Inferring gene coexpression networks with Biclustering based on Scatter Search

Juan A. Nepomuceno
Dpt. Lenguajes y Sistemas Informáticos
University of Seville, Spain
janepo@us.es

Alicia Troncoso, Jesús S. Aguilar–Ruiz
Area of Computer Science
Pablo de Olavide University, Spain
{ali,aguilar}@upo.es

***Abstract*—The identification of regulatory modules is one of the most important tasks in order to discover disease markers. This paper presents a methodology to infer coexpression networks based on local patterns in gene expression data matrix. In the proposed algorithm two steps can clearly be differentiated. Firstly, a Biclustering procedure that uses a Scatter Search schema to find biclusters and, secondly, a network extraction procedure based on linear correlations among the genes of the previously obtained bicluster. Experimental results from Yeast cell Cycle are reported where three different algorithms have been applied. Also, a possible understanding of one of the obtained networks has been presented from a biological point of view.**

***Keywords*-Gene Expression Data, Gene Coexpression Networks, Biclustering, Scatter Search.**

I. Introduction

The process whereby gene codify proteins is known as *Gene Expression* process. *DNA microarrays* experiments enable us to measure the gene expression level of thousand of gene simultaneously. After several preprocessing steps, the expression matrix of a microarray experiment is a numerical matrix where rows represent gene and columns the experimental conditions of the experiment: different samples, different times points, different patients, etc. Each number of the matrix is the gene expression level of a gen under a specific condition. This technology provides a huge volume of biological information that must be analyzed to be became in knowledge using Data Mining techniques.

Biclustering is an Unsupervised Data Mining technique that searches for local patterns in the gene expression data matrix. Traditional Clustering is not adequate with this kind of data because gene are considered along all experimental conditions and interesting local relations could not be contemplated. Several Biclustering algorithms have been proposed and recently published surveys can be found in [1], [2], [3]. In the context of microarray analysis, Biclustering was firstly considered by Cheng and Church in 2000. Cheng and Church algorithm [4] is a greedy iterative search method and consists in building a bicluster adding or removing rows or columns iteratively, thus, improving its quality which is measured with the *Mean Squared Residue* (MSR). Recently, some approaches based on metaheuristics have been proposed. Some of them use MSR as part of their merit function and are based on Scatter Search metaheuristics [5]. Other have similar schemes but using others kind of quality criteria in their fitness function [6], [7].

The identification of regulatory modules is one of the most important tasks in order to identify *disease markers*. For example, in cancer tissue classification, subnetwork marker approaches have been proven to be superior than single gene marker approaches [8]. Therefore, the identification of gene interaction network is one of the main open problems. Gene networks inferring techniques are based on the idea that if two genes show similar expression profiles, they are supposed to follow the same regulatory regime [9]. This idea assumes that co-expression gene means co-regulation gene. Biclustering algorithms discover local dependencies among genes analyzed in microarray experiments. This idea of local co-expression can be used to infer gene interaction networks. Gene interaction networks are inferred in [10] using a multiobjective evolutionary biclustering algorithm. In [8], Biclustering is used as inner step of a method based on computing gene subnetwork marker in order to classify cancer tissues.

A methodology to infer gene interaction networks with biclusters obtained by a Biclustering algorithm based on Scatter Search is presented in this paper. It is organized as follows: Section 2 presents the proposed methodology. Experimental results and the analysis of some of the obtained networks are reported in Section 3. Finally, Section 4 outlines the main conclusions of the paper and future works.

II. Description of the Methodology

The proposed methodology to infer gene coexpression networks combines a Biclustering algorithm with a network extraction procedure. The Biclustering algorithm reports a bicluster through the optimization of a fitness function. A gene interaction network is built using the information of linear correlations among genes belonging to this bicluster. The Biclustering algorithm is based on Scatter Search schemes. Scatter Search is a population-based metaheuristic where a set of individuals that represent trial solutions evolves in order to find optimal solutions of the problem. This metaheuristic emphasizes systematic process against random procedure as for in example Genetic Algorithms.

