14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

A DIFFERENTIAL BICLUSTERING ALGORITHM FOR COMPARATIVE ANALYSIS OF GENE EXPRESSION

[1]Alain B. Tchagang, [2]Ahmed H. Tewfik, [3]Amy P.N. Skubitz and [4]Keith Skubitz

Dept. of [1]Biomedical Engineering, [2]Electrical and Computer Engineering, [3]Lab. Pathology and Medicine, and [4]Medicine
University of Minnesota, 55455, Minneapolis, USA, phone: + (1) 612 625 6024, fax: + (1) 612 625 4583,
Email: *{tcha0003, tewfik, skubi002, skubi001}@umn.edu*

ABSTRACT

Convergences and divergences among related organisms (S.cerevisiae and C.albicans for example) or same organisms (healthy and disease tissues for example) can often be traced to the differential expression of specific group of genes. Yet, algorithms to characterize such differences and similarities using gene expression data are not well developed. Given two related organisms A and B, we introduce and develop a differential biclustering algorithm, that aims at finding convergent biclusters, divergent biclusters, partially conserved biclusters, and split conserved biclusters. A convergent bicluster is a group of genes with similar functions that are conserved in A and B. A divergent bicluster is a group of genes with similar function in A (or B) but which play different role in B (or A). Partially conserved biclusters and split conserved biclusters capture more complicated relationships between the behavior and functions of the genes in A and B. Uncovering such patterns can elucidate new insides about how related organisms have evolved or the role played by some group of genes during the development of some diseases. Our differential biclustering algorithm consists of two steps. The first step consists of using a parallel biclustering algorithm to uncover all valid biclusters with coherent evolutions in each set of data. The second step consists of performing a differential analysis on the set of biclusters identified in step one, yielding sets of convergent, divergent, partially conserved and split conserved biclusters.

1. INTRODUCTION

Prior computational methodologies for comparative analysis of large scale gene expression data have focused primarily on evolutionarily distant model organisms, for which large sets of expression data are available [1-4]. A generalization of the singular value decomposition for example was applied in [1] for a comparative analysis of the cell cycle datasets from saccharomyces cerevisiae and human. Such studies have primarily emphasized the analysis of co-regulated patterns, rather than differences in expression patterns. They have demonstrated that conservation of co-expression can improve functional gene annotation [2, 3]. On the other hand, algorithms to characterize differences and similarities of related organisms using gene expression data are not well developed.

In recent years, for example, the C.albicans genome was sequenced [5], revealing that almost two-thirds of its ~6000 open reading frames are orthologous to S.cerevisiae genes. Microarray studies were performed by several groups characterizing the C.albicans genome-wide expression program under a range of conditions [6-11]. The availability of large sets of expression data in both S.cerevisiae and C.albicans, which are related organisms that span a significant evolutionary distance, provides a useful framework to develop and test computational tools for analyzing differences and similarities of related organism using their gene expression data. In [12] for example, a differential clustering algorithm was developed for comparative analysis of gene expression data.

In this study, given two related organisms A and B for example, we propose and develop a novel methodology called differential biclustering algorithm that aims at finding convergent biclusters, divergent biclusters, partially conserved biclusters, and split conserved biclusters. A convergent bicluster is a group of genes with similar functions that are conserved in A and B. A divergent bicluster is a group of genes with similar function in A (or B) but which play different role in B (or A). Partially conserved biclusters and split conserved biclusters capture more complicated relationships between the behavior and functions of the genes in A and B and are defined below. Uncovering such patterns can elucidate new insides about how related organisms have evolved or the role played by some group of genes during the development of some diseases. When the proposed differential biclustering methodology is applied to the gene expression data of healthy and disease tissues of the same organism under the same set of conditions, it can be used to identify group of genes that are related to the development of the disease. More precisely, the evolution of a disease into an organism can be traced to a group of divergent biclusters, e.g., a group of genes that are functionally related in the healthy tissue but which play different functions or roles in diseased tissues. Therefore, by analyzing the differences and similarities of the genomic properties of such tissues, those specific genes and the genetic pathways in which they are involved can be identified. Thus biological analysis and experimentation could then confirm the biological significance of the candidate group of genes, and the role they play during the early stage, developmental stage, and late stage of the disease.

Unlike prior comparative gene expression data approaches that are based on a global comparison, our methodology provides a comprehensive framework for a local comparative analysis of gene expression data via a parallel biclustering approach developed in [13].

