A Multiobjective Evolutionary Conceptual Clustering Methodology for Gene Annotation Within Structural Databases: A Case of Study on the *Gene Ontology* Database

Rocío C. Romero-Zaliz, Cristina Rubio-Escudero, J. Perren Cobb, Francisco Herrera, Óscar Cordón, and Igor Zwir

***Abstract*—Current tools and techniques devoted to examine the content of large databases are often hampered by their inability to support searches based on criteria that are meaningful to their users. These shortcomings are particularly evident in data banks storing representations of structural data such as biological networks. Conceptual clustering techniques have demonstrated to be appropriate for uncovering relationships between features that characterize objects in structural data. However, typical conceptual clustering approaches normally recover the most obvious relations, but fail to discover the less frequent but more informative underlying data associations. The combination of evolutionary algorithms with multiobjective and multimodal optimization techniques constitutes a suitable tool for solving this problem. We propose a novel conceptual clustering methodology termed *evolutionary multiobjective conceptual clustering* (EMO-CC), relying on the NSGA-II multiobjective (MO) genetic algorithm. We apply this methodology to identify conceptual models in structural databases generated from gene ontologies. These models can explain and predict phenotypes in the immunoinflammatory response problem, similar to those provided by gene expression or other genetic markers. The analysis of these results reveals that our approach uncovers cohesive clusters, even those comprising a small number of observations explained by several features, which allows describing objects and their interactions from different perspectives and at different levels of detail.**

***Index Terms*—Conceptual clustering, database annotation, evolutionary algorithms (EAs), gene expression profiles, gene ontology (GO), knowledge discovery, multiobjective (MO) optimization.**

This work was supported in part by the Spanish Ministry of Science and Technology under Project TIC-2003-00877, Project BIO2004-0270E, and Project TIN2006-12879. The work of I. Zwir was supported by the Howard Hughes Medical Institute, Washington University School of Medicine, St. Louis, MI.

R. Romero-Zaliz, C. Rubio-Escudero, and F. Herrera are with the Department of Computer Science and Artificial Intelligence, University of Granada, E-18071 Granada, Spain (e-mail: rocio@decsai.ugr.es; crubio@decsai.ugr.es; herrera@decsai.ugr.es).

J. P. Cobb is with the Cellular Injury and Adaptation Laboratory, Washington University School of Medicine, St. Louis, MO 63110 USA (e-mail: cobb@wustl.edu).

Ó. Cordón is with the Department of Computer Science and Artificial Intelligence, University of Granada, E-18071 Granada, Spain. He is also with the European Center for Soft Computing, 33600 Mieres, Asturias, Spain (e-mail: ocordon@decsai.ugr.es; oscar.cordon@softcomputing.es).

I. Zwir is with the Department of Computer Science and Artificial Intelligence, University of Granada, E-18071 Granada, Spain. He is also with the Howard Hughes Medical Institute, Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MI 63110 USA (e-mail: igor@decsai.ugr.es; zwir@borcim.wustl.edu).

I. Introduction

THE INCREASED availability of repositories containing representations of complex objects in spatial databases, such as satellite maps, or temporal databases, including microarray time series, regulatory networks, or metabolic pathways, permits access to vast amounts of data where these objects may be observed [1]–[3]. However, the underlying object representations used in these databases are typically based on computational convenience of database implementers and their tendency to increase the amount of stored data [4]. Current tools and techniques devoted to examine the contents of these large databases are often hampered by their inability to support searches based on criteria that are meaningful to the users of those repositories. In particular, and in spite of the recent renewed interest in knowledge discovery techniques (or data mining), there is a potential dearth of data analysis methods intended to facilitate the understanding of the represented objects and related systems by their most representative features and those relationships derived from these features (i.e., structural data [5]). Plain databases cannot deal with this structural information. For example, images often stored in spatial databases are composed of small pieces of geometrical objects (e.g., triangles or squares) that encode complex relationships between them, including nested or composite relative locations [e.g., square on triangle, Fig. 1 (a1)–(a2)]. These types of relationships normally exceed the simple presence/absence of the underlying elements (e.g., triangle and square). Indeed, plain data are difficult to generalize into more abstract concepts (e.g., object on triangle) resulting from frequent patterns found in the database [Fig. 1 (d2)].

Structural data, in contrast to plain data, can be viewed as a graph containing nodes representing objects. Subgraph partitions of the dataset are termed *substructures* [5] [Fig. 1(b)–(d)]. Each object in a substructure is described by its most representative features, which are encoded as nodes linked to other nodes by edges corresponding to their relationships. Conceptual clustering techniques have been successfully applied to structural data to uncover concepts that explain underlying objects by searching through a predefined space of potential hypothesis (i.e., substructures that represent associations of features) for those that best fits the training examples [6].

The formulation of conceptual clustering as a search problem, in a graph-based structure, would result, however, in the generation of many substructures with small extent, as it is easier

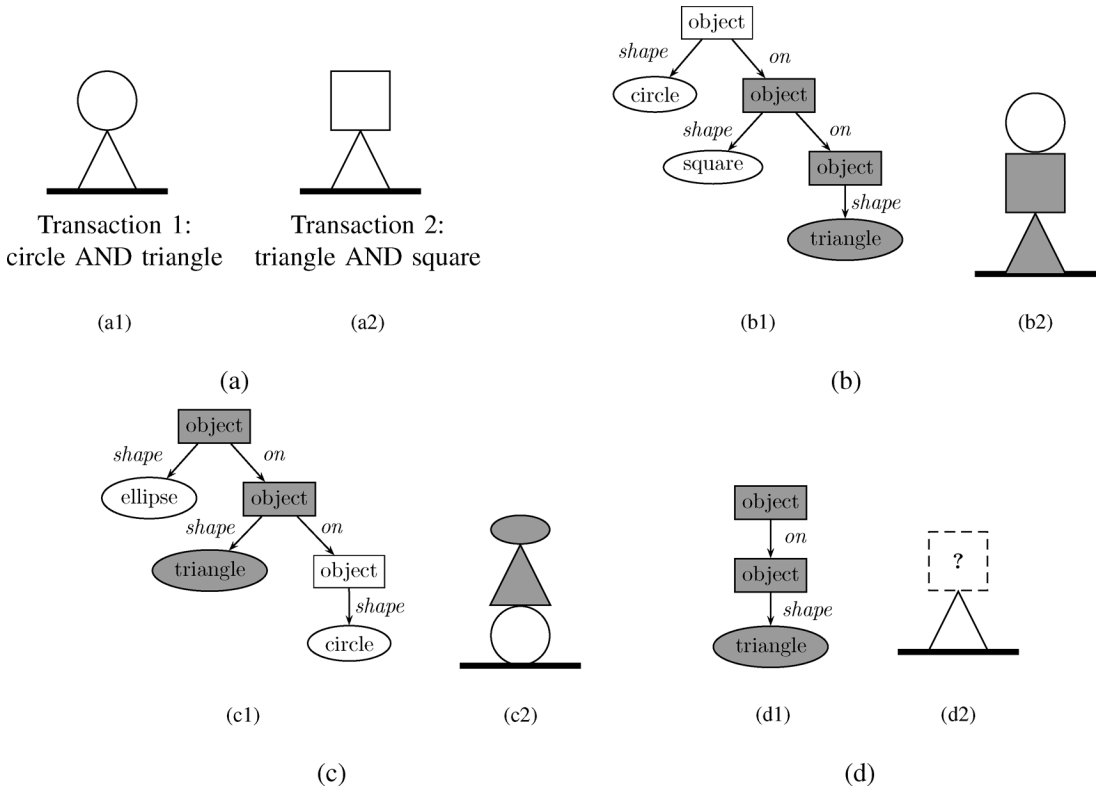

Fig. 1. Differentiating plain and structural databases: example of geometrical observations. (a) Plain codification of two observations as typical transactions encoding a presence/absence relationship between data. (b) and (c) Structural codification also encoding positional relationships. (b1)–(c1) Tree-graphs corresponding to computational representations in a structural database. (b2)–(c2) Geometrical interpretation of the represented observations. The color-coded parts of the trees show repeated instances that generate substructures. (d) A generalized substructure learned from (b)–(c) that cannot be encoded by (a).

to explain or model smaller data subsets than those that constitute a significant portion of the dataset. For this reason, any successful methodology should also consider additional criteria to extract better defined concepts based on the complexity of the substructure being explained, the number of retrieved substructures, and their diversity [5], [7]–[9]. These are conflicting criteria that can be approached as an optimization problem, close in spirit to minimum description length (MDL) methods [10], which are based on the aggregation of the various objectives into a global measure of cluster quality. The basic challenge with this approach is the potential bias caused by weighting the objectives [7], [11], which always derives from the convergence to solutions corresponding to single or limited regions of the search space. This problem is noteworthy because typical data mining approaches, particularly in computational biology, tend to emphasize consensus or most frequent patterns [12]. These consensus patterns often conceal rather than reveal novel and useful knowledge about the problem, retrieving only already known or irrelevant information that discourages the use of computational methods [13], [14]. Consequently, there is a need for new methods that can provide even less frequent but more descriptive substructures that reflect problem descriptions from different angles [15].

In this paper, we propose a conceptual clustering methodology termed EMO-CC for *evolutionary multiobjective conceptual clustering* that uses multiobjective and multimodal optimization techniques to retrieve meaningful substructures from structural databases. The EMO-CC methodology uses an efficient search process based on evolutionary algorithms (EAs) [16]–[18] relying on the NSGA-II algorithm [19], which inspects large data spaces that otherwise would be intractable. Indeed, it explores hierarchically organized databases, which can contain data defined at different levels of specificity. EMO-CC identifies optimal clusters corresponding to different substructures lying in the Pareto optimal frontier [7], [17]. This frontier is composed of a collection of multiobjective optima in the sense that their solutions are not worse than any other substructure for the objectives being considered (i.e., nondominated) [17]. This approach is less biased than aggregating various objectives into a weighted function. The clusters obtained by EMO-CC are composed of solutions belonging to different neighborhoods, where each cluster represents a local optimum in a multimodal problem [20]. The methodology optimizes the number of substructures being retrieved based on a flexible compression of the database and provides annotations for the uncovered substructures [5]. Finally, EMO-CC applies an unsupervised classification approach to predict new members of previously discovered substructures [6].

We apply EMO-CC to the discovery of meaningful substructures containing genes sharing common sets of features (i.e., GO terms) in the gene ontology (GO) database [3], which is composed of biological processes, cellular components and molecular functions defined at different levels of specificity. These substructures can explain/predict gene expression profiles. We consider gene profiles that reflect differences in gene expression over time, treatment and patient, corresponding to an inflam-

matory response study performed on human volunteers treated with intravenous endotoxin compared to a placebo [21]. Understanding the inflammation process is critical because the body uses inflammation to protect itself from an infection or an injury (e.g., crashes, massive bleeding, or a serious burn) which, in extreme cases (e.g., car accidents or gun shootings), can lead to massive organ malfunction and death. Moreover, the majority of the deaths are caused due to these problems [21].

To validate our approach, we perform a two-way analysis: 1) we evaluate the performance of our proposal using standard metrics for evolutionary multiobjective algorithms and 2) we use biological information from gene expression measured from blood samples by Affymetrix microarrays to independently explain and summarize the obtained results. Indeed, we perform comparisons between EMO-CC and well-known conceptual clustering and bioinformatic techniques. The obtained results suggest that EMO-CC is a useful tool for extracting novel biological information that provides insights into the analysis of gene expression data.

This work is organized as follows. Section II presents our preliminaries, including descriptions of conceptual clustering and unsupervised methods; characterization of the GO project, and a brief summary of multiobjective optimization techniques. Section III describes the EMO-CC methodology. Section IV introduces the inflammatory response problem and the gene expression profiles. Section V shows the results obtained by EMO-CC applied to the GO database and compares these results with four other methods: two of them are clustering approaches, while the other two are state-of-the-art GO tools. Finally, Section VI concludes with the discussion.

