Decision Making Association Rules for Recognition of Differential Gene Expression Profiles

C. Rubio-Escudero[1], Coral del Val[1], O. Cordón[1,2], and I. Zwir[1,3]

[1] Department of Computer Science and Artificial Intelligence, University of Granada, Spain
[2] European Center for Soft Computing, Mieres, Spain
[3] Howard Hughes Medical Institute, Washington University School of Medicine, St. Louis, MO
{crubio, delval, ocordon, zwir}@decsai.ugr.es

Abstract. The rapid development of methods that select over/under expressed genes from RNA microarray experiments have not yet satisfied the need for tools that identify differential profiles that distinguish between experimental conditions such as time, treatment and phenotype. We evaluate several microarray analysis methods and study their performance, finding that none of the methods alone identifies all observable differential profiles, nor subsumes the results obtained by the other methods. Therefore, we propose a machine learning based methodology that identifies and combines the abilities of microarray analysis methods to recognize differential profiles. We encode the results of this methodology in decision making association rules able to decide which method or method-aggregation is optimal to retrieve a set of genes exhibiting a common profile. These solutions are optimal in the sense that they constitute partial ordered subsets of all method-aggregations bounded by the most specific and the most sensitive available solution. This methodology was successfully applied to a study of inflammation and host response to injury data set derived from the analysis of longitudinal blood microarray profiles of human volunteers treated with intravenous endotoxin compared to placebo. Our approach was able to uncover a cohesive set of differentially expressed genes and novel members exhibiting previously studied differential profiles. This guideline serves as a means to support decisions on new microarray problems.

1 Background

Advances in molecular biology and computational techniques permit the systematical study of molecular processes that underlie biological systems [1]. Particularly, microarray technology has revolutionized modern biomedical research by its capacity to monitor changes in RNA abundance for thousands of genes simultaneously [2].

To address the statistical challenge of analyzing these large data sets, new methods have emerged ([3], [4], [5], [6], [7]). However, there is a dearth of computational methods to facilitate understanding of differential gene expression profiles (e.g., profiles that change over time and/or over treatments and/or over patients) and to decide which is the most reliable method to identify differences across profiles.

We develop a detailed evaluation of the performance of several commonly used statistical methods to identify differential expression profiles. We found that the application of these methods return different results applied over the same set of data: the methods do not identify all observable differential profiles (genes exhibiting a

E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 1137 – 1149, 2006.

common behavior throughout experimental conditions). Moreover, none of the methods subsume the results obtained by the other methods.

Our study reveals how some methods are able to recognize some differential profiles and not others and that some of the not retrieved profiles might contain significant genes for the experiment under study. Therefore, we propose a methodology that combines the properties of each method into a set of decision making association rules ([8], [9], [10]) devoted to discover optimal aggregations of microarray analysis methods in an effort to identify differential gene expression profiles. The association rules allow users to query for the most appropriate method or aggregation of them to retrieve significant genes based on the differential profiles they exhibit.

To create such set of decision association rules we perform the following steps over a set of microarray gene expression data (Fig. 1). First, we extract from the data set all genes which behave in a different way from one experimental condition to the others (i.e., genes that change over time, treatments and phenotype). We apply several classical microarray analysis methods (T-Tests [11], Permutation Tests [6], Analysis of Variance [5] and Repeated Measures ANOVA [12]). Second, we create a database containing distinct types of differential profiles over time, experiment and subjects from previously recovered genes.

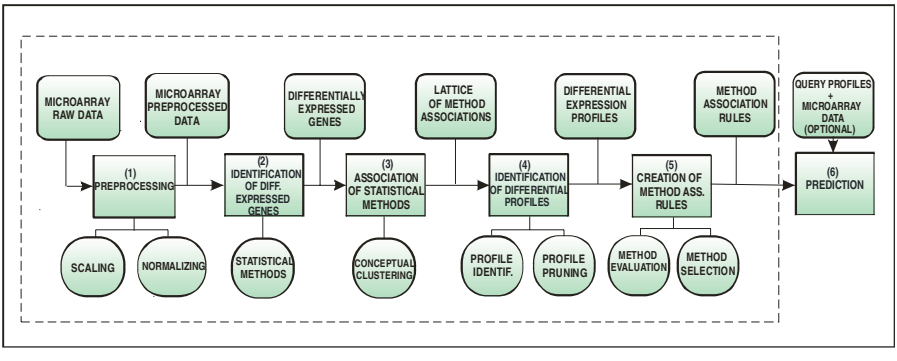

Fig. 1. Graphical representation of the methodology. The squared boxes represent the phases of the methodology, the round cornered boxes correspond to the input/output data at each step, and the ellipses the operations performed at each phase.

