Classification of Gene Expression Profiles: Comparison of K-means and Expectation Maximization Algorithms

Cristina Rubio-Escudero
Dpto. Lenguajes y Sistemas Informáticos
Universidad de Sevilla
crubioescudero@us.es

Francisco Martínez-Álvarez
Área de Lenguajes y Sistemas Informáticos
Universidad Pablo de Olavide
fmaralv@upo.es

Rocío Romero-Zaliz
DECSAI
Universidad de Granada
rocio@decsai.ugr.es

Igor Zwir
DECSAI
Universidad de Granada
zwir@decsai.ugr.es

Abstract

Biomedical research has been revolutionized by high-throughput techniques and the enormous amount of data they are able to generate. In particular technology has the capacity to monitor changes in RNA abundance for thousands of genes simultaneously. The interest shown over microarray analysis methods has rapidly raised. Clustering is widely used in the analysis of microarray data to group genes of interest targeted from microarray experiments on the basis of similarity of expression patterns. In this work we apply two clustering algorithms, K*-means and Expectation Maximization to particular a problem and we compare the groupings obtained on the basis of the cohesiveness of the gene products associated to the genes in each cluster.*

1. Introduction

Advances in molecular biology and new computational techniques permit the systematical study of molecular processes that underlie biological systems [8]. Particularly, microarray technology has revolutionized modern biomedical research by its capacity to monitor changes in RNA abundance for thousands of genes simultaneously [3]. The greatest challenge in microarray technology development is the analytical process followed to successfully analyze data acquired from microarray experiments.

Once the genes of interest from DNA microarray data are targeted a common first step is to cluster the data on the basis of similarity of expression patterns since genes the same expression patterns are likely to be involved in the same regulatory processes [14]. Though in theory there is a big step from simple correlation analysis to gene interaction networks, several papers indicate that the clustering of gene expression data does result in groups of genes that have related functions [6]. Therefore, clustering genes of known functions with poorly characterized genes provides a means of gaining insights into the functions of the latter [1]. However, the extent to which clustering reveals useful information about the system under study depends on the extent to which the clustering method successfully groups intrinsically related elements. A remedy then is to integrate known biological knowledge into the clustering procedure itself. One of the most widely used sources of biological knowledge is the Gene Ontology project [2], which stores one of the most powerful characterization of genes based on gene products.

In this work we compare the results of two different clustering algorithms when grouping microarray data: the K-means [7], a classic clustering algorithm based on Euclidean distance which is widely used on data from microarray experiments [9], and the Expectation Maximization (*EM*), proposed by Lauritzen in 1995 [10] as a variation the K-means. The main novelty of this technique is to obtain the previously unknown *Probability Distribution Function* (PDF) [15] of the complete dataset. The algorithms will be applied to an specific problem, the inflammation and host response to injury in humans. Clusters resulting from both algorithms are compared based on the cohesiveness of the gene products associated to the genes in each cluster.

To obtain such gene product information we will use an algorithm termed *EMO-CC* (Evolutionary Multi-Objective Conceptual Clustering), proposed in Romero-Zaliz *et al.,* [11], which retrieves meaningful substructures from network databases using multi-objective and multi-modal opti-

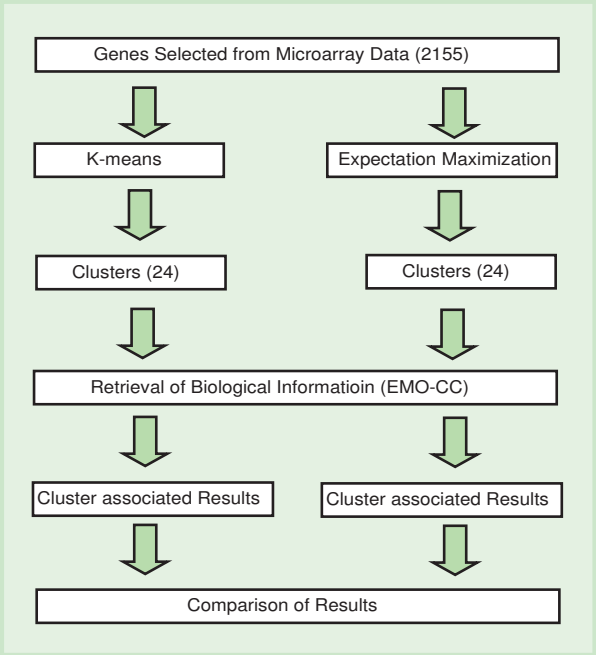

Figure 1. Outline of the work structure.

mization techniques. We will compare the results obtained by application of both *K*-means and Expectation Maximization in their specificity and their capability to retrieve immuno-inflammatory related terms.