II. Preliminaries

In this section, we provide the methodological and problem background used in this work. First, we briefly supply a general framework for conceptual clustering algorithms, and introduce two methods used to mine structural databases. These methods include the SUBDUE conceptual clustering method [22] and the APRIORI unsupervised method [23]. Second, we describe the GO project and its structural database, and introduce two state-of-the-art GO methods: FatiGO [24] and Onto-Express (OE) [25]. These four methods are selected to be compared with our approach. We also provide a brief survey of evolutionary and multiobjective optimization. Finally, we characterize the multiobjective optimization problem and define a set of metrics used to evaluate the quality of the results obtained by EMO-CC in comparison with the other methods.

A. Conceptual Clustering

Cluster analysis, or simply clustering, is a data mining technique often used to identify various groupings or taxonomies in databases [26]. Most existing methods for clustering are designed for plain feature-value data. However, sometimes we need to represent structural data that do not only contain descriptions of individual observations, but also relationships between these observations [5]. Therefore, mining structural databases entails addressing both the uncertainty of which observations should be placed together, as well as which distinct relationships among features best characterize different sets of observations [6]. This is more problematic since, *a priori*, we do not know which features are meaningful for a given relationship. Typical clustering techniques are not designed to deal with this [27], even when combined with global feature extraction methods such as principal component analysis or stepwise descendant methods [28], [29]. In contrast, conceptual clustering techniques have been successfully applied to structural databases to uncover concepts that are embedded in subsets of structural data or substructures [5].

While most machine learning techniques applied directly or indirectly to structural databases exhibit methodological differences, they share a five-step framework, even though they use distinct metrics, heuristics, or probability interpretations [5], [30].

- **Database representation**. Structural data can be viewed as a graph containing nodes representing features, linked to other nodes by edges corresponding to their relations. A substructure consists of a subgraph of structural data, which represents an object or concept embedded in the data [5]. These data can be efficiently organized by taking advantage of a naturally occurring structure over the feature space, which consists of a general to specific ordering of possible substructures (i.e., a direct acyclic graph (DAG) [31]).
- **Structure learning**. This process consists of searching through the DAG space for potential substructures, and returning either the best one found or an optimal sample of them. If the number of substructures is superexponential in the number of nodes, different heuristic methods can be applied for this learning process (e.g., greedy [32], hill climbing [32], and genetic algorithms [33]).
- **Cluster evaluation**. The substructure quality is measured by optimizing several criteria, including complexity, where harboring more features always increases the inferential power; support, where a large coverage of the dataset produces good generality; and diversity, where minimal overlapping between clusters generates more distinct clusters and descriptions from different angles [6]. The basic challenge with this approach consists of fixing the potential bias and inflexibility caused by combining these criteria in a weighted sum formula [7], [10].
- **Database compression**. The database compression provides simple representations of the objects in a database. This procedure is often done by selecting the best substructures and replacing their instances by single vertices. However, it may be the case that these summarized substructures need to be decompressed or recompressed when they are combined with different or independent data sources [22].
- **Inference**. New observations can be predicted from previously learned substructures by using classifiers that optimize their matching based on distance [34] or probabilistic metrics [6]. When designed for labeled data, the approach is referred to as supervised learning (as opposed to unsupervised learning) [6].

Here, we exemplify two different methods, one originally designed to work with structural databases.

1) SUBDUE: This method [22] is a typical example of a conceptual clustering approach that finds repeated substructures in databases represented as graphs. SUBDUE starts by looking

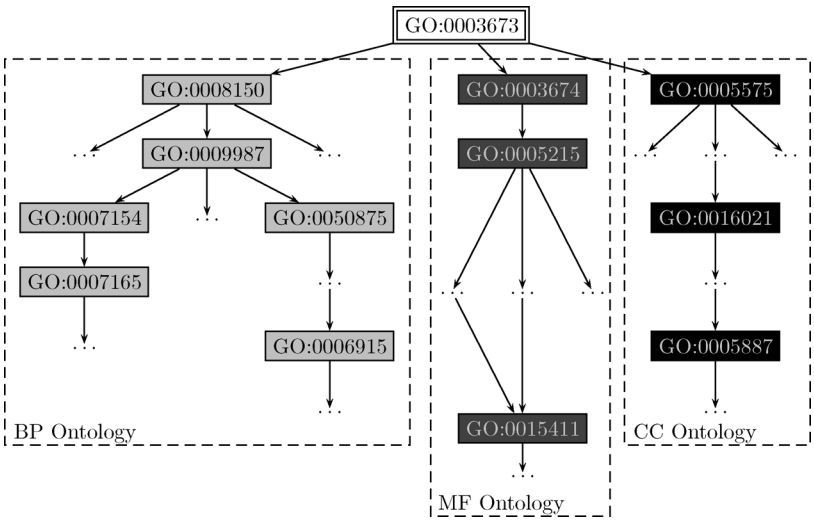

Fig. 2. The GO project ontology. The GO database is composed of three subontologies, which are shown at different colors starting from the root node GO:0003673: Biological process (BP), molecular function (MF), and cellular component (CC).

for the substructure that best compresses the graph using the MDL principle [10], which states that the best description of a dataset is the one that minimizes the description length of the entire dataset. After finding the first substructure, SUBDUE compresses the graph and iterates repeating the same process. SUBDUE uses a computationally constrained beam search strategy to find substructures. The algorithm starts with a single vertex as the initial substructure and at each iteration expands it, using new instances, to explore possible extension edges and potentially generate new substructures. This substructures are recursively considered for expansion.

2) APRIORI: This method uses a classic algorithm for learning association rules [23]. It is designed to operate on databases containing transactions (e.g., collections of items bought by customers, or items of a website access). The algorithm attempts to find subsets of items (e.g., sets of retail transactions of each listing individual items purchased) shared by at least a minimum number of observations. This approach is similar to the computation of biclusters, as it allows simultaneous clusterings of features and transactions [35]. APRIORI uses a bottom-up approach, where frequent subsets are extended one item at a time and evaluated by their supporting observations. The algorithm terminates when no further successful extensions are found. APRIORI uses breadth-first search and a hash tree structure to efficiently search for the best substructures. This algorithm was originally designed to work with plain data, and here adapted to manage structural databases.

B. The Gene Ontology (GO) Project

The GO project [3] stores one of the most powerful characterizations of genes. It uses three structured vocabularies (i.e., ontologies) to describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner [3]. The GO terms are organized as hierarchical networks, where each level corresponds to a different specificity definition of such terms (i.e., higher level terms are more general than lower level terms, Fig. 2). From the computational point-of-view, these networks are organized as structures termed DAGs, which are one way routed graphs that can be represented as trees.

There are many tools developed to extract relevant GO terms from a group of given genes. We select two of the most representative GO clustering methods for comparison with our approach: FatiGO [24] and OE [25].

1) FatiGO: This method carries out a clustering process that assigns a ranking of GO terms to a query, which is often composed of coexpressed genes. GO terms are related to human, mouse, rat, arabidopsis, fly, worm and yeast genes, and proteins. FatiGO implements Fisher's exact test for 2×2 contingency tables to compare two groups of genes and to extract a list of GO terms whose distribution among the groups is significantly different. The results of the test are corrected for multiple-testing to obtain an adjusted p-value. These results are displayed in HTML and text format, which includes a tree representation of GO terms associated with the query and the number of genes annotated with a specific GO term [24].

2) Onto-Express (OE): This method maps queries based on coregulated genes into functional profiles, each one built based on individual GO terms. The significance of functional profiles is calculated by using the binomial distribution for each functional category, which allows distinctions between significant biological processes and random events [25].

C. Evolutionary and Multiobjective (MO) Optimization

Evolutionary algorithms (EAs) [17] are often used to solve knowledge discovery or data mining problems. Several EAs have been successfully applied in classical clustering problems including hard and fuzzy c-means functional optimization [36] and estimation of optimal number of clusters [37]. Moreover, genetic algorithms in combination with multiobjective optimization techniques have been used for selecting features in an unsupervised fashion [38] and for developing multiclassifiers [39]. Linguistic and association rules also incorporated evolutionary techniques for their optimization and searching processes [40], [41]. Indeed, biclustering techniques, often

used in bioinformatics [42], use an appropriate combination of EAs and multiobjective optimization [43].

We incorporate some of the former features successfully applied to knowledge discovery to develop a novel evolutionary method focused on conceptual clustering data. Here, we describe the multiobjective optimization problem and the general notation used throughout this work.

1) Notation: Let us consider, without loss of generality, a multiobjective minimization problem with m decision variables (parameters) and n objectives

$$\begin{aligned} \text{Minimize} \quad & y = f(x) = (f_1(x), \cdots, f_n(x)) \in Y \\ \text{where} \quad & x = (x_1, \ldots, x_m) \in X \end{aligned} \tag{1}$$

where x is called the **decision vector**, X is the **parameter space**, y is the **objective vector**, Y is the **objective space**, and $f_1 \ldots f_n$ are the objective functions. A decision vector $a \in X$ is said to **dominate** a decision vector $b \in X$ (also written as $a \prec b$) if, and only if

$$\begin{aligned} &\forall i \in \{1, \ldots, n\} : \quad f_i(a) \leq f_i(b) \quad \wedge \\ &\exists j \in \{1, \ldots, n\} : \quad f_j(a) < f_j(b) \end{aligned} \tag{2}$$

It is also customary to write any of the following:

- $a \preceq b$ if and only if $a \prec b$ and $f(a) = f(b)$, or $a \npreceq b$;
- a is **nondominated** by b;
- a belongs to a **Pareto-optimal** set, if it is not dominated by b.

2) Evaluation Metrics: We evaluate the performance of MO algorithms by applying a set of metrics that analyze the objective space of solutions [18], [44], [45]. We also use two alternative metrics to perform pairwise comparisons between algorithms [9], [45].

Given a set of pairwise nondominated decision vectors $X' \subseteq X$; a Pareto-optimal set $\bar{X} \subseteq X$; $Y', \bar{Y} \subseteq Y$ the sets of objective vectors that correspond to X' and $\bar{X}$, respectively; a neighborhood parameter $\sigma^* > 0$ (to be chosen appropriately); and a distance metric $\|\cdot\|^*$, we define the following metrics.

- The function $\mathcal{M}_2^*$ evaluates the distribution of the solutions in combination with the number of nondominated solutions

$$\mathcal{M}_2^*(Y') = \frac{1}{|Y' - 1|} \sum_{p' \in X'} \left| \{ q' \in X'; \|p' - q'\|^* > \sigma^* \} \right|. \tag{3}$$

- The function $\mathcal{M}_3^*$ evaluates the extent of the front described by Y'

$$\mathcal{M}_3^*(Y') = \sqrt{\sum_{i=1}^{n} \max \left\{ \|p_i' - q_i'\| ; p', q' \in Y' \right\}}. \tag{4}$$

Finally, we use two other binary metrics for comparing two Pareto sets.

- The metric $\mathcal{C}(X', X'')$ [45] measures the dominance relationship between the set of nondominated solutions $X' \subseteq X$ on another set of nondominated solutions $X'' \subseteq X$. The value $\mathcal{C}(X', X'') = 1$ means that all solutions in X'' are dominated by solutions in X'. The opposite, $\mathcal{C}(X', X'') = 0$, represents the situation where none of the solutions in X'' are covered by the set X'. Note that both $\mathcal{C}(X', X'')$ and $\mathcal{C}(X'', X')$ have to be considered, since $\mathcal{C}(X', X'')$ is not necessarily equal to $\mathcal{C}(X'', X')$

$$\mathcal{C}(X', X'') = \frac{\left| \{ a'' \in X''; \exists a' \in X' : a' \preceq a'' \} \right|}{|X''|}. \tag{5}$$

- The metric $\mathcal{ND}(X', X'')$ [9] compares two sets of nondominated solutions and provides the number of solutions of $X' \subseteq X$ neither equal nor dominated by any member of $X'' \subseteq X$. Once again, both $\mathcal{ND}(X', X'')$ and $\mathcal{ND}(X'', X')$ must be calculated

$$\begin{aligned} \mathcal{ND}(X', X'') = \big| \{ & a' \in X' \wedge a' \notin X'' : \\ & (\forall a'' \in X'' : a'' \npreceq a' \} \big|. \end{aligned} \tag{6}$$

There is a clear difference between the metric $\mathcal{ND}$ and the previous metric $\mathcal{C}$. The former metric counts the number of novel solutions belonging to a Pareto set that are not included in the other, while the latter shows the dominance relationship between two sets of solutions. These metrics are applied and customized for the biological problem in Section IV.

