# Computational prediction of protein interaction networks through supervised classification techniques

Fiona Browne, Haiying Wang* and Huiru Zheng

School of Computing and Mathematics,
University of Ulster at Jordanstown,
Northern Ireland, UK
E-mail: browne-f@ulster.ac.uk
E-mail: hy.wang@ulster.ac.uk
E-mail: h.zheng@ulster.ac.uk
*Corresponding author

Francisco Azuaje

Research Centre for Public Health (CRP-Santé)
1AB rue Thomas Edison, Strassen L-1445, Luxembourg
E-mail: fj.azuaje@ieee.org

**Abstract:** This paper implements integrative methods to predict Pairwise (PW) and Module-Based (MB) protein interactions in *Saccharomyces cerevisiae*. The predictive ability of combining diverse sets of relatively strong and weak predictive datasets is investigated. Different classification techniques: Naive Bayesian (NB), Multilayer Perceptron (MLP) and K-Nearest Neighbors (KNN) were evaluated. The assessment demonstrated that as the predictive power of single-source datasets became weaker, MLP and NB performed better than KNN. Generation of PPI maps for *S. cerevisiae* and beyond will be improved with new, higher-quality datasets with increased interactome coverage and the integration of classification methods.

**Keywords:** Protein-Protein Interactions; PPIs; module-based interactions; machine and statistical learning; functional data; feature encoding; dataset integration; computational systems biology.

**Reference** to this paper should be made as follows: Browne, F., Wang, H., Zheng, H. and Azuaje, F. (2008) 'Computational prediction of protein interaction networks through supervised classification techniques', *Int. J. Functional Informatics and Personalised Medicine*, Vol. 1, No. 2, pp.205–221.

**Biographical notes:** Fiona Browne is a PhD student at the University of Ulster at Jordanstown. She received a BSc Hons (I) in Computer Science from the University of Ulster in 2004. Her research interests include bioinformatics, artificial intelligence, supervised machine learning and data mining. She has published research papers at international journals and conference proceedings.

Haiying Wang received the BEng and MSc Degrees in Optical Electronics Engineering from Zhejiang University, Hangzhou, China, in 1987 and 1989

respectively. He was a senior engineer in Applied Electronics at the Fujian Electronic Technology Institute, Fuzhou, China and received the PhD Degree on Artificial Intelligence in Biomedicine from the University of Ulster, Jordansotown, UK, in 2004. He is currently a Lecturer in the School of Computing and Mathematics at the University of Ulster. His research focuses on artificial intelligence, machine learning, pattern discovery and visualisation, XML, and their applications in medical informatics and bioinformatics.

Huiru Zheng received the BEng Degree in Biomedical Engineering from Zhejiang University, China in 1989, the MSc Degree in Communication and Electronic System from Fuzhou University, China in 1992, and the PhD Degree on Data Mining and Bioinformatics from the University of Ulster, UK in 2003. Before she joined the University of Ulster, she was working in Fuzhou University as an Assistant Lecturer (1992), Lecturer (1995) and Associate Professor (2000). Her research interests include biomedical engineering, medical informatics, bioinformatics, data mining and artificial intelligence. She has 50+ publications in journals and conferences in these areas. She is currently a Lecturer in the Faculty of Computing and Engineering at the University of Ulster.

Francisco Azuaje received the BSc Degree in Electronic Engineering from Simon Bolivar University, Caracas, Venezuela, in 1995. He received the PhD in artificial intelligence and medical informatics from the University of Ulster, UK, in 2000. Before joining the CRP-Sante in Luxembourg, he was a reader with the University of Ulster, and a Lecturer with Trinity College Dublin, Ireland. He has extensively published in journals, books and conference proceedings relating to the areas of bioinformatics, artificial intelligence, and medical informatics. He has co-edited three books in the areas of bioinformatics and medical informatics. He is a senior member of the IEEE.

## 1 Introduction

Protein-Protein Interactions (PPIs) play an important role in biological processes. Most proteins perform their functions by interacting with other proteins. Several large-scale PPI maps have been produced for *Saccharomyces cerevisiae* and other organisms, such as *Drosophila melanogaster* and *Homo sapiens* from experimental high-throughput methods (Uetz et al., 2000; Gavin et al., 2006; Krogan et al., 2006). However, even the best studied model organisms contain a large number of proteins whose functions are currently unknown. This highlights the continued need for computational methods to help to direct scientists in the search of novel interactions.

### *1.1 High-throughput experimental methods*

The development of high-throughput experimental methods and completion of genome sequencing projects have accelerated the pace of discovery of PPI. Common experimental methods include the yeast-two hybrid screen, Tandem Affinity Purification (TAP), Mass Spectrometry (MS) and protein chips. However, these data are often incomplete, contradictory and noisy with thousands or tens of thousands interactions yet unknown. Experimental methods can only identify a subset of the interactions that occur in an organism. Therefore, coverage (i.e., the area of the genome covered by protein

pairs) of the interactome (the collection of all the PPI that occur within a cell) is limited (Browne et al., 2006). Methods such as the yeast-two hybrid system exhibit high False Positive (FP) and false negative interaction rates.

Traditional small scale experiments for PPI prediction produce more accurate results compared to single source high-throughput methods. However, they are limited in terms of coverage of the interactome. Due to the inadequacies exhibited by both the traditional, experimental and computational methods, we and others argue that more advanced computational integrative methods are essential to predict PPI (Lu et al., 2005).

### *1.2 Datasets*

Protein interaction datasets are obtained from a variety of different experimental and computational sources. These datasets have intrinsic problems such as missing data and relatively high levels of false negative and FPs predictions. Data from a single (experimental or computational) predictive source should be viewed with caution. The inference of a PPI network could be improved by integrating different genomic features (i.e., datasets) (Jansen et al., 2003; Lu et al., 2005; Troyanskaya et al., 2003). When multiple diverse datasets support a prediction, then the confidence in this prediction increases. Different features (e.g., protein abundance) may cover different areas of the interactome. It has been demonstrated that the predictive integration of datasets can increase predictive coverage and reduce FP predictions (Jansen et al., 2004). The integration of datasets is not a trivial task. There is no standard approach to meaningfully represent predictive features from the datasets.

