International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:7, 2007

A Phenomic Algorithm for Reconstruction of Gene Networks

Rio G. L. D'Souza, K. Chandra Sekaran, and A. Kandasamy

***Abstract*—**The goal of Gene Expression Analysis is to understand the processes that underlie the regulatory networks and pathways controlling inter-cellular and intra-cellular activities. In recent times microarray datasets are extensively used for this purpose. The scope of such analysis has broadened in recent times towards reconstruction of gene networks and other holistic approaches of Systems Biology. Evolutionary methods are proving to be successful in such problems and a number of such methods have been proposed. However all these methods are based on processing of genotypic information. Towards this end, there is a need to develop evolutionary methods that address phenotypic interactions together with genotypic interactions. We present a novel evolutionary approach, called Phenomic algorithm, wherein the focus is on phenotypic interaction. We use the expression profiles of genes to model the interactions between them at the phenotypic level. We apply this algorithm to the yeast sporulation dataset and show that the algorithm can identify gene networks with relative ease.

***Keywords*—**Evolutionary computing, Gene expression analysis, Gene networks, Microarray data analysis, Phenomic algorithms.

I. Introduction

MICROARRAY datasets are obtained through carefully planned experiments conducted by biologists on microarray slides. These datasets typically have a small number of records but a large number of attributes. This situation is opposite to that of traditional datasets where the number of records is usually large, whereas the attributes are relatively few. Moreover dimensionality reduction techniques are not very effective because they affect the final results to a very large extent. Hence innovative methods are adopted to determine useful patterns from of the dataset [1].

Since microarray datasets are unlike traditional datasets, traditional data mining techniques do not work here. Clustering is performed to identify genes and subsets of genes which can be used to differentiate two or more cell-types, stages in disease development, or sub-types of a disease.

Gene networks represent relationships between genes, based on observations of how the expression level of each gene affects the expression levels of the others [2]. In order to reverse engineer and elucidate such relationships from gene expression measurements, it is necessary to compare expression patterns of genes. The challenge is to carry out the huge number of comparisons in a coordinated manner such that all relationships of interest are discovered. Evolutionary methods have been used for such a reconstruction of gene networks with some success by others [3].

Gene expression data could be used to model interaction of genes and hence could overcome the problem of objective function characterization which plagues evolutionary algorithms. Is it possible to use the gene expression patterns to model the interaction between individuals? In this research work, we show that it is possible and this leads to a method of assigning fitness to individuals without requiring an explicit objective function.

The rest of this paper is organized as follows: In Section 2 the related work done by others is reviewed. In Section 3 the biological background of the problem is explained. In Section 4 the computational background of the problem is explained. Section 5 is devoted to a discussion about the methodology adopted by the phenomic algorithm. The problem is precisely stated and the details of methods and techniques to be adopted are explained. Section 6 presents the results and discussion. Section 7 concludes the paper and is followed by a list of references.

II. Related Work

Contemporary gene expression analysis research encompasses work done during the last ten years to identify and analyze gene expression patterns. Before that gene expression analysis was technically limited to a handful of genes per study. The methods used then included Northern blot and real-time PCR, which were limited in number of genes, but individual measurements were fairly accurate [4].

Several high-throughput technologies were developed to overcome the limitations of traditional methods. These high-throughput methods allow more genes to be studied (typically in thousands) but the measurement of each gene is usually less accurate than those resulting from traditional methods. However this is a small price to pay since the information gained from measuring the expression of thousands of genes simultaneously is considered significant [5].

The microarray revolution was kicked off by Schena and

Rio G. L. D'Souza is with the Department of Computer Science and Engg, St Joseph Engineering College, Vamanjoor Post, Mangalore 575028, India (phone: 91-824-2263753, fax: 91-824-2263751, mobile: 91-9449470561, e-mail: rio@ieee.org)

K. Chandra Sekaran is with the Department of Computer Science and Engg, NITK - Surathkal, Srinivasanagar Post, Mangalore, India (e-mail: kchandrain@yahoo.co.in.).

A. Kandasamy is with the Department of Mathematical and Computational Sciences, NITK - Surathkal, Srinivasanagar Post, Mangalore, India (e-mail: kandy@nitk.ac.in).

other researchers [6] through their seminal paper on DNA microarrays. Since then researchers have tried out both traditional as well as non-traditional approaches to analyze microarray data. Among traditional approaches are the statistical approaches followed by Kuo [7], Troyanskaya [8], Baggerly [9], and Didier [10]. Most of these were univariate analyses based on calculation of p values and ANOVA. Phillips [11] stressed the need for multivariate analyses in such complex domains.