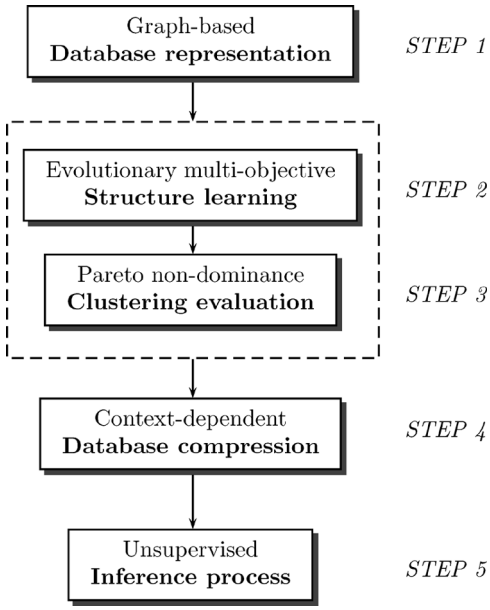

Fig. 3. The EMO-CC methodology. The different steps of EMO-CC are developed based on the typical phases of a conceptual clustering method. The dashed box represents the search and evaluation iterative process carried out by the multiobjective EA.

III. Methodology: Evolutionary Multiobjective Conceptual Clustering (EMO-CC)

In this section, we describe the EMO-CC methodology in terms of each of the steps of the conceptual clustering framework previously introduced (Fig. 3).

A. Graph-Based Database Representation (STEP 1)

The input of the EMO-CC methodology is a graph-based database, which includes feature-values that usually map to nodes, and relationships between them that map to edges. For example, given a database corresponding to a geometrical domain [Fig. 1(b)–(c)], the nodes of the graph correspond to geometric properties, (e.g., *circle*), while the edges of the graph correspond to the relationships between them (e.g., *on*).

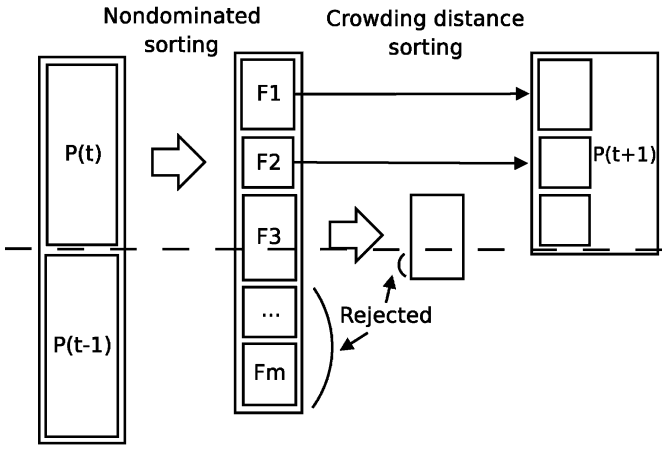

Fig. 4. The EMO-CC structure learning procedure based on the NSGA-II algorithm [19] corresponding to the iterative process of STEPS 2 and 3.

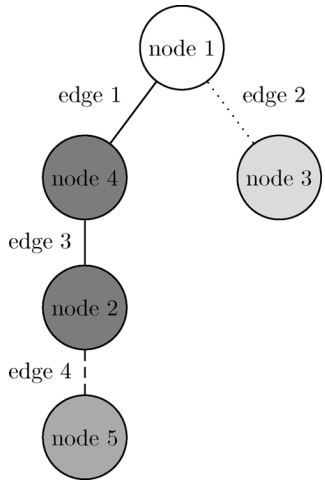

Fig. 5. Example of an EMO-CC chromosome. This representation encodes each node and each edge of the tree with a tag, which corresponds to the type of feature being described. We show node-tags using different colors (e.g., blue, white, yellow) and edge-tags using different line styles (e.g., solid, dotted, dashed). Each node and edge also has an associated tag-value that indicates the value of such feature (e.g., node 1, edge 3).

B. Evolutionary Multiobjective Structure Learning (STEP 2)

The main objective of EMO-CC is to identify optimal substructures in a structural database by searching in the feature space using an efficient multiobjective EA. To do so, we apply an evolutionary search on the space of a graph-represented database relying on a multiobjective EA termed nondominated Sorting GA-II (NSGA-II) developed by Deb *et al.* [19], [46]. A short description provided by Deb *et al.* [19] is the following.

The step-by-step procedure shows that NSGA-II algorithm is simple and straightforward. First, a combined population $R_t = P_t \cup Q_t$ is formed. The population R_t is of size $2N$. Then, the population R_t is sorted according to nondomination. Since all previous and current population members are included in R_t, elitism is ensured. Now, solutions belonging to the best nondominated set F_1 are of best solutions in the combined population and must be emphasized more than any other solution in the combined population. If the size of F_1 is smaller then N, we definitely choose all members of the set F_1 for the new population P_{t+1}. The remaining members of the population P_{t+1} are chosen from subsequent nondominated fronts in the order of their ranking. Thus, solutions from the set F_2 are chosen next, followed by solutions from the set F_3, and so on. This procedure is continued until no more sets can be accommodated. Say that the set F_i is the last nondominated set beyond which no other set can be accommodated. In general, the count of solutions in all sets from F_1 to F_i would be larger than the population size. To choose exactly N population members, we sort the solutions of the *last* front F_i using the crowded-comparison operator $\prec_n$ in descending order and choose the best solutions needed to fill all population slots. The NSGA-II procedure is also shown in Fig. 4. The new population P_{t+1} of size N is now used for selection, crossover, and mutation to create a new population Q_{t+1} of size N. It is important to note that we use a binary tournament selection operator but the selection criterion is now based on the crowded-comparison operator $\prec_n$. Since this operator requires both the rank and crowded distance of each solution in the population, we calculate these quantities while forming the population P_{t+1}, as shown in the above algorithm.

The components of the EMO-CC structure learning process are described as follows:

1) Chromosome Representation: EMO-CC encodes only feasible substructures in each chromosome. We implement chromosomes as trees, which is the typical representation used in genetic programming (GP) [47]. The GP evolutionary approach and its multiobjective variants [48], [49] have been widely used to solve many different real-world problems, including system identification [50], information retrieval [51], [52], or data mining [53], achieving successful results. This chromosome representation encodes each node and each edge of the tree with a tag, describing the type of feature, and an associated tag-value that indicates the value of such feature (Fig. 5). The initial population consists of a set of chromosomes, each one generated by choosing a random observation from the input database and representing it as a subtree. The set of all nondominated chromosomes of the final population represents an optimal partition of the given data.

2) Genetic Operators: EMO-CC applies crossover and mutation operators with a given probability over the chromosomes composing the GP population. The crossover operator is performed by swapping two random subtrees, which is a classical choice in GP. The mutation operators used in our GP implementation are also classical and straightforward.

- *Delete a leaf*, where a random leaf of the tree is selected and deleted along with the edge that connects it to the tree.
- *Change a node*, where a random node is selected and replaced by another node belonging to the set of nodes constrained by the same tag.
- *Add a leaf*, where a random leaf is created and connected to the tree by a new edge.

Each type of node has associated a different tag in the chromosome representation that constrains crossover and mutation operators (e.g., in Fig. 5 blue nodes are only allowed to be compared with other blue nodes, and solid edges with other solid edges). Therefore, the crossover operation can be applied if, and only if, the root of both exchanged subtrees has the same tag,

thus allowing to maintain feasible offspring chromosomes. This particular chromosome representation with constraints is known as *type-based* GP [47].

3) Selection: EMO-CC employs a crowded binary tournament selection operator [17], [19] preserving the diversity among nondominated solution. Assuming that every individual in the population has two attributes: nondomination rank (i_{rank}) and crowding distance (i_{distance}), the crowded operator $\prec_n$ is defined as

$$i \prec_n j \quad if (i_{\text{rank}} < j_{\text{rank}}) \; or \; ((i_{\text{rank}} = j_{\text{rank}}) \; and \; (i_{\text{distance}} > j_{\text{distance}})) \quad (7)$$

where the quantity i_{distance} serves as an estimate of the perimeter of the cuboid formed by using the nearest neighbors as the vertices (call this the crowding distance) and i_{rank} specifies the level of nondomination of the solution (e.g., level 0 for nondominated solutions, level 1 for solutions dominated by only one solution). That is, between two solutions with differing nondomination ranks, we prefer the solution with the lower (i.e., better) rank. Otherwise, if both solutions belong to the same front, then we prefer the solution that is located in a lesser crowded region. See [19] for a complete description.

4) Fitness Functions: We consider that good substructures are those that maximize the *complexity* and the *support* objectives. On the one hand, the complexity of a substructure s is associated with its size (i.e., the number of nodes and edges that compose the substructure), which corresponds to the size of the tree represented in the chromosome

$$\text{Complexity(s)} = \#\text{nodes(s)} + \#\text{edges(s)}. \quad (8)$$

On the other hand, the support of a substructure s is calculated as the number of database instances that occur in the substructure

$$\text{Support(s)} = \#\text{instances(s)} \quad (9)$$

where an instance of a substructure occurs in an observation of the dataset if the instance tree is a subtree of the observation tree. These are conflicting objectives since the more complex the substructure, the smaller its support [7], [8].

C. Pareto Nondominance Substructure Evaluation (STEP 3)

We evaluate the quality of the substructures based on a multiobjective strategy that retrieves cohesive and well supported solutions [7], [17]. Indeed, we use a multimodal optimization approach to uncover diverse results [54]. To do so, we search for a set of solutions that are nondominated, based on their complexity and support, in the sense that there is no other solution that is superior in all of the objectives (i.e., Pareto optimal front) in a neighborhood [17]. Therefore, optimal solutions corresponding to different search spaces do not compete among each other. We implement this multimodal strategy using niches [7], [8], which are groups of substructures covering a common set of instances. The scope of a niche is calculated by using the Jaccard's index [55]

$$\text{jaccard}(s_i, s_j) = \frac{s_i \cap s_j}{s_i \cup s_j} \quad (10)$$

where s_i and s_j are substructures.

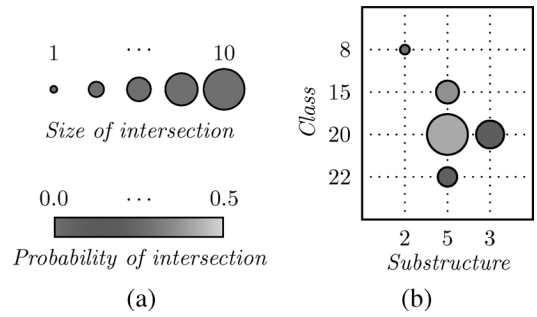

Fig. 6. Validating and compressing substructures. (a) We evaluate the ability of a substructure to describe an independent class by using the probability of intersection (red: high; green: low) among their recognized instances. This metric slightly differs from the size of the intersection (size of the circle), and allows identification of more cohesive relationships. (b) A class can be explained and summarized by more than one substructure, which are then compressed and become indistinguishable for the given class. However, other substructures remain as diverse explanations of the same class (e.g., substructures 3 and 5, both of which describe class #20).

