An assessment of machine and statistical learning approaches to inferring networks of protein-protein interactions

Fiona Browne, Haiying Wang, Huiru Zheng, Francisco Azuaje*

School of Computing and Mathematics, University of Ulster at Jordanstown, Northern Ireland, UK

Abstract

Protein-protein interactions (PPI) play a key role in many biological systems. Over the past few years, an explosion in availability of functional biological data obtained from high-throughput technologies to infer PPI has been observed. However, results obtained from such experiments show high rates of false positives and false negatives predictions as well as systematic predictive bias. Recent research has revealed that several machine and statistical learning methods applied to integrate relatively weak, diverse sources of large-scale functional data may provide improved predictive accuracy and coverage of PPI. In this paper we describe the effects of applying different computational, integrative methods to predict PPI in *Saccharomyces cerevisiae*. We investigated the predictive ability of combining different sets of relatively strong and weak predictive datasets. We analysed several genomic datasets ranging from mRNA co-expression to marginal essentiality. Moreover, we expanded an existing multi-source dataset from *S. cerevisiae* by constructing a new set of putative interactions extracted from Gene Ontology (GO)-driven annotations in the Saccharomyces Genome Database. Different classification techniques: Simple Naive Bayesian (SNB), Multilayer Perceptron (MLP) and K-Nearest Neighbors (KNN) were evaluated. Relatively simple classification methods (i.e. less computing intensive and mathematically complex), such as SNB, have been proven to be proficient at predicting PPI. SNB produced the “highest” predictive quality 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. This assessment also demonstrates the strong predictive power of GO-driven models, which offered predictive performance above 0.90 using the different machine learning and statistical techniques. As the predictive power of single-source datasets became weaker MLP and SNB performed better than KNN. Moreover, predictive performance saturation may be reached independently of the classification models applied, which may be explained by predictive bias and incompleteness of existing “Gold Standards”. More comprehensive and accurate PPI maps will be produced for *S. cerevisiae* and beyond with the emergence of large-scale datasets of better predictive quality and the integration of intelligent classification methods.

***Keywords*:** protein-protein interactions, machine and statistical learning, ‘omic’ datasets, functional genomics, computational systems biology.

1 Introduction

Most proteins perform their functions by interacting with other proteins. Information about the networks of interactions within a cell can greatly increase our understanding of protein

[*] Corresponding author: fj.azuaje@ulster.ac.uk

function and cellular processes. Notwithstanding the advances of post-genome biology and bioinformatics, there is a need to improve our knowledge of protein-protein interactions (PPI). This has inspired efforts to map interactions on a proteome-wide scale. 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 [1-6].

The completion of genome sequencing projects followed by the development of high-throughput technologies has accelerated the pace of discovery of PPI. This has resulted in an enormous accumulation of biological data. Common experimental methods include the yeast-two hybrid screen, tandem affinity purification, mass spectrometry and protein chips [1-7]. However, the data obtained by these methods are often noisy, incomplete and contradictory [7] (i.e. weak predictive data sources) 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. Methods such as the yeast-two hybrid system exhibit high false positive and false negative interaction rates.

Traditional data integration methods for PPI prediction produce more accurate results compared to single source high-throughput methods. 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 [12,15,16].

Integration of diverse genomic datasets has been proven to improve the prediction of PPI [8-9]. By integrating different datasets the coverage of the interactome increases with possible overlapping of datasets. When the same PPI prediction is agreed on by two or more different datasets the predictions may be considered to be more reliable. Datasets that do not directly measure PPI, such as sequence, structural and diverse functional genomic information, can also be used to predict PPI. For example, the application of gene co-expression to infer PPI is based on the hypothesis that proteins found in the same complex often interact, and proteins 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.

In this study we apply computational techniques to integrate diverse sources of information to infer PPI. When predicting PPI using computational methods, a number of factors have been taken into account. These include choosing a Gold Standard (GSTD) and different predictive (classification) techniques. A GSTD is a dataset consisting of a number of known interacting and non-interacting protein pairs, which are used to train classifiers and estimate their predictive ability. Careful consideration is required when choosing datasets and the GSTD as they will affect the validity and reliability of the predictions. Selecting a GSTD can be problematic. For example, what does a GSTD specifically measure? In this study the GSTD include “positive” cases representing pairs of proteins found in the same complex. Another problem is to select a GSTD that has an adequate coverage of the interactome. The specific task of selecting a negative GSTD (i.e. non-interacting protein pairs) also represents a significant challenge. In this study, non-interacting proteins are based on the assumption that protein pairs assigned to different cellular compartments are unlikely to become interacting pairs. The difficulty in defining a negative class (i.e. non-interacting proteins) for a GSTD is one of the root causes for the poor or, in some cases, overestimated performance of machine learning algorithms in the prediction of PPI [10-11]. There are no universal GSTD available within the field of functional genomics and systems biology. The quality of the statistical and machine learning methods will depend on the relevance and validity of the GSTD to the prediction problem under study.

The selection of classification techniques is also a critical task. Moreover, some classifiers that perform well in other problem domains may not perform as well within the realm of PPI. This is due to the type of data a classifier can handle. Classifiers exhibit systematic bias (i.e. a method produces solutions that highly favour a limited number of specific situations or circumstances) or logical assumptions (for example, independence between datasets which can lead to systematic errors). Therefore, it is essential to rigorously assess available classification models for correctly inferring PPI.