The rest of this paper is organized as follows. In paragraph 2 we provide a brief description of the parallel biclustering algorithm. In paragraph 3 and 4, we introduce and develop the differential biclustering algorithm. In paragraph 5, we illustrate our methodology by performing an analysis on normal and cancer ovarian dataset.

2. PARALLEL BICLUSTERING ALGORITHM

2.1 Definition

Let us consider the N x M gene expression matrix $A = [a_{ij}]$, with set of rows or genes $G = \{g_1, \ldots, g_N\}$, and the set of conditions or columns $C = \{c_1, \ldots, c_M\}$. The element a_{ij} corresponds to a value representing the relation between row i and column j, which is the expression level of gene i under condition j. Thus we will also define the gene expression matrix as: $A = \{G, C\}$. Given the gene expression matrix A as defined above, we define a bicluster $B_k = \{I_k, J_k\}$ as a subset of A, with I_k being a subset of G, and J_k a subset of C. Therefore, the specific problem addressed by biclustering algorithms is to identify the set of biclusters $B_k = \{I_k\ J_k\}$ such that each bicluster B_k satisfies some specific characteristics of homogeneity. There exist four types of biclusters that have been identified in the literature [14]: biclusters with constant values, biclusters with constant values on rows or columns, biclusters with coherent values, and biclusters with coherent evolutions. Uncovering such patterns from a set of gene expression data can provide a starting point for elucidating genetic pathways.

2.2 Parallel Biclustering

Introduced in [13] by Tewfik and Tchagang, the parallel biclustering algorithm aims at finding all biclusters with coherent evolution (subgroups of genes that are up-regulated or down-regulated coherently across subgroups of conditions) from a set of data in a timely manner without solving any optimization problem, and all the biclusters it identifies have no imperfections. The parallel biclustering algorithm consists of two steps: a pre-processing step followed by a bicluster identification step.

The pre-processing step in particular, starts with a data conditioning routine that strictly speaking is not part of the proposed algorithm. Its main purpose is to deal with the noise in the *DNA* microarray data as well as missing values. The actual bicluster identification step consists of two sub-steps. For all valid numbers K of conditions, where $K \geq K_{min}$, and K_{min} is the pre-specified minimum number of conditions in a valid bicluster, the procedure will enumerate all combinations of K conditions from the given N conditions in the *DNA* microarray data that could potentially appear in a valid bicluster. For each subset of K conditions, it then uses a row sort procedure that allows focusing on the coherent evolutions of gene expression levels, rather than the raw or processed expression levels. The output of this step is a matrix that contains the rank of each of the K conditions for each row (gene) when the expression levels of each gene are ordered in a non-decreasing manner. Finally, the main bicluster identification routine identifies all valid coherent evolution patterns involving all genes and a set of K conditions *simultaneously* through a fast row sorting procedure. Note that this allows the algorithm to identify all the possible valid biclusters *without* an exhaustive enumeration of all possible $K!$ permutations of the K conditions. The procedure will also yield biclusters of genes where a subset of genes are coherently up-regulated and another subset coherently down-regulated across the K conditions. We refer the reader to [13] for more information.

3. DIFFERENTIAL BICLUSTERING ALGORITHM

3.1 Definition

Two biclusters $B_k = \{I_k, J_k\}$ and $B_l = \{I_l, J_l\}$ are said to overlap if $B_k \cap B_l \neq \emptyset$, that is: $I_k \cap I_l \neq \emptyset$ and $J_k \cap J_l \neq \emptyset$, where Ø is the empty set. The cardinality of a set, denoted by the operator *Card (.)*, is the number of its elements. Thus $Card(\emptyset) = 0$. The overlapping coefficient S_{kl} of two biclusters $B_k = \{I_k, J_k\}$ and $B_l = \{I_l, J_l\}$ can be defined using the following equation: $S_{kl} = Card(I_k \cap I_l) \times Card(J_k \cap J_l)$. S_{kl} can be viewed as the area covered by the intersection of B_k, and B_l. If $l = k$ then, $S_{kl} = S_{kk} = Card(I_k \cap I_k) \times Card(J_k \cap J_k) = Card(I_k) \times Card(J_k)$ which is the area covered by the bicluster B_k.

3.2 Differential Biclustering

The Differential biclustering algorithm consists of two steps. The first step consists of using the parallel biclustering algorithm mentioned above and fully developed in [13] to uncover all valid biclusters with coherent evolutions in each set of data. The second step consists of performing a differential analysis on the set of biclusters identified during step one. The differential analysis step allows us to uncover the set of convergent biclusters, the set of divergent biclusters, the set of partially conserved biclusters, and the set of split conserved biclusters (Fig. 1). We shall define mathematically convergent, divergent, partially conserved, and split conserved biclusters below in terms of constraints on the size of the intersections of their corresponding lists of genes and conditions.

