A novel method for pathway classification based on gene expression data and known gene networks

CHANCHAL KUMAR, JAGIR R HUSSAN
Technology Incubation Center
IBM Software Labs, India
Embassy Golf Links Business Park, Block-C, Floor-7, Off Indiranagar-Koramangala
Intermediate Ring Road, Bangalore-560071
INDIA

Abstract: - Microarrays experiments provide genome-wide expression pattern across many different conditions that provide a landscape of gene expression in a particular cellular state. The information derived from this high throughput technology is proving to be highly constructive in understanding the functioning of organism on the molecular level. The regulation of gene expression is in turn achieved through genetic regulatory systems structured by networks of interactions between DNA, RNA, proteins, and small molecules. Using these two data sources in conjunction with appropriate mathematical formalism we can model the abstract problem of classification of pathways pertaining to an observed phenotype.

We present an approach for identifying the class/phenotype of a putative pathway. The basic idea is to create a compendium of pathways, which are based on perturbed cellular states and use the compendium for classification of uncharacterized pathways. We propose to integrate information contained in high throughput gene expression data with the known gene regulatory pathways to devise scoring schema and suitable data structures to create the compendium. For classification we applied an augmented graph theoretic algorithm. The significance of the approach has been ascertained by an appropriate statistical method. We also illustrate that the model developed is nonparametric and generic. The insight derived from current approach suggests that our method is only limited by current state of knowledge about gene networks and analysis of gene expression data. More robust results can be achieved as the knowledge base becomes rich and more accurate.

Key-Words: microarrays; gene networks; graphs; sub-graph isomorphism; graph matching; inexact graph matching; attribute relational graph

1 Introduction

High throughput techniques like microarray experiments measures mRNA expression levels of thousands of genes in parallel, this seems to be a valuable tool for understanding the underlying genetic machinery of a living cell [1]. Inferential and descriptive statistical methods have been applied to microarray data to uncover patterns of gene expression and behavior of genetic markers in diseases like cancer [2,3]. Reverse engineering of gene networks using gene expression data based on machine learning algorithms has gained momentum in recent years [4]. Concurrent to these computational approaches there are ongoing efforts which, based on elaborate in-vivo and in-vitro experimentation like ChIP, Y2H screens and Protein chips are providing us experimentally verified facts about mechanism of genetic and protein interactions, genetic regulation and gene expression in various prokaryotes and eukaryotes [5,6]. Serious attempts have been made to understand the underlying topology of the biochemical networks and to establish structure-function kind of relationship in organisms based on the topology and other properties of these networks [7,8]. Attempts to understand the regulatory and transcriptional mechanisms on system level are underway and knowledge from different domains are being incorporated to develop in-silico models of cells and organisms

[9,10]. All these studies have contributed significantly to our present knowledge about living organisms and their mode of function. But still there are many open-ended questions to be answered. The key focus of postgenomic biomedical research is to undertake a holistic and systemic approach towards understanding the complex machinery called "cell". In this paper we explain the approach we have taken to combine knowledge from gene expression data along with known gene networks, which allows us to build a compendium of reference pathways. The compendium is used in conjunction with an augmented graph theoretic algorithm to classify putative pathways. The approach presented is generic though we have verified it using the data for yeast, *Saccharomyces cerevisiae*.

Our approach is based on the hypothesis that the use of available knowledge on biological networks coupled with inference derived from large-scale gene expression data will aid in development of reliable and robust models and methods, which can be applied to gain valuable insights for various applications. Few efforts have been made to combine available knowledge on biological networks and gene expression data [11,12]. Most of these efforts were centered on analysis of gene expression data based on the network structure or in characterizing pathways, which are best described by the gene expression profile. Still, they illustrate that an integrated approach that combines knowledge from gene expression data and biochemical networks can aid us in decoding the phenomenon of life. The recent review on network biology consolidates the same [13]. Our work builds up on these concepts. Starting from the known gene networks, we extract measures from microarray data to assign score to the genetic relationships in these pathways. This procedure yields us model pathways, which are stored as compendium and used in classification of a putative pathway. The classification strategy is based on an "Inexact sub-graph isomorphism algorithm" which has been customized to solve the problem in hand. The scoring function can be modified to reflect any desired property as long as it is supported by gene expression data and network characteristics. The scoring function proposed in this paper has been designed keeping in mind that the regulatory relationships among genes are linear as well as non-linear.