Numerous statistical and machine learning methods have been used to integrate diverse sources of data for PPI prediction. Jansen *et al.* [8] and Troyanskaya *et al.* [9], for instance, applied a Bayesian networks (BN) approach to predicting PPI by integrating genomic data. These studies produced accurate PPI networks providing a comprehensive view of the *S. cerevisiae* interactome. Barutcuoglu *et al.* [15] have recently developed a probabilistic, query-based system to discover pathway-specific networks by integrating diverse genome-wide data. This system is based on BN and was validated by accurately recovering networks for 31 known biological processes in *S. cerevisiae*. Research by Jansen *et al.* [8] was consequently extended by Lu *et al.* [12]. Lu *et al.* [12] focused on assessing the predictive limits of genomic data integration. Simple Naïve Bayesian (SNB) was used to integrate sixteen diverse datasets. As with a previous study [8], relatively high predictive accuracies were obtained. However, the addition of relatively weaker datasets only marginally improved the predictive power of the models. The Random Forest (RF) machine learning method has been applied in studies by Chen *et al.* [13] and Qi *et al.* [24]. The RF classifier used in [13] this study predicted PPI with an average sensitivity of around 80% and specificity below 65% [13]. Support Vector Machines (SVM) [14] have also become a powerful approach for PPI prediction. Using a number of classifiers (RF, RF integrated with K-Nearest neighbor (KNN), Naive Bays, Decision Tree, Logistic Regression and SVM), Qi *et al.* [16] showed the effect of dataset selection and encoding on the PPI predictive performance. Despite the relative success of these methods, predictive variability and potential systematic bias indicate that there is still a need for improvement in terms of predictive quality and computational efficiency. Moreover, with regard to predictive performance, it has been suggested that relatively simple classification methods (e.g. KNN) may achieve high predictive performance in comparison to more sophisticated approaches, such as SVM.

It is difficult to compare and contrast the results of these particular studies as each investigation uses different GSTD and datasets. Recent empirical assessments include research by Lu *et al.* [12], for instance, who focused on integrating diverse sources of information based on the application of SNB. However, it is crucial to perform more rigorous and comprehensive empirical assessments to determine the differences between other computational integrative prediction methods in well-studied organisms such as *S. cerevisiae.* This is an essential step toward the design, adaptation and integration of prediction strategies in relatively more complex organisms.

In this research, three classification models from the field of machine and statistical learning have been assessed. These classifiers range in terms of computational complexity and learning approaches. Some of these classifiers have been previously applied to predict PPI, whereas other classifiers have never been individually applied. Thus, this study will provide a comprehensive assessment of representative prediction models. Neural Networks were implemented by Lee *et al.* [25] to predict PPI. In their study, using 10-fold cross-validation, a classifier accuracy of 96%, a sensitivity of 98%, and specificity of 96% were on average achieved [25]. To the best of our knowledge the KNN has not been assessed in the prediction of PPI. We investigated the predictive ability of combining different sets of relatively strong and weak datasets. Unlike the assessment reported by Qui *et al.* [16], our study analyses and integrates eight different datasets (including a new dataset constructed by the authors) using different machine and statistical learning techniques. Our work will focus on PPI prediction in *S. cervisiae* and integrates a wide range of datasets, consisting of physical, genetic and biological content of genes. Current research on integrating relatively weaker, high-throughput data sources has shown that these datasets only marginally increase the predictive power when different datasets are integrated together [12]. This paper also shows how a relatively strong, functional annotation-driven dataset affects the predictive power of the models when integrated with other large-scale datasets. Such a new set of putative interactions inferred from functional similarity information extracted from a database annotated to the Gene Ontology (GO) was constructed to implement predictive integration with other datasets using different prediction models. The GO-driven PPI dataset, which from now on is referred to as GOSEM, was also compared with another GO-driven PPI prediction method proposed in [12]. These GO-driven PPI prediction methods differ in terms of the way they estimate similarity between GO terms and between genes to predict PPI associations.

Throughout this paper datasets may be referred to as being *strong* and *weak*. We define a strong dataset as a predictive resource that contains a relatively small number of false positive predictions and false negative putative interacting pairs in relation to the GSTD. This type of datasets covers a larger proportion of the interactome compared to a "weaker" dataset. A weak dataset contains more false positive and false negative predictions and is more limited in coverage compared to a stronger dataset in relation to the GSTD. An analysis of how different classifiers differ in predicting PPI when integrating diverse data sources was also implemented. This paper discusses the results and reasons for the differences and similarities between the techniques. We compare the results of these classifiers to determine if relatively simpler classifiers may outperform more complex classifiers. We determine and discuss the most effective and reliable prediction models and assess different optimal combinations of datasets (features). 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 PPI prediction approaches and outline current and future work.

2 Data sources and gold standard

Seven different functional datasets obtained from Lu *et al.* [12] along with our GO-driven PPI dataset were analysed and integrated to predict PPI (see description below). Each of the seven datasets has been used in studies previously performed by Lu *et al.* [12] and Jansen *et al.* [8]. In the study performed by Lu *et al.* [12] sixteen different datasets were analysed. Based on their study we chose the "top" seven datasets. These datasets were defined as "top" as they cover at least half a million (~20%) Opening Read Frame (ORF) pairs in the Gold Standard (derived from the MIPS complex catalogue database). Moreover, these datasets showed the highest overall predictive performances in Lu *et al.* [12]. Our study differs from these contributions [8, 12] by assessing different classification models (KNN and MLP), and by introducing a new set of putative PPI. 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. Table 1 provides a brief summary of functional datasets that were assessed for PPI integrative prediction.

2.1 MRNA co-expression (COE)

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 [17]. 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 [17]. Pearson's correlation values were calculated for each gene pair. The results range from 0 to 1.

2.2 MIPS functional catalogue (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) [18]. The FunCat is an annotation scheme that contains data on the functional description of proteins from prokaryotes, unicellular eukaryotes, plants and animals [18]. The FunCat is separate from the MIPS complex catalogue, which represents the GSTD in this study. Mews *et al.* [18] provide additional information on the FunCat annotation database. The calculation of similarity between gene pairs for the FunCat dataset and the traditional GO-driven frequency-based similarity dataset are the same. Section 2.3 outlines this similarity estimation procedure. The results within this dataset range from 0 to 7.