Consequently, we reformulate the dominance criterion, where a substructure s_i dominates another substructure s_j

$$s_i \prec s_j \quad iff \quad \begin{array}{l}(\text{jaccard}(s_i, s_j) > \delta) \text{ and} \\ (f_1(s_i) \geq f_1(s_j) \text{ and } f_2(s_i) \geq f_2(s_j)) \text{ and} \\ (f_1(s_i) > f_1(s_j) \text{ or } f_2(s_i) > f_2(s_j))\end{array} \quad (11)$$

with f_1 and f_2 being the observed functions that measure the complexity and support of the substructures, respectively; and $\delta = 0.5$ is the niche size. This is the less biased initialization value, which is calculated as a tradeoff between redundant and smaller number of accepted substructures [7]. The use of this modified dominance criterion means that they cover different instances in the same class (i.e., that they are largely nonoverlapping).

D. Context-Dependent Database Compression (STEP 4)

EMO-CC identifies diverse but nonredundant substructures that explain classes derived from additional information. This constitutes a compression process [22] based on circumstantial queries that allows flexible and context dependent summarization of the substructures. This step is divided into two processes. 1) *Validating substructures*, where we reevaluate the quality of the substructures identified by a conceptual clustering method by their ability to explain a set of classes derived from an independent experiment. We use the hypergeometric measurement in a test that represents the probability of observing at least k instances from a specific class c, derived from an independent experiment, within an identified substructure s of size n [56]

$$PI(c, s) = 1 - \sum_{i=0}^{k-1} \frac{\binom{v}{i}\binom{w-v}{n-i}}{\binom{w}{n}} \quad (12)$$

where v is the total number of instances in c, and w is the total number of instances in the database. The lower the probability of intersection PI, the better the quality of the cluster intersection, and thus, the greater the confidence in the detected substructure (Fig. 6). This approach differs from supervised learning methods, which are just focused on learning substructures based

COMPRESS (S set of substructures, C set of classes, γ threshold)

1: $W \leftarrow \emptyset$
2: **for all** $c \in C$ **do**
3: $Q \leftarrow \{s \in S \mid PI(s,c) < \gamma\}$
4: $Q' \leftarrow \{q \in Q \mid (\neg\exists x \in Q \mid (x \prec q))\}$
5: Create a graph structure Y using each $q \in Q'$ as a node.
6: **for all** $x, y \in Q', x \neq y$ **do**
7: Connect node $a \in Y$, corresponding to substructure x, with node $b \in Y$, corresponding to substructure y, if each of the GO terms describing substructure x has a GO term in substructure y belonging to the same branch in the GO hierarchy.
8: **end for**
9: $Q'' \leftarrow \emptyset$
10: **for all** Y_i *connected graph* $\in Y$ **do**
11: $Q'' \leftarrow Q'' \cup \{c \in Y_i \mid (\forall x \in Y_i \mid c \text{ is more general than } x)\}$
12: **end for**
13: $W \leftarrow W \cup Q''$
14: **end for**
15: Return W

Fig. 7. Context-dependent database compression algorithm. The quality of explaining an external class by a set of substructures S is evaluated by using a statistical test based on the hypergeometric measure (PI). Irrelevant substructures are filtered (3–4) and relevant substructures are compressed (5–12) to provide more clear and understandable explanations.

on their ability of discriminating among output classes. 2) *Filtering and compressing validated substructures*, where we select substructures that explain independent classes with a significant PI [27] [Fig. 7, (3)]. Then, we compress a set of substructures if they can explain the same phenomena by applying the nondominance relationship to the subset of instances covered by each substructure [(11), Fig. 7, (4–12)]. The nondominance relationship needs to be recalculated (i.e., both specificity and sensitivity objectives) since some of the instances covered by each substructure do not belong to the given class, thus, we remove them from the set. These compressed substructures can be indistinguishable for a specific class in a particular experiment, but they can be reversed and regrouped in a different fashion under another experimental conditions.

E. Unsupervised Inference Process (STEP 5)

The EMO-CC methodology classifies new instances based on a set of optimal substructures resulting from previous steps of the algorithm. We use a k-nearest neighbor unsupervised classifier to explain new instances by their similarity with one or more substructures. Thus, we classify a query observation x based on a set of n substructures previously learned as follows:

$$\begin{aligned} knn(x,S) &= knn(x,s_1,\ldots,s_n) \\ &= s_i \in S/\mu(x,s_i) \\ &= \max\{\mu(x,s_1),\ldots,\mu(x,s_n)\}, \quad \forall i \end{aligned} \tag{13}$$

where μ represents the degree of matching between an observation x and a substructures s. The degree of matching μ is equal to the substructure complexity (8), where x can belong to one or more substructures with a different membership degree [34].

IV. The Inflammatory Response Problem

The host's 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. No single research center or small group of centers have the resources to delineate the integrated response of this complete biological system, which involves multiple molecular and genetic interactions that vary in time. This study, in part carried out at the Cellular Injury and Adaptation Laboratory, Washington University School of Medicine, is a piece of a large-scale research project devoted to profile leukocyte gene expression and plasma proteins of burn and trauma patients [21]. Prior to initiating studies in actual patients, it was proposed that the human endotoxin model could serve as a starting point and test bed for these subsequent studies. Our proposal will help to promote the identification of significant relationships, which regulate the integration of this complex biological system, with the expectation that this understanding will ultimately impact the treatment of hospitalized patients.

We analyzed 48 GeneChips HG-U133A v2.0 from Affymetrix, Inc., derived from samples taken from human blood of eight patients: four treated with intravenous endotoxin (i.e., patients 1–4) and four with a placebo (i.e., patients 5–8), and expression retrieved over time at hours 0, 2, 4, 6, 9, and 24. The analysis is performed in three steps: 1) we identify 1770 significantly expressed genes that change their expression using a very sensitive approach that combines several statistical methods (e.g., t-tests, permutation tests, analysis of variance and repeated measures ANOVA) [57], [58], which is a reasonable number of genes for a general-purpose inflammation process [21]; 2) we arrange the expression levels of the extracted genes by linking patient 1 hr 0, hr 2, hr 4, hr 6, hr 24, patient 2 hr 0, hr 2, hr 4, hr 6, hr 24, patient 3 hr 0, hr 2, hr 4, hr 6, hr 24, patient 4 hr 0, hr 2, hr 4, hr 6, hr 24 (control expression is arranged in a similar fashion); and 3) we separately cluster treated and control prearranged expression levels into profiles and identify 24 pairwise differential profiles that change over time, treatment, or patient [58] (Fig. 8). For example, differential gene expression among treatment (e.g., Fig. 8, row: 1 column: 1, row: 1 column: 2) and patients (e.g., Fig. 8, row: 1 column: 4, row: 3 column: 1) is explicitly illustrated in several profiles. Using the selected representation, we can distinguish differential profiles that reveal even subtle biological variabilities that are usually difficult to identify by averaging gene expression. We use the GO annotations corresponding to these 1770 genes as our input database, which are provided by the GO project described in Section II-B. These differential profiles will be used as a reference partition in the subsequent analysis of EMO-CC's results.

In the following sections, we provide details for the database representation (STEP 1), and the learning process (STEP 2) of the EMO-CC methodology (Fig. 3) for the inflammatory response problem.

A. Graph-Based Database Representation (STEP 1)

The database representation used for the GO domain can be viewed as a database containing different features, where each feature has nested values denoting descriptions at different

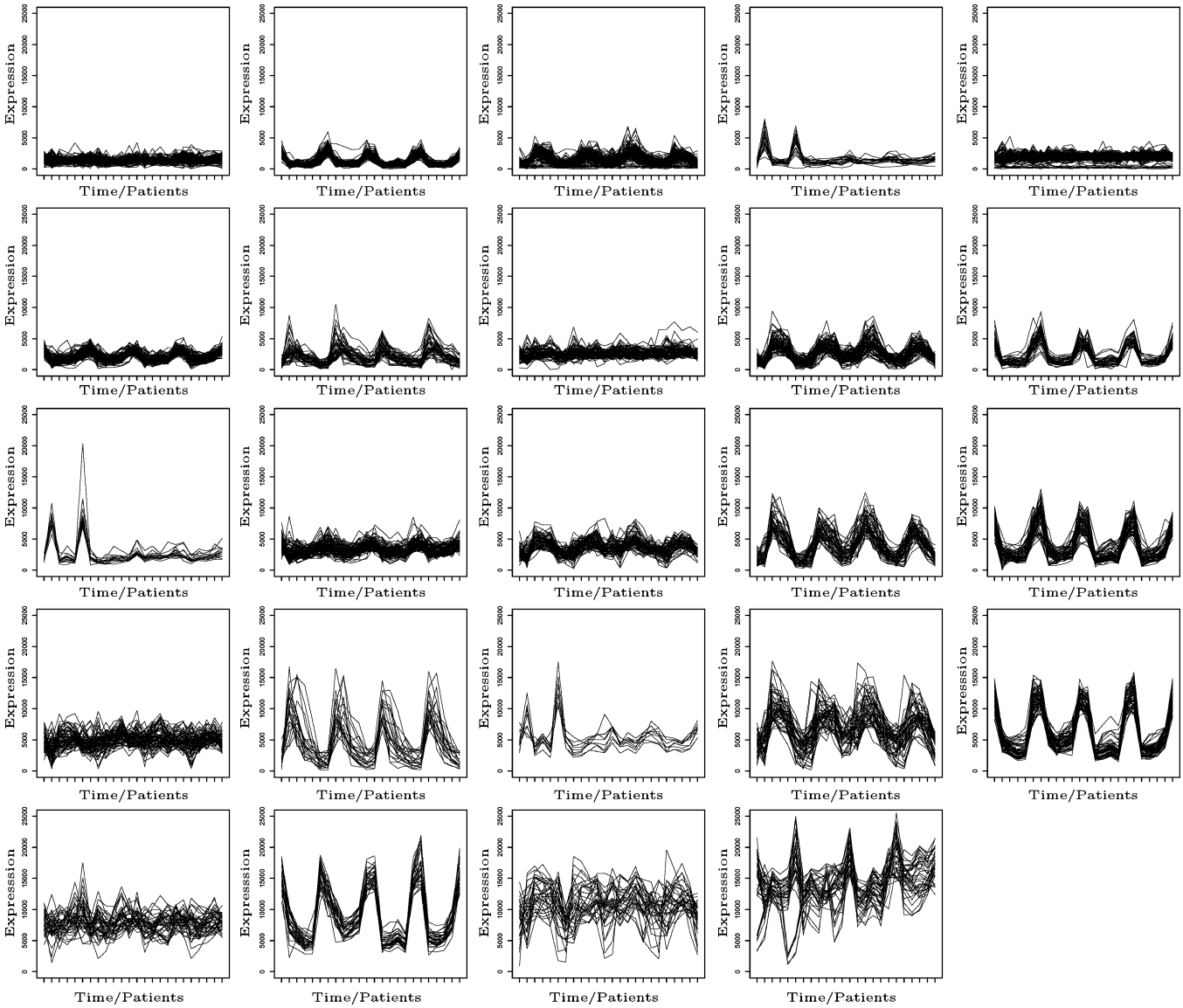

Fig. 8. Gene expression profiles from the inflammatory response problem. We account for eight patients: four treated with intravenous endotoxin (i.e., patients 1–4) and four with a placebo (i.e., patients 5–8), and expression retrieved over time at hours 0, 2, 4, 6, 9 and 24, each one corresponding to different GeneChips and its replicas. The expression profiles are represented separately for each experimental group (i.e., treatment and control), and patients are arranged individually. Each profile is represented by 24 time points: patient 1 hr 0,. . ., patient 1 hr 24,. . ., patient 4 hr 0,. . ., patient 4 hr 24 (horizontal axis). The vertical axis corresponds to the gene expression level. Only expression profiles for the treatment group are shown. Differential gene expression among patients is explicitly illustrated in several profiles (e.g., row: 1, column: 4, row: 3, column: 1). All of these profiles derived from treated patients have their counterpart in control profiles (here not shown), from which they are differentiated.