The pseudocode of the method is presented in Algorithm 1. The input data are the matrix of gene expression of the

Algorithm 1 Methodology for inferring gene co-expression networks.

```
INPUT microarray $M$, correlation parameter $0 \leq p \leq 1$, penalization
    factor for fitness function $M_1$ and $M_2$.
OUTPUT gene interaction network $N$.

begin
  BICLUSTERING PROCEDURE:
  Generating Initial Population
  Building Reference Set: $RefSet$
  while ($i < 20$) do
    while ($RefSet$ is not stable) do
      Combination Method($RefSet$)
      Reference Set Update Method
    end while
    Rebuilding Reference Set
  end while
  $B \leftarrow$ the best bicluster from $RefSet$
  NETWORK EXTRACTION PROCEDURE:
  Let be $G = \{g_1, \dots g_k\}$ set of gene of $B$
  Calculation of $M = (m_{i,j})_{0 \leq i,j \leq k}$ correlation matrix where $m_{ij} =
  \rho(g_i, g_j)$
  Network Generation using M: pre-Network $N_0$
  Modification of $M$ using correlation parameter $p$
  if ($|m_{i,j}| \leq p$) then
    $m_{i,j} = 0$
  end if
  Network Generation using M: post-Network $N$
  Result $\leftarrow$ N.
end
```

microarray experiment, a correlation parameter ($0 \leq p \leq 1$) to extract the network from the bicluster and two parameters M_1 and M_2 in the fitness function in order to control how the bicluster resultant is. The output is the gene interaction network inferred.

A. Biclustering Procedure

The Biclustering algorithm presented in this paper is based on the ones presented in [5], [6], [7]. Thus, a Scatter Search scheme is used to solve an optimization problem whose result is a bicluster.

The optimization process consists in the evolution of a set called *Reference Set* and at the end of the process the best bicluster of this set will be the output. Firstly, the initial population of biclusters (possible solutions) is generated as diverse as possible (according to a concept of distance). The *Reference Set* set is initially built with the best solutions from the population, according to the value of their fitness function, and the most scattered ones from the population regarding the previous best solutions. This set is updated by using the *Combination Method* until it does not change. When the *Reference Set* is stable, that is, after applying the combination, it contains the same solutions that the reference set at the previous iteration, then it is rebuilt again. That is, the building of the *Reference Set* is based on quality and diversity, but its updating is only guided by quality. Thus, diversity is introduced in the evolutionary process when the initial population is generated and, mainly, when the reference set is rebuilt at each step. The quality is due to the Combination Method and in some versions is due to a local search method which improves the solutions in each step too.

Possible solutions, biclusters, are codified as binary strings which represent whether a gene or a condition of the microarray matrix belongs or not to the bicluster. The optimization is established respect the value of a fitness function. Given a bicluster B the fitness function to determine its quality is:

$$f(B) = Q(B) + M_1 \left(\frac{1}{nG}\right) + M_2 \left(\frac{1}{nC}\right) \qquad (1)$$

where $Q(B)$ is the function to determinate the quality of patterns inside the bicluster, nG and nC are the number of genes and conditions, respectively and M_1 and M_2 are penalty factors to control the volume of the bicluster.

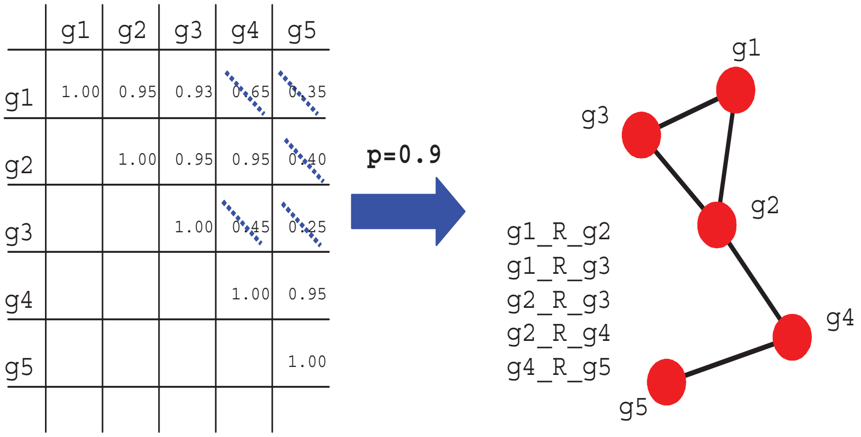
$Q(B)$ establishes the homogeneity of patterns among genes from the bicluster. Several criteria have been considered: the MSR [5] and the *average correlation* [6], [7].

