we ran similar tests as above using a *H. sapiens* PPI data set. *H. sapiens* positive and negative protein interaction sets were used to evaluate TCSS, Resnik, Lin, Jiang, Schlicker and SimGIC methods. The evaluation was done using BMA and MAX

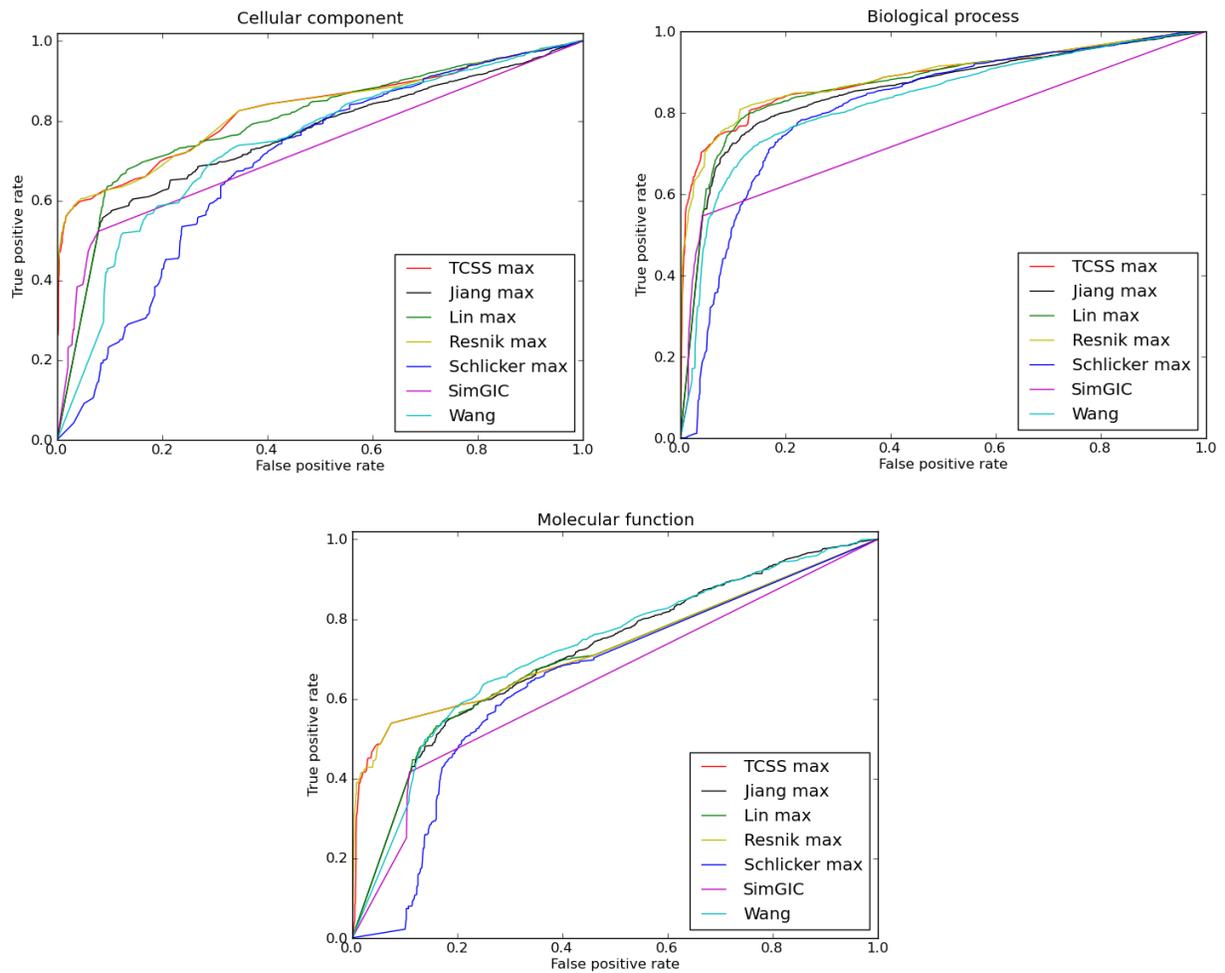


Figure 2.3: ROC curves for *S. cerevisiae* PPI dataset. ROC evaluations of semantic similarity measures at different cutoffs based on the *S. cerevisiae* PPI dataset derived from DIP are shown. The evaluation was performed using cellular component, biological process, molecular function ontologies of GO. Maximum (max) approach for combining multiple annotations was used on dataset without (IEA-) electronic annotations. TCSS and Resnik show the best ROC profiles for all three ontologies.

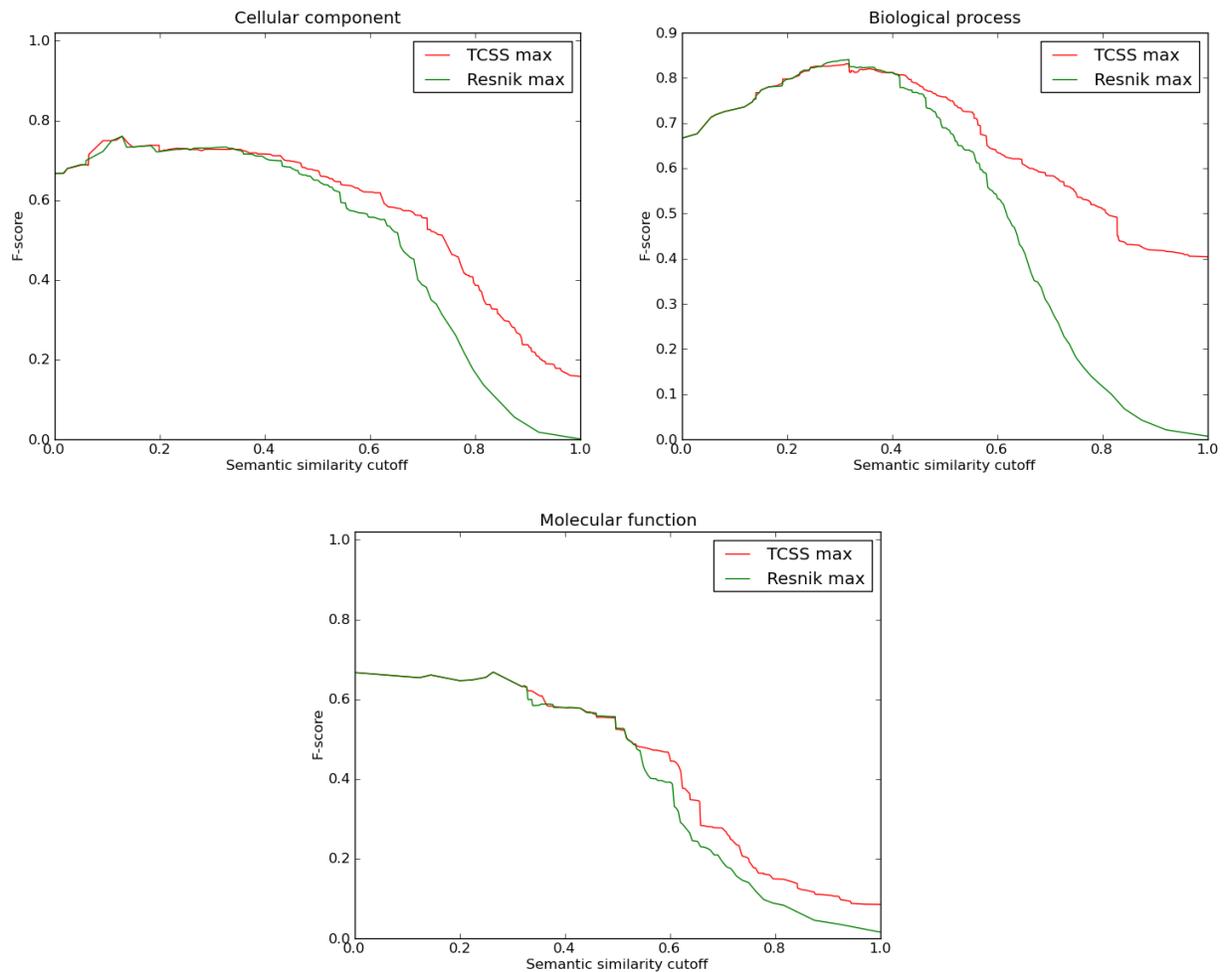


Figure 2.4: F-score curves for *S. cerevisiae* PPI dataset. F_1 score (harmonic mean of precision and recall) evaluations of semantic similarity measures at different cutoffs based on the *S. cerevisiae* PPI dataset derived from DIP are shown. The evaluation was performed using cellular component, biological process, and molecular function ontologies of GO. Maximum (max) approach for combining multiple annotations was used on a dataset with only manual annotations (no electronic annotations (IEA-)). F_1 score reaches its best value at 1 and worst at 0. TCSS does better than Resnik for semantic similarity cutoff scores in all three ontologies.

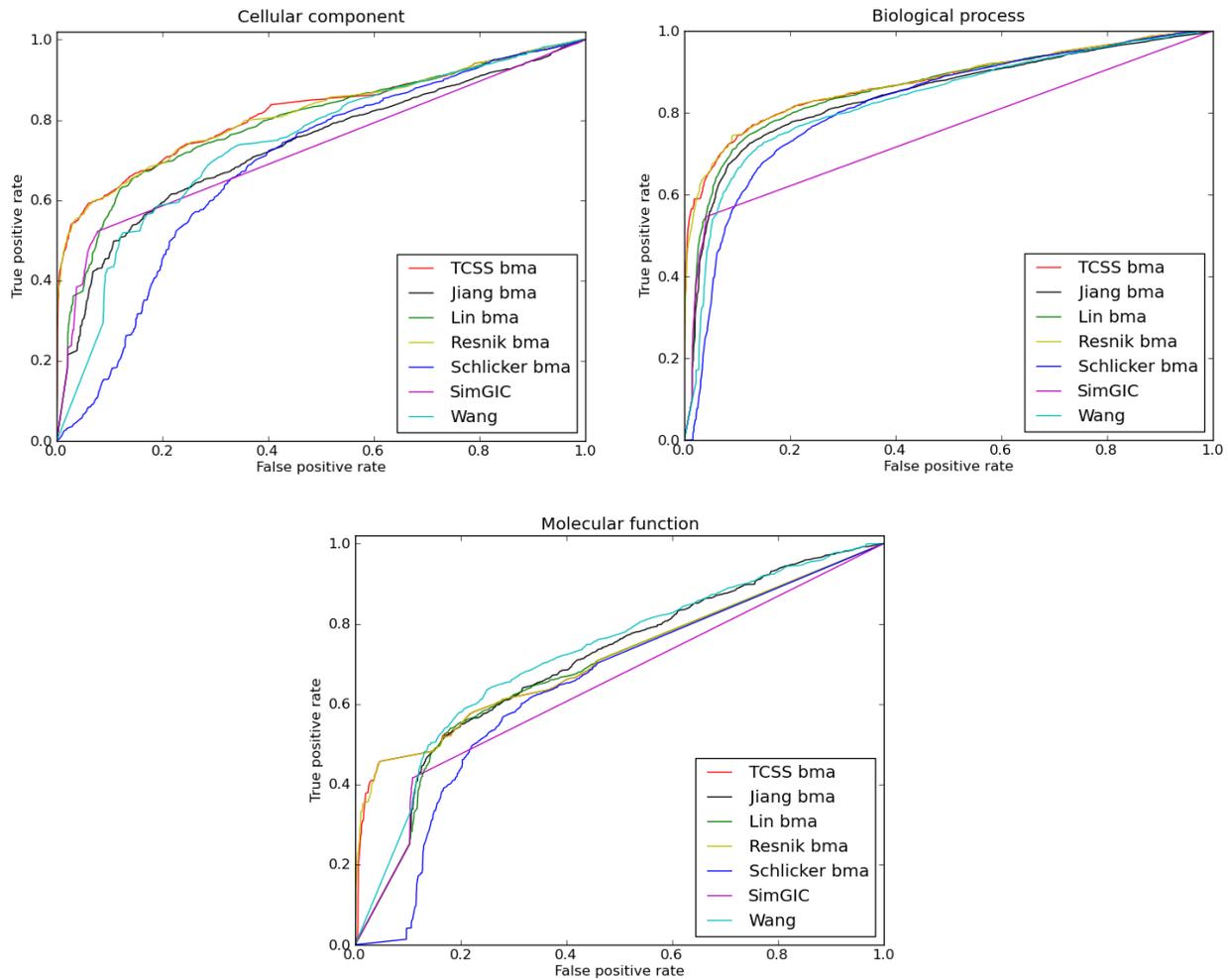


Figure 2.5: ROC curves for *S. cerevisiae* PPI dataset (IEA-). ROC evaluations of semantic similarity measures at different cutoffs based on the *S. cerevisiae* PPI dataset derived from DIP are shown. The evaluation was performed using cellular component, biological process, molecular function ontology of GO. Maximum (max) approach for combining multiple annotations was used on dataset without (IEA-) electronic annotations. TCSS and Resnik show the best ROC profiles for all three ontologies.

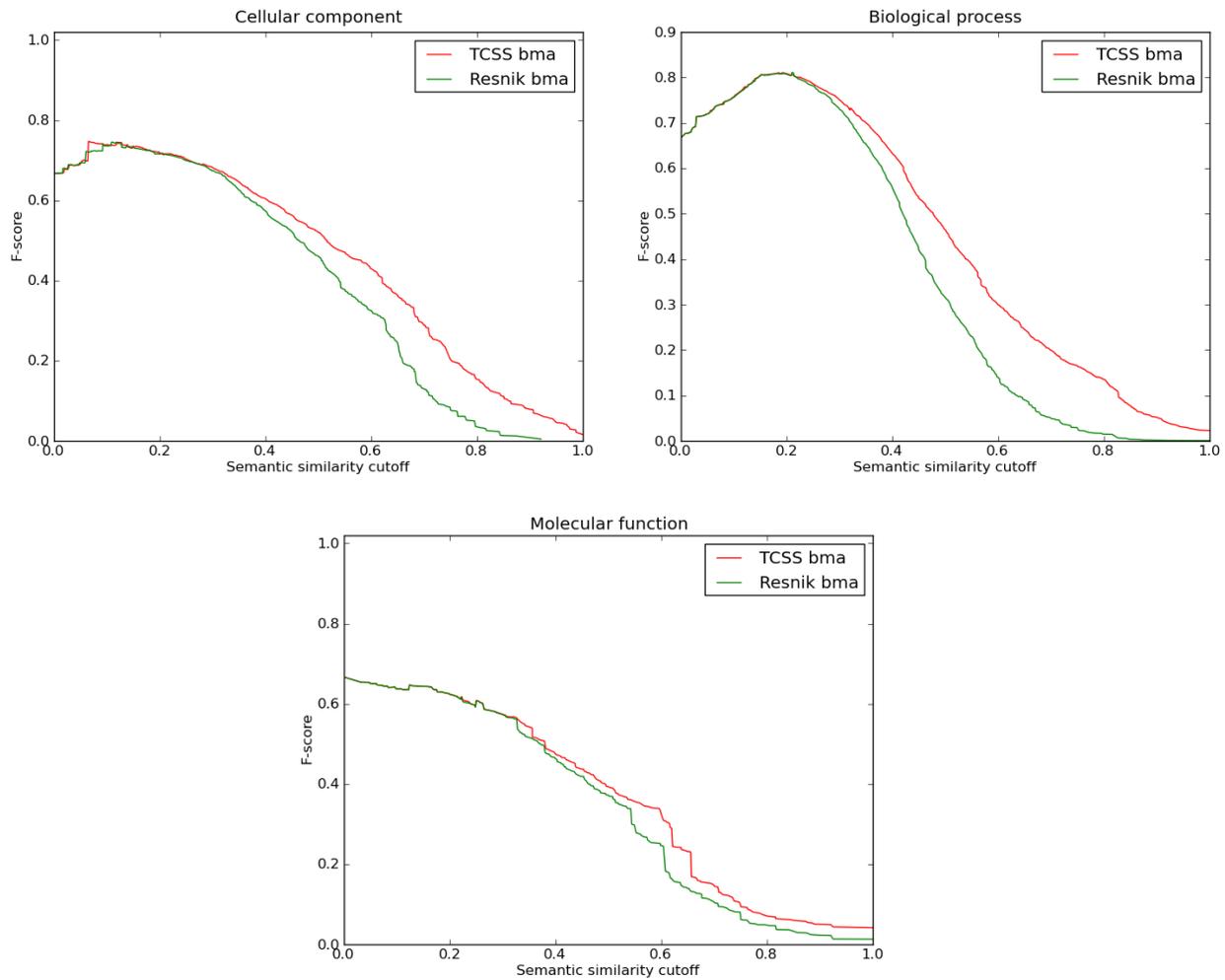


Figure 2.6: F_1 -score curves for *S. cerevisiae* PPI dataset (IEA-). F_1 score (harmonic mean of precision and recall) evaluations of TCSS and Resnik semantic similarity measures at different cutoffs based on the *S. cerevisiae* PPI dataset derived from DIP are shown. The evaluation was performed using cellular component, biological process, molecular function ontology of GO. Best-match average (bma) approach for combining multiple annotations was used on dataset without (IEA-) electronic annotations. F_1 score reaches its best value at 1 and worst at 0. TCSS does better than Resnik for semantic similarity cutoff scores for all three ontologies.

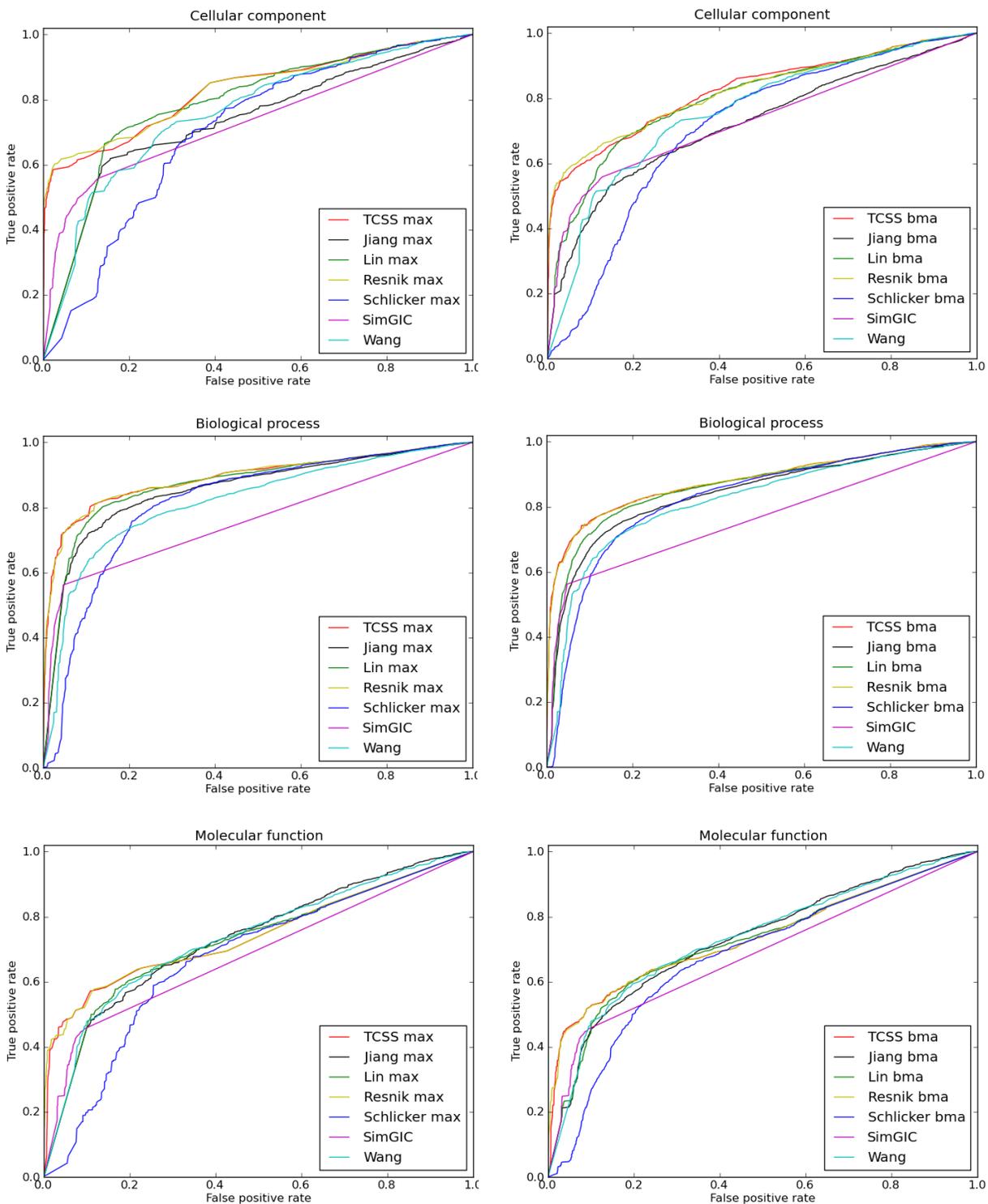


Figure 2.7: ROC curves for *S. cerevisiae* PPI dataset (IEA+). ROC evaluations of semantic similarity measures at different cutoffs based on the *S. cerevisiae* PPI dataset derived from DIP are shown. The evaluation was performed using cellular component, biological process, molecular function ontology of GO. Best-match average (bma) and maximum (max) approaches for combining multiple annotations are used on dataset with (IEA+) electronic annotations. TCSS & Resnik show best ROC profiles for all three ontologies.

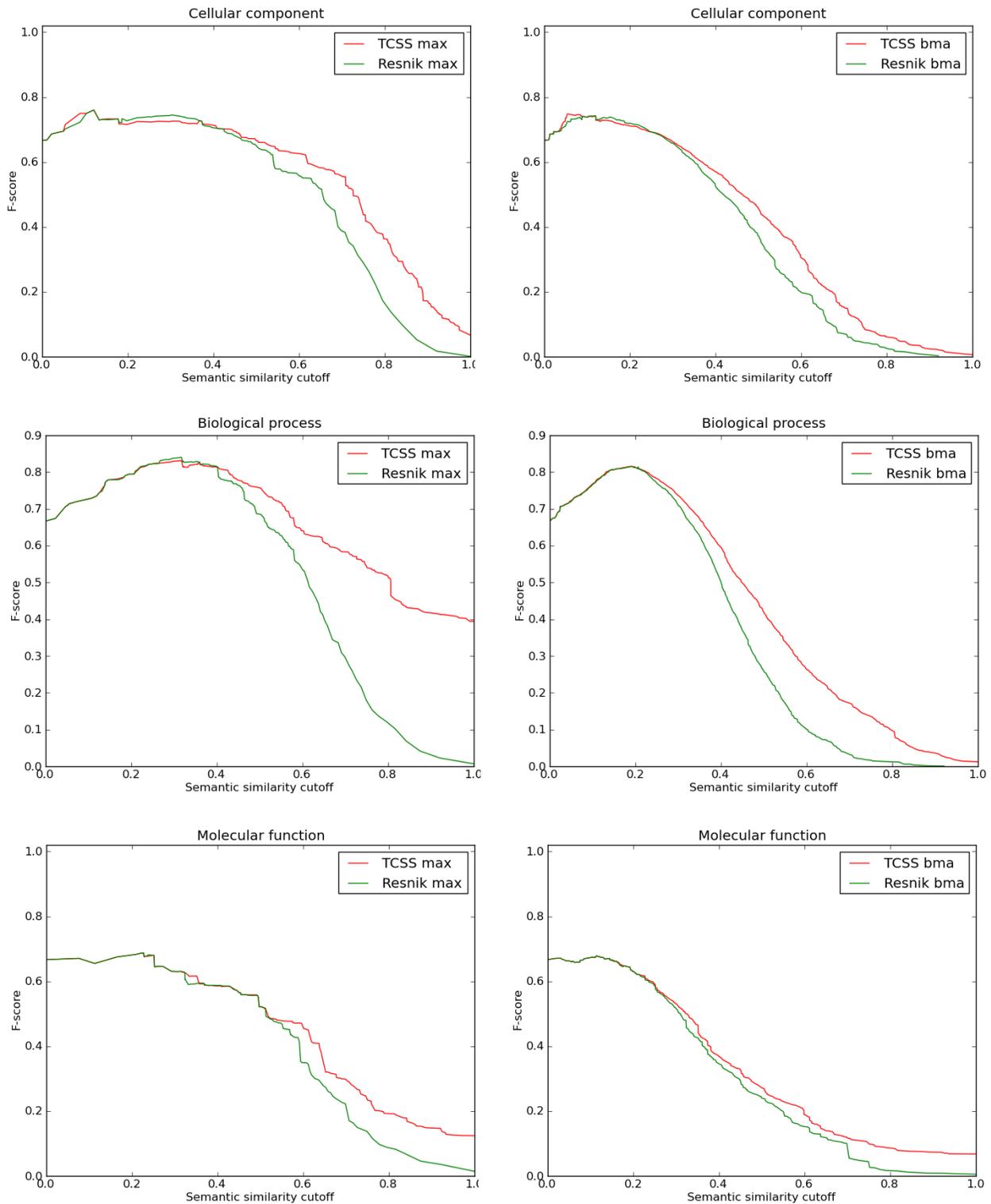


Figure 2.8: F_1 -score curves for *S. cerevisiae* PPI dataset (IEA+). F_1 score (harmonic mean of precision and recall) evaluations of TCSS and Resnik semantic similarity measures at different cutoffs based on the *S. cerevisiae* PPI dataset derived from DIP are shown. The evaluation was performed using cellular component, biological process, molecular function ontology of GO. Best-match average (bma) and maximum (max) approaches for combining multiple annotations was used on dataset with (IEA+) electronic annotations. F_1 score reaches its best value at 1 and worst at 0. TCSS does better than Resnik for semantic similarity cutoff scores in all three ontologies.

Best-match average						Maximum					
IEA-			IEA+			IEA-			IEA+		
CC	BP	MF	CC	BP	MF	CC	BP	MF	CC	BP	MF
3.44	1.51	2.42	1.28	1.64	4.0	2.7	1.48	2.0	1.53	1.58	1.50

Table 2.4: Improvement in F_1 score for *H. sapiens* PPI dataset. Average improvement in F_1 scores achieved by TCSS over Resnik for maximum and best-match average approaches. TCSS does 2.2 times better than Resnik for cellular component (CC), 1.5 times for biological process (BP), and 2.5 times for molecular function (MF) ontologies on average.

approaches for combining multiple GO annotations on IEA+/- datasets (Additional file 1: Supp. figs. 6-9, Supp. tab. 1). Table 2.4 shows the improvement in F_1 scores achieved by TCSS over Resnik. On average TCSS performed 2.2 times better than Resnik for CC, 1.5 times for BP, and 2.5 times for MF ontologies.

Correlation with gene expression

To test how our method performs in another application scenario, we tested its correlation with gene expression data. Two gene products that have similar function are more likely to have similar expression profiles and be annotated to similar GO terms (Sevilla *et al.*, 2005). Therefore, a comparison of the similarity between gene expression of two gene products with the semantic similarity scores obtained by different measures can be used as a performance test. Gene expression profiles of randomly selected *S. cerevisiae* gene pairs were evaluated against the above mentioned semantic similarity methods. The evaluation was performed as above using the BMA/MAX approaches of combining multiple GO annotations on IEA+ dataset. TCSS showed the best correlation between gene expression and semantic similarity with all three GO ontologies (Figures 2.144(a), 2.13).