Third, we create decision making association rules, where the antecedents are differential profiles and the consequents are methods or aggregations of them capable to identify the profiles. Fourth, we arrange the association rules into a lattice, where the rules are ordered from the most general (top) to the most specific solution (bottom). We use this structure to evaluate the performance of the rules by analyzing their specificity, sensitivity and cost, applying multiobjective optimization techniques. Fifth, we use a selected set of optimal rules as a framework to support new decisions about the applicability of microarray analysis methods to retrieve differential gene expression profiles.

2 Results

The results are obtained from the application of our procedure to a data set derived from longitudinal blood expression profiles of human volunteers treated with intravenous endotoxin compared to placebo. The motivation of these experiments is to provide insight to the host response to injury as part of a Large-scale Collaborative Research Project sponsored by the National Institute of General Medical Sciences (www.gluegrant.org) [13]. 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. The data were acquired from blood samples collected from eight normal 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 after the i.v. infusion and hybridized with GeneChips® HG-U133A v2.0 from Affymetrix Inc., which contains 22216 probe sets, analyzing the expression level of 18400 transcripts and variants, including 14500 well-characterized human genes.

2.1 Accuracy of the Statistical Methods

We investigate the performance of several commonly used statistical methods in identifying differential expression profiles that change over time, treatments and phenotype. We name T-Test as M^1, T-Test considering time as M^2, Permutation Test as M^3, Permutation Test considering time as M^4, ANOVA over treatment as M^5, ANOVA over time as M^6, ANOVA over treatment and time as M^7, RMANOVA over treatment as M^8, RMANOVA over time as M^9 and RMANOVA over treatment and time as M^{10}, where considering time refers to the fact that the tests have been specifically applied to find differences between time points. For our set of data, we found that these methods do not identify all observable distinct profiles. Moreover, none of them subsumes the results obtained by other methods (Table 1). Different methods retrieve different amounts of probe sets (e.g., the application of M^1 over the microarray dataset retrieves 962 genes as differentially expressed, whereas M^5 retrieves 1734 genes, and M^3 retrieves 612 genes). The concordance rates between the sets of genes retrieved also varies widely, indicating that none of the methods subsumes the others (Table 1)(e.g., from the genes retrieved by M^3, only 31.11% are also retrieved by M^5, and 52.29% by M^1.

2.2 Statistical Methods and Differential Profiles

We found that there is a relationship between the statistical methods and the differential profiles they are able to identify, having differential profiles identified by some methods and not by others. This type of relation is what we encode in the set of decision making association rules that we obtain from the application of our methodology. In our particular problem, there are genes highly related with the inflammation problem which exhibit profiles that would not be retrieved applying some of the classic microarray analysis methods individually. That is the case of probe set 206011_at, which is related in behavior and in function (apoptosis-related cysteine peptidase) to

probe sets 211367_s_at and 211368_s_at (Fig. 2(a)), stated as relevant for the inflammation problem in [13]. For these particular probe set, the isolated application of classical methods such as M^1 or M^3 with either default *p-value or* false discovery, rate, depending on what each method uses, would not retrieve such probe set as differentially expressed. The same situation applies to probe sets 202076_at and 210538_s_at, related both in behavior and in function (inhibitor of apoptosis protein 2 and 1 respectively) (Fig. 2(b)).

Table 1. Intersection of the results between methods recognizing differentially expressed genes. The number in each cell represents the ratio of coincidence between genes retrieved by the statistical method in the column and in the row relative to the total number of genes recovered by the method in the row $((Row \bigcap Column)/Row)$.

