Optimal Selection of Microarray Analysis Methods Using a Conceptual Clustering Algorithm

C. Rubio-Escudero[1], R. Romero-Záliz[1], O. Cordón[1], O. Harari[1],
C. del Val[1], and I. Zwir[1,2]

[1] Department of Computer Science and Artificial Intelligence,
C/Daniel Saucedo Aranda s/n, Granada 18071, Spain
{crubio, rocio, ocordon, oharari, delval, zwir}@decsai.ugr.es
[2] Howard Hughes Medical Institute, Washington University School of Medicine,
St. Louis, MO

Abstract. The rapid development of methods that select over/under expressed genes from microarray experiments have not yet matched the need for tools that identify informational profiles that differentiate between experimental conditions such as time, treatment and phenotype. Uncertainty arises when methods devoted to identify significantly expressed genes are evaluated: do all microarray analysis methods yield similar results from the same input dataset? do different microarray datasets require distinct analysis methods?. We performed a detailed evaluation of several microarray analysis methods, finding that none of these methods alone identifies all observable differential profiles, nor subsumes the results obtained by the other methods. Consequently, we propose a procedure that, given certain user-defined preferences, generates an optimal suite of statistical methods. These solutions are optimal in the sense that they constitute partial ordered subsets of all possible method-associations bounded by both, the most specific and the most sensitive available solution.

1 Introduction

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] and many others). 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 patient) and to decide which is the most reliable method to identify differences across profiles.

We investigated the performance of several commonly used statistical methods, including T-Tests [4], Permutation Tests [5], Analysis of Variance [6] and Repeated Measures ANOVA [7], in identifying differential expression profiles that change over time, treatments and phenotype. We found that these methods do not identify all ob-

F. Rothlauf et al. (Eds.): EvoWorkshops 2006, LNCS 3907, pp. 172 – 183, 2006.

servable distinct profiles. Moreover, none of them subsumes the results obtained by the other methods.

In view of these results, we propose a conceptual clustering method [8], [9], [10], devoted to discover optimal associations of microarray analysis methods in an effort to identify differential gene expression profiles.

2 Methods

We propose a conceptual clustering approach [8], [9], [10] devoted to identify optimal associations among microarray analysis methods in an effort to identify differential expression profiles (Fig. 1). This approach consists of six phases: (1) preprocessing of the dataset; (2) identification of differentially expressed genes by application of several statistical methods; (3) arrangement of a lattice structure containing all possible associations of the statistical methods applied; (4) association of differentially expressed genes into differential profiles by clustering genes that change their expression over time, patient and/or treatment; (5) evaluation of the performance of the method-associations based on their specificity and sensitivity in the identification of previously detected differential profiles, using multiobjective optimization techniques [11], [12]. We create a set of method association rules based on the learned mappings of differential profiles into method-associations, [13]; (6) finally, we are able to predict optimal method-associations to identify differential profiles in new microarray datasets by use of the method association rules.

2.1 Identification of Differentially Expressed Genes

We perform the retrieval of differentially expressed genes from one experimental condition to the other/s by application of several statistical techniques [3], [14], harboring Student's T-Test proposed in [4], including some of the variants the method poses to distinguish changes in the abundance of RNA occurring over both treatment and time; Permutation Test described in [5], also including a time approach; Analysis of Variance described in [6]; and Longitudinal Data approach by using Repeated Measures Analysis of Variance described in [7].