Correlation with EC, Pfam, and sequence similarity

The Collaborative Evaluation of GO-based Semantic Similarity Measures (CESSM) website was developed by Pesquita *et al.* (2008) to evaluate semantic similarity measures on a standard set of data and benchmarks: correlation of similarity measure with similarity of sequence, Pfam domains and Enzyme Commission (EC) numbers. We compared TCSS against Resnik, Schlicker, Jiang, Lin and SimGIC using CESSM for both MAX and BMA approaches on IEA- dataset. TCSS showed the best (or one of the best) correlation with EC similarity for all three ontologies (Figure 2.14(b), Additional file 1: Supp. fig. 17). For Pfam similarity with MAX approach, TCSS is best for CC and MF ontologies and SimGIC showed better correlation than TCSS for BP ontology

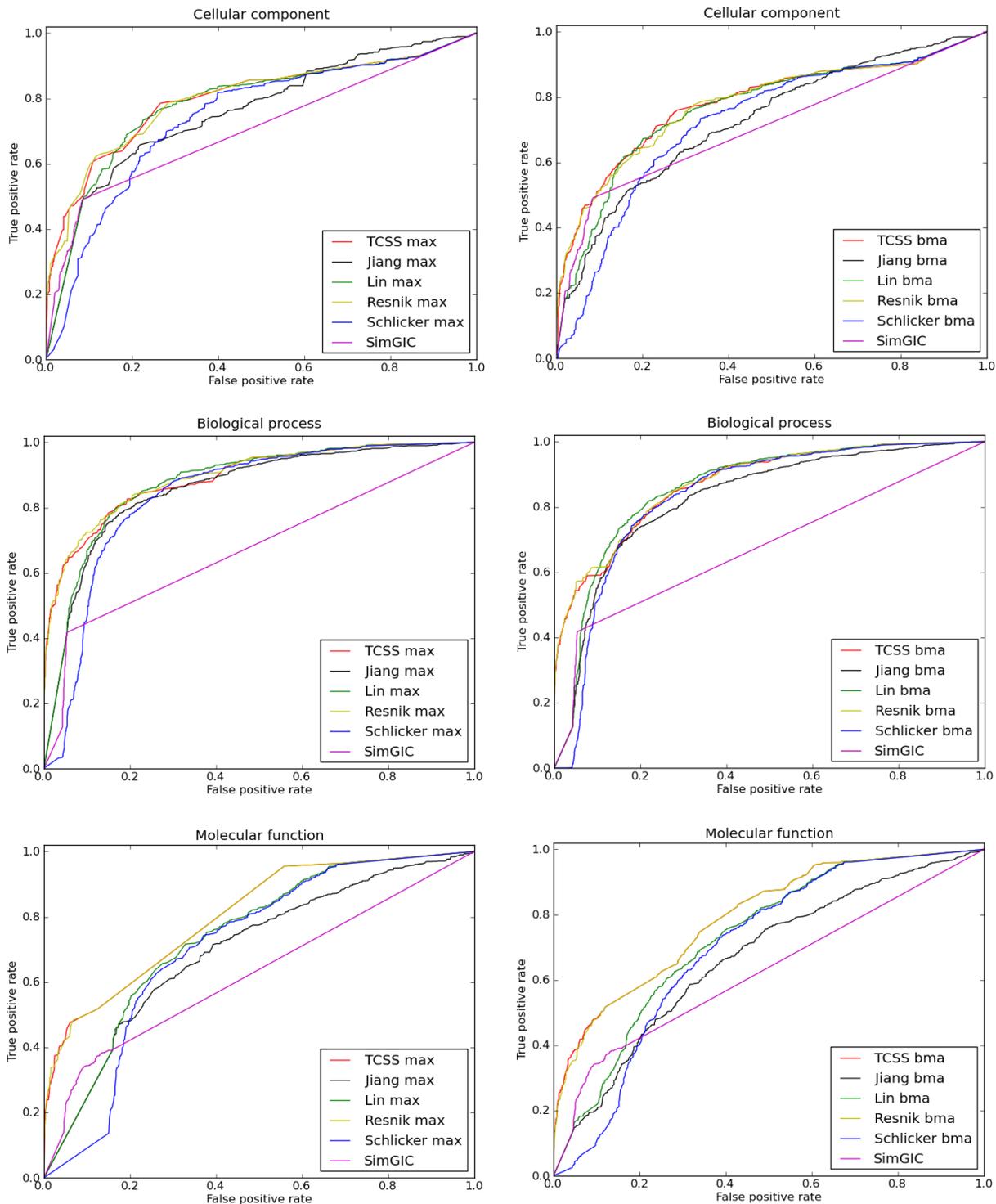


Figure 2.9: ROC curves for *H. sapiens* PPI dataset (IEA-). ROC evaluations of semantic similarity measures at different cutoffs based on the *H. sapiens* PPI dataset derived from DIP are shown. The evaluation was performed using cellular component, biological process, molecular function ontology of GO. Maximum (max) approach for combining multiple annotations was used on dataset without (IEA-) electronic annotations. TCSS and Resnik show the best ROC profiles for all three ontologies.

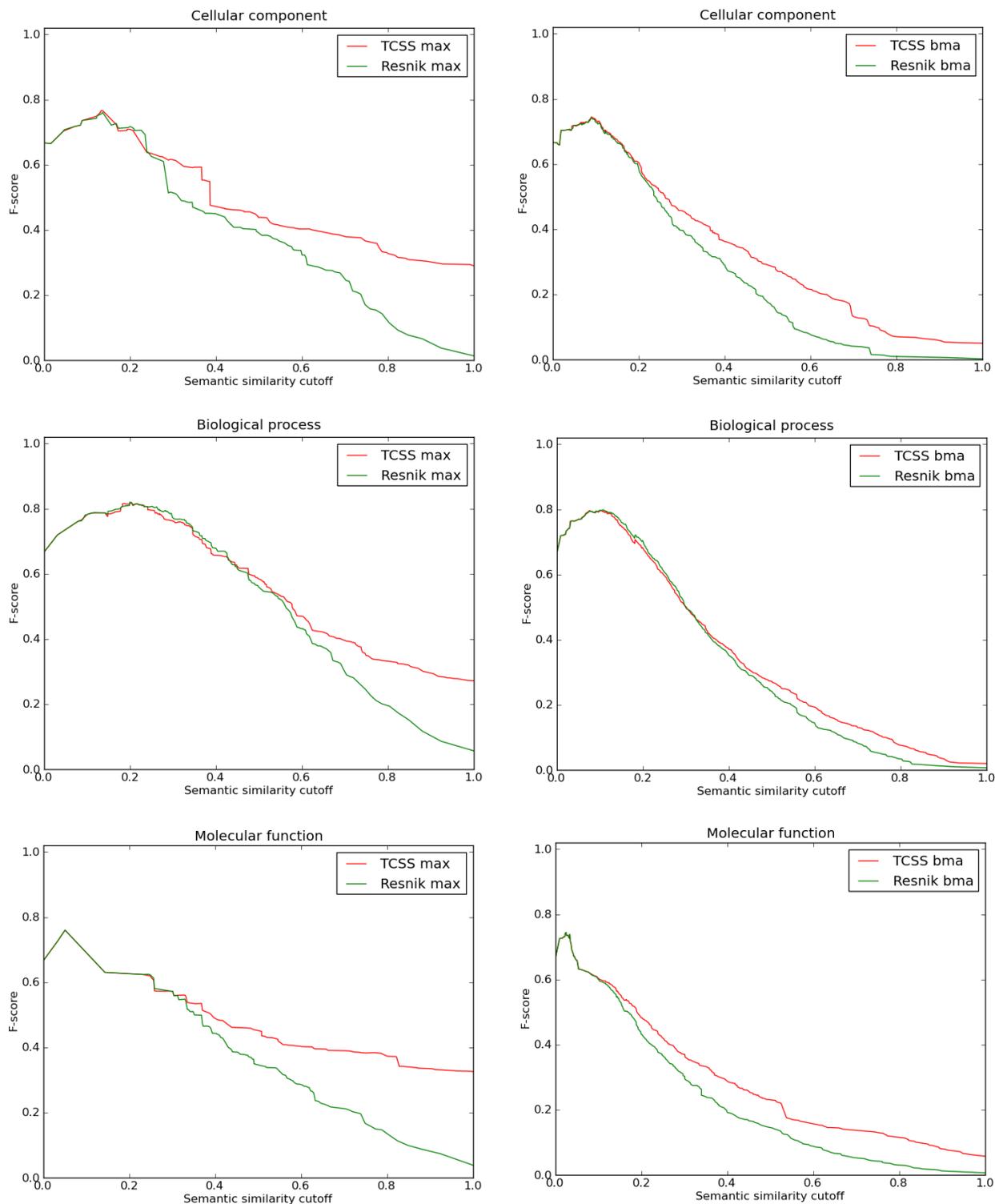


Figure 2.10: F_1 -score curves for *H. sapiens* PPI dataset (IEA-). F_1 score (harmonic mean of precision and recall) evaluations of TCSS and Resnik semantic similarity measures at different cutoffs based on the *H. sapiens* PPI dataset derived from DIP are shown. The evaluation was performed using cellular component, biological process, molecular function ontology of GO. Best-match average (bma) approach for combining multiple annotations was used on dataset without (IEA-) electronic annotations. F_1 score reaches its best value at 1 and worst at 0. TCSS does better than Resnik for semantic similarity cutoff scores in all three ontologies.

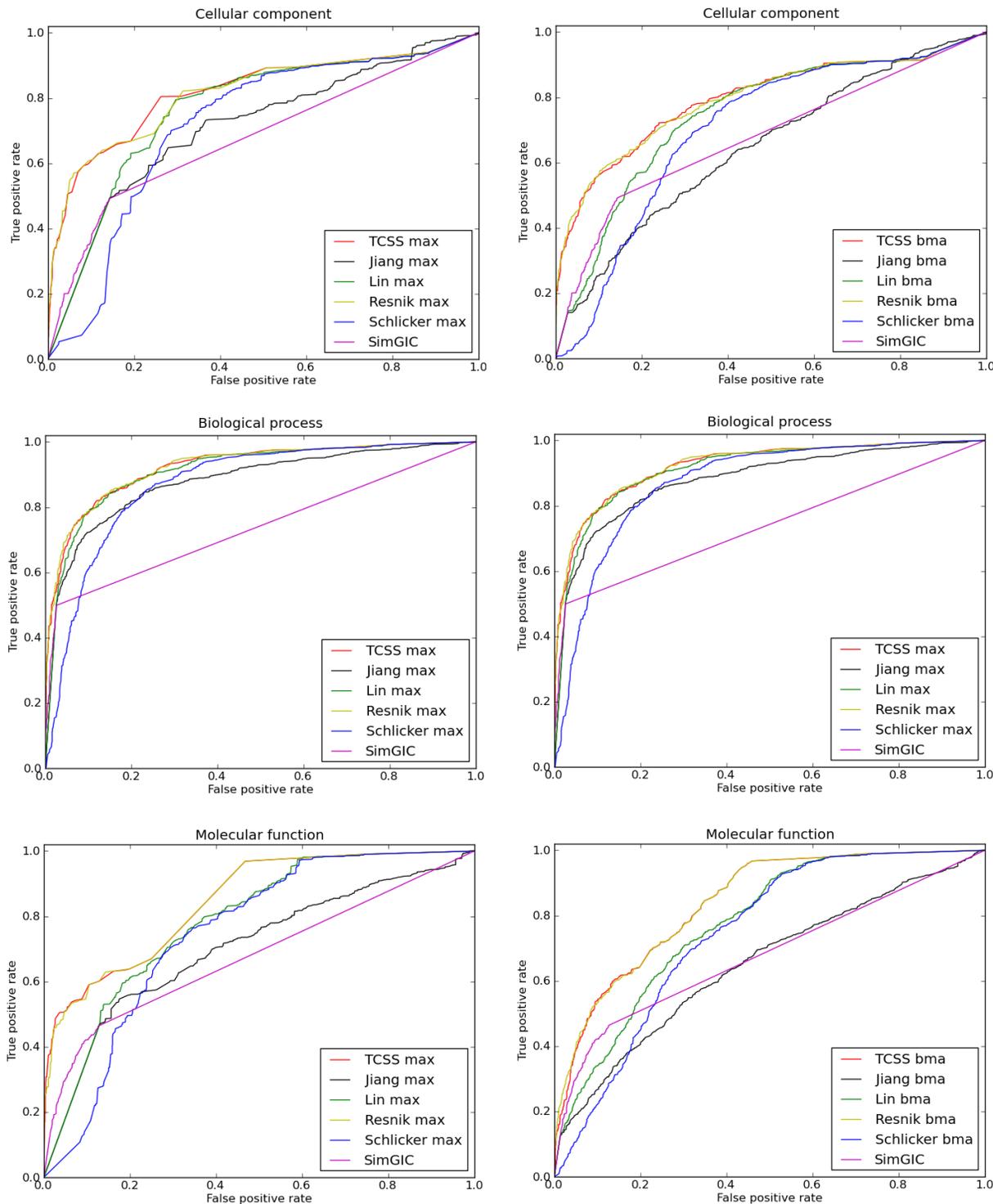


Figure 2.11: ROC curves for *H. sapiens* PPI dataset (IEA+). ROC evaluations of semantic similarity measures at different cutoffs based on the *H. sapiense* PPI dataset derived from DIP are shown. The evaluation was performed using cellular component, biological process, molecular function ontology of GO. Best-match average (bma) and maximum (max) approaches for combining multiple annotations are used on dataset with (IEA+) electronic annotations. TCSS & Resnik show best ROC profiles for all three ontologies.

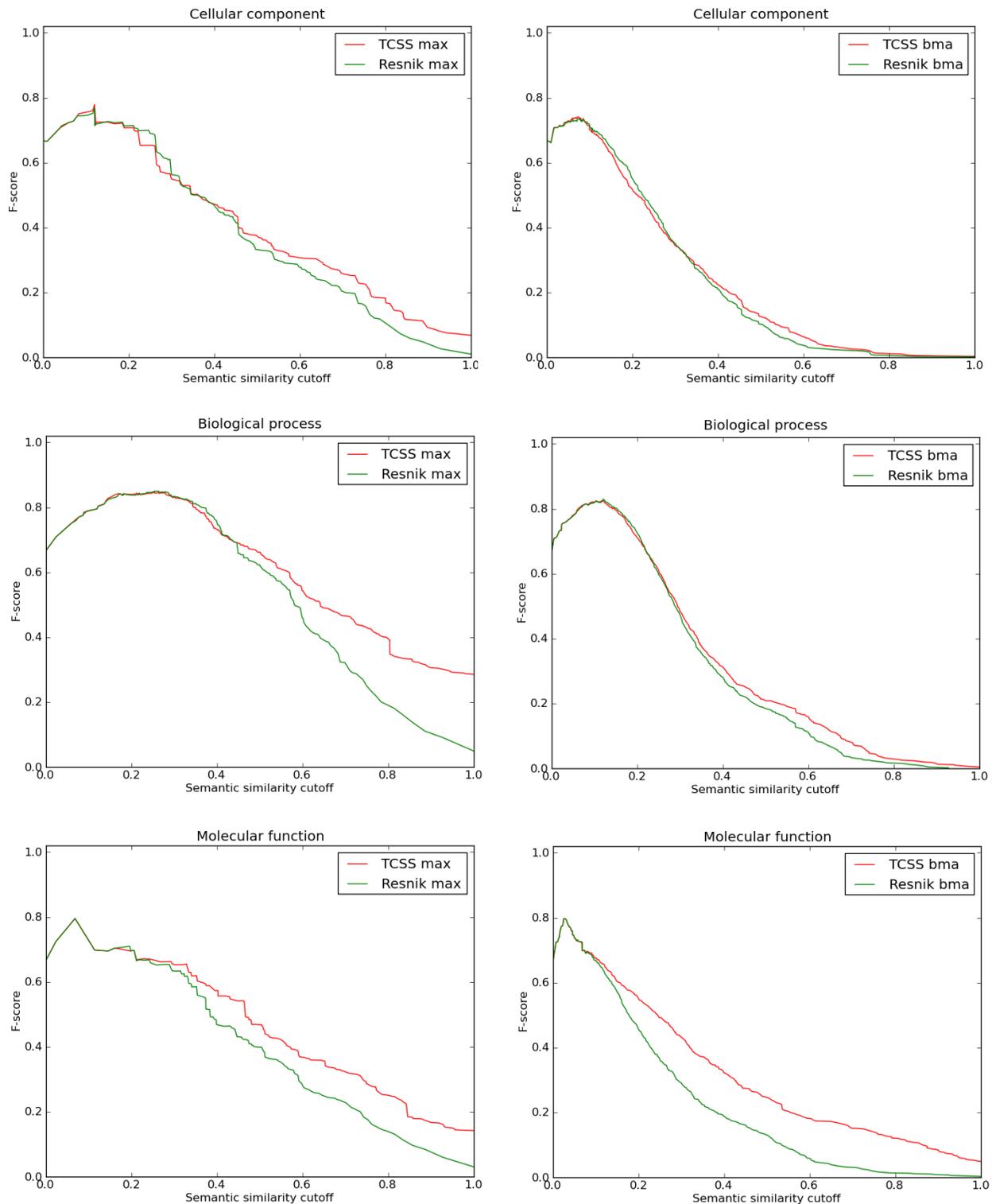


Figure 2.12: F_1 -score curves for *H. sapiens* PPI dataset (IEA+). F_1 score (harmonic mean of precision and recall) evaluations of TCSS and Resnik semantic similarity measures at different cutoffs based on the *H. sapiens* PPI dataset derived from DIP are shown. The evaluation was performed using cellular component, biological process, molecular function ontology of GO. Best-match average (bma) and maximum (max) approaches for combining multiple annotations was used on dataset with (IEA+) electronic annotations. F_1 score reaches its best value at 1 and worst at 0. TCSS does better than Resnik for semantic similarity cutoff scores in all three ontologies.

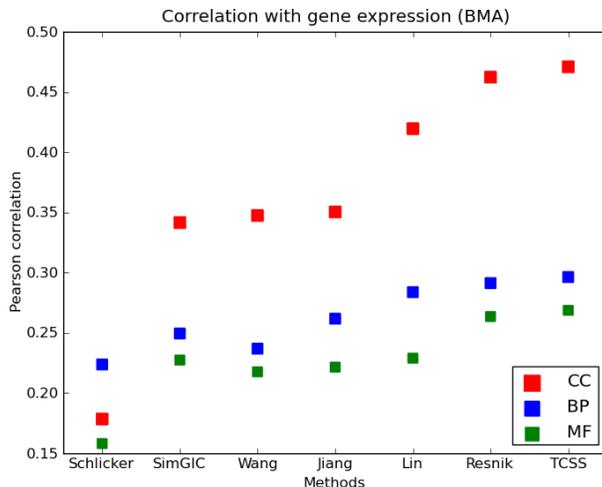


Figure 2.13: Correlation with gene expression. Pearson correlation between gene expression similarity and semantic similarity on *S. cerevisiae* dataset are shown. The evaluation was performed for cellular component, biological process, and molecular function ontologies of GO. Best-match average (bma) approach for combining multiple GO annotations was used. TCSS showed best correlation with gene expression in all three ontologies.

(Figure 2.14(c)). SimGIC better correlates with sequence similarity than other methods in all three ontologies (Figures 2.14(d), 2.15).

2.6 Discussion

We present a new algorithm (TCSS) for calculating semantic similarity and tested its performance against other methods. TCSS shows an average improvement of 4.6 times in F_1 scores over Resnik, the next best method, on our *S. cerevisiae* PPI test and 2 times on our *H. sapiens* PPI test. This clearly indicates the advantage of using TCSS to retrieve positive protein interactions and hold back negative interactions over Resnik’s method. We compared TCSS using both the BMA and MAX approaches for combining multiple GO annotations, and found that MAX generally works best for PPI datasets. The use of the MAX function to score PPIs, instead of an ‘average’ function, makes sense because proteins in PPIs only need to be in close proximity (similar cellular component terms) or in a similar biological process once, among all possible combinations annotation terms, to be biologically relevant. Therefore, the MAX approach is unlikely to overestimate true PPIs. However, there may be application scenarios (e.g to compute a more general measure of functional similarity) where the MAX approach could lead to over-estimation and BMA would be a better choice. In these cases, TCSS can be modified to use the BMA method instead of MAX. For example,

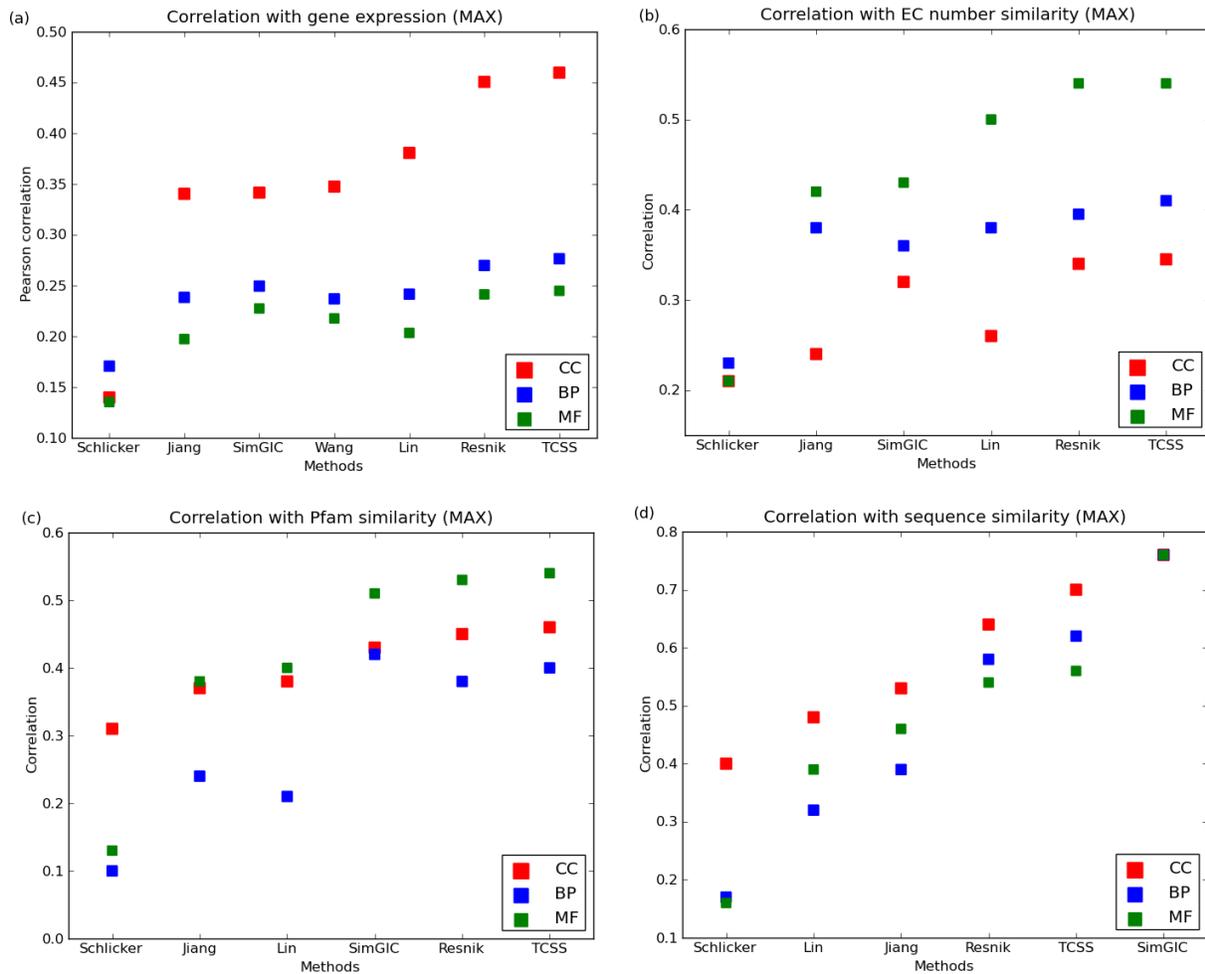


Figure 2.14: Correlation with gene expression and CESSM dataset. (a) Pearson correlation between gene expression similarity and semantic similarity on a *S. cerevisiae* dataset containing 5000 randomly selected protein pairs are shown. (b - d) Correlation between semantic similarity and sequence, enzyme commission (EC), protein family (Pfam) similarity using online CESSM tool. The evaluation was performed for cellular component (CC), biological process (BP), and molecular function (MF) ontologies of GO using maximum (max) approach for combining multiple GO annotations.

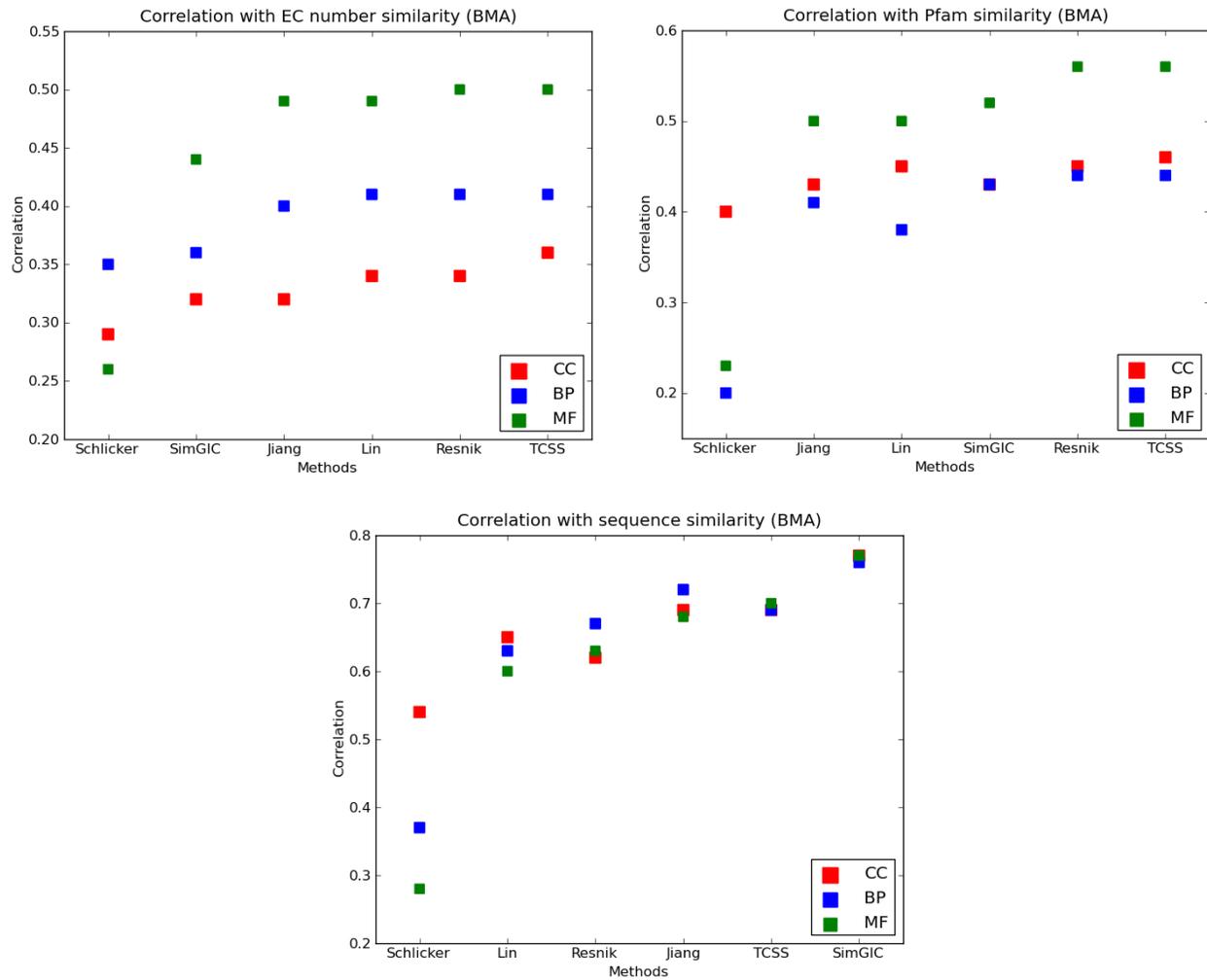


Figure 2.15: Correlation with CESSM dataset. Correlation between semantic similarity and sequence, enzyme commission (EC), protein family (Pfam) similarity using online CESSM tool. The evaluation was performed for cellular component (CC), biological process (BP), and molecular function ontologies (MF) of GO. Best-match average (bma) approach for combining multiple GO annotations was used on the dataset without (IEA-) electronic annotations. TCSS showed best correlation with EC & Pfam similarity for CC ontology and same as Resnik’s for MF and BP ontologies.

TCSS shows worse correlation with Pfam similarity than SimGIC on the biological process ontology test, but becomes better when using BMA (Figure 2.15). Also, it is evident from the correlation of semantic similarity with gene expression similarity that TCSS is more likely to assign a higher score to gene products if they also exhibit similar gene expression. Tests using the CESSM benchmark dataset were in favor of TCSS for EC number similarity and Pfam similarity. SimGIC does better than TCSS in the sequence similarity correlation test. One reason for this could be that SimGIC scores gene products with shared annotation terms and gene products annotated to same term are more likely to be part of the same gene family.