%	M^1	M^2	M^3	M^4	M^5	M^6	M^7	M^8	M^9	M^{10}
M^1	--	92.20	52.29	75.05	96.48	69.23	85.55	70.06	61.33	50.52
M^2	56.06	--	34.07	57.84	85.27	59.54	71.11	62.64	50.57	42.98
M^3	82.19	88.07	--	96.24	94.77	57.35	78.75	72.87	56.86	46.73
M^4	67.22	85.19	54.84	--	95.16	55.49	73.65	70.20	51.49	42.83
M^5	55.20	77.80	33.45	58.94	--	50.28	66.72	66.38	46.42	38.93
M^6	59.04	83.51	31.11	52.84	77.30	--	89.63	56.56	60.64	49.38
M^7	58.36	79.79	34.18	56.10	82.05	71.70	--	62.34	57.23	49.07
M^8	57.36	84.34	37.96	64.17	95.96	54.30	74.80	--	49.62	40.51
M^9	62.10	84.21	36.63	58.21	84.74	72.00	84.95	61.36	--	72.31
M^{10}	59.56	83.34	35.05	56.37	82.72	68.26	84.80	58.34	84.19	--

In contrast, some other available methods retrieve profiles that do not differ between the considered experimental conditions. For example, ANOVA, perhaps based on the violation of statistical constraints ([14]), retrieves a 43% of genes lacking an observable change with the default parameter values. The increase of the specificity of these parameters generates severe effects on the sensitivity of other true changes.

These findings reveal that there are desired and undesired differential profiles termed positive and negatively, respectively. For example, some profiles exhibiting similarly arranged patterns but shifted over time may be relevant for a specific experiment but not for other. In addition, we also found that methods applied to microarray profiles are focused on identifying differences among expression patterns over treatment and/or time since biological replicates are averaged in the same experimental group. However, we might also need to detect differences among subjects.

We create a database with all possible differential profiles derived from genes retrieved in our inflammation problem by all available methods (i.e., 28 differential profiles). This database contains differential profiles that can be labeled as positive or negative according to their interest to be retrieved for a particular study. To validate biologically these profiles we calculate the coincidence the coincidence between our retrieved differential profiles and external information provided by the Gene Ontology database ([15]) showing that genes sharing behaviour are related in function ([16]).

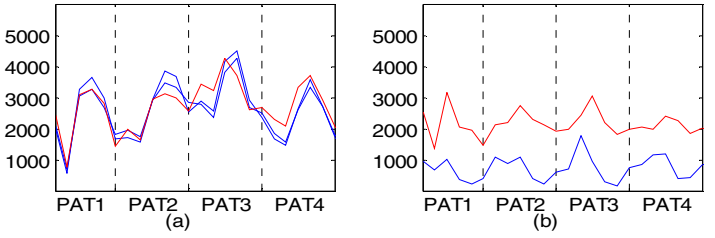

Fig. 2. Probe sets in blue are stated as relevant for the inflammation problem in ([13]). Probe sets in red are detected by application of our methodology but not by applying some classical microarray analysis methods individually. In (a) the probe set in red, 206011_at, is related to probe sets 211367_s_at and 211368_s_at (blue) both in expression throughout time and in function (apoptosis-related cysteine peptidase). In (b) we see the same situation between 202076_at in red and 210538_s_at in blue, which have correlated level of expression throughout time and share their function (inhibitor of apoptosis protein 2 and 1 respectively).

The temporal expression data in our database can be averaged or sequentially represented for each biological replicate. Originally, the database was built based on the inflammatory response patterns, which is based on a very robust microarray experiment ([13]). Now, it is being updated with experiments provided from different sources such as the Ventilator Associated Pneumonia (unpublished results).

The application of our methodology to the database of differential profiles allowed the optimal retrieval of the desired differential profiles. For example, if we were interested in probe exhibiting any of 27 of the profiles in the database and not exhibiting one of the profiles, we are able to do applying $M^5 \cup M^6$ with specificity and sensitivity levels of 94% and 92% respectively. In Fig. 3 we show the method-aggregations from the optimal rules to retrieve individually each of the 28 profiles in our database.

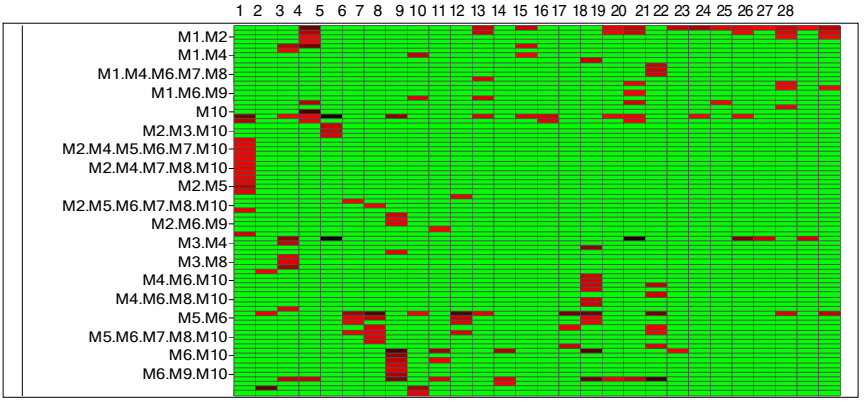

Fig. 3. Microarray analysis methods are able to retrieve some differential profiles and not others. Rows correspond to method-aggregations using the union operator and columns to each of the 28 individual differential profiles from our example. The coloring scheme corresponds to the sensitivity in retrieving the differential profile: from green, the lowest, to black, the highest.