Datasets that do not directly measure PPI, such as sequence, structural and diverse functional genomics information can also be used to predict PPI. For example, the application of gene CO-Expression (COE) to infer PPI is based on the hypothesis that proteins found in the same complex are often co-expressed. Therefore, dataset selection is crucial for the prediction of PPI in order to improve interactome coverage and accuracy.

### *1.3 Computational PPI prediction approaches*

Pairwise (PW) interaction prediction has been the most prominent PPI network prediction principle reported to date (Myers et al., 2005). A PW interaction involves the interaction between two proteins. Limited research has been performed in the area of supervised Module-Based (MB) interaction network prediction. A MB prediction approach aims to detect whether (or not) a group of proteins (rather than a pair of proteins) belongs to the same protein complex. Definitions of known protein complexes and proteins that exist within these complexes can be obtained from the publicly available sources, such as the MIPS complex catalogue (Mewes et al., 2002) for different organisms. Most cellular activities involve groups of genes or gene products that behave in a coordinated way to perform a specific biological process (Myers et al., 2005). With the availability of high-throughput large-scale datasets, a wealth of information is available to discover PPI networks. The bulk of this data is currently used for the prediction of pair-wise interactions. Exploiting the full potential of diverse PPI data, one could discover MB networks, which represent higher-level representations of PPI networks (Myers et al., 2005). A study by Myers et al. (2005) used a Bayesian Network (BN) and a probabilistic graph search algorithm to analyse MB and PW predictions. The datasets in their study included genetic and physical interaction data obtained from the BIND and GRID

databases, gene expression data, cellular localisation data, protein complex data and sequence data for the *S. cerevisiae* organism. The researchers discovered that MB predictions tend to be more precise than PW predictions.

### *1.4 Gold Standards (GS)*

Selecting a GS is an essential task in PW and MB prediction. A GS is a dataset consisting of a number of known interacting and non-interacting proteins, which are used to train classifiers or to estimate their predictive ability. Using the MB approach, construction of the GS theoretically represents a more complex conceptual problem in comparison to PW prediction. The quality of the prediction methods will depend on the relevance and validity of the GS to the prediction problem under study. In this study, a GS consists of 'positive' cases representing groups of (two or more) proteins found in the same protein complex. The task of selecting a negative GS (i.e., non-interacting protein pairs) represents a significant challenge. In this study, non-interacting proteins are based on the assumption that protein pairs or groups of proteins belonging to different subcellular locations and complexes are unlikely to be interacting proteins. The difficulty in defining a negative class for a GS is one of the root causes for the poor or, in some cases, overestimated performance of machine learning algorithms in the prediction of PPI (Jansen et al., 2004). Another problem is to select a GS that has an adequate coverage of the interactome.

### *1.5 Related research*

Computational methods are required to integrate disparate high-throughput biological data for PW and MB interaction prediction. Therefore, selection of appropriate machine or statistical learning techniques is vital. Classifiers that perform well in other problem domains may not perform as well within the realm of PW or MB prediction. Classifiers exhibit systematic bias (i.e., a method produces solutions that highly favour a limited number of specific situations) or are based on major assumptions (e.g., independence between datasets). Such potential bias and assumptions may lead to systematic prediction errors (Browne et al., 2006). In this study we evaluated three computational classification techniques to integrate diverse sources of information for PW and MB prediction. Traditional statistical and machine learning methods have been applied to integrate diverse sources of data for PPI prediction. For example Jansen et al. (2003) and Troyanskaya et al. (2003), both applied a BN approach to predicting PPI by integrating genomic data. These frameworks performed a probabilistic weighting of diverse data sources. The data sources individually were weak predictors of PPI. However, when integrated, these studies produced accurate PPI networks providing a comprehensive view of the *S. cerevisiae* interactome. Barutcuoglu et al. (2006) recently developed a probabilistic, query-based system to discover MB interaction networks by integrating diverse genome-wide data sources. This system is based on BN and was validated by accurately recovering networks for 31 known biological processes in *S. cerevisiae*. The research by Jansen et al. (2003) was consequently extended by Lu et al. (2005). Lu et al. (2005) focused on assessing the predictive limits of genomic data integration. Naive Bayesian (NB) was used to integrate 16 diverse datasets. As with a previous study, relatively high predictive accuracies were obtained. However, the addition of relatively weaker datasets only marginally improved the predictive power of the models.

The NB classifier 'naively' assumes conditional independence among datasets (features). Interestingly, Lu et al. (2005) provided evidence of no conditional independence existing between the datasets used within their study. However, as high-throughput technologies continue to emerge, datasets produced will overlap with one another. Recent research by Xia et al. (2006) has demonstrated how logistic regression outperforms the bench mark NB in predicting the PPI network of the helical membrane protein interactome in *S. cervisiae*. The Random Forest (RF) machine learning method has been utilised by Qi et al. (2006) to predict a PPI network in *S. cerevisiae*. The RF classifier used in this study predicted PPI with an average sensitivity of around 80% and specificity below 65% Qi et al. (2006), demonstrated the effect of dataset selection and encoding on the PPI predictive performance using different classification techniques (RF, RF integrated with KNN, NB, Decision Tree, Logistic Regression and Support Vector Machines). This study found that RF performed robustly in general among the six classifiers in inferring PPI in all three subtasks. This was due to the ability of RF to integrate diverse sources of data. RF does not assume feature independence and is robust against noise and missing values. In a study by Browne et al. (2006) NB, Multi-Layer Perceptron (MLP) and K-Nearest Neighbour (KNN) were evaluated in the task of PPI prediction using the PW approach. The classifier NB achieved the 'highest' predictive performance by obtaining an area under Receiver Operating Characteristic (ROC) curve (AUC) value of 0.99. The lowest AUC value of 0.90 was obtained by the KNN classifier. A recent study by Collins et al. (2007) merged two affinity purification/MS studies (Gavin et al., 2006; Krogan et al., 2006) in *S. cerevisiae* into a single reliable collection of experimentally-based PPIs by analysing the primary affinity purification data using a novel Purification Enrichment (PE) scoring system. Using a GS obtained from the Munich Database of Interacting Proteins (MIPS) (Mewes et al., 2002) and Saccharomyces Genome Database (SGD) their study demonstrated that the consolidated dataset is of greater accuracy than the individual sets and is comparable to PPIs defined using more reliable, small-scale experimental methodologies. By applying the PE metric and a stringent cut-off, a set of 9074 interactions were identified (including 4456 which were not among the 12,122 interactions) (Collins et al., 2007).