Machine learning techniques have been used to perform unsupervised as well as supervised learning from microarray data. The use of hierarchical clustering, self-organizing maps, and multidimensional scaling can be found in Ramaswamy [12], Nikkila [13], Fuller [14] and Eisen [15]. Such unsupervised learning techniques are of limited value since the sample size in microarray data is usually small. Supervised learning techniques are increasingly employed, as seen in the use of support vector machines [16] and artificial neural networks [17]. The problem of analyzing temporal gene expression data was first introduced, when a tutorial was presented in the Pacific Symposium on Biocomputing by D'haesseleer, Liang and Somogyi [3].

As an indication of the current trend, several researchers have compared rank-based gene selection methods with genetic algorithms. An effective parallel genetic algorithm has been presented by Liu, Iba, and Ishizuka [18]. Though they prove the efficacy of genetic algorithms in finding optimal gene sets, their method can only be applied to few classes. An extension of their work in terms of better fitness functions and multi-class classification is worth pursuing. Another work by Ooi and Tan [19] addresses the problem of multi-class prediction by employing an elaborate genetic algorithm combined with a maximum likelihood classifier.

Deutsch [20] has presented a work wherein a novel means of reducing the number of high-ranked genes is explored. It uses an evolutionary algorithm known as replication algorithm which is used in quantum simulations and protein folding. The combination of genetic algorithms and artificial neural network models can result in powerful solutions. Creighton and Hanash [21] have used data mining to discover association rules in gene expression datasets. Though the method used by them is based on the classical A-priori algorithm, it nevertheless could be an ideal application for the use of genetic algorithms.

Recently interest has developed in reverse engineering a gene network from its activity profiles. In a first attempt, a simple method was introduced that showed that reverse engineering is possible in principle [22]. A more systematic and general algorithm was developed by Liang et al. [23], using mutual information to identify a minimal set of inputs that uniquely define the output for each gene at the next time step. Akutsu et al. [24], [25] proposed simplified reverse engineering algorithms and rigorously examined the input data requirements.

D'haeseleer et al. [26] conclude their thesis on reverse engineering of gene networks with the following suggestion: "Since it is the ultimate goal to identify the causative relationships between gene products that determine important phenotypic parameters, top priority should be given to develop reverse engineering methods that provide significant predictions. Alternative computational approaches should be applied to given data sets, and their predictions tested in targeted experiments to identify the most reliable methods."

III. Biological Background

Ever since their inception, two types of microarray technologies have dominated. These are the complementary DNA (cDNA) microarrays and the oligonucleotide microarrays. Despite differences in their experiment protocols, both technologies measure the expression level of genes.

The range of applications of microarrays is potentially vast: they have been used to study expression profiles of genes in areas of development, the study of progression of a disease, survival upon onset of disease, and response to various drug compounds.

Applications which depend on the clustering of microarray data are the study of gene function, gene regulation, cellular processes, and subtypes of cells. Genes with similar expression patterns (coexpressed genes) can be clustered together with similar cellular functions. Coexpressed genes in the same cluster are likely to be involved in the same cellular processes. A strong correlation of expression patterns between genes indicates co-regulation. This can lead to an elucidation of regulatory networks. Clustering of different samples on the basis of corresponding expression profiles can also reveal subtypes of cells, which may otherwise be difficult to identify.

It is possible to divide clustering tasks into three categories: gene-based clustering, sample-based clustering, and subspace clustering [27]. In gene-based clustering, the genes are treated as objects of which the samples are the features. The aim would be to find clusters of co-expressed genes based on their expression patterns. In sample-based clustering, the samples are treated as the objects of which the genes are the features. The aim, in this case, would be to partition the samples into homogenous groups, where each group would correspond to some macroscopic phenotype. The third category, subspace clustering, considers both genes as well as samples symmetrically, so that either may be treated as objects or samples. The aim would be to find subsets of genes that participate in any cellular process which takes place only in a subset of samples.