Our approach also recovers probe sets with related behavior to other probe sets with already known profiles which might have related functionalities ([16]). For instance, probe set 206011_at is related in behavior and in function (apoptosis-related cysteine peptidase) to probe sets 211367_s_at and 211368_s_at (Fig. 2(a)), stated as relevant for the inflammation problem in [13]. For these particular probe set, the isolated application of classical methods such as M^1 or M^3 with the default *p-value* and false discovery rate would not retrieve such probe set as differentially expressed. We retrieve such probe set applying the rule that implies the method aggregation $M^7 \cup M^{10}$ with values (1, 0.25, *0.8*) for sensitivity, specificity and cost respectively. The same situation applies to probe sets 202076_at and 210538_s_at, related both in behavior and in function (inhibitor of apoptosis protein 2 and 1 respectively) (Fig. 2(b)). It is retrieved applying the rule of methods $M^3 \cup M^6$ with values *(0.93, 0.35, 0.8)*.

In addition, the representation used in the inflammation problem (Fig. 4) allows us to independently examine the gene behavior in each subject, helping to uncover individual tendencies among biological replicates that could represent conditions not previously considered such as gender or age (e.g., differential profile #15 (Fig. 5), where some of the probe sets from patient 1 exhibit a very different behavior than the rest of the patients).

We illustrate the obtained association rules for Profile #19 from our database (Table 2) and the Pareto-optimal front for the three objectives corresponding to the selected rules (Fig. 6).

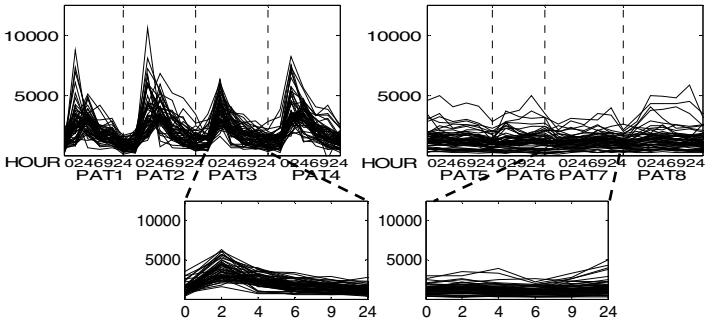

Fig. 4. Profile #19: the expression profiles have been represented separately for each subject the experimental group and patients are arranged individually

3 Methods

Most machine learning techniques are applied to mine into datasets to discover concepts involving objects which share a common methodological framework, even though they employ distinct metrics, heuristics or probability interpretations ([17], [18]): (1) *identification of a database*, different data types can be efficiently organized by taking advantage of a naturally occurring structure over feature space. (2) *learning rules from the database*, searching through the feature space for potential relationships

among data, and either returning the best one found or an optimal sample of them. This learning process would result in the generation of many rules with small extent, as it is easier to explain or match small data subsets than those that constitute a significant portion of the dataset. For this reason, any successful methodology should also consider additional criteria ([19]) to extract broader or more comprehensive rules as a multiobjective optimization problem, based on their specificity, sensitivity and cost as a measures of the rule quality. (3) *Inference*, where new observations can be predicted from previously learned rules by using classifiers that optimize their matching to the rules based on distance ([18]) or probabilistic metrics ([20], [21]).

We propose a method, inspired on conceptual clustering and optimization techniques ([9], [10], [17]), that identifies a database of gene profiles that change their expression over time and/or over treatments and/or over subjects, learns associations rules and make decisions about the microarray analysis method or the best aggregation of methods capable of detecting a desired set of differential profiles, and finally uses these rules to make decisions based on new situations (Fig.1).