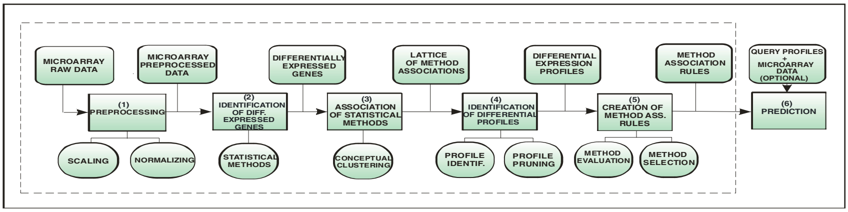
MICROARRAY RAW DATA
MICROARRAY PREPROCESSED DATA
DIFFERENTIALLY EXPRESSED GENES
LATTICE OF METHOD ASSOCIATIONS
DIFFERENTIAL EXPRESSION PROFILES
METHOD ASSOCIATION RULES
QUERY PROFILES + MICROARRAY DATA (OPTIONAL)
(1) PREPROCESSING
(2) IDENTIFICATION OF DIFF. EXPRESSED GENES
(3) ASSOCIATION OF STATISTICAL METHODS
(4) IDENTIFICATION OF DIFFERENTIAL PROFILES
(5) CREATION OF METHOD ASS. RULES
(6) PREDICTION
SCALING
NORMALIZING
STATISTICAL METHODS
CONCEPTUAL CLUSTERING
PROFILE IDENTIF.
PROFILE PRUNING
METHOD EVALUATION
METHOD SELECTION

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.

2.2 Detection of Method-Associations

We arrange a lattice containing all potential associations of the statistical methods used to retrieve differentially expressed genes (Fig. 2). The methods are associated as:

$$M = \{M^1, M^2, M^n, M^1 \oplus M^2, M^1 \oplus M^3, \ldots, M^1 \oplus M^2 \oplus \ldots \oplus M^n\}, \qquad (1)$$

where $\oplus$ is a classical set operator (e.g., the union ($\bigcup$) or the intersection ($\bigcap$)) applied to the sets of genes retrieved by each method, and M^1 corresponds to T-Test, M^2 to T-Test considering time, M^3 Permutation Test, M^4 Permutation Test con sidering time, M^5 ANOVA over treatment, M^6 ANOVA over time, M^7 ANOVA over treatment and time, M^8 RMANOVA over treatment, M^9 RMANOVA over time and M^{10} RMANOVA over treatment and time.

The lattice containing all potential method-associations, *M,* is structured from top (i.e., intersection of all methods) to bottom (i.e., union of all methods) [15]. Each node in the lattice ($M^i \in M$) is applied to the microarray dataset (*D*) retrieving the set of differentially expressed genes that are recognized by the method or method-associations in such node $(M^i(D))$.

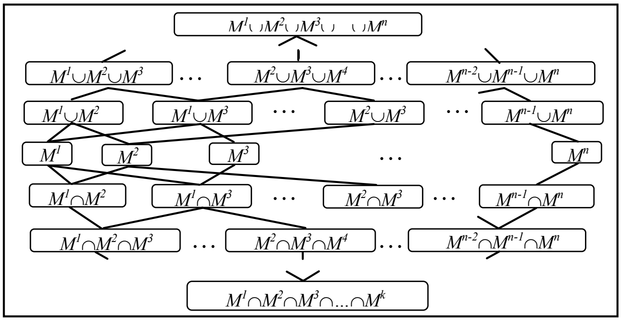

Fig. 2. Lattice structure containing all statistical methods potential associations

2.3 Identification of Differential Profiles

The set of genes previously identified in Section 2.2 serves as a means to create differential expression profiles (i.e., sets of genes with coordinate changes in RNA abundance) between treatment P_T, control P_C and subject. The applied representation (Fig. 3) allows us to identify different pattern behavior among patients inside the same experimental group, since this information may be missed if patients in the same experimental group were not plotted individually.

We clustered separately genes in treatment and control groups. Therefore, genes belonging to a cluster in treatment, P_T, can fit in more than one cluster in control, P_C, and vice versa. We apply the *K*-means clustering algorithm [16] and identify differential profiles denoted as $(P_T P_C)$, which are pairwise relationships between profiles, P_T

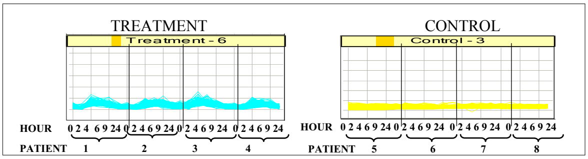

Fig. 3. The expression profiles have been represented separately for each experimental group and patients arranged individually

and P_C, from treatment and control experiments, respectively. This relationship is defined as the significant intersection of genes between P_T and P_C, which is constrained by a threshold based on the typical statistical power of 80%.

