World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:3, No:5, 2009

An Integrative Bayesian Approach to Supporting the Prediction of Protein-Protein Interactions: A Case Study in Human Heart Failure

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

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/7161/pdf

***Abstract*—**Recent years have seen a growing trend towards the integration of multiple information sources to support large-scale prediction of protein-protein interaction (PPI) networks in model organisms. Despite advances in computational approaches, the combination of multiple "omic" datasets representing the same type of data, e.g. different gene expression datasets, has not been rigorously studied. Furthermore, there is a need to further investigate the inference capability of powerful approaches, such as fully-connected Bayesian networks, in the context of the prediction of PPI networks. This paper addresses these limitations by proposing a Bayesian approach to integrate multiple datasets, some of which encode the same type of "omic" data to support the identification of PPI networks. The case study reported involved the combination of three gene expression datasets relevant to human heart failure (HF). In comparison with two traditional methods, Naive Bayesian and maximum likelihood ratio approaches, the proposed technique can accurately identify known PPI and can be applied to infer potentially novel interactions.

***Keywords*—**Bayesian network, Classification, Data integration, Protein interaction networks.

I. Introduction

PROTEIN-protein interactions (PPI) are crucial for many biological functions within the cell. These processes include signal transduction, protein folding, cell cycle control, DNA replication and transport [1]. The systematic identification of PPI networks will increase our understanding of health and disease, and will assist in the identification of novel biomarkers and therapeutic approaches. This has motivated efforts to map PPI on a proteomic-wide large-scale [2]. For example, PPI maps have been produced for yeast [3, 4], fruit fly [5] and human [6] through the use of experimental high-throughput technologies, including yeast two-hybrid (Y2H), Mass Spectrometry (MS) and Tandem Affinity Purification (TAP). However, these data are highly noisy or incomplete [7]. The limitations of using predictions obtained from a single information source (either experimental or computational) have been widely discussed [8].

In an attempt to obtain a better understanding of interactomes (the complete set of PPI in an organism), researchers [9, 10] have applied different computational approaches to integrate available PPI data for organisms to aid in describing interactomes (all possible PPI in an organism). It has been found that the integration of diverse "omic" features could significantly improve the inference of PPI networks [9-12]. Some of the advantages of computationally integrating data sources for PPI inference are: when two or more diverse datasets support a prediction, confidence in the PPI prediction increases; diverse datasets may cover different areas of the interactome; and integrating diverse datasets may increase predictive coverage of the interactome [12].

Different computational methods have been proposed to combine diverse heterogeneous datasets for the prediction of PPI networks [13-15]. For instance, a recent study in [14] provided an integrated analysis of human PPI using a Naïve Bayesian (NB) classifier. Four types of data were employed: homology derived PPI, gene co-expression, shared biological function and domain interaction. The Human Protein Reference Databases (HPRD) [16]was employed as the Gold Standard (GS), i.e. the set of known PPI used to implement and test the prediction model. Experimental methods confirmed protein interactions predicted by the framework. Scott *et al.* [13] constructed a probabilistic framework to integrate diverse features including co-expression, localization (proteins found in the same sub-cellular location are more likely to be interacting than proteins found in different sub-cellular locations), domain-domain interactions. A total of 37,606 PPI was predicted, 80% of these predicted PPI were not found in other human PPI databases. A recent study by Qi *et al.* [15] addressed the limitations imposed by missing data and feature redundancy for inferring PPI in human. A "mixture-of-features" framework was employed to predict

This work was supported in part by a grant from EUFP6, CARDIOWORKBENCH project.

F. Browne is with the School of Computing and Mathematics, University of Ulster BT37 0QB UK (e-mail: browne-f@ulster.ac.uk).

H. Zheng is with the School of Computing and Mathematics, the Computer Science Research Institute, University of Ulster BT37 0QB UK (e-mail: h.zheng@ulster.ac.uk).

H. Wang is with the School of Computing and Mathematics, the Computer Science Research Institute, University of Ulster BT37 0QB UK (corresponding author phone: +44 (0)2890368908; fax: +44 (0)2890366068; e-mail: hy.wang@ulster.ac.uk).

F. Azuaje is with the Public Research Centre for Health (CRP-Santé), L-1445 Luxembourg (e-mail: francisco.azuaje@crp-sante.. lu).

