Gene Interactions Sub-networks and Soft Computing

Ranajit Das and Sushmita Mitra

Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India
ranajit_r@isical.ac.in, sushmita@isical.ac.in

Abstract. Analysis of gene interaction networks is crucial for understanding fundamental cellular processes involving growth, development, hormone secretion and cellular communication. A gene interaction network comprises of proteins and genes binding to each other, and acting as a complex input-output system for controlling cellular functions. A small set of genes take part in a cellular process of interest, while a single gene may be involved in more than one cellular process at the same time. Soft computing is a consortium of methodologies that works synergistically and provides flexible information processing capability for handling real life ambiguous situations. The tools include fuzzy sets, evolutionary computing, neurocomputing, and their hybridizations. We discuss some existing literature pertaining to the use of soft computing and other classical methodologies in the reverse engineering of gene interaction networks. As a case study we describe here a soft computing based strategy for biclustering and the use of rank correlation, for extracting rank correlated gene interaction sub-networks from microarray data. Experimental results on time series gene expression data from *Yeast* were biologically validated based on standard databases and information from literature.

Keywords: Soft Computing, bioinformatics, multi-objective evolutionary biclustering, transcriptional regulatory network extraction, gene expression profile, rank correlation, gene interaction network.

1 Introduction

With the current development in microarray technology (gene chips), today researchers in Bioinformatics have, at their disposal, expression data of thousand of genes of different organisms under various experimental conditions. This had led to complete-genome expression profiling of several organisms. The latest Affymetrix gene chips contain 750,000 unique 25-mer oligonucleotide features constituting more than 28,000 mouse gene-level probe sets. This DNA microarray technology forms an indispensable tool for exploring transcriptional regulatory networks from the system level and is useful when one dwells into the cellular environment to investigate various complex interactions [1]. Biological networks connect genes, gene products (in the form of protein complexes) or their groups to one another. A network of co-regulated genes may form gene clusters that can encode proteins, which interact amongst themselves and take part in common biological processes. Clustering of gene expression profiles have been employed to identify co-expressed groups of genes [2] as well as to extract gene interaction/gene regulatory networks [3].

Sharing of the regulatory mechanism amongst genes, in an organism, is predominantly responsible for their co-expression. Genes with similar expression profiles are

A. Bargiela, W. Pedrycz (Eds.): Human-Centric Information Processing, SCI 182, pp. 313–327.
springerlink.com

very likely to be regulators of one another or be regulated by some other common parent gene [4]. Often, it is noted that during few conditions a small set of genes are co-regulated and co-expressed, their behavior being almost independent for rest of the conditions. The genes share local rather than global similar patterns in their gene expression profiles. Generally, group of genes are identified in the form of biclusters using continuous columns biclustering because biological processes start and terminate over a continuous interval of time [5, 6]. The aim of biclustering is to bring out such local structure inherent in the gene expression data matrix. It refers to the clustering of both rows (genes) and columns (conditions) of a data matrix (gene expression matrix), simultaneously, during knowledge discovery about local patterns from microarray data [7].

The genome, comprising the set of all genes in an organism along with their expressions values, is considered to be a switching network, with its vertices denoting the proteins or molecules and the directed edges representing their various interactions and inter-dependence. Such networks relate genes, gene products or their groups (like protein complexes or protein families) to each other. A directed edge (or arc) connects one node (or vertex) to another. Consider the graph depicted in Fig. 1. Mathematically a network can be expressed as a graph $G = \{V, E\}$, where V represents the set of N vertices $\{V_1, V_2, ..., V_N\}$ while E represents the set of edges that connect two elements in V.

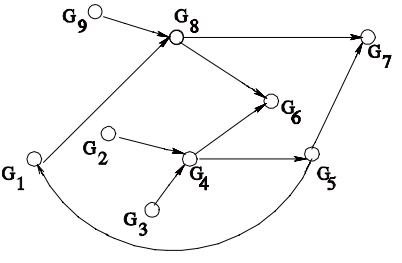

Fig. 1. A sample gene interaction network with nine nodes and ten edges

In this chapter we provide an overview on the extraction of gene interaction networks followed by a study involving a rank correlation-based multi-objective evolutionary technique for the extraction of simple gene interaction sub-networks from microarray data. Use of soft computing, with biclustering, is described in this connection. Preprocessing, involving the discretization of the rank correlation matrix (using quantile partitioning) and subsequent elimination of weak correlation links, is employed to retain strongly rank correlated (positive or negative) gene interaction pairs. Experimental results on *Yeast* data are validated in terms of a gene ontology (GO) study. The rest of the chapter is organized as follows. Section 2 introduces the basics of biological networks and gene interaction networks. Section 3 describes some classical reverse engineering approaches for generating gene interaction networks using time series gene expression data. In Section 4, the existing literature pertaining to the use of soft computing in the extraction of gene interaction networks is compiled. As a case study,

the use of multi-objective evolutionary biclustering and rank correlation for the extraction of Gene interaction sub-network is described in Section 5. The effectiveness of the discussed methodology is also demonstrated therein, using time-series gene expression data from *Yeast*. The article is concluded in Section 6.