2 Methods

2.1 Expression data

For calculation of measures, we used microarray data used and reported by Hughes *et al.* [14]. The data set contained 276 single gene deletion experiments of *Saccharomyces cerevisiae* mutant strains. For each relationship between two entities in the network the score was calculated from the gene expression values of the two entities. For details of the score metric refer section 2.4. The dataset was reformatted in the form of N × M matrix where, N=number of genes, M=number of samples and N > M. For subsequent discussions we refer to expression profile of a gene X as expression vector $\boldsymbol{X}$ such that $\boldsymbol{X} = \{x_1, x_2, x_3, \ldots, x_M\}$ where, x_i is the expression value of gene **X** in experiment *i*.

2.2 Data preprocessing

For each cellular network a matrix (n × m) of expression data was created from the original dataset provided by Hughes *et al.,* 2000. We used Gene Ontology(GO) information from SGD to create these 7 matrices. First we used SGD Gene Ontology Term Mapper to cluster 287 deletion experiments (containing genes and uncharacterized ORF's) under different groups according to the cellular (biological) processes they shared. Second, the obtained clusters of microarray experiments were mapped to the seven pathways. The cluster that was nearest in terms of definition to a particular pathway was chosen as the set of experiments that will be included to create the data matrix. Thus we choose 7 GO clusters (containing deletion experiments) to create the matrix for 7 pathways. Finally, the data matrix had rows as the genes that were present in the pathway and columns as experiments contained in the nearest GO cluster.

2.3 Pathway construction

For illustrating the approach we collected 7 cellular networks of *S. cerevisiae* from literature sources and pathway database KEGG [15,16,17,18]. The criterion for selecting these networks was that they had a corresponding entry in the gene expression data that we used to score the genetic relationships. The entities in these networks are genes, transcription factors (TFs),

protein and their complexes. The relation between any two entities is directed. The detail of the seven pathways taken is presented in table 1.

Model Graphs	Description
Cell cycle	Transcriptional regulatory network in yeast cell cycle
MAPK	Pheromone response pathway in yeast cell
PKC	Protein kinase C(PKC) pathway activated by cell surface stress during formation of mating projection in yeast cell
Galactose	Galactose utilization pathway in yeast cell
HOG	The HOG pathway in response to hypertonic stress
Filament	The filamentous growth pathway to promote haploid invasive growth in rich medium low in nitrogen
Respiration	Aerobic respiration in yeast cell

Table 1: Description of the pathways stored as compendium

2.4 Score calculation

For scoring the relationships among entities in a pathway we used two measures based on correlation and mutual information. For the calculation of correlation and mutual information we used the data set from Hughes *et al.* [14].

2.4.1 Calculating correlation

Correlation was calculated between two genes X and Y from their respective expression vectors ***X*** and ***Y*** (with ***M*** data points) using the Pearson correlation **(*ρ*)** metric (1).

$$\rho(X,Y) = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{x_j-\mu_x}{\sigma_x}\right)\left(\frac{y_j-\mu_y}{\sigma_y}\right) \quad (1)$$

Where μ_i and σ_i denote mean and standard deviation of data for gene *i.* We took the absolute values of $\rho(X,Y)$ for our analysis. This absolute correlation was defined as (2)

$$\rho_{norm} = |\rho(X,Y)| \quad (2)$$

According to Hughes *et al.* [14] the absolute value of $\rho(X,Y)$ is used because conditions that produce perfectly anticorrelated response are sometimes considered as redundant from information theory point of view.

2.4.2 Calculating mutual information

The mutual information, *I(X,Y)* between two genes X and Y was calculated using equation (3).

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \quad (3)$$

Where, $H(I)$ = Entropy for single gene I.

And, $H(X,Y)$=Joint entropy of gene X and Y.