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/7161/pdf

PPIs. Knowledge from biological experts was incorporated to aid in the prediction of PPI. This approach obtained better precision-recall results in comparison to other classifiers including NB, support vector machines and random forest. Furthermore, 18 potentially novel interactions were identified.

Recent studies have mainly focused on the development of different computational techniques to integrate diverse datasets extracted from different "omic" features for the prediction of PPI networks. For instance: gene co-expression, shared biological function, homology derived PPI. The NB approach is perhaps one of most widely investigated models [9, 12, 17]. However, limited research has been performed where a fully-connected Bayesian approach is applied to integrate data sources for the inference of PPI networks. Furthermore, the combination of multiple datasets derived from sources with the same type of data hasn't been rigorously studied. Recently, Rhodes *et al.* [14] proposed a maximum likelihood ratio (MLR) approach to analyzing multiple datasets representing gene co-expression values only, i.e. only the maximum likelihood ratio per gene co-expression data source per protein pair was considered.

The aims of our study are: to investigate a Bayesian Network (BN) approach to the integration of multiple datasets, including datasets derived from the same type of data, and b) to study its potential relevance in the inference of a disease-specific PPI network. To address these aims, we present a case study in human Heart Failure (HF). Heart failure (HF) is one of the main causes of death in the world [18]. Dilated cardiomyopathy (DCM) is a common cause of HF in Western countries [19]. Gene expression studies have provided useful insights into the causation of DCM [19], therefore three gene co-expression datasets obtained from different human DCM studies were investigated in this study. The MLR approach applied by Rhodes *et al.* [14] and the NB method are employed as comparative approaches to the BN method. Different methods to measure their predictive performance were applied: Receiver Operating Characteristic ROC curves; Partial ROC curves; and True Positive (TP) / False Positive (FP) ratio.

The remainder of this paper is organized as follows. Section II briefly describes the data sources integrated in this study followed by a description of the methodologies applied. Section IV presents the results. The paper concludes with discussion of the results, the limitations of this study and future research.

II. Gold Standard and Prediction Features

In this study, the GS reference data set has been constructed with information extracted from the HPRD [16]. In the study by Rhodes *et al.* [14], four types of "omic" data were integrated for the prediction of human PPI. These include: homology derived PPI, gene co-expression, shared biological function and domain interaction. We also investigated these types of data to compare our approach to the MLP approach employed by Rhodes *et al.* [14] and the NB approach [9, 12]. Three diverse predictive features: homology derived PPI, shared biological function and domain interaction have been obtained from the study by Qi *et al.* [15]. Furthermore, in order to support the prediction of PPI networks relevant to human heart failure, three gene expression data sets associated with DCM were obtained from the Gene Expression Omnibus (GEO) [20]. Predictive features extracted from all these datasets were analyzed and integrated in the task of PPI inference. A description of the GS and predictive features analyzed in this study are described below.

A. Gold Standard

A GS is a reference dataset that contains protein pairs that are known to interact (i.e. positive cases) and non-interacting protein pairs (i.e. negative cases). The selection of a GS is an important task as it may be employed to measure the reliability of the genomic features or to validate computational PPI predictions. In this study, the positive GS has been generated from PPI information extracted from the HPRD [16]. Previous studies including [13-15] have used the HPRD as a data source to construct GS. The HPRD contains information on PPI which have been manually curated from literature by expert biologists. There is no direct information on protein pairs that do not interact. Therefore, the negative GS in this study has been constructed by generating random protein pairs and removing protein pairs found in the positive GS. This GS has been applied in the previous study [15] and contains a total of 13,184 protein pairs in the positive GS and 109,667 protein pairs in the negative GS.

B. Features

Six "omic" features have been assessed for the integrative prediction of PPI in human. A description of each feature is provided below. The feature name has been shortened for easier representation throughout the paper. A summary of these features can also be found in table I.