2.3 Traditional GO-driven frequency-based similarity (GOFREQ)

Using the same hypothesis as outlined in FunCat, information on the involvement of pairs of proteins in specific biological process was extracted from a GO-driven annotation database: Saccharomyces Genome Database (SGD). The GO aims to deliver a shared, structured and controlled vocabulary that can be applied to any organism. GO consists of three independent hierarchies: molecular function, biological process, cellular component. Within these hierarchies, terms are interrelated forming a directed acyclic graph. Ashburner *et al.* [19] provide a detailed description of the GO construction and applications. The values generated by this dataset range from 0 to 7.

The GOFREQ and FunCat datasets were both constructed by calculating the similarity values between gene products annotated in the GO biological process hierarchy and FunCat respectively. Outlined below are the steps involved in quantifying the functional similarity between two gene products using similarity information extracted from FunCat and SGD.

2.3.1 Traditional method for calculating similarity values from SGD and FunCat

Given two proteins that share a specific set of lowest common ancestor nodes in the classification structure, one can count the total number of protein pairs, *n,* that share the same set of annotation terms. For all the protein pairs in *S. cerevisiae* (~18 million), it was counted how many of these pairs share the exact functional terms. This resulted in a count ranging between 0 and 18 million. A smaller count reflects a more specific functional description of the two proteins, which suggests a higher functional similarity and more chance of belonging in the same cellular complex. A larger count indicates a less specific functional relationship between the proteins; therefore, there is less chance that the proteins belong to the same cellular complex. These datasets were divided into four frequency-based similarity bins 1-9, 10-99, 100-1000, 1000-10000 to 10000-infinity [12].

2.4 GO-driven semantic similarity (GOSEM)

The Lin's semantic similarity technique [20] was also 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 [19,20]. The Lin's formula for estimating between-term similarity is:

$$sim\left(c_i, c_j\right) = \frac{2 \times \max\limits_{c \in S\left(c_i, c_j\right)} \left[\log\left(\mathrm{P}\left(c_j\right)\right)\right]}{\log\left(\mathrm{P}\left(c_i\right)\right) + \log\left(\mathrm{P}\left(c_j\right)\right)} \tag{1}$$

Where S represents the set of parental terms shared by terms c_i and c_j; $\max$ represents the maximum operator, $\mathrm{P}(c)$ represents the probability of finding c or any of its parents in the SGD [20].

Between-gene similarity was calculated by aggregating the similarity values obtained between the annotation terms of the genes. Given a pair of gene products, g_k and g_p, sets of annotations A_k and A_p comprising of m and n terms, the between-gene similarity, $SIM\left(g_k, g_p\right)$, may be defined as the average inter-set similarity between terms from A_i to A_j [20]:

$$SIM\left(g_k, g_p\right) = \frac{1}{m \times n} \times \sum_{c_i \in A_k, c_j \in A_p} sim\left(c_i, c_j\right) \tag{2}$$

Where $sim\left(c_i, c_j\right)$ can be calculated using equation (1) [20].

2.5 Co-essentiality (ESS)

This dataset is derived from the MIPS complex catalogue and also from transposon and gene deletion experiments [18]. 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 either essential or non-essential but not both. This is because a deletion mutant of either one protein should produce the same phenotype, and their mutual deletion would impair the function of the same complex. 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 non-essential 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. Mews *et al.* [18] provide more detailed information about this dataset.

2.6 Marginal essentiality (MES)

This is a quantitative measure of the importance of a non-essential gene to a cell. It is based on the 'marginal benefit' hypothesis that many non-essential 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 [21]. Marginal essentiality relates to many of the topological characteristics of PPI networks. In particular, proteins with a greater degree of marginal essentiality tend to be network hubs (i.e. with many interactions) and tend to have a shorter characteristic path length to their neighbors [21]. Protein pairs are defined as interacting if two proteins have a higher combined marginal essentiality. 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 [21]. Non-essential 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 [21]. Yu *et al.* [21] provide information on the genomic analysis of essentiality within PPI networks. The values generated by this dataset range from -0.9 to – 27.

2.7 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 [22]. 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. Greenbaum *et al.* [22] detail the relationship between of mRNA expression and protein abundance data. The values generated by this dataset range from 0 to 20

2.8 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 [22]. EXP is an approximation of absolute expression levels of mRNA within a cell. The values generated by this dataset range from 0 to 10.

2.9 Gold standard (GSTD)

To validate PPI computational predictions, it is essential to have a reference dataset that contains known positive (proteins that are both in the same complex) and negative (non-interacting) protein pair cases. Such a knowledge reference is known as a Gold Standard and is used to label the protein pairs in the prediction model construction and evaluation. The GSTD used in this study is constructed under the assumption that if two proteins are known to be in the same complex then they can be defined as interacting pairs. There is no direct information on proteins that do not interact. However, one may assume that pairs of proteins belonging to different cellular compartments are less likely to interact than those belonging to the same compartment. In this study non-interacting protein pairs were derived from pairing proteins from different subcellular complexes as described in [8]. Both the positive (i.e. interacting) and negative (i.e. non-interacting) sets were obtained from the MIPS complex catalogue. 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 [12]. There are 8250 protein pairs in the positive GSTD and 2,708,622 in the negative GSTD. Only protein pairs that were contained in a single complex were selected (minimum size of complex: 5 proteins). We assume that protein pairs that are found within these sub-classes interact. This GSTD was also used in [8, 12].