For calculating *I(X,Y),* the expression values were discretized, using histogram based technique discussed in [19]. As accounted by Michaels *et al.* [20] the mutual information depends on the distribution of individual datasets. Therefore, for a coherent analysis, we normalized *I (X, Y)* to the maximal entropy of each of the contributing expression vectors (numerical range: 0-1), giving a high value for highly correlated genes, independent of the individual entropies. The normalized entropy was defined as:

$$I_{norm} = I(X,Y) / \max\{H(X), H(Y)\} \quad (4)$$

2.4.3 Combining scores

The reason for calculating two different scores to signify a relationship was the known fact that genetic (regulatory) relationships are not always linear. While correlation based method can detect linear relationships among expression patterns, information theory approach serves the purpose to uncover relational features that are not readily discovered by correlation. The two scores ρ_{norm} and I_{norm} were combined to yield a composite score which we define as:

$$\xi_{total} = \alpha \times \rho_{norm} + \beta \times I_{norm} \quad (5)$$

Where, α and β are tuning parameters which were both fixed at $\alpha = \beta = 1$ for our analysis. The score ξ_{total} defines the weight of the relation X→Y between two entities X and Y in the network. As, $0 \le \rho_{norm} \le 1$ and $0 \le I_{norm} \le 1$ so, as per equation (5), $0 \le \xi_{total} \le 2$. For each of the seven selected networks (see table1) weights to all the relations of type X→Y were assigned using equation (5). These networks were subsequently stored as weighted directed graphs, each we call model graph (G_M). The model graphs are stored as a compendium against which we match a hypothetical Input graph (G_I), iteratively using a modified sub-graph isomorphism approach (refer section 3.2.1). The input graph is assigned the class label of the best matching model graph based on edit costs operations (refer section 3.2.2).

3 Classification

The main focus of our approach was to devise a classification methodology that can classify a hypothetical pathway based on known biochemical and gene regulatory networks and available gene expression data. We followed following steps to accomplish the goal:

1) Extract relevant biochemical networks from literature and database.
2) Score the relationships among entities in these networks by measures derived from gene expression data.
3) Create a compendium of such networks and store those using appropriate data structures.
4) Devise an appropriate algorithm that can classify a new pathway based on compendium of pathways.

In the preceding section we explained steps 1 through 3. In the following we discuss the data structures, algorithm and the classification strategy in detail. For the sake of simplicity we refer to pathways and networks synonymously in all contexts.

3.1 Data structures

The algorithm used for classification works on labeled graphs. The data structure we choose for representing the model graphs (G_M) and the hypothetical input graph (G_I) was Attribute Relational Graphs (ARG). In an ARG the nodes and edges are assigned labels. Due to their representational power ARG's are widely used in various applications in computer vision and pattern recognition. Figure 1 shows an ARG representation of a subnetwork of galactose utilization pathway.

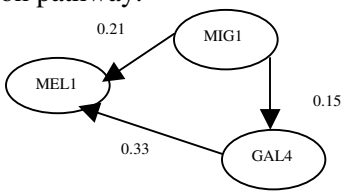

Fig.1: An ARG representation for a subnetwork of galactose utilization pathway

3.2 Inexact graph matching

In many applications complex structures are classified, detected, or compared to each other by means of an appropriate matching scheme. We propose that classification of a pathway can be modeled as a graph matching problem. A graph is an ideal data structure to represent a cellular network. We used a modified version of "inexact graph matching algorithm" which had been earlier applied for finding best possible match between two graphs[21]. The algorithm was implemented assuming that the input graphs are *ARG's*.