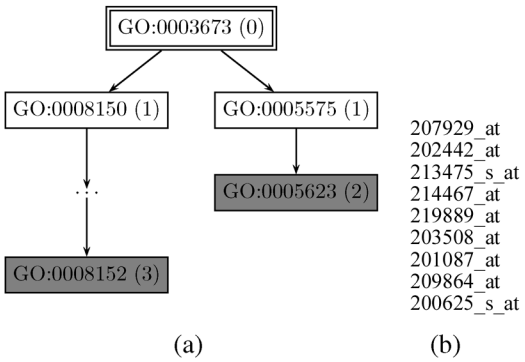

Fig. 9. An example of a chromosome representing a substructure (specificity= 0.6769, support= 0.0051). (a) A tree representation of a substructure, where the gray boxes are the most specific GO terms, and the levels of the terms in the GO hierarchy are shown in parenthesis. (b) The list of genes that corresponds to the substructure.

TABLE I
PARAMETERS FOR THE GO DOMAIN

Parameter	Value
Population Size	200
Number of Evaluations	20000
Crossover probability	0.6
Mutation probability	0.2

levels of specificity. Therefore, in identifying which distinct relationships among features best characterize different sets of observations, we have to consider, not only the process of grouping distinct type of features (e.g., biological process GO:0007165 and GO:0050785, representing a signal transduction process and an advanced glycation end-product receptor activity, respectively, and cellular component GO:0016021, representing an integral to membrane situation), but also

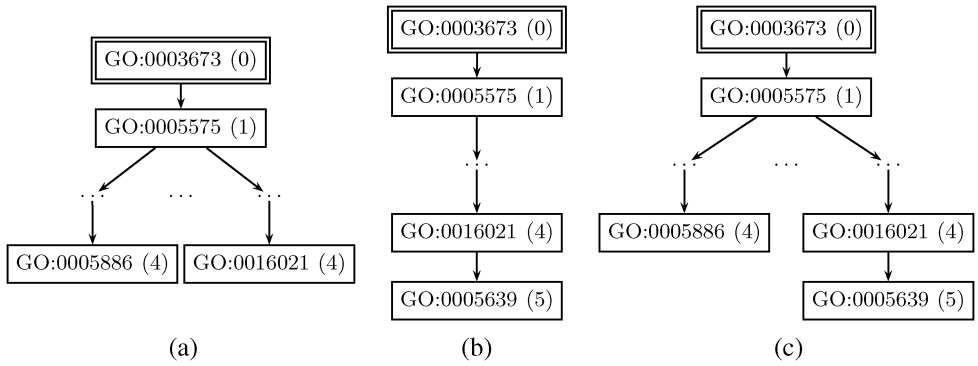

Fig. 10. Relationship between substructures and observations. (a) *Substructure #1* and (b) *Substructure #2* both represent an observation (c). In this example, *substructure #2* is more specific than *substructure #1* and, therefore, more complex, since the leaf nodes from the former belong to level 5, while those of the latter belong to level 4. The double frame box corresponds to the root of the GO, the boxes indicate cellular component terms, and the number in parenthesis correspond to the level of the nodes in the GO hierarchy.

TABLE II
COMPARATIVE EVALUATION OF THE SOLUTIONS IDENTIFIED BY APRIORI, SUBDUE, AND EMO-CC FOR THE GO DOMAIN BY USING DIFFERENT METRICS: (A) $\mathcal{M}_3^*$. (B) $\mathcal{M}_2^*$. (C) $\mathcal{C}$. (D) $\mathcal{ND}$

(a)

	$\mathcal{M}_3^*$
APRIORI	1.1853
SUBDUE	1.1978
EMO-CC average ($stdev$)	1.2334 (0.0047)

(b)

	$\mathcal{M}_2^*$	$\lvert X\rvert$
APRIORI	6	7
SUBDUE	17.56	19
EMO-CC average ($stdev$)	186.8680 (12.2065)	188 (12.19)

(c)

$\mathcal{C}(X', X'')$	APRIORI	SUBDUE	EMO-CC average ($stdev$)
APRIORI	-	0.00000	0.00000 (0.00000)
SUBDUE	0.00000	-	0.00050 (0.00160)
EMO-CC average ($stdev$)	0.00000 (0.00000)	0.08421 (0.04438)	-

(d)

$\mathcal{ND}(X', X'')$	APRIORI	SUBDUE	EMO-CC average ($stdev$)
APRIORI	-	1	1.20 (0.42)
SUBDUE	13	-	1.60 (1.17)
EMO-CC average ($stdev$)	181.80 (11.99)	171.80 (11.62)	-

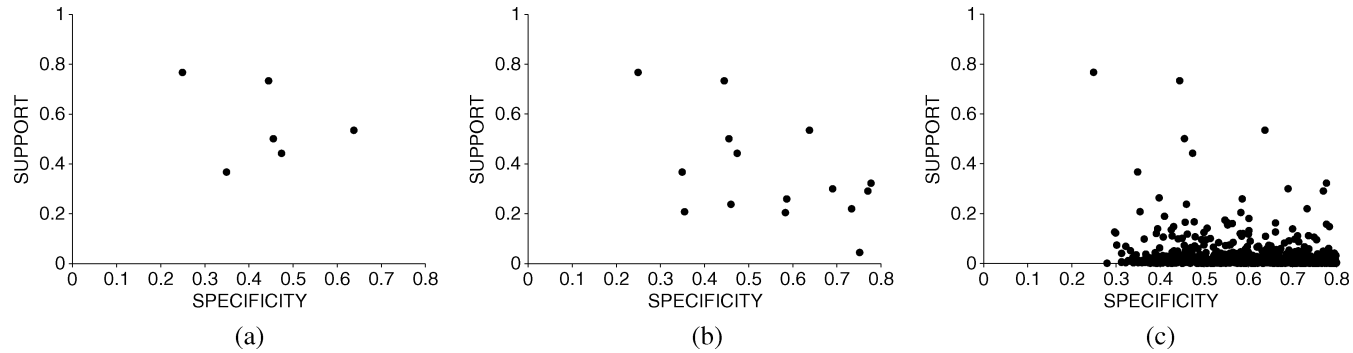

Fig. 11. The Pareto fronts obtained by different methods. Each dot represents a solution with the support given by its value on the y axis, and the specificity given by its value on the x axis. Nondominated solutions reported by: (a) APRIORI, (b) SUBDUE, and (c) EMO-CC.

defining at which level of specificity they have to be represented. This is even more problematic since several values of the same type of feature may be useful for describing a set of observations, and thus, represented in a substructure [e.g., biological process GO:0007165 (level 4) and GO:0050785 (level 3)]. Consequently, to address the problem of the multilevel definition of a feature we redefine an instance as the particular subset of values that constitutes a prefix tree[1] of a database observation. Then, an instance of a substructure occurs in an observation of the database if a subgraph of the prefix tree that represents that instance matches with the observation tree.

[1]Tree t' is a prefix tree of t if t can be obtained from t' by appending zero or more subtrees to some of the nodes in t'. Notice that any tree t is a prefix of itself.

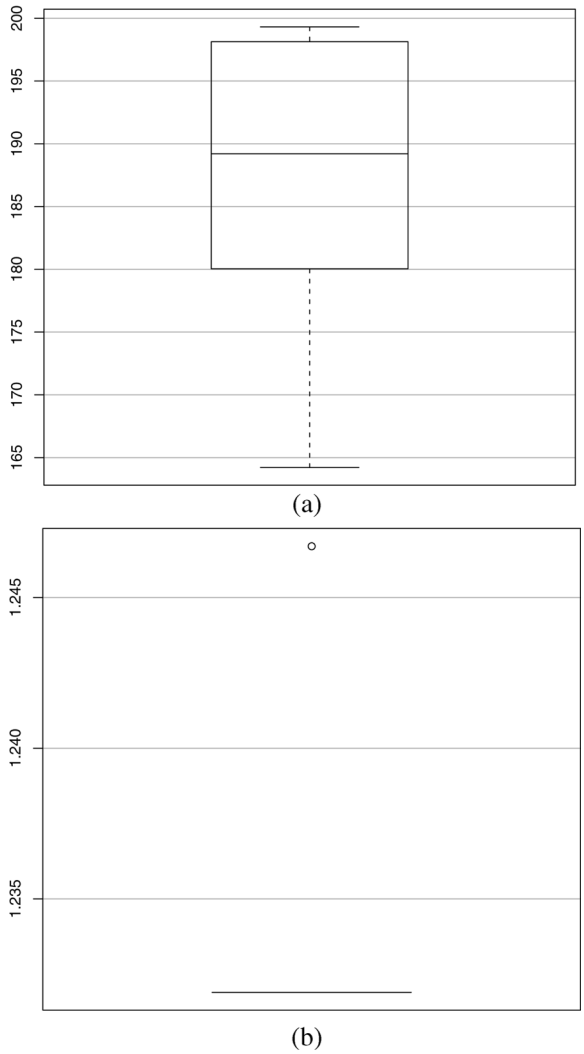

Fig. 12. Boxplots of (a) $\mathcal{M}_2^*$ and (b) $\mathcal{M}_3^*$ metrics for EMO-CC in the GO domain. The boxplots show the resulting values for each metric in different runs of EMO-CC. The smaller the boxes in the boxplot, the more homogeneous results over ten runs.

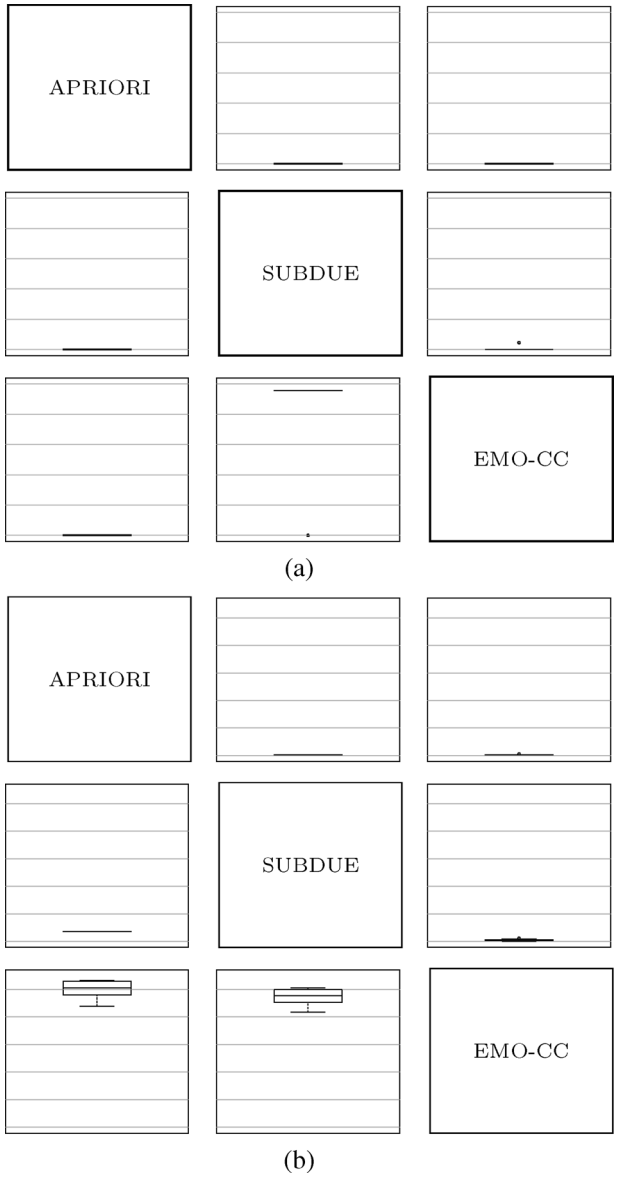

Fig. 13. Boxplots of (a) $\mathcal{C}$ and (b) $\mathcal{ND}$ for the EMO-CC in the GO domain. The boxplots show the resulting values for each metric in different runs of EMO-CC. The smaller the boxes in the boxplot, the more homogeneous results over ten runs.

The substructure tree contains tagged nodes with the type of feature (e.g., biological process), its corresponding value (e.g., GO:0007165), and the edges representing relationships between features (e.g., is_a).