Scatter plots of the semantic similarity scores obtained by TCSS (MAX) and Resnik (MAX) methods clearly indicate that a significant number of positive interactions are under-scored by Resnik (Figure 2.16) in all three ontologies (p-values by Kolmogorov-Smirnov test: Cellular component: $6.4e-59$, Biological process: $3.4e-163$, Molecular function: $1.6e-15$). Given below are some biological examples selected from these scatter plots in support of our claim:

- **Cellular component:** Rpl10p is a *S. cerevisiae* protein responsible for joining of the 40S and 60S ribosomal subunits (Stark *et al.*, 2006). It has been found to interact (Krogan *et al.*, 2006; Hofer *et al.*, 2007; West *et al.*, 2005; Eisinger *et al.*, 1997) with Sqt1p, an essential protein involved in a late step of 60S ribosomal subunit assembly or modification (Stark *et al.*, 2006) using affinity capture-mass spectrometry (MS), affinity capture-western and two-hybrid experimental methods. RPL10 is annotated to the 'cytosolic large ribosomal subunit' term and Sqt1p is annotated to the 'cytosolic ribosome' term (The Gene Ontology Consortium, 2000). The score assigned by Resnik (MAX) to the Rpl10p-Sqt1p interaction is 0.4 which is low considering that both the proteins are in similar cellular components and the 'cytosolic large ribosomal subunit' term is the child term of 'cytosolic ribosome' in GO. The same interaction gets a score of 0.78 by TCSS (MAX), which categorizes it as a high confidence interaction, due to the normalization step on the 'cytosolic ribosome' sub-graph.
- **Biological process:** The Nth1p-Dcs1p protein-protein interaction was experimentally shown by Yu *et al.* (2008); Uetz *et al.* (2000) using two-hybrid experiments. Both Nth1p and Dcs1p proteins share the 'vacuolar protein catabolic process' term in GO (The Gene Ontology Consortium, 2000). The score assigned by Resnik (MAX) to the Nth1p-Dcs1p interaction is 0.45 which is low considering that both proteins are part of the same biological process. The same interaction gets a score of 1 by TCSS (MAX), due to the normalization on 'vacuolar protein

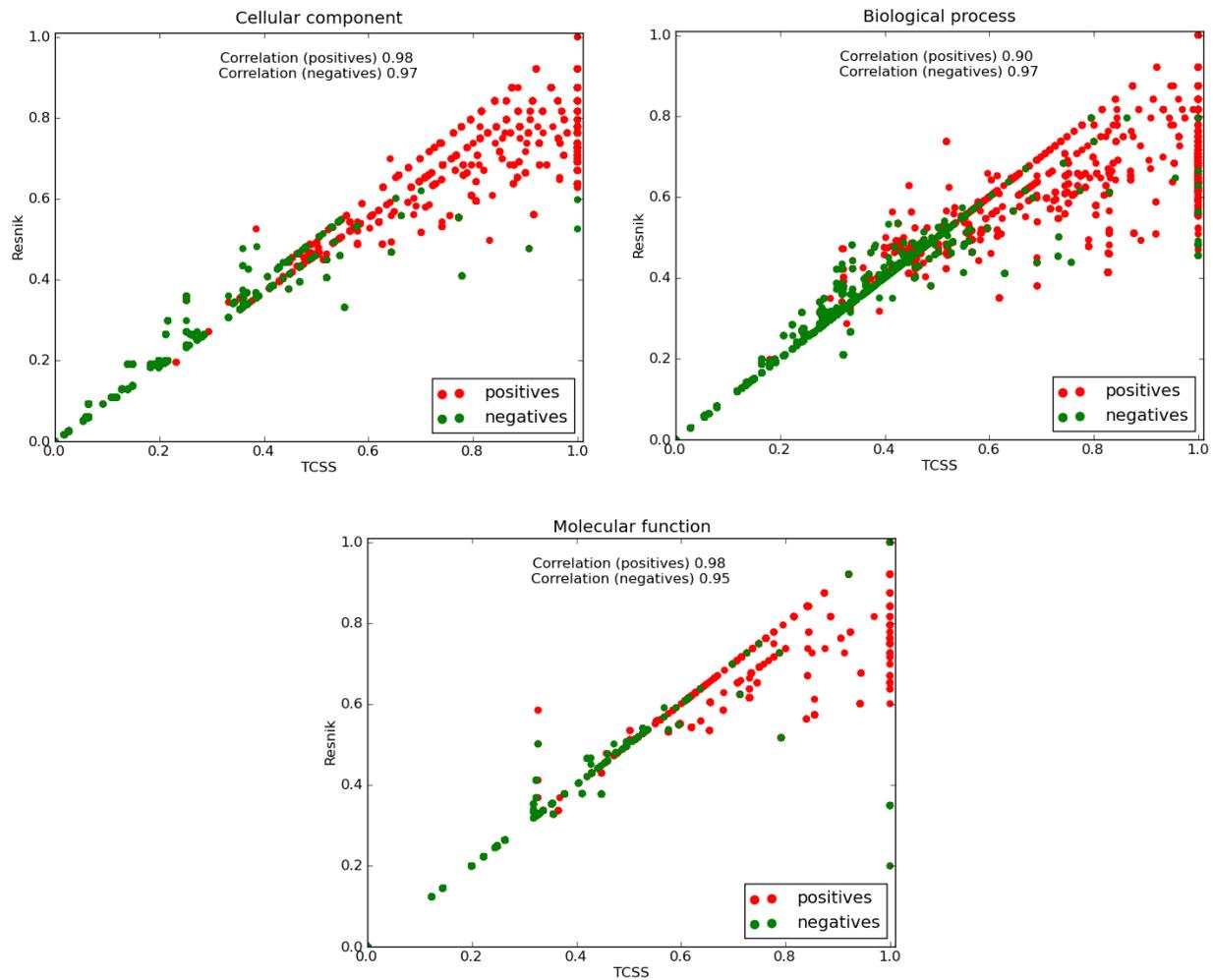


Figure 2.16: Comparison of our topological clustering method and Resnik (MAX) as scoring positive and negative PPIs. The scatter plot of semantic similarity scores for positive (red) and negative (green) interactions. Semantic similarity scores range between 0.0 and 1.0 for both methods, with 1.0 being the best. A significant number of positive interactions are under-scored by Resnik (Max) in all three ontologies compared to TCSS.

catabolic process' sub-graph, thus categorizing it as a high confidence interaction.

- **Molecular function:** Mft1p and Hpr1p are the subunits of the nuclear THO complex. THO complex is involved in transcription elongation, mitotic recombination and telomere maintenance (Stark *et al.*, 2006). Mft1p-Hpr1p interaction has been shown by affinity capture-MS and affinity capture-western experimental techniques (Strasser *et al.*, 2002; Krogan *et al.*, 2006; Gavin *et al.*, 2006; Chavez *et al.*, 2000). Both Mft1p and Hpr1p are annotated to the 'nucleic acid binding' term of GO (The Gene Ontology Consortium, 2000). This interaction is assigned a score of 0.2 by Resnik (MAX) because the term nucleic acid binding is fairly general. This score is low considering that both the proteins are part of a same GO term. The same interaction is assigned a score of 1 by TCSS (MAX), due to the normalization step on the 'nucleic acid binding' sub-graph. 'Nucleic acid binding' is a general molecular function term with a shallow hierarchy.

Future directions for TCSS development include testing if the GO graph edge type (e.g. is-a, part-of) can provide additional information that will lead to improved performance and also testing the method more rigorously with other data sets.

2.7 Conclusions

We present a new semantic similarity algorithm, Topological Clustering Semantic Similarity, designed to use the GO for PPI confidence assessment. It partitions the GO DAG into non-overlapping sub-graphs, using a topological clustering method, and computes semantic similarity normalized within each sub-graph. We evaluated TCSS against other methods for measuring semantic similarity between GO terms annotated to proteins involved in protein-protein interactions from *S. cerevisiae* and *H. sapiens*. We also tested the correlation between multiple semantic similarity scoring methods with gene expression, protein sequence, EC, and Pfam similarity. Performance tests were generally in favor of TCSS in all three GO ontologies: cellular component, biological process and molecular function. This new method will be useful as an evidence source in PPI prediction or in confidence assessment of PPI datasets.

Implementation

TCSS

The algorithm was implemented using the Python programming language (Van Rossum and Drake Jr, 1995). An important step in our algorithm is to determine the size of sub-graphs. This is determined by thresholding the topological information content (ICT) of terms in a given ontology. The cutoff is chosen to maximize performance (AUC and F_1 measures) on a given benchmark/test. The relationship between AUC and topology cutoff follows a U - shaped curve with a global maximum for all three ontologies. Average F-score shows a general upward trend with topology cutoffs (Figures 2.17, 2.18, 2.20, 2.21). A topology cutoff must be computed for each test before we compute semantic similarity scores, which is a practical disadvantage of our method, though we expect cutoffs to be useful generally for a type of data and an organism once computed. Topology cutoffs for different datasets are as follows:

- *S. cerevisiae* PPI dataset: 2.4 for CC, 3.6 for BP, and 3.2 for MF (Figure 2.19)
- *H. sapiens* PPI dataset: 3.0 for CC, 4.0 for BP, and 3.6 for MF (Figure 2.22)
- Expression dataset: 2.4 for CC, 3.6 for BP, and 3.2 for MF
- CESSM dataset: 3.4 for CC, 3.2 for BP, and 3.0 for MF

Our results are resilient in the immediate cutoff range of ± 0.1 for all three ontologies.

Other methods

Semantic similarity measurement methods proposed by Resnik (1995) (Resnik), Lin (1998) (Lin), Schlicker *et al.* (2006) (simRel) (Schlicker), Jiang and Conrath (1997) (Jiang), and Pesquita *et al.* (2007) (SimGIC) were implemented as mentioned in respective publications. The GOSemSim (Yu *et al.*, 2010) implementation in R was used Wang *et al.* (2007) (Wang).

ROC and F-measure

Different measures used for analyzing the performance of our algorithm are as follows:

- True positive rate (TPR), also known as Recall:

$$TPR = \frac{TP}{TP + FN} \quad (2.12)$$

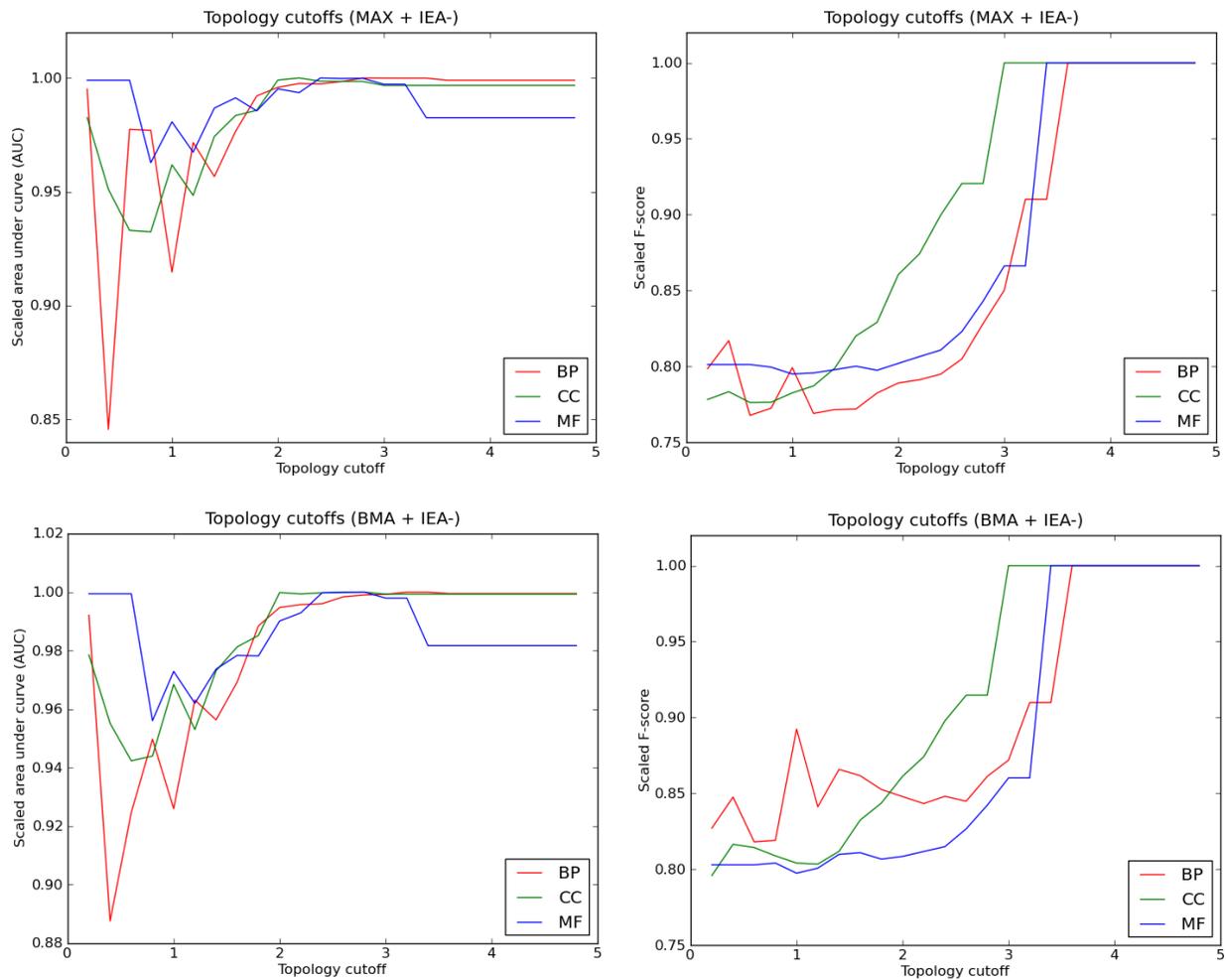


Figure 2.17: Effect of topology cutoff on (ROC) AUC and F-score for *S. cerevisiae* PPI dataset (IEA-). Change in AUC (TPR/FPR ROC) values and average F-scores with respect to topology cutoffs under different settings. BMA stands for best-match average approach of combining multiple annotations and MAX stands for maximum approach. Test was conducted separately for cellular component (CC), biological process (BP), and molecular function (MF) ontologies without IEA (IEA-) annotations.

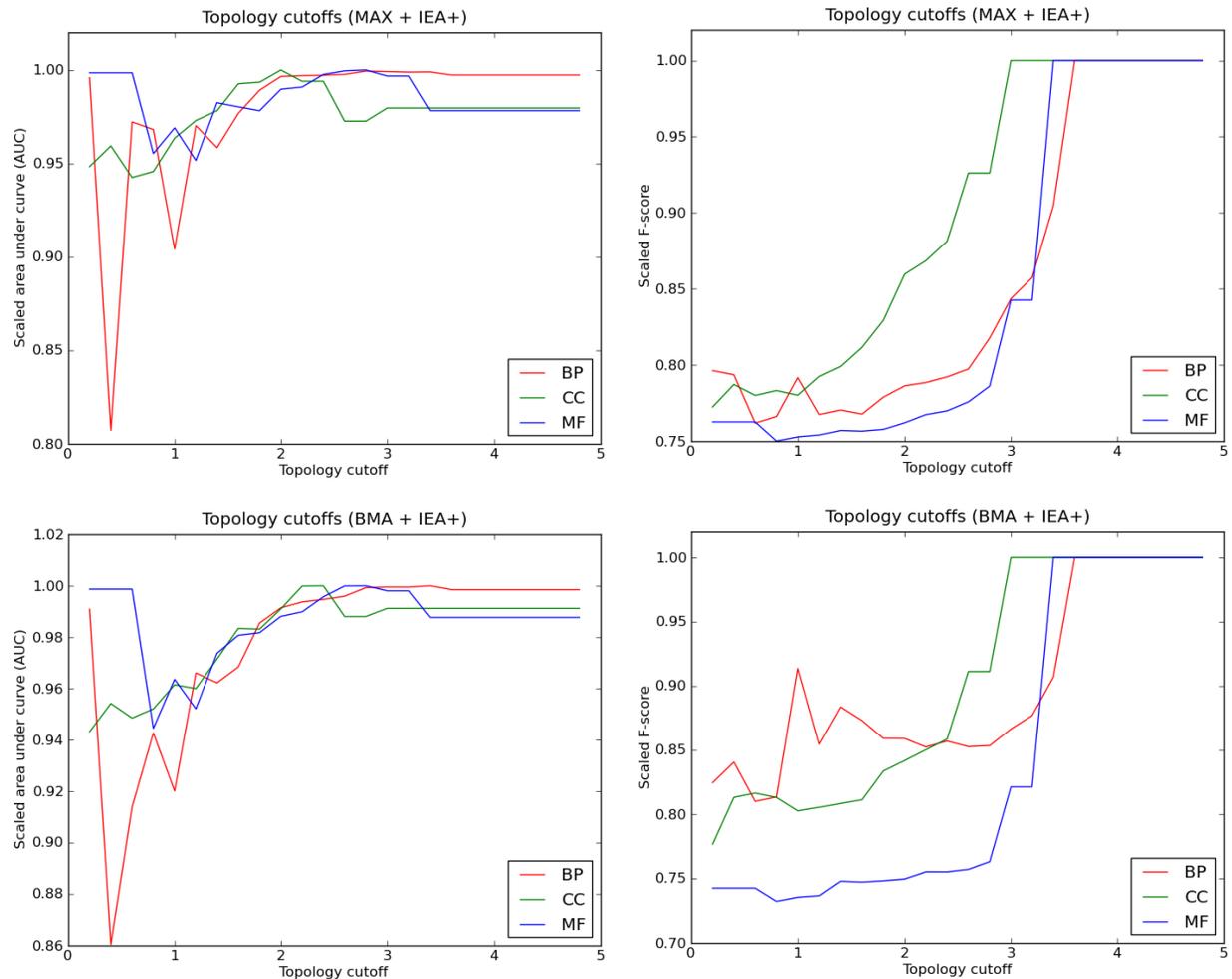


Figure 2.18: Effect of topology cutoff on (ROC) AUC and F-score for *S. cerevisiae* PPI dataset (IEA+). Change in AUC (TPR/FPR ROC) values and average F-scores with respect to topology cutoffs under different settings. BMA stands for best-match average approach of combining multiple annotations and MAX stands for maximum approach. Test was conducted separately for cellular component (CC), biological process (BP), and molecular function (MF) ontologies with IEA (IEA+) annotations.

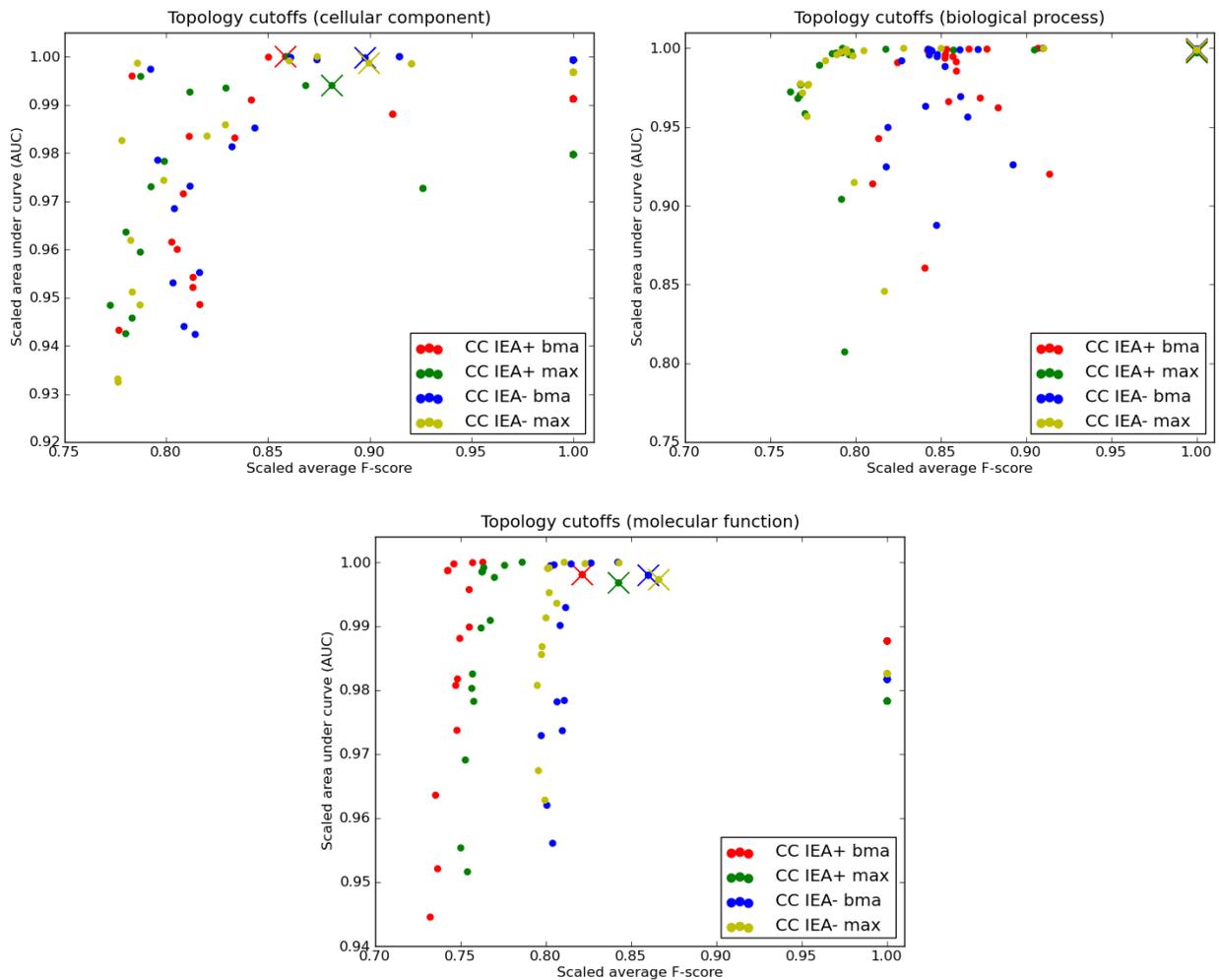


Figure 2.19: Topology cutoff for *S. cerevisiae* PPI dataset. Topology cutoffs for cellular component (CC), biological process (BP), and molecular function (MF) ontologies were determined by evaluating AUC values and average F-scores at different cutoffs. The topology cutoff where both the AUC and average F-score are maximized under different conditions is picked. Test was done with best-match average (bma) and maximum (max) approaches of combining multiple annotations on datasets with (IEA+) and without (IEA-) electronic annotations. Topology cutoff value chosen for CC is 2.4, BP is 3.6, and MF is 3.2 (marked by "X").

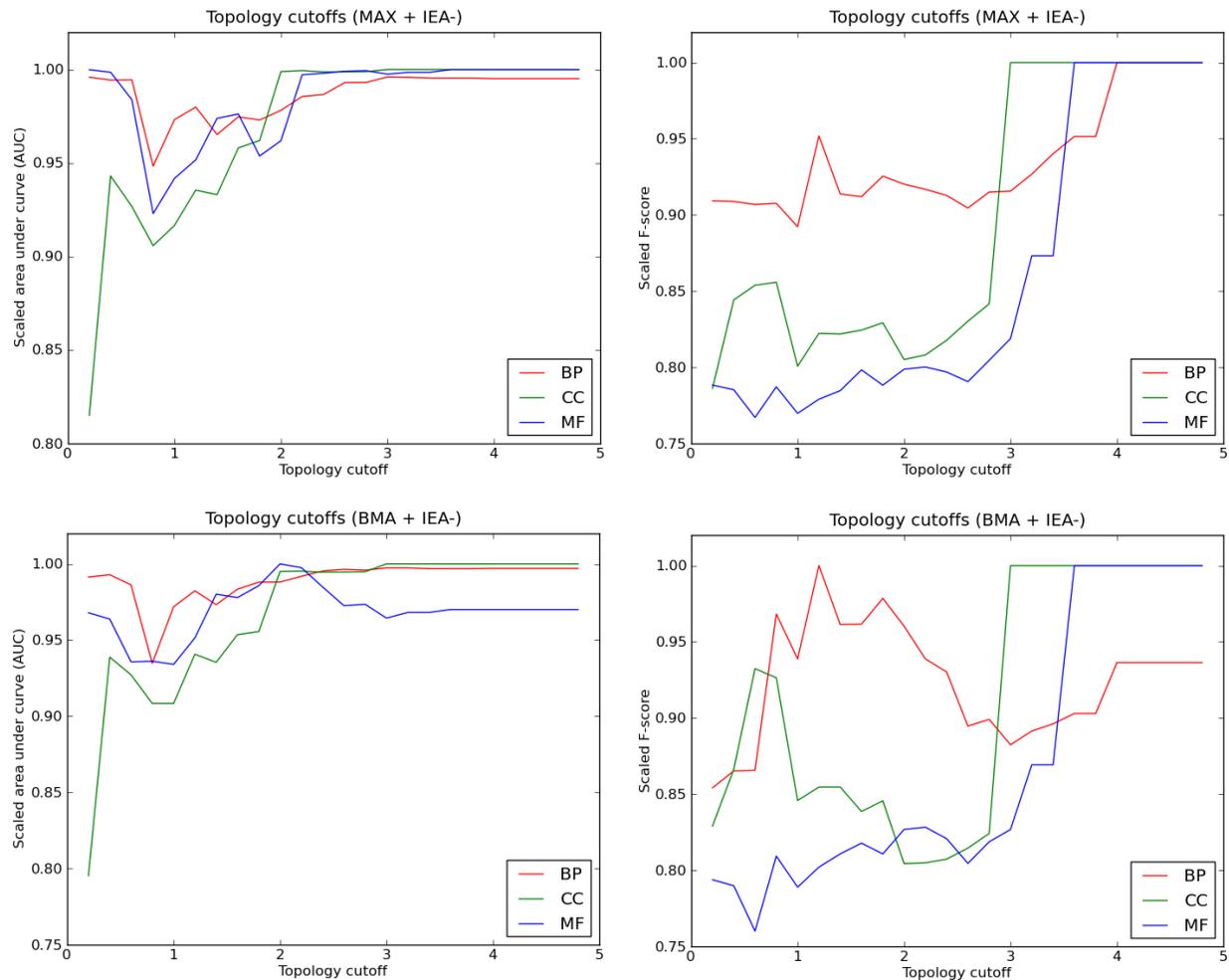


Figure 2.20: Effect of topology cutoff on (ROC) AUC and F-score for *H. Sapiens* PPI dataset (IEA-). Change in AUC (TPR/FPR ROC) values and average F-scores with respect to topology cutoffs under different settings. BMA stands for best-match average approach of combining multiple annotations and MAX stands for maximum approach. Test was conducted separately for cellular component (CC), biological process (BP), and molecular function (MF) ontologies without IEA (IEA-) annotations.

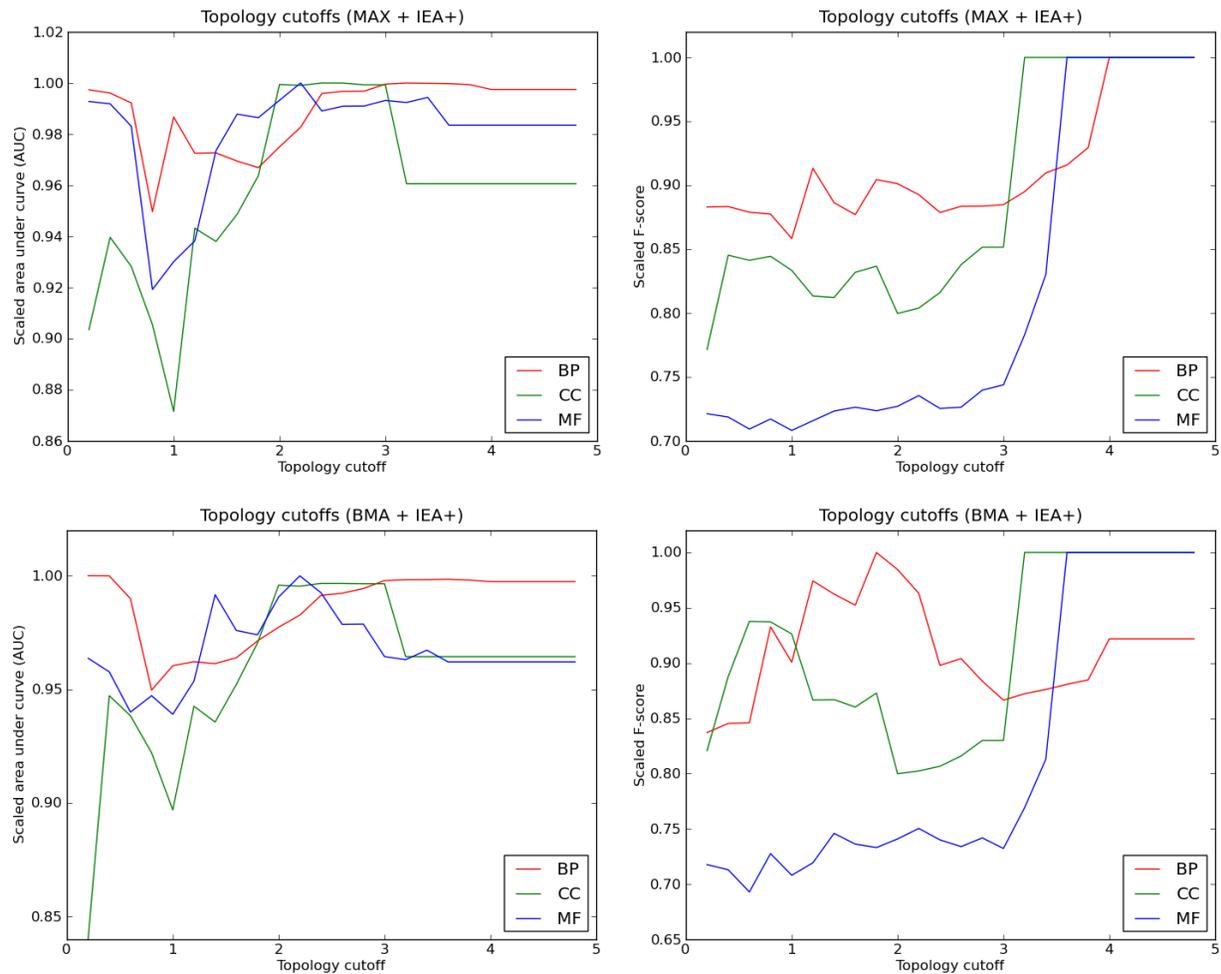


Figure 2.21: Effect of topology cutoff on (ROC) AUC and F-score for *H. Sapiens* PPI dataset (IEA+). Change in AUC (TPR/FPR ROC) values and average F-scores with respect to topology cutoffs under different settings. BMA stands for best-match average approach of combining multiple annotations and MAX stands for maximum approach. Test was conducted separately for cellular component (CC), biological process (BP), and molecular function (MF) ontologies with IEA (IEA+) annotations.