2.4 Creation of Method Association Rule

We create a set of method association rules that, given a set of differential profiles queried by the user, suggests the most appropriate method-associations capable to retrieve them. The method association rules are created based on the lattice structure from Section 2.2, containing all potential method-associations, and the set of all possible differential profiles P from Section 2.4 defined as $P = \{(P_T P_C)_1,....,(P_T P_C)_l\}$ where $(P_T P_C)_j \in P$ represents each of the differential profiles present in P.

2.4.1 Method-Association Performance Evaluation

We evaluate the performance of the method-associations $M^i \in M$ for the query profiles $X^S = (x_1,..,x_s)$, over two objectives: *specificity* and *sensitivity*

$$Specificity = TN/(TP+FN) \qquad Sensitivity = TP/(TP+FN)\,, \tag{2}$$

where TP stands for True Positives (i.e., genes exhibiting profile $x_u \in X^S$, which have been successfully retrieved by the applied method-association M^i), TN stands for True Negatives (i.e., genes exhibiting profile $x_u \notin X^S$ and not retrieved by M^i), FP stands for False Positives (i.e., genes exhibiting profile $x_u \notin X^S$ and retrieved by M^i) and FN stands for False Negatives (i.e., genes exhibiting profile $x_u \in X^S$ and not retrieved by M^i). These four factors are calculated as:

$$TP = \frac{\varphi^u \cap \eta^i}{\varphi^u} \; TN = \frac{(D-\varphi^u)\cap(D-\eta^i)}{(D-\varphi^u)} \; FP = \frac{(D-\varphi^u)\cap\eta^i}{(D-\varphi^u)} \; FN = \frac{\varphi^u \cap (D-\eta^i)}{\varphi^u}, \tag{3}$$

where φ^u represents the genes in the microarray set D that exhibit the queried profile $x_u \in X^S$, and $\eta^i = M^i(D)$, the genes from D retrieved by the method-association M^i.

2.4.2 Method-Association Selection

We evaluate the method-associations in M based on their specificity and sensitivity. These two objectives are always conflicting, so we use a multiobjective optimization technique to maximize them, allowing us to detect all optimal methods-associations in M for the query profiles X^S [11], [12]. We define objectives $(O_{1,}O_2)$ corresponding to specificity and sensitivity respectively.

2.4.3 Creation of a Set of Method Association Rules

We use the non-dominated method-associations described in Section 2.4.2 to create the method association rules $R = \{R^1, \ldots, R^k\}$ where $R^f \in R$ is defined as:

$$R^f : \text{IF } x_1 \text{ IS } (P_T P_C)_1^f \text{ AND}, \ldots, \text{AND } x_s \text{ IS } (P_T P_C)_s^f \text{ THEN } z^f \text{ IS } M^i \text{ WITH } C^f , \tag{4}$$

where $(x_1, \ldots, x_s)$ are the profiles X^S queried by the user; $(P_T P_C)_1^f$,..., $(P_T P_C)_s^f \in P$; $z^f \in M$ is the appropriate method-association to retrieve X^S according to rule R^f ; and C^f denotes a measure of the specificity/sensitivity levels for z^f , defined as:

$$C^f = \frac{(w_1 * O_1(M^i)) + (w_2 * O_2(M^i))}{w_1 + w_2}, \tag{5}$$

where w_1 and w_2 are the weights associated to (O_1, O_2) respectively. These values are provided by the user based on the relevance of each of these objectives for the particular study. If no values are given, the standard (0.5, 0.5) are used.