We use the GO database and compatibilize the terms with descriptions provided by Affymetrix (i.e., the type of microarrays used in this study [59]), where each observation of the database has the following features.

- *Name:* Affymetrix identifier for each gene in HG-U133A v2.0 set of arrays.
- *Biological process:* List of biological processes where a gene product is involved. This list is indexed by a list of GO codes [e.g., GO:0007067 (mitosis), GO:0008152 (metabolic process)]. The processes are broad biological goals that are accomplished by ordered assemblies of molecular functions.
- *Molecular function:* List of biological functions of gene products, which are indexed by a list of GO codes [e.g., GO:0030246 (carbohydrate binding), GO:0016887 (ATPase activity)]. These functions are tasks performed by individual gene products.

TABLE III
CLASS #13 AND SUBSTRUCTURE INTERSECTIONS

Substructure	Size	Intersection	Probability of Intersection
179	7	5	2.20×10^{-6}
536	69	12	1.52×10^{-5}
759	42	10	1.43×10^{-6}
256	22	6	1.91×10^{-4}
89	104	14	5.79×10^{-5}
380	18	6	5.43×10^{-5}
607	179	18	2.37×10^{-4}

- *Cellular component:* List of cellular components indicating location of gene products, which are indexed by a list of GO codes [e.g., GO:0005634 (nucleus), GO:0019012 (virion)]. These components are subcellular structures, locations, and macromolecular complexes.

TABLE IV
SUBSTRUCTURES DERIVED FROM THE GO DATABASE BY EMO-CC EXPLAINING CLASS #13 GENE EXPRESSION PROFILE ($PI < 3.1 \times 10^{-4}$)

#Substructure	Biological process	Molecular function	Cellular component
179	GO:0006915 apoptosis (level: 6)		GO:0005887 integral to plasma membrane (level: 4)
536	GO:0007165 signal transduction (level: 4)		GO:0016021 integral to membrane (level: 3)
759	GO:0007165 signal transduction (level: 4)		GO:0005887 integral to plasma membrane (level: 4)
89	GO:0007154 cell communication (level: 3)		GO:0016021 integral to membrane (level: 3)
256	GO:0007154 cell communication (level: 3) GO:0050875 cellular physiological process (level: 3)		GO:0016021 integral to membrane (level: 3)
380	GO:0007165 signal transduction (level: 4) GO:0050875 cellular physiological process (level: 3)		GO:0016021 integral to membrane (level: 3)
607		GO:0004871 signal transducer activity (level: 2)	GO:0016021 integral to membrane (level: 3)

B. Multiobjective GP Structure Learning (STEP 2)

The chromosome representation used in the GO domain is a tree-like structure (Fig. 9). Each node of this tree corresponds to a GO term, and each edge corresponds to a `is_a` or `part_of` relationship.

The complexity of the substructures in the GO domain is not linearly dependent on its size (8). This happens because the GO ontology is composed of terms that can be located at different levels in the hierarchy. For example, a substructure (substructure #1) is less specific than another substructure (substructure #2), if the leaf nodes from the former belong to a lower level (level 4) than the latter (level 5) (Fig. 10). However, by calculating the complexity as the number of edges plus nodes of each substructure, the first substructure reaches a higher evaluation value (i.e., $\text{complexity} = 8$) than the second (i.e., $\text{complexity} = 7$). Thus, we redefine the complexity as *specificity*, extending the original objective by including not only the size of the substructure measured by the number of nodes and edges, but also the accuracy of the substructure in modeling the covered instances

$$\text{Specificity}(s) = \frac{\sum_{i=1}^{k}\left(1 - \sum_{u=1}^{l}\frac{\text{dist}(\text{node}_{ui}, s)}{\text{level}(\text{node}_{ui})}\right)}{k} \qquad (14)$$

where k is the number of instances occurring in substructure s, l is the number of leaf-nodes in the ith instance occurring in substructure s, and node_{ui} is a leaf-node of the ith instance occurring in substructure s. The distance dist between a node and a substructure is calculated as the number of edges between the given node and its closest ancestor in the GO hierarchy appearing in the given substructure. The level of a node is calculated as the length of the shortest path to the root node. For example, to obtain a $\text{specificity} = 1$ all nodes of the instance must appear in the substructure and their distances to the substructure must be zero.

V. EXPERIMENTS AND ANALYSIS OF RESULTS

The structural database used for the GO domain is composed of 1770 significantly expressed genes, extracted from the set of the total genes available in a GeneChip (i.e., approximately 22 000), and their GO associated terms. The population of the EA is initialized by 50% of randomly chosen subtrees from the database, and by another 50% of random trees. This randomization procedure is needed to avoid the potential bias introduced in the search process using only a subset of GO terms instead of the complete GO database.

We execute EMO-CC ten times with different seeds and a set of parameters that maximizes the computational performance (Table I). We analyze the sensitivity of the parameters, increasing the population (e.g., from 200 to 800) and changing the operator probabilities (e.g., crossover from 0.6 to 0.9 and mutation from 0.1 to 0.3). The similar results obtained by this analysis suggests that the NSGA-II has a robust behavior. Then, we used the average of the ten runs to report the results evaluated by the metrics $\mathcal{M}_2^*$, $\mathcal{M}_3^*$, $\mathcal{C}$, and $\mathcal{ND}$.

In the following sections, we show the experimental results obtained by EMO-CC in the inflammatory response problem (STEPS 3–5). In the first section (STEP 3), we compare EMO-CC with two other methods: the conceptual clustering method SUBDUE [22], and the APRIORI unsupervised method [60], which is adapted for using structural data. In the second subsection (STEP 4), we perform a context-dependent database compression of the learned substructures that can explain gene expression. Also, we introduce a comparison with two other

state-of-the-art GO mining methods, FatiGO [24] and OE [25]. Finally, the last section (STEP 5) shows the results obtained in the inference process, allowing to predict new substructure members.

A. Pareto Nondominance Clustering Evaluation (STEP 3)

Since both APRIORI and SUBDUE methods are not MO algorithms, we remove from the final set of solutions of both methods those solutions that are dominated, to provide a comparable set of substructures. For APRIORI, we also transform the original structural database into a plain repository by adding all parent terms for each GO term used in the biological application. The results for a single run are reported. We show the union of the results obtained by SUBDUE from three runs, each one using a different optimization criteria, including: support (i.e., the number of instances occurring in a substructure), complexity (i.e., the size of the substructure calculated as the number of bits needed to encode the adjacency matrix corresponding to the graph [22]), and a weighted sum metric that combines the latter two (i.e., MDL [10]), which is the default option of SUBDUE. The results obtained by APRIORI and SUBDUE are compared with each of the Pareto sets found by EMO-CC, when using the MO evaluation metrics.

The substructures recovered by EMO-CC obtain a better coverage of the Pareto front extent than SUBDUE and APRIORI, as reported by metric $\mathcal{M}_3^*$ [Table II(a)]. The results also reveal that EMO-CC obtains the most diverse substructures, as evaluated by metric $\mathcal{M}_2^*$ [Table II(b)] and illustrated by the Pareto fronts (Fig. 11). Moreover, EMO-CC obtains robust results when evaluated for ten runs (Fig. 12).

In addition, we apply two other metrics, $\mathcal{C}$ and $\mathcal{ND}$, to compare the Pareto sets of the different methods. To do so, we modify both metrics replacing the classical nondominance criterion by the one introduced in Section III-C (10) to account for diversity. The obtained results reveal that there is no solution found by EMO-CC that is dominated by APRIORI, and only one solution obtained by SUBDUE dominates a solution belonging to the EMO-CC Pareto set, as described by metric $\mathcal{C}$ [Table II(c) and Fig. 13(a)]. Moreover, EMO-CC discovers more nondominated solutions than both APRIORI and SUBDUE, as evaluated by metric $\mathcal{ND}$ [Table II(d) and Fig. 13(b)]. The difference between the values reported by the $\mathcal{ND}$ metric from EMO-CC and those from APRIORI and SUBDUE [i.e., 181.89 and 171.80 versus 1.20 and 1.60 from Table II(d)] suggests that EMO-CC retrieves almost all solutions identified by the other methods and covers a wide set of all optimal solutions in the GO domain. Moreover, both APRIORI and SUBDUE obtain a limited number of nondominated solutions in comparison with the EMO-CC methodology (Fig. 11 and Appendix Table VII, available at http://www.ieee-explore.ieee.org). Besides, EMO-CC extracts more diverse solutions, in the objective space, than those found by APRIORI and SUBDUE. Particularly, our approach retrieves substructures of the Pareto optimal front containing few instances but harboring several features (i.e., cohesive substructures), which were undetected by the other methods. Moreover, EMO-CC finds diverse solutions in the variable space due to the niching strategy used in the nondominance measure (Section III-C).

TABLE V
RESULTS OBTAINED FROM THE FATIGO METHOD. (A) BIOLOGICAL PROCESS (BP). (B) MOLECULAR FUNCTION (MF). (C) CELLULAR COMPONENT (CC)

(a)

Biological process	#genes	Percentage
cell communication	28	37.83%
regulation of biological process	23	31.08%
establishment of localization	11	14.86%
defense response	10	13.51%
cell organization and biogenesis	10	13.51%
response to external stimulus	9	12.16%
response to stress	9	12.16%
protein localization	7	9.45%
death	6	8.10%
cell adhesion	5	6.75%
response to biotic stimulus	4	5.40%
cell proliferation	3	4.05%
cell activation	3	4.05%
cell homeostasis	2	2.70%

(b)

Molecular function	#genes	Percentage
protein binding	34	0.45%
receptor activity	11	0.14%
ion binding	9	0.12%
nucleotide binding	7	0.09%
transferase activity	6	0.08%
nucleic acid binding	6	0.08%
oxidoreductase activity	5	0.06%
hydrolase activity	3	0.04%
cofactor binding	2	0.02%
small protein conjugating enzyme activity	2	0.02%
GTPase regulator activity	2	0.02%
ligase activity	2	0.02%
transcriptional activator activity	2	0.02%
receptor signaling protein activity	2	0.02%

(c)

Cellular component	#genes	Percentage
cell	45	0.60%
membrane-bound organelle	16	0.21%
extracellular region	5	0.06%
non-membrane-bound organelle	5	0.06%
organelle	4	0.05%

The examination of the results obtained by APRIORI and SUBDUE suggests that their deficiencies can be attributed to (i) the linearization of the database in the APRIORI method, which constrains the data representation; (ii) the thresholds used in APRIORI, which discard substructures with few members, even if they cohesively share several features; and (iii) the inflexibility caused by weighting the evaluation objectives in SUBDUE (i.e., complexity and support) into a single function, which can constrain the set of solutions to a single or limited region of the search space.