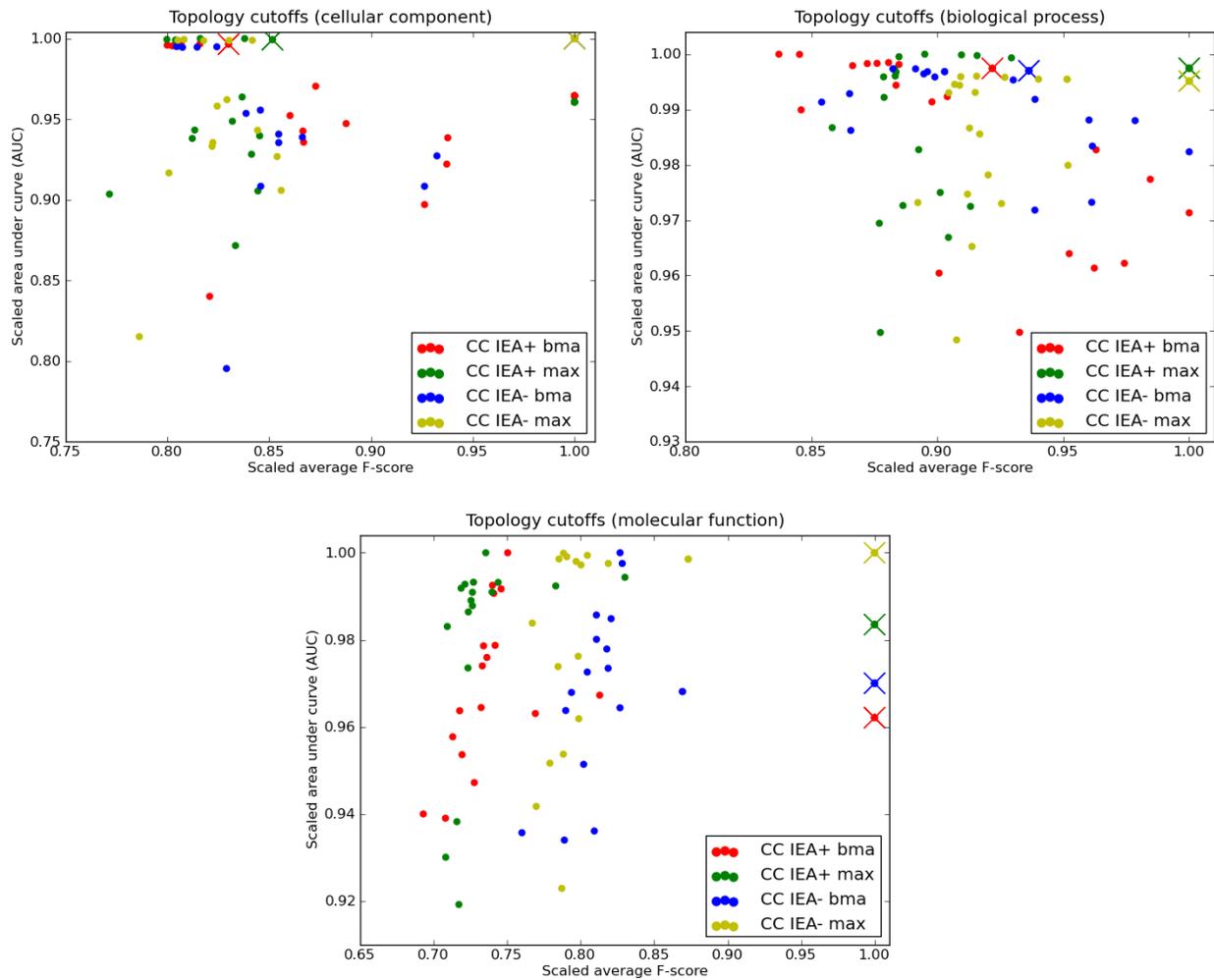


Figure 2.22: Topology cutoff for *H. Sapiens* PPI dataset. Topology cutoffs for cellular component (CC), biological process (BP), and molecular function (MF) ontologies were determined by evaluating AUC values and average F-scores at different cutoffs. The topology cutoff where both the AUC and average F-score are maximized under different conditions is picked. Test was done with best-match average (bma) and maximum (max) approaches of combining multiple annotations on datasets with (IEA+) and without (IEA-) electronic annotations. Topology cutoff value chosen for CC is 3.0, BP is 4.0, and MF is 3.6 (marked by "X").

- False positive rate (FPR):

$$FPR = \frac{FP}{FP + TN} \quad (2.13)$$

- Precision (P):

$$P = \frac{TP}{TP + FP} \quad (2.14)$$

- F₁measure (F):

$$F = 2 \frac{P \times TPR}{P + TPR} \quad (2.15)$$

- Improvement in F₁ score is calculated as the average improvement at different semantic similarity cutoffs.
- Area under curve (AUC) was calculated using the trapezoidal rule.

where TP , FP , TN , FN are true positive, false positive, true negative, and false negative, respectively.

Correlation with gene expression

Average gene expression Pearson correlation was calculated for the *S. cerevisiae* positive and negative interaction dataset using Fisher's z transformation (Faller, 1981).

$$z_n = 1/2 \ln \left(\frac{1 + r_n}{1 - r_n} \right) \quad (2.16)$$

where r_n is the Pearson correlation between two genes for the n^{th} experiment. Then, an appropriate estimate of the true mean is calculated as,

$$\bar{z}_n = N^{-1} \sum_{i=1}^N z_i \quad (2.17)$$

where N is the total number of experiments. Then, by inversion, average correlation is calculated as,

$$\bar{r}_n = \frac{e^{2\bar{z}_n} - 1}{e^{2\bar{z}_n} + 1} \quad (2.18)$$

CESSM evaluation

TCSS, Schlicker, Jiang, SimGIC and Resnik methods were used to find the semantic similarity between protein pairs provided by CESSM. Correlation between semantic similarity scores and

sequence, Pfam, EC similarity for these methods was calculated using the CESSM online tool. Wang was not used here due to the difficulty in modifying the R implementation to use the datasets provided by the CESSM website.

Chapter 3

Predicting *in-vivo* SH3 domain mediated protein interactions in yeast

This work was published in Bioinformatics, 15;32(12):1865-72: Jain, S. and Bader, GD. (2016), Predicting physiologically relevant SH3 domain mediated protein-protein interactions in yeast.

Author contributions: I collected the data, developed and implemented the method and performed the analyses. Gary D. Bader supervised and advised this project.

3.1 Abstract

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. A novel method for predicting physiologically relevant (or *in vivo*) SH3 domain-peptide mediated protein-protein interactions in *S. cerevisiae* 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. **Domain-Motif Mediated Interaction Prediction (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>.

3.2 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 Schlessinger, 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 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.* (2012b), 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 works in the field of ensemble PPI prediction by 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.

3.3 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 Bayes integration to score the PWM targets. Another naïve Bayes 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 3.1 shows the work flow of our PRM mediated PPI prediction pipeline.

3.4 Methods

3.4.1 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. PWMs contain a weight for each alphabet symbol i at each position j in the motif. Weight can be described as a log-odds score of a probabilistic model against a background (Pizzi *et al.*, 2011).

$$M(i, j) = \log \frac{P(i, j)}{B(i)} \quad (3.1)$$

where $B(i)$ is the background probability of amino acid i in the proteome and $P(i, j)$ is the proba-

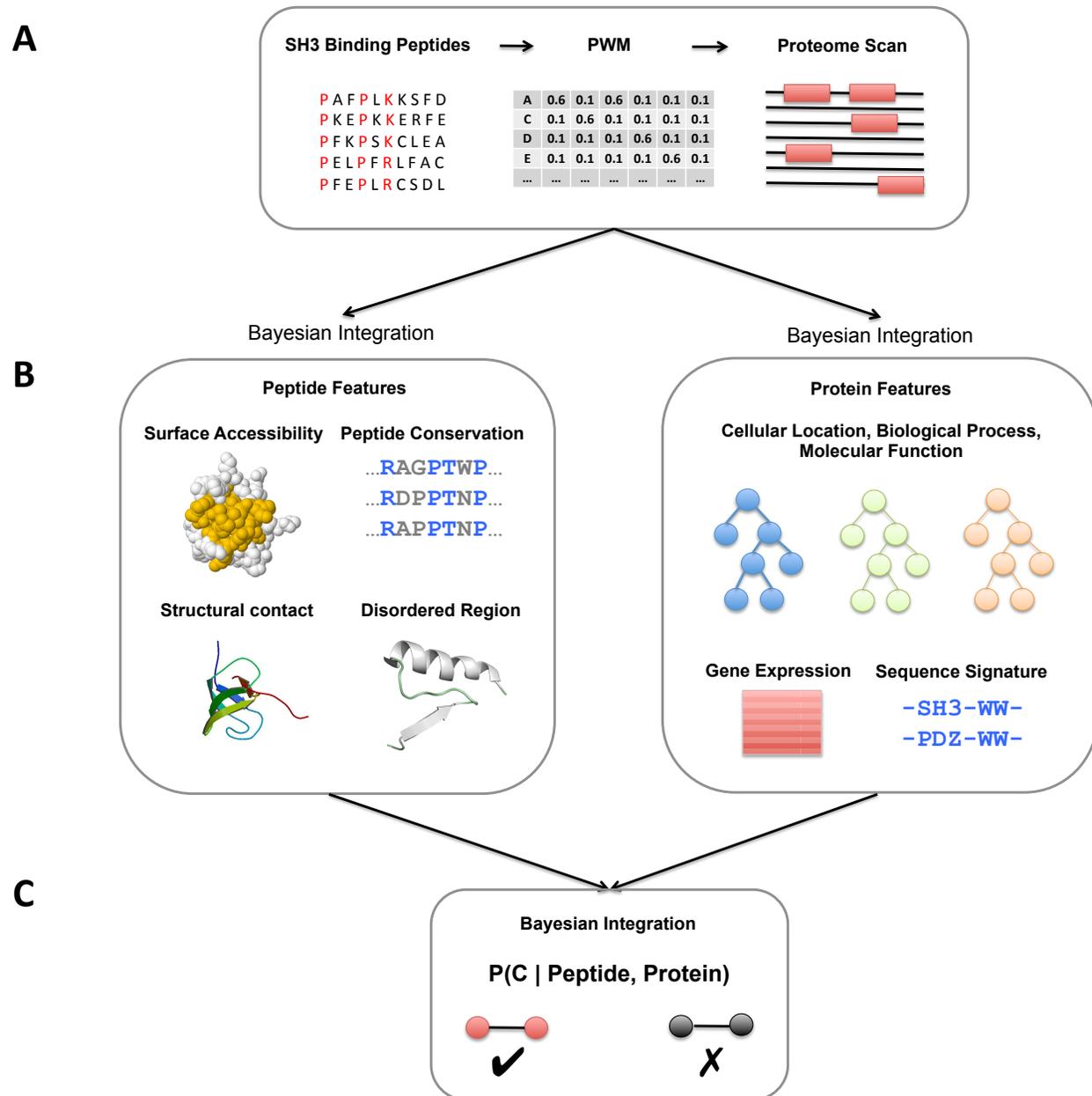


Figure 3.1: Work flow of PRM mediated PPI prediction pipeline. (A) Proteome is scanned using a PWM built using experimentally derived binding peptides (e.g. from phage display) of a given SH3 domain for potential interactors. (B) Separate naïve Bayes classifiers for peptide and protein features. (C) Integration of classifiers for predicting interacting and non-interacting protein pairs.

bility of amino acid i at position j .

$$P(i, j) = \frac{\text{count}(i, j)}{N} \quad (3.2)$$

where $\text{count}(i, j)$ is the empirical count of amino acid i at position j and N is the count of all the amino acids at position j . Low information content positions or columns at the edges of PWMs are removed to improve signal of the core motif. The information content of each position in the motif is calculated as (Erill, 2012),

$$IC(j) = \left[-\sum_{i=1}^m B(i) \log B(i) \right] - \left[-\sum_{i=1}^m P(i, j) \log P(i, j) \right] \quad (3.3)$$

where $IC(j)$ is the mutual information content of j^{th} position in the motif. Information content ratio is then calculated as,

$$ICR(j) = \frac{IC(j)}{IC_{max}} \quad (3.4)$$

Amino acid positions on both ends of the motif with $ICR(j) \leq 0.4$ are removed. Trimmed PWMs are used to scan a protein sequence to find matches of the weighted pattern above a threshold score (k). For a protein sequence ($S = s_1s_2s_3\dots$) the match score ($W(s)$) of any m amino acid long segment is the sum of individual amino acid weights in the PWM (Pizzi *et al.*, 2011).

$$W(s_j\dots s_{j+m-1}) = \sum_{j=1}^m M(s_j, j) \quad (3.5)$$

where $M(s_i, j)$ is the log-odds score of amino acid s_i at position j in the PWM. The number of statistically significant matches are controlled by converting match score thresholds to p-values. For a given PWM the relationship between its match scores and p-values is defined such that in the background distribution match scores $W(s) \geq k$ (Pizzi *et al.*, 2011; Wu *et al.*, 2000). Not all amino acid positions within a motif are significant. For example, in class 1 SH3 binding motif [R/K]xxPxxP, positions 1, 4, and 7 are more significant than others. Amino acid positions with $(IC(j)) \geq 0.5$ within the trimmed PWMs are identified as significant. These significant amino acid positions are used in calculation of disordered region, surface accessibility, and peptide conservation scores.

3.4.2 Peptide features

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. Disordered region (DR) score is calculated as the fraction of disordered amino acids at significant positions in the binding site.

$$DR = \frac{\sum_i p_i}{N} = \begin{cases} 1 & \text{if amino acid } i \text{ is disordered} \\ 0 & \text{otherwise} \end{cases} \quad (3.6)$$

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.

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). Surface accessibility (SR) score is then calculated as the fraction of exposed amino acid residues at significant positions in the binding site.

$$SA = \frac{\sum_i p_i}{N} = \begin{cases} 1 & \text{if } RSA \geq 25\% \\ 0 & \text{otherwise} \end{cases} \quad (3.7)$$

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

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 (Katoch *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). Peptide conservation (PC) score is defined as average conservation score of significant amino acid residues in the binding site.

$$PC = \frac{\sum_i p_i}{N} \quad (3.8)$$

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.

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. Structural contact (SC) score is defined as the average contact area of significant amino acid residues in the binding site.

$$SC = \max_j \frac{\sum_i c_{ij}}{N} \quad (3.9)$$

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.

3.4.3 Protein features

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.

$$CC = TCSS(a, b, ontology = C, cutoff = 2.4) \quad (3.10)$$

$$BP = TCSS(a, b, ontology = P, cutoff = 3.5) \quad (3.11)$$

$$MF = TCSS(a, b, ontology = F, cutoff = 3.3) \quad (3.12)$$

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

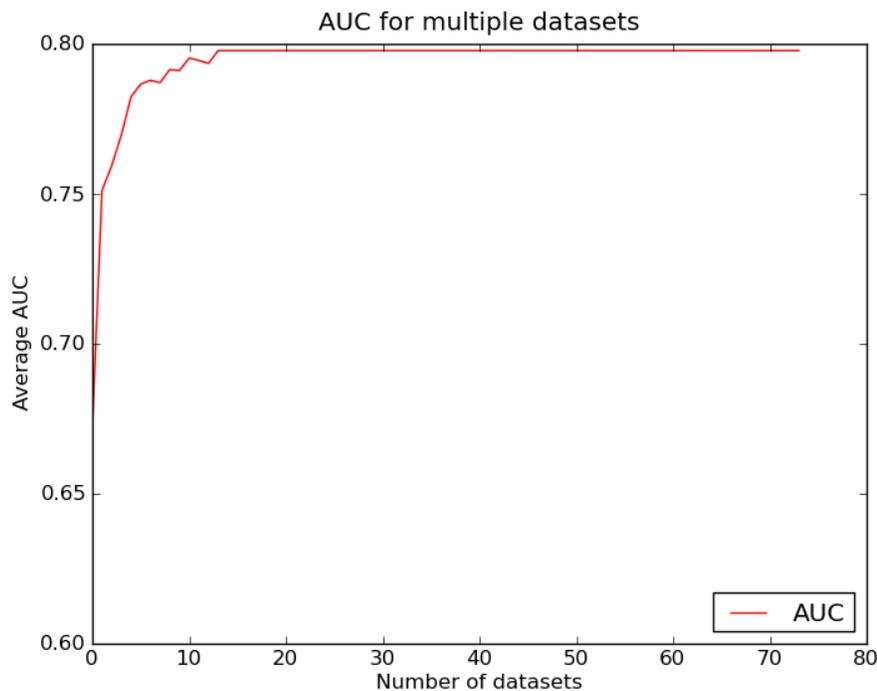


Figure 3.2: Change in average area under the curve (AUC) with the number of yeast gene expression datasets used for predicting PPIs. This figure was generated by randomly selecting (repeated 100 times) yeast gene expression datasets in incremental fashion and doing receiver operating characteristic (ROC) analysis.

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 3.2. 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)

$$EX = 1 - \frac{e^{2\bar{z}} + 1}{e^{2\bar{z}} - 1} \quad (3.13)$$

$$\bar{z} = N^{-1} \sum_{i=1}^N \frac{1}{2} \ln \left(\frac{1 + r_i}{1 - r_i} \right) \quad (3.14)$$

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

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).

$$SS = \sum_{ij} -\log_2 \left(\frac{p_{ij}}{p_i p_j} \right) \quad (3.15)$$

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.

3.4.4 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 Bayes 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. (Mitchell, 1997). For a protein pair described by a set of features $X = \langle X_1, X_2, \dots, X_n \rangle$ a naïve Bayes PPI prediction model is defined as,

$$\begin{aligned}
\arg \max_Y P(Y|X) &= \arg \max_Y \frac{P(X|Y)P(Y)}{P(X)} \\
&= \arg \max_Y P(Y) \prod_i P(X_i|Y) \\
\arg \max_Y \log P(Y|X) &= \arg \max_Y \log P(Y) + \sum_i \log P(X_i|Y)
\end{aligned} \tag{3.16}$$

where $P(Y)$ is the class prior probability and $P(X_i|Y)$ is the class-conditional probability of feature $X_i \in X$. As there are only two classes $Y \in \{\text{interacting, non-interacting}\}$ therefore class priors are estimated by treating $P(Y)$ as a multinomial (or categorical) distribution $P(Y) = \Pi_Y$. All continuous peptide and protein features are discretized by binning and modeled using a multinomial probability distribution $P(X_i|Y) = \text{Multi}(X_i; \theta_{iY}) \propto \Theta_{iY}^{X_i}$. Putting it all together, the naïve Bayes model is defined as,

$$\arg \max_Y \log P(Y|X) = \arg \max_Y \log \Pi_Y + \sum_i \log \Theta_{iY}^{X_i} \tag{3.17}$$

where model parameters Π_Y and $\Theta_{iY}^{X_i}$ are learned from the training data set. 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),

$$P(Y|M_{pep}, M_{pro}) = \frac{P(Y)P(M_{pep}, M_{pro}|Y)}{P(M_{pep}, M_{pro})} \tag{3.18}$$

as M_{pep} and M_{pro} are independent therefore, they are conditionally independent given the class Y ,

$$P(M_{pep}, M_{pro}|Y) = P(M_{pep}|Y)P(M_{pro}|Y) \tag{3.19}$$

substituting $P(M_{pep}, M_{pro}|Y)$ in equation (3.18),

$$P(Y|M_{pep}, M_{pro}) = \frac{P(Y)P(M_{pep}|Y)P(M_{pro}|Y)}{P(M_{pep}, M_{pro})} \quad (3.20)$$

re-writing $P(M_{pep}|Y)$ and $P(M_{pro}|Y)$ using Bayes theorem,

$$\begin{aligned} P(Y|M_{pep}, M_{pro}) &= \frac{P(Y)P(Y|M_{pep})P(M_{pep})P(Y|M_{pro})P(M_{pro})}{P(Y)P(Y)P(M_{pep}, M_{pro})} \\ &= \frac{P(M_{pep})P(M_{pro})}{P(M_{pep}, M_{pro})} \times \frac{P(Y|M_{pep})P(Y|M_{pro})}{P(Y)} \\ &= \alpha \frac{P(Y|M_{pep})P(Y|M_{pro})}{P(Y)} \end{aligned} \quad (3.21)$$

$\alpha = \frac{P(M_{pep})P(M_{pro})}{P(M_{pep}, M_{pro})}$ is a class independent term and thus can be treated as normalization constant to ensure $\sum_i P(Y_i|M_{pep}, M_{pro}) = 1$.

3.5 Results

3.5.1 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**).

Peptide classifier positive set (**P1**)

MUSI (Kim *et al.*, 2011) is used to identify multiple binding specificities of the 864 unique peptides (sequence length less than 25 amino acids) belonging to 1238 SH3-peptide PPIs from the MINT database (Licata *et al.*, 2012). This resulted in three generic PWMs capturing major known SH3 domain binding motif classes RxxPxxP, PxxPxR, and PxxP (Figure 3.3).

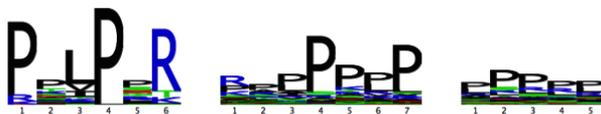


Figure 3.3: SH3 domain binding motifs in MINT database

All 864 peptides were scored using the three PWMs and only those with scores greater than the stringent p-value threshold of $1e - 05$ were retained. This filtering resulted in a set of 683 interactions. Further, interactions with missing feature information are removed thus resulting in a high confidence positive set of 628 SH3 domain-peptide mediated interactions.

Peptide classifier negative set (N1)

The negative dataset consists of randomly selected protein pairs with one member containing a SH3 domain and the other a 10 – 17 amino acid long randomly selected proteome sequence. Peptide sequences are scored using positive PWMs from the P1 dataset and only those with scores below the p-value threshold of 0.05 are retained (Figure 3.4).



Figure 3.4: Negative peptide set motif

Also, the protein pairs are not part of known interactions from the iRefIndex (version 13.0) database (Razick *et al.*, 2008). Positive (P1) and negative (N1) data sets are balanced with complete feature information.

Protein classifier positive set (P2)

5,795 pairwise yeast PPIs are retrieved from iRefIndex using its web interface iRefWeb (Turner *et al.*, 2010). iRefIndex consolidates PPIs from 10 major public databases and provides many filters to create a high confidence PPI set. The interactions retrieved from iRefWeb are all physical, experimental, from a single organism, supported by at least two publications and have a MI (MINT-Inspired) score ≥ 0.5 . A high confidence set of 5,215 interactions was created after removing instances with missing protein feature information.

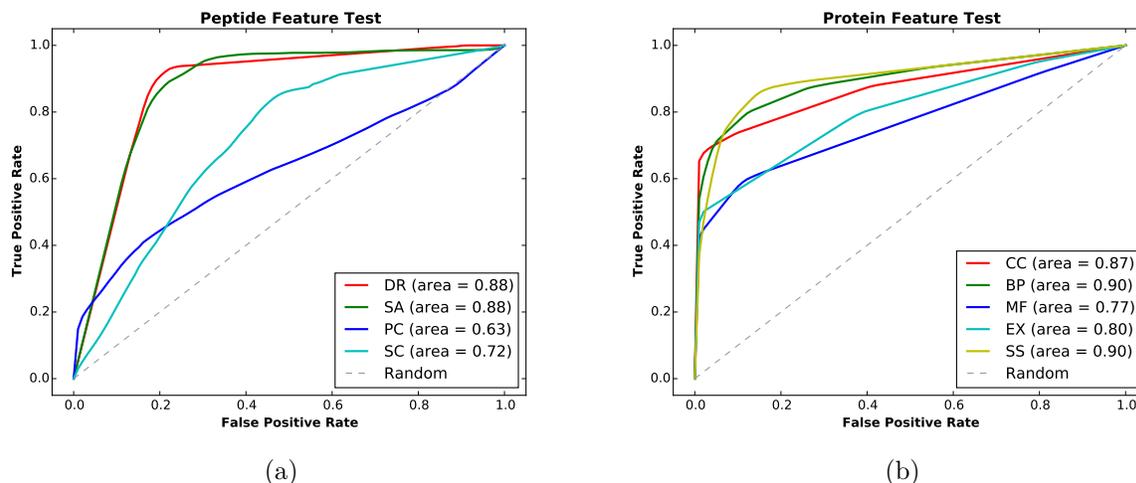


Figure 3.5: Prediction efficacy of individual (a) peptide features: disordered region (DR), surface accessibility (SA), peptide conservation (PC), structural contact (SC); and (b) protein features: cellular component (CC), biological process (BP), molecular function (MF), gene expression (EX), sequence signature (SS).

Protein classifier negative set (N2)

5,215 randomly selected protein pairs which are not known yeast interactions (over 117 thousand) from iRefIndex and have complete feature information.

3.5.2 Feature selection

Figure 3.5 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 3.6). 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).

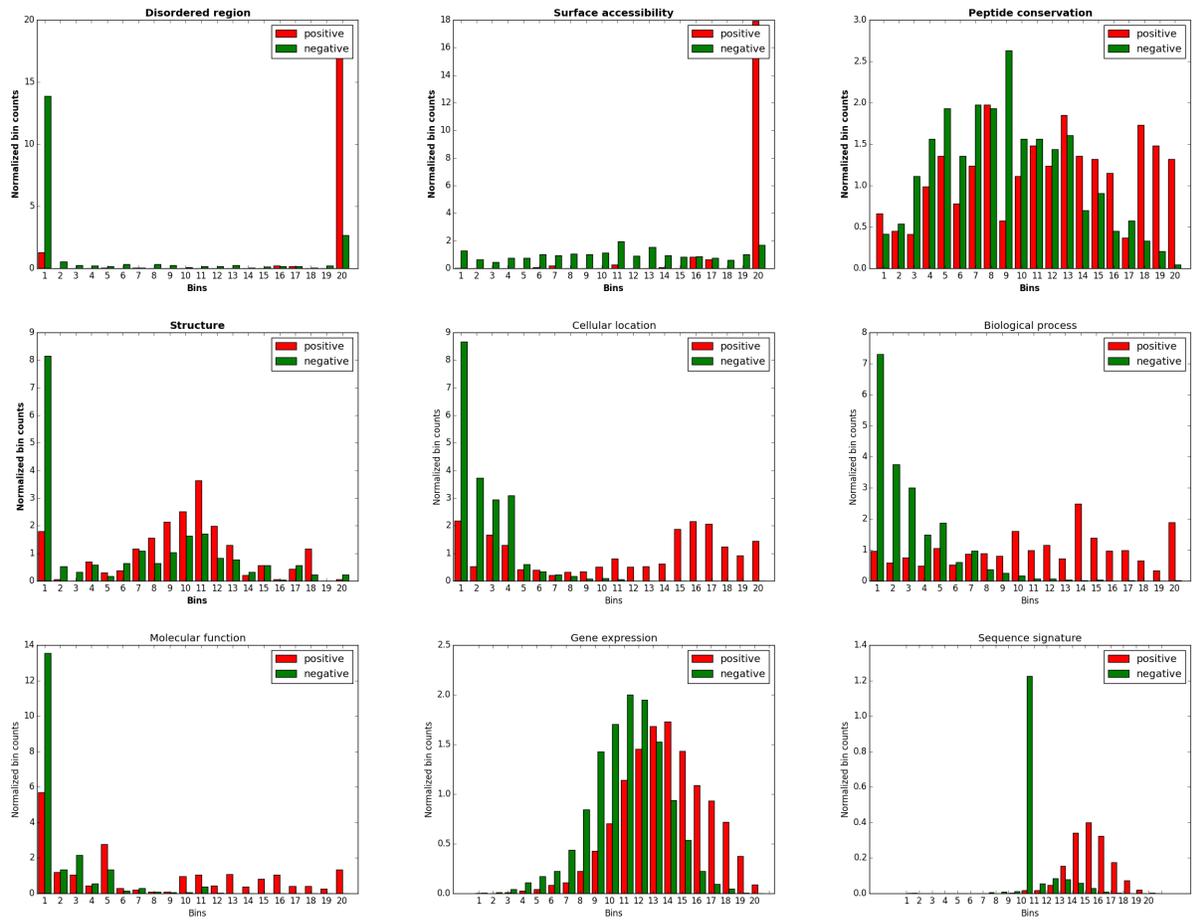


Figure 3.6: Distribution of positive and negative dataset score for peptide and protein features.