B. Network Extraction Procedure

This procedure constitutes a pruning phase to eliminate not relevan connections. The set of gene belonging to the obtained bicluster is considered to build the correlation matrix. This matrix is a square matrix where each element, $m_{i,j}$, is the linear correlation between two gene of the set: $m_{i,j} = \rho(g_i, g_j)$. Note that, $\rho_{i,j} = \rho_{j,i}$, $0 \leq \rho_{i,j} \leq 1$ and that correlation equal to 0 implies independency and equal to 1 implies the same tendency. It must be noted that the expression profile of every gene is only considered respect the conditions that determine the bicluster. The value of the linear correlation expresses a local dependence among genes determined by the bicluster. A gene interaction graph can be built with the information from the correlation matrix. This matrix is processed in order to only contemplate highly correlated gene. Every element with a value under a parameter is rejected to build the graph. Two different networks can be built: before and after the correlation matrix is processed, we call them pre-network and post-network respectively. After matrix is processed, the network built is the procedure result. Figure 1 represents a toy example of the network generation step of the procedure.

III. Experimental Results

The data set used for experiments has been the same that in [10], *Yeast cell cycle CDC28* data, that studies *yeast Saccharomyces cerevisiae* along temporal conditions of two completes cycles of cell cycle. The preprocessing steps have been made in the same way: gene with more than half missing values have been eliminated and the remaining ones have been estimated. After that, the microarray matrix is composed by 6131 genes and 17 conditions that represent samples each 10 minutes in the microarray experiment. Three different Biclustering algorithms based on Scatter



	g1	g2	g3	g4	g5
g1	1.00	0.95	0.93	~~0.65~~	~~0.35~~
g2		1.00	0.95	0.95	~~0.40~~
g3			1.00	~~0.45~~	~~0.25~~
g4				1.00	0.95
g5					1.00

Figure 1. An example of the Network Generation Step where a set of five gene determines a network structure.

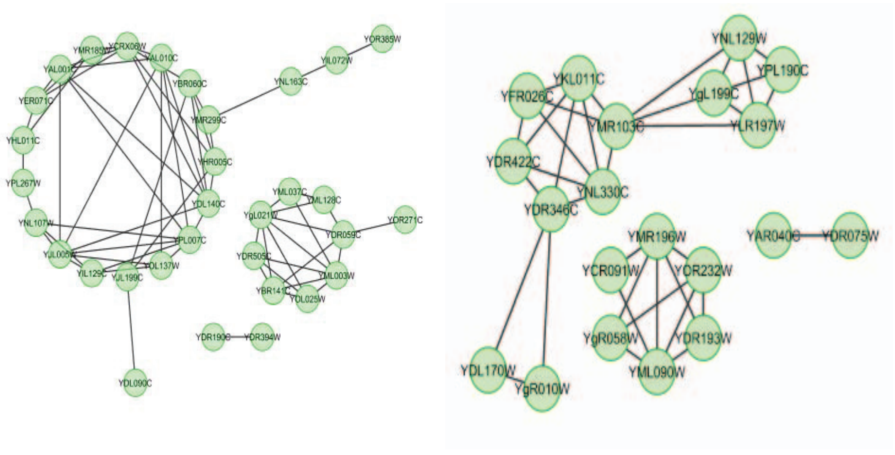


Figure 2. Two networks obtained for SSH: $(M_1 = 1, M_2 = 1)$ on the left side and $(M_1 = 10, M_2 = 10)$ on the right side.