2.5 Prediction Using Method Association Rules

The prediction phase works at two levels depending on the given input. If the input is a microarray data set *D*', our methodology will provide the differential expression profiles *P'* in the data set along with the optimal method-associations to retrieve such profiles. It might be the case that some of the differential profiles *P'* uncovered from *D*' were not included in the set of differential profiles *P* already learned by the methodology. Consequently, the information provided as input will be used to update *P* and *R*. If the input is a set of query profiles X^S , the output will consist of the optimal method-association M^h for X^S at a certain C^f value. To obtain these outputs, we apply *matching* and *inference* operations to the method association rule set [17].

Given an association rule set $R = \{R^1, \ldots, R^k\}$, for the differential profiles provided as the query set $X^S = (x_1, \ldots, x_s)$, we define the matching degree Q of $x_u \in X^S$ with the *if-part* of the association rule R^f as:

$$Q(x_{u,} (P_T P_C)_u^f) = 1 - \left\| \overline{x_u} - \overline{(P_T P_C)_u^f} \right\|, \tag{6}$$

with $\| \ \|$ being the Euclidean distance, and $\overline{(P_T P_C)}$ the centroids of the profiles.

Therefore, given a set of query profiles X^S , we define the strength of activation of the *if-part* of the rule R^f as:

$$R^f(X^S) = \min(Q(x_1, (P_T P_C)_1^f), \ldots, Q(x_s, (P_T P_C)_s^f)) . \tag{7}$$

Let $h^f(R^f(X^S), C^f)$ denote the *degree of association* of the query profiles X^S with the method-association M^i according to rule R^f and the specificity/sensitivity level C^f . This degree is obtained by applying a product operator between $R^f(X^S)$ and C^f . The optimal method-association for the queried profiles X^S is defined as:

$$M^i / h^i(\ R^i(X^S), C^i) = \max_{f \in k}\ h^f(\ R^f(X^S), C^f)\,. \tag{8}$$

3 Results

We apply our procedure to a data set derived from longitudinal blood expression profiles of human volunteers treated with intravenous endotoxin compared to placebo. We expect to identify molecular pathways that provide insight into the host response over time to systemic inflammatory insults, as part of a Large-scale Collaborative Research Project sponsored by the National Institute of General Medical Sciences (www.gluegrant.org) [18].

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) [18]. 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 Affymetryx Inc., containing a set of 22283 genes.

3.1 Identification of Differentially Expressed Genes

The statistical methods harbored have been applied using the standard *p-value* $\sigma = 0.05$. The number of differentially expressed genes retrieved by each of the methods from the original set of genes is M^1-10942 genes, M^2-7841, M^3-3904, M^4-8023, M^5-13151, M^6-4588, M^7-6070, M^8-8557, M^9-3995, M^{10}-3367. These values show the number of significant genes retrieved by each of the statistical methods ranges in a wide rank. Moreover, the concordance rates also vary widely, in-

Table 1. Coincidence between methods in the retrieval of genes. The number in each cell represents a ratio of coincidence between genes retrieved by the statistical method in that column and the genes retrieved by the statistical method in that row relative to the total number of genes retrieved 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	--

dicating 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).

3.2 Association of Statistical Methods

The lattice arranged in this particular work contains all potential combinations of union and intersection of the ten statistical methods applied. Thus, M' is defined as

$$M = \{M^1, M^2, ..., M^{10}, M^1 \oplus M^2, M^1 \oplus M^3, \ldots,$$
$$M^2 \oplus M^3, \ldots, M^1 \oplus M^2 \oplus M^3, \ldots, M^1 \oplus M^2 \oplus M^3 \oplus \ldots \oplus M^9 \oplus M^{10}\}$$

We found that there is a relationship between the statistical methods and the differential profiles they are able to identify (see Section 2.2), having differential profiles identified by some methods and not by others. For example, the differential profile in (Fig. 4(a)) harbors 29 genes in our dataset D and is only retrieved by those statistical methods that take into account the time factor (e.g., M^2, which retrieves more than 90% of these genes). This happens because the statistical methods that consider the treatment vs. control factor make an average of the expression values from patients 1 and 2 with those of patients 3 and 4 by considering them as replicas. Consequently, the differential behavior between them is lost.