1) *Gene Ontology Biological Process (GOBP):* the GOBP data source has been obtained through extracting similarity information between gene products annotated to the Gene Ontology (GO) biological process terms [21]. It is assumed that two proteins that function in the same biological process are more likely to interact than proteins from different processes [14]. The similarity between protein pairs was calculated by determining how many times both proteins are in the same functional class of the GO biological process hierarchy. The GOBP dataset contains a total of 58,538 protein pairs with similarity values ranging from 0 to 6. As the similarity value increases, the likelihood of an interacting protein pair also increases.
2) *Homology (HOM):* a protein exhibiting a function in one organism may exhibit the same function in a different organism [15]. The homology relationship between human proteins and yeast proteins are based on sequence alignment scores from PSI-BLAST [15]. Publically available PPI datasets were downloaded from the Database of Interacting Proteins (DIP) [22]. The HOM dataset contains 122,851 protein pairs with values ranging from 0 to 317.

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/7161/pdf

TABLE I
A DESCRIPTION OF THE GENOMIC FEATURES. THE FIRST COLUMN PROVIDES THE NAME OF THE FEATURE FOLLOWED BY THE SOURCE, THE BIOLOGICAL ASSUMPTION, COVERAGE OF THE GS AND A REFERENCE

Feature	Source	Assumption	Coverage of GS	Ref
GOBP	Gene Ontology Biological Process	Proteins found in the same biological process are more likely to interact than proteins found in different biological process.	**Ovlp** + / - 10465/48073	[21]
COE1	Gene co-expression	Interacting proteins often have similar co-expression patterns, therefore interacting proteins are more likely to be co-expressed	**Ovlp** + / - 13169/109406	[20]
COE2	Gene co-expression	Interacting proteins often have similar co-expression patterns, therefore interacting proteins are more likely to be co-expressed.	**Ovlp** + / - 6148/35881	[20]
COE3	Gene co-expression	Interacting proteins often have similar co-expression patterns, therefore interacting proteins are more likely to be co-expressed.	**Ovlp** + / - 8424/61111	[20]
HOM	Homology	The function of proteins in model organisms often retains the same function in Human. Therefore a pair of interacting orthologs from a model organism is likely to be interacting in the Human organism.	**Ovlp** + / - 13184/109667	[15]
DOM	Domain	Proteins usually involve interactions between protein domains.	**Ovlp** + / - 10456/49819	[15]
GS	Human Protein Reference Database	Protein complex membership.	**GSP** 13184 **GSN** 109667	[16]

3) Domain interaction (DOM): The study in [14] suggests that novel PPI could be predicted by identifying pairs of domains over-represented in known interacting proteins [15]. Based on the calculation of the hypergeometric distribution of domain co-occurrence in protein interaction pairs [15], the number of interaction domains shared by a protein pair was computed. The DOM data source consists of 60,275 protein pairs with values which range from 0 to 25. It is assumed that the higher the DOM value, the higher the likelihood of an interaction.

4) *Gene Co-expression (COE1, COE2 and COE3):* Protein pairs that interact often have similar gene expression patterns [14]. Therefore, protein pairs which are co-expressed may be more likely to interact than protein pairs that are not co-expressed. Three microarray datasets analyzed in this case study have been obtained from the GEO [20], accession numbers: GDS1362, GDS2205 and GDS2206 (referred to as COE1, COE2 and COE3 throughout the paper). These data sources were selected to highlight the performance of the BN in comparison to the NB and MLP approach when data sources of the same data type are integrated. COE1 was obtained from an oligo array experiment consisting of 12 samples: 5 samples were obtained from non-failing heart patients and 7 samples from DCM heart patients. COE2 was extracted from a cDNA array consisting of 28 samples: 15 samples were obtained from non-failing heart patients and 13 samples from DCM heart patients. COE 3 was obtained from an oligo array containing 37 samples: 7 samples were obtained from non-failing heart patients, 20 from DCM heart patients. Samples were removed from the gene co-expression datasets if 50% or more of the co-expression values were missing. Each gene co-expression dataset was normalized per chip and then per gene. Values were normalized between -1 and 1 by calculating the mean and standard deviation of the row (per chip) and then the column (per gene). The values in the COE1, COE2 and COE3 datasets were obtained by calculating the Pearson's correlation co-efficient values between pairs of proteins. The COE1 contains 122,575 co-expression values, COE2 a total of 42,029 co-expression values and COE3 69,535 co-expression values. The closer a gene co-expression value is to 1, the likelihood of an interaction increases.