2 Biological Networks

Biological pathways can be conveniently represented as networks and broadly classified as *metabolic pathways*, *signal transduction pathways* and *gene interaction networks*. The repository of information about various biological pathway data is available in some databases like BioCyc[1] [8], EcoCyc [9], What Is There (WIT) system[2], RegulonDB [10], *etc*. GO [11] and the KEGG Orthology [12] promote the use of controlled vocabulary to facilitate computational analysis. These databases can be integrated with various computational methods to get an insight into complex biological functions. They can help in (i) reconstructing biochemical pathways from the complete genome sequence, and (ii) predicting gene interaction networks. The proper understanding of gene interaction networks is essential for the understanding of fundamental cellular processes involving growth and decay, development, secretion of hormones, cellular communication, etc. During transcription of gene expression specific groups of genes may be made active by certain signals, which on activation, may regulate similar biological processes. The genes may also be regulators of each other's transcription.

The metabolic pathways facilitate mass generation, energy production, information transfer and cell-fate specification, in a cell or micro-organism; they are seamlessly integrated through a complex network of cellular constituents and reactions. Such a metabolic network consists of nodes, *i.e.*, substrates (genes or proteins), which are interconnected through links, *i.e.*, metabolic reactions in which enzymes provide the catalytic scaffolds [13].

Signal transduction is the process by which a cell converts one kind of signal or stimulus into another by a series of steps, causing functional changes inside the cell. The signal may pass from one cell to another (Hormone-Receptor concept), from extracellular environment to inside the cell (through plasma membrane) or from one compartment inside the cell to another compartment (*i.e.*, from cytoplasm to nucleus). A signal transduction pathway can be considered as a biological network of biomolecules connected by various kinds of interactions (protein-protein interactions, protein-ion interactions, *etc.*) among them.

Analyzing various types of messenger RNAs (mRNAs) produced by a cell and quantifying them, one can determine the gene or set of genes that get transcribed under particular experimental conditions. A cell dynamically responds to both environmental stimuli and its own changing requirements in a highly complicated and tightly regulated process. This process helps one to monitor the required increase or decrease of the expression levels of particular genes. The control and regulation of gene expression could be caused by various external factors, occurring at different stages of

[1] http://www.biocyc.org/

[2] http://wit.integratedgenomics.com/

the cellular information flow from DNA, RNA to protein, like in mRNA splicing, translational control and/or post-translational control. Nevertheless, the one involving the initiation of transcription has been most widely studied in literature [14, 15, 16].

A gene regulatory network (GRN) determines which subset of genes is expressed, up to what level, and in response to what conditions of the cellular environment. While the metabolic networks form the basis for the net accumulation of biomolecules in living organisms, the regulatory networks modulate their action – thereby leading to physiological and morphological changes. However, one should note that any apparent similarity of expression profiles between two genes may not always mean that they may regulate each other but may signify (i) indirect co-regulation by other genes, (ii) direct regulation of one gene by the other, or (iii) a mere coincidence involving no causal relationship. An integration of additional biologically relevant knowledge may, therefore, provide constraints on suitable identification of groups of co-regulated genes.

3 Reverse Engineering of Genetic Interaction Networks

Reconstruction of interactions in gene regulatory networks, from gene expression data, is termed reverse engineering. Some of the techniques, typically used for the purpose, include the generalized Bayesian networks [17, 18], Boolean networks [19, 20, 21, 22], linear and non-linear ordinary differential equations (ODEs) [14, 23, 24]. Boolean networks are binary models with genes taking on values one (or zero) to represent active (or inactive) states [22]. However these ignore the effect of genes at intermediate levels, and result in information loss during discretization. Bayesian networks are graph models that estimate complicated multivariate joint probability distributions through local probabilities [17]. Reverse engineering with Bayesian learning [18] enabled the generation of gene regulatory interactions from simulated gene expression data. Dynamic Bayesian networks (DBNs) were subsequently used for inferring the relationship amongst genes from time-series gene expression data [25, 26].

Gene regulatory relationships were extracted for cell cycle-regulated genes in *yeast*, with the activation or inhibition between gene pairs being represented as events [27]. Matching of corresponding events was followed by a sequence alignment of the event strings. Regulatory relationships have also been deduced from the correlation of co-expressions, between a DNA-binding transcription regulator and its target gene, by using a probabilistic expression model [28]. However, correlation matching alone is deemed unsuitable to effectively distinguish between regulators and target genes. It is also difficult to discern whether the correlated target is directly or indirectly regulated. Hence additional information like protein-DNA binding has been integrated into transcriptional regulatory networks [29] for validating direct regulator-target interaction.

Co-regulated genes are often functionally, *i.e.*, physically, spatially and/or genetically associated. In real life, however, the genes may be co-regulated only across a subset of all observed experimental conditions. In other words, a small number of genes participate in a cellular process of interest while a gene may be simultaneously active in more than one cellular process. It is here, where biclustering (or coclustering) becomes more appropriate than standard clustering, for the purpose of modeling regulatory pathways. Here we perform simultaneous clustering of both rows (genes)

and columns (conditions) of the gene expression matrix, for knowledge discovery in maximal subgroups of local patterns [30, 31]. An algorithm *cMonkey* has been developed [32], to detect putatively co-regulated gene groupings by integrating biclustering of gene expressions and various functional associations with the *de novo* detection of sequence motifs.