Table 1 - Summary of functional datasets assessed for PPI integrative prediction

Dataset	Source	Assumption	No, Of Protein Pairs	Reference
COE	Microarrays	Proteins in the same complex are often co-expressed	**Num / Total:** 2,682,887/18,773,128 **Ovlp + / -:** 7,614/2,675,273	[17]
FunCat	MIP FunCat	Proteins acting in the same biological process are more likely to interact than two proteins involved in different processes	**Num / Total:** 1,321,629/6,161,805 Ovlp + / -: 8,051/1,313,579	[18]
GOSEM	GO-driven annotations (SGD)	Proteins acting in the same biological process are likely to interact together.	**Num / Total:** 655,417/2,878,800 Ovlp + / -: 7,520/647,060	[19]
GOFREQ	GO-driven annotations (SGD)	Proteins acting in the same biological process are likely to interact together.	**Num / Total:** 655,417/2,878,800 Ovlp + / -: 7,520/647,060	[19]
ESS	MIPS complex database, transposon and gene deletion experiments	Protein can be experimentally characterised as either essential or non-essential. An indicator that the proteins are both members of the same complex.	**Num / Total:** 649,210/8,130,528 Ovlp + / -: 2,150/647,060	[18]
MES	Large-scale phenotypic experiments	With a higher combined marginal essentiality, proteins are more likely to interact	**Num / Total:** 2,595,937/17,775,703 Ovlp + / -: 7,738/2,588,199	[21]
EXP	Microarrays, Affymetrix chips	Two proteins interacting should be present in stoichiometrically similar amounts	**Num / Total:** 2,703,788/19,303,791 Ovlp + / -: 7,786/2,696,002	[22]
APA	Gel electrophoresis and mass spectrometry	Two proteins interacting should be present in stoichiometrically similar amounts.	**Num / Total:** 1,519,747/7474,911 Ovlp + / -: 5,192/1,514,555	[22]
GSTDS	MIPS complex database	Protein complex membership.	GSTD+ 8,250 GSTD- 2,708,622	[8]

Num / Total – Overlap of Gene pairs from dataset with GSTD / Total Number of Gene Pairs in dataset.
Ovlp +/- Number of overlaps with GSTD+/GSTD-

3 Methods

Three different machine and statistical learning methods were chosen to integrate the eight diverse datasets. These classifiers range in complexity and type. We built several SNB, KNN and MLP prediction models. The aim is to show how these methods differ in predictive accuracy when integrating the different datasets. Below we briefly outline the different machine and statistical learning methods. Each classifier was obtained from the WEKA toolbox [23] and a 10-fold cross-validation was performed to estimate the predictive performance (i.e. specificity and sensitivity). 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 GSTD. Such estimations were summarised and validated with Receiver Operating Characteristic curves (Section 3.4).

3.1 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 PPI by Qi *et al.* [16, 24]. Although in these investigations KNN was combined with RF. However, this classifier has not been evaluated individually to predict PPI. In this investigation, KNN classified gene pairs 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. An empirical analysis was carried out to determine the optimal number of nearest neighbours. *K* was set to 1, 3, 5 and 10. When *K* was set at 3, this produced the best predictive quality results.

3.2 Multilayer perceptron (MLP)

Previous research by Lee *et al.* [25] applied MLP to predict PPI using three datasets and a GSTD derived from the MIPS complex database. This research produced high predictive effectiveness (96%) results (measured as the AUC, area under the *receiver operating characteristic curve*). In terms of mathematical complexity we regard, MLP as the most complex of the three classifiers. MLP is a non-linear classification approach and that is trained using the back propagation algorithm. In this investigation the network consists of three layers, an input layer –where the eight datasets are placed; a hidden layer -to which all input nodes are connected and an 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. Further information on the MLP model can be found in [26].

3.3 SNB

Several papers have reported the application of the Bayes rule for the prediction of PPI [8,9,12]. The classifier, SNB, has previously been used by Lu *et al.* [12] to combine diverse genomic features. Therefore, SNB is being used as a benchmark to compare the other classifiers of varying complexity. SNB offers a simple approach and is based on the Bayes rule of conditional probability. It is regarded “Naïve” as it “naively” assumes independence between features (datasets). However, due to this assumption, the predictive power of SNB may be reduced if a dataset is highly correlated with an existing dataset. SNB is considered simple as it uses the normal distribution to model numeric attributes by calculating the mean standard deviation for each class. This technique can handle diverse heterogeneous sources of data. Although this technique is relatively simple in terms of mathematical complexity, relatively high prediction accuracies have been obtained by several studies [27]. Detailed information on SNB can be found in [12, 28]

3.4 ROC (receiver operating characteristic) curves

ROC analysis investigates the accuracy of a model’s ability to separate the positive from negative cases. For this study, ROC curves were chosen to evaluate the predictive models as they capture in a single graph the trade off between sensitivity and specificity over its entire range of the dataset. The predictive quality of a classifier is assessed by measuring the sensitivity and specificity. *TP*, *TN*, *FP* and *FN* are the counts of true positives, true negatives, false positives and false negatives obtained from the cross-validation analysis. The formula used to calculate sensitivity and specificity are detailed below:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

A predictive dataset will produce a ROC curve that rises steeply to the left hand side of the graph and has a large area under the curve. The threshold values of the ROC curves displayed within this study are dependant upon the dataset values, for example similarity values or co-expression values. In this paper the machine and statistical learning methods are ranked according to the AUC obtained. The AUC values are estimated from the 10-fold cross-validation procedure. A perfect classifier will have an AUC value of 1.0. A prediction model based on random assignments of pairs of proteins to classes would give an AUC equal to 0.5.