In this work, we focus on gene-based clustering. But the methods that we develop could be used for subspace clustering also, provided that the sample space is processed in a symmetrical way. In order to focus on phenotypic interactions, we have opted to go closer to nature rather than resorting to computing shortcuts. This strategy of going closer to nature by incorporating gene expression into a messy genetic algorithm is adopted in [28]. Computing shortcuts might bring short-term benefits but do not fare well in realizing long-term strategies [29]. Nature has evolved

strategies which ensure success in the long-term and it is such techniques that we wish to mimic.

IV. Computational Background

Several researchers have proposed evolutionary algorithms for gene expression analysis, with varying degrees of success [30]. The curse of the elusive objective function is highlighted in many cases. The evolutionary algorithm needs to attach a fitness to each individual solution. The fitness of an individual is the manifestation of its phenotypic composition [31], [32].

The phenotype is dependent on the genotype. An objective function is generally used here to characterize this dependence. But in the absence of an appropriate objective function, researchers use several roundabout means to elicit the fitness. We intend to follow a more direct and intuitive path here. By embedding the expression of a gene within the individual, we have a ready reference for determining fitness.

When constructing gene networks, we study the relationship between genes. As shown in Fig. 1, if gi and gj are objects representing two such genes, their expression patterns across m samples may be written as $g_i = \{w_{ik} \mid 1 \le k \le m\}$ and $g_j = \{w_{jk} \mid 1 \le k \le m\}$. The proximity between genes can be expressed in terms of a correlation coefficient. Some correlation coefficients which are frequently used in gene expression studies are given by Eq. (1) and Eq. (2), where w_{ij} represents the expression level of the ith gene in the jth sample and μ_{gi} represents the average of expression levels of the ith gene over all the samples.

$$Euclidean(g_i, g_j) = \sqrt{\sum_{k=1}^{m} (w_{ik} - w_{jk})^2} \qquad (1)$$

Each gene object is normalized with zero mean and variance before calculating the distance. Since Euclidean distance does not capture the profiles of the expression patterns, another measure that is used is pearson correlation coefficient [7], [33].

$$Pear(g_i, g_j) = \frac{\sum_{k=1}^{m} (w_{ik} - \mu_{g_i})(w_{jk} - \mu_{g_j})}{\sqrt{\sum_{k=1}^{m} (w_{ik} - \mu_{g_i})^2} \sqrt{\sum_{k=1}^{m} (w_{jk} - \mu_{g_j})^2}} \qquad (2)$$

Once the proximity measure for the genes is defined, the gene interactions such as "meet", "know", "like", "dislike" can be defined as operations on genes gi and gj, as follows:

$meet(g_i, g_j)$ returns TRUE iff g_i and g_j were partners, at least once.

$know(g_i, g_j)$ returns TRUE iff the proximity measure for g_i and g_j is known.

$like(g_i, g_j)$ returns TRUE iff proximity measure for g_i and g_j is less than or equal to D.

$dislike(g_i, g_j)$ returns TRUE iff proximity measure for g_i and g_j is more than D.

These operations determine the character of the phenotypic interactions that take place between gene objects. By storing links between genes that "like" each other it is possible to elucidate the relationships that are required for reconstructing the gene network.

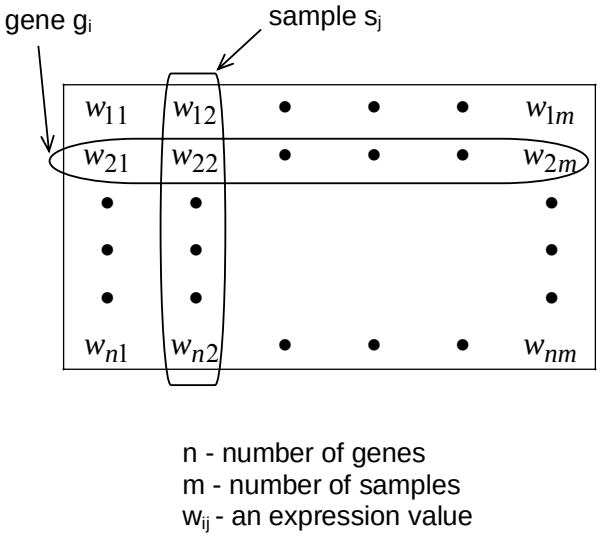

Fig. 1 Expression matrix representing a microarray dataset and corresponding notation

V. The Phenomic Algorithm

Gene expression is a crucial link between the genotype and the phenotype and hence plays a significant role in evolution. The genes express through transcription into mRNA which is translated into proteins. Proteins are the workhorses of the cell and are responsible for all the intra-cellular, inter-cellular and extra-cellular processes. Thus the genotype realizes the phenotype through gene expression.