4 Role of Soft Computing

In addition to the combinatorial approach, soft computing is gradually opening up several possibilities by generating low-cost (computational cost both in terms of space and time complexity), low-precision (approximate), good solutions. Soft computing is a consortium of methodologies that works synergistically and provides flexible information processing capability for handling real life ambiguous situations [33]. The tools include fuzzy sets, evolutionary computing, neurocomputing, and their hybridizations. Typically, they require little a priori knowledge about the underlying system, and the model can be derived directly from the data. Since the work deals with huge amounts of incomplete or ambiguous data, (i) the uncertainty handling capacity of fuzzy sets, (ii) the learning ability of artificial neural networks (ANNs) to discover hidden regularities within the data, and (iii) the searching potential of evolutionary strategies (like genetic algorithms) to explore the large pattern space, are typically utilized [34].

The human mind expresses higher level of perceptions using vague, non-crisp concepts. So for developing really intelligent methods for approximate reasoning about similar concepts accessible for intelligent systems, languages need to be developed. One way out while searching for solutions to these tasks is the use of Granular Computing. Granular computing [35] (GC) is useful in finding meaningful patterns in data by expressing and processing chunks of information (granules). The solutions involving GC become feasible because they specify non-Boolean or non-crisp specifications to a satisfactory degree and can be, more often than not, efficiently constructed than those involving detailed, purely numeric solutions. GC may thus be informally defined as a general computing theory for effectively using granules in the form of classes, clusters, subsets or groups, *etc.* and intervals for developing efficient computational models for complex applications involving huge amount of data, information and knowledge [36].

A problem that we conceive of is generally cast into frameworks, which facilitate the observations about clusters of objects with some commonality and eventually lead to the effective formulation of the problem and its solution with considerable acuity [35]. Such frameworks are ideal for problems involving pattern recognition, feature selection and reduction, knowledge discovery and bioinformatics. Identification of relevant features of objects contained in information granules help us to formulate hypotheses about the significance of the objects, construct new granules and refine the information, use GC to measure the distance among complex granules, *etc.* GC brings together the existing formalisms of set theory, fuzzy sets, and rough sets under a common platform by clearly visualizing some fundamental similarities and synergies.

The modeling of imprecise and qualitative knowledge, as well as the transmission and handling of uncertainty at various stages are possible through the use of fuzzy

sets. Fuzzy logic is capable of supporting, to a reasonable extent, human type reasoning in natural form. Fuzzy Adaptive Resonance Theory (FART) associated matrix method has been developed [37] to cluster gene expression profiles of *Saccharomyces cerevisiae* (yeast) responding under oxidative stresses, followed by the extraction of genetic networks from them. The inferred genetic interactions are quantitatively evaluated, and validated in terms of the KEGG metabolic map, BRITE[3] protein interaction map and related literature. The number of clusters is controlled by the vigilance parameter of FART. Fuzzy rules of an activator-repressor model of gene interactions were used [38] to transform expression values into qualitative descriptors. A new multiscale fuzzy c-means clustering method was designed to model gene interactions between regulatory pathways, across different conditions and at different levels of detail [39].

The adaptivity of artificial neural networks (ANNs) to learn from data-rich environments and their robustness to noise make them good candidates for modeling genetic interactions from gene expressions. Some such connectionist models employed for extracting genetic regulatory effects include perceptrons [40, 41], self-organizing maps [42, 43], and recurrent neural networks (RNNs) [44, 45]. The RNN was used to model the dynamics of gene expression in the *lambda phage*[4] regulatory system [44].

Use of genetic algorithm (GAs) for reconstructing genetic networks has been reported in literature [46, 47]. The mutation and crossover operators help to intelligently guide the GA in the complex search space. Typically the GA searches for the most likely genetic networks that best fit the data, considering the set of genes to be included in the network along with the strength of their interactions. Gene interaction networks were inferred from microarray data [48], using GAs for interactive reverse engineering. However the combinatorial complexity is expected to be unmanageable in real-world problems, involving a large number of genes [49].

Hybrid techniques like neuro-fuzzy computing have found applications in the realm of genetic networks as well. ANNs and fuzzy logic have been employed to form a framework for inferring gene interaction networks. Knowledge-based neural networks, which incorporated prior knowledge about gene interactions, were used by Kasabov [50] for the reverse engineering of genetic networks. A hybrid methodology for this purpose has been developed [51] by combining ANN, fuzzy sets and multi-objective GAs.