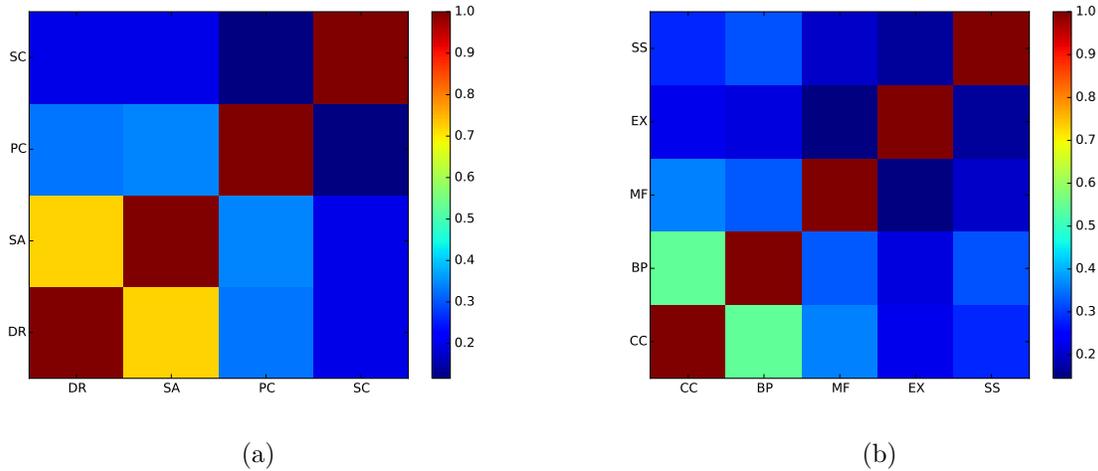


Figure 3.7: Maximal information coefficients for (a) Peptide feature set: disordered region (DR), surface accessibility (SA), peptide conservation (PC), structural contact (SC). (b) Protein feature set: cellular component (CC), biological process (BP), molecular function (MF), gene expression (EX), sequence signature (SS)

An important assumption behind a naïve Bayes classifier is that the features are independent of each other. Highly correlated features can negatively impact the performance of a naïve Bayes classifier (Ratanamahatana and Gunopulos, 2003). Mutual information is one of the methods for measuring dependence between two variables. Mutual information can capture both linear and non-linear relationships.

$$MI(X, Y) = \sum_{y \in Y} \sum_{x \in X} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \quad (3.22)$$

where $P(x, y)$ is the joint probability distribution and $P(x)$ and $P(y)$ are the marginal probability distributions. Mutual information score lies within the range $[0, \infty]$. Maximal information coefficient (MIC) technique calculates normalized mutual information scores within the range $[0, 1]$ where, a score of 0 indicates complete independence and 1 total dependence between two variables (Albanese *et al.*, 2012; Reshef *et al.*, 2011). Figure 3.7 shows the MICs for peptide and protein features. Peptide features: disordered region (DR) and surface accessibility (SA) and protein features: cellular component (CC) and biological process (BP) have MICs of 0.72 and 0.5 respectively.

To analyze the impact of correlation between DR and SA in peptide feature set and CC and BP in protein feature set on the performance of naïve Bayes classifier we built four different classifiers without one of the correlated features: (-)DR, (-)SA, (-)CC, and (-)BP and compared their performance with classifiers built using all features (ALL) using different statistics. Moreover, to

Model	AUROC	AUPRC	BRIER	F ₁ -score	MCC	ACC
ALL	0.94	0.93	0.09	0.86	0.73	0.86
(-) DR	0.93	0.92	0.09	0.64	0.45	0.68
(-) SA	0.94	0.93	0.09	0.65	0.46	0.69
(-) PC	0.92	0.9	0.1	0.84	0.69	0.84
(-) SC	0.92	0.92	0.1	0.87	0.73	0.86
(-) DR, SA	0.78	0.77	0.19	0.47	0.26	0.57
(-) DR, PC	0.91	0.88	0.1	0.69	0.47	0.71
(-) DR, SC	0.92	0.91	0.11	0.54	0.34	0.61
(-) SA, PC	0.93	0.91	0.1	0.72	0.52	0.74
(-) SA, SC	0.92	0.91	0.11	0.55	0.35	0.62
(-) PC, SC	0.9	0.91	0.11	0.86	0.72	0.86
(-) DR, SA, PC	0.72	0.68	0.21	0.33	0.0	0.5
(-) DR, SA, SC	0.64	0.7	0.23	0.48	0.27	0.57
(-) DR, PC, SC	0.88	0.9	0.12	0.33	0.0	0.5
(-) SA, PC, SC	0.88	0.9	0.11	0.33	0.0	0.5

Table 3.1: Peptide classifier: area under ROC curve (AUROC), area under precision-recall curve (AUPRC), Brier score (BRIER), F₁-score, Matthews correlation coefficient (MCC) and accuracy (ACC) for different models.

identify the feature subset which maximizes the performance of both classifiers we compared all possible feature combinations. We computed average area under ROC curve (AUROC), area under precision-recall curve (AUPRC), Brier score (BRIER), F₁-score, Matthews correlation coefficient (MCC) and accuracy (ACC) of 10-fold cross-validation protocol to determine the performance of different models. The peptide classifier was trained and tested using P1 & N1 datasets and the protein classifier using P2 & N2. F₁-score, MCC and ACC are reported at threshold score ≥ 0.9 . All measures except the Brier score are directly proportional to performance i.e. the higher the score for a model, the better the performance. On the other hand, the lower the Brier score for a model, the better the performance. Except MCC, which lies within the range $[-1, 1]$, other measures are within $[0, 1]$ range. It is clear from the Tables 3.1 and 3.2 that removing any of the individual features or any of the combinations do not improve the performance of either classifier. Even removing one of the correlated features does not improve the performance. For the peptide classifier, F₁-score, MCC, and ACC drop sharply for (-)DR and (-)SA models. Similarly, for the protein classifier, the performance degrades when either BP or CC are removed.

3.5.3 Model evaluation

Blind validation protocols is used to assess the predictive power of peptide M_{pep} and protein M_{pro} naïve Bayes classifiers. The majority of interactions in the P1 dataset are from two peptide array

Model	AUROC	AUPRC	BRIER	F ₁ -score	MCC	ACC
ALL	0.98	0.98	0.06	0.9	0.81	0.9
(-) CC	0.97	0.98	0.06	0.89	0.8	0.89
(-) BP	0.97	0.98	0.06	0.89	0.8	0.89
(-) MF	0.97	0.98	0.06	0.9	0.81	0.9
(-) EX	0.97	0.98	0.07	0.89	0.8	0.89
(-) SS	0.95	0.96	0.08	0.88	0.78	0.88
(-) CC, BP	0.96	0.97	0.07	0.84	0.72	0.84
(-) CC, MF	0.97	0.97	0.07	0.88	0.78	0.88
(-) CC, EX	0.97	0.97	0.07	0.87	0.76	0.87
(-) CC, SS	0.94	0.95	0.09	0.85	0.73	0.85
(-) BP, MF	0.97	0.97	0.07	0.88	0.78	0.88
(-) BP, EX	0.97	0.97	0.07	0.86	0.76	0.87
(-) BP, SS	0.93	0.95	0.09	0.86	0.76	0.87
(-) MF, EX	0.97	0.98	0.07	0.88	0.78	0.88
(-) MF, SS	0.94	0.96	0.09	0.87	0.76	0.87
(-) EX, SS	0.93	0.95	0.09	0.86	0.75	0.86
(-) CC, BP, MF	0.94	0.95	0.09	0.79	0.64	0.79
(-) CC, BP, EX	0.94	0.95	0.09	0.82	0.68	0.82
(-) CC, BP, SS	0.88	0.91	0.12	0.81	0.67	0.81
(-) CC, MF, EX	0.96	0.97	0.08	0.84	0.72	0.85
(-) CC, MF, SS	0.93	0.94	0.1	0.82	0.69	0.82
(-) CC, EX, SS	0.91	0.94	0.11	0.79	0.65	0.8
(-) BP, MF, EX	0.96	0.97	0.07	0.84	0.73	0.85
(-) BP, MF, SS	0.91	0.94	0.1	0.85	0.73	0.85
(-) BP, EX, SS	0.91	0.93	0.11	0.84	0.72	0.85
(-) MF, EX, SS	0.92	0.94	0.1	0.85	0.73	0.85
(-) CC, BP, MF, EX	0.9	0.92	0.12	0.69	0.51	0.71
(-) CC, BP, MF, SS	0.8	0.86	0.16	0.72	0.56	0.74
(-) CC, BP, EX, SS	0.77	0.84	0.18	0.66	0.48	0.69
(-) CC, MF, EX, SS	0.9	0.92	0.12	0.77	0.62	0.78
(-) BP, MF, EX, SS	0.87	0.91	0.12	0.8	0.67	0.81

Table 3.2: Protein classifier: area under ROC curve (AUROC), area under precision-recall curve (AUPRC), Brier score (BRIER), F₁-score, Matthews correlation coefficient (MCC) and accuracy (ACC) for different models.

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 3.3: 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 ≥ 0.9 , accuracy (ACC), F₁-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 F₁-score are reported at threshold score ≥ 0.9 .

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 *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 3.3. The AUROC for peptide classifier 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

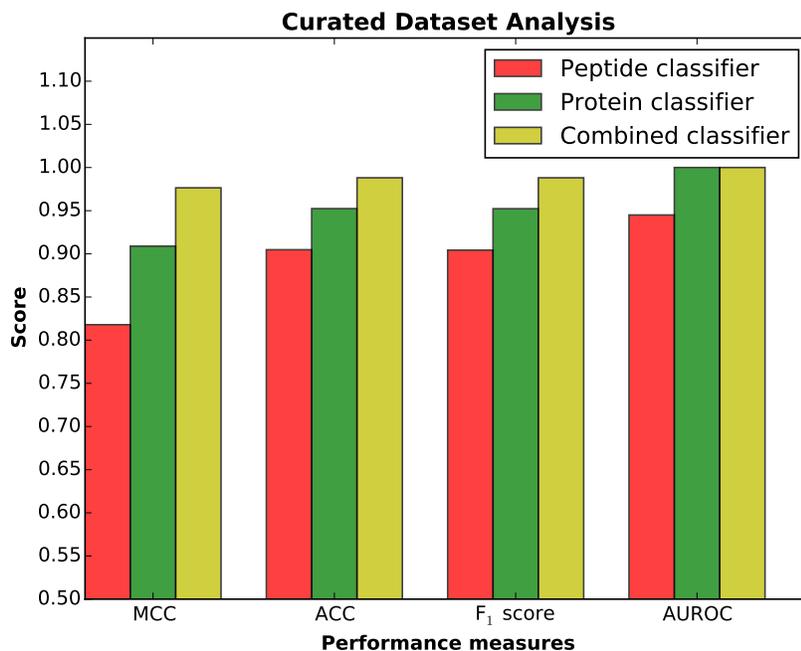


Figure 3.8: Performance of peptide, protein, and combined classifiers on the curated SH3 domain mediated PPI set. (Note: small size of curated validation dataset prevents the variance from being estimated.)

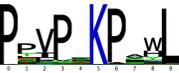
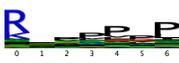
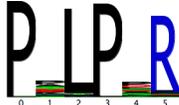
with curated set) and P2 & N2 datasets, respectively. A randomized negative test set is created in the same way as N1. Results from different statistical measures are summarized in Figure 3.8. The combined classifier outperforms both the peptide and protein classifiers on the curated set.

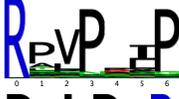
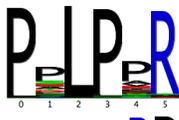
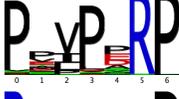
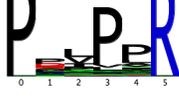
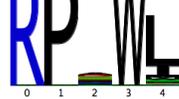
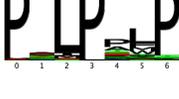
3.5.4 SH3 domain mediated PPI predictions

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 3.4.1 (Table 3.4). 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. 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 ≥ 0.9 . Biological pathway enrichment (KEGG (Kanehisa, 2002) and Reactome (Fabregat *et al.*, 2015)) 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 Alan, 2000) (Table 3.5). 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 3.6). These examples support our results and suggest that our predicted interactions are biologically relevant.

Table 3.4: List of yeast SH3 domains from Tonikian *et al.* (2009) and their binding motifs (trimmed) with significant amino acid positions within those motifs.

Domain id	Phage logo	Significant positions
P11710		0, 1, 2, 3, 4
P15891		0, 2, 3, 5, 6, 8, 9
P29366-1		0, 2, 4, 5
P29366-2_PXXP		0, 1, 2, 3, 5, 6
P32790-1_classI		0, 4, 6
P32790-2_classII		0, 1, 2, 3, 4, 5, 6, 8
P32790-3		0, 1, 2, 3, 5
P32793		0, 2, 3, 5
P36006		0, 4, 5, 8

P38041		0, 2, 3, 4, 5, 6
P38753		0, 3, 5
P38822-1		0, 2, 3, 4, 5, 6, 7
P38822-2		3, 4
P39743_ClassI		0, 1, 2, 3, 6
P39743_ClassII		0, 2, 3, 5
P39969		0, 3, 4, 5, 6
P40073		0, 1, 3, 4
P43603		0, 2, 3, 5
P47068_classIIcombined		0, 2, 3, 5, 6
P53281_classI		0, 1, 5, 6
P53281_classII		0, 3, 5
P80667_classIIA		0, 2, 3, 5, 6, 7
P80667_classIIB		0, 1, 3, 4
Q04439		0, 4, 5, 8
Q05080		0, 2, 3, 6

P-value	Term ID	Term name	Proteins
0.00113	KEGG:04011	MAPK signaling pathway - yeast	P24583, P32917, Q03497, P08018, P41832, P32491
0.0375	KEGG:04144	Endocytosis	P34216, P25604, P35197, P40343, Q12446
0.00077	KEGG:04070	Phosphatidylinositol signaling system	P24583, P34756, P50942, Q12271
0.0169	KEGG:00562	Inositol phosphate metabolism	P34756, P50942, Q12271
0.00698	REAC:5733237	Innate Immune System	Q03306, Q03497, P08018, Q12236, Q12446, P32491
0.00316	REAC:5733336	Fc epsilon receptor (FCERI) signaling	Q03306, P08018, Q12236, P32491
0.00000197	REAC:5733138	Signal Transduction	P24583, Q03306, Q04739, Q03497, P40450, P32521, P41832, Q12236, Q12446, P48582, P32873, P32491
2.97E-09	REAC:5733143	Signaling by Rho GTPases	P24583, Q03306, Q03497, P40450, P32521, P41832, Q12236, Q12446, P48582, P32873
1.65E-08	REAC:5733142	RHO GTPase Effectors	P24583, Q03306, Q03497, P40450, P41832, Q12236, Q12446, P48582
0.05	REAC:5733141	RHO GTPases activate PKNs	P24583, Q03306, Q12236
0.0337	REAC:5733628	Signaling by ERBB4	Q03306, Q12236, P32491
0.0337	REAC:5733629	Signaling by SCF-KIT	Q03306, Q12236, P32491
0.0314	REAC:5733228	Signalling by NGF	Q03306, P32521, Q12236, P32873, P32491
0.0123	REAC:5733461	Costimulation by the CD28 family	Q03306, Q03497, Q12236
0.0123	REAC:5733460	CD28 co-stimulation	Q03306, Q03497, Q12236

Table 3.5: Enrichment analysis of predicted high confidence interactors.

P-value	Term ID	Term name	Proteins
0.04	REAC:5733141	RHO GTPases activate PKNs	Q03306, Q12236
0.000817	REAC:5733234	Signaling by ERBB2	Q03306, Q12236, P32491
0.000817	REAC:5733232	Signaling by EGFR	Q03306, Q12236, P32491
0.000817	REAC:5733230	Signaling by PDGF	Q03306, Q12236, P32491
0.000161	REAC:5733628	Signaling by ERBB4	Q03306, Q12236, P32491
0.000344	REAC:5733311	VEGFA-VEGFR2 Pathway	Q03306, Q12236, P32491
0.00337	REAC:5733625	PIP3 activates AKT signaling	Q03306, Q12236
0.000473	REAC:5733336	Fc epsilon receptor (FCERI) signaling	Q03306, Q12236, P32491
0.00376	REAC:5733190	IGF1R signaling cascade	Q03306, Q12236, P32491
0.00337	REAC:5733185	Activation of AKT2	Q03306, Q12236
0.000817	REAC:5733242	Signaling by FGFR	Q03306, Q12236, P32491
0.00337	REAC:5733405	Downstream TCR signaling	Q03306, Q12236
0.00337	REAC:5733635	CD28 dependent PI3K/Akt signaling	Q03306, Q12236
0.000161	REAC:5733629	Signaling by SCF-KIT	Q03306, Q12236, P32491
0.00376	REAC:5733187	IRS-mediated signalling	Q03306, Q12236, P32491

Table 3.6: Enrichment analysis of predicted MYO3 interactors.

3.6 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>.

Chapter 4

Predicting *in-vivo* SH3 domain mediated protein interactions in human

This chapter is the basis of a manuscript which will be submitted for publication: Jain, S., Teyra, J., Huang, H., Sidhu SS and Bader, GD. (2018), Predicting in-vivo relevant SH3 domain mediated protein-protein interactions in human.

Author contributions: I collected the data, developed and implemented the method and performed the analyses. Gary D. Bader supervised and advised this project. Joan Teyra, Haiming Huang and Sachdev S. Sidhu performed the phage display experiments and gave us early data access.

4.1 Abstract

SH3 domains mediate many intracellular signaling processes and are critical for cell functioning. These domains bind to proline rich regions which can be identified using high-throughput experimental screens such as phage display. SH3 binding motifs identified by phage display can be used to computationally predict domain mediated protein-protein interactions. The existing landscape of computational approaches for predicting protein interactions is either limited by their inability to predict peptide recognition module mediated interactions or do not consider many known constraints governing these interactions. A novel method of predicting SH3 domain-peptide mediated

protein-protein interactions in humans using phage display data is presented. This method builds upon our previously published work of combining multiple binding site and full length protein features using naïve Bayes models for predicting PRM mediated interactions. In this work, we present a novel algorithm for predicting protein interactions using network topology and show that it outperforms existing approaches. We have also extended the semi-supervised training framework of multinomial naïve Bayes classifier developed for text classification to Gaussian naïve Bayes models for PPI prediction to overcome limited availability of labeled data. Domain-Motif Mediated Interaction Prediction (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-human>. The DoMo-Pred command line tool is implemented using Python 2.7 and C++.

4.2 Introduction

Protein-protein interactions (PPIs) are the building blocks of complex cellular processes. They are critical for cell functioning and their knowledge help us better understand cellular dynamics. They are the physical associations between protein pairs in a specific biological context (Jain and Bader, 2016). Experimental methods for detecting PPIs can be classified into two broad categories: small scale techniques like affinity chromatography, co-immunoprecipitation, affinity blotting (Phizicky and Fields, 1995) and high-throughput techniques such as mass-spectrometry combined with affinity-purification, phage display, yeast two-hybrid, and next-generation sequencing (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). PPIs detected using physical techniques are largely accurate but it is difficult to adapt these techniques for whole genome analysis. On the other hand, high-throughput methods can be used to detect PPIs on genomic scale but they are resource intensive and also suffer from the problem of false positives. This led to the development of computational approaches using experimental data to predict high confidence novel interactions on genome-wide scale (Skrabanek *et al.*, 2008). Computational approaches can model complex systems are inexpensive to apply to whole genome analysis and this motivates their further development (Jain and Bader, 2016).

Domains are the functional units in proteins. A protein can have multiple domains of same or different families. Domains mediate PPIs by physically interacting with other domains or peptide sequences. We are interested in interactions involving peptide recognition modules (PRMs), in particular Src homology 3 (SH3), which are important in many cellular signaling processes. These

domains bind to linear sequence motifs of 10 – 15 amino acids within proteins (Pawson and Nash, 2003). SH3 domains are approximately 60 amino acids long and fold into a beta-barrel structure composed of five to six anti-parallel beta strands. The SH3 domain has a flat, hydrophobic surface which consists of three shallow pockets with conserved aromatic residues (Zafra-Ruano and Luque, 2012). They often bind to proline-rich regions with PxxP as the minimal consensus target sequence. Based on the positioning of positively charged residues around the core PxxP motif, SH3 domains are divided into two classes. Class I SH3 domains bind to [R/K]xxPxxP and class II bind to PxxPx[R/K] motifs (Mayer, 2001; Saksela and Permi, 2012). However, studies have shown that class boundaries are not rigid and SH3 domains can bind to peptide sequences without the core PxxP motif like PxxDY, RxxPxxxP, PxxxPR, and PxRPxR (Carducci *et al.*, 2012; Mongiovi *et al.*, 1999; Tian *et al.*, 2006; Moncalián *et al.*, 2006). They can also bind to proline-free regions containing combinations of arginine, lysine, and/or tyrosine residues such as RxxK and RKxxYxxY (Tong *et al.*, 2002; Carducci *et al.*, 2012). 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.

This work focuses on developing a computational model to predict SH3 domain mediated PPIs in humans using peptides identified from phage display experiments. As discussed in our previous work Jain and Bader (2016), the straightforward approach of constructing a position weight matrices (PWMs) from phage peptides and scanning the whole proteome for potential targets lacks the contextual information needed to reduce false positives. For example, the predicted binding site might not be accessible or the two proteins might never be co-localized in the cell. Contextual information can be grouped into two categories, one for predicting high confidence binding site (peptide features) and the other working with full length proteins for predicting PPIs in general (protein features). Efforts have been made to predict PPIs by combining different full length protein evidence sources but they cannot be used to predict PRM mediated PPIs (Jansen *et al.*, 2003; Rhodes *et al.*, 2005; Li *et al.*, 2008; Zhang *et al.*, 2012b). Lam *et al.* (2010) combined comparative and structural genomic features with phage display data to predict high confidence binding sites but they failed to incorporate many cellular constraints like interacting proteins must be in close proximity and should be part of same process (Jain and Bader, 2016). Chen *et al.* (2015) combined

limited number of peptide and protein features for predicting PRM mediated PPIs in humans. Apart from using an outdated protein feature set from the earlier works in the field of ensemble PPI prediction by Jansen *et al.* (2003) their method is also not compatible with high-throughput binding peptide data, such as from phage display (Jain and Bader, 2016). Recently, we published DoMo-Pred, a novel method for predicting SH3 domain-peptide mediated PPIs in yeast by combining large number of peptide and protein features using multiple Bayesian models (Jain and Bader, 2016). Here, we extend DoMo-Pred to include protein expression and network topology features. We developed a novel algorithm for predicting PPIs using network topology and extended the semi-supervised training regime of multinomial naïve Bayes classifier developed for text classification to Gaussian naïve Bayes models for PPI prediction.

4.3 Approach

PPIs in general are governed by many cellular and structural constraints. For example, two proteins can physically interact only if they are expressed at the same time and are co-localized. Also, a true binding site should be accessible and is more likely to be conserved when compared to random sequence. Therefore, the potential binding sites identified by scanning the whole proteome using SH3 domain specific PWMs are filtered using a combination of peptide (or binding site) and full length protein features. Peptide features: disorder, surface accessibility, peptide conservation and structural contact are combined using a semi-supervised naïve Bayes model to score domain binding sites. Another semi-supervised naïve Bayes model is used to combine protein features: cellular location, biological process, molecular function, gene expression, sequence signature, protein expression, and network topology to score the same PWM targets. Scores from both the models are combined using Bayes theorem. Figure 4.1 shows the workflow of our PRM mediated PPI prediction pipeline.

4.4 Methods

4.4.1 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 is the size of alphabet (20 amino acids for protein sequences) and n is the motif length. PWMs contain a weight for each alphabet symbol i at each position j in

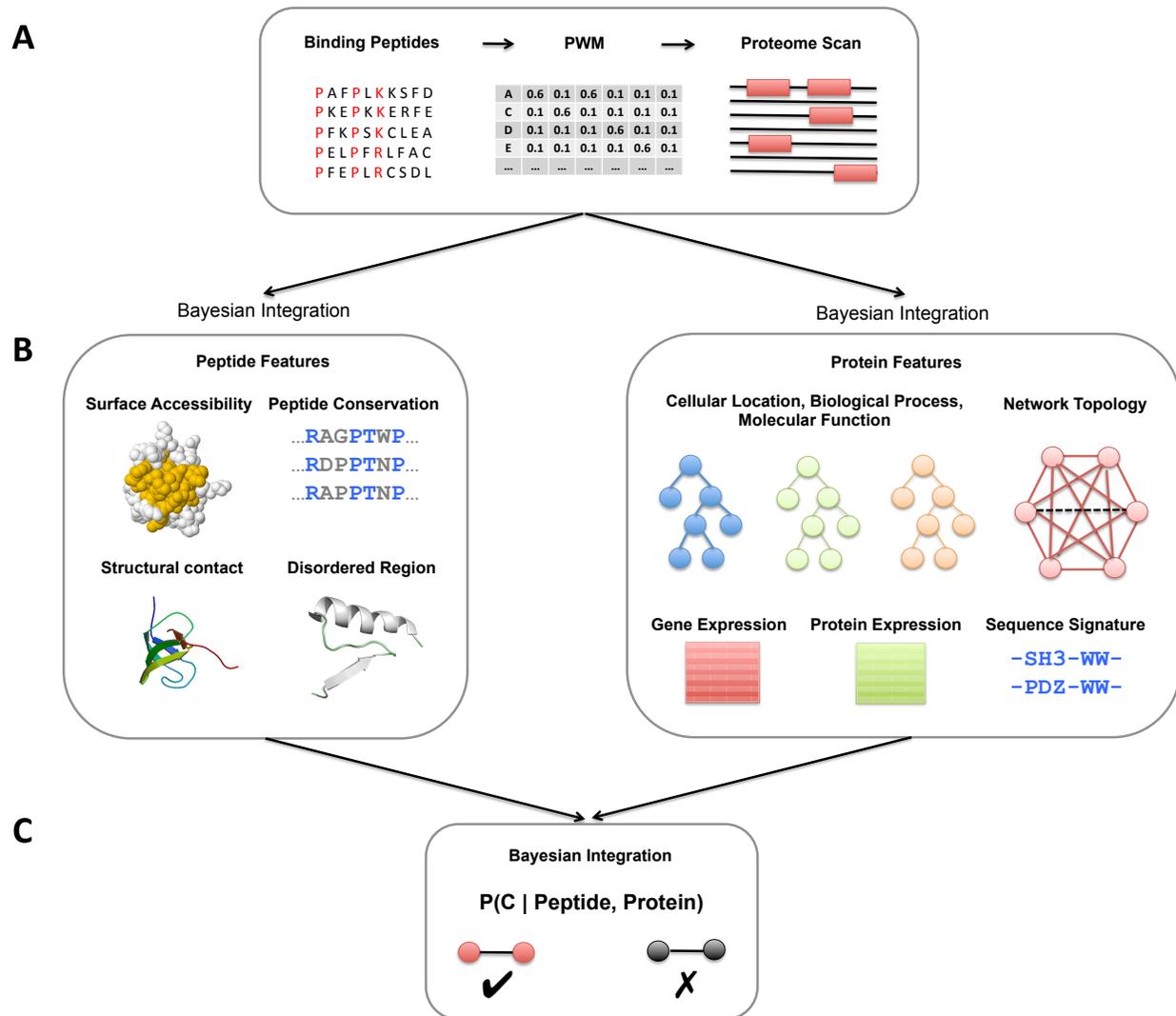


Figure 4.1: Work flow of PRM mediated PPI prediction pipeline. (A) Proteome is scanned using a PWM built using experimentally derived binding peptides (e.g. from phage display) of a given SH3 domain for potential interactors. (B) Separate naïve Bayes classifiers for peptide and protein features. (C) Integration of classifiers for predicting interacting and non-interacting protein pairs.

the motif as a log-odds score of a probabilistic model against the background probability distribution of amino acids in proteome ($B(i)$) (Pizzi *et al.*, 2011). For a given set of peptides binding to a SH3 domain the PWM for that SH3 domain is constructed using the empirical count of amino acid i at position j .

$$M(i, j) = \log \frac{\text{count}(i, j)}{N \times B(i)} \quad (4.1)$$

where N is the count of all the amino acids at position j . PWMs are trimmed at either ends to improve the signal of core motif by removing low information content positions. Trimmed PWMs are then used to scan a protein sequence to find matches of the weighted pattern above a p-value threshold score. Not all amino acid positions within a motif are significant. For example, in class 1 SH3 binding motif [R/K]xxPxxP, positions 1, 4, and 7 are more significant than others. Amino acid positions with information content ≥ 0.5 within the trimmed PWMs are identified as significant. These significant amino acid positions are used in calculation of disordered region, surface accessibility, and peptide conservation scores (Jain and Bader, 2016).