3.5 Analysis of statistical significance

In this study, analysis of variance (ANOVA) and the paired samples Student *t*-test (two-tailed) analysis were applied to determine if significant differences existed between the predictive models in terms of accuracy. ANOVA tells us whether the factors (e.g. classification techniques) significantly contribute to the variations observed in the prediction outcomes. The accuracy values obtained from the cross-validation procedures (10 values) were used in the ANOVA and *t*-test. ANOVA and the *t*-test were performed using the statistical package SPSS version 11.0 [29]. A significant difference is observed when the *p* value obtained is less than 0.05.

4 Results and discussion

In this section we present and discuss the results obtained from the assessment of the three representative predictive models using the individual *S. cervisiae* datasets. Firstly, we contrasted the results obtained from the three individual computational techniques to integrate diverse sources of information to infer PPI. Secondly, we focus on showing how a relatively strong, functional annotation-driven dataset, GOSEM, affects the predictive power when integrated with other large-scale experimental datasets. The classifiers were built using the measurements obtained from each dataset as the inputs to the models. The "true" categorisations for each input (i.e. protein pair) were obtained from the GSTD.

4.1 Performance of machine and statistical learning methods: single source models

The ROC curves in Figure 1 depict the predictive power of individual datasets using each machine and statistical learning technique. Table 2 exhibits the AUC values obtained by the predictive models for the individual datasets. By performing ANOVA analysis, significant differences were observed between the classification models in terms of their predictive accuracy based on different individual datasets ($p < 0.001$).

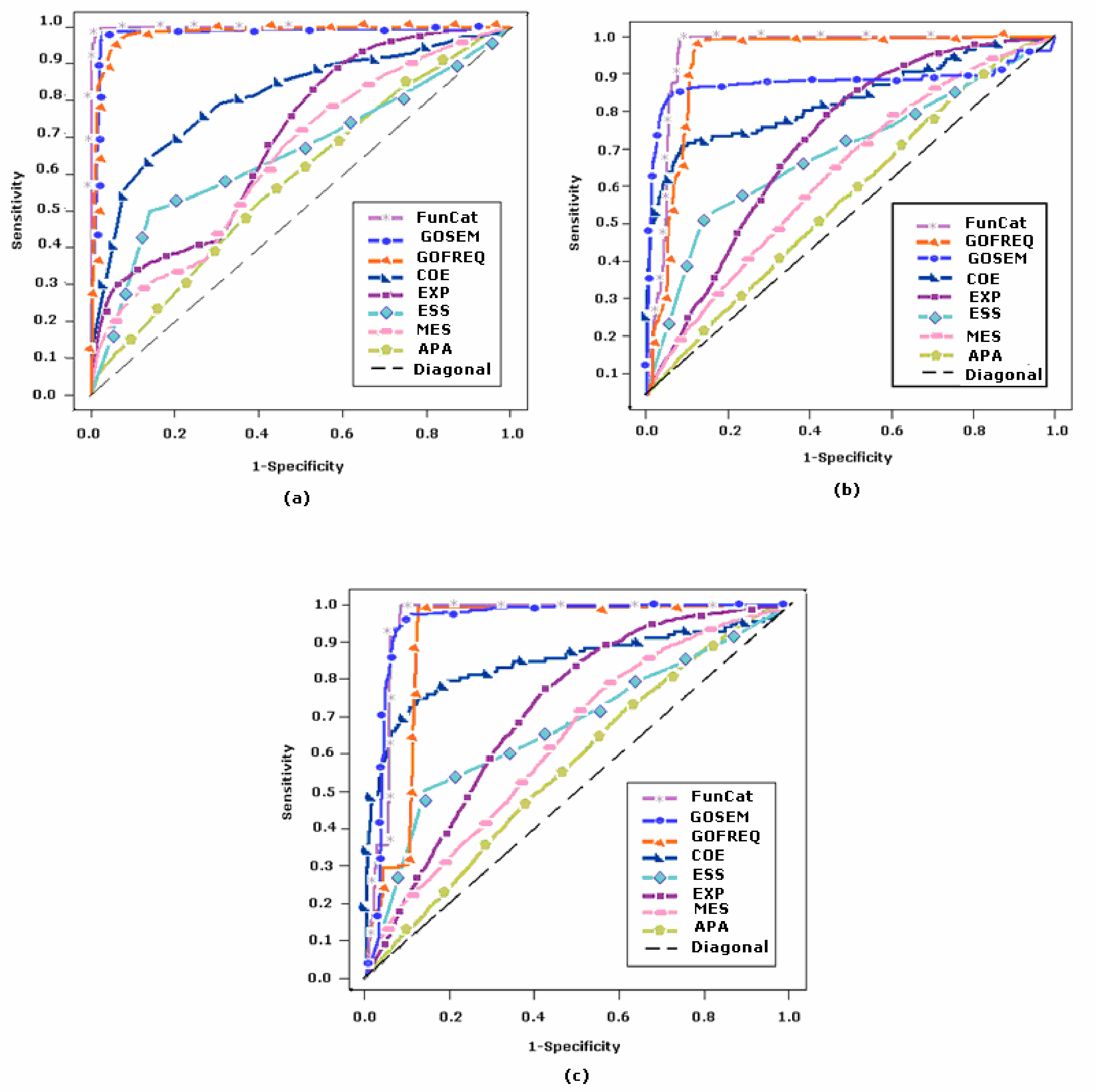

Figure 1. The panels a-c each represent the predictive power of individual features (datasets) using each predictive model. The results are displayed in a ROC curve where we plot Sensitivity against (1-Specificity). The predictive models are displayed as follows (a) KNN. (b) MLP. (c) SNB.

Based on the AUC values from Table 2, it is observed that each predictive model ranks the FunCat dataset as the “strongest” dataset and the APA dataset as a relatively “weakest” dataset. The FunCat dataset may be defined as “strong” as it overlaps with 3000 more protein pairs in the positive GSTD compared to the APA dataset. The GSTD contains information on protein complex membership and the FunCat dataset encodes information on the similarity between gene pairs. Protein abundance is less related to the specific task of protein complex membership. These factors could help explain why APA was found to be a relatively “weak” performing dataset.