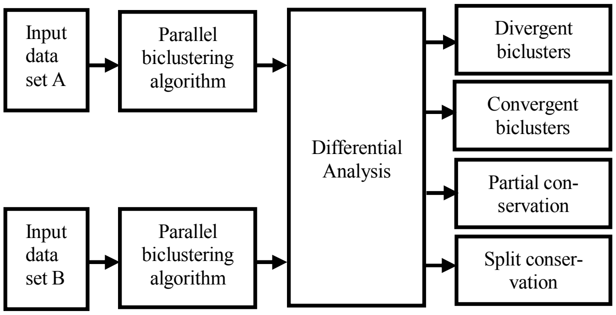

Figure 1: Illustration of the differential biclustering algorithm

Let K be the cardinality of the set of all valid biclusters $B_{Ak} = \{I_{Ak}, J_{Ak}\}$, discovered in the gene expression data matrix of A, and L the cardinality of the set of all valid biclusters $B_{Bl} = \{I_{Bl}, J_{Bl}\}$, discovered in the gene expression data matrix of B, using the parallel biclustering algorithm, with $1 \leq k \leq K$, and $1 \leq l \leq L$. We treat A and B as two different entities. They can correspond to the gene expression datasets of two different or related species (S.cerevisiae and C.albicans for example), datasets coming from the same organism (healthy and disease tissues, for example), or datasets characterizing different types of a disease (epithelial ovarian cancer and clear cell ovarian cancer, for example), and etc. The datasets must contain the expression level of the same genes under the same experimental conditions.

3.1. Convergent Biclusters (full conservation)

A bicluster $B_{Ak} = \{I_{Ak}, J_{Ak}\}$ in A is said to be fully conserved in B, if there exists a bicluster $B_{Bl} = \{I_{Bl}, J_{Bl}\}$ in B such that the following relations are true: $B_{Ak} \cap B_{Bl} = B_{Ak}$, that is, $I_{Ak} \cap I_{Bl} = I_{Ak}$ and $J_{Ak} \cap J_{Bl} = J_{Ak}$. Likewise, a bicluster $B_{Bl} = \{I_{Bl}, J_{Bl}\}$ in B is said to be fully conserved in A, if there exists a bicluster $B_{Ak} = \{I_{Ak}, J_{Ak}\}$ in A such that the following relations are true. $B_{Bl} \cap B_{Ak} = B_{Bl}$, that is: $I_{Bl} \cap I_{Ak} = I_{Bl}$ and $J_{Bl} \cap J_{Ak} = J_{Bl}$

As illustrated in figure (2), this set of biclusters corresponds to some of the patterns that stay co-regulated in both entities. They represent the subsets of genes that continue to work together in A and in B under the same subsets of conditions. This can indicate the conservation of some biological functions, or of some genetic pathways. Therefore, they can be used to understand how both entities are related.

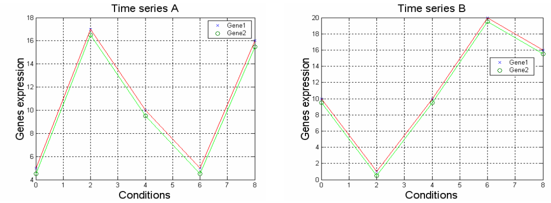

Figure 2: illustration of the fully convergent biclusters

3.2. Divergent Biclusters

A bicluster $B_{Ak} = \{I_{Ak}, J_{Ak}\}$ in A is said to be fully divergent in B if for all biclusters $B_{Bl} = \{I_{Bl}, J_{Bl}\}$ in B, $B_{Ak} \cap B_{Bl} = \emptyset$, that is: $I_{Ak} \cap I_{Bl} = \emptyset$ and $J_{Ak} \cap J_{Bl} = \emptyset$. Likewise, a bicluster $B_{Bl} = \{I_{Bl}, J_{Bl}\}$ in B is said to be fully divergent in A if for all biclusters $B_{Ak} = \{I_{Ak}, J_{Ak}\}$ in A, $B_{Bl} \cap B_{Ak} = \emptyset$, that is: $I_{Bl} \cap I_{Ak} = \emptyset$ and $J_{Bl} \cap J_{Ak} = \emptyset$

As illustrated in figure (3), the set of divergent biclusters represents the subsets of genes that are co-regulated in A (or B), and completely not co-regulated in B (or A) under the same subsets of conditions. This can indicate the absence of some biological functions or the suppression of some genetic pathways. Thus they play a significant role in understanding how different the two entities are.