B. Context-Dependent Database Compression Using Gene Expression Profiles (STEP 4)

We use 24 gene expression profiles (Fig. 8), which constitute the independent classes used for validating the substructures detected by the three methods previously described, or, in other words, which can be explained by these substructures. For example, class #13 constitutes a differential gene expression profile that changes between treatment and control gene expression (Fig. 15). This class is described by several substructures identified by EMO-CC, including substructure #89, #179, and #256,

TABLE VI
RESULTS OBTAINED FROM THE ONTO-EXPRESS (OE) METHOD.
(A) BIOLOGICAL PROCESS (BP). (B) MOLECULAR FUNCTION (MF). (C) CELLULAR COMPONENT (CC)

(a)

Biological process	#genes	Percentage
immune response	13	17.55%
cell surface receptor linked signal transduction	7	12.15%
cytokine and chemokine mediated signaling pathway	3	4.05%
positive regulation of I-kappaB kinase/NF-kappaB cascade	5	6.75%
inflammatory response	6	8.10%
response to pathogenic bacteria	2	2.70%
signal transduction	16	21.60%
membrane fusion	2	2.70%
intracellular signaling cascade	7	12.15%
response to virus	3	4.05%
JAK-STAT cascade	2	2.70%
apoptosis	6	8.10%
cell motility	4	5.40%
cell-cell signaling	5	6.75%
cellular defense response	2	2.70%
transmembrane receptor protein tyrosine kinase signaling pathway	2	2.70%
cell adhesion	5	6.75%
positive regulation of cell proliferation	2	2.70%
protein transport	3	4.05%
protein complex assembly	2	2.70%
ubiquitin cycle	3	4.05%
small GTPase mediated signal transduction	2	2.70%
cell proliferation	3	4.05%
electron transport	2	2.70%
protein amino acid phosphorylation	3	4.05%
proteolysis	2	2.70%
regulation of transcription, DNA-dependent	7	9.45%
development	2	2.70%
transcription	4	5.40%

(b)

Molecular function	#genes	Percentage
IgG binding	2	2.70%
interferon-gamma receptor activity	2	2.70%
transmembrane receptor activity	6	8.10%
cytokine binding	2	2.70%
protein binding	27	36.48%
hematopoietin/interferon-class (D200-domain) cytokine receptor signal transducer activity	2	2.70%
hematopoietin/interferon-class (D200-domain) cytokine receptor activity	3	4.05%
ubiquitin conjugating enzyme activity	2	2.70%
receptor activity	11	14.86%
ubiquitin-protein ligase activity	3	4.05%
signal transducer activity	4	5.40%
ligase activity	3	4.05%
growth factor activity	2	2.70%
structural molecule activity	2	2.70%
actin binding	2	2.70%
oxidoreductase activity	3	4.05%
GTP binding	2	2.70%
transcription factor activity	5	6.75%
ATP binding	3	4.05%
DNA binding	2	2.70%
nucleotide binding	5	6.75%
RNA binding	2	2.70%
calcium ion binding	3	4.05%
transferase activity	3	4.05%
metal ion binding	6	8.10%
zinc ion binding	5	6.75%

(c)

Cellular component	#genes	Percentage
integral to plasma membrane	16	21.62%
membrane	26	35.13%
plasma membrane	8	10.81%
integral to membrane	15	20.27%
extracellular space	4	5.40%
extracellular matrix (sensu Metazoa)	2	2.70%
endoplasmic reticulum	3	4.05%
Golgi stack	2	2.70%
cytoplasm	5	6.75%
nucleus	11	14.86%
mitochondrion	2	2.70%

at different coincidence levels represented by the PI (12) between classes and substructures [Fig. 16(c) and Table III, $PI < 3.1 \times 10^{-4}$]. Substructure #89 describes class #13 based on a cell communication biological process located at the integral to

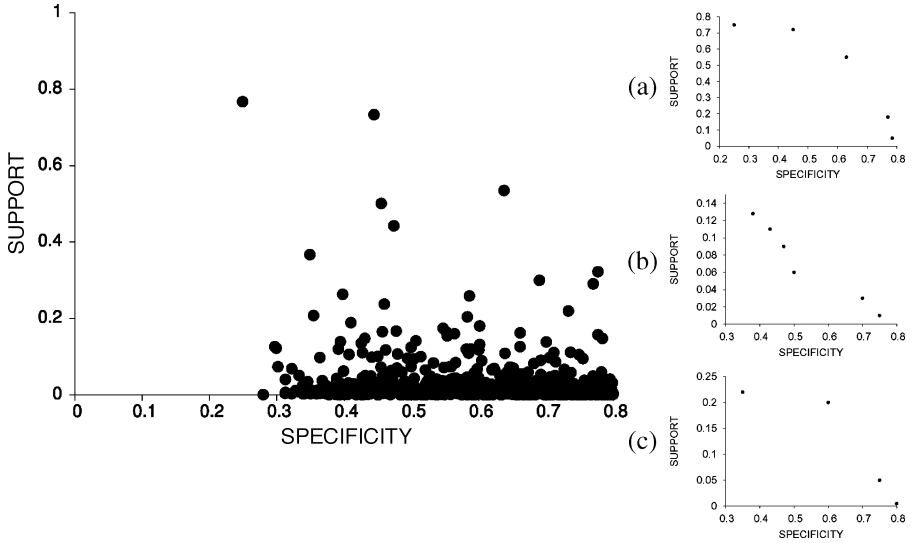

Fig. 14. The nondominated solutions obtained by EMO-CC. (a)–(c) Each subplot shows nondominated solutions in different neighborhoods, which do not compete with each other, as a consequence of the multimodal policy followed by EMO-CC.

the plasma membrane or, in a more general case, at the integral to membrane cellular component (Table IV). A slightly different description is provided by substructure #256, which includes a cellular physiological process (Table IV). A different example is given by substructure #179, which describes an apoptosis process (i.e., a form of programed cell death) located at the integral to plasma membrane (Table IV). Significantly, these descriptions are based on different types of features (e.g., biological process and cellular components) that belong to different levels of the GO hierarchy (e.g., level 6 or level 4). These diverse substructures are optimal in the sense that they belong to the Pareto optimal set composed of specific and sensitive descriptions (Fig. 11).

We compare the performance of EMO-CC for extracting biologically valid substructures with APRIORI and SUBDUE. We have already seen that EMO-CC subsumes those solutions obtained by the other methods and provides novel and diverse optimal solutions (i.e., belonging to the Pareto optimal set) by the evaluation of several quantitative metrics. A qualitative evaluation of these methods reveals that EMO-CC obtains more specific substructures than the other methods for those substructures discovered in common. Moreover, the matching among substructures retrieved by EMO-CC and the independently obtained classes derived from the expression profiles is better than the one achieved by the other methods. For example, substructure #5 identified by APRIORI matches with class #15 with a PI of 2.1738×10^{-4}, while the corresponding PI for substructure #811 retrieved by EMO-CC is 6.9854×10^{-6} (Fig. 16).

We compress those substructures that explain the same expression profile to provide a summarized description of this phenomenon. The 24 expression profiles can be explained by 45 substructures of GO terms (Appendix Table VII, available at http://www.ieeexplore.ieee.org). For example, substructures #89 and #256, which explain class #13 are compressed because they are indistinguishable for this class (Table IV). However, substructure #179 describes it from a very different point-of-view and it is preserved as a diverse solution. This compression is dynamic because substructures are regrouped in a context-de-

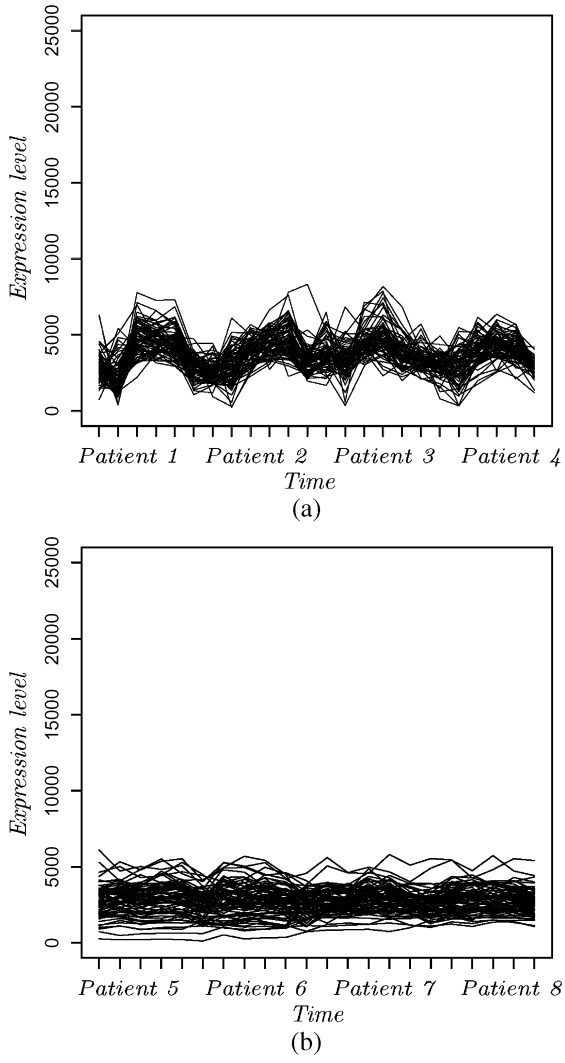

Fig. 15. Class #13 differential expression profile encodes genes with different behavior between treatment and control, with a similar pattern among patients. (a) Gene expression corresponding to treated patients. (b) Gene expression from patients belonging to the control group.

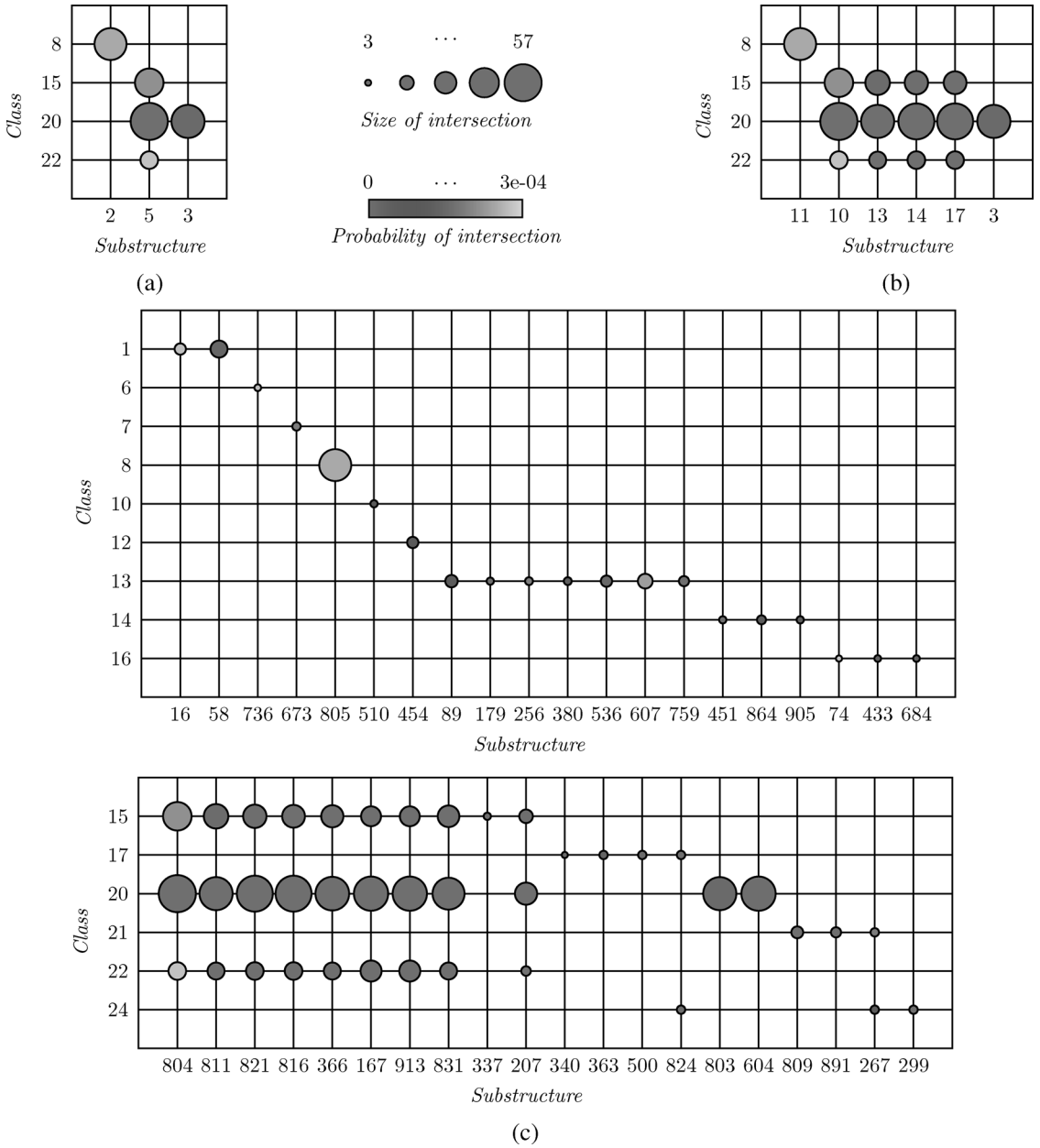

Fig. 16. Description of gene expression profiles that are explained by GO substructures. Each intersection is represented by a circle, where the size corresponds to the number of elements in common between a class and a substructure, and the color illustrates the probability of intersection (green: low, red: high). (a) APRIORI. (b) SUBDUE. (c) A subset of all EMO-CC intersections. The complete graph is shown in the Appendix (Fig. 21, available at http://www.ieeexplore.ieee.org).

pendent fashion, where the context corresponds to an explained class, and a different classification can produce a distinct substructure association (e.g., substructures #89 and #256 are indistinguishable for class #13, while it may not be the case for other classes of microarray or clinical experiments). An emergent property of current explanations provided by the substructures retrieved by EMO-CC consists of their usefulness for differentiating even subtle expression patterns (Fig. 17). Notably, this classification is performed based on external information provided by the GO database, instead of the levels of expression.