Gene expression data embodies the characteristics of the phenotype and can be modeled as individuals in a genetic algorithm. Since the focus is on interactions between genes and their role in diseases, each gene is modeled as an individual whose expression pattern across samples represents that gene's phenotypic tendencies.

Evolutionary methods which are used in the analysis of microarray datasets ignore the inherent advantages due to the ready availability of the expression of genes. An approach based on phenotypic features which exploits these inherent advantages is the basis for developing this new evolutionary algorithm. Since the focus is on phenotypic features, rather than genetic features, the new name "phenomic algorithm" is proposed. The following are the main characteristics of the phenomic algorithm:

1. Modeling genes as individuals: The expression pattern of each of the genes is embedded within an object that represents an individual in the evolutionary algorithm. This type of a representation is the first step in gene-based clustering algorithms [27].

2. Simulating gene interaction: An environment for expression and interaction of individuals is simulated. Through random replication a population of individuals is

created and a host of interactions are performed on randomly chosen partners from this population. Interaction between partners would be similar to those encountered in nature. Partners would meet, know, like, or dislike each other.

3. Enforcing natural processes: The passage of time in nature affects individuals in the population. As interaction proceeds they learn more and more about each other. But new generations must be given a chance to bring in new relationships. This is done through consolidation of the population from time to time. Consolidation involves identification of individuals that have been replicated and removal of multiple copies. Only one copy of each type is left behind. The individuals that are removed are replaced by randomly choosing new individuals from the rest of the gene-pool.

4. Conservation of memes and phenotypic characteristics: Memes are patterns of gene interactions which give rise to functions and such patterns are conserved across generations. The principle of the extended phenotype [34], [35] envisages phenotypes which are not restricted to the boundaries of individuals. Several species collaborate in stable patterns which are part of food chains and similar artifacts of ecosystems. Such patterns need to be conserved since they may be part of the final solution.

The flowchart shown in Fig. 2 represents a randomized search strategy incorporating some of the ideas given above. In the flowchart, initially a microarray dataset is divided into n segments. The initial population is obtained by replicating the initial segment as many times as necessary to obtain N individuals. This replication is randomized and hence some individuals may have a larger representation in the population.

During the evolutionary phase, the individuals interact and it is during these interactions that the relationships between genes are captured. The actual nature of information captured depends on the application. For example, when constructing gene networks, one would be interested in capturing the links between genes. A link can be construed between two genes which are close to each other according to some distance measure. The distance measure used could be euclidean distance, pearson correlation coefficient or spearman correlation coefficient. These have been routinely used in gene expression studies as proximity measures [27], [36].

After a predetermined number of randomized interaction cycles, the population is consolidated to remove replicated individuals. This process is akin to death of some of the population and is done on a random basis. The links embedded in dying individuals are carried over to the survivors.

A birth process at this stage brings in new individuals from the rest of the microarray data segments (one by one). Again there is randomized replication and the process repeats over and over again. All segments are ultimately absorbed and will have interacted and consolidated so that the final population hold the results of the process. As seen from experimental results, this algorithm is able to discover links between genes when applied to gene expression data.