Throughout this paper we use the following key terms: dataset, cases and features. Datasets are used as inputs into prediction models. A dataset contains n number of cases. Each case is made up of *m* number of features (each one representing, for example Marginal Essentiality (MES), protein abundance). For example, using the PW approach a case will encode information on two proteins. Features may be referred to as being strong and weak. We define a strong dataset as a predictive resource that covers a larger proportion of the GS (compared to weaker sources) and containing a relatively small number of FP predictions and false negative interacting pairs in relation to the GS.

This paper discusses key predictive performance differences between the techniques. We also compared the results of these classifiers to determine if relatively simpler classifiers may outperform more complex classifiers. We investigated the most effective and reliable prediction models. A discussion on the impact individual predictive features have on prediction accuracy is presented. Finally, we conclude the paper with some recommendations for the design and application of PW and MB prediction approaches and outline current and future work.

## 2 Data sources and Gold Standard

In this study, seven diverse data sources (encoding different input features) were analysed and integrated for PW prediction. For the MB approach only six diverse features were generated and integrated. The PE dataset obtained from the study by Collins et al. (2007) was not integrated as we were unable to generate a sufficient amount of data from this data source. Five of the features 2.1–2.5 were used in a previous study by Lu et al. (2005). The GO-driven PPI dataset was obtained from our previous study (Browne et al., 2006) and the PE dataset integrated for PW prediction was acquired from a recent study by Collins et al. (2007). Both the MIPS Functional Catalogue (FUNCAT) and GO-driven semantic similarity (GOSEM) are high quality annotation datasets and independent from the GS; however, they are limited in terms of coverage of the interactome. In this research we integrate diverse features to improve coverage. A brief description on how these datasets were obtained, along with the rationale for applying them are presented below. The dataset names have been shortened for easier representation within the paper.

### *2.1 mRNA CO-Expression*

This dataset is based on the assumption that proteins found in the same complex interact, and proteins belonging to the same complex are often co-expressed. This dataset has been constructed from publicly-available expression data (Cho et al., 1998). It represents the time course of expression fluctuations during the yeast cell cycle and the Rosetta compendium, consisting of the expression profiles of 300 deletion mutants and cells under different chemical treatments (Cho et al., 1998). Pearson's correlation values were calculated for each gene pair. The results range from –0.9 to 1.

### *2.2 FunCat*

It is assumed that two proteins acting in the same biological process are more likely to interact than two proteins involved in different processes. Therefore, two proteins are defined to be interacting if they belong to the same biological process. Non-interaction proteins are defined as two proteins that do belong to the same biological process as defined by the Functional Catalogue of MIPS (FunCat) (Mewes et al., 2002). The FunCat is an annotation scheme that contains data on the functional description of proteins from prokaryotes, unicellular eukaryotes, plants and animals (Mewes et al., 2002). The FunCat is separate from the MIPS complex catalogue, which represents the GS in this study. The results within this dataset range from 0 to 7. Please refer to Jansen et al. (2003) for detail on the similarity metric.

### *2.3 GOSEM*

The Lin's semantic similarity technique was used to compute the similarity between GO terms and gene products annotated in the SGD. The gene-pair similarity values provide the PPI predictions in the GOSEM dataset. This similarity method uses both the information content of shared GO term parents, and that of the query GO terms used to annotate a gene. This similarity is based on the number of times each term, or any child term, occurs in the GO corpus (SGD) and is expressed as a probability. Lin's technique estimates the similarity between two terms as the ratio between the information content

of the minimum subsumer. Pairs of genes described by more general (less specific) GO terms, will tend to show similarity values closer to zero. The value produced is a normalised similarity value between 0 and 1. A more detailed description of Lin's semantic similarity technique and their relationship with other functional properties can be found in Azuaje et al. (2005).

### *2.4 Co-essentiality (ESS)*

This dataset is derived from the MIPS complex catalogue and also from transposon and gene deletion experiments (Mewes et al., 2002). The hypothesis is that proteins can be experimentally characterised as Either Essential (EE) or Non-essential (NN), which may be used an indicator that the proteins are both members of the same complex. If two proteins exist in the same complex they are EE or NN but not both. In this dataset if both protein pairs are EE or NN then they are assumed to interact together. However, if they are a mixture of essential and NN proteins then the protein pair is said not to interact. In this dataset, NN, are represented by 0, mixture of NN and EE are represented by 1 and EE are represented by 2.