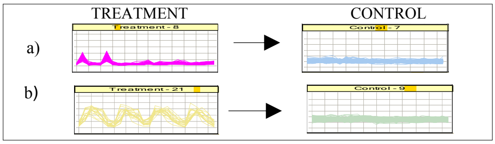

Fig. 4. Examples of differential profiles only identified by some of the statistical methods

3.3 Identification of Differential Profiles

The expression profiles have been represented separately for each experimental group (Section 2.3), and patients arranged individually. In our current problem, with eight patients, four treated with intravenous endotoxin (i.e., patients 1 to 4) and four with placebo (i.e., patients 5 to 8), and data retrieved over time at hours 0, 2, 4, 6, 9 and 24, each profile is represented by 24 consecutive time points (see Fig. 5).

The differential profiles extracted from the treatment group show different levels of expression change. For example, there are sets of genes sharing very high variations in the levels of expression (e.g., profiles 15, 19, 21, and 22 in Fig. 5). In addition, some other profiles show differential characteristics for the patients (e.g., profiles 8 and 16 in Fig. 5). In the control group, the profiles are more homogeneous than in the treatment group.

Typically, testing the coincidence among different data sources and clustering methods serves as a tool to investigate the validity of the identified groupings [19]. We follow this guideline to increase the confidence in the obtained differential profiles. Therefore, we calculate the coincidence between our retrieved differential profi-

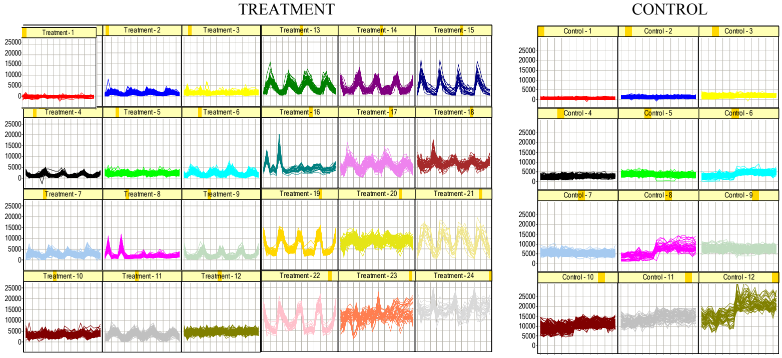

Fig. 5. Representation of the differential profiles obtained separately for the treatment and control groups using the statistical methods applied in the current work

les and external information provided by the Gene Ontology database [20]. To address this problem we developed an evolutionary multiobjective conceptual clustering methodology (R.R.Z., C.R.E., O.C., J.P.C., and I.Z., manuscript in preparation) that extracts clusters composed of features such as biological processes, molecular functions and cellular components defined at different specificity levels, and compare these clusters with our differential profiles by using a coincidence index test based on the hypergeometric distribution [9], [10], [19].

3.4 Creation of Method Association Rules

We have arbitrarily selected six profiles (i.e., $(P_T P_C)_1,\ldots,(P_T P_C)_6$) identifying a total of 1395 genes in our dataset D and plotted as treatment clusters 2, 3, 4, 5, 10 and 12 in Fig. 5. These profiles represent genes exhibiting non-uniform behavior for distinct patients in the treatment group, and genes with changes in a level of expression smaller than 5000. We applied our methodology to find the optimal method-associations M^i to retrieve them.

3.4.1 Method Association Performance Evaluation

The results of the evaluation of the method-associations contained in the lattice M' for the differential profiles are shown in Table 2, where the information relative to the sensitivity and specificity levels for the application of the most representative method-associations over D is also specified. On the one hand, we observe that the union set of the genes obtained by seven of the statistical methods evaluated (i.e., methods $M^2,M^3,M^5,M^6,M^7,M^8,M^{10}$) contains the 1395 genes desired (i.e., sensitivity value of 1) but with a low level of specificity (i.e., value of 0.369). On the other hand, the intersection set of genes obtained by the same seven statistical methods has a very low level of sensitivity (i.e., only 95 out of the 1395 genes were retrieved), whereas the value for specificity is very high. In between these two extremes we see some other method-associations which evaluation reveal trade-off solutions between the specificity and sensitivity objectives (Table 2).