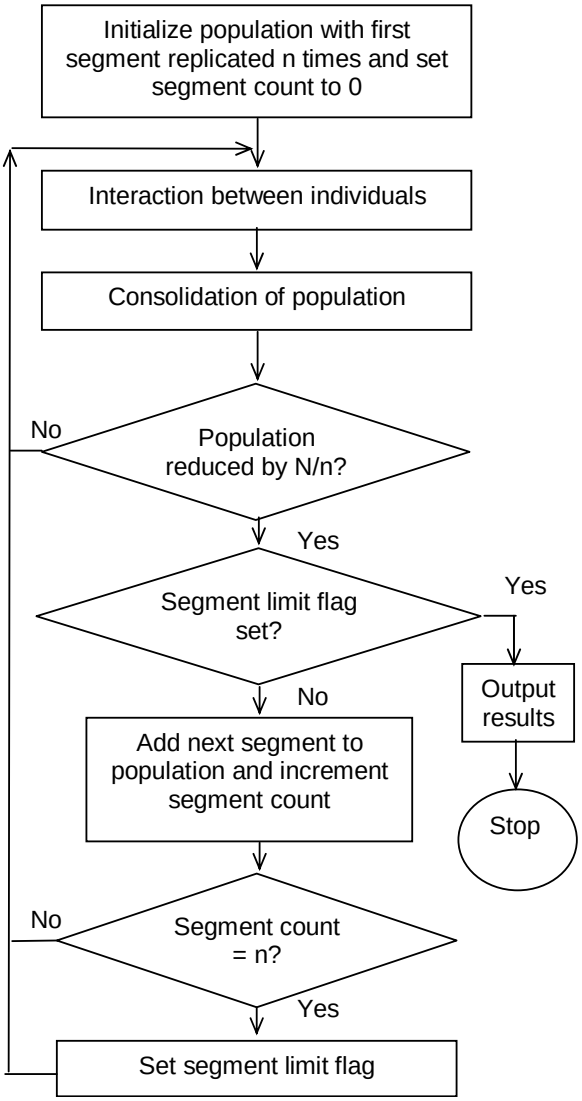

Fig. 2 Flowchart depicting the basic phenomic algorithm

VI. Results and Discussion

We have applied the phenomic algorithm to the problem of reconstruction of gene networks. Specifically, we have chosen the yeast sporulation dataset used by DeRisi [37] which represents the temporal expression patterns of 6118 genes of Saccharomyces cerevisiae. The dataset is available at http://gepas.bioinfo.cipf.es/data/sporulation/sporulation.txt.

We have followed the preprocessing strategy adopted by DeRisi [37] and chosen only those genes that have a 2.2-fold change in mRNA levels. Thus only 698 genes are chosen for further processing.

The pearson correlation coefficient, as given in Eq. (2), is used to determine the distance between gene profiles. Only those gene relationships which are closer than a preset distance threshold are considered significant. For various values of distance threshold d, the gene networks obtained are

shown in Fig. 3, Fig. 4, and Fig. 5.

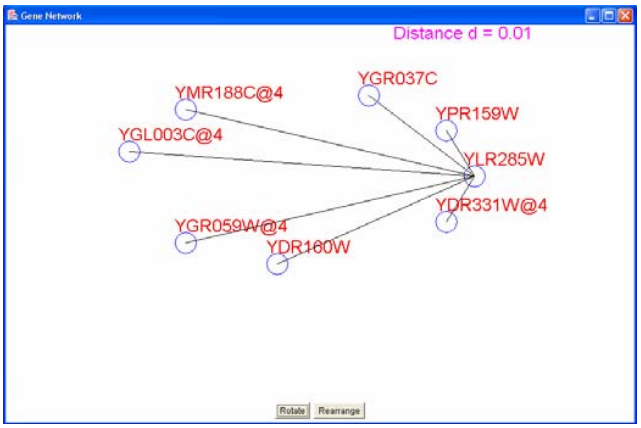

Fig. 3 A gene network derived from the yeast sporulation dataset, with a distance threshold d=0.01

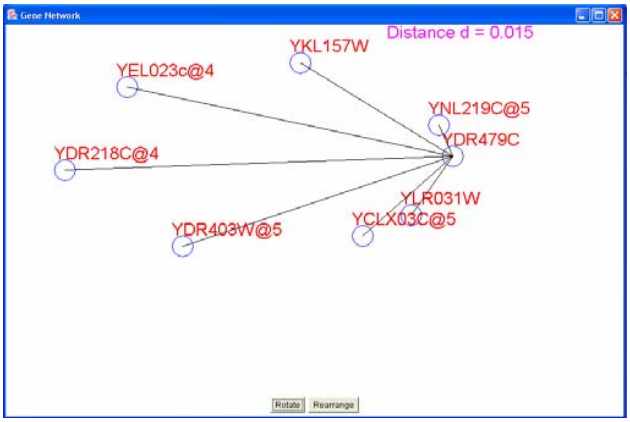

Fig. 4 A gene network derived from the yeast sporulation dataset, with a distance threshold d=0.015