### *2.5 Marginal essentiality*

This is a quantitative measure of the importance of a NN gene to a cell. It is based on the 'marginal benefit' hypothesis that many NN genes make significant but small contributions to the fitness (i.e., health and performance of a cell) of a cell. This dataset was obtained through quantitatively combining the results from four large-scale phenotypic experiments (e.g., growth rate inhibition from knockouts), that examined different aspects of the impact of a protein on cell fitness (Yu et al., 2004). Protein pairs are defined as interacting if two proteins have a higher combined MES. It has been suggested that essential proteins have an average degree (number of links per protein) of 18.7, a clustering co-efficient of 0.182, a characteristic path length (average distance between proteins) of 3.84 and a diameter (maximum inter-protein distance) of 10 (Yu et al., 2004). NN proteins are defined as having an average degree of 7.4, a clustering co-efficient of 0.095, a characteristic path of 4.49 and a diameter of 11 (Yu et al., 2004). The values generated range from –0.9 to –27.

### *2.6 Absolute Protein Abundance (APA)*

APA datasets have been scaled and merged from available yeast protein-abundance datasets. Protein abundance was obtained through a number of experimental methods: gel electrophoresis and several mass spectrometry approaches with varying degrees of accuracy. These datasets were merged and made available by Greenbaum et al. (2002). Protein abundance can be calculated by counting the number of proteins within a cell. If the concentration of protein and their interactions are true contributory forces in the cell then it is important to know the corresponding protein quantities. The hypothesis applied to this dataset states that two proteins interacting should be present in stoichiometrically (the calculation of the quantities of reactants and products in a chemical reaction) similar amounts. The paper by Greenbaum et al. (2002) details the relationship between of mRNA expression and protein abundance data. The values generated by this dataset range from 0 to 20.

### 2.7 Absolute mRNA expression (EXP)

For PPI, EXP uses a similar assumption as in Section 2.7. EXP has often been used as a surrogate for APA. Substantial agreement between these two datasets has been found (Greenbaum et al., 2002). EXP is an approximation of absolute expression levels of mRNA within a cell. The values generated by this dataset range from 0 to 10.

### 2.8 Purification enrichment

This dataset contains PPI data from two studies. The affinity purification experimental method was utilised in both studies to produce the data. A single dataset was constructed by merging the experimentally-based PPI's using a novel purification enrichment scoring systems by (Collins et al., 2007) The values generated from this dataset range from 0 to 26.

### 2.9 The construction of GS

In this study we constructed GSs for PW and MB approaches. In the MB approach, three GSs were produced for cases described by three, four and five proteins. Both the positive (i.e., interacting) and negative (i.e., non-interacting) sets were based on information obtained from the MIPS protein complex catalogue (minimum size of complex: five proteins) (Mewes et al., 2002). This catalogue was chosen as it contains lists of known protein complexes based on data collected from validated, small-scale studies obtained from biomedical literature (Mewes et al., 2002). Our PW GS contains 10,802 protein pairs in the positive and 330,000 protein pairs in the negative GS. Our MB GSs contain 11,000 protein groups in the positive and 330,000 protein groups in the negative GS.

## 3 Methods

Different PPI prediction models based on NB, KNN and MLP classifiers were chosen to integrate the diverse datasets. We aimed to show how these methods differ in predictive accuracy when using different feature encoding methods, integration methods and diverse features. Each classifier was built using the YALE toolbox (Mierswa et al., 2006) and 10-fold cross-validations (unless otherwise indicated e.g., Leave-One-Out (LOO) validation due to small data size) were performed to estimate the predictive performance. The values in the datasets were linearly normalised between 0 and 1, which represented the inputs to the prediction models. The predictive performance of the classifiers was measured using the known class assignments derived from the GS.

Unlike previous research, our does not focus on supervised PPI PW prediction (Lu et al., 2005; Browne et al., 2006). We propose and evaluate different encoding strategies, as well as supervised classification models, for MB prediction. For the MB prediction method we contrasted two feature encoding techniques: the feature mean and median-based techniques. These features and encoding techniques are described below.

### *3.1 Integration techniques*

In this study two types of techniques were applied to integrate the diverse features.

- *'Full' integration*. A dataset will only consist of cases containing complete descriptions of the features.
- *'Incomplete' integration*. A dataset will consist of cases that may contain incomplete feature descriptions. In this study, when a case contains incomplete feature descriptions, cases with less than four feature descriptions were removed.

To the best of our knowledge, these types of integration methods have not previously been contrasted for the task of MB prediction.

### *3.2 Feature encoding techniques*

PW interaction data were transformed and extended to generate cases for MB datasets. All unique PW combinations between the proteins were determined. For each unique PW combination, a feature value was obtained. For example, using the feature FUNCAT and a case consisting of a group of four proteins, six FUNCAT values would be obtained. We then calculated the mean and the median of these values to produce single values for each feature in the MB datasets. For cases where only some or no feature values were obtained for each PW combination, the feature value was depicted as missing.

### *3.3 Classification techniques*

*KNN*. In terms of mathematical complexity we regard traditional KNN as the simplest method assessed in this investigation. KNN has previously been used in the prediction of PW PPI by (Browne et al., 2006). KNN classified cases as interacting or non-interacting by taking each new instance (test dataset) and comparing it with existing instances (in training dataset) using the Euclidean distance metric. In this research, $K$ was set at 3, as this produced the best predictive performance results.

*MLP*. Is a non-linear neural network classification approach that is trained using the back propagation algorithm. In this investigation the network has three layers: The input layer –where the datasets are input; 1 hidden layer and the output layer. The results reported in this paper were obtained by setting the learning rate at 0.3, and the momentum at 0.2. The number of training epochs was equal to 500. Within the hidden layer, the number of hidden nodes was the defined to be equal to the average value of the number of features and classes. In terms of mathematical complexity we regard, MLP as the most complex of the three classifiers.

*NB*. Has previously been evaluated by (Lu et al., 2005) to combine diverse genomic features. Therefore, NB is being evaluated here as a benchmark to compare the other classifiers of varying complexity. NB is based on the Bayes rule of conditional probability. Given the class membership of an instance, the probability of observing a combination of variables is the product of the probabilities for the individual variables. NB is regarded 'Naïve' as it 'naively' assumes independence between features (datasets). Due to this assumption, the predictive power of NB may be reduced if a dataset is highly correlated with an existing dataset. NB can handle diverse heterogeneous sources of data.