III. METHODOLOGY

Previous studies including [9, 12] have employed the NB technique to integrate diverse predictive features for PPI inference. However, the NB assumes conditional independence between features. The study in [12] suggested that subtle correlations or dependencies between features may have an adverse effect upon the NB classification performance. This study applies a fully-connected Bayesian (FCB) approach to infer PPI networks, whereby dependencies between features are taken into consideration. By modeling relationships between multiple "omic" features we aim to illustrate the improvement in terms of classification performance in comparison to the NB and MLR approaches.

A. A Bayesian Network Approach to Data Integration

To determine the likelihood of an interaction between a pair of proteins, the FCB approach takes into consideration the predictive features available. The inference of PPI may be defined as a binary classification problem. An interacting protein pair can be represented as a "*positive case" (pos)* and a non-interacting protein pair as a "*negative case*"*(neg)*. The feature values employed to predict PPI can be represented as $f_1, f_2 \ldots, f_n$. To determine the probability that two proteins

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/7161/pdf

are interacting given predictive evidence, the posterior odds of an interaction are calculated as follows:

$$P(pos \mid f_1, f_2, ..., f_n) = \frac{L(f_1, f_2, ..., f_n)}{L(f_1, f_2, ..., f_n) + P(pos) / P(neg)} \quad (1)$$

The $P(pos)$ and $P(neg)$ represent the prior probability of interacting and non-interacting protein pairs respectively. The likelihood-ratio can be calculated as:

$$L(f_1, f_2, ..., f_n) = \frac{P(f_1, f_2, ..., f_n \mid pos)}{P(f_1, f_2, ..., f_n \mid neg)} \quad (2)$$

In cases whereby dependencies between features exist (for instance statistical or biological relationships), a fully connected network is employed. The likelihood-ratio is calculated by the determination of all possible state combinations of these features. Although this process can be computationally intensive; the over-estimation of likelihood ratios is prevented. Furthermore, it has been observed that the performance of a Bayesian classifier that assumes independence between features may degenerate when non-independent features are included [23]. Features which are conditionally independent are integrated using a NB approach. The likelihood-ratio can be expressed as:

$$\begin{aligned} &L(f_1, f_2, \ldots, f_n) \\ &= \frac{P(f_1, f_2, \ldots, f_M \mid pos)}{P(f_1, f_2, \ldots, f_M \mid neg)} \times \prod_{i=M+1}^{n} \frac{P(f_i \mid pos)}{P(f_i \mid neg)} \\ &= L(f_1, f_2, \ldots, f_M) \times \prod_{i=M+1}^{n} L(f_i) \end{aligned} \quad (3)$$

The $f_1, ..., f_M$ represents values obtained from features that are not conditionally independent and $f_{M+1}, ..., f_n$ represents values obtained from features that are independent. The likelihood-ratio can be calculated as the simple product of likelihood-ratios obtained from FCB and NB approaches. The $P(f_i \mid pos)$ and $P(f_i \mid neg)$ values are obtained from the feature values f_i that overlap with the positive and negative cases in the GS respectively.

B. Inferring Protein-Protein Interactions

As illustrated in Equation 2, there is a link between the posterior probability of an interaction and the likelihood ratio. For instance, the posterior probability increases monotonically with the calculated likelihood ratio. Thus, Bayesian classification can be employed by using the combined likelihood-ratio scores. If the posterior probability is greater than 0.5 then the protein pair is predicted as interacting. The study in [13] suggests in order to obtain a posterior probability of an interacting protein greater than 0.5, the likelihood-ratio should be larger than 400.

To evaluate the predictive performance of the FCB approach 10-fold cross validation (CV) has been applied to train and test the classifier. In 10-fold CV, the dataset is firstly divided into 10 equal partitions. At each stage of the CV, nine partitions are used train the classifier and one partition is applied for testing the classification performance. At each run of the CV, likelihood-ratios are estimated from the data in the training partitions. Associated likelihood-ratios for each protein pair are obtained for cases in the test partition. At the end of the CV process, inferred protein predictions are validated against protein pairs in the GS. From this, the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) are calculated.

C. Evaluating Predictive Performance

Three evaluation methods have been applied to assess the predictive quality of the FCB, MLP and NB approaches. These three measures are described below.