Table 2. The predictive AUC values obtained by each machine and statistical learning method using individual datasets.

Datasets	SNB	KNN	MLP
APA	0.56	0.54	0.55
COE	0.85	0.79	0.82
ESS	0.67	0.67	0.68
EXP	0.71	0.74	0.70
GOFREQ	0.91	0.98	0.92
GOSEM	0.95	0.97	0.87
MES	0.63	0.59	0.61
FunCat	0.96	0.99	0.95

From these individual dataset results we can conclude that KNN obtains marginally higher AUC values when the "strongest" datasets are used. For example, KNN achieves an AUC value of 0.99 using only the FunCat dataset. The predictive models SNB and MLP perform consistently throughout with all the datasets individually and obtain marginally higher AUC values when relatively weaker datasets such as MES, ESS and APA are used as inputs to the models.

The next step to investigating the predictive quality of the models was to perform different integrations of the datasets. All the machine and statistical learning techniques produced relatively very high AUC values (between 0.95 and 0.99) when combining all the datasets as inputs to the prediction models. Table 3 summarises the different combination of datasets used in Table 4. Table 4 shows the AUC values for each machine and statistical learning method with different data integration schemes. Using these AUC values the three models can be ranked in order of effectiveness: SNB, MLP followed by KNN. By performing the ANOVA analysis we also found a significant difference between the predictive accuracies of these methods ($p < 0.001$).

Interestingly from Table 4 it is observed that by combining relatively weaker datasets (APA, COE, ESS, EXP, MES) and applying MLP (which we regard as the most complex method within this paper in terms of mathematical complexity) and SNB produced the highest AUC values (0.90 and 0.88 respectively). Significant differences between all the classifiers in terms of their overall accuracies were observed (ANOVA, $p < 0.001$).

From the results in Table 2, the predictive model KNN obtains an AUC value equal to 0.99 using only FunCat dataset. However, from Table 4, a decrease in predicative quality (i.e. an AUC value of 0.95 was obtained) is viewed when all the datasets were integrated using KNN. A significant difference in terms of predictive accuracy was observed between these two integrations ($p = 0.002$).

Table 3. Datasets involved in integration type

Description	Datasets
All datasets	FunCat+GOFREQ+GOSEM+EXP+ESS+COE+APA
Strong datasets I	FunCat +GOSEM+COE+ESS+GOFREQ
Strong datasets II	FunCat +COE+ESS
Strong datasets III	COE+ESS
Weaker datasets + Strong datasets III	MES+EXP+ESS+APA+COE
+ symbolizes the integration of the datasets	

Table 4. AUC values for each machine and statistical learning method with different integrations of the datasets.

Data integration scheme	SNB	KNN	MLP
All datasets	0.99	0.95	0.99
All datasets excluding GOSEM	0.98	0.93	0.96
Strong datasets I	0.99	0.91	0.98
Strongest datasets II + GOFREQ	0.98	0.95	0.97
Strongest datasets II+ GOSEM	0.99	0.95	0.99
Strongest datasets II	0.97	0.97	0.93
Strong datasets III + GOSEM	0.98	0.95	0.98
Weaker datasets + Strong datasets III	0.88	0.8	0.90
Weaker datasets + Strong datasets III + GOFREQ	0.96	0.83	0.96
Weaker datasets + Strong datasets III + GOSEM	0.98	0.96	0.98

4.2 Performance of machine and statistical learning methods: multiple source models

The relatively strong, functional annotation-driven dataset, GOSEM, was integrated with other datasets using different prediction models. Table 4 indicates that the addition of the GOSEM dataset has a positive impact on the AUC values obtained, for different integration schemes and different predictive models. When comparing the integration of all datasets before and after the integration of the GOSEM, a marginal increase in the predictive performance for the models SNB, MLP and KNN was achieved. For each classifier, a *t*-test was performed to determine if there was a significant difference due to the addition of the GOSEM dataset to the integration of all datasets. The incorporation of GOSEM did not significantly improve the predictive accuracy of MLP (*t*-test, $p = 0.262$) and KNN (*t*-test, $p = 0.509$). However, the integration of GOSEM caused a significant improvement in the prediction accuracy of the SNB (*t*-test, $p < 0.001$). Significance differences in terms of accuracies were also detected between the three classification techniques when integrating all the datasets (ANOVA, $p < 0.001$).

The integration of GOSEM improved the prediction performance (estimated as AUC values) with all the datasets (FunCat, GOFREQ, APA, COE, ESS, EXP, MES) under all the machine and statistical learning methods, in comparison to the models excluding GOSEM (Table 4). Interestingly SNB reaches the maximum AUC value (0.99) when relatively strong datasets (FunCat, GOSEM, COE, ESS, GOFREQ) are integrated only. The addition of relatively weaker datasets (APA and MES) do not have a significant impact (*t*-test, $p = 0.136$) on the AUC value when integrated with these relatively stronger datasets. This same trend was observed for MLP (*t*-test, $p = 0.129$) and KNN (*t*-test, $p = 0.104$).

These results confirm that the GOSEM dataset is a relatively strong dataset. Depending on predictive model, it ranks as the second or third “strongest” dataset in terms of AUC values obtained. GOSEM achieves these high AUC values due to the quality of the information it represents. GOSEM encodes relationships between gene pairs on the basis on their involvement in biological processes using a GO-driven annotation database.