3.4.2 Method Association Selection

Once the method-associations *M* have been evaluated, we search for the non-dominance relations in their applications to the microarray dataset *D*. The decision is based on the levels of specificity and sensitivity in Table 2. The Pareto optimal front conformed by this set of non-dominated method-associations is represented in Fig. 6.

Table 2. Specificity and sensitivity values for the method-associations. The non-dominated solutions are pointed out with a star.

Methods	Specificity	Sensitivity
M^2	0.611	0.707
M^3	0.826	0.205
M^5	0.448	0.785
* M^6	0.813	0.447
* M^7	0.747	0.587
M^8	0.625	0.537
* M^{10}	0.859	0.322
$M^2 \cap M^3$	0.803	0.432
* $M^2 \cup M^3$	0.618	0.866
* Union of ($M^2, M^3, M^5, M^6, M^7, M^8, M^{10}$)	0.3690	1
* Intersection of ($M^2, M^3, M^5, M^6, M^7, M^8, M^{10}$)	0.983	0.066

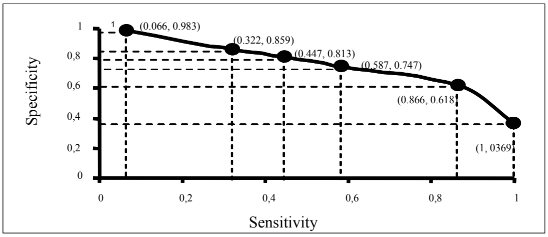

Fig. 6. Results of the evaluation of the method-associations contained in the lattice *M*' for the six selected differential profiles

3.4.3 Creation of Method Association Rules

The set of method association rules is created based on the evaluated profiles (i.e., $(P_T P_C)_1, \ldots, (P_T P_C)_6$), and the method-associations M^i present in the Pareto optimal front of non-dominated solutions. The weights (w_1, w_2) associated to the objectives (O_1, O_2) are set to (0.5, 0.5) to calculate the specificity/sensitivity measure C^f . We illustrate two association rules extracted from the evaluation of M' over the former profiles, which have the following form:

$$R^1 : \text{IF } x_1 \text{ IS } (P_T P_C)_1^1 \text{ AND}, \ldots, \text{AND } x_6 \text{ IS } (P_T P_C)_6^1 \text{ THEN } Z^1 \text{ IS } M^6 \text{WITH } C^1$$

where C^f is calculated based on the specificity/sensitivity levels obtained on the application of such method over $(P_T P_C)_1$,…, $(P_T P_C)_6$ profiles (Table 2):

$$C^1 = (0.5 * 0.813) + (0.5 * 0.447) / (0.5 + 0.5) = 0.631$$

and: R^2 : IF x_1 IS $(P_T P_C)_1^2$ AND,…,AND x_6 IS $(P_T P_C)_6^2$ THEN Z^2 IS $M^2 \cup M^3$ WITH C^2

where C^2 is defined as: $C^2 = (0.5 * 0.618) + (0.5 * 0.866) / (0.5 + 0.5) = 0.742$

3.5 Prediction Using Method Association Rules

To evaluate the ability of our computational approach to retrieve differential profiles, we have randomly selected 100 query sets X^S containing a random number of differential profiles from the 24 actually available. Using the method association rules created, and averaging the results, we obtained an 86.92% of overall performance measurement [21] as a particular correlation coefficient implementation.5 Prediction using method association rules

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. Despite the multiplicity of methods devoted to identify differentially expressed genes, there is a dearth of computational methods intended to optimize use of a particular method or suite of methods. Our motivation was to address two frequently asked questions: 1) do all methods retrieve the same results with the same set of input data, and 2) are the results from methods which retrieve a smaller amount of genes subsumed in the results of methods retrieving a larger amount of genes? We have shown herein how commonly used statistical methods yield different results for the same data input: each statistical method applied neither identifies all observable differential profiles, nor subsumes the results obtained by the other methods (see Tables 1 and 2). Our method also addresses another common conundrum, specifically the need for computational methods to facilitate understanding of differential gene expression profiles, to establish comparisons among them, and to decide which the most reliable method to identify informational profiles is. In this context we propose a procedure that generates optimal associations of microarray analysis methods for the set of data being analyzed, based on the differential expression profiles exhibited by the genes in the dataset.