4.4.2 Peptide features

As discussed in Jain and Bader (2016), disordered region, surface accessibility, peptide conservation, and structural contact features are used to assess the confidence of binding site predicted from PWM scan. SH3 domains often bind to proline rich regions containing PXXP, [R/K]xxPxxP, PxxPx[R/K] motifs. As presence of prolines disrupts the secondary structure of a protein therefore, a true SH3 domain binding site is more likely to be disordered. Disordered (DR) score is defined as the fraction of significant amino acid residues predicted being disordered by DISOPRED (Ward *et al.*, 2004) tool as shown in equation 4.2. Also, any SH3 domain binding peptide should be accessible i.e. it should not be buried inside the folded protein. Equation 4.3 estimates the surface accessibility (SA) score of a peptide using the relative solvent accessibility (RSA) of significant amino acids as estimated by SABLE (Adamczak *et al.*, 2004).

$$\text{DR} = \frac{\sum_i p_i}{N} = \begin{cases} 1 & \text{if amino acid } i \text{ is disordered} \\ 0 & \text{otherwise} \end{cases} \quad (4.2)$$

$$SA = \frac{\sum_i p_i}{N} = \begin{cases} 1 & \text{if } RSA \geq 25\% \\ 0 & \text{otherwise} \end{cases} \quad (4.3)$$

where p_i is the disordered or surface accessible i^{th} amino acid residue and N is the number of significant amino acids in the binding site.

Biologically relevant peptides are more likely to be conserved across different organisms. For measuring the conservation of peptides, orthologs of human protein sequences in *P. troglodytes*, *M. musculus*, *M. putorius*, *L. africana*, *O. cuniculus*, *M. lucifugus*, *M. domestica*, *S. scrofa*, *S. harrisii*, *X. tropicalis*, *L. chalumnae*, *O. niloticus*, *T. chinensis*, *X. maculatus*, *T. rubripes*, *M. gallopavo*, *T. guttata*, *O. latipes* and *O. anatinus* (organisms with different taxonomic order and $\geq 10,000$ human orthologs) are identified using INPARANOID (Remm *et al.*, 2001) and aligned with MAFFT (Kato *et al.*, 2002). 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 and converted into the peptide conservation (PC) score,

$$PC = \phi\left(\frac{\sum_i con(i)}{N}\right) \quad (4.4)$$

where $con(i)$ is the AL2CO conservation z-score of i^{th} amino acid residue of SH3 domain binding peptide and ϕ is the normal cumulative distribution function. N is the number of significant amino acids in the binding site.

2-D contact maps of eight human SH3-peptide co-complex PDB structures (1AGZ, 1BBZ, 1IO6, 3UA7, 4CC2, 4EIK, 4J9F, 4LN2) are used as base models. Query domain and peptide sequences are aligned with all base models to calculate the contact distance between aligned residues. Then the structural contact (SC) score is defined as maximum average contact area

$$SC = \max_j \frac{\sum_i c_{ij}}{N} \quad (4.5)$$

4.4.3 Protein features

As discussed in Jain and Bader (2016), full length protein features: cellular location, biological process, molecular function, gene expression, and sequence signature are used to independently assess the interaction potential of proteins involved in PRM mediated PPIs. True physical interactions are more likely to occur between co-localized proteins belonging to same biological process with

similar molecular function. The Gene Ontology (GO) describes cellular compartment (CC), biological process (BP), and molecular function (MF) of proteins using a set of hierarchical terms and the relationship between these terms can be quantified by topological clustering semantic similarity (TCSS) measure. For a given protein pair, CC, BP, and MF scores are defined as,

$$CC = TCSS(a, b, ontology = C, cutoff = 3.4) \quad (4.6)$$

$$BP = TCSS(a, b, ontology = P, cutoff = 4.0) \quad (4.7)$$

$$MF = TCSS(a, b, ontology = F, cutoff = 3.6) \quad (4.8)$$

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

Cell systems are optimized to co-express functionally related genes and if the genes are functionally related, their protein products are more likely to physically interact. Therefore, for a given protein pair, correlation coefficients from 140 human gene expression datasets from GeneMANIA (Warde-Farley *et al.*, 2010) are combined using Fisher’s z transformation (Faller, 1981; Jain and Bader, 2010) and the result is normalized to lie within the range $(0, 1)$ using a logistic function as shown below,

$$EX = \frac{1}{1 + e^{-5\bar{r}}} \quad (4.9)$$

$$\bar{r} = \frac{e^{2\bar{z}} + 1}{e^{2\bar{z}} - 1} \quad (4.10)$$

$$\bar{z} = N^{-1} \sum_{i=1}^N \frac{1}{2} \ln \left(\frac{1 + r_i}{1 - r_i} \right) \quad (4.11)$$

where N is the number of gene expression datasets and r_i is the Pearson correlation of the i^{th} dataset.

Sequence composition of known protein interactions can be used as an indicator for predicting novel PPIs. Information content of co-occurring InterPro (Apweiler *et al.*, 2001) sequence signatures is used for scoring interactions, as described by Jain and Bader (2016); Sprinzak and Margalit (2001).

$$SS = \sum_{ij} -\log_2 \left(\frac{p_{ij}}{p_i p_j} \right) \quad (4.12)$$

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.

4.4.4 Protein expression

Recently, Kim and co-workers published a draft map of human proteome with expression profiles of 17,294 protein coding genes. They did in-depth proteomic profiling of 30 histologically normal human samples, including 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells using high-resolution Fourier-transform mass spectrometry (Kim *et al.*, 2014). As discussed previously, gene expression profiles are used to assess the likelihood that protein products of co-expressed genes also interact physically. With the availability of protein expression data, we hypothesized and proved that protein expression pattern is a better predictor of protein-protein interactions than gene expression profiles (Kim *et al.*, 2014). For a given protein pair we correlate their normalized spectral count profiles across all 30 cells/tissues in our data using Pearson correlation. The Pearson correlation coefficients are normalized to lie within the range $(0, 1)$ using a logistic function,

$$\text{PX} = \frac{1}{1 + e^{-5r}} \quad (4.13)$$

where r is the Pearson correlation coefficient of spectral count profiles for a given protein pair.

4.4.5 Network Topology

Networks or graphs provide a powerful computational framework to represent and analyze complex biological systems. They include transcriptional regulatory networks, metabolic networks, signal transduction networks, and PPI networks. Development of high-throughput technologies for detecting protein interactions have created large-scale PPI networks where, nodes correspond to proteins and undirected edges represent physical interactions amongst them. Much work has been done in defining the relationship between the PPI network topology and biological function (Sharan *et al.*, 2007). We are interested in predicting PPIs using network topology. Goldberg and Roth (2003) exploited the neighborhood cohesiveness property of small-world networks to assess confidence of PPIs in high-throughput experimental network. They showed that true edges in a PPI have higher neighborhood cohesiveness as compared to false edges. Conversely, an edge in a PPI can be qualified to be true positive if it shows higher degree of neighborhood cohesiveness. Bader *et al.* (2004)

proposed that interacting proteins sharing interactors are more likely to be biologically relevant. Yu *et al.* (2006) predicted interactions in protein networks by completing defective cliques. They used the matrix model interpretation of the results from large-scale experiments, which states that two proteins interacting with the same protein clusters are likely to interact with each other. Yu *et al.*'s method is biased towards complexes and Goldberg's neighborhood cohesiveness approach does not consider graph properties such as edge density, edge connectivity, mean degree and others.

Network TOPology (NTOP) Algorithm

We propose a machine learning based algorithm for predicting PPIs using topological information of known networks. PPI networks have the properties of small-world network: presence of cliques or near-cliques, over-abundance of hub nodes, and random connectivity at longer distances (Bader *et al.*, 2004). Therefore, for a given a graph $G = (V, E)$ we compute the following small-world network descriptors:

- **Edge density** is defined as the fraction of edges (interactions) and all possible edges in a graph. For undirected PPI networks edge density (ED) is calculated as,

$$ED = \frac{2|E|}{|V|(|V| - 1)} \quad (4.14)$$

- **Mean degree** is defined as the fraction of twice the number of edges in G and the total number of vertices in G ,

$$MD = \frac{2|E|}{|V|} \quad (4.15)$$

- **Edge connectivity** of an undirected graph G is defined as the minimum number of edges (E) whose deletion from G disconnects it. A graph is said to be k -edge-connected if it remains connected when fewer than k edges are removed.

- **Transitivity** is the measure of clustering in a graph. Small-world networks tend to have tightly knit groups of nodes. Therefore, true interactions in a PPI network are more likely to be present in densely connected regions of the network. Transitivity is calculated by counting the number of triangles in G ,

$$TA = \frac{3 \times \text{number of triangles in the graph}}{\text{number of connected triples in the graph}} \quad (4.16)$$

- **Mutual clustering coefficient** is a measure of neighborhood cohesiveness around the edge of interest in G . Watts and Strogatz (1998) showed that small-world networks have high clustering coefficients. Therefore, clustering coefficients of true interactions in a PPI network must be greater than that of randomly established links (Goldberg and Roth, 2003). One of the methods for computing mutual clustering coefficient as defined by Goldberg and Roth (2003) is,

$$E_{ab} = \frac{|N_a \cap N_b|}{\min(|N_a|, |N_b|)} \quad (4.17)$$

where N_a and N_b are the neighbors of nodes a and b respectively, in G .

Feature vector: For a given pair of proteins A & B in a PPI network G a feature vector is created as,

1. Compute the neighbors N_A and N_B of nodes A & B respectively, in graph G .
2. Compute the sub-graphs S_{N_A} , S_{N_B} and $S_{N_A \cup N_B}$ from graph G .
3. For each sub-graph in step 2 compute edge density, mean degree, edge connectivity, and transitivity. Also, compute mutual clustering coefficient for sub-graph $S_{N_A \cup N_B}$.
4. Concatenate the graph features into a single vector of size 13 to be used by a logistic regression model in next step.

Logistic regression model for PPI prediction learns the function of the form $f : X \rightarrow P(Y|X)$ where $X = (X_1, X_2, \dots, X_n)$ is a vector containing discrete or continuous variables (features) and $Y = k$ is "interacting" or "non-interacting" class (Mitchell, 1997). Logistic regression uses a sigmoid function to parameterize the probability distribution $P(Y|X)$. The parameterized form used by logistic regression classifier is,

$$\arg \max_k P(Y = k|X) = \arg \max_k \frac{1}{1 + e^{-(\Theta_o + \sum_{i=1}^n \Theta_i X_i)}} \quad (4.18)$$

where the model parameters Θ_i are learned from training set of 9,29 high confidence human protein interactions from iRefIndex (Razick *et al.*, 2008) downloaded through its web interface iRefWeb (Turner *et al.*, 2010). These interactions are all pairwise, physical, experimental, from a single organism and have a MI (MINT-Inspired) score ≥ 0.8 . To remove any bias in the dataset due to the presence of hub proteins, only single randomly selected interaction involving the hub protein

Method	MCC	ACC	F ₁ -score
NTOP	0.57	0.78	0.78
Mutual clustering coefficient	0.45	0.68	0.65
Defective cliques	0.46	0.70	0.68

Table 4.1: Matthews correlation coefficient (MCC), accuracy (ACC) and F₁-score for NTOP, mutual clustering coefficient proposed by Goldberg and Roth (2003) and predicting PPIs by completing defective cliques proposed by Yu *et al.* (2006). Best value for each measure is shown in bold.

is retained. A negative interaction (or non-interaction) set is built using an equal number of low confidence interactions from iRefWeb. These interactions have MI (MINT-Inspired) score < 0.3 . Similar to the positive set any bias due to hub proteins is removed.

Feature selection and model evaluation

The recursive feature elimination (RFE) algorithm is used to select an optimal subset of features for the logistic regression model. RFE is an iterative procedure of ranking the features and eliminating smallest ranked feature in each iteration. It is a greedy approach and is an instance of backward feature elimination (Guyon *et al.*, 2002). RFE along with 10-fold cross-validation was used to rank the individual features and find a subset which optimize the classifier performance. All the 13 features when used together maximizes the performance of logistic regression classifier.

NTOP achieved an accuracy of 0.71, F₁-score of 0.71 and Matthews correlation coefficient score of 0.43 in a 10-fold cross-validation protocol. We compared the performance of NTOP algorithm for predicting PPIs to that of mutual clustering coefficient approach proposed by Goldberg and Roth (2003) and completing defective cliques by Yu *et al.* (2006) on a blind set (no overlap between training and test sets) of 544 high confidence human protein interactions from iRefWeb (P2) and equal number of experimentally verified non-interaction set from Negatome (N2) (Blohm *et al.*, 2013) (refer to section 3.7 for details). Table 4.1 summarizes the result of blind validation test. NTOP outperforms both mutual clustering coefficient and completing defective cliques techniques. For a given protein pair network topology (NT) score is the probability estimate of it belonging to the interacting class.

4.4.6 Semi-supervised training of naïve Bayes model

As discussed earlier, PRM mediated PPI prediction algorithm computes a set of features involving domain-peptide and another using full length proteins. Assuming that peptide and protein features

are independent from each other, two separate naïve Bayes models M_{pep} for peptide features and M_{pro} for protein features are built to independently estimate the probability that a given protein pair interacts. A naïve Bayes model simplifies this problem by assuming independence between different types of biological evidence. For a protein pair described by a set of features ($X_i = X_1, X_2, \dots, X_n$) a naïve Bayes PPI prediction model is defined as,

$$\begin{aligned} \arg \max_k P(Y = k|X_i) &= \arg \max_k \frac{P(X_i|Y = k)P(Y = k)}{P(X_i)} \\ &= \arg \max_k P(Y = k) \prod_i P(X_i|Y = k) \end{aligned} \quad (4.19)$$

where $P(Y = k)$ is the class prior probability and $P(X_i|Y)$ is the class-conditional probability. Class priors are estimated by treating $P(Y = k)$ as a multinomial distribution $P(Y) = \Pi_k$ as there are only two classes $k \in \{\text{interacting, non-interacting}\}$. All peptide and protein features X_i are continuous therefore the probability distributions $P(X_i|Y)$ are modeled as Gaussian $P(X_i|Y = k) = \mathcal{N}(X_i; \mu_{ik}, \sigma_{ik})$. Putting it all together, the Gaussian naïve Bayes (GNB) model is defined as,

$$\arg \max_k P(Y = k|X_i) = \arg \max_k \Pi_k \prod_i \mathcal{N}(X_i; \mu_{ik}, \sigma_{ik}) \quad (4.20)$$

where $\mathcal{N}(X_i; \mu_{ik}, \sigma_{ik})$ is the probability density of a Gaussian (or normal) distribution,

$$\mathcal{N}(X_i; \mu_{ik}, \sigma_{ik}) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp^{-\frac{(X_i - \mu_{ik})^2}{2\sigma_{ik}^2}} \quad (4.21)$$

where μ_{ik} is the mean and σ_{ik}^2 is the variance of i^{th} feature in class k . Mean, variance, and priors are estimated using a training set of positive and negative interactions (labeled dataset) in PPI prediction setting. They are defined as,

$$\mu_{ik} = \frac{\sum_{j=1}^{N_k} x_{ik}^j}{N_k} \quad (4.22)$$

$$\sigma_{ik}^2 = \frac{\sum_{j=1}^{N_k} (x_{ik}^j - \mu_{ik})^2}{N_k - 1} \quad (4.23)$$

$$\Pi_k = \frac{|N_k|}{\sum_k N_k} \quad (4.24)$$

where x_{ik}^j is the j^{th} observation of X_i belonging to class k . N_k is the number of examples in class k .

Nigam *et al.* (2000) proposed a text classification algorithm for learning a naïve Bayes classifier using both labeled and unlabeled documents. They showed that the accuracy of a multinomial naïve Bayes text classifier can be improved by augmenting a small set of labeled training documents with a large set of unlabeled documents using Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). The EM technique iteratively computes maximum-likelihood estimates when the data has missing values. In semi-supervised training setting, the missing values correspond to the missing labels of examples in unlabeled dataset. The E-step of the EM algorithm estimates the class probabilities (or expectations) of unlabeled data given the latest iteration of model parameters. M-step maximizes the likelihood of model parameters using previously estimated probabilities of unlabeled data and labeled data (Nigam *et al.*, 2000). The training of naïve Bayes classifier using unlabeled data proceeds as follows:

1. Train the naïve Bayes classifier with labeled dataset.
2. Compute the class probabilities of examples in unlabeled set using the naïve Bayes classifier trained in step 1 (E-step).
3. Re-estimate the classifier parameters using both labeled and unlabeled data. Class probabilities estimated for unlabeled examples in step 1 are treated as true class labels i.e. in a two class setting the unlabeled example is assigned to both the classes with probability estimated in previous step (M-step).
4. Go back to step 2 until convergence.

Convergence of EM algorithm is measured as the change in complete log probability of training data (labeled & unlabeled) and the prior. Complete probability of training data is the sum of posterior probabilities of labeled (S^l) and unlabeled data (S^u).

$$\ell = \log(P(Y)) + \sum_{i \in S^u} \log \sum_Y P(Y = k) P((X_i | Y = k)) + \sum_{i \in S^l} \log(P(Y = k) P((X_i | Y = k)) \quad (4.25)$$

EM algorithm converges when $\Delta\ell = (\ell_{\text{present}} - \ell_{\text{previous}}) < \epsilon$, where ϵ is a threshold set to 0.01.

Machine learning based methods for PPI predictions rely on positive and negative datasets during the training phase. In most cases, including this work, a positive set is created using experimentally determined high confidence protein interactions. These interactions are readily available from a number of PPI databases (Ceol *et al.*, 2007; Razick *et al.*, 2008; Salwinski *et al.*, 2004). On the other hand, it is difficult to find experimentally detected negative interactions (or non-interacting protein pairs), as they are rarely published. Also, experimentally detected PPIs, especially through high-throughput screens, suffer from false positives. Therefore, we hypothesized that using a high confidence set of experimentally detected positive and negative interactions as labeled data (S^l) and a larger unlabeled set (S^u) during the training phase could improve the performance of classifier. We extended Nigam *et al.*'s semi-supervised training approach to Gaussian naïve Bayes models by replacing mean, variance and priors in equations 4.22, 4.23 and 4.24 with weighted mean, weighted variance and weighted priors respectively,

$$\bar{\mu}_{ik} = \frac{\sum_{j=1}^{N_k} \delta_k^j x_{ik}^j}{\sum_{j=1}^{N_k} \delta_k^j} \quad (4.26)$$

$$\bar{\sigma}_{ik}^2 = \frac{\sum_{j=1}^{N_k} \delta_k^j (x_{ik}^j - \bar{\mu}_{ik})^2}{\sum_{j=1}^{N_k} \delta_k^j - \left(\frac{\sum_{j=1}^{N_k} (\delta_k^j)^2}{\sum_{j=1}^{N_k} \delta_k^j} \right)} \quad (4.27)$$

$$\bar{\Pi}_k = \frac{\sum_{j=1}^{N_k} \delta_k^j}{\sum_k N_k} \quad (4.28)$$

where δ_k^j is the weight of j^{th} example of class k and it is defined as,

$$\delta_k^j = \begin{cases} 1 \text{ or } 0 & \text{if } j \in S^l \\ p(j, k) & \text{if } j \in S^u \end{cases} \quad (4.29)$$

where $p(j, k)$ is the probability of j^{th} unlabeled example belonging to class k as computed in the E step of EM algorithm. For labeled examples, δ_k^j is either 1 or 0. The EM algorithm iterates over E & M steps till the classifier parameters improve, which is measured by the change in the sum of log probability of labeled data, unlabeled data, and the prior (Nigam *et al.*, 2000). Both the peptide (M_{pep}) and protein (M_{pro}) classifiers are trained using labeled and unlabeled datasets (see model training for more details). Assuming that peptide and protein features are independent of each other, the posterior probabilities $P(Y|M_{pep})$ and $P(Y|M_{pro})$ are combined using Bayes' theorem

(Jain and Bader, 2016),

$$\begin{aligned} P(Y|M_{pep}, M_{pro}) &= \frac{P(Y)P(M_{pep}, M_{pro}|Y)}{P(M_{pep}, M_{pro})} \\ &= \alpha \frac{P(Y|M_{pep})P(Y|M_{pro})}{P(Y)} \end{aligned} \quad (4.30)$$

where α is the normalizing constant.

4.5 Results

4.5.1 Model training

Peptide classifier training and validation set

400 unique peptides (sequence length upto 15 amino acids) belonging to 1,074 SH3-peptide mediated PPIs were downloaded from MINT (Licata *et al.*, 2012), DOMINO (Ceol *et al.*, 2007) databases and a literature curated list from Carducci *et al.* (2012). MUSI (Kim *et al.*, 2011) was used to divide the peptide set into two generic PWMs capturing major known SH3 domain binding motif classes [R/K]xxPxxP and PxxPxR. These PWMs were then used to scan all 1,074 interactions to identify significant amino acid positions and trim the binding sites. 847 out of 1,074 interactions have complete feature information. The vast majority of these interactions (793 out of 847) are only supported by *in vitro* experiments, like phage display or peptide chips. They are not supported by any evidence that indicates their occurrence in the cell (i.e. *in vivo*). However, 583 out of 793 interactions are supported by experimental techniques other than phage display and peptide chips at the protein level in iRefWeb database (Turner *et al.*, 2010). These interactions are used as labeled training set (**L1**). The remaining 54 interactions out of 847 with complete feature information are supported by *in vivo* experiments and used as a blind validation set (**V1**).

There are more than 200 SH3 domain containing human proteins but the high confidence SH3 domain-peptide labeled set (L1) covers only 6 SH3 domain containing proteins. Also, the training set only represents class I and II canonical binding motifs whereas it is known that SH3 domains have other binding modes. Therefore, to increase the coverage, unlabeled data is used in a semi-supervised training setting. 210 interactions out of 793 which are not supported by any *in vivo* or small scale experiment are treated as unlabeled (**U1**). Another, 2,500 novel SH3 domain-peptide mediated PPIs predicted using high-throughput peptide chip technology and sequence matching by

Carducci *et al.* (2012) are added to the unlabeled set as these interactions are not supported by *in vivo* evidence. By adding unlabeled data the coverage is increased to 59 human SH3 domain containing proteins.

The negative dataset is created from randomly selected protein pairs with one member containing a randomly selected human SH3 domain and the other a randomly selected 10 – 15 amino acid long peptide sequence. Positive PWMs from the P1 dataset are used to scan the peptide sequences and only those with scores below the p-value threshold of 0.05 are retained. Also, the protein pairs are not part of known interactions from the iRefIndex (version 13.0) database (Razick *et al.*, 2008). To balance the positive training and validation sets equal number of negative interactions are added to L1, V1 and U1 datasets.

Protein classifier training and validation set

544 high confidence pairwise human PPIs are randomly selected from 4,652 interactions retrieved from iRefIndex (Razick *et al.*, 2008) using its web interface iRefWeb. iRefIndex consolidates PPIs from 10 major public databases and provides many filters to create a high confidence PPI set. The interactions retrieved from iRefWeb are all physical, experimental, from a single organism, supported by at least two publications and have a MI (MINT-Inspired) score ≥ 0.9 . The number of interactions is limited to 544 because the high confidence negative interaction set has only that many interactions with complete feature information. 1,048 non-interacting human protein pairs are retrieved from Negatome (Blohm *et al.*, 2013). Negatome uses manual literature curation and 3-D structure analysis to compile a high confidence list of non-interacting protein pairs. A high confidence set of 544 non-interactions is created after removing instances with missing protein feature information (labeled set: **L2**).

Trabuco *et al.* (2012) published a set of negative interactions computationally derived from two-hybrid screens. As these interactions are not experimentally verified they are not part of the labeled dataset. 4,000 derived negatives and equal number of positive interactions from iRefWeb are used to create an unlabeled set for semi-supervised training (**U2**). To make an unbiased assessment of the classifier 1,000 interactions (with no missing information) from the core subset of Database of Interacting Proteins (DIP) that do not overlap with L2 & U2 training sets are used (Salwinski *et al.*, 2004). DIP interactions are based on different filtering criteria compared to the MINT-inspired score used to select the iRefIndex. 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

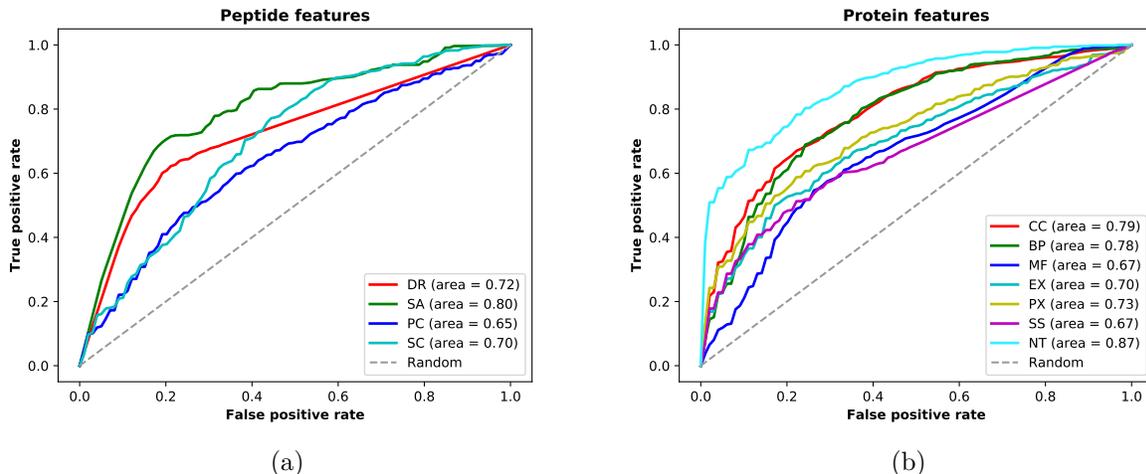


Figure 4.2: Prediction efficacy of individual (a) peptide features: disordered region (DR), surface accessibility (SA), peptide conservation (PC), structural contact (SC). (b) protein features: cellular component (CC), biological process (BP), molecular function (MF), gene expression (EX), protein expression (PX), sequence signature (SS), network topology (NT).

of interaction between paralogs (PVM), probable domain-domain interactions between protein pairs (DPV), and comparison with expression profiles (EPR) (Salwinski *et al.*, 2004). Equal number of derived negatives are added to the blind validation set (**V2**).

4.5.2 Feature selection

Figure 4.2 shows the prediction efficacy of individual features for peptide and protein classifiers. All the peptide and protein features perform better than random in separating positives from negatives. Surface accessibility (SA) is ranked best, whereas the prediction efficacy of peptide conservation (PC) is least among the peptide features. Surprisingly, the performance of disordered region (DR) is comparatively weak considering the preference of SH3 domains for amino acids in these regions. Approximately, 62% of human proteins are either intrinsically disordered or have intrinsically disordered regions (Deiana and Giansanti, 2016) and this could lead to higher disorder score for peptides in the negative dataset thus, affecting its prediction performance. In the protein feature set, network topology (NT) outperforms biological process (BP), cellular component (CC), and sequence signature (SS) molecular function (MF), gene expression (EX) and protein expression (PX). Weak performance of molecular function can be attributed to the fact that proteins same molecular function might belong to completely different biological processes. As expected, protein expression performs better than gene expression in discriminating positives from negatives, as ex-

pression of a gene does not necessarily lead to a protein product for interaction (Vogel and Marcotte, 2012).