3.2.1 Classification strategy

We created a compendium of pathways pertaining to perturbed cellular states based on the gene expression data of the key genes involved in the pathways. In order, to classify a putative pathway, we match the input pathway against this compendium. The algorithm applied was the modified "inexact graph matching algorithm" which falls under the category of subgraph isomorphism problem. This algorithm is in turn inspired by Ullman's algorithm and error-tolerant subgraph isomorphism procedure[22,23]. Details of the algorithm are available in original paper. In order to compare the input graph to the compendium of model graphs and decide which of the models is most similar to the input, it is necessary to define a distance measure for graphs. Given two ARG's/graphs G_M and G_I, the goal is to find the best matching between their nodes that leads to the smallest matching error. This smallest error between the two graphs can be viewed as the distinguishing distance between them. To compute the matching error, we compute the dissimilarity between each pair of matched nodes, plus the dissimilarity between (corresponding) edges. The matching error is calculated on the basis of edit operations which have been defined keeping in mind their biological relevance. Similar to the error-tolerant subgraph isomorphism problem where edit operations are used to define graph edit distance, we define edit operations on the input graph such that they are transformed to match a model graph [23]. To each edit operations, a certain cost is assigned. We consider the following edit operations for an input graph: {*vertex label distortion, edge label distortion, missing edges, reversed edges*}. The details of edit costs are provided in section 3.2.2. The input graph (pathway) is simply assigned class label of the best matching model graph (pathway) based on minimum matching error. The steps in compendium creation and pathway classification are illustrated in figure 2.

3.2.2 Edit costs/penalty

As mentioned earlier, an important factor in the application of the graph-matching algorithm is the definitions of costs of edit operations. For the subsequent discussions we define input graph G_I and model graph G_M as two ARG's such that, $G_I = (V_I, E_I, \omega_I, \lambda_I)$ and $G_M = (V_M, E_M, \omega_M, \lambda_M)$.For vertices X_I, Y_I of input graph G_I ($X_I, Y_I \subseteq V_I$) and vertices X_M, Y_M of model graph G_M ($X_M, Y_M \subseteq V_M$). In case the algorithm returns a mapping between nodes as $X_I \Leftrightarrow X_M$ and $Y_I \Leftrightarrow Y_M$. (*Where* $\Leftrightarrow$ *denotes a mapping from input graph* G_I *to model graph* G_M). We need to score the mapping based on (i) node matching, and (ii) corr-esponding edge matching. The edit cost for node matching is :

$$[\{\omega_I(X_I) - \omega_M(X_M)\} + \{\omega_I(Y_I) - \omega_M(Y_M)\}].$$

We consider three cases of edit costs for edge matching (Note*: node matching is common to all the three cases*):

Case 1: When there exists an edge $X_I \rightarrow Y_I$ in G_I and a corresponding edge $X_M \rightarrow Y_M$ in G_M. Then the edit cost is $\lambda_I(X_I, Y_I) - \lambda_M(X_M, Y_M)$.

Case 2: When there exists an edge $X_I \rightarrow Y_I$ in G_I and no corresponding edge between X_M, Y_M in G_M i.e. $(X_M, Y_M) \not\subset E_M$. Then edit cost is a constant penalty of 2.