NB ignores missing values while determining priors. Although this technique is relatively simple in terms of mathematical complexity, relatively high prediction accuracies have been obtained by several PW PPI studies (Lu et al., 2005; Browne et al., 2006).

### *3.4 ROC curves and analysis of statistical significance*

ROC curves were used to graphically plot the sensitivity vs. (1 – specificity) obtained by the classifiers as the discrimination threshold is varied. The fraction of True Positives (TP) and FPs were obtained from the 10-fold cross-validation analysis. AUC is the measurement of the total area under the ROC curve. In this study, the average AUC value is obtained from the 10-fold cross-validation.

Analysis of Variance (ANOVA) and the paired samples Student *t*-test (two-tailed) analysis were applied to determine if significant differences existed between the classification models in terms of predictive quality (AUC values) in this study. ANOVA tells us whether the factors (e.g., classification techniques) significantly contribute to the variations observed in the prediction outcomes. The AUC values obtained from the cross-validation procedures (ten values) were used in the ANOVA and *t*-test. ANOVA and the *t*-test were performed using the statistical package SPSS version 11.5 (SPSS Inc., 2005). Results are presented using $F$ values and $p$ values. A significant difference is observed when the $p$ value obtained was less than 0.05.

## 4 Results

In this section we present and discuss results obtained using the PW and MB prediction approaches in *S. cerevisiae*. Each case in a PW dataset consists of two proteins and seven features. Three different MB approaches were evaluated. Each case in the MB_3 datasets contains three proteins, a case in the MB_4 datasets contains four proteins and a case in the MB_5 datasets contains five proteins. In the MB approach each case consists of six features (as described in Section 2). For each MB approach, two datasets are constructed using the mean and median feature encoding techniques. In this section, we compared the prediction performance (in terms of AUC values obtained by the classifiers) of three computational classification techniques: KNN, MLP and NB for PW and MB interaction prediction. Each classifier performs the task of predictive integration of diverse features described in Section 2. We contrasted prediction performance when datasets contain cases with full description of the features compared to cases where incomplete feature descriptions were available. Each classifier was built using measurements obtained from each dataset as the inputs to the models. The 'true' categorisations for each case (i.e. protein pair or groups of proteins) were obtained from the GS.

### *4.1 Classification performance: PW approach*

Table 1 exhibits the AUC values obtained by the predictive models based on single- source features using the different classification approaches. Three classifiers were applied to each individual feature (i.e., single input classifiers). Significant differences ($F = 9.8$, $p = 0.001$) were observed between the classification models using ANOVA in terms of their predictive performance (AUC values) when applied to the

GOSEM feature. Observing the AUC values from Table 1, classifiers constructed using the FUNCAT and GOSEM features obtained higher AUC values compared to classifiers constructed using the features APA and MES. The ROC curves in Figure 1 depict the predictive power of NB classification models based on individual input features. Panel (a) in Figure 1 represents the PW prediction approach.

The predictive performance of the classifiers constructed by the integration of diverse features was evaluated. The AUC values obtained by each classifier constructed using datasets containing full and incomplete feature descriptions can be found in Tables 2 and 3. The ROC curve in Figure 2(a) depicts the predictive power of each classification method when applied to datasets containing incomplete features (for PW interaction prediction). In Table 2 it can be viewed when using the PW approach, all classifiers obtained high AUC values (0.97–0.98) when built using the PW_Full dataset. Using Table 3, no significant difference was observed (in terms of AUC value) between the classifiers NB, MLP and KNN using ANOVA ($F = 2.314$, $p = 0.188$). Classifiers constructed using the PW_Incomplete dataset obtained marginally lower AUC values (0.89–0.97) compared to classifiers constructed using the PW_Full dataset. Both MLP and NB are the best classifiers (in terms of AUC values) when applied to the PW_Full and PW_Incomplete datasets.

**Table 1** AUC values obtained from all classifiers and datasets using individual input features

| | | *Features* | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Datasets* | *Classifiers* | *APA* | *COE* | *EXP* | *FUNCAT* | *GOSEM* | *MES* | *PE* |
| PW | KNN | 0.56 | 0.78 | 0.63 | 0.98 | 0.96 | 0.63 | 0.88 |
| | MLP | 0.56 | 0.83 | 0.64 | 0.93 | 0.95 | 0.62 | 0.93 |
| | NB | 0.57 | 0.84 | 0.66 | 0.93 | 0.95 | 0.65 | 0.91 |
| MB_3_Mean | KNN | 0.75 | 0.62 | 0.83 | 0.98 | 0.91 | 0.54 | |
| | MLP | 0.63 | 0.74 | 0.79 | 0.98 | 0.94 | 0.64 | |
| | NB | 0.64 | 0.75 | 0.80 | 0.99 | 0.95 | 0.63 | |
| MB_3_Median | KNN | 0.71 | 0.76 | 0.80 | 0.94 | 0.97 | 0.76 | |
| | MLP | 0.62 | 0.73 | 0.78 | 0.95 | 0.94 | 0.65 | |
| | NB | 0.63 | 0.75 | 0.79 | 0.95 | 0.95 | 0.63 | |
| MB_4_Mean | KNN | 0.6 | 0.66 | 0.7 | 0.99 | 0.89 | 0.6 | |
| | MLP | 0.64 | 0.77 | 0.84 | 0.99 | 0.9 | 0.69 | |
| | NB | 0.69 | 0.78 | 0.87 | 0.99 | 0.93 | 0.69 | |
| MB_4_Median | KNN | 0.65 | 0.69 | 0.77 | 0.97 | 0.9 | 0.6 | |
| | MLP | 0.65 | 0.78 | 0.85 | 0.97 | 0.94 | 0.69 | |
| | NB | 0.69 | 0.79 | 0.86 | 0.96 | 0.96 | 0.69 | |
| MB_5_Mean (LOO) | KNN | 0.82 | 0.97 | 0.84 | 0.99 | 0.99 | 0.84 | |
| | MLP | 0.57 | 0.83 | 0.65 | 0.99 | 0.93 | 0.49 | |
| | NB | 0.64 | 0.86 | 0.65 | 0.99 | 0.99 | 0.52 | |
| MB_5_Median (LOO) | KNN | 0.86 | 0.93 | 0.86 | 0.92 | 0.99 | 0.90 | |
| | MLP | 0.51 | 0.82 | 0.69 | 0.90 | 0.99 | 0.56 | |
| | NB | 0.58 | 0.77 | 0.70 | 0.90 | 0.99 | 0.60 | |