5 Extraction of Gene Interaction Network: A Multi-objective Evolutionary Approach

Biological networks involving gene pairs, which demonstrate transcription factor (TF)-target relationship, is an important research problem. A gene interaction network is a complex structure comprising various gene products activating or repressing other gene products. A gene that regulates other genes is termed the transcription factor,

[3] KEGG BRITE Database is a collection of hierarchical classifications representing knowledge on various aspects of biological systems. http://www.genome.jp/kegg/brite.html

[4] Enterobacteria phage λ (lambda phage) is a temperate bacteriophage that infects the bacteria *Escherichia coli.*

while the gene being regulated is called its target. The presence of a TF, can alternatively switch "ON" some genes in the network while others remain "OFF", orchestrating many genes simultaneously. The proper understanding of gene interaction networks is essential for the understanding of fundamental cellular processes involving growth and decay, development, secretion of hormones, *etc.* During transcription of gene expression specific groups of genes may be made active by certain signals, which on activation may regulate similar biological processes. The genes may also be regulators of each other's transcription. Target genes sharing common TFs demonstrate similar gene expression patterns along time [14, 52]. Analysis of similar expression profiles brings out several complex relationships between co-regulated gene pairs, including co-expression, time shifted, and inverted relationships [53].

We describe a methodology for modeling the relationship between a transcription factor and its target's expression level variation over time in the framework of the generated biclusters. The extraction of the relationship between the gene pair is biologically more meaningful and computationally less expensive as a bicluster is a subset of highly correlated genes and conditions. Rank correlation provides a similarity measure, which retains the relevant information necessary for computing pairwise correlation between gene pairs. The relationship is presented in terms of rules, where a TF is connected to its regulated target gene. These rules are subsequently mapped to generate parts of the entire regulatory network. It may be noted that intra-pathway gene interactions, responsible for a particular biological function and possibly within a bicluster, are generally stronger than any inter-pathway interactions.

The goal in genetic networks is to identify possible direct excitatory and/or inhibitory connections between genes, gene products and proteins, when the time-steps are close enough. Otherwise, indirect connections, through a third gene, needs to be established. Sometimes additional biological knowledge, such as gene ontology[5] and transcription factors, is included.

Most real-world search and optimization problems typically involve multiple objectives. A solution that is better with respect to one objective requires a compromise in other objectives. In problems with more than one conflicting objective there exists no single optimum solution. Rather, there exists a set of solutions, which are all optimal involving trade-offs between conflicting objectives. Unlike single-objective optimization problems, the multi-objective evolutionary algorithms (MOEA) tries to optimize two or more conflicting characteristics represented by fitness functions. Modeling this situation with single-objective GA would amount to heuristic determination of a number of parameters involved in expressing such a scalar-combination-type fitness function. MOEA, on the other hand, generates a set of Pareto-optimal solutions, which simultaneously optimize the conflicting requirements of the multiple fitness functions. Among the different multi-objective algorithms, it is observed that non-dominated sorting genetic algorithm (NSGA-II) possesses all the features required for a good MOEA. It has been shown that this can converge to the global Pareto front, while simultaneously maintaining the diversity of population. More details on the characteristics of NSGA-II, like non-domination, crowding distance and crowding selection operator can be found in [54].

[5] A shared, controlled vocabulary that is being developed to cover all organisms, in terms of molecular function, biological process and cellular component. http://www.geneontology.org

Biclustering refers to the simultaneous clustering and redundant feature reduction involving both attributes and samples. This results in the extraction of biologically more meaningful, less sparse partitions from high-dimensional data, and exhibit similar characteristics. The partitions are known as biclusters. Biclustering has been applied to gene expressions from cancerous tissues [31], mainly for identifying co-regulated genes, gene functional annotation, and sample classification. A bicluster can be defined as a pair (g, c), where $g \subseteq \{1,..., m\}$ represents a subset of genes and $c \subseteq \{1,..., n\}$ represents a subset of conditions (or time points). The optimization task [30] involves finding the maximum-sized bicluster not exceeding a certain homogeneity constraint mentioned below. The size (or volume) f(g, c) of a bicluster is defined as the number of cells in the gene expression matrix E (with values e_{ij}) that are covered by it. The homogeneity $G(g, c)$ is expressed as a mean squared residue score. More details on the biclustering scheme can be obtained in [54].

The Multi-objective GA (NSGA II), in association with the local search procedure discussed in [54], was used for the generation of the set of biclusters. The algorithm followed is discussed in details in [54]. The maximal set of genes and conditions representing size were generated keeping the "homogeneity" criteria of the biclusters intact. Since these two characteristics of biclusters are conflicting to each other, multi-objective optimization was employed to model them. To optimize this conflicting pair, the fitness function f_1 (corresponding to size) is always maximized while function f_2 (reflecting ratio of means square residual error and the threshold) is maximized as long as the residue is below the threshold, δ.

Like GC biclustering also contains some condensed information pertaining to correlation/co-regulation among subset(s) of genes. So, this helps in the extraction of gene interaction sub-networks, which appear to be more understandable to the human end-user.