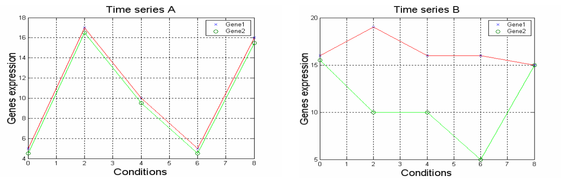

Figure 3: illustration of the fully divergent biclusters

3.3. Partially and Split Conserved Biclusters

The partially conserved biclusters and the split conserved biclusters are illustrated by figure (4) and figure (5) respectively.

A bicluster $B_{Ak} = \{I_{Ak}, J_{Ak}\}$ in A is said to be partially conserved in B if there exist two distinct subsets of conditions in B: J_{Bl1} and J_{Bl2} such that the set of genes I_{Ak} in A are still co-regulated under J_{Bl1} and not co-regulated at all under J_{Bl2} in B, with $J_{Bl1} \cup J_{Bl2} = J_{Ak}$.. In other words, the set of genes I_{Ak} in A are conserved under the set of conditions J_{Bl1} and diverge under the set of conditions J_{Bl2} in B.

Also, a bicluster $B_{Ak} = \{I_{Ak}, J_{Ak}\}$ in A is said to be split conserved in B if there exists L_1 distinct subsets of genes I_{Bl1} and conditions J_{Bl1} in B, such that the set of genes in I_{Bl1} are co-regulated under the set of conditions in J_{Bl1} with $\cup I_{Bl1} = I_{Ak}$, $\cup J_{Bl1} = J_{Ak}$, and $1 \leq l_1 \leq L_1$.

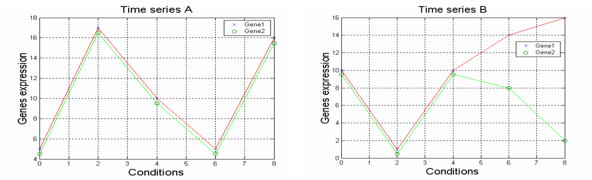

Figure 4: illustration of the partial conserved biclusters

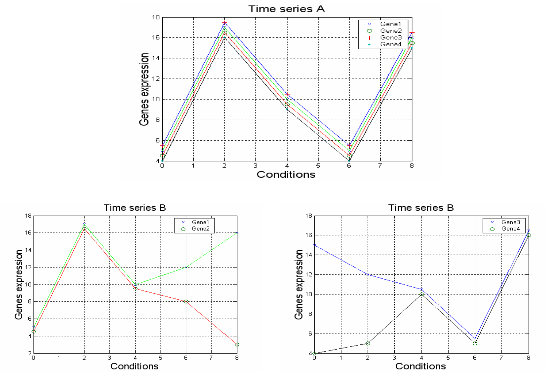

Figure 5: illustration of the split conserved biclusters

4. DIFFERENTIAL BICLUSTERING ALGORITHM OUTPUT MODELING AND IDENTIFICATION

4.1. Output Modelling

From the above definitions one way to model the relationship among the given two sets of biclusters is to construct the bellow differential matrix M of size KxL, of equation (1) below. In equation (1), the rows of M are the set of valid biclusters in A, and the columns of M the set of valid biclusters in B.

$$M = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1l} & \dots & m_{1L} \\ m_{21} & m_{22} & \dots & m_{2l} & \dots & m_{2L} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{k1} & m_{k2} & \dots & m_{kl} & \dots & m_{kL} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{K1} & m_{K2} & \dots & m_{Kl} & \dots & m_{KL} \end{bmatrix}. \qquad (1)$$

The entries of the differential matrix M are positive integers and are defined by equation (2) if $card(I_{Ak} \cap I_{Bl}) \geq I_{min}$ and $card(J_{Ak} \cap J_{Bl}) \geq J_{min}$, and zero otherwise.

$$m_{kl} = Card(I_{Ak} \cap I_{Bl}) x Card(J_{Ak} \cap J_{Bl}) \ . \qquad (2)$$

In equation (2), m_{kl} is the overlapping coefficient or the area covered by the intersection of the two biclusters $B_{Ak} = \{I_{Ak}, J_{Ak}\}$ and $B_{Bl} = \{I_{Bl}, J_{Bl}\}$ considered, and I_{min} and J_{min} the minimum number of genes and conditions in their intersection defined by the user.