This paper is structured as follows. In section 2 we describe the experiment under study. In section 3 we the algorithms applied in this study and we can find in section 4 the results obtained by application of the algorithms. Section 5 highlights the conclusion obtained from this work. An outline of steps followed in the preparation of this paper the paper structure can be seen in Fig.1

2. Problem Description: Inflammation and the Host Response to Injury

The problem under study deals with inflammation and the host response to injury. Understanding the inflammation process is critical because the body uses inflammation to protect itself from infection or injury (e.g., crushes, massive bleeding, or a serious burn). The host response to trauma and burns is a collection of biological and pathological processes that depends critically upon the regulation of the human immuno-inflammatory response [4].

The data were acquired from blood samples collected from eight human volunteers, four treated with intravenous endotoxin (i.e., patients 1 to 4) and four with placebo (i.e., patients 5 to 8). Complementary RNA was generated from circulating leukocytes at 0, 2, 4, 6, 9 and 24 hours with GeneChips HG-U133A v2.0 from Affymetrix Inc., which contains 22216 probe sets (a probe set codes part of a gene), analyzing the expression level of 18400 transcripts and variants, including 14500 well-characterized human genes. Analysis of the set of gene expression profiles obtained from this experiment is complex, given the number of samples taken and variance due to treatment, time, and subject phenotype. Therefore, we believe this problem is typical and informative as a microarray case study.

3. Methods

We compare the results of two different clustering algorithms when grouping microarray data: the *K*-means [7], a classic clustering algorithm and the Expectation Maximization (EM) algorithm, proposed by Lauritzen in 1995 [10] as a variation the *K*-means. They are applied to the inflammation and host response to injury in humans described in Section 2. The comparison is based on the coherence of the clusters obtained from both algorithms with the gene products associated to the records in each cluster.To obtain such gene product information we use an algorithm termed *EMO-CC* (Evolutionary Multi-Objective Conceptual Clustering), proposed in Romero-Zaliz *et al.,* [11].

One of the main steps in microarray data analysis is to decide the set of probe sets (a probe set codes part of a gene) from the ones analyzed by the microarray which show a significant behavior for the problem under study. In our particular problem, this set is made out of probe sets which exhibit different behavior between the treatment and control experimental conditions, among subjects and among time points. From the 22216 probe sets present in the GeneChips HG-U133A v2.0 microarray, 2155 are selected as the set of probe sets exhibiting different behavior between treatment and control experimental conditions, among subjects and among time points [12, 13]. Therefore, any further analysis (clustering probe sets) will be made over this dataset only taking into account the treatment group (subjects inoculated with an endotoxin).

3.1. *K*-means Clustering Algorithm

We apply a classic clustering algorithm, *K*-means [7], for identification of gene expression patterns in the inflammation problem data set. The *K*-means clustering can be described as a partitioning method which groups the observations in your data into non-overlapping clusters. The algorithm runs an iterative process, where each record is assigned to the closest centroid. New centroids are calculated for the resulting clusters and the records are reassigned to the closest centroid. The process automatically stops once a steady state has been reached.

The similarity measure chosen has been the Euclidean Distance, a classical distance measure, since distance measures have exhibited a better behavior than correlation based measures for gene grouping [12, 13].

3.2. Expectation Maximization Clustering Algorithm

The Expectation Maximization (*EM*) algorithm, proposed by Lauritzen in 1995 [10], is a variation the *K*-means. The main novelty of this technique is to obtain the previously unknown *Probability Distribution Function* (PDF) [15] of the complete dataset.