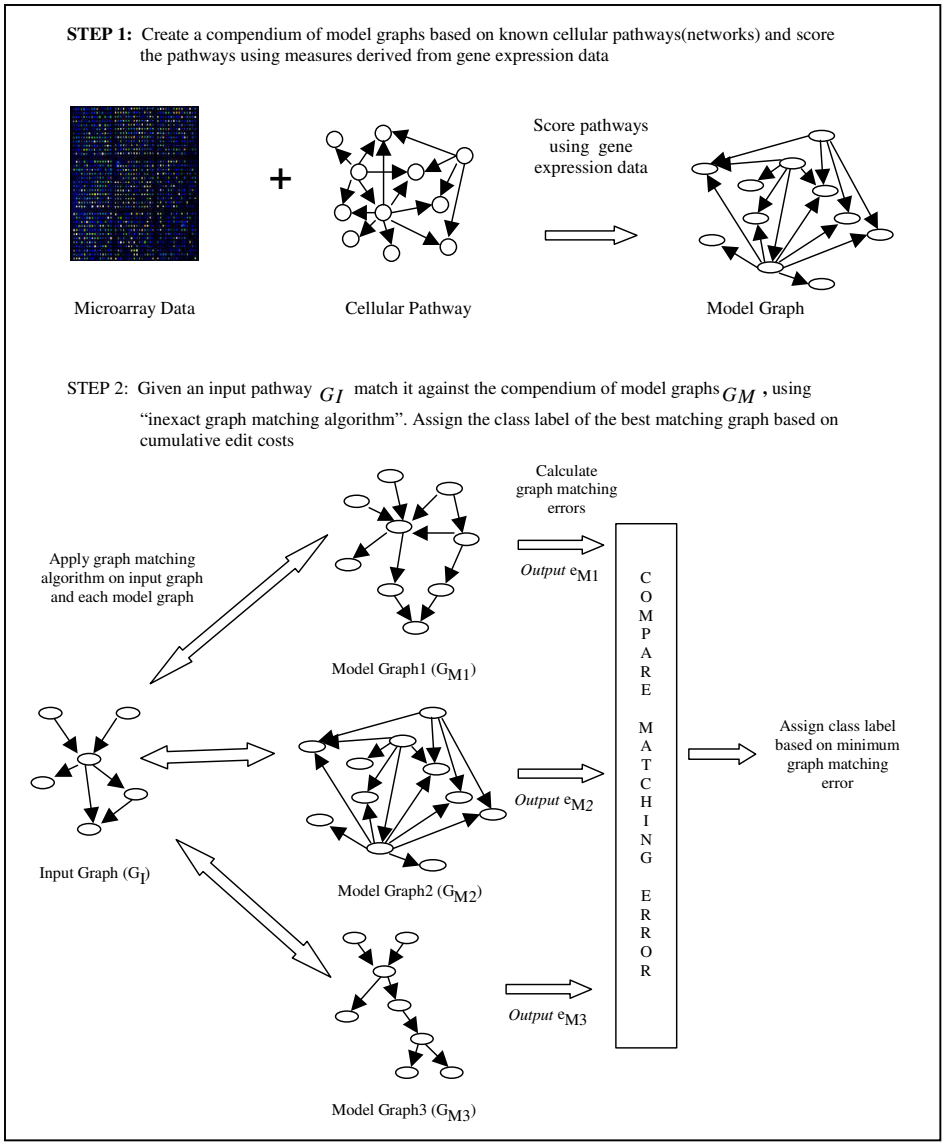

Fig.2: Illustration of the compendium creation process and classification strategy

Case 3: When for edge $X_I \rightarrow Y_I$ in G_I there exists an edge $Y_M \rightarrow X_M$ in G_M (i.e. reversed edge). Then edit cost is a constant penalty of 2.

3.2.3 Classification rule

Let $M = \{G_{M1}, G_{M2}, G_{M3}, \ldots, G_{Mn}\}$ be the set of model graphs in the compendium and $C = \{c_1, c_2, c_3 \ldots c_n\}$, be the set of class labels of model graphs. For the compendium we had created, all the model graphs were assigned unique labels, which is the ideal case but this definition may change depending upon the nature of analysis. Given an input graph G_I compare the graph iteratively to all the model graphs G_{Mj} and calculate a corresponding graph matching error score $e_{Mj}, \forall j = 1,2,3,\ldots,n$. The e_{Mj} is in turn a combination of two errors:

a) η_{error} = node matching error

b) ε_{error} = edge matching error (corresponding to the matched nodes).

These errors were further normalized to the scale (0-1) as described:

$$\eta_{error}(norm) = \eta_{error} / |V_I| \quad (6)$$

Where, $|V_I|$ =number of nodes in G_I

$$\varepsilon_{error}(norm) = \varepsilon_{error} / |E_I| \quad (7)$$

Where, $|E_I|$ =number of edges in G_I

Finally, graph-matching error eMj was calculated as:

$$e_{Mj} = 10 \times \eta_{error}(norm) + \varepsilon_{error}(norm) \quad (8)$$

Thus, according to equation (8), $0 \le e_{Mj} \le 11$ .The class label of graph G_I is assigned based on equation (9).

$$class(G_I) = class\{\arg min(e_{M1}, \ldots, e_{Mj}, \ldots e_{Mn})\} \quad (9)$$

4 Results and Discussions

We studied our approach from two broad perspectives. Following is a description of the results, which demonstrate:

(i) suitability of our approach to model the problem of classification.

(ii) assessment of the reliability of the error metric from statistical methods.