Search have been considered for the experiments. SSH [5] uses MSR function as quality criteria of each bicluster. SScorr [6] uses the average correlation as quality criteria. Finally, SScorrM which uses the average correlation and works with a local search procedure during the optimization process.

Table I reports the experiments carried out. The results for executions with values of the parameters $(M_1 = 1, M_2 = 1)$, $(M_1 = 1, M_2 = 10)$ and $(M_1 = 10, M_2 = 10)$ respectively for 10 biclusters for SSH, SScorr and SScorrM algorithms are presented. Low levels of M_1 and M_2 implies more importance to the homogeneity among genes that the volume of the bicluster, however, high level of them implies the opposite situation. The values shown in the Table are the average for these 10 executions. The Biclustering procedure is an evolutionary process and the results vary for each execution. The information presented in the Table is the obtained bicluster (number of genes and number of conditions), the pre-Network (the network that can be built before the pruning phase using the correlation parameter) and post-Network which is the network reported as the result of the proposed method. For each network, the number of vertices and the number of edges are shown. The value chosen for the correlation parameter has been $p = 0.9$ in order to obtain genes connected with strong local patterns. Note that the number of genes of biclusters is the same that the number of vertices of pre-networks. The reason is that every gene is considered to build the pre-network. However, the network reported by the method, that is, the post-network, has a lower number of genes because only highly locally correlated gene are considered and there are very connections that disappear. It can be observed that for SSH there are post-networks with a huge number of edges. The reason is that some biclusters have only two conditions and these kind of biclusters are not good enough to infer a network as two conditions determine a huge number of gene with exactly the same behavior, and therefore, correlation equal to one, but this situation does not represent a good local pattern and the pre and post network

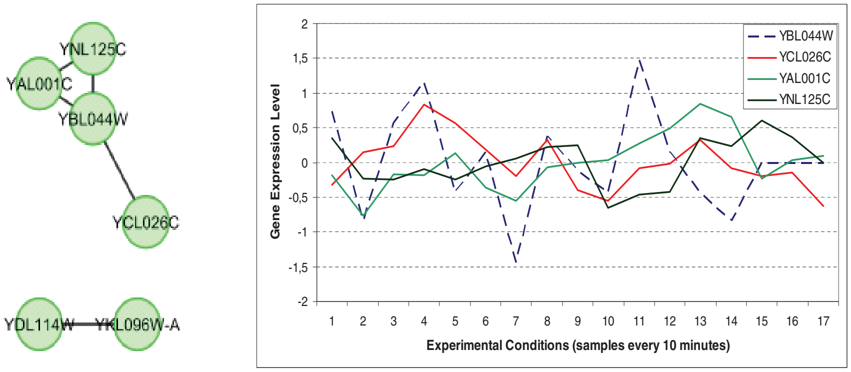


Figure 3. A network obtained for SScorrM with ($M_1 = 1$,$M_2 = 10$ and the behavior of a group of its gene in the microarray experiment.)

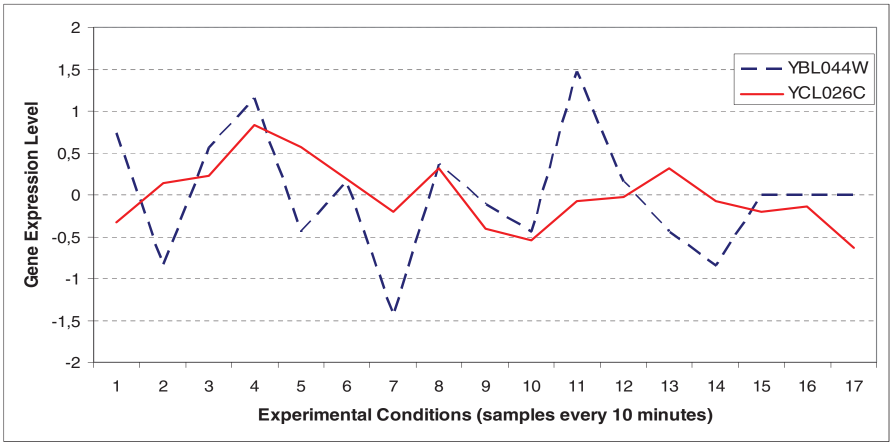


Figure 4. Detail of the proposed hypothesis: $YBL044W$ as transcription factor of the regulation of gene $YCL026$.