This PDF can be approximated as a linear combination of NC components, defined from certain parameters $\Theta = \cup\Theta_j, \forall_j = 1...NC$ that have to be found.

$$P(x) = \sum_{j=1}^{NC} \pi_j p(x; \Theta_j) \quad (1)$$

$$\sum_{j=1}^{NC} \pi_j = 1 \quad (2)$$

where π_j are the *a priori* probability of each cluster, $P(x)$ denotes the arbitrary PDF and $p(x; \Theta_j)$ the PDF of each j component. Each cluster corresponds to its data samples, which belong to a single density that are combined. PDF of arbitrary shapes can be estimated by using *T*-Student, Bernouilli, Poisson, normal or log-normal functions. In this research, the normal distribution has been used as shape of the PDF.

3.3. EMO-CC

We make use of a conceptual clustering methodology termed *Evolutionary Multi-Objective Conceptual Clustering (EMO-CC)*, [11] relying on the NSGA-II multi-objective (MO) genetic algorithm [5], that focuses primarily on the discovery of objects identified by their most representative features lying in the set of all optimal solutions of a multi-objective optimization problem. We apply this methodology to identify conceptual models in structural databases generated from gene ontologies. These models can explain and predict phenotypes in the inflammation and host response to injury problem, similar to models provided by gene expression or other genetic markers.

We apply *EMO-CC* to the *Gene Ontology* database (i.e., the GO Project, [2]) *EMO-CC* to recover optimal substructures containing genes sharing a common set of terms, which are defined at different levels of specificity and correspond to different networks. Therefore, each cluster obtained by application of *K*-means or Expectation Maximization algorithm is used as a query for applying the *EMO-CC* algorithm on the Gene Ontology database. We retrieve substructures containing gene related terms shared by members of the cluster and comparison of *K*-means and Expectation Maximization is based on such substructures, both in their specificity and in their capability to retrieve immuno-inflammatory related terms.

4. Results

The two clustering algorithms compared in this paper need the number of resulting clusters, k, as an input parameter. This number k has been estimated for the treatment group in the inflammation and host response to injury dataset (see Section 2) was calculated by means of the silhouette function. Its maximum mean value $(0, 47)$ was reached when $k = 24$ after evaluating $k = 1...120$.

The analysis of the relation between clusters obtained applying *K*-means and clusters obtained applying *EM* can be seen in Table 1. We can see how some of the levels of intersection between *K*-means and *EM* are very high for some clusters: *K*-means #12 and *EM* #22 have an intersection level of 97,06%, only one probe set out of 34 has been classified different by them. The same happens with *K*-means #24 and *EM* #8, with an intersection level of 95,24%, with only two probe sets out of 42 classified differently. A total of 16 clusters out of the 24 are similarly classified with an intersection level greater than 70%. However, some of the clusters show low levels of intersection, as it happens with *K*-means #14 with 53 probe sets, which is classified by *EM* in clusters #3(15 probe sets), #13(15 probe sets), #20(4 probe sets) and #23 (19 probe sets). The same situation applies to *K*-means #6 with 145 probe sets, which is classified by *EM* in clusters #1(49 probe sets), #5(65 probe sets), #9(30 probe sets) and #23 (19 probe sets). In this case partition of *K*-means #6 by *EM* has resulted in three groups with a significant number of probe sets (33,8% 44,82% and 20,7% of the 145 respectively). It is noteworthy how the probe sets obtained in three groups obtained from *K*-means are clustered together with probe sets from other groups by *EM*.

The comparison of both clustering techniques has been made applying the *EMO-CC* algorithm. Each cluster obtained by application of *K*-means or Expectation Maximization algorithm is used as a query for applying the *EMO-CC* algorithm on the Gene Ontology database [2]. We have retrieved substructures containing gene related terms shared by members of the cluster and comparison of *K*-means and Expectation Maximization is based on such substructures, both in their specificity and in their capability to retrieve immuno-inflammatory related terms. To perform the comparison we have considered two different types of cluster classification of *K*-means in relation to *EM*. On the one hand, clusters with a high percentage of genes classified together and only some of them are excluded from the main