5.1 Correlation between Gene Pairs

In this section we demonstrate the efficacy of a rank correlation-based approach for the extraction of gene interaction networks. A small number of genes participate in a cellular process of interest, being expressed over few conditions. Co-regulated genes are often found to have similar patterns in their gene expression profiles locally, rather than globally. The genes share similar sub-profiles, over a few time points, instead of the complete gene expression profiles. Thus, considering the global correlation amongst genes, *i.e.*, computation of correlation amongst genes employing the complete gene expression data matrix, would not reveal proper relationship between two of them. The *Spearman rank correlation* provides such a local similarity measure between the two time-series curves, since it is shape-based. The expression profile e of a gene may be represented over a series of n time points. Since the genes in a bicluster are co-expressed, the concept of correlation has been used to quantify their similarity. Instead of the commonly used similarity measures like the Euclidean distance or the Pearson correlation the *Spearman rank correlation* (RC) have been employed due to its robustness towards outliers and measurement errors [55, 56].

Moreover, RC does not assume a Gaussian distribution of points. $RC(e_1, e_2)$ between gene expression profile pair e_1 and e_2 provides a shape-based similarity measure between the two time-series curves, sampled at e_{1i} and e_{2i} over n time intervals. This is expressed as

$$\mathrm{RC}(\mathrm{e}_1, \mathrm{e}_2) = 1 - \frac{6}{\mathrm{n}(\mathrm{n}^2 - 1)} \sum_{\mathrm{i}} [\mathrm{r}_{\mathrm{e1}}(\mathrm{e}_{1\mathrm{i}}) - \mathrm{r}_{\mathrm{e2}}(\mathrm{e}_{2\mathrm{i}})]^2, \quad (1)$$

where $r_{e1}(e_{1i})$ is the rank of e_{1i}. Here an extended version of the RC has been used which takes into account the resolving of ties, *i.e.,* $e_{1j} = e_{1i}$ for $i \neq j$. The RC satisfies $-1 \leq RC(e_1, e_2) \leq 1$ for all e_1, e_2.

The first preprocessing step is to filter correlation coefficients, which contribute minimally towards regulation. This is because often an exhaustive search of the possible interactions between genes is intractable. Next those coefficients are selected whose absolute values are above a detection threshold, suggesting greater correlation amongst the gene pairs. In this way we focus on a few highly connected genes that possibly link the remaining sparsely connected genes. The correlation range $[RC_{\max}, RC_{\min}]$ is divided into three partitions each, using *quantiles* [57] so that the influence of noise is lessened. Only strong and positive (negative) interactions are selected. Thereafter, a network connecting the various genes is generated.

5.2 The Algorithm

The main steps of the procedure are outlined as follows:

I) Extraction of biclusters by multi-objective genetic algorithm.
II) Determination of pairwise rank correlation between gene pairs.
III) Discretization of the correlation matrix for eliminating the weaker interactions.
IV) Network generation from connectivity matrix (Section 5.3.1)
V) Biological validation (as discussed in Section 5.3.2).

5.3 Experimental Results

Data from the budding yeast *S.cerevisiae* is employed for extracting the gene interaction sub-networks.

5.3.1 Network Extraction

Yeast cell-cycle CDC28 data [58] is a collection of 6178 genes (attributes) for 17 conditions (time points), taken at 10-minute time intervals covering nearly two cycles. The synchronization of the yeast cell cultures was done using the so-called CDC28 arrest. The experiments were performed using Affymetrix oligonucleotide array. The missing values present in the data set were imputed according to the methodology

provided in [59][6]. At first pairwise rank correlation coefficients between gene pairs are computed by eqn. 1 to generate the network architecture from the extracted biclusters. Quantile partitioning is employed next, to choose the strong positive as well as negative correlation links. In this way, the top $\frac{1}{3}$ of the positive and negative links is chosen to be connected in a network. A sample network consisting of three biclusters of sizes 7, 10, and 14, respectively, are shown in Fig. 2. A transcription factor is connected to its target gene by an arrow when such a TF-Target pair is found to exist within any of the biclusters. Gene pairs connected by solid lines depict positive correlation, while those connected by dashed lines are negatively correlated. TFs external to the network, but having targets within the network, are connected to their corresponding targets by dotted arrows. As an example, the TF YHR084W (encircled with solid lines) is a member of the network of 10 genes and has targets in all the three networks. An external TF YJL056C (encircled with dotted lines) has targets in networks of 7 and 10 genes. The biclusters were biologically validated from gene ontology study, based on the statistically significant GO annotation database[7].

5.3.2 Biological Validation

During the prediction of regulatory networks [60] the genes YHR084W and YLR351C were reported to form a TF-Target pair. We also obtained the summary of the TF-Target pair YHR084W-YLR351C (Fig. 2) in terms of *Molecular Function, Biological Process and Cellular Component* from the *Saccharomyces Genome Database* (SGD)[8]. From our calculations we have also confirmed that an interaction exists between the target and its TF. It is reported in the database that the biological process involving protein YLR351C is not fully understood as yet and YHR084W has transcription factor activity. It becomes more difficult when one attempts to extract some biologically meaningful information involving these two entities. From such scanty information our method has been able to identify that there exists a link between a TF and its target. From their cellular components we model, as an efficacy of the biclustering, the transcription of YLR351C by YHR084W occurring inside the nucleus, and then the regular translation mechanism follows. In likewise manner for the TF-Target pair of YPL075W and YJR045C (Fig. 2) reported in [59], we obtained their summary from SGD and found YPL075W to be transcriptional activator of genes involved in glycolysis while YJR045C has *ATPase*, enzyme regulator and protein transporter activity. Again we were able to predict that YPL075W is involved in the transcription of YJR045C and would go into the glycolysis process.