**Figure 1** ROC curves obtained by classifier NB when built using individual features from datasets: (a) PW; (b) MB_3_Median; (c) MB_4_Median and (d) MB_5_Median (see online version for colours)

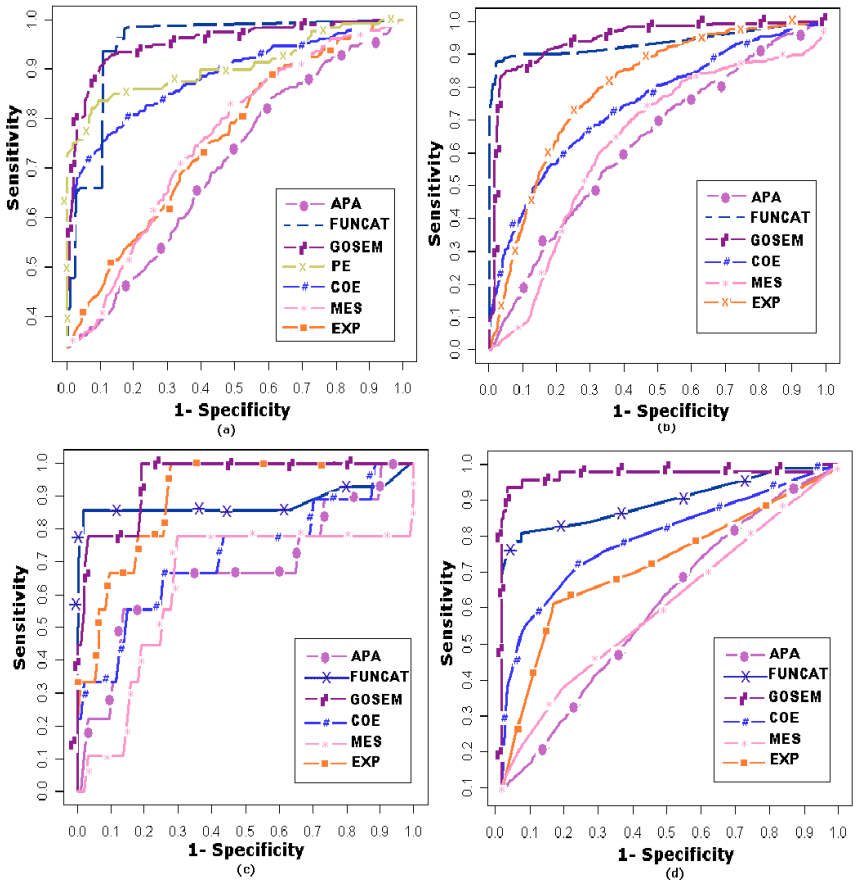


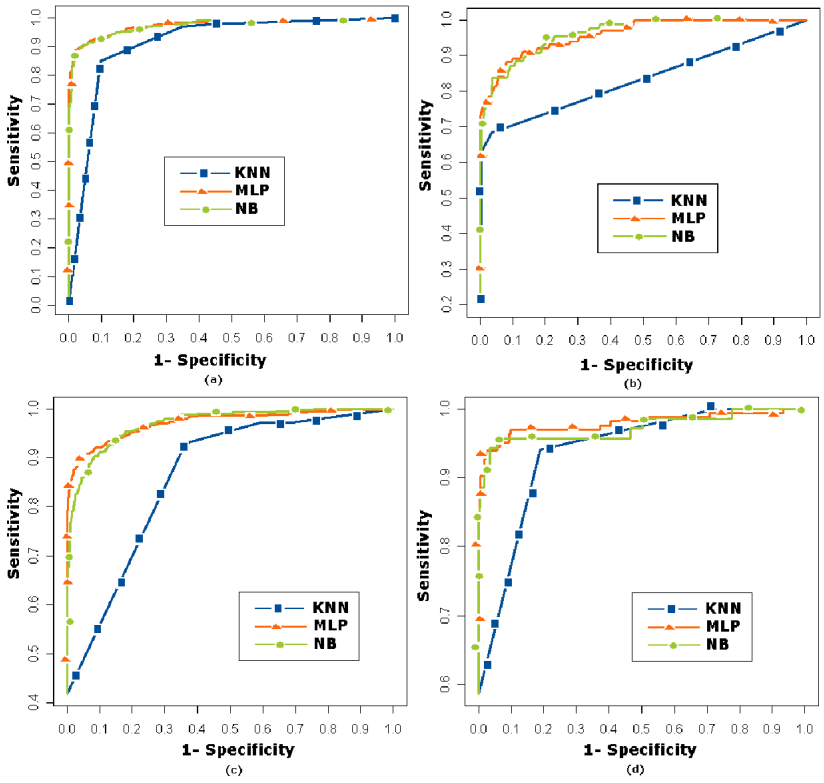

**Table 2** AUC results obtained by classifiers integrating features with full descriptions

| *Datasets* | *Classifiers* | | |
|---|---|---|---|
| | *KNN* | *MLP* | *NB* |
| PW_Full | 0.97 | 0.98 | 0.98 |
| MB_3_Mean_Full | 0.97 | 0.99 | 0.99 |
| MB_3_Median_Full | 0.97 | 0.97 | 0.97 |
| MB_4_Mean_Full | 0.97 | 0.99 | 0.99 |
| MB_4_Median_Full | 0.98 | 0.99 | 0.98 |
| MB_5_Mean_Full (LOO) | 0.99 | 0.99 | 0.99 |
| MB_5_Median_Full (LOO) | 0.93 | 0.92 | 0.98 |