ROC Analysis: A ROC curve is used to capture in a single graph the trade-off between sensitivity and specificity over the entire range of a dataset. In this study, a ROC curve is employed to graphically illustrate the performance of the BN plotted for different likelihood-ratio thresholds (TH). The TP, TN, FP, and FN are calculated using the 10-fold CV analysis. The sensitivity and specificity are calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

A predictive dataset will obtain a ROC curve that rises steeply to the left hand side of the graph and has a large area under the curve (AUC). AUC values are estimated from the 10-fold CV procedure.

Partial ROC Curve: Due to the imbalance of the dataset applied to infer PPI in human, the ROC measurement may produce an overly optimistic view of the classification performance. Therefore, it may be informative to report AUC values obtained from a partial ROC to assess the predictive performance. The partial ROC AUC measures predictions made which exceed a minimum TH of 400. The AUC values for the partial ROC method are referred to as AUC400 throughout this paper.

True to False Positive (TP/FP) ratio: The true to false positive (TP/FP) ratio is plotted against the TH of likelihood ratio as a measure of the probability of a positive interaction. This measure has been previously employed in [9] and described below:

$$\frac{TP}{FP}\Big|_{L=TH} = \sum_{L=TH} \frac{N_{pos}(L)}{N_{neg}(L)} \quad (6)$$

The $N_{pos}(L)$ and $N_{neg}(L)$ are the number of interacting and non-interacting protein pairs in the GS with a given likelihood ratio of *L*.

IV. Results

A. Estimation of Likelihood Ratios

Likelihood ratios for each individual dataset was estimated based on the overlap of protein pairs between the dataset and the GS as shown in columns TP and TN in tables II to VII. The number of categories each dataset was discretized into is shown in the first column in these tables. The conditional probabilities of the given values and the corresponding likelihood ratio are provided in the last three columns. As shown in these tables, when using a single dataset, most protein pairs have a very low likelihood ratio, well below the minimum TH cut-off (400) which is required for the posterior probability of an interaction to be greater than 0.5. Hence, based on the information from one single dataset, none of the protein pairs can be predicted as a true positive with probability greater than 0.5. This highlights the importance of dataset integration for the prediction of PPI relevant to the development of DCM.

TABLE II
THE ESTIMATE OF LIKELIHOOD RATIOS FOR COE1 DATASET. THE DATASET WAS DISCRETIZED INTO 4 CATEGORIES AS SHOWN IN THE FIRST COLUMN

Range	Gold-standard Overlap		p(value/pos)	p(value/neg)	LR
	TP	TN			
[-1.0, -0.7]	388	3,499	0.03	0.03	0.92
[-0.7, 0.2]	7,305	67,878	0.5	0.62	0.89
[0.2, 0.9]	5,453	37,963	0.41	0.35	1.19
[0.9, 1.0]	23	66	1.75E-03	6.03E-04	2.90

TABLE III
THE ESTIMATE OF LIKELIHOOD RATIOS FOR COE2 DATASET. THE DATASET WAS BINNED INTO 5 CATEGORIES AS SHOWN IN THE FIRST COLUMN

Range	Gold-standard Overlap		p(value/pos)	p(value/neg)	LR
	TP	TN			
[-1.0, 0.2]	3,525	23,953	0.57	0.668	0.89
[0.2, 0.7]	2,341	11,259	0.38	0.314	1.21
[0.7, 0.8]	199	532	0.03	0.0148	2.18
[0.8, 0.9]	69	128	0.01	3.57E-02	3.15
[0.9, 1.0]	14	9	2.28E-03	2.51E-04	9.08

TABLE IV
THE ESTIMATE OF LIKELIHOOD RATIOS FOR COE3 DATASET. THE DATASET WAS BIINNED INTO 3 CATEGORIES AS SHOWN IN THE FIRST COLUMN

Range	Gold-standard Overlap		p(value/pos)	p(value/neg)	LR
	TP	TN			
[-1.0, 0.2]	6,340	47,081	0.75	0.77	0.98
[0.2, 0.8]	2,069	14,010	0.25	0.23	1.07
[0.8, 1.0]	15	20	1.78E-03	3.27E-04	5.44

TABLE V
THE ESTIMATE OF LIKELIHOOD RATIOS FOR GOBP DATASET. THE DATASET WAS DISCRETIZED INTO 4 CATEGORIES AS SHOWN IN THE FIRST COLUMN