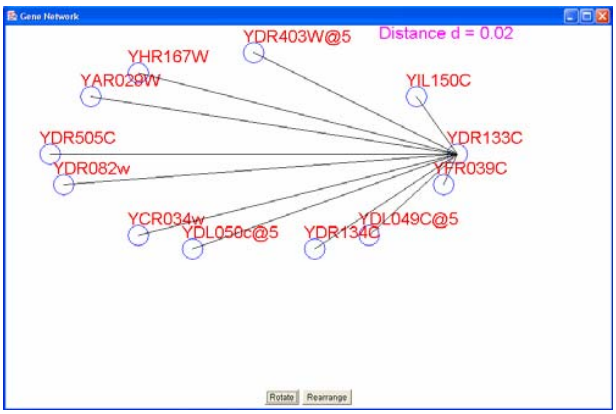

Fig. 5 A gene network derived from the yeast sporulation dataset, with a distance threshold d=0.02

The gene networks derived by the phenomic algorithm show relationships between genes that have been recorded in the Saccharomyces Genome Database at http://www.yeastgenome.org/ [38]. When distance threshold d is more than 0.05, the phenomic algorithm finds networks with a large number of relationships. At that stage additional information is required if significant relationships have to be identified.

VII. Conclusion

We have presented a new evolutionary algorithm that focuses on the phenotypic, rather than genetic, features of an individual. The relationships between genes are elicited by allowing interaction between individuals in a virtual environment that simulates the survival of the fittest. These relationships form the links between genes in the gene network. The algorithm was applied to yeast sporulation data and the resulting gene networks were found to be biologically relevant.

Gene networks represent complex relationships among genes, and several types of gene networks have been proposed that focus on one type of interaction or the other. Hence we have gene regulatory networks, metabolic networks, protein-protein interaction networks, etc. Currently we are working on extending the phenomic algorithm by bringing in additional information resources [39] at the interaction stage, so that other types of relationships between genes can be derived.

References

[1] A. Schulze, and J. Downward, "Navigating gene expression using microarrays - a technology review," Nature Cell Biology, Vol 3, pp. E190-E195, Aug 2001.
[2] L. A. Soinov, M. A. Krestyaninova, and A. Brazma, "Towards reconstruction of gene networks from expression data by supervised learning," Genome Biology, 4(1), pp. R6, 2003.
[3] P. D'haeseleer, S. Liang, and R. Somogyi, "Gene expression analysis and genetic network modeling: Tutorial," Pacific Symposium on Biocomputing (PSB '99), 1999.
[4] W. P. Kuo, E. Kim, J. Trimarchi J, et al., "A primer on gene expression and microarrays for machine learning researchers," Jour. of Biomedical Informatics, 37 (2004), pp. 293-303, 2004.
[5] N. L. W. van Hal, O. Vorst, A. M. M. L. van Houwelingen, et al., "The application of DNA microarrays in gene expression analysis," Jour. of Biotechnology, 78, pp. 271–280, 2000.
[6] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with complementary DNA microarray," Science, 270 (5235), pp. 467-470, 1995.
[7] W. P. Kuo, E. Mendez, C. Chen, et al., "Functional relationships between gene pairs in oral squamous cell carcinoma," Proc. AMIA Symp. 2003, pp. 371-375, 2003.
[8] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman, "Nonparametric methods for identifying differentially expressed genes in microarray data," Bioinformatics, 18(11), pp. 1454–1461, 2002.
[9] K. A. Baggerly, K. R. Coombes, K. R. Hess, D. N. Stivers, L. V. Abruzzo, and W. Zhang, "Identifying differentially expressed genes in cDNA microarray experiments," Jour. Comput. Biol., 8(6), pp. 639–659, 2001.
[10] G. Didier, P. Brezellec, E. Remy, and A. Henaut, "GeneANOVA - gene expression analysis of variance," Bioinformatics, 18(3), pp. 490–491, 2002.
[11] T. J. Phillips, and J. K. Belknap, "Complex-trait genetics: emergence of multivariate strategies," Nat. Rev. Neurosci., 3(6), pp. 478–485, 2002.
[12] S. Ramaswamy, P. Tamayo, et al., "Multiclass cancer diagnosis using tumor gene expression signatures," Proc. Natl. Acad. Sci. USA, 98(26), pp. 15149–15154, 2001.
[13] J. Nikkila, P. Toronen, S. Kaski, J. Venna, E. Castren, and G. Wong, "Analysis and visualization of gene expression data using self-organizing maps," Neural Netw., 15(8–9), pp. 953–966, 2002.
[14] G. N. Fuller, K. R. Hess, C. H. Rhee, et al., "Molecular classification of human diffuse gliomas by multidimensional scaling analysis of gene