	Expectation Maximization																									
***K*-means**		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**14**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**Intersection**
	1	25	16	0	255	7	0	14	0	0	1	0	0	0	0	0	5	0	4	0	0	0	0	0	0	77,98%
	2	0	0	0	14	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	104	75,36%
	3	3	26	0	0	0	0	0	0	0	0	0	0	0	0	0	140	0	2	0	0	0	4	0	0	80,00%
	4	1	0	0	2	0	0	38	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	88,37%
	5	207	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	0	6	0	0	0	15	0	0	89,22%
	6	49	0	0	0	65	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	44,83%
	7	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	6	2	0	55,56%
	8	15	0	0	0	0	0	0	0	32	2	0	0	0	0	0	0	0	1	0	6	0	54	1	0	48,65%
	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	1	0	5	0	40	44	0	44,44%
	10	0	0	0	0	46	0	0	0	2	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	86,79%
	11	0	0	0	0	0	0	4	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	85,19%
	12	0	0	0	0	0	0	0	0	73	2	3	0	0	0	25	0	0	0	0	11	0	0	0	0	64,04%
	13	0	0	0	0	0	0	0	0	0	0	0	0	10	0	5	0	0	0	0	71	0	0	2	0	80,68%
	14	0	0	15	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	4	0	0	19	0	35,85%
	15	0	0	0	0	0	0	0	0	0	0	66	0	0	0	4	0	0	0	4	0	0	0	0	0	89,19%
	16	0	0	1	0	0	0	0	0	0	0	0	0	26	1	59	0	0	0	0	1	2	0	0	0	65,56%
	17	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	1	5	0	0	0	0	0	0	73,91%
	18	0	0	0	0	0	0	0	0	0	3	0	0	0	0	14	0	0	0	0	0	2	0	0	0	73,68%
	19	0	0	1	0	0	0	0	0	0	0	0	0	10	32	0	0	9	0	0	0	0	0	0	0	61,54%
	20	0	0	0	0	0	0	0	0	0	0	7	4	0	0	0	0	0	0	54	0	5	0	0	0	77,14%
	21	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	43	0	0	0	84,31%
	22	0	0	0	0	0	0	0	1	0	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	97,06%
	23	0	0	0	0	0	34	0	0	0	0	0	0	0	3	0	0	4	0	0	0	0	0	0	0	82,93%
	24	0	0	0	0	0	2	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95,24%

Table 1. Relation of the clusters obtained by application of K-means and Expectation Maximization.

group, as it the case of *K*-means #5 and *EM* #1 and *K*-means #17 and *EM* #3, where 89,22% and 73,91% of the probe sets are grouped together respectively. We are also interested in clusters obtained by application of *EM* grouping probe sets from different *K*-means clusters, as it is the case of *K*-means #12 and #16 and *EM* #14.

For each of the clusters compared we now show tables containing the common substructures obtained by each of the clusters including the gene product terms classified in Biological Process, Molecular Function and Cellular component with their relatives specificity values (specificity as a *P*-value measured between 0 and 1, with acceptable when smaller than $0,05$ and with better values when closer to 0).

Comparison of clusters *K*-means #5 and *EM* #1 has been summarized in Table 2. We can see how for common substructures (substructures containing the same or almost the same gene products), the specificity values obtained by clusters grouped by *EM* are better (closer to 0) than the ones obtained by clusters grouped by *K*-means. For instance, the substructure associated to protein metabolism and regulation of biological process has an specificity value of $8,02E-07$ for the *EM* and a value of $0,01602896$ for the *K*-means. However, the majority of substructures retrieved by *EM* are also retrieved by *K*-means.

We can see the results of comparing *K*-means #17 and *EM* #3 in Table 3

The situation is very similar to the comparison of *K*-means #5 and *EM* #1. All substructures retrieved by *EM* are also retrieved by *K*-means but the first obtains better results regarding the specificity values. We can see how the majority of substructures retrieved by *K*-means have worst (higher) specificity values than substructures retrieved by *EM*. However, there are some cases where *K*-means clusters have better specificity values than *K*-means. That is the case of the substructure related to signal transduction, with a specificity value of $1,20E-05$ *K*-means #12 and $4,16E-06$ for *K*-means #16 and a value of $0,001010838$ for *EM* #15. It is also remarkable the fact that there are some interesting substructures which are retrieved by *EM* and not by *K*-means clusters, as it is the case of the substructure associated to ion transport, regulation of biological process and immune response or the ones associated to regulation of transcription.