Range	Gold-standard Overlap		p(value/pos)	p(value/neg)	LR
	TP	TN			
[0, 1]	7,990	46,364	0.76	0.96	0.79
[1, 2]	1,768	1,535	0.17	0.03	5.29
[2, 3]	535	162	0.05	3.37E-03	15.17
[3, 4]	169	12	0.02	2.50E-04	64.71

TABLE VI
THE ESTIMATE OF LIKELIHOOD RATIOS FOR DOM DATASET. THE DATASET WAS BINNEDZED INTO 4 CATEGORIES AS SHOWN IN THE FIRST COLUMN

Range	Gold-standard Overlap		p(value/pos)	p(value/neg)	LR
	TP	TN			
[0, 6.7]	3,411	35,653	0.33	0.72	0.46
[6.7, 8.6]	498	916	0.05	0.02	2.59
[8.6, 9]	456	3,589	0.04	0.07	0.61
[9, 25]	6,091	9,661	0.58	0.19	3.00

TABLE VII
THE ESTIMATE OF LIKELIHOOD RATIOS FOR HOM DATASET. THE DATASET WAS DISCRETIZED INTO 4 CATEGORIES AS SHOWN IN THE FIRST COLUMN

Range	Gold-standard Overlap		p(value/pos)	p(value/neg)	LR
	TP	TN			
[0.0, 0.66]	12,524	109,318	0.950	0.997	0.95
[0.66, 4.13]	429	297	0.0325	0.00271	12.02
[4.13, 11.39]	130	41	0.00986	0.000374	26.37
[11.39, 317]	101	11	0.00766	0.000100	76.38

B. Data Integration based on Likelihood Ratios

In this section the following three approaches, namely MLR, NB and FCB, were used to combine the three gene co-expression datasets (COE1, COE2, COE3); GOBP dataset; HOM dataset and DOM dataset. The results were compared in terms of AUC values and the ratio of TP/FP. To evaluate the predictive performance, 10-fold cross validation was carried out.

- MLR – similar to the approach applied by Rhodes *et al.* [14] the maximum likelihood ratio obtained per co-

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/7161/pdf

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/7161/pdf

expression data type per protein pair was integrated with the features GOBP; DOM; HOM using NB. For instance if COE2 obtains the highest likelihood ratio for a protein pair compared to COE1 and COE3, then the COE2 likelihood is employed only.

- NB - obtained when the features COE1; COE2; COE3; GOBP; DOM; HOM were integrated using NB approach.
- FCB – repsresent the results obtained when the COE datasets COE1; COE2; COE3 are integrated using a fully connected Bayesian approach and GOBP; DOM; HOM are integrated using the NB approach.

Fig. 1 shows the relationship between the TP/FP ratio and likelihood ratio for the three data integration models. By integrating three expression dataset (*COE1*, *COE2* and *COE3*) with FCB approach, the TP/FP ratio is improved significantly, in particular, when the TH becomes large (greater than 400).

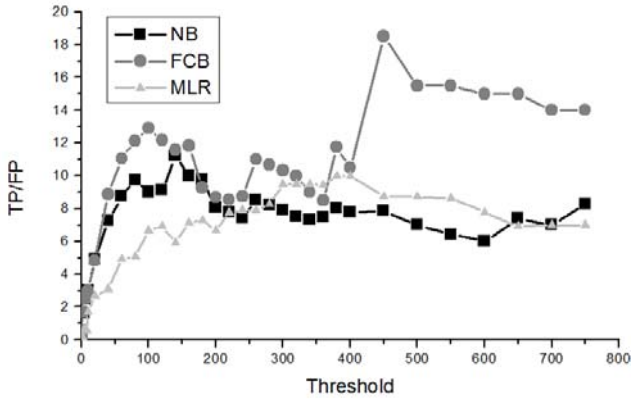


Fig. 1 Comparison of three data integration models in terms of the ratio of TP/FP. The selected TH of likelihood ratio is plotted on the x-axis, and the corresponding TP/FP ratio is shown on the y-axis. A protein pair is predicted as positive if its combined likelihood ratio is greater than a given TH. The TP/FP ratio was computed as the ratio of the number of positives and negatives in the gold-standard given a particular likelihood ratio