**Table 3** AUC results obtained by classifiers integrating features with incomplete descriptions

| *Datasets* | *Classifiers* | | |
|---|---|---|---|
| | *KNN* | *MLP* | *NB* |
| PW_Incomplete | 0.89 | 0.97 | 0.97 |
| MB_3_Mean_Incomplete | 0.83 | 0.94 | 0.96 |
| MB_3_Median_Incomplete | 0.87 | 0.93 | 0.96 |
| MB_4_Mean_Incomplete | 0.70 | 0.95 | 0.97 |
| MB_4_Median_Incomplete | 0.69 | 0.95 | 0.97 |
| MB_5_Mean_Incomplete | 0.79 | 0.94 | 0.97 |
| MB_5_Median_Incomplete | 0.58 | 0.89 | 0.96 |

**Figure 2** Comparison of classifiers constructed from datasets: (a) PW_Incomplete; (b) MB_3_Mean_Incomplete; (c) MB_4_Mean_Incomplete and (d) MB_5_Mean Incomplete (see online version for colours)




## *4.2 Classification performance: MB approach*

In this section we use the MB approach to evaluate the predictive performance of the classifiers NB, MLP and KNN when applied to individual input features and integrated diverse features. The performances of the classifiers are contrasted when built using datasets consisting of cases with complete and incomplete feature descriptions.

#### 4.2.1 Classification performance: Single-source features in the MB approach

The three classifiers: MLP, SNB and KNN were applied to individual features from the MB_3_Mean dataset. Significant differences between the classifiers were observed using ANOVA ($F = 7.75$, $p = 0.002$), when the classifiers were built using FUNCAT. When the three classifiers were applied to individual features from the MB_3_Median dataset no significant difference was observed using ANOVA ($F = 2.2$, $p = 0.121$) between the classifiers when applied to the COE feature. High AUC values were obtained by all classifiers when constructed from the FUNCAT and GOSEM datasets (Table 1). Significant differences were observed between the three classifiers when built using the same individual features by from the MB_4_Mean and MB_4_Median datasets. However, no significant differences were observed when the classifiers were built using FUNCAT ($F = 0.99$, $p = 0.906$) by ANOVA. The classifiers MLP and NB produced marginally higher AUC values when built using individual features from both MB_4_Mean and MB_4_Median datasets compared to KNN.

#### 4.2.2 Classification performance: Integrated features in the MB approach

High AUC values were obtained (0.97–0.99) by all classifiers when applied to the MB_3_Mean_Full dataset (Table 2). MLP and NB produce AUC values marginally higher than KNN. By performing the ANOVA we discovered a significant difference ($F = 6.14$, $p < 0.05$) between MLP, SNB and KNN. From Table 2 it can be observed that all classifiers obtained an AUC value of 0.97 when constructed from the MB_3_Median_Full dataset. Significant differences ($F = 77.3$, $p < 0.05$) were obtained using AVONA between classifiers when built using the MB_3_Mean and MB_3_Median incomplete datasets. The classifier MLP obtains high AUC values (0.99), when constructed using the MB_4_Mean_Full and the MB_4_Median_Full dataset (Table 2).

Using the AUC values in Table 3, it is observed that the classifiers vary in AUC values obtained (0.69–0.97) when built using the MB_4_Median_Incomplete dataset. Significant differences ($F = 96.6$, $p < 0.05$) were obtained between the three classifiers when constructed using both the MB_4_Mean and MB_4_Median incomplete datasets using ANOVA. The Panels (b)–(d) in Figure 2 graphically show the classifiers SNB and MLP obtain higher AUC values compared to KNN. The classifier NB obtains a higher AUC value when constructed using all MB_5 datasets compared to MLP and KNN. Significant differences ($F = 20.4$, $p < 0.05$) are observed between the classifiers when built using these datasets using ANOVA.

#### 4.2.3 Classification performance: Contrast between full and incomplete in the MB approach

For all MB approaches (MB_3, MB_4, MB_5), classifiers built with datasets containing cases with full feature descriptions obtain higher AUC values compared to classifiers constructed using datasets containing incomplete feature descriptions.

#### 4.2.4 Classification performance: Contrast between feature encoding techniques

Using the MB approach and the MB_3_Full datasets, it was observed that the classifier MLP obtains a marginally higher AUC value when built using the mean feature encoding

technique compared to the median encoding technique. A significant difference ($t = 2.88$, $p = 0.018$) was observed when the MLP was built using the MB_3_Mean_Full and MB_3_Median_Full dataset using the *t*-test.

Using the MB approach and the MB_3_Incomplete datasets, it was observed that the KNN classifier obtained a higher AUC value when built using the MB_3_Median_ Incomplete dataset compared to the MB_3_Mean_Incomplete dataset. Using the *t*-test a significant difference ($t = 5.86$, $p < 0.05$) was observed between the KNN classifier built using the MB_3_Median_Incomplete and MB_3_Mean_Incomplete dataset.

## 5 Conclusion and future work

In this paper we have addressed the problem of network interaction prediction for both PW and MB approaches. PPI play an important role in biological systems. Initial research (Myers et al., 2005) suggests that MB interaction predictions are an important area in predicting PPI.