expression profiles parallels morphology-based classification, correlates with survival, and reveals clinically relevant novel glioma subsets," Brain Pathol., 12(1), pp. 108–116, 2002.
[15] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," Proc. Natl. Acad. Sci. USA, 95(25), pp. 14863–14868, 1998.
[16] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," Bioinformatics, 16(10), pp. 906–914, 2000.
[17] E. Keedwell, and A. Narayanan, "Genetic algorithms for gene expression analysis," EvoBio 2002, Springer Verlag LNCS 2611, pp. 76-86, 2003.
[18] J. Liu, H. Iba, and M. Ishizuka, "Selecting informative genes with parallel genetic algorithms in tissue classification," Genome Informatics, 12, pp. 14-23, 2001.
[19] C. H. Ooi, and P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," Bioinformatics, 19, pp. 37-44, 2003.
[20] J. M. Deutsch, "Evolutionary algorithms for finding optimal gene sets in microarray prediction," Bioinformatics, 19, pp. 45-52, 2003.
[21] C. Creighton and S. Hanash, "Mining gene expression databases for association rules," Bioinformatics, 19, pp. 79-86, 2003.
[22] R. Somogyi, S. Fuhrman, M. Askenazi, and A. Wuensche, "The gene expression matrix: towards the extraction of genetic network architectures," Proc. of Second World Cong. of Nonlinear Analysts (WCNA96), 30(3), pp. 1815-1824, 1997.
[23] S. Liang, S. Fuhrman, and R. Somogyi, "REVEAL, a general reverse engineering algorithm for inference of genetic network architectures," Pacific Symp. on Biocomputing, 3, pp. 18-29, 1998.
[24] T. Akutsu, S. Miyano, and S. Kuhara, "Identification of genetic networks from a small number of gene expression patterns under the boolean network model," Pacific Symp. on Biocomputing, 4, pp. 17-28, 1999.
[25] T. Akutsu, S. Miyano, and S. Kuhara, "Algorithms for inferring qualitative models of biological networks," Pacific Symp. on Biocomputing, 2000.
[26] P. D'haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: from co-expression clustering to reverse engineering," Bioinformatics, 16(8), pp. 707-726, 2000.
[27] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: a survey," IEEE Trans. on Knowledge and Data Engg., Vol 16, No 11, pp. 1370-1386, 2004.
[28] H. Kargupta, "The gene expression messy genetic algorithm," Proc. Of IEEE Intl. Conf. on Evolutionary Computation, 1996.
[29] J. Falqueto, J. M. Barreto, and P. S. da Silva Borges, "Amplification of perspectives in the use of evolutionary computation," BIBE 2000, pp. 150, IEEE Int'l. Symp. on Bioinformatics and Biomedical Engg., 2000.
[30] G. B. Fogel, and D. W. Corne (Editors), *Evolutionary computation in bioinformatics*, Morgan Kaufmann, 2003.
[31] G. Kampis, "A Causal Model of Evolution," Proc. of 4th Asia-Pacific Conf. on Simulated Evol. and Learning (SEAL 02), pp. 836-840, 2002.
[32] R. Dawkins, *The blind watchmaker*, Penguin Books, 1988.
[33] D. Stekel, *Microarray bioinformatics*, Cambridge University Press, 2003.
[34] R. Dawkins, *The selfish gene*, Oxford University Press, 1976.
[35] R. Dawkins, *The extended phenotype*, Oxford University Press, 1982.
[36] P. Baldi, and G. W. Hatfield, *DNA microarrays and gene expression*, Cambridge University Press, 2002.
[37] S. Chu, J. DeRisi, M. Eisen, et al., "The transcriptional program of sporulation in budding yeast," Science, 282, pp. 699-705, 1998.
[38] SGD project. "Saccharomyces Genome Database" http://www.yeastgenome.org/ (15/9/2007).
[39] Z. Lubovac, and B. Olsson, "Towards reverse engineering of genetic regulatory networks," Technical Report No. HS-IDA-TR-03-003, University of Skovde, Sweden, 2003.