Graphically, one can also model the relationship between the two sets of biclusters by a labelled weighted graph, as shown in Fig. (6). The vertices of the graph are the individual biclusters in each set, the edges the relationship that exists between different biclusters, the weight their overlapping coefficient, which indicates by how much they overlap and which correspond to the differential matrix M defined above. The vertices are labelled with the set of genes, conditions, and their respective attributes. The edges are labelled with the set of genes, conditions, and attributes that belong to the intersection of the corresponding two biclusters.

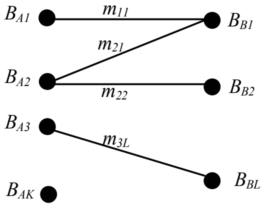

Figure 6: Illustration of the graph representation

4.2. Output Identification

Using the labelled weighted graph modelling approach, one can identify the set of fully convergent, fully divergent, partially, and split conserved biclusters as follows.

4.2.1. Identification of fully Divergent Biclusters

From the above graph modelling approach the set of fully divergent biclusters will correspond to the set of isolated nodes or vertices of the graph, example of B_{AK} in figure (6).

In the differential matrix M defined above, the bicluster B_{Ak} in A will be fully divergent in B if all the elements of the k^{th} row of M are zeros. Likewise, the bicluster B_{Bl} in B will be fully divergent in A if all the elements of the l^{th} column of M are zeros.

4.2.2. Identification of fully convergent Biclusters

The bicluster $B_{Ak} = \{I_{Ak}, J_{Ak}\}$ in A will be fully conserved in B if equation (3) is true.

$$S_{Ak} = Card(I_{Ak}) \times Card(J_{Ak}) = sup(M(k,:)). \qquad (3)$$

In equation (3), S_{Ak} is the area covered by the bicluster $B_{Ak} = \{I_{Ak}, J_{Ak}\}$ in A, $M(k,:)$ is the set of the weight of all the edges that are connected to node or bicluster $B_{Ak} = \{I_{Ak}, J_{Ak}\}$, which is the k^{th} row of the differential matrix M defined above.

Likewise, the bicluster $B_{Bl} = \{I_{Bl}, J_{Bl}\}$ in B will be fully conserved in A if equation (4) is true.

$$S_{Bl} = Card(I_{Bl}) \times Card(J_{Bl}) = sup(M(:,l)). \qquad (4)$$

In equation (4), S_{Bl} is the area covered by the bicluster $B_{Bl} = \{I_{Bl}, J_{Bl}\}$ in B, $M(:,l)$ is the set of the weight of all the edges that are connected to node or bicluster $B_{Bl} = \{I_{Bl}, J_{Bl}\}$, which is the l^{th} column of the differential matrix M defined above.

4.2.3. Identification of Partially and Split Conserved Biclusters

A general approach to identify the set of partially and split conserved biclusters will be to consider it as being the remaining of the set of all biclusters minus the set of fully convergent biclusters and fully divergent biclusters.

5. RESULTS

We applied our approach to tissues provided by the University of Minnesota Cancer Center's Tissue Procurement Facility. Bulk tumor and normal tissues were identified, dissected, and snap-frozen in liquid nitrogen within 15 to 30 minutes of resection from the patient. Tissue sections were made from each sample, stained with hematoxylin and eosin (H&E), and examined independently by two pathologists to confirm the pathological state of each sample. The tissue samples consisted of 50 normal ovaries, 20 serous papillary ovarian carcinoma tumors, 17 metastases of serous papillary ovarian carcinoma to the omentum, and 372 other tissue samples from 21 different sites, such as kidney, breast, or lung. All tissue samples underwent stringent quality control measures to verify the integrity of the *RNA* before use in gene array experiments. Gene expression was determined by Gene Logic Inc. using Affymetrix HU_95 arrays containing 12,000 known genes and 48,000 expressed sequence tags. The gene expression matrix was normalized using Affymetrix (*M.A.S. 4.0.1*), and the log floor data transform with a floor value of 1 was performed. Because of missing values, 5 metastases of serous papillary ovarian carcinoma tissues were removed, and about 74 genes were eliminated because they all had missing values. Thus the final gene expression matrix used for simulation contained: 44 normal ovaries, 17 serous papillary ovarian carcinoma tumors, and 15 metastases of serous papillary ovarian carcinoma to the omentum, about 12626 genes among which ~12000 are known genes.