The generation of the optimal method-associations is based on a set of previously obtained method association rules between differential profiles and the optimal method-associations to identify them. The methodology proposed is valid for either providing the optimal method-associations for a set of query profiles, or identifying all differential profiles in a given set of microarray data, suggesting the optimal method-associations for them and updating the set of possible profiles used for prediction. Although we have applied our procedure to a time-course structured experiment, we have to take into account that time-course experiments constitute more general cases than simpler microarray problems where time is not a factor and microarray samples are taken as single data points. Therefore, the methodology presented is also useful for simpler microarray experiments with single data points.

This approach presents various advantages over the standard analytical methods usually applied to microarray experiments. First, it permits combining the results of independent analytical methods for microarray experiments. Our proposal consists of a conceptual clustering technique that combines the advantages of the methods applied. The combination of the union and intersection operators also provides the possibility of querying negative samples (i.e., genes which exhibit a given profiles but not others). Second, it permits interaction with the user in the selection of differentially expressed profiles, where the user provides the differential profiles queried from the set of microarray data and receives the optimal combination of statistical methods to retrieve the genes exhibiting those profiles. Third, the representation used for the profiles is optimal, as plotting the patients sequentially presents advantages over the traditional one, where all biological replicates (i.e., patients in the same experimental group) are combined in just one set of values. The main advantage of this representation is that we can examine the behavior of the genes independently in each patient, making it possible for us to recognize different behaviors of genes across the patients in the same experimental group. These differences can help us to discover the influence of biological conditions not previously considered in the experiment such as gender or age. Finally, the system provides solutions based on a trade-off of specificity vs. sensitivity, whereas other methods evaluate their solutions over one measure, usually a ratio of False Positives and the total number of genes retrieved [4], [5]. As a result of this trade-off, the procedure provides as output all non-dominated solutions in terms of specificity and sensitivity by application of multiobjective techniques.

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 results biologically significant for the experts, proper management of complex experiments harboring experimental conditions, time-series and patients. Therefore, it sets up a robust platform for the analysis of all 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. 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.
4. 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.
5. 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.

6. 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.
7. Der,G. and Everitt,B.S. (2001) Handbook of Statistical Analyses using SAS. Chapman and Hall/CRC.
8. Cheeseman,P. and Oldford,R.W. (1994) Selecting models from data : artificial intelligence and statistics IV. Springer-Verlag, New York.
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 (in press).
11. Chankong,V. and Haimes,Y.Y. (1983) Multiobjective decision making theory and methodology. North-Holland.
12. Deb,K. (2001) Multi-objective optimization using evolutionary algorithms. John Wiley & Sons, Chichester, New York.
13. 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
14. Kooperberg,C., Sipione,S., LeBlanc,M., Strand, A.D., Cattaneo,E. and Olson, J.M. (2002) Evaluating test statistics to select interesting genes in microarray experiments. Hum. Mol. Genet., 11(19), 2223-2232.
15. Mitchell,T. (1997) Machine Learning. McGraw Hill.
16. Duda, R. O., and Hart, P. E. (1973) Pattern Classification and Scene Analysis. John Wiley & Sons, New York, USA.
17. 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.
18. 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, in press.
19. 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.
20. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat Genet, 25, 25-29.
21. Benitez-Bellon,E., Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) Evaluation of thresholds for the detection of binding sites for regulatory proteins in Escherichia coli K12 DNA. Genome Biol. 3(3)):RESEARCH0013.