are exactly the same. SSH found biclusters with a low number of conditions due to the quality criteria used in the fitness function (MSR). In general, note that when the values of parameters M_1 and M_2 increase, the biclusters have more volume (number of genes by number of conditions) and less homogeneity. Hence, more connections will be disappear in the pruning phase. SScorrM discovers biclusters with highly correlated genes and for this reason the post-network is not so different of the pre-network built without considering the pruning phase. Note that for execution of SScorrM with $M_1 = 10$ and $M_2 = 10$ the difference between the number of edges in the pre and post network is very high. However, for $M_1 = 1$ and $M_2 = 1$ is the opposite. The reason is that the terms of the fitness function related to the volume of biclusters has a bigger weight than the ones related to the patterns underlying in the bicluster. Therefore, a great number of genes with a low correlation is obtained.

Figure 2 presents two networks obtained by the proposed methodology. The one on the left side was obtained with values $M_1 = 1$ and $M_2 = 1$ in one of the executions for SSH. This network has 32 vertices and 63 edges. This value is coherent with the ones presented in Table I. As it was mentioned before, the numbers in the Table are the average values and several executions obtained biclusters that are not enough good to build a network, in particular, the ones which have two conditions. The network drawn in the Figure was obtained from a bicluster with 33 genes and 4 conditions. The pre-network had 528 edges. Only 63 edges represent strong locally connections among genes. The network has two components and a module with two isolated gene. The network on the right side of Figure 2 is a network obtained for a SSH execution with $M_1 = 10$ and $M_2 = 10$. It has 20 vertices and 36 edges and it was obtained from a bicluster with 26 genes and 5 conditions. The pre-network had 325 edges. It can be observed that two circular components with 6 genes exist and one of them is connected with another one with a 4–gene module. Moreover, it has a module with two isolated genes.

Figure 3 presents a network obtained for SScorrM with value $M_1 = 1$ and $M_2 = 10$ with a low number of genes. This network has only 6 vertices and 5 edges. The bicluster used to build it had 6 genes and 15 conditions.

Table I
INFORMATION ABOUT EXPERIMENTS WITH SSH, SSCORR AND SSCORRM ALGORITHMS.

		Biclusters		**pre-Networks**		**post-Networks**	
		gene	conditions	vertices	edges	vertices	edges
SSH	$(M_1 = 1, M_2 = 1)$	69.2	3.4	69.2	2659.5	67.7	1175.3
	$(M_1 = 1, M_2 = 10)$	63.1	3.2	63.1	2117.9	62.7	1038.9
	$(M_1 = 10, M_2 = 10)$	74.3	3.3	74.3	3761.3	70.6	2848.5
SScorr	$(M_1 = 1, M_2 = 1)$	252.9	7.8	252.9	29403	195.2	11272.5
	$(M_1 = 1, M_2 = 10)$	259.5	15.1	259.5	34389.9	48.4	44.1
	$(M_1 = 10, M_2 = 10)$	300.9	14.5	300.9	45718	70.9	66.3
SScorrM	$(M_1 = 1, M_2 = 1)$	15.8	7.5	15.8	121.2	15.8	105.4
	$(M_1 = 1, M_2 = 10)$	10.2	13.9	10.2	53	7.7	11.2
	$(M_1 = 10, M_2 = 10)$	64.8	13.9	64.8	2942.9	31.1	53

The pre-network had 15 edges. This network has a module with two gene connected each other and a module with 4 genes: *YAL001C*, *YNL125C*, *YBL044W* and *YCL026C*. The gen *YBL044W* is the central node of the module and it can be analyzed as the mean gen of this group. On the right side of the Figure the expression level of these four genes is represented. Note that they are represented along every conditions of the microarray, therefore this graphic is not the representation of the bicluster related. *YBL044W* is drawn with a dashed line and the rest with solid lines. It can be observed that the gene mentioned is expressed with higher values that the rest and it drives the behavior of the other three connected genes. A hypothesis to study is to consider *YBL044W* the transcription factor and the other three genes would be the corresponding regulated genes. Note that the hypothesis has been formulated considering the topology of the network. It can be observed that from to 120 (number 12 on the x axis) to 150 the solid lines do not follow the tendency of the dashed one. Hence, the hypothesis is that this regulation process does not occur in this moment of the cell cycle.