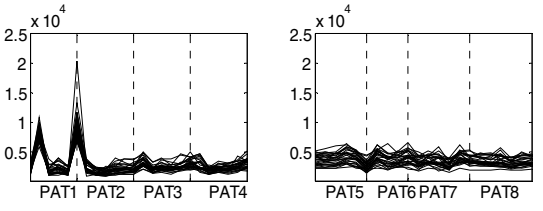

Fig. 5. Profile #15: patient 1 behaves different than patients 2, 3 and 4 for the treatment group

Table 2. Set of decision making association rules generated to retrieve Profile #19. The axes (X,Y,Z) represent the number of methods, specificity and sensitivity for each of the 35 solutions generated.

RULES	Sensitivity	Specificity	Cost
R^1:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^1 IS M^1	0.878378	0.0675676	0.9
R2:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^2 IS M^2	0.972973	0.045512	0.9
R^3:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^3 IS M^7	0.905405	0.0475177	0.9
R^4:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^4 IS $M^1 \cap M^2$	0.864865	0.0721533	0.8
R^5:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^5 IS $M^2 \cup M^{10}$	1	0.0430733	0.8
R^7:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^7 IS $M^1 \cap M^{10}$	0.594595	0.090535	0.8
R^8:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^8 IS $M^2 \cap M^7$	0.891892	0.0538776	0.8
R^9:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^9 IS $M^3 \cap M^9$	0.472973	0.100575	0.8
R^{10}:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^{10} IS $M^3 \cap M^{10}$	0.405405	0.104895	0.8
R^{11}:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^{11} IS $M^1 \cap M^2 \cap M^7$	0.797297	0.0732919	0.7
R^{12}:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^{12} IS $M^1 \cap M^2 \cap M^9$	0.662162	0.0853659	0.7
R^{13}:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^{13} IS $M^1 \cap M^3 \cap M^7 \cap M^{10}$	0.459459	0.112211	0.6
R^{14}:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^{14} IS $M^3 \cap M^6 \cap M^9 \cap M^{10}$	0.364865	0.135	0.6
R^{15}:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^{15} IS $M^1 \cap M^3 \cap M^6 \cap M^9 \cap M^{10}$	0.364865	0.140625	0.6
R^{16}:IF x_1 IS $(P_TP_C)_{19}$ THEN Z^{16} IS $M^1 \cap M^3 \cap M^4 \cap M^6 \cap M^9 \cap M^{10}$	0.364865	0.141361	0.4

3.1 Identification of the Database

Our database is composed of differential profiles obtained from the probe sets differentially expressed retrieved from the expression datasets. The probe sets are obtained using several classical microarray analysis methods. These methods include Student's T-Test proposed in [11], with variants that distinguish changes in the abundance of RNA occurring not only over treatment but also over time; Permutation Test described in ([6]), also including a time approach; Analysis of Variance described in ([5]); and Longitudinal Data approach using Repeated Measures Analysis of Variance described in ([12]).

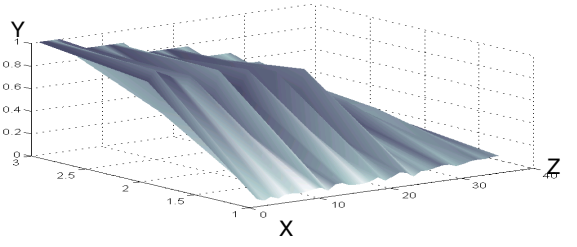

Fig. 6. Pareto-front representation for the set of rules generated to retrieve Profile #19. The axes (*X,Y,Z*) represent the number of methods, specificity and sensitivity for each of the 35 solutions generated.

The probe sets identified by the statistical methods serve as a means to create differential expression profiles (i.e., sets of genes with coordinate changes in RNA abundance) expressed from one experimental condition to the others (i.e., probe sets that change over time, treatments and phenotype). We group separately probe sets for different experimental conditions, treatment and P_T control P_C by applying the *K*-means clustering algorithm ([22]), which takes three input parameters: first, number of resulting clusters *K*, which is estimated by application of the Davies-Bouldin validity index ([23]); second, the similarity measure applied, Euclidean distance, which yields the best results in the clustering of this problem and third, the initialization strategy, random generation of the cluster centroids.

Particularly, in our inflammation problem, we consider the temporal expression data sequentially represented for each biological replicate (i.e., patients in the same experimental group) instead of averaging them to uncover phenotype differences.