In Figure 2 each panel represents the predictive performance response of each classification method for two data integration schemes: a) all datasets, and b) the 5 top strongest datasets (FunCat, GOFREQ, GOSEM, ESS, COE). In relation to SNB, Figure 2 shows results that are consistent with the results obtained by Lu *et al.* [12]. In general an improvement in prediction performances of SNB and MLP were obtained when integrating all the datasets. But in the case of KNN the integration of all datasets did not cause an improvement of the prediction performance compared to the outcomes derived from the integration of the strongest datasets only.

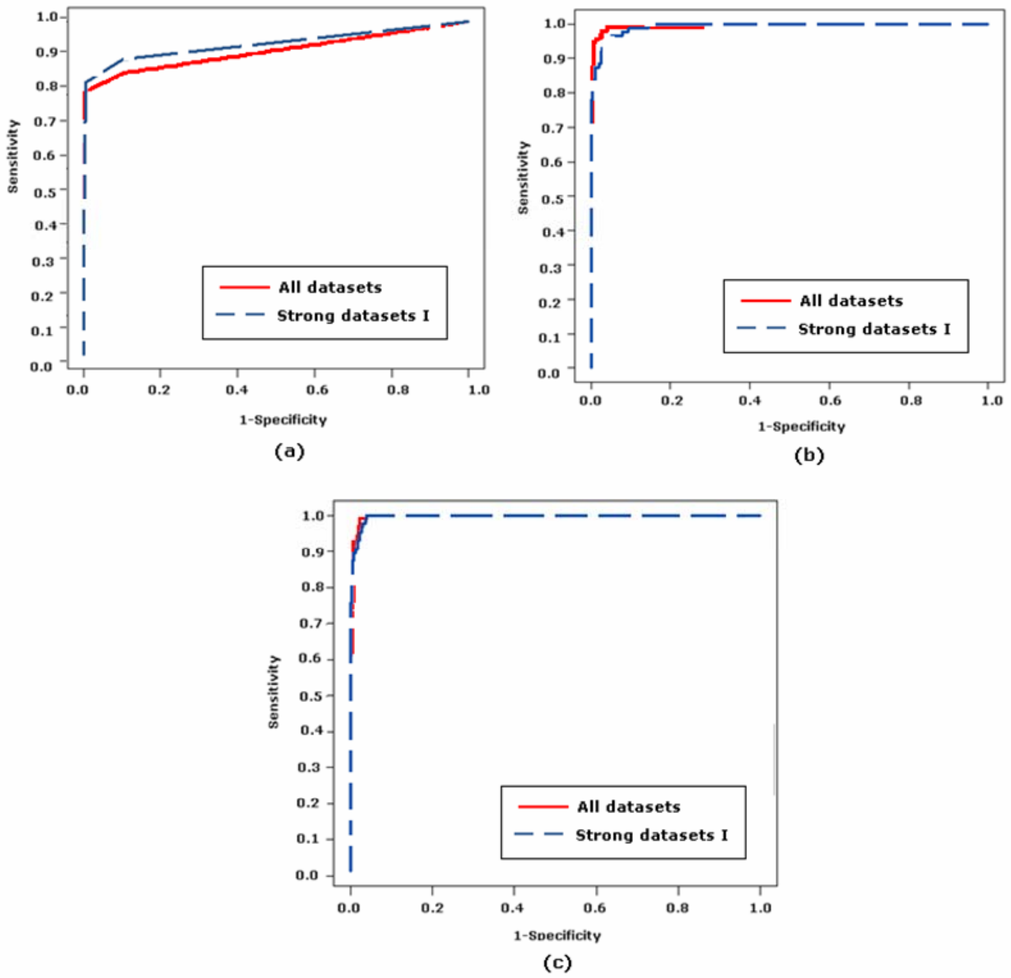

Figure 2. ROC curves obtained from integrating all eight features together, compared with integrating only the strongest 5 datasets. Each panel represents a different machine and statistical learning method (a)KNN, (b) MLP (c) SNB.

5 Conclusions

This paper presented a comprehensive assessment of representative predictive models for inferring PPI. It highlights the diversity of predictive responses depending on the classification technique and combination of inputs features (i.e. datasets). This focused on (1) a comparison of three different predictive models and the assessment of differences in their performance; (2) using seven existing datasets and constructing a new set of putative interactions (GOSEM) extracted from a GO-driven annotation database; and (3) a detailed comparison of different data integration schemes. The classification techniques evaluated are representative approaches to PPI prediction. SNB and MLP were previously evaluated in [12, 25], but without incorporating the GOSEM dataset. To the best of our knowledge KNN has not been rigorously assessed and compared against SNB and MLP.

This investigation has proven that the relatively strong, functional annotation-driven dataset, GOSEM, may support the improvement of the predictive power when integrated with other large-scale datasets. Table 4 suggests that GOSEM may improve PPI classification performance in comparison to GOFREQ, which is a traditional method for inferring PPI from GO-driven databases. This was demonstrated in the case of SNB and MLP, but further investigations are required to assess potential predictive power differences between these two

GO-driven PPI prediction techniques. Both GOSEM and GOFREQ are expected to be strong datasets because they encode functional relationships between gene pairs on the basis on their involvement in biological processes.

Consistent with previous studies carried out by Lu *et al.* [12] we found that the integration of weak features and stronger features may not have significant impact on the predictive performance of all machine and statistical learning models.

All classifiers showed high predictive quality (i.e. AUC values ranging from 0.90 to 0.99) when integrating all datasets. SNB and MLP were the best predictive methods in terms of AUC values. Both classifiers obtained an AUC value of 0.99 when all datasets were integrated. This could be due to ability of SNB to combine highly heterogeneous genomic features. MLP is known to be robust to noisy dataset. This was proven as MLP was the strongest prediction model when only the weak datasets were integrated together (AUC value equal to 0.90). These factors could contribute to the relatively high prediction quality results obtained by MLP.