We combined the biclustering technique with a sensitivity analysis of the results by varying the thresholds used in the approach to define divergent biclusters. We identified 55 upregulated in ovarian cancer tissues compared to normal ovarian tissue and the other 372 non-ovarian tissues. This set included all 40 candidate biomarkers listed in [15]. We also identified 25 genes that were downregulated in ovarian cancer tissue and the other 372 non-ovarian tissues compared to normal ovarian tissue. Note that this category has never been studied before. The well separated histograms of the gene expression patterns in normal and non-ovarian tissues and cancerous ovarian tissues of many of the newly identified candidate genes make them more promising biomarkers than previously reported candidates. Immunohistochemistry analysis and reverse transcriptase polymerase chain reaction screening of all candidate biomarkers are in progress and will be reported at the conference.

6. CONCLUSION

In this study, we introduce and develop a differential biclustering algorithm for local comparison of gene expression data. The proposed algorithm can be used to identify similarities and differences among related organisms or to identify group of genes that are involved in the development of some disease. Analysis of a set of normal and cancer ovarian data using the proposed algorithm shows interesting patterns that are either specific to normal ovarian tissue or ovarian cancer tissue. Investigation of the differences and similarities of the S.cerevisiae and the C.albicans which are two related organism using our differential biclustering algorithm approach is currently under progress.

7. REFERENCES

[1]- Alter O, Brown PO, Botstein D (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression datasets of two different organisms. Proc Natl Acad Sci USA 100: 3351–3356.

[2]- Stuart JM, Segal E, et al. (2003) A gene coexpression network for global discovery of conserved genetic modules. Science 302: 249–255.

[3]- Bergmann S, Ihmels J, Barkai N (2004) Similarities and differences in genome-wide expression data of six organisms. PLoS Biol 2: e9. DOI: 10.1371/journal.pbio.0020009.

[4]- McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, et al. (2004) Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nat Genet 36: 197–204.

[5]- Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, et al. (2004) The diploid genome sequence of Candida albicans. Proc Natl Acad Sci U S A 101:7329–7334.

[6]- Heckman DS, Geiser DM, Eidell BR, Staffer RL, Kardos NL, et al. (2001) Molecular evidence for the early colonization of land by fungi and plants. Science 293: 1129–1133.

[7]- Karababa M, Coste AT, Rognon B, Bille J, Sanglard D (2004) Comparison of gene expression profiling between Candida albicans azole-resistant clinical isolates and laboratory strains exposed to drugs inducing multidrug transporters. Antimicrob Agents Chemother 48: 3064–3079.

[8]- Enjalbert B, Nantel A, Whiteway M (2003) Stress-induced gene expression in Candida albicans: Absence of a general stress response. Mol Biol Cell 14:1460–1467.

[9]- Bennett RJ, Uhl MA, Miller MG, Johnson AD (2003) Identification and characterization of a Candida albicans mating pheromone. Mol Cell Biol 23:8189–8201.

[10]- Fradin C, De Groot P, MacCallum D, Schaller, et al. (2005) Granulocytes govern the transcriptional response, morphology, and proliferation of Candida albicans in human blood. Mol Microbiol 56:397–415.

[11]- Lorenz MC, Bender JA, Fink GR (2004) Transcriptional response of Candida albicans upon internalization by macrophages. Eukaryot Cell 3: 1076–1087.

[12]-Jan Ihmels, Sven Bergmann, Judith Berman, Naama Barkai, Comparative Gene Expression Analysis by a Differential Clustering Approach: Application to the Candida albicans Transcription Program, PLOS Genetics, Septembre 2005, volume 1, issue 3, e39.

[13] - A. H. Tewfik, A. B. Tchagang, and L. Vertatschitsch " Parallel Identification of Gene Biclusters with Coherent Evolution", IEEE Transaction on Signal Processing, Special Issue on Genomics Signal Processing, accepted for publication June 2006.

[14] - S. C. Madeira, A. L. Oliveira, "Biclustering Algorithms for Biological Data Analysis: A Survey", IEEE Transactions on computational Biology and Bioinformatics, Vol. 1, No. 1, Jan-March 2004.

[15]- Hibbs, K., Skubitz, K.M. Pambuccian, S., Casey, R.C., Burleson, K.M, Oegema, T., Jr., Thiele, J.J., Grindle, S.M., Bliss, R., and Skubitz, A.P.N. (2004) Differential gene expression in ovarian carcinoma: Identification of potential biomarkers. *American Journal of Pathology* 165(2):397-414.