4.1 Influence of network size on classification

We wanted to ascertain whether the size of input network influences the performance of the algorithm. In order to obtain hints about the efficacy of our procedure, we created a set of networks (called test networks) with varying number of nodes (***n*** = 6,10,12). The networks were specifically created keeping in mind that they should be largely similar to one of the model graphs and partially similar to few other model graphs. This was done to ascertain the coherence of results. We created 3 test graphs with high degree of matching to "Galactose" model graph and partial matching to "Respiration" model graph. These test networks were iteratively compared to the model graphs, the result of the comparison is presented in table 2.

Model Graphs	Test graph 1	Test graph 2	Test graph 3
Cell cycle	10.8167	10.8065	No match
MAPK	10.8096	10.7879	No match
PKC	10.7514	10.8636	No Match
Galactose	2.5310	2.8814	2.7487
HOG	10.8388	10.8942	10.7874
Filament	10.8493	10.8636	10.9140
Respiration	5.9363	8.9352	6.7902

Table 2: Error table of the comparison of model graphs with input graphs of varying size (error is on scale of 0 -11)

From the data in table 2 it's evident that matching error is least for Galactose graph in each category, and the result for Respiration graph is the second best, which was expected according to the design of analysis. Also, the error rates in the case of Galactose graph are comparable in the three cases. Hence, we could ascertain that the result of the algorithm is independent of the size of the input graph. If there is a match from G_I to G_M then our approach is able to extract the same. The No-match entry in column 4 signifies that no valid mappings were extracted for the matching of test graph3 with the corresponding model graphs, which is consistent according to our observation. As the Test graph3 was largely different from Cell Cycle, MAPK and PKC model graphs.

4.2 Statistical significance of results based on error metric

As described earlier the pivot of our classification strategy is to find the minimum matching error, which in turn serves as a criterion for class assignment. We tried to ascertain that the matching errors generated by the algorithm were truly a reflection of the degree of match between two graphs and not the artifact of the overall procedure. Here we present the approach to measure statistical significance of the error metric. Figure 3 illustrates the strategy we deployed to access the statistical significance of the results based on error metric (function). We created an ensemble of random networks. The strategy used was, for each model graph (G_{M_j}) we first create a reference network ($G_{\Re_j}$), which is a slightly distorted sub network of G_{M_j}. This distortion was introduced by either adding few extraneous nodes and edges that were not originally present in G_{M_j}. This was done to ensure that we get an error measure > 0. Subsequently we create 15 random networks (G_{rand_i}) for i =1, 2, 3,…, 15 which have identical nodes as in $G_{\Re_j}$ but differ in topology and degree distributions. Thus we have a total of 7 reference networks ($G_{\Re_j}$) corresponding to 7 model graphs (G_{M_j}) and a total of 105 random networks (G_{rand_i}). Let we define matching error between $G_{\Re_j}$ and G_{M_j} as