A naïve Bayes classifier works under the assumption of feature independence. Therefore, a highly correlated set of features could degrade its performance (Ratanamahatana and Gunopulos, 2003). The maximal information coefficient (MIC) technique measures the linear and non-linear relationship between two variables by calculating their normalized mutual information. Mutual information quantifies the reduction in uncertainty of one variable given the information of another. MIC score ranges from 0 for complete independence to 1 for total dependence between two variables (Albanese *et al.*, 2012; Jain and Bader, 2016). Figure 4.3 shows the MIC plot for peptide and protein features. None of the feature pairs in peptide and protein classifiers show significant correlation. Maximum MIC score of 0.37 is observed between disordered region (DR) and surface accessibility (SA) features in peptide classifier and 0.36 for cellular component (CC) and biological process (BP) in protein classifier. Further, to identify the feature subset which maximizes the performance of both classifiers recursive feature elimination (RFE) algorithm with a support vector classifier (SVC) is used. RFE-SVC routine starts with computing weights for all the features using a SVC model and then recursively eliminates the feature with smallest absolute weight. The RFE routine is performed in a 5-fold cross-validation loop to select the optimal number of features using model accuracy as a benchmark. Figure 4.4 shows that both peptide and protein classifiers achieve maximum accuracy when all the features are used.

4.5.3 Model evaluation

Peptide classifier is evaluated using a blind validation protocol where the classifier is trained on the labeled set L1 and the unlabeled set U1 and tested with validation set V1. The accuracy of the peptide classifier increased from 0.80 to 0.83 when unlabeled data is added to the training procedure. Table 4.2 shows the incremental increase in MCC, accuracy and F_1 -score with the increase in number of unlabeled training examples. Similarly, the protein classifier is trained on the labeled set L2 and the unlabeled set U2 and tested with validation set V2. MCC, accuracy and F_1 -score of the protein model also showed an increasing trend with the increase in number of unlabeled training examples (table 4.3). Accuracy of the protein classifier increased from 0.86 to 0.92 when unlabeled data is added. As discussed earlier, PRM mediated PPI prediction pipeline combines both the binding site (domain-peptide) and full length protein features to identify high confidence interactions with true binding sites. Validation set V1 with 54 interactions with complete feature information that are

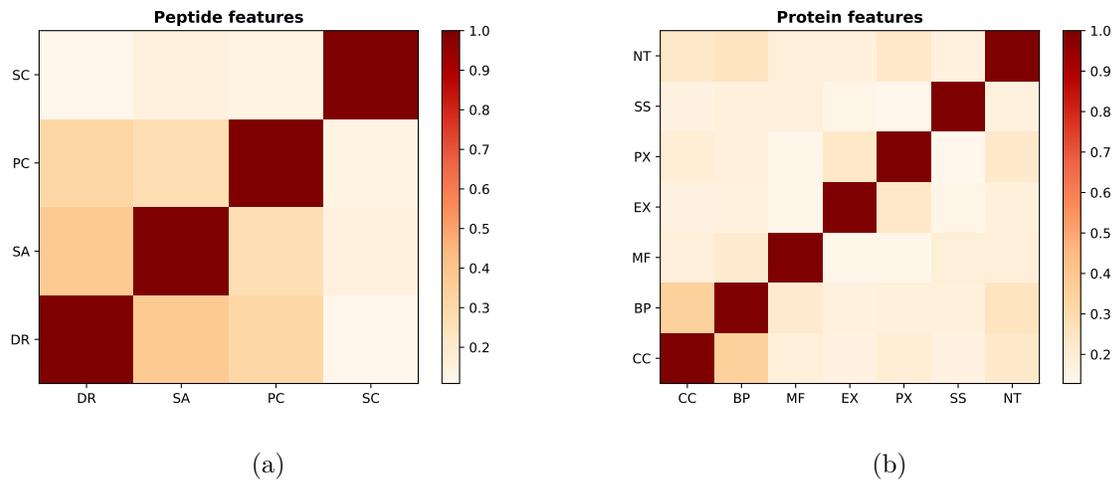


Figure 4.3: Maximal information coefficients for (a) Peptide feature set: disordered region (DR), surface accessibility (SA), peptide conservation (PC), structural contact (SC). (b) Protein feature set: cellular component (CC), biological process (BP), molecular function (MF), gene expression (EX), protein expression (PX), sequence signature (SS), network topology (NT).

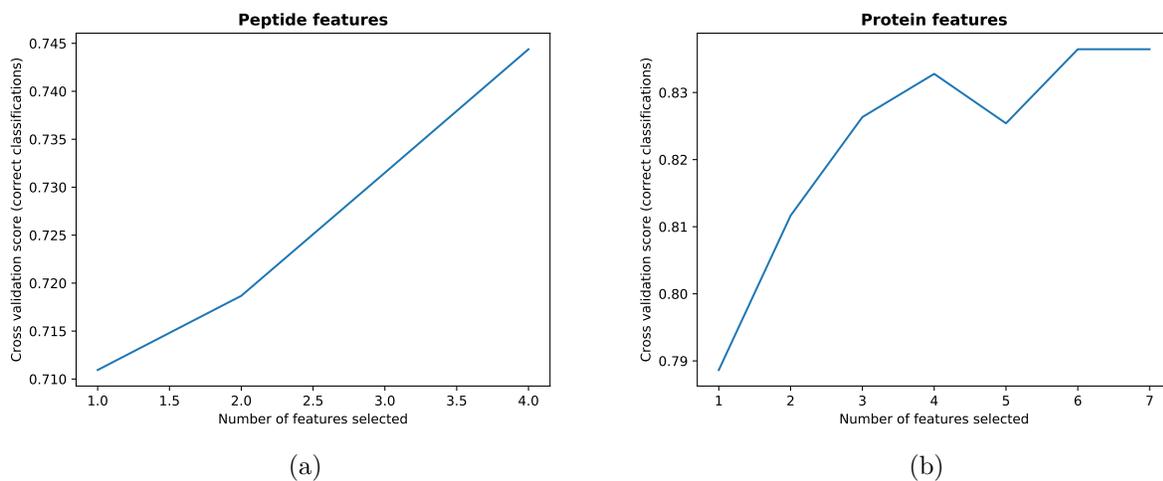


Figure 4.4: Recursive feature elimination plots (a) Peptide classifier. (b) Protein classifier.

Peptide classifier			
Unlabeled data	MCC	ACC	F ₁ -score
0	0.63	0.80	0.79
500	0.66	0.81	0.81
1000	0.67	0.82	0.82
1500	0.68	0.83	0.83
2000	0.69	0.83	0.83
2500	0.69	0.83	0.83

Table 4.2: Peptide classifier: Matthews correlation coefficient (MCC), accuracy (ACC), and F1-score for different models with increasing number of unlabeled data.

Protein classifier			
Unlabeled data	MCC	ACC	F ₁ -score
0	0.75	0.87	0.86
500	0.78	0.88	0.88
1000	0.79	0.89	0.89
1500	0.81	0.90	0.90
2000	0.82	0.91	0.91
2500	0.83	0.91	0.91
3000	0.83	0.91	0.91
3500	0.83	0.91	0.91
4000	0.83	0.92	0.92

Table 4.3: Protein classifier: Matthews correlation coefficient (MCC), accuracy (ACC), and F1-score for different models with increasing number of unlabeled data.

supported by *in vivo* experiments and equal number of negative interactions is used to compare the combined classifier with that of individual peptide and protein classifiers. Figure 4.5 shows that by combining peptide and protein classifiers high confidence PRM mediated PPIs can be predicted along with binding sites. Moreover, using only peptide features as was used by MOTIPS (Lam *et al.*, 2010) are not sufficient and adding full length protein information improves the PRM mediated PPI predictions as shown by different measures.

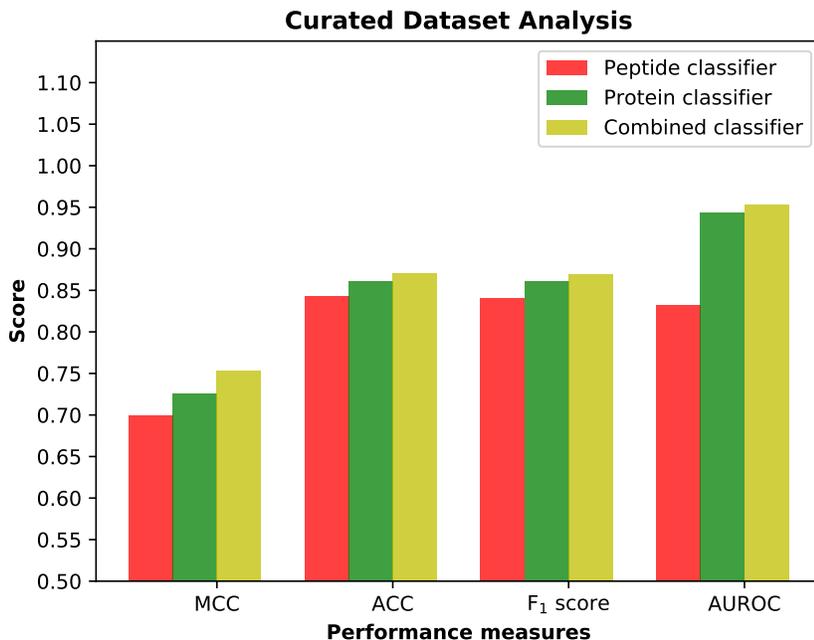


Figure 4.5: Performance of peptide, protein and combined classifiers on the curated SH3 domain mediated PPI set. (Note: small size of curated validation dataset prevents the variance from being estimated.)

4.5.4 Discussion¹

Recently, Teyra *et al.* (2017) comprehensively surveyed the specificity landscape of human SH3 domains in an unbiased manner using peptide-phage display and deep sequencing. Based on more than 70,000 unique binding peptides, 154 specificity profiles for 115 SH3 domains were obtained, which revealed that roughly half of the SH3 domains exhibit non-canonical specificities and collectively recognize a wide variety of peptide motifs, most of which were previously unknown. 154 binding specificities for 115 SH3 domains were organized into nine classes based on similarities in peptide binding preferences using PWM logos. Class I (Fig. 4.6(A)) and class II (Fig. 4.6(B)) domains are defined as those recognizing peptides containing the PxxP core and an R/K residue either N- or C-terminal to the core, respectively. Domains able to recognize both class I and class II peptides were placed in class I/II (Fig. 4.6(C)). They identified 25 domains showing alternative class I-like or class II-like specificities that were grouped in six additional classes (III to VIII). Specificities where the proximal Pro residue was not required were classified as class III (-6RxxxxxP0(+), three domains, Fig. 4.7(D)) or class IV (0PxxxxR+5(-), three domains, Fig. 4.7(E)) if they resembled

¹This section is derived from our published work in Structure: Teyra, J., Huang, H., Jain, S. *et al.* (2017). I collected human SH3 domains and analyzed phage display data.

A Class I: $^{-3}\text{RxxP}^0\text{xxP}^{+3}(+)$			B Class II: $^0\text{PxxP}^{+3}\text{xR}^{+5}(-)$			C Class I/II		
Domain	n	Logo	Domain	n	Logo	Domain	n	Logos
GRAP2-2/2	62		NCK1-2/3	1064		PACSIN2-1/1	158	
PACSIN1-1/1	216		SH3RF2-3/3	33		PLCG2-1/1	850	
NCK2-1/3-SV2/2	48		SH3RF1-3/4	763		PIK3R2-1/1	303	
SRC-1/1-IS1/2	133		SH3RF3-1/2	668		PLCG1-1/1	1262	
SH3GLB1-1/1	134		ARHGAP26-1/1	815		ARHGAP10-1/1-SV2/2	288	
SNX18-1/1	69		ARHGAP42-1/1	288		SH3YL1-1/1	3302	
ARHGAP32-1/1	103		AMPH-1/1	1039		OSTF1-1/1	872	
SORBS1-2/3	176		GRAP-2/2	195		VAV3-1/2	152	
MYO1E-1/1-SV1/2	223		SH3D19-5/5	163		LCK-1/1	1033	
BLK-1/1	1165		ARHGFE7-1/1	896		SORBS1-3/3	636	
YES1-1/1	1253		UBASH3B-1/1	33		RP5-862P8.2-1/1	421	
RIMBP2-1/3	172		LYN-1/1	2484		VAV1-2/2	1136	
FYB-2/2	51		SH3GL1-1/1	43		VAV1-1/2	51	
RIMBP3B-1/1	278		PPP1R13B-1/1	304		SH3KBP1-1/3	1564	
BZRAP1-1/3	489		ITSN2-1/5	188		ITSN1-3/5	485	
RIMBP2-3/3	1009		ITSN1-1/5-IS3/3	4007		ITSN2-3/5	173	
BZRAP1-3/3	165		NOSTRIN-1/1-SV1/2	2052		SH3RF1-1/4	403	
SH3D21-3/3	41		SORBS2-3/3	47		FGR-1/1	957	
			AHI1-1/1	143		FYN-1/1	1375	
			ITSN1-5/5	130		SRC-1/1-IS2/2	1364	
			ITSN2-5/5	236		GRB2-1/2	328	
			CTTN-1/1	265		NCK1-1/3	752	
			BZRAP1-2/3	475		NCK2-1/3-SV1/2	535	
			CRK-1/2	2469				

Figure 4.6: A total of 154 peptide-binding specificities for 115 SH3 domains were grouped in nine classes, as follows: (A) class I, (B) class II, (C) class I/II, (D) class III, (E) class IV, (F) class V, (G) class VI, (H) class VII, (I) class VIII, and (J) class IX. Each panel contains the list of SH3 domains for a particular class with the defining motif shown at the top. Each row contains the SH3 domain name, the number of unique peptide ligands isolated by phage display (n), and the sequence logos derived from the frequencies of amino acids in aligned peptide ligand sequences.

D Class III: $^6\text{RxxxxxP}^0(+)$			G Class VI: $^{-3}\text{xxxP}^0\text{xR}^{+2}(-)$			J Class IX: atypical		
Domain	n	Logo	Domain	n	Logo	Domain	n	Logo
STAC2-1/1	281		ASAP1-1/1	4743		GRAP2-2/2	275	
BTK-1/1	478		ASAP2-1/1	2001		STAM-1/1	102	
TEC-1/1	1270		NPHP1-1/1	78		EPS8-1/1	1547	
E Class IV: $^0\text{PxxxxR}^{+5}(-)$			H Class VII: $^0\text{PxK}^{+2}\text{P}^{+3}(-)$			SH3KBP1-2/3		
MAP3K9-1/2	72		BCAR1-1/1	620		ITK-1/1	66	
CD2AP-1/3	2245		NEDD9-1/1	350		SH3PXD2A-3/5	40	
CD2AP-2/3	1442		I Class VIII: $^0\text{PxxP}^{+3}\text{xxP}^{+6}(-)$			SH3PXD2A-1/5	89	
F Class V: $^{-3}\text{RxxP}^0\text{xxx}^{+3}(+)$			ARHGAP4-1/1	76		SH3PXD2B-1/4	222	
SNX33-1/1	124		SORBS2-1/3	118		NOSTRIN-1/1-SV1/2	1453	
SNX9-1/1	593		SORBS2-2/3	374		STAM-1/1	183	
NPHP1-1/1	1905		PTK6-1/1	60		PTK6-1/1	54	
TXK-1/1	197		SORBS2-2/3	123		DNMBP-6/6	3062	
BIN1-1/1	294		SRGAP1-1/1	107		BAIAP2L2-1/1	38	
SH3PXD2A-5/5	363		ABL2-1/1	458		ABL2-1/1	551	
SH3RF1-4/4	96		J Class IX: atypical (continued)			BAIAP2-1/1	196	
SH3PXD2B-3/4	70		GRB2-2/2	70		AC093799-1-1/1	56	
ASAP2-1/1	235		LASP1-1/1-SV2/2	47		ARHGAP12-1/1	360	
			LNK1-1/1-SV3/3	52		ARHGAP27-1/1	98	
			SH3D19-4/5	191		ARHGAP9-1/1	135	
			SH3RF1-2/4	67		ARHGAP4-1/1	52	
			STAM-1/1	96		ARHGAP12-1/1	89	
			SH3GL1-1/1	58		GRAP-2/2	104	
			SRC-1/1-IS1/2	71				
			ITSN1-2/5	79				
			ITSN2-2/5	27				
			OBSCN-1/1	109				

Figure 4.7: A total of 154 peptide-binding specificities for 115 SH3 domains were grouped in nine classes, as follows: (A) class I, (B) class II, (C) class I/II, (D) class III, (E) class IV, (F) class V, (G) class VI, (H) class VII, (I) class VIII, and (J) class IX. Each panel contains the list of SH3 domains for a particular class with the defining motif shown at the top. Each row contains the SH3 domain name, the number of unique peptide ligands isolated by phage display (n), and the sequence logos derived from the frequencies of amino acids in aligned peptide ligand sequences.

class I or II, respectively. Conversely, if the distal Pro residue was not required, the specificities were classified as class V (-3RxxP0xxx+3(+), nine domains, Fig. 4.7(F)) or class VI (-3xxxP0xR+2(-), three domains, Fig. 4.7(G)) if they resembled class I or II, respectively. Class II-like domains that substituted the flanking Arg residue with a Lys residue inserted within the PxxP core were classified as class VII (0PxK+2P+3(-), two domains, Fig. 4.7(H)). Class VII peptides bound in a minus orientation because of the positively-charged residue embedded in the PxxP core for class II domains, 0PxK+2P+3xR+5(-), and because this orientation is observed in all known structures of SH3 domains in complex with a peptide conforming to a 0Px[R/K]+2P+3(-) motif. Domains that recognize peptides with an extra Pro residue in place of the canonical flanking Arg residue were placed in class VIII (0PxxP+3xxP+6(-), seven domains, Fig. 4.7(I)). Finally, many SH3 domains exhibited a variety of atypical specificities which will require additional study to characterize their binding properties and define new classes, thus these were grouped for convenience into a single class IX (35 domains, Fig. 4.7(J)). These PWMs are then used to predict SH3 domain-peptide interactions using the combined classifier. 2,359 high confidence (probability > 0.9) unique SH3 mediated PPIs with 3,097 binding sites are predicted. Proteins involved in these interactions are found to be enriched in Reactome (Croft *et al.*, 2013) biological pathways such as signaling, cell-cell communication, and phagocytosis (Figure 4.8). These pathways are known to be mediated by SH3 domains thus providing an independent source of verification.

4.6 Conclusion

We developed a novel method for predicting SH3 domain mediated physiologically relevant PPIs in human. This method combines diverse binding site (peptide) features, including presence in a disordered region of the protein, surface accessibility, conservation across different human 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 and protein expressions, network topology, and sequence signature. Two separate Bayesian models are used to combine peptide and protein features using a semi-supervised training framework and their respective posterior probabilities are further combined using Bayes rule for predicting high confidence interactions. We have also developed a novel algorithm for predicting PPIs using network topology and successfully extended the existing semi-supervised training framework for text classification to Gaussian naïve Bayes model for PPI prediction problem. The combination of peptide and protein models achieved

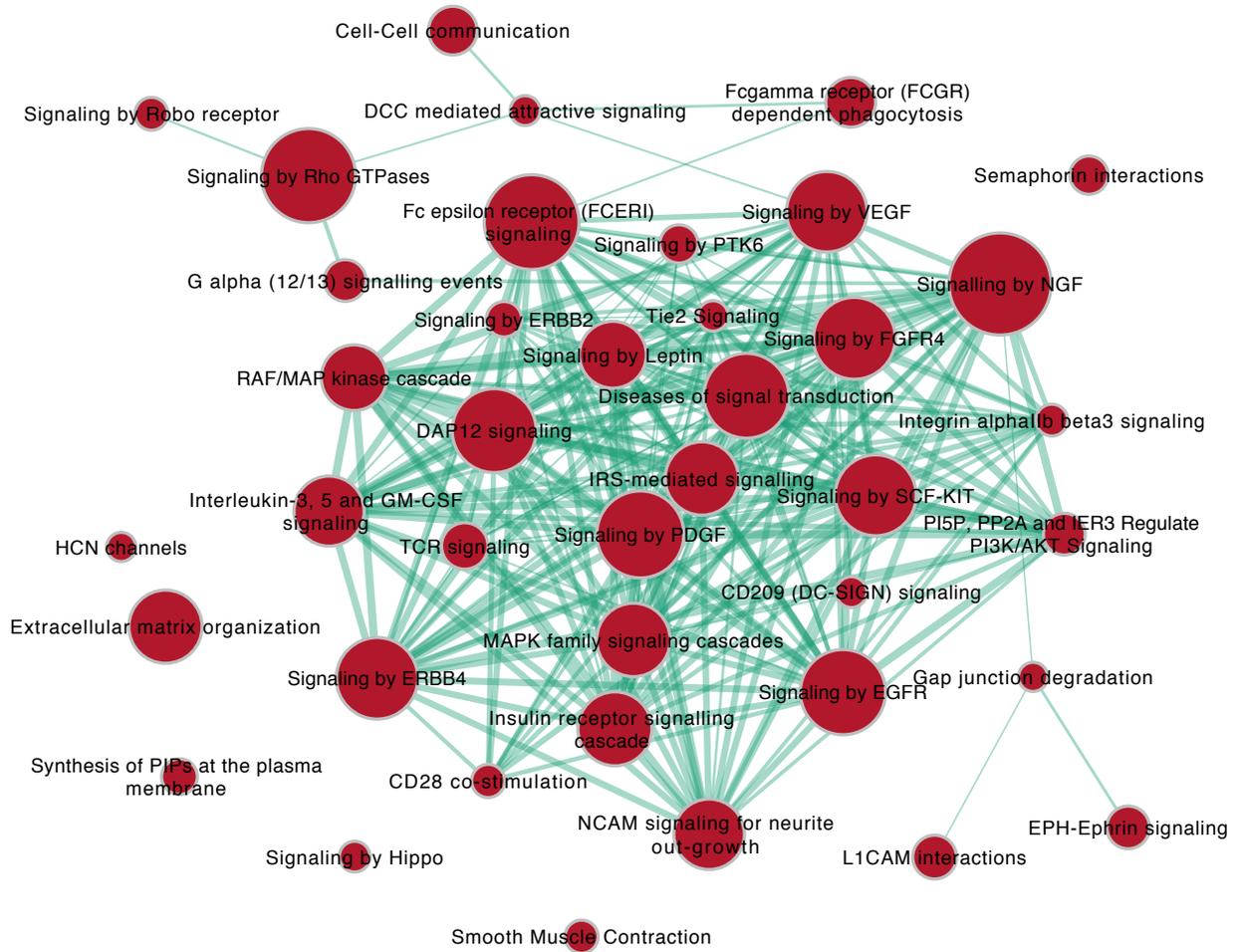


Figure 4.8: Enrichment map of proteins involved in SH3 domain mediated PPIs predicted by DoMo-Pred.

a higher AUROC score of 0.97 compared to individual models on a benchmark dataset. Surface accessibility data from the peptide feature set and network topology 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 and given binding peptide and feature data it can be used to predict other PRM mediated PPIs in different organisms. Additional features such as text mining derived protein relationships, evolutionary conservation of protein interactions, and correlated mutations within the binding sites can be added to our framework. Future development includes testing this method on other PRMs in different organisms and a web based PRM mediated PPI interaction prediction tool.

Implementation

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

Acknowledgements

We thank Joan Teyra, Haiming Huang, and Sachdev S. Sidhu for providing us with human SH3 phage display data.

Chapter 5

Summary and future directions

My thesis focuses on building computational models for predicting peptide recognition module (PRM) mediated protein-protein interactions.

5.1 Summary of our major contributions

5.1.1 Cellular location, biological process, molecular function

This work was published in BMC Bioinformatics: An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology (Jain and Bader, 2010).

Semantic similarity measures are useful to assess the physiological relevance of protein-protein interactions (PPIs). They quantify similarity between proteins based on their function using annotation systems like the Gene Ontology (GO) (The Gene Ontology Consortium, 2000). Proteins that interact in the cell are likely to be in similar locations or involved in similar biological processes compared to proteins that do not interact. Thus the more semantically similar the gene function annotations are among the interacting proteins, more likely the interaction is physiologically relevant. However, most semantic similarity measures used for PPI confidence assessment do not consider the unequal depth of term hierarchies in different classes of cellular location, molecular function, and biological process ontologies of GO and thus may over- or under-estimate similarity. We developed an algorithm, Topological Clustering Semantic Similarity (TCSS), to compute semantic similarity between GO terms annotated to proteins in interaction datasets. Our algorithm, considers unequal depth of biological knowledge representation in different branches of the GO graph. The central idea is to divide the GO graph into sub-graphs and score PPIs higher if participating proteins belong

to the same sub-graph as compared to if they belong to different sub-graphs. The TCSS algorithm performs better than other semantic similarity measurement techniques that we evaluated in terms of their performance on distinguishing true from false protein interactions, and correlation with gene expression and protein families.

5.1.2 Gene expression

This work was published in Bioinformatics: Predicting physiologically relevant sh3 domain mediated protein-protein interactions in yeast (Jain and Bader, 2016).

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. We showed that by combining multiple GEPs improves the performance of a PPI predictor. In case of yeast, we combined correlation coefficients from 86 gene expression profiles and 140 for human from GeneMANIA (Warde-Farley *et al.*, 2010) for a given pair of genes using Fisher's z transformation to improve the efficacy of gene expression as a measure for PPI prediction (Faller, 1981; Jain and Bader, 2010).

5.1.3 Protein expression

This work was part of a collaboration that was published in Nature: A draft map of the human proteome (Kim et al., 2014). I did the PPI network analysis and proposed the use of protein expression data for predicting PPIs.

The availability of human genome sequence has transformed biomedical research over the past decade. However, an equivalent map for the human proteome with direct measurements of proteins and peptides was not available until Kim and co-workers used high-resolution Fourier-transform mass spectrometry to produce a draft map of human proteome. They did in-depth proteomic profiling of 30 histologically normal human samples, including 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells, which resulted in identification of proteins encoded by 17,294

genes. As described in previous section, gene expression profiles across various experimental conditions or tissues have been utilized to investigate the likelihood of co-expressed genes to physically interact at the protein level. With the availability of protein expression data, we hypothesized that protein expression pattern should be a better predictor of protein-protein interactions than gene expression measured at the mRNA level. We correlated normalized spectral count profiles for each available protein pair across all 30 cells/tissues in our data using Pearson correlation and compared this to known protein-protein interactions. We then repeated this analysis using correlations obtained from 111 published gene expression data sets. Our analysis clearly showed that the human proteome profile correlation outperforms gene expression profile correlation for predicting protein-protein interactions, even if all gene expression data sets are combined and used as a single predictor.

5.1.4 Network topology (NTOP)

This work was part of the manuscript: Predicting in-vivo SH3 domain mediated interactions in human (Jain et al., to be submitted).

Networks or graphs provide a powerful computational framework to represent and analyze complex biological systems. They include transcriptional regulatory networks, metabolic networks, signal transduction networks, and PPI networks. Development of high-throughput technologies for detecting protein interactions have created large-scale PPI networks where, nodes correspond to proteins and undirected edges represent physical interactions amongst them. We propose a machine learning based algorithm for predicting PPIs using topological information of known networks. We combined network properties, such as edge density, edge connectivity, mean degree, transitivity, and mutual clustering coefficient using a logistic regression model and outperformed the existing network topology based PPI prediction methods. Network topology was then used as one of the features in our PRM mediated PPI prediction pipeline.

5.1.5 Semi-supervised training

This work was part of the manuscript: Predicting in-vivo SH3 domain mediated interactions in human (Jain et al., to be submitted).

Machine learning based methods for PPI predictions rely on positive and negative datasets during

training phase. In most cases, including this work, the positive set is created using experimentally determined high confidence protein interactions. These interactions are readily available from a number of PPI databases (Ceol *et al.*, 2007; Razick *et al.*, 2008; Salwinski *et al.*, 2004). On the other hand, it is difficult to find experimentally detected negative interactions (or non-interacting protein pairs), as they are rarely published. Also, experimentally detected PPIs, especially through high-throughput screens, suffer from false positives. Therefore, we hypothesized that using a high confidence set of experimentally detected positive and negative interactions as labeled data and a larger unlabeled set during the training phase could improve the performance of classifier. We extended Nigam *et al.* (2000)’s semi-supervised training approach to Gaussian naïve Bayes (GNB) models and were able to improve the performance of our PRM mediated PPI prediction model. The proposed semi-supervised GNB framework can be applied to any supervised classification problem including generic PPI prediction problem.