One can arrive at similar kind of conclusions, for the rest TF-Target pairs, with a certain definite degree of confidence. As relevant literature in this area are really very sparse a large number negative results is only expected. Our algorithm has not yet detected any false positive or false negative TF-Target pairs, which is consistent with the information available either in the literature or in the databases.

---

[6] LSimpute: accurate estimation of missing values in microarray data with least squares methods.

[7] http://db.yeastgenome.org/cgi-bin/GO/goTermFinder

[8] A scientific database of the molecular biology and genetics of the yeast *Saccharomyces cerevisiae* - http://db.yeastgenome.org/

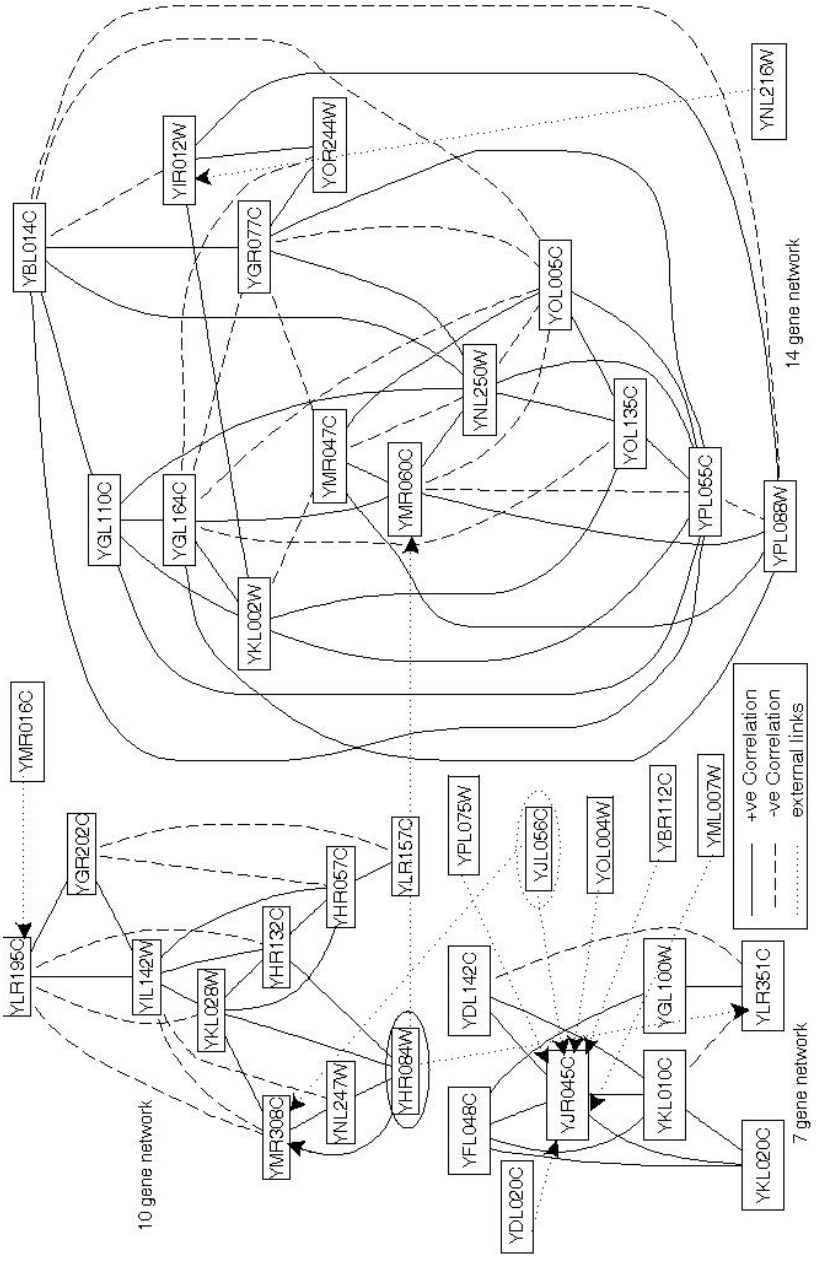

Fig. 2. Network (bicluster) of 10 genes connected by transcription factor YHR084W to networks (biclusters) of 7 and 14 genes

6 Conclusions and Discussion

In this chapter we have described the extraction of gene interaction networks. This was followed by a soft-computing approach to reverse engineering. Multi-objective evolutionary biclustering selected the co-regulated partitions. Subsequently, rank correlated gene pairs were extracted as a part of the gene interaction subnetworks.