Figure 4 presents the behavior of two genes from the previous subnetwork analyzed with more details. Gen *YCL026C* is represented by the solid line and gen *YBL044W* by the dashed line. It can be clearly observed that *YBL044W* is expressed more intensively than *YCL026C* and it drives the tendency of two genes in the most of the time of the experiment. From the begin to the minute 30 they have different tendency but from 30 to 40 they express in the same way. From minutes 40 to 90 both genes show the same behavior but with different intensity. From minutes 90 until 130, the gene candidate to be the transcription factor (dashed line) shows the same behavior that the other one but a short period of time before and with more intensity, that is, with higher values of the expression level. From this point to the end both genes show different tendency.

IV. CONCLUSIONS

In this paper a methodology to infer genes coexpression networks has been presented. As previous step, a Biclustering algorithm has been used in order to obtain biclusters. Later, a network extraction procedure based on the linear correlations among genes of such biclusters has been provided. Experiments have been reported from three different Biclustering algorithms and a detailed analysis of one of the obtained networks has been shown.

Future works will focused on some improvements for the proposed methodology with regards to the network extraction procedure and the comparison with other gene networks algorithms.

V. ACKNOWLEDGMENTS

The financial support given by the Ministry of Science and Innovation, projects TIN2007-68084-C02, PCI2006-A7-0575, and by the Junta de Andalucía, project P07-TIC-02611 is acknowledged.

REFERENCES

[1] Madeira, S., Oliveira, A.: Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE Transactions on Computational Biology and Bioinformatics **1**(1) (2004) 24–45

[2] Tanay, A., Sharan, R., Shamir, R.: Biclustering Algorithms: A Survey. Handbook of Computational Molecular Biology **9** (2005) 26–1

[3] Busygin, S., Prokopyev, O., Pardalos, P.: Biclustering in data mining. Computers and Operations Research **35**(9) (2008) 2964–2987

[4] Cheng, Y., Church, G.: Biclustering of Expression Data. Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology **8** (2000) 93–103

[5] Nepomuceno, J.A., Troncoso, A., Aguilar-Ruiz, J.S.: An overlapping control-biclustering algorithm from gene expression data. In: ISDA 2009: Ninth International Conference on Intelligent Systems Design and Applications, Pisa, Italy , November 30-December 2, 2009. (2009) 1239–1244

[6] Nepomuceno, J.A., Troncoso, A., Aguilar-Ruiz, J.S.: Evolutionary metaheuristic for biclustering based on linear correlations among genes. In: SAC 2010: Proceedings of the 2010 ACM Symposium on Applied Computing (SAC), Sierre, Switzerland, March 22-26, 2010. (2010) 1143–1147

[7] Nepomuceno, J., Troncoso, A., Aguilar-Ruiz, J.: Biclustering of gene expression data by correlation-based scatter search. BioData Mining **4**(1) (2011) 3

[8] Dao, P., Colak, R., Salari, R., Moser, F., Davicioni, E., Schönhuth, A., Ester, M.: Inferring cancer subnetwork markers using density-constrained biclustering. Bioinformatics **26**(18) (2010) i625–i631

[9] Nepomuceno-Chamorro, I., Aguilar-Ruiz, J., Riquelme, J.: Inferring gene regression networks with model trees. BMC Bioinformatics **11**(1) (2010) 517

[10] Mitra, S., Das, R., Banka, H., Mukhopadhyay, S.: Gene interaction - an evolutionary biclustering approach. Information Fusion **10**(3) (2009) 242–249