5.1.6 Domain-Motif Mediated Interaction Prediction (DoMo-Pred)

First version of DoMo-Pred was published in Bioinformatics: Predicting physiologically relevant sh3 domain mediated protein-protein interactions in yeast (Jain and Bader, 2016). Latest version of DoMo-Pred is part of the manuscript: Predicting in-vivo SH3 domain mediated interactions in human (Jain et al., to be submitted).

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. We developed a novel method for predicting physiologically relevant SH3 domain-peptide mediated protein-protein interactions in *S. cerevisiae* and *H. sapiens* using phage display data. 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 se-

quence based and protein pair based evidence sources. By combining different peptide and protein features using multiple Bayesian models we were able to predict high confidence interactions.

5.2 Future directions

5.2.1 Additional features and other domains

DoMo-Pred tool uses a combination of peptide (or binding site) and full length protein features for predicting PRM mediated PPIs. Surface accessibility of binding site, its presence in disordered region of a protein, conservation and structural contact are combined with protein features, such as cellular proximity of domain containing and binding site proteins, biological process, molecular function, correlated gene and protein expression profiles, sequence composition, and network topology. The feature landscape used in DoMo-Pred is quite comprehensive but more features can be added to the model to improve its performance and coverage.

Automated information extraction from literature through text mining can be used to predict PRM mediated PPIs or PPIs in general. Proteins that are often cited together in sentences in journal articles are more likely to interact. The simplest way to extract PPIs from the literature is to detect the co-occurrence of protein names in a text. The STRING database parses scientific documents for statistically relevant co-occurrences of gene names (Snel *et al.*, 2000). Ramani *et al.* (2005) developed natural-language processing and literature-mining algorithms to recover interactions among human proteins from Medline abstracts. Daraselia *et al.* (2004) proposed MedScan, a full-sentence based PPI information extraction system from Medline abstracts. Donaldson *et al.* (2003) proposed a support vector machine based information extraction system trained to recognize abstracts describing biomolecular interactions. While text mining is error-prone when combined with other features it could improve the classifier's performance (Reimand *et al.*, 2012). Other PPI evidence sources which can be used are evolutionary conservation of PPIs across different species and correlation between evolutionary mutations in binding sites (Jothi *et al.*, 2006). These evidence sources are promising but have their own sets of challenges. For example, the coverage of conserved PPIs across different species is quite low. Further research into these evidence sources could help in improving the model. Moreover, improvements in some of the existing DoMo-Pred features, such as developing new models for combining multiple gene expression profiles, can help in improving the classifier performance.

In this thesis, the focus was on the SH3 domains but the framework can be easily extended

to other domains binding to short linear motifs, such as WW, PDZ, 14-3-3, DEP, and G-alpha. Extending DoMo-Pred algorithm to other domains will help us in generating a high-resolution (i.e. with binding site information) PPI network which in turn will help in better understanding different biological processes and human diseases.

5.2.2 PRM-mediated protein-protein interaction networks in human disease

PRM-mediated PPIs are involved in many important biological processes including signaling systems and human diseases (Reimand *et al.*, 2012). For example, mutations in PDZ domains disrupts cell polarity which in turn plays a role in tumor metastasis and immune deficiencies (Dev, 2004; Doorbar, 2006). SH2 and SH3 domain mediated PPIs are required for the transmission of signals initiated by tyrosine kinases and are involved in cancers, such as acute lymphocytic leukemias and HER-2/Neu in breast and ovarian cancer (Smithgall, 1995). Therefore, a systematic analysis of high-confidence, high-resolution protein interaction networks will help explain disease mechanisms and identify new therapeutic targets. In this thesis, the focus was on developing these high-confidence and high-resolution networks and the next logical step is to use them for generating insights into the genetic basis of specific diseases and their treatments. PRM mediated PPI networks can help us in understanding network rewiring due to disease DNA mutations which could lead to gain or loss of an interaction. Mapping a large number of disease-associated mutations and their network effects will also enable a better understanding of the relationship between genomes and networks (Reimand *et al.*, 2012).

Bibliography

- 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.
- Adler, P., Kolde, R., Kull, M., Tkachenko, A., Peterson, H., Reimand, J., and Vilo, J. (2009a). Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome biology*, **10**, R139.
- Adler, P., Peterson, H., Agius, P., Reimand, J., and Vilo, J. (2009b). Ranking genes by their co-expression to subsets of pathway members. *Annals of the New York Academy of Sciences*, **1158**, 1–13.
- Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, **422**(6928), 198–207.
- Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., and Furlanello, C. (2012). Minerva and minepy: a c engine for the mine suite and its r, python and matlab wrappers. *Bioinformatics*, **29**(3), 407–408.
- Aloy, P. and Russell, R. B. (2003). Interprets: protein interaction prediction through tertiary structure. *Bioinformatics*, **19**(1), 161–162.
- Alterovitz, G., Xiang, M., Hill, D. P., Lomax, J., Liu, J., Cherkassky, M., Dreyfuss, J., Mungall, C., Harris, M. A., Dolan, M. E., Blake, J. A., and Ramoni, M. F. (2010). Ontology engineering. *Nature Biotechnology*, **28**, 128–130.
- 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.
- Aytuna, A. S., Gursoy, A., and Keskin, O. (2005). Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, **21**(12), 2850–2855.
- Azuaje, F., Al-Shahrour, F., and Dopazo, J. (2006). Ontology-driven approaches to analyzing data in functional genomics. *Bioinformatics and Drug Discovery*, pages 67–86.
- Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, **22**(1), 78–85.

- Beltrao, P. and Serrano, L. (2005). Comparative genomics and disorder prediction identify biologically relevant sh3 protein interactions. *PLoS Comput Biol*, **1**(3), e26.
- Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21 Suppl 1**, i38–i46.
- 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. L. and Alan, H. (2000). Rho gtpases and their effector proteins. *Biochemical Journal*, **348**(2), 241–255.
- Blohm, P., Frishman, G., Smialowski, P., Goebels, F., Wachinger, B., Ruepp, A., and Frishman, D. (2013). Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic acids research*, page gkt1079.
- 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.
- Breiman, L. and Schapire, E. (2001). Random forests. In *Machine Learning*.
- Brown, K. R. and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics*, **21**(9), 2076–2082.
- Carducci, M., Perfetto, L., Briganti, L., Paoluzi, S., Costa, S., Zerweck, J., Schutkowski, M., Castagnoli, L., and Cesareni, G. (2012). The protein interaction network mediated by human sh3 domains. *Biotechnology advances*, **30**(1), 4–15.
- Ceol, A., Chatr-aryamontri, A., Santonico, E., Sacco, R., Castagnoli, L., and Cesareni, G. (2007). Domino: a database of domain-peptide interactions. *Nucleic Acids Res*, **35**(Database issue), D557–D560.
- Chavez, S., Beilharz, T., Rondon, A. G., Erdjument-Bromage, H., Tempst, P., Svejstrup, J. Q., Lithgow, T., and Aguilera, A. (2000). A protein complex containing tho2, hpr1, mft1 and a novel protein, thp2, connects transcription elongation with mitotic recombination in *saccharomyces cerevisiae*. *The EMBO journal*, **19**(21), 5824–5834.
- 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. *PLoS computational biology*, **11**, e1004248.
- Chen, X.-W. and Liu, M. (2005). Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, **21**(24), 4394–4400.
- Chen, Y. and Xu, D. (2005). Genome-scale protein function prediction in yeast *Saccharomyces cerevisiae* through integrating multiple sources of high-throughput data. *Pac Symp Biocomput*, **1**, 471–482.
- Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D., and Siani-Rose, M. A. (2004). A knowledge-based clustering algorithm driven by gene ontology. *J Biopharm Stat*, **14**(3), 687–700.

- Chien, C. T., Bartel, P. L., Sternglanz, R., and Fields, S. (1991). The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci U S A*, **88**(21), 9578–9582.
- Christie, K. R., Weng, S., Balakrishnan, R., Costanzo, M. C., Dolinski, K., Dwight, S. S., Engel, S. R., Feierbach, B., Fisk, D. G., Hirschman, J. E., *et al.* (2004). Saccharomyces genome database (sgd) provides tools to identify and analyze sequences from saccharomyces cerevisiae and related sequences from other organisms. *Nucleic acids research*, **32**(suppl_1), D311–D314.
- Consortium, U. *et al.* (2010). The universal protein resource (uniprot) in 2010. *Nucleic acids research*, **38**(suppl 1), D142–D148.
- Couto, F. M., Silva, M. J., and Coutinho, P. M. (2005). Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors. In *CIKM 05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 343–344, New York, NY, USA. ACM.
- Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., *et al.* (2013). The reactome pathway knowledgebase. *Nucleic acids research*, **42**(D1), D472–D477.
- Dalby, P. A., Hoess, R. H., and Degrado, W. F. (2000). Evolution of binding affinity in a ww domain probed by phage display. *Protein Science*, **9**(12), 2366–2376.
- Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, **23**(9), 324–328.
- Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, **20**(5), 604–611.
- 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.
- Deiana, A. and Giansanti, A. (2016). Variants of intrinsic disorder in the human proteome. *arXiv preprint arXiv:1611.06072*.
- del Pozo, A., Pazos, F., and Valencia, A. (2008). Defining functional distances over gene ontology. *BMC Bioinformatics*, **9**, 50.
- Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D’Eustachio, P., Schaefer, C., Luciano, J., Schacherer, F., Martinez-Flores, I., Hu, Z., Jimenez-Jacinto, V., Joshi-Tope, G., Kandasamy, K., Lopez-Fuentes, A. C., Mi, H., Pichler, E., Rodchenkov, I., Splendiani, A., Tkachev, S., Zucker, J., Gopinath, G., Rajasimha, H., Ramakrishnan, R., Shah, I., Syed, M., Anwar, N., Babur, O., Blinov, M., Brauner, E., Corwin, D., Donaldson, S., Gibbons, F., Goldberg, R., Hornbeck, P., Luna, A., Murray-Rust, P., Neumann, E., Ruebenacker, O., Reubenacker,

- O., Samwald, M., van Iersel, M., Wimalaratne, S., Allen, K., Braun, B., Whirl-Carrillo, M., Cheung, K.-H., Dahlquist, K., Finney, A., Gillespie, M., Glass, E., Gong, L., Haw, R., Honig, M., Hubaut, O., Kane, D., Krupa, S., Kutmon, M., Leonard, J., Marks, D., Merberg, D., Petri, V., Pico, A., Ravenscroft, D., Ren, L., Shah, N., Sunshine, M., Tang, R., Whaley, R., Letovksy, S., Buetow, K. H., Rzhetsky, A., Schachter, V., Sobral, B. S., Dogrusoz, U., McWeeney, S., Aladjem, M., Birney, E., Collado-Vides, J., Goto, S., Hucka, M., Le NovÅšre, N., Maltsev, N., Pandey, A., Thomas, P., Wingender, E., Karp, P. D., Sander, C., and Bader, G. D. (2010). The biopax community standard for pathway data sharing. *Nature biotechnology*, **28**, 935–942.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.
- Dev, K. K. (2004). Making protein interactions druggable: targeting pdz domains. *Nature reviews Drug discovery*, **3**(12), 1047–1056.
- Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G. D., Michalickova, K., Pawson, T., and Hogue, C. W. V. (2003). Prebind and textomy—mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, **4**, 11.
- Doorbar, J. (2006). Molecular biology of human papillomavirus infection and cervical cancer. *Clinical science*, **110**(5), 525–541.
- Eisinger, D. P., Dick, F. A., Denke, E., and Trumpower, B. L. (1997). Sqt1, which encodes an essential wd domain protein of *saccharomyces cerevisiae*, suppresses dominant-negative mutations of the ribosomal protein gene *qsr1*. *Molecular and cellular biology*, **17**(9), 5146–5155.
- 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.
- Eom, J.-H. and Zhang, B.-T. (2006). Prediction of protein interaction with neural network-based feature association rule mining. In *Proceedings of the 13th international conference on Neural information processing - Volume Part III*, ICONIP’06, pages 30–39, Berlin, Heidelberg. Springer-Verlag.
- Erill, I. (2012). Information theory and biological sequences: insights from an evolutionary perspective. *International Journal of Evolution Equations*, **7**(3/4), 247.
- Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., *et al.* (2015). The reactome pathway knowledgebase. *Nucleic acids research*, **44**(D1), D481–D487.
- Faller, A. (1981). An average correlation coefficient. *Journal of Applied Metereology*, **203-205**, 20.
- Fernandez-Ballester, G., Beltrao, P., Gonzalez, J. M., Song, Y.-H., Wilmanns, M., Valencia, A., and Serrano, L. (2009). Structure-based prediction of the *saccharomyces cerevisiae* sh3-ligand interactions. *Journal of molecular biology*, **388**, 902–916.

- Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, **340**(6230), 245–246.
- Fukuhara, N. and Kawabata, T. (2008). HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. *Nucleic Acids Res*, **36**(Web Server issue), W185–W189.
- Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., *et al.* (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**(7084), 631–636.
- 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.
- Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- Gilchrist, M. A., Salter, L. A., and Wagner, A. (2004). A statistical framework for combining and interpreting proteomic datasets. *Bioinformatics*, **20**(5), 689–700.
- Goldberg, D. S. and Roth, F. P. (2003). Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A*, **100**(8), 4372–4376.
- 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.
- Guo, X., Liu, R., Shriver, C. D., Hu, H., and Liebman, M. N. (2006). Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, **22**(8), 967–973.
- Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res*, **36**(9), 3025–3030.
- Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, **46**(1-3), 389–422.
- Hofer, A., Bussiere, C., and Johnson, A. W. (2007). Mutational analysis of the ribosomal protein rpl10 from yeast. *Journal of Biological Chemistry*, **282**(45), 32630–32639.
- 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.
- Hue, M., Riffle, M., Vert, J.-P., and Noble, W. S. (2010). Large-scale prediction of protein-protein interactions from structures. *BMC bioinformatics*, **11**(1), 144.
- 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.
- Jaimovich, A., Elidan, G., Margalit, H., and Friedman, N. (2006). Towards an integrated protein-protein interaction network: a relational Markov network approach. *J Comput Biol*, **13**(2), 145–164.
- 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.
- Jain, S. and Bader, G. D. (2016). Predicting physiologically relevant sh3 domain mediated protein-protein interactions in yeast. *Bioinformatics*, **32**(12), 1865–1872.
- 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.
- Jensen, L. J., Gupta, R., Staerfeldt, H.-H., and Brunak, S. (2003). Prediction of human protein function according to gene ontology categories. *Bioinformatics*, **19**(5), 635–642.
- Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*, pages 9008+.
- Jin, J., Xie, X., Chen, C., Park, J. G., Stark, C., James, D. A., Olhovskiy, M., Linding, R., Mao, Y., and Pawson, T. (2009). Eukaryotic protein domains as functional units of cellular evolution. *Science signaling*, **2**(98), ra76–ra76.
- Jothi, R., Cherukuri, P. F., Tasneem, A., and Przytycka, T. M. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *Journal of molecular biology*, **362**(4), 861–875.
- 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.
- Kaufmann, K., Shen, N., Mizoue, L., and Meiler, J. (2011). A physical model for pdz-domain/peptide interactions. *Journal of molecular modeling*, **17**, 315–324.
- 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.

- Kim, T., Tyndel, M. S., Huang, H., Sidhu, S. S., Bader, G. D., Gfeller, D., and Kim, P. M. (2011). MUSI: an integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets. *Nucleic acids research*, page gkr1294.
- 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.
- Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., *et al.* (2006). Global landscape of protein complexes in the yeast *saccharomyces cerevisiae*. *Nature*, **440**(7084), 637–643.
- 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.
- Lee, H.-J. and Zheng, J. J. (2010). PdZ domains and their binding partners: structure, specificity, and modification. *Cell communication and signaling : CCS*, **8**, 8.
- Lee, I., Date, S. V., Adai, A. T., and Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *Science*, **306**(5701), 1555–1558.
- Lei, Z. and Dai, Y. (2006). Assessing protein similarity with gene ontology and its use in subnuclear localization prediction. *BMC Bioinformatics*, **7**, 491.
- 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.
- Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardoza, A. P., Santonico, E., Castagnoli, L., and Cesareni, G. (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res*, **40**(Database issue), D857–D861.
- Lin, D. (1998). An information-theoretic definition of similarity. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann.
- Lin, N., Wu, B., Jansen, R., Gerstein, M., and Zhao, H. (2004). Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, **5**, 154.
- 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.
- Lu, L., Lu, H., and Skolnick, J. (2002). MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, **49**(3), 350–364.

- 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.
- Martin, S., Roe, D., and Faulon, J.-L. (2005). Predicting protein-protein interactions using signature products. *Bioinformatics*, **21**(2), 218–226.
- Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D’Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic acids research*, **37**, D619–D622.
- 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.
- Mohamed, T. P., Carbonell, J. G., and Ganapathiraju, M. K. (2010). Active learning for human protein-protein interaction prediction. *BMC Bioinformatics*, **11 Suppl 1**, S57.
- Moncalián, G., Cárdenes, N., Deribe, Y. L., Spínola-Amilibia, M., Dikic, I., and Bravo, J. (2006). Atypical polyproline recognition by the cms n-terminal src homology 3 domain. *Journal of Biological Chemistry*, **281**(50), 38845–38853.
- Mongioví, A. M., Romano, P. R., Panni, S., Mendoza, M., Wong, W. T., Musacchio, A., Cesareni, G., and Di Fiore, P. P. (1999). A novel peptide–sh3 interaction. *The EMBO journal*, **18**(19), 5300–5309.
- 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.
- Najafabadi, H. S. and Salavati, R. (2008). Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biol*, **9**(5), R87.
- Nanni, L. and Lumini, A. (2006). An ensemble of k-local hyperplanes for predicting protein-protein interactions. *Bioinformatics*, **22**(10), 1207–1210.
- Nariai, N., Kolaczyk, E. D., and Kasif, S. (2007). Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS One*, **2**(3), e337.
- 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.

- Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine learning*, **39**(2-3), 103–134.
- 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.
- Overbeek, R., Fonstein, M., DSouza, M., Pusch, G. D., and Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, **96**(6), 2896–2901.
- Ozbabacan, S. E. A., Engin, H. B., Gursoy, A., and Keskin, O. (2011). Transient protein–protein interactions. *Protein Engineering Design and Selection*, **24**(9), 635–648.
- Patil, A. and Nakamura, H. (2005). Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics*, **6**, 100.
- 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.
- Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96**(8), 4285–4288.
- Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., and Orengo, C. (2010). Transient protein-protein interactions: structural, functional, and network properties. *Structure*, **18**(10), 1233–1243.
- Pesquita, C., Faria, D., Bastos, H., Falcao, A. O., and Couto, F. M. (2007). Evaluating GO-based semantic similarity measures. In *ISMB/ECCB 2007 SIG Meeting Program Materials, International Society for Computational Biology*.
- Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E. N., Falcao, A. O., and Couto, F. M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, **9 Suppl 5**, S4.
- Pesquita, C., Faria, D., Falcao, A. O., Lord, P., and Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, **5**(7), e1000443.
- Phizicky, E. M. and Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol Rev*, **59**(1), 94–123.
- Pires, J. R., Hong, X., Brockmann, C., Volkmer-Engert, R., Schneider-Mergener, J., Oschkinat, H., and Erdmann, R. (2003). The scpex13p sh3 domain exposes two distinct binding sites for pex5p and pex14p. *Journal of molecular biology*, **326**, 1427–1435.

- Pitre, S., Dehne, F., Chan, A., Cheetham, J., Duong, A., Emili, A., Gebbia, M., Greenblatt, J., Jessulat, M., Krogan, N., Luo, X., and Golshani, A. (2006). PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, **7**, 365.
- Pitre, S., Alamgir, M., Green, J. R., Dumontier, M., Dehne, F., and Golshani, A. (2008). Computational methods for predicting protein-protein interactions. *Adv Biochem Eng Biotechnol*, **110**, 247–267.
- 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.
- Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and SÄ©raphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, **24**(3), 218–229.
- Qi, Y., Klein-Seetharaman, J., and Bar-Joseph, Z. (2005). Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac Symp Biocomput*, **1**, 531–542.
- 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.
- Ramani, A. K., Bunescu, R. C., Mooney, R. J., and Marcotte, E. M. (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol*, **6**(5), R40.
- Ratanamahatana, C. a. and Gunopulos, D. (2003). Feature selection for the naive bayesian classifier using decision trees. *Applied artificial intelligence*, **17**(5-6), 475–487.
- Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC bioinformatics*, **9**(1), 405.
- Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g:Profiler – a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*, **35**(suppl 2), W193–W200.
- 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.
- Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *science*, **334**(6062), 1518–1524.
- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

- 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.
- Roy, S., Martinez, D., Platero, H., Lane, T., and Werner-Washburne, M. (2009). Exploiting amino acid composition for predicting protein-protein interactions. *PLoS One*, **4**(11), e7813.
- Saksela, K. and Permi, P. (2012). Sh3 domain ligand binding: What’s the consensus and where’s the specificity? *FEBS letters*, **586**(17), 2609–2614.
- Salah, Z., Alian, A., and Aqeilan, R. I. (2012). Ww domain-containing proteins: retrospectives and the future. *Frontiers in bioscience (Landmark edition)*, **17**, 331–348.
- 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.
- Sanchez, I. E., Beltrao, P., Stricher, F., Schymkowitz, J., Ferkinghoff-Borg, J., Rousseau, F., and Serrano, L. (2008). Genome-wide prediction of sh2 domain targets using structural information and the foldx algorithm. *PLoS computational biology*, **4**, e1000052.
- Schlessinger, J. (1994). SH2/SH3 signaling proteins. *Current opinion in genetics & development*, **4**(1), 25–30.
- Schlicker, A. and Albrecht, M. (2008). FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res*, **36**(Database issue), D434–D439.
- Schlicker, A., Domingues, F. S., Rahnenfuhrer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, **7**, 302.
- Scott, M. S. and Barton, G. J. (2007). Probabilistic prediction and ranking of human protein-protein interactions. *BMC Bioinformatics*, **8**, 239.
- Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martinez-Cruz, L. A., Corrales, F. J., and Rubio, A. (2005). Correlation between gene expression and go semantic similarity. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **2**(4), 330–338.
- Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol Syst Biol*, **3**, 88.
- Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A*, **104**(11), 4337–4341.
- Shen, R., Chinnaiyan, A. M., and Ghosh, D. (2008). Pathway analysis reveals functional convergence of gene expression profiles in breast cancer. *BMC Med Genomics*, **1**, 28.
- Sidhu, S. S., Bader, G. D., and Boone, C. (2003). Functional genomics of intracellular peptide recognition domains with combinatorial biology methods. *Curr Opin Chem Biol*, **7**(1), 97–102.

- 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.
- Smith, C. A. and Kortemme, T. (2010). Structure-based prediction of the peptide sequence space recognized by natural and synthetic pdz domains. *Journal of molecular biology*, **402**, 460–474.
- Smith, G. P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, **228**(4705), 1315–1317.
- Smithgall, T. E. (1995). Sh2 and sh3 domains: potential targets for anti-cancer drug design. *Journal of pharmacological and toxicological methods*, **34**(3), 125–132.
- Snel, B., Lehmann, G., Bork, P., and Huynen, M. A. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res*, **28**(18), 3442–3444.
- 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.
- Songyang, Z., Fanning, A. S., Fu, C., Xu, J., Marfatia, S. M., Chishti, A. H., Crompton, A., Chan, A. C., Anderson, J. M., and Cantley, L. C. (1997). Recognition of unique carboxyl-terminal motifs by distinct pdz domains. *Science (New York, N.Y.)*, **275**, 73–77.
- Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, **311**(4), 681–692.
- Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic acids research*, **34**(suppl_1), D535–D539.
- Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobisch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**(6), 957–968.
- Stiffler, M. A., Chen, J. R., Grantcharova, V. P., Lei, Y., Fuchs, D., Allen, J. E., Zaslavskaya, L. A., and MacBeath, G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. *Science*, **317**(5836), 364–369.
- Strasser, K., Masuda, S., Mason, P., Pfannstiel, J., Oppizzi, M., Rodriguez-Navarro, S., Rondon, A. G., Aguilera, A., Struhl, K., Reed, R., *et al.* (2002). Trex is a conserved complex coupling transcription with messenger rna export. *Nature*, **417**(6886), 304–308.
- Tamames, J., Casari, G., Ouzounis, C., and Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J Mol Evol*, **44**(1), 66–73.
- Tao, Y., Sam, L., Li, J., Friedman, C., and Lussier, Y. A. (2007). Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics*, **23**(13), i529–i538.

- Teyra, J., Sidhu, S. S., and Kim, P. M. (2012). Elucidation of the binding preferences of peptide recognition modules: Sh3 and pdz domains. *FEBS letters*, **586**, 2631–2637.
- Teyra, J., Huang, H., Jain, S., Guan, X., Dong, A., Liu, Y., Tempel, W., Min, J., Tong, Y., Kim, P. M., Bader, G. D., and Sidhu, S. S. (2017). Comprehensive analysis of the human sh3 domain family reveals a wide variety of non-canonical specificities. *Structure (London, England : 1993)*, **25**, 1598–1610.e3.
- The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nat Genet*, **25**(1), 25–29.
- Tian, L., Chen, L., McClafferty, H., Sailer, C. A., Ruth, P., Knaus, H.-G., and Shipston, M. J. (2006). A noncanonical sh3 domain binding motif links bk channels to the actin cytoskeleton via the sh3 adapter cortactin. *The FASEB journal*, **20**(14), 2588–2590.
- 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., Boone, C., and Sidhu, S. S. (2007). Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries. *Nat Protoc*, **2**(6), 1368–1386.
- 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.
- Trabuco, L. G., Betts, M. J., and Russell, R. B. (2012). Negative protein–protein interaction datasets derived from large-scale two-hybrid experiments. *Methods*, **58**(4), 343–348.
- Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., Morrison, K., Donaldson, I. M., and Wodak, S. J. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford)*, **2010**, baq023.
- 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.
- Van Rossum, G. and Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

- 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.
- von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**(6887), 399–403.
- Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of go terms. *Bioinformatics*, **23**(10), 1274–1281.
- Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004). The disopred server for the prediction of protein disorder. *Bioinformatics*, **20**(13), 2138–2139.
- 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.
- Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *nature*, **393**(6684), 440–442.
- West, M., Hedges, J. B., Chen, A., and Johnson, A. W. (2005). Defining the order in which nmd3p and rpl10p load onto nascent 60s ribosomal subunits. *Molecular and cellular biology*, **25**(9), 3802–3813.
- Wintjens, R., Wieruszkeski, J. M., Drobecq, H., Rousselot-Pailley, P., BuÅ©e, L., Lippens, G., and Landrieu, I. (2001). 1h nmr study on the binding of pin1 trp-trp domain with phosphothreonine peptides. *The Journal of biological chemistry*, **276**, 25150–25156.
- Wu, H., Su, Z., Mao, F., Olman, V., and Xu, Y. (2005). Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic Acids Res*, **33**(9), 2822–2837.
- 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.
- Xia, K., Dong, D., and Han, J.-D. J. (2006). IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics*, **7**, 508.
- 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).
- Xu, T., Du, L., and Zhou, Y. (2008). Evaluation of go-based functional similarity measures using *S. cerevisiae* protein interaction and expression profile data. *BMC Bioinformatics*, **9**, 472.
- Yamanishi, Y., Vert, J.-P., and Kanehisa, M. (2004). Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, **20 Suppl 1**, i363–i370.
- Yu, C.-Y., Chou, L.-C., and Chang, D. T.-H. (2010). Predicting protein-protein interactions in unbalanced data using the primary structure of proteins. *BMC Bioinformatics*, **11**, 167.

- Yu, H., Gao, L., Tu, K., and Guo, Z. (2005). Broadly predicting specific gene functions with expression similarity and taxonomy similarity. *Gene*, **352**, 75–81.
- Yu, H., Paccanaro, A., Trifonov, V., and Gerstein, M. (2006). Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, **22**(7), 823–829.
- Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabási, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**(5898), 104–110.
- Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X., *et al.* (2011). Next-generation sequencing to generate interactome datasets. *Nature methods*, **8**(6), 478–480.
- Zafra-Ruano, A. and Luque, I. (2012). Interfacial water molecules in sh3 interactions: Getting the full picture on polyproline recognition by protein–protein interaction domains. *FEBS letters*, **586**(17), 2619–2630.
- Zhang, L. V., Wong, S. L., King, O. D., and Roth, F. P. (2004). Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, **5**, 38.
- Zhang, P., Zhang, J., Sheng, H., Russo, J. J., Osborne, B., and Buetow, K. (2006). Gene functional similarity search tool (GFSST). *BMC Bioinformatics*, **7**, 135.
- Zhang, Q. C., Petrey, D., Garzón, J. I., Deng, L., and Honig, B. (2012a). PrePPI: a structure-informed database of protein–protein interactions. *Nucleic acids research*, page gks1231.
- Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., *et al.* (2012b). Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature*, **490**(7421), 556–560.