We identify differential profiles by applying a coincidence index (*CI*) based on the hypergeometric distribution (*p-value* <0.05), which determines the statistical significance of overlap between pairwise profile association in treatment and control conditions ([16]):

$$CI(P_T, P_C) = 1 - (\sum_{q=0}^{p} \binom{h}{q}\binom{q-h}{n-q} / \binom{g}{h}) \quad (1)$$

that gives the chance probability of observing at least p candidates from a set P_T of size h within another set P_C of size n, in a universe of g candidates. Therefore, probe sets belonging to a cluster in treatment, P_T, can fit in more than one cluster in control,

P_C, and vice versa. We define a differential profile by a triplet ($P_T P_C G$), which represents a set of genes G with similar behavior in treatment P_T and control P_C experiments. The available profiles can be labeled as positive or negative examples for the decision process according to their relevance for a desired analysis.

3.2 Learning Association Rules

We create a set of decision making association rules from the database that, given a desired differential profile or set of profiles, suggests the most appropriate method-aggregations to recognize it. The rules are created to retrieve all possible combinations of differential profiles $P = \{(P_T P_C G)_1, ..., (P_T P_C G)_l\}$ present in our database.

We say that a method M^i is able to retrieve a differential profile $(P_T P_C G)_j$ if it identifies a sufficient number of the genes exhibiting that profile in the data set. That is, $M^i(G) > t$, where t satisfies a statistical power of 80%. Then, an association rule is defined as:

$$R: \text{IF } X \text{ IS } (P_T P_C G)_j \text{ THEN } Z \text{ IS } M^i, \tag{2}$$

where X is the profile queried by the user and Z is a latent class returned from M^i, which represents a method or a method-aggregation. The antecedent of the rule is activated by considering the degree of matching between a query and a profile, both of which are represented by their centroids as a fuzzy set prototype ([19]). We use the Euclidean distance normalized in the unit interval to account for this matching. We extend the antecedent to encode several profiles linked by using typical AND-operations in fuzzy rules (e.g., *T-norms* including the MINIMUM or the PRODUCT ([24]). The consequent of the rules is composed of a single or a method-aggregation (e.g., $M^i \& M^h (G) > t$). The potential method-aggregations are defined as:

$$M = \{M^1, M^2, M^n, M^1 \oplus M^2, M^1 \oplus M^3, \ldots, M^1 \oplus M^2 \oplus \ldots \oplus M^n\} \tag{3}$$

where $\oplus$ is a classical set operator (e.g., the union ($\bigcup$) or the intersection ($\bigcap$)). The association rules are arranged into a lattice (Fig. 7) for one or a set of desired profiles, and structured from top (i.e., intersection of all methods, increasing Type I error) to bottom (i.e., union of all methods increasing error of Type II) [21].

3.3 Selecting Association Rules

We evaluate a rule by the ability of a method-aggregation to recognize a desired positive and not to detect an undesired negative differential profile, as well as the number of methods being considered in the consequent. We explicitly perform a multiobjective evaluation of the performance of the rules by considering three objectives: *specificity, sensitivity* and *cost*:

$$\begin{gathered} Specificity = TN/(TP + FN) \;\; Sensitivity = TP/(TP + FN) \\ Cost = 1 - (\#Methods / Max(Methods)), \end{gathered} \tag{4}$$

where *TP* stands for *True Positive, TN* stands for *True Negative, FP* stands for *False Positive, FN* stands for *False Negative, #Methods* is the number of methods included in the consequent and *Max(Methods)* is the total number of methods available.

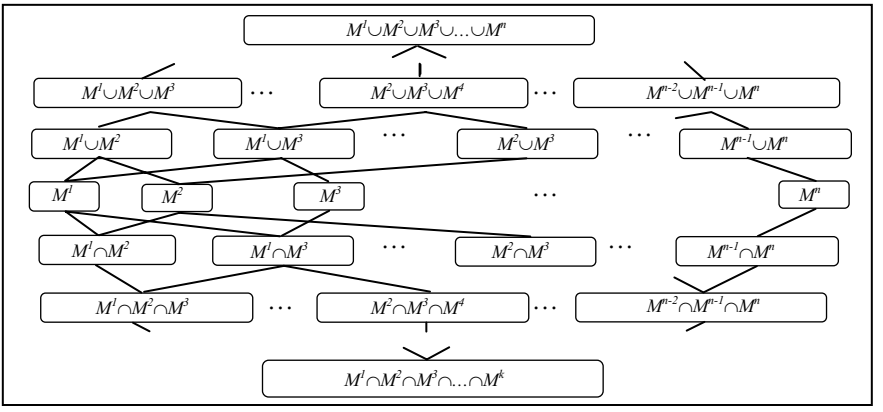

Fig. 7. Lattice structure containing possible association rules

We obtain a set of optimal rules by calculating a trade-off between the opposing objectives that is estimated by selecting a set of solutions that are non-dominated, in the sense that there is no solution that is superior to them in all objectives (i.e., Pareto optimal frontier ([25])). The dominance relationship in a maximization problem of at least two objectives is defined as:

$$a > b \ \ iff \ \ \forall i O_i(a) \geq O_i(b) \exists j O_j(a) > O_j(b) \tag{5}$$

where the O_i and O_j are either one or another defined objective.

3.4 Inferring from Association Rules

We use the set of non-dominated rules and the corresponding metrics derived from the multiobjective optimization process to update the rule defined in equation (2) :

$$R: \text{IF } X \text{ IS } (P_T P_C G)_j \text{ THEN } Z \text{ IS } M^i \text{ WITH } C \tag{6}$$

where C is the confidence of the rule, defined as a weighted sum of the sensitivity, specificity and cost. The rules are fired as typical fuzzy classification rules ([26]):

$$INFERENCE\ (R_1(X),..., R_n(X)) = i,\ i \in \{1,.., n\} \tag{7}$$

where:

$$R_i(X) = T - conorm\{R_1(X),..., R_n(X)\}, \tag{8}$$

and,

$$R_k(X) = \alpha_k \times C_k,\ \forall k \in \{1,..., n\} \tag{9}$$

with α_k and C_k being the degree of matching of the antecedent and the confidence value of the rule k when the profile $(P_T P_C G)_j$ is evaluated, and the *T-conorm* is the fuzzy operation defined as the MAXIMUM.

4 Discussion

The emergence of microarray technology as a standard tool for biomedical research has necessarily led to the rapid development of specific analytical methods to handle these large data sets. Based on what we learnt from studying the performance of classical microarray analysis methods, different methods yield different results for the same set of input data, and some methods are more capable to retrieve certain differential profiles than others, we create a set of decision making association rules between methods or aggregation of them, and differential profiles, that will help us in the decision of which microarray analysis methods to apply on new data sets in order to retrieve the genes exhibiting the desired differential profile.

Our method addressed the need for computational methods to facilitate understanding of differential gene expression profiles, to establish comparisons among them, and to decide which is the most reliable method to identify informational profiles. The proposed methodology is valid for either providing the optimal method-aggregations for a query profiles, or identifying all differential profiles in a given set of microarray data, suggesting the optimal method-aggregations for them and updating the set of possible profiles used for prediction. Although we have applied our procedure to time-course structured experiments, they constitute more general cases of simpler microarray problems where microarray samples are taken as single data points. Therefore, the methodology presented is also useful for simpler microarray experiments with single data points.

Our approach presents various advantages over the standard analytical methods for microarray experiments. First, our proposal consists of machine learning techniques that combine the properties of the methods applied. Second, it permits interaction with the user: given the differential profile queried from the set of data obtains the optimal combination of statistical methods to retrieve the genes exhibiting such profile. Third, the representation used for the profiles, allowing us to examine the behavior of the genes independently in each subject, facilitates the identification of different behaviors of genes across the subjects in the same experimental group. Finally, the system provides solutions based on a trade-off of specificity, sensitivity and cost and the number of methods applied, whereas other methods evaluate their solutions only over one measure, usually a ratio between False Positives and the total number of genes retrieved ([6], [11]).

The computational procedure we propose solves many of the problems actually present in the process of analyzing a microarray experiment, such as the decision of analytical methodology to follow, extraction of biologically significant results, proper management of complex experiments harboring experimental conditions, time-series and intersubject variation. Therefore, it provides a robust platform for the analysis of many types of microarray experiments, from the simplest experimental design to the most complex, providing accurate and reliable results.

References

1. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.
2. Brown,P. and Botstein,D. (1999) Exploring the new world of the genome with DNA microarrays. Nature Genet., 21 (Suppl.), 33-37.