The ROC curves of the predictions are presented in Fig. 2, in which the y-axis represents sensitivity (TP/TP+FN), and x-axis shows the value of 1-specificity (FP/TN+FP). Surprisingly, in Fig. 2(B) all three integration schemes achieve similar prediction performances. The advantage of FCB approach is not evident. However, a close examination reveals that, due to the high imbalance of the dataset employed (for instance for every "interacting" protein pair there are 29 "non-interacting" protein pairs in the GS), the ROC measurement may produce an overly optimistic view of the classification performance. In order to obtain a more realistic picture of the performance obtained from three data integration approaches, we also calculated the partial ROC. This measurement may be more informative to assess the predictive performance obtained from a highly skewed dataset, as shown in table VIII. Clearly, the FCB approach achieves better results based on the calculation of AUC400. Furthermore, the partial ROC curves plotted in Fig. 2(A) illustrate a steeper curve obtained for the FCB integration scheme in comparison to the curves from the NB and MLR integration schemes.

TABLE VIII
AUC AND AUC400 VALUES OBTAINED WHEN THE FEATURES COE1, COE2, COE3, DOM, HOM AND GOBP WHERE INTEGRATED USING NB, MLR AND FCB TECHNIQUES

Measures	NB	MLR	FCB
AUC	0.705	0.703	0.699
AUC400	1.54E-07	1.38E-07	2.08E-07

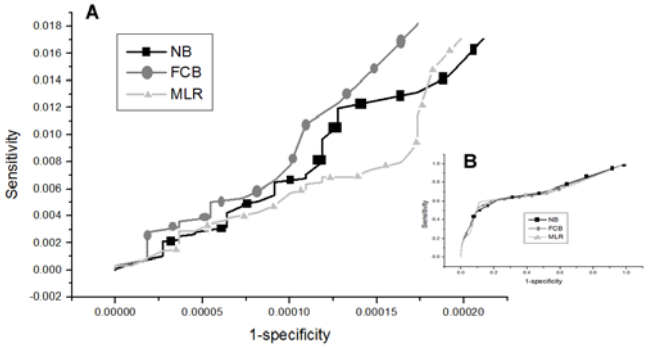


Fig. 2 (A) Illustrates the partial ROC curves generated with the three data integration schemes. (B) Illustrates the ROC curves generated with three data integration schemes

V. DISCUSSION AND CONCLUSION

Limited research has been performed in applying a FCB approach to integrate diverse "omic" features for the prediction of PPI in human. Furthermore, the integration of multiple "omic" data has yet to be thoroughly addressed. This paper presented a FCB approach for the integration of multiple datasets, (with some datasets containing the same type of "omic" data) for the inference of disease-specific PPI networks. We demonstrated the application of the FCB approach in the prediction of PPI networks relevant to development of DCM. A previous study by Rhodes *et al.* [14] utilized four types of "omic" data for the prediction of PPI. Therefore, the same "omic" data types were employed in this paper for comparative purposes. Furthermore, Rhodes *et al.* [14] proposed a MLR approach when dealing with multiple datasets of the same "omic" type (for instance, multiple gene co-expression datasets). The case study presented in this paper employed three gene co-expression datasets (COE1, COE2, COE3) relevant to human HF along with three other datasets: DOM; HOM; GOBP to reconstruct a PPI network relevant to the development of DCM. By modeling relationships between multiple datasets of the same "omic" type, an improvement in prediction performance was achieved in terms of AUC400 and the ratio of TP/FP by the FCB approach in comparison to the MLR and NB approaches.

Currently the proposed FCB technique was tested on a relatively small number of datasets with the same type of "omic" data. The behavior of FCB when dealing with large number of datasets deserves further investigation. In addition, the GS for true positives used in this study was generated from HPRD [16]. The study of the impact of using other sources, such as GRID [24] as a GS provides one direction for future research. Furthermore, the task of constructing a

negative GS presents a significant challenge. It is difficult to experimentally validate that two proteins are not interacting. In this study, we randomly selected protein pairs to generate the negative GS. However, in future work other negative GS construction methods could be applied. For instance, the study in [12] uses sub-cellular membership information to construct the negative GS.