For both the PW and MB approach, classifiers built using FUNCAT and GOSEM obtain high AUC values. This is due to the high quality of annotation data in both the FUNCAT and GOSEM. These data are independent of the GS and have been gathered manually from small scale traditional experiments, which are thought to be more accurate than high-throughput data. However, these features are limited in coverage of the interactome. For example, for all possible 21,658,071 protein pairs in *S. cerevisiae* (6582 ORFs from MIPS) (Lu et al., 2005), the FUNCAT covers 6,161805 and MES covers 17,775705 proteins pairs. When these two features are integrated, coverage is increased to 18,166007 protein pairs. Therefore, by integrating diverse features, both predictive accuracy (AUC) and coverage could be improved.

Individually, the FUNCAT and GOSEM obtained the highest AUC values obtained. However, the other five features are important as they too obtained high AUC values. For example when the classifier KNN was built using the EXP feature from the MB_3_Mean dataset, an AUC value of 0.75 was obtained. EXP also increases the coverage of protein pairs in *S. cerevisiae* from 9,040603 to 19,509381 protein pairs when EXP is integrated with FUNCAT and GOSEM.

Our research has demonstrated that supervised statistical and machine learning techniques can be successfully applied to the problem areas of PW and MB interaction prediction. Using the PW approach, the best classifiers to inferring PPI are MLP and NB. For the MB approach, the classifier NB consistently produced high AUC values and obtained higher AUC values compared to KNN and MLP when constructed using dataset with incomplete descriptions of features.

The construction of GS in this research provides a useful facility to researchers when applying statistical and machine learning techniques to infer PW and MB interaction networks.

For both PW and MB approaches, the introduction of missing values into the datasets has caused a decrease in prediction performance in terms of AUC values obtained by all classifiers.

Feature encoding (using the mean and median techniques) has an impact on AUC values obtained by classifiers. For example, NB and MLP classifiers constructed using the MB_3 datasets obtained higher AUC values when the mean feature encoding technique was implemented compared to the median encoding technique.

To improve the predictive quality and biological relevance of integrative prediction models, we aim to expand and improve the selection of input datasets, construction of GS and combination of predictive models. Comparative assessments and alternative integrative prediction models (e.g., SVM classifiers) will be extended to *S. cerevisiae* and other organisms, such as *D. melanogaster* and *H. sapiens*.

## References

Azuaje, F., Wang, H. and Bodenreider, O. (2005) 'Ontology-driven similarity approaches to supporting gene functional assessment', *Proceedings of the ISMB'2005 SIG Meeting on Bio-ontologies*, Detroit, USA, pp.9–10

Barutcuoglu, Z., Schapire, R. and Troyanskaya, O. (2006) 'Hierarchical multi-label prediction of gene function', *Bioinformatics*, pp.830–836.

Browne, F., Wang, H., Zheng, H. and Azuaje, H. (2006) 'An assessment of machine and statistical learning approaches to inferring networks of protein-protein interactions', *JIB*, Vol. 2, p.41.

Cho, R., Campbell, M., Winzeler, E., Steinmetz, L. and Conway, A. (1998) 'A genome-wide transcriptional analysis of the mitotic cell cycle', *Molecular Cell*, pp.65–73.

Collins, S.R., Kemmeren, P., Zhao, X.C., Greenblatt, J.F., Spencer, F., Holstege, F.C., Weissman, J.S. and Krogan, N.J. (2007) 'Towards a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae', *Molecular and Cellular Proteomics*, Vol. 6 pp.439–450.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S. and Dümpelfeld, B. (2006) 'Proteome survey reveals modularity of the yeast cell machinery', *Nature*, pp.631–636.

Greenbaum, D., Jansen, R. and Gerstein, M. (2002) 'Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts', *Bioinformatics*, pp.586–596.

Jansen, R. and Gerstein, M. (2004) 'Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction', *Current Opinion in Microbiology*, Vol. 5 pp.535–545.

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*, pp.449–453.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrin-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A. and Greenblatt, J.F. (2006) 'Global landscape of protein complexes in the yeast Saccharomyces cerevisiae', *Nature*, pp.637–643.

Lu, L.J., Xia, Y., Paccanaro, A., Yu, H. and Gerstein, M. (2005) 'Assessing the limits of genomic data integration for predicting protein networks', *Genome Research*, Vol. 15, pp.945–953.

Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. and Weil, B. (2002) 'MIPS: a database for genomes and protein sequences', *Nucleic Acids Research*, Vol. 30, pp.31–34.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. and Euler, T. (2006) 'YALE: rapid prototyping for complex data mining tasks', *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, PA, USA, pp.935–940.

Myers, C.L., Robson, D., Wible, A., Hibbs, M.A., Chiriac, C., Theesfeld, C.L., Dolinski, K. and Troyanskaya, O.G. (2005) 'Discovery of biological networks from diverse functional genomic data', *Genome Biology.*, Vol. 6, p.R114.

Qi, Y., Bar-Joseph, Z. and Klein-Seetharaman, J. (2006) 'Evaluation of different biological data and computational classification methods for use in protein interaction prediction', *Proteins: Structure, Function, and Bioinformatics*, pp.490–500.

SPSS Inc. (2005) *SPSS Base 10.0 for Windows User's Guide*, SPSS Inc., Chicago, IL.

Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B. and Botstein, D. (2003) 'A Bayesian framework for combining heterogeneous data sources for gene function prediction', *PNAS*, Vol. 14, pp.8348–8353.

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*, pp.623–627.

Xia, Y., Lu, L.J. and Gerstein, M. (2006) 'Integrated prediction of the helical membrane protein interactome in Yeast', *Journal of Molecular Biology*, Vol. 357, pp.339–349.

Yu, H., Greenbaum, D., Lu, H., Zhu, X. and Gerstein, M. (2004) 'Genomic analysis of essentiality within protein networks', *Science*, Vol. 20, No. 6, pp.227–231.