Biologically relevant small biclusters were obtained, using time-series gene expression data from *Yeast*. These were validated using the statistically significant GO annotation database. The pairwise rank correlation coefficients among gene pairs were computed by eqn. 1 followed by the quantile partitioning to select the strong positive as well as negative correlation links. The strongly correlated genes were then chosen to be connected in a network. The TF-Target gene pairs in the network, shown in Fig. 2, were found to exhibit strong correlations. We tried to model the interaction among them from information available in the literature/databases *viz.*, SGD. We have also analyzed the expression profiles of the regulator and the regulated genes, which revealed several complex (time shifted, inverted, and simultaneous, *etc.*) relationships between them. The sparse nature of gene regulatory networks was reflected well on choosing Spearman rank correlation as the similarity measure.

References

[1] Mitra, S., Pedrycz, W. (eds.): Special Issue on Bioinformatics. Pattern Recognition 39 (2006)
[2] Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of National Academy of Sciences USA 95, 14863–14868 (1998)
[3] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M.: Systematic determination of genetic network architecture. Nature Genetics 22, 281–285 (1999)
[4] Gasch, A.P., Eisen, M.B.: Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. Genome Biology 3:research 0059.1-0059.22 (2002)
[5] Ji, L., Tan, K.L.: Identifying time-lagged gene clusters using gene expression data. Bioinformatics 21, 509–516 (2005)
[6] Madeira, S.C., Oliveira, A.L.: A Linear Time Biclustering Algorithm for Time Series Gene Expression Data. In: Casadio, R., Myers, G. (eds.) WABI 2005. LNCS (LNBI), vol. 3692, pp. 39–52. Springer, Heidelberg (2005)
[7] Cheng, Y., Church, G.M.: Biclustering of gene expression data. In: Proceedings of ISMB 2000, pp. 93–103 (2000)
[8] Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N., Kunin, V., Lopez-Bigas, N.: Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Research 19, 6083–6089 (2005)
[9] Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M., Karp, P.D.: EcoCyc: A comprehensive database resource for Escherichia coli. Nucleic Acids Research 33, 334–337 (2005)
[10] Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J., Martinez-Antonio, A., Collado-Vides, J.: RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Research 34, D394–D397 (2006)

[11] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology, the gene ontology consortium. Nature Genetics 25, 25–29 (2000)

[12] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M.: KEGG: Kyoto Encyclopedia of Genes Genomes. Nucleic Acids Research 27, 29–34 (1999)

[13] Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.-L.: The large scale organization of metabolic networks. Nature 407, 651–654 (2000)

[14] de Jong, H.: Modeling and simulation of genetic regulatory systems: A literature review. Journal of Computational Biology 9, 67–103 (2002)

[15] D'haeseleer, P., Liang, S., Somogyi, R.: Genetic network inference: From co-expression clustering to reverse engineering. Bioinformatics 16, 707–726 (2000)

[16] Thieffry, D., Huerta, A.M., P´erez-Rueda, E., Collado-Vides, J.: From specific gene regulation to genomic networks: A global analysis of transcriptional regulation in Escherichia coli. BioEssays 20, 433–440 (1998)

[17] Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. Journal of Computational Biology 7, 601–620 (2000)

[18] Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics 19, 2271–2282 (2003)

[19] Akutsu, T., Miyano, S., Kuhara, S.: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: Proceedings of Pacific Symposium on Biocomputing, pp. 17–28 (1999)

[20] Liang, S., Somogyi, F.S.: Somogyi Reveal: a general reverse engineering algorithm for inference of genetic network architectures. In: Proceedings of Pacific Symposium on Biocomputing, pp. 18–29 (1998)

[21] Martin, S., Zhang, Z., Martino, A., Faulon, J.-L.: Boolean dynamics of genetic regulatory networks inferred from microarray time series data. Bioinformatics 23, 866–874 (2007)

[22] Shmulevich, I., Dougherty, E.R., Kim, S., Zhang, W.: Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 18, 261–274 (2002)

[23] Gardner, T.S., di Bernardo, D., Lorenz, D., Collins, J.J.: Inferring genetic networks and identifying compound mode of action via expression profiling. Science 4, 102–105 (2003)

[24] de Jong, H., Page, M.: Search for steady states of piecewise-linear differential equation models of genetic regulatory networks. IEEE Transactions on Computational Biology and Bioinformatics 5, 208–222 (2008)

[25] Zou, M., Conzen, S.D.: A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21, 71–79 (2005)

[26] Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., Jarvis, E.D.: Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics 20, 3594–3603 (2004)

[27] Kwon, A.T., Hoos, H.H., Ng, R.: Inference of transcriptional regulation relationships from gene expression data. Bioinformatics 19, 905–912 (2003)

[28] Segal, E., Taskar, B., Gasch, A., Friedman, N., Koller, D.: Rich probabilistic models for gene expression. Bioinformatics 17, S243–S252 (2001)