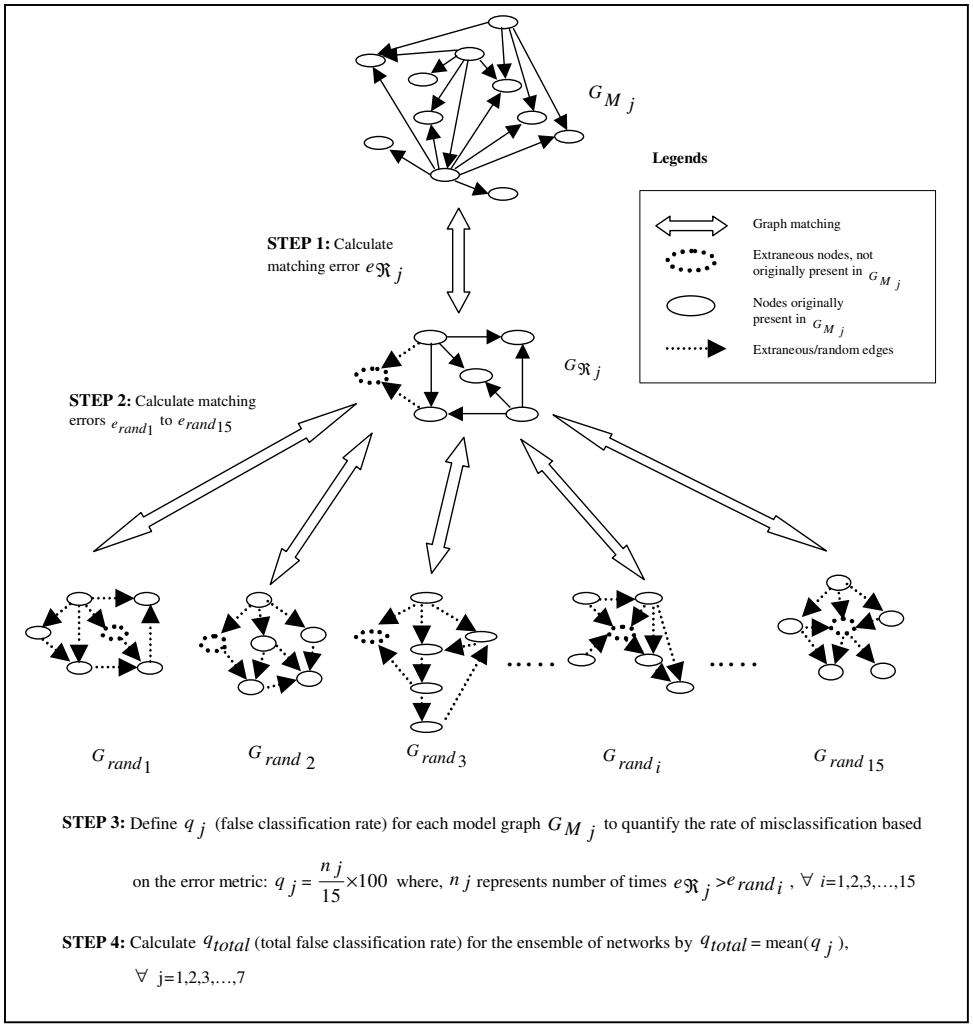

Fig.3: Illustration of approach taken to ascertain the statistical significance of the results based on the defined error metric

$^{e}\Re_j$ and the matching error between $G_{rand\,i}$ and $G_{M\,j}$ as $e_{rand\,i}$ $\forall i = 1,2,3,...,15$. So, we define a measure q_j (false classification rate) for each model graph $G_{M\,j}$ to quantify the misclassification based on the error metric:

$$q_j = \frac{n_j}{15} \times 100 \qquad (10)$$

Where, n_j represents number of times $^{e}\Re_j > e_{rand\,i}$, $\forall i = 1,2,3,...,15$

The total false classification rate for the ensemble of networks is calculated by:

$$q_{total} = mean\,(q_j),\ \forall j = 1,2,3,...,7 \qquad (11)$$

Table 3 shows the false classification rate for each of the seven model graphs. The q_{total} calculated by this method was 0.95%. This result illustrates that the error metric we have defined is able to distinguish between significant and non-significant graph matching in more than 99% of the cases.

Model Graphs	**False classification rate** (q_j)
Cell cycle	6.667
MAPK	0.00
PKC	0.00
Galactose	0.00
HOG	0.00
Filament	0.00
Respiration	0.00

Table 3: The false classification rate (%) for the compendium of pathways

5 Conclusion

We have demonstrated a new method for pathway classification, which integrates knowledge about known biological networks and gene expression data. While the basic principle of scoring pathways is similar to other approaches [11,12], we use the scored pathways to create a compendium. This compendium is used in conjunction with an augmented graph-theoretic matching algorithm to classify a putative pathway. A metric based on graph edit operations was defined to calculate the degree of similarity between two graphs (pathways). Though our choice of edit cost and the corresponding metric was heuristic, the preliminary/empirical results demonstrate that the algorithm is able to extract significant matching and accurately assigns the class/phenotype label for an uncharacterized pathway. We state that the results presented in this paper are not exhaustive, but they suggest that such integrated approaches are inevitable for gaining better insights into living systems. In this approach we defined the error metric as a linear function of two errors (distance measures) but other functional combination of these distance measures should be explored or if sufficiently many pathways are known in advance, this knowledge may be utilized to learn an appropriate error metric (function) by employing machine learning methods. The scoring schema was devised for immediate relationships between two entities but in the context of gene regulatory networks cascade effect on downstream genes can also be integrated. Currently this area of research is still in its infancy, we excluded the cascade effect from the current analysis. We also propose that this approach is generic and can be suitably modified to suit other applications. More information from other domains of biomedical research can be integrated in the scoring schema and application specific edit operations/costs can be defined. Specifically, knowledge about cis-regulatory elements and macromolecular (binding) interactions (protein-protein and protein-DNA) can be utilized to explore the model on a wider scale. We believe that our approach will benefit from advances in the field of functional genomics, network biology and continued breakthroughs in experimental and computational biology.