Previous research by Lee *et al.* [25] also concluded that MLP was a powerful predictor of PPI obtaining a classification accuracy value of 96%. Lee *et al.* [25] used a GSTD based on MIPS complex database and three different datasets containing information on functional similarity between genes (unrelated to GO-driven similarity method applied in this study), co-localisation of gene pairs and topological properties of PPI networks. The results obtained from our investigation using the MLP in general agree with the results reported by Lee *et al.* [25]. However, in some cases (e.g. integration of all datasets) our results may represent an improvement in terms of prediction performance (higher AUC values).

The classifier KNN produced satisfactory results when integrating all datasets (AUC values of 0.95). However, it was observed that KNN was relatively slower and more processor intensive than the other methods when dealing with relatively large datasets, such as COE, which contains over 1 million gene pairs.

This study also indicates that a predictive saturation could be reached by the prediction models available. This means that in this particular investigation the addition of more datasets may not necessarily improve the predicative performance of the machine and statistical learning methods. However, this will strongly depend on the selection of the GSTD. Therefore, other assessments including alternative GSTDs, data sources (e.g. PPI extracted from the literature and other high-throughput experimental source) and model organisms are required.

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 GSTD and combination of predictive models. Comparative assessments and alternative integrative prediction models (using for example, SVM classifiers and probabilistic models) will be extended to *S. cerevisiae* and more complex organisms, such as *Drosophila melanogaster* and *Homo sapiens*. Investigations of how noise and incompleteness of the interaction data could affect the different machine learning approaches will also be carried out as part of future work.

References

[1] P. Uetz L. Giot G. Cagney, T.A. Mansfield, R.S. Judson *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 403(1):623-627. 2000

[2] T. Ito, T. Chiba and M. Yoshida. Exploring the protein interactome using comprehensive two-hybrid projects. NCBI,19(10), 2001

[3] A. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. NCBI, 414(6868):123-124. 2002.

[4] A. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, *et al.* Proteome survey reveals modularity of the yeast cell machinery. Nature, 440(7084): 631-6, 2006

[5] Y. Ho, A. Gruhler, A. Heilbut, G. Bader, L. Moore, *et al.* Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature, 414:180-183. 2002

[6] N. Krogan G. Cagney, H. Yu, G. Zhong, X. Guo, *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature, 440(7084): 637-43, 2006

[7] C. von Mering, R. Krause, B. Snel, M. Cornell, S. Oliver, S. Fields, and P Bork. Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 417(6887):399-403. 2002.

[8] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, and M. Gerstein, M. A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. Science, 302(5644):449-453. 2003

[9] O.G. Troyanskaya, K. Dolinski, A.B. Owen, R.B. Altman, and D Botstein. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). PNAS,100(14):8348-8353.2003

[10] R. Jansen and M. Gerstein. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Current Opinion in Microbiology, 7(5): 535-545. 2004.

[11] A. Ben-Hur and S. Noble. Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinformatics 7(Suppl 1)**:**S1, 2006.

[12] L.J. Lu, Y. Xia,, A. Paccanaro, H. Yu, M. Gerstein. Assessing the Limits of Genomic Data Integration for Predicting Protein Networks. Genome Res, 15:945-953, 2005

[13] X. Chen, and M. Liu. Prediction of protein–protein interactions using random decision forest framework. Bioinformatics, 21(24):4394-4400, 2005

[14] S. Lo, C. Cai, Y. Chen and M. Chung. Effect of training datasets on support vector machine prediction of protein-protein interactions, Proteomics, 5(4): 876-84. 2005

[15] Z. Barutcuoglu, R. Schapire and O. Troyanskaya. Hierarchical multi-label prediction of gene function, Bioinformatics, 1(22):830-6. 2006

[16] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction, Proteins: Structure, Function, and Bioinformatics, 63(3): 490 - 500. 2006

[17] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart and R. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell, 2(1) 65-73. 1998

[18] H.W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd and B Weil. MIPS: a database for genomes and protein sequences. 30(1):31-34. 2002

[19] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, M.J Cherry, A.P. Davis, K. Dolinski, S.S Dwight, J.T. Eppig, M.A. Harris, D.P Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin and G Sherlock. Gene Ontology: tool for the unification of biology. Nature, 25(1) 2000

[20] F. Azuaje, H. Wang, O. Bodenreider. Ontology-driven similarity approaches to supporting gene functional assessment. Proc.Of The Eighth Annual Bio-Ontologies Meeting. 2005

[21] H. Yu, D. Greenbaum, H. Lu, X. Zhu, and M. Gerstein. Genomic analysis of essentiality within protein networks, Science,20(6):227-231. 2004

[22] D. Greenbaum, R. Jansen and M. Gerstein. 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, 18(4):586-596, 2002

[23] I.H. Witten and E. Frank. Data mining Practical machine learning tools and techniques.2nd Edition. Elsevier. 2005

[24] Y. Qi, J. Klien-Seetharaman and Z. Bar-Joseph. Random Forest Similarity for Protein-Protein interaction Prediction from Multiple Sources. Pacific Symposium on Biocomputing. 10:531-542. 2005

[25] Min Su Lee, Seung Soo Park and Min Kyung Kim. A protein interaction verification system based on a neural network algorithm. IEEE Computational Systems Bioinformatics Conference – Workshops (CSBW'05), 151-154, 2005

[26] C.M. Bishop. Neural Networks for pattern recognition. Clarendon Press, Oxford. 1995

[27] A McCallum and K. Nigam. A comparison of event models for Naïve Bayes text classification. AAAI/ICML-98 Workshop on learning for text categorization, pp 41-48. 1998

[28] R.E Schapire. The strength of weak learn ability. Machine learning, 5:197-227. 1990

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