3. Inza, I., Larranaga, P., Blanco, R., Cerrolaza, A.J. (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med. 31(2):91-103.
4. Pan,W., Lin.J. and Le.C. (2001) A mixture model approach to detecting differentially expressed genes with microarray data. Funct. Integr. Genomics, 3(3), 117-124.
5. Park,T., Yi,S.G., Lee,S., Lee,S.Y., Yoo,D.H., Ahn, J.I. and Lee, Y.S. (2003) Statistical tests for identifying differentially expressed genes in time-course microarray experiments. Bioinformatics, 19(6), 694-703.
6. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA. 98, 5116-5121.
7. Vaquerizas, J.M., Conde, L., Yankilevich, P., Cabezon, A., Minguez, P., Diaz-Uriarte, R., Al-Shahrour, F., Herrero, J., Dopazo, J. (2005) GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. Nucleic Acids Res. 1;33(Web Server issue):W616-20.
8. Agrawal,R., Imielinski,T., Swami,A.N. (1993) Mining association rules between sets of items in large databases. In Buneman, P., Jajodia, S., eds.: Proceedings of the ACM SIGMOD. International Conference on Management of Data, Washington, D.C., 207—216.
9. Zwir,I., Shin,D., Kato,A., Nishino,K., Latifi,K., Solomon,F., Hare,J.M., Huang,H. and Groisman,E.A. (2005a) Dissecting the PhoP regulatory network of Escherichia coli and Salmonella enterica. Proc Natl Acad Sci, 102, 2862-2867.
10. Zwir,I., Huang,H. and Groisman,E.A. (2005b) Analysis of Differentially-Regulated Genes within a Regulatory Network by GPS Genome Navigation, Bioinformatics 21(22):4073-83.
11. Li,C. and Wong,W.H. (2003) DNA-Chip Analyzer (dChip). In Parmigiani,G., Garrett,E.S., Irizarry,R. and Zeger,S.L. (eds), The analysis of gene expression data: methods and software. Springer.
12. Der,G. and Everitt,B.S. (2001) Handbook of Statistical Analyses using SAS. Chapman and Hall/CRC.
13. Calvano,S.E., Xiao,W., Richards,D.R., Feliciano,R.M., Baker, H.V., Cho, R.J., Chen, R.O., Brownstein,B.H., Cobb,J.P., Tschoeke,S.K., Miller-Graziano,C., Moldawer,L.L., Mindrinos, M.N., Davis, R.W., Tompkins,R.G. and Lowry,S.F. (2005) The Inflammation and Host Response to Injury Large Scale Collaborative Research Program. A Network-Based Analysis of Systemic Inflammation in Humans. Nature, 13,437(7061):1032-7.
14. Gao, X., Song, PX. (2005) Nonparametric tests for differential gene expression and interaction effects in multi-factorial microarray experiments. BMC Bioinformatics, 21,6:186
15. Romero-Zaliz, R., Rubio-Escudero, C., Cordón, O., Harari, O., del Val, C., Zwir, I. Mining Structural Databases : An Evolutionary Multi-Objective Conceptual Clustering Methodology. Applications of Evolutionary Computing, LNCS 3907, 2006.
16. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture, Nat Genet, 22, 281-285.
17. Cheeseman,P. and Oldford,R.W. (1994) Selecting models from data : artificial intelligence and statistics IV. Springer-Verlag, New York.
18. Cooper, G. and Herskovits, E. (1992) Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9:309--347.
19. Ruspini, E. (2002) Introduction to Longitudinal Research. Edited by M. Bulmer, Social Research Today. London: Routledge.
20. Bezdek, J.C. (1998) Pattern Analysis. In Pedrycz, W., Bonissone, P.P. and Ruspini, E.H. (eds), Handbook of Fuzzy Computation. Institute of Physics, Bristol, F6.1.1-F6.6.20.
21. Mitchell, T. (1997) Machine Learning. McGraw Hill.

22. Duda, R. O., and Hart, P. E. (1973) Pattern Classification and Scene Analysis. John Wiley & Sons, New York, USA.
23. Davies D.L. and Bouldin D.W. (1979) A cluster separation measure. IEEE Trans. On Pattern Analysis and Machine Intelligence, (1)2: 224-227.
24. Klir, G.J. and Yuan, B. (2005) Fuzzy Sets and Fuzzy Logic: Theroy and Applications. Prentice-Hall.
25. Deb, K. (2001) Multi-objective optimization using evolutionary algorithms. John Wiley & Sons, Chichester, New York.
26. Cordón O., del Jesus, M.J., Herrera, F. (1999) A Proposal on Reasoning Methods in Fuzzy Rule-Based Classification Systems. International Journal of Approximate Reasoning, 20: 21-45.