The last comparison to show, *K*-means #12-#16 and *EM* #14, is summarized in Table 4.

5. Conclusions

In this work we have compared the results of two different clustering algorithms when grouping microarray data: the *K*-means [7], a classic clustering algorithm based on Euclidean distance which is widely used on data from microarray experiments [9] and the Expectation Maximization algorithm (*EM*), proposed by Lauritzen in 1995 [10] as a variation the *K*-means. They have been applied to the inflammation and host response to injury in humans (see Section 2). The comparison has been based on the of the clusters obtained from both algorithms with the gene products associated to the records in each cluster. To obtain such gene product information we have applied the *EMO-CC* (Evolutionary Multi-Objective Conceptual Clustering), proposed in Romero-Zaliz *et al.,* [11], which retrieves meaningful substructures from network databases using multi-objective and multi-modal optimization techniques.

Clusters have been obtained for the treatment group of the experiment under study by application of both algorithms, resulting the optimal number of cluster $k = 24$ calculated by means of the silhouette function. In Table 1 we can see the relation between the two clusterings performed. Some of the levels of intersection between *K*-means and *EM* resulted very high for some of the clusters, and therefore the substructures retrieved by them from the gene product

Biological Process	Molecular Function	Cellular Component	Specificity	Cluster
immune response	binding antioxidant activity	cellular component	0,01286757	EM #1
immune response	binding antioxidant activity	cellular component	0,04208912	Km #5
protein metabolism regulation of biological process	nucleotide binding protein binding	cellular component	8,02E-07	EM #1
protein metabolism regulation of biological process	nucleotide binding protein binding	cellular component	0,01602896	Km #5
protein targeting	enzyme regulator activity protein binding	intracellular	1,18E-05	EM #1
protein targeting	enzyme regulator activity protein binding	intracellular	0,000101943	Km #5
protein targeting regulation of transcription DNA-dependent immune response	molecular function	integral to membrane cell fraction	3,97E-05	EM #1
protein targeting regulation of transcription DNA-dependent immune response	molecular function	integral to membrane cell fraction	0,006271776	Km #5

Table 2. Comparison of cluster K-means (#5) and Expectation Maximization (#1).

Biological Process	Molecular Function	Cellular Component	Specificity	Cluster
protein metabolism cellular process	translation regulator activity	cellular component	2,22E-07	EM #3
protein metabolism cellular process	translation regulator activity	cellular component	0,007641355	Km #17
protein targeting	enzyme regulator activity protein binding	intracellular	1,18E-05	EM #3
protein targeting	catalytic activity protein binding zinc ion binding	cellular component	0,04743564	Km #17
protein transport	catalytic activity; nucleotide binding	cytoplasm	0,000976312	EM #3
protein transport	catalytic activity nucleotide binding	cytoplasm	0,009011493	Km #17
regulation of biological process	transporter activity transition metal ion binding catalytic activity	cellular component	3,16E-05	EM #3
regulation of biological process	transporter activity transition metal ion binding catalytic activity	cellular component	0,04729886	Km #17
signal transduction protein targeting regulation of biological process	catalytic activity nucleotide binding protein binding	cellular component	1,05E-06	EM #3
signal transduction protein targeting regulation of biological process	catalytic activity nucleotide binding protein binding	cellular component	0,000795458	Km #17

Table 3. Comparison of cluster K-means (#17) and Expectation Maximization (#3).

Biological Process	Molecular Function	Cellular Component	Specificity	Cluster
immune response	antioxidant activity protein binding	cellular component	0,0174212	EM #15
immune response	signal transducer activity protein binding	extracellular region	0,0374719	Km #16
protein metabolism	molecular function	cellular component	5,53E-05	EM #15
protein metabolism immune response	molecular function	cellular component	3,30E-06	Km #12
protein metabolism immune response	molecular function	cellular component	8,96E-08	Km #16
protein folding	nucleotide binding catalytic activity RNA binding	cellular component	0,04578355	EM #15
cell communication	catalytic activity nucleotide binding transition metal ion binding	cellular component	0,02786016	EM #15
signal transduction	binding; catalytic activity	integral to membrane	0,001010838	EM #15
protein transport	catalytic activity nucleotide binding	cytoplasm	0,004275726	EM #15

Table 4. Comparison of cluster K-means (#12 and #16) and Expectation Maximization (#15).

database are almost the same. A total of 16 clusters out of the 24 were similarly classified with an intersection level greater than 70%. However, some of the clusters showed low levels of intersection.