[29] Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B.: E Fraenkel, Jaahhola TS, Young RA, Gifford DK Computational discovery of gene modules and regulatory networks. Nat. Biotechnol. 21, 1337–1342 (2003)
[30] Cheng, Y., Church, G.M.: Biclustering of gene expression data. In: Proceedings of ISMB 2000, pp. 93–103 (2000)
[31] Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: A survey. IEEE Transactions on Computational Biology and Bioinformatics 1, 24–45 (2004)
[32] Reiss, D.J., Baliga, N.S., Bonneau, R.: Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinformatics 7, 280 (2006)
[33] Zadeh, L.A.: Fuzzy logic, neural networks, and soft computing. Communications of the ACM 37, 77–84 (1994)
[34] Mitra, S., Acharya, T.: Data Mining: Multimedia, Soft Computing, and Bioinformatics. John Wiley, New York (2003)
[35] Pedrycz, W., Skowron, A., Kreinovich, V.: Handbook of Granular Computing. Wiley, England (2008)
[36] Bargiela, A., Pedrycz, W.: Toward a Theory of Granular Computing for Human-Centered Information Processing. IEEE Transactions on Fuzzy Systems 16, 320–330 (2008)
[37] Takahashi, H., Tomida, S., Kobayashi, T., Honda, H.: Inference of common genetic network using fuzzy adaptive resonance theory associated matrix method. Journal of Bioscience and Bioengineering 96, 154–160 (2003)
[38] Woolf, P.J., Wang, Y.: A fuzzy logic approach to analyzing gene expression data. Physiol Genomics 3, 9–15 (2000)
[39] Du, P., Gong, J., Wurtele, E.S., Dickerson, J.A.: Modeling gene expression networks using fuzzy logic. IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics 35, 1351–1359 (2005)
[40] Kim, S., Dougherty, E.R., Chen, Y., Sivakumar, K., Meltzer, P., Trent, J.M., Bittner, M.: Multivariate Measurement of Gene Expression Relationships. Genomics 67, 201–209 (2000)
[41] Huang, J., Shimizu, H., Shioya, S.: Clustering gene expression pattern and extracting relationship in gene network based on artificial neural networks. J Biosci. Bioeng. 96, 421–428 (2003)
[42] Ressom, H., Wang, D., Natarajan, P.: Clustering gene expression data using adaptive double self-organizing map. Physiol. Genomics 14, 35–46 (2003)
[43] Toronen, P., Kolehmainen, M., Wong, G., Castren, E.: Analysis of gene expression data using self-organizing maps. FEBS Lett. 451, 142–146 (1999)
[44] Vohradsky, J.: Neural network model of gene expression. FASEB Journal 15, 846–854 (2001)
[45] Weaver, D.C., Workman, C.T., Stormo, G.D.: Modelling regulatory networks with weight matrices. In: Proceedings of Pacific Symposium on Biocomputing, pp. 112–123 (1999)
[46] Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., Tomita, M.: Dynamic modeling of genetic networks using genetic algorithm and S-system. Bioinformatics 19, 643–650 (2003)
[47] Xiong, M., Li, J., Fang, X.: Identification of genetic networks. Genetics 166, 1037–1052 (2004)
[48] Iba, H., Mimura, A.: Inference of a gene regulatory network by means of interactive evolutionary computing. Information Science 145, 225–236 (2002)

[49] Keedwell, E., Narayanan, A.: Discovering gene networks with a neural-genetic hybrid. IEEE Transactions on Computational Biology and Bioinformatics 2, 231–242 (2005)
[50] Kasabov, N.K.: Knowledge-based neural networks for gene expression data analysis modelling and profile discovery. Biosilico 2, 253–261 (2004)
[51] Cotik, V., Zaliz, R.R., Zwir, I.: A hybrid promoter analysis methodology for prokaryotic genomes. Fuzzy Sets and Systems 152, 83–102 (2005)
[52] Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernardo, D.: How to infer gene networks from expression profiles. Molecular Systems Biology 3, 1–10 (2007)
[53] Zhang, Y., Zha, H., Chu, C.H.: A time-series biclustering algorithm for revealing co-regulated genes. In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC 2005), pp. 1–6 (2005)
[54] Mitra, S., Banka, H.: Multi-objective evolutionary biclustering of gene expression data. Pattern Recognition 39, 2464–2477 (2006)
[55] Balasubramaniyan, R., Hllermeier, E., Weskamp, N., Kamper, J.: Clustering of gene expression data using a local shape-based similarity measure. Bioinformatics 21, 1069–1077 (2005)
[56] Das, R., Mitra, S., Banka, H., Mukhopadhyay, S.: Evolutionary biclustering with correlation for gene interaction networks. In: Ghosh, A., De, R.K., Pal, S.K. (eds.) PReMI 2007. LNCS, vol. 4815, pp. 416–424. Springer, Heidelberg (2007)
[57] Davies, G.R., Yoder, D.: Business Statistics. John Wiley & Sons, Inc., London (1937)
[58] Cho, R.J., Campbell, M.J., Winzeler, L.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W.: A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell 2, 65–73 (1998)
[59] Bo, T.H., Dysvik, B., Jonassen, I.: Lsimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Research 32, 1–8 (2004)
[60] Qian, J., Lin, J., Luscombe, N.M., Yu, H., Gerstein, M.: Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics 19, 1917–1926 (2003)