This study contributes to the development of computational approaches to supporting the integration of diverse sources of genomic and proteomics information for comprehensive large-scale prediction of PPI networks.

REFERENCES

[1] C. Royer, "Protein-protein interactions," Outline of the Thermodynamic and Structural Principles Governing the Ways that Proteins Interact with Other Proteins. Previously Published in the Biophysics Textbook Online (BTOL), 1999.

[2] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield et al., "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." Nature, vol. 403, pp. 623-627, 2000.

[3] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, et al., "Proteome survey reveals modularity of the yeast cell machinery," Nature, vol. 440, pp. 631-636, 2006.

[4] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, et al., "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," Nature, vol. 440, pp. 637-643, Mar 30. 2006.

[5] M. Middendorf, E. Ziv and C. H. Wiggins, "Inferring network mechanisms: The Drosophila melanogaster protein interaction network," Proceedings of the National Academy of Sciences of the United States of Americal, vol. 102, pp. 3192-3197, 2005.

[6] R. M. Ewing, P. Chu, F. Elisma, H. Li, et al., "Large-scale mapping of human protein–protein interactions by mass spectrometry," Molecular Systems Biology, vol. 3, 2007.

[7] C. von Mering, R. Krause, B. Snel, M. Cornell, S. Oliver, et al.,"Comparative assessment of large-scale data sets of protein-protein interactions". Nature 417(6887), pp. 399-403, 2002.

[8] H. Ge, A. Walhout and M. Vidal, "Integrating 'omic'information: a bridge between genomics and systems biology," Trends Genet., vol. 19, pp. 551-560, 2003.

[9] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, et al., "A Bayesian networks approach for predicting protein-protein interactions from genomic data," Science, vol. 302, pp. 449-453, 2003.

[10] 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)," Proceedings of the National Academy of Sciences, vol. 100, pp. 8348-8353, 2003.

[11] E. M. Marcotte, "Detecting Protein Function and Protein-Protein Interactions from Genome Sequences," Science, vol. 285, pp. 751-753, 1999.

[12] L. J. Lu, Y. Xia, A. Paccanaro, H. Yu and M. Gerstein, "Assessing the limits of genomic data integration for predicting protein networks," Genome Res., vol. 15, pp. 945, 2005.

[13] M. S. Scott and G. J. Barton, "Probabilistic prediction and ranking of human protein-protein interactions," BMC Bioinformatics, vol. 8, pp. 239, 2007.

[14] D. Rhodes R., S. Tomlins A., S. Varambally, V. Mahavisno, et al., "Probabilistic model of the human protein-protein interaction network". Nature 23(8), pp. 951-959, 2005.

[15] Y. Qi, J. Klein-Seetharaman and Z. Bar-Joseph, "A mixture of feature experts approach for protein-protein interaction prediction," BMC Bioinformatics, vol. 8 Suppl 10, pp. S6, 2007.

[16] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, et al., "Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans," Genome Res., vol. 13, pp. 2363, 2003.

[17] 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, vol. 63, pp. 490 - 500, 2006.

[18] American Heart Association (AHA) American Heart Association, "Heart diseases and stroke Statistics–2007 update," 2007

[19] A. Camargo and F. Azuaje, "Linking Gene Expression and Functional Network Data in Human Heart Failure," PLoS ONE, vol. 2, 2007.

[20] "Gene Expression Omnibus" [http://www.ncbi.nlm.nih.gov/geo/]

[21] M. Ashburner, C. Ball and J. Blake, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium Database resources of the National Center for Biotechnology Information," Nucleic Acids Res., vol. 34, 2006.

[22] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie and D. Eisenberg, "The Database of Interacting Proteins: 2004 update," Nucleic Acids Res., vol. 32, pp. 449-451, 2004.

[23] C. J. Needham, J. R. Bradford, A. J. Bulpitt and D. R. Westhead, "A primer on learning in Bayesian networks for computational biology," PLoS Comput Biol, vol. 3, pp. e129, 2007.

[24] B. J. Breitkreutz, C. Stark and M. Tyers, "The GRID: the General Repository for Interaction Datasets," Genome Biol., vol. 4, pp. R23, 2003.

Open Science Index, Mathematical and Computational Sciences Vol:3, No:5, 2009 publications.waset.org/7161/pdf