Acknowledgements

We are grateful to Rajendra K. Bera for his guidance and inputs during research. We thank Albee Jhoney for his constant support and motivation, which helped tremendously in completion of the project. We also thank Dinesh A Venkateswaran and Deepak K Gangadhar for providing valuable comments on the manuscript.

References:

[1] DeRisi, J.L. *et al.,* Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, Vol.278, 1997, pp.680-686.

[2] Golub, T.R. *et al.,* Molecular classification of cancer: class discovery and class prediction by gene expression modeling, *Science*, Vol.286, 1999, pp.531-537.

[3] Alizadeh, A.A. *et al.,* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, Vol.40, 2000, pp.503-511.

[4] Friedman, N., Inferring cellular networks using probabilistic graphical models, *Science*, Vol.303, 2004, pp.799-805.

[5] Tong, A.H. *et al.,* Global Mapping of the yeast genetic interaction network, *Science*, Vol.303, 2004, pp.808-813.

[6] Uetz, P. *et al.*, A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae, Nature*, Vol.403, 2000, pp.623–627.

[7] Milo, R. *et al.*, Network Motifs: Simple Building Blocks of complex networks, *Science*, Vol.298, 2002, pp.824-827.

[8] Stelling, J. *et al.*, Metabolic networks structure determines key aspects of functionality and regulation, *Nature*, Vol.420, 2002,pp.190-193.

[9] Kitano, H., System Biology: A brief overview, *Science*, Vol.295, 2002, pp.1662-1664.

[10] Forster, J. *et al.*, Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network, *Genome Research*, Vol.13, 2003, pp.244-253.

[11] Hanisch, D. *et al.,* Co-clustering of biological networks and gene expression data, *Bioinformatics*, Vol.18, 2002, pp.145s-154s.

[12] Zien, A. *et al.*, Analysis of gene expression data with pathway scores, In Altman, R. *et al.* (eds), *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology,* La Jolla, CA, 2000, pp.407-417. AAAI.

[13] Barabasi, A.L. and Oltvai, Z.N., Network Biology: Understanding the cell's functional organization, *Nature Review Genetics*, Vol.5, 2004, pp.101-113.

[14] Hughes, T.R. *et al.*, Functional Discovery via a compendium of expression profiles, *Cell*, Vol.102, 2000 pp.109-126.

[15] Simon, I. *et al.*, Serial regulation of transcriptional regulators in the yeast cell cycle, *Cell*, Vol.106, 2001, pp.697-708.

[16] Roberts, C.J. *et al.*, Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles, *Science*, Vol.287, 2000, pp.873-880.

[17] Ideker, T. *et al.*, Integrated genomic and proteomic analyses of a systematically perturbed metabolic network, *Science*, Vol.292, 2001, pp.929-934.

[18] Kanehisa, M. and Goto S., KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Research*, Vol.28, 2000, pp.27-30.

[19] Butte A.J. and Kohane, I.S., Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements, *Pac. Symp. Biocomputing*, 5, 2000, pp.415–426.

[20] Michaels, G.S. *et al.*, Cluster analysis and data visualization of large-scale gene expression data, *Pac. Symp. Biocomput.*, 3, 1998, pp.42–53.

[21] Hlaoui, A., and Wang, S., A new algorithm for inexact graph matching, *Proceedings of the 16th International Conference on Pattern Recognition*, August 11-15, 2002, pp.40180-40183.

[22] Ullman, J.R., An algorithm for subgraph isomorphism, *Journal of the ACM*, Vol.23, 1976, pp.31-42.

[23] Messmer, B. and Bunke, H., A New Algorithm for Error-Tolerant Subgraph Isomorphism Detection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.20, 1998, pp. 493-505, 1998.