In order to compare the clusters we considered two different types of cluster classification of *K*-means in relation to *EM*. On the one hand, clusters with a high percentage of genes classified together and only some of them excluded from the main group, as it the case of *K*-means #5 and *EM* #1 and *K*-means #17 and *EM* #3. We can see in Table 2 and Table 3 respectively how the substructures retrieved by both of them are very similar and the specificity levels obtained by *EM* clusters are higher than the ones obtained by *K-means*. Therefore *EM* is removing probe sets which are not related to the other records in the cluster. On the other hand we are interested in clusters obtained by application of *EM* grouping probe sets from different *K*-means clusters, as it is the case of *K*-means #12 and #16 and *EM* #14. We can see how in general the substructures retrieved by *EM* have better specificity values and furthermore, new substructures are retrieved by *EM* with good levels of specificty (less than $0,05$). Therefore we can conclude that *EM* is breaking up clusters created by *K*-means with successful results in the gene product related information.

Finally, we can conclude that the Expectation Maximization algorithm performs better than the *K*-means algorithm for the analysis of microarray data as seen in the comparison related to gene product information.

6. Acknowledgments

This work was supported in part by the Spanish Ministry of Science and Technology under project TIN-2006-12879 and in part by The Consejeria de Innovacion, Investigacion y Ciencia de la Junta de Andalucia under project TIC-02788.

References

[1] D. Allison, X. Cui, G. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7:55–65, 2006.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, 2000. 1061-4036 Journal Article.

[3] P. Brown and D. Botstein. Exploring the new world of the genome with dna microarrays. *Nature Genet.*, 21(Suppl.):33–37, 1999.

[4] S. E. Calvano, W. Xiao, D. R. Richards, R. M. Felciano, H. V. Baker, R. J. Cho, R. O. Chen, B. H. Brownstein, J. P. Cobb, S. K. Tschoeke, C. Miller-Graziano, L. L. Moldawer, M. N. Mindrinos, R. W. Davis, R. G. Tompkins, S. F. Lowry, and I. A. Large Scale Collab Res Program. A network-based analysis of systemic inflammation in humans. *Nature*, 437, 2005.

[5] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6:182–197, 2002.

[6] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.

[7] R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* 1973.

[8] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[9] A. Guiller, A. Bellido, A. Coutelle, and L. Madec. Spatial genetic pattern in the land helix aspersa inferred from a 'centre-based clustering' procedure. *Genet Res*, 88(1):27–44, 2006.

[10] S. L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19(2):191–201, 1995.

[11] R. C. Romero-Zaliz, C. Rubio-Escudero, J. P. Cobb, F. Herrera, C. O., and I. Zwir. A multi-objective evolutionary conceptual clustering methodology for gene annotation within structural databases: A case of study on the gene ontology database. *IEEE Transactions on Evolutionary Computation*, page Accepted for future publication.

[12] C. Rubio-Escudero. *Fusion of Knowledge towards Identification of Genetic Profiles in the Systemic Inflammation Problem. Ph.D Thesis*. Unversidad de Granada, 2007.

[13] C. Rubio-Escudero, O. Harari, O. Cordn, and I. Zwir. Modeling genetic networks: comparison of static and dynamic models. In S. B. Heidelberg., editor, *5TH European Conference on evolutionary computation, machine learning and data mining y bioinformatics.*, volume 4447/2007, pages 78–89, Valencia, Spain, 2007. Springer.

[14] M. Tan, E. Smith, J. Broach, and C. Floudas. Microarray data mining: A novel optimization-based approach to uncover biologically coherent structures. *BMC Bioinformatics*, in press.

[15] S. Zacks. *The theory of statistical inference*. Wiley, 1946.