In addition to previous methods, we also compared the performance of EMO-CC with two other state-of-the-art methods typically used for GO analysis. To do so, we investigate FatiGO and OE by running them three times to extract the GO terms associated with genes expressed in the inflammatory response problem. This happens because both methods need separate runs for each type of feature [i.e., biological process (BP), molecular function (MF), cellular component (CC)]. Indeed, FatiGO needs the specification of a predefined level of the GO hierarchy, thus, we used the default level 3. Both methods organize their results by a ranked percentage of matching between a query, which in this case corresponds to the class #13 gene expression profile, and the retrieved GO terms (Tables V and VI). To standardize the PI between GO terms and the query set, we recalculate it for all methods by using the relevant GO occurrences in the 1770 genes with significant expression for the inflammatory problem.

The best ranked result provided by FatiGO for the BP ontology is the term GO:0007154 cell communication, which

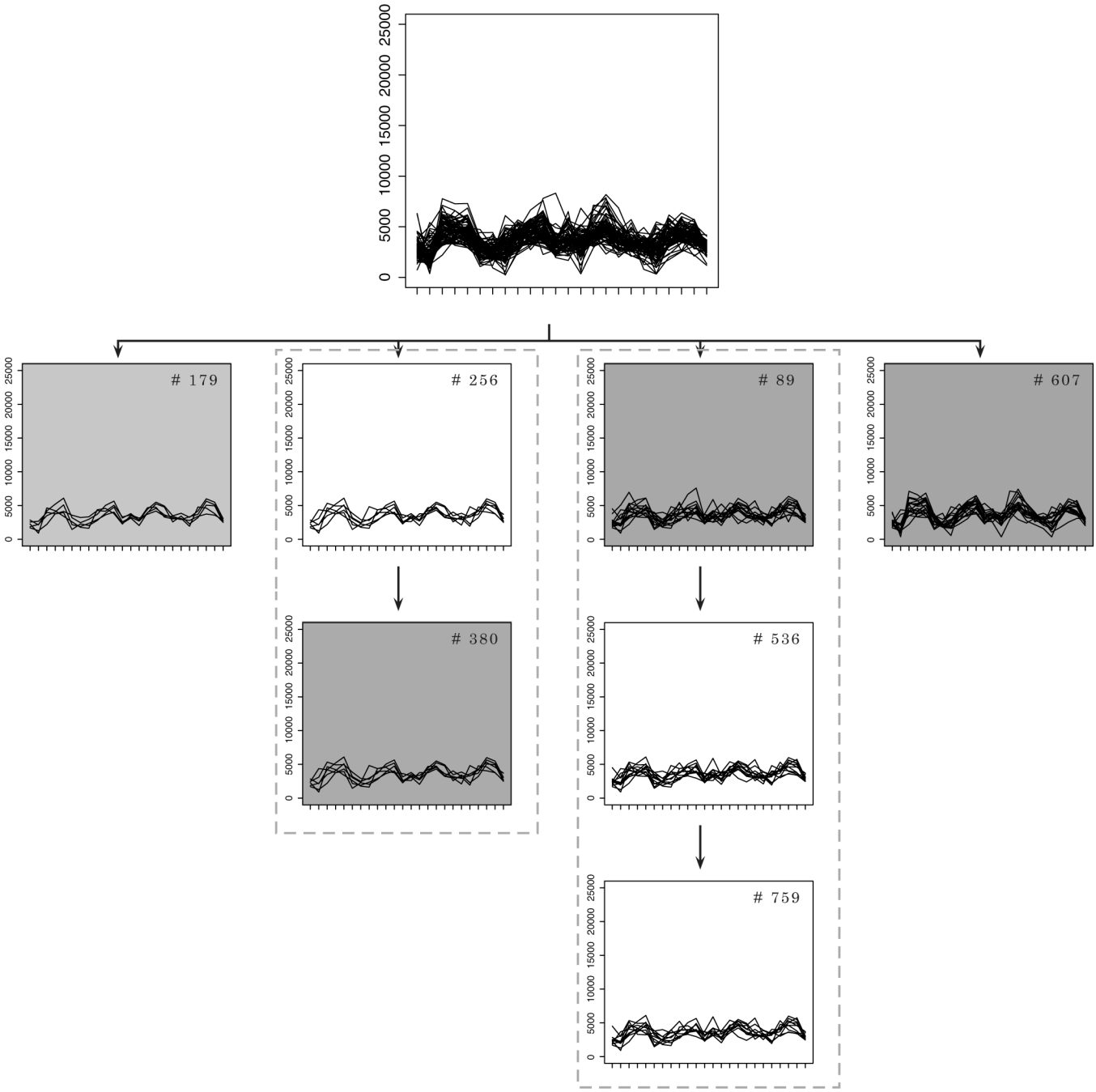

Fig. 17. Compressed substructures that explain class #13 expression profile (Fig. 8, row: 3 column: 3). Class #13 is explained by seven substructures (color-coded subgraphs show compression of substructures from Table IV). These substructures are arranged by parental order in the GO database and compressed, dissecting similar expression patterns based on independent information provided by GO.

achieves a PI of 6.9×10^{-4} (Table V). EMO-CC identifies two substructures containing this term (Table IV): substructure #89, containing terms GO:0007154 cell communication and GO:0016021 integral to membrane; and substructure #256, containing terms GO:0007154 cell communication, GO:0050875 cellular physiological process and GO:0016021 integral to membrane. Both substructures explain class #13 but with more accurate PIs: 5.79×10^{-5} and 1.91×10^{-4}, respectively. Indeed, the second ranked solution for the BP ontology obtained by FatiGO is the term GO:0050789 regulation of biological process with a PI of 0.46. EMO-CC uncovers two substructures including this term: substructure #469, containing terms GO:0050789 regulation of BP and GO:0007154 cell communication; and substructure #34, containing terms GO:0050789 regulation of biological process and GO:00016020 membrane. Both of them with a $PI \leq 0.005$, suggesting better explanations than the ones provided by FatiGO. The analysis of the ranked terms in MF and CC ontologies presents similar results.

The most relevant solutions obtained by OE for the BP ontology is the term GO:0006955 immune response, which achieves a PI of 3.4×10^{-3} (Table VI). EMO-CC achieves similar results with substructure #250. Indeed, EMO-CC identifies

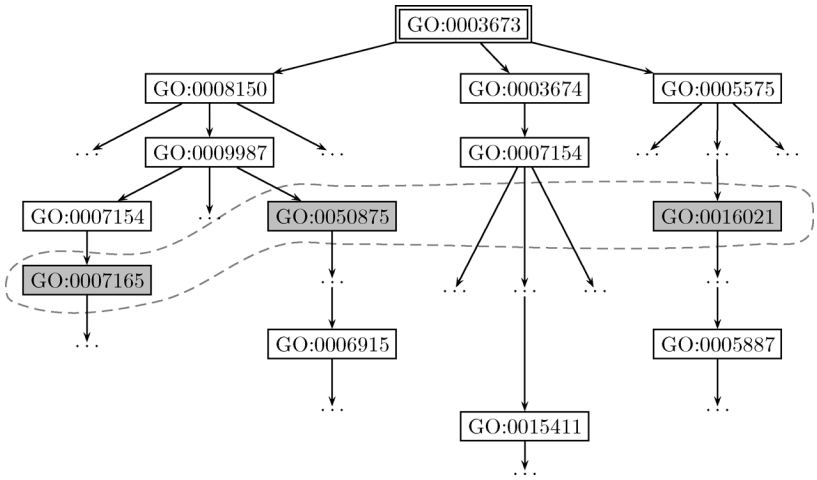

Fig. 18. Example of a novel annotation uncovered by EMO-CC (dashed lines) based on substructure #380. The tree represents the GO hierarchy with the three subontologies as the main branches (GO:0008150 "biological process," GO:0003674 "molecular function," and GO:0005575 "cellular component"). This annotation include GO terms from different subontologies and defined at different levels of specificity.

three other nondominated substructures that include several GO terms: substructure #536, containing terms GO:0007165 signal transduction and GO:0016021 integral to membrane; substructure #759, containing terms GO:0007165 signal transduction and GO:0005887 integral to plasma membrane; and substructure #380, containing terms GO:0007165 signal transduction, GO:0050875 cellular physiological process and GO:0016021 integral to membrane. All of these substructures are better explanations than the single term recovered by OE achieving a PI of 1.5210×10^{-5}, 1.43×10^{-6}, and 5.43×10^{-5}, respectively. Curiously, OE retrieves the term GO:0006954 inflammatory response, which is pertinent for this study, with a PI of 0.62 for substructure #13. In addition to this weak explanation, EMO-CC finds that this term in conjunction with term GO:0005622 intracellular accurately explain class #17 with a PI of 1.21×10^{-5}. Again, we obtain similar results for the other ontologies considered in this study.

Summarizing, EMO-CC discovers substructures composed of different subontologies defined at distinct levels of specificity (Fig. 14 and Appendix Table VII, available at http://www.ieeexplore.ieee.org). Neither FatiGO nor OE are able to provide a comprehensive strategy to encode this information in their solutions. Instead, they report an exhaustive lists of all individual terms at each of the levels queried by the user. Moreover, these methods supply individual results that always conceal relevant relationships between them. For example, OE reports two solutions that explain substructure #13 consisting of term GO:0007242 intracellular signaling cascade and term GO:0007165 signal transduction, but missed the parental relationship between them (Table VI). In contrast, EMO-CC provides optimal and diverse substructures within an appropriate inclusive order that can explain independent experiments (Fig. 17).

The substructures identified by EMO-CC can be considered new annotations (Fig. 18 and Table IV). These annotations include different types of features defined at distinct hierarchically organized levels of specificity, which can be used to uncover new members to the underlying substructures based on the similarity with the corresponding GO terms. Consequently, this guideline can be used to indirectly classify new members of an expression class, as we will see in the next section.

Finally, we validate the GO substructures obtained by EMO-CC using a high-quality hand-curated database termed Ingenuity Pathways Knowledge Base [61], which is, at the moment, a gold-standard for metabolic pathways. We queried this database with the web-based entry tool developed by Ingenuity Pathways Analysis (IPA) [61]. For example, by using the list of genes from class #13, the best description identified by IPA (score 45, focus genes 21) functionally corresponds to an inflammatory network *Inflammatory Disease* (Appendix Table VIII and Appendix Fig. 22, available at http://www.ieeexplore.ieee.org). Moreover, *Inflammatory Disease* is the prevalent function of this network with p-values between $1.15 \times 10^{-5} - 8.83 \times 10^{3}$ (Appendix Table IX, available at http://www.ieeexplore.ieee.org), suggesting that class #13 and the EMO-CC substructures that explain it constitute a meaningful biological association.

C. Unsupervised Classifier Inference Process (STEP 5)

The EMO-CC methodology classifies new instances by their similarity with one or more substructures using a k-nearest neighbor unsupervised classifier. We evaluate the performance of the proposed inference process by the following procedure: 1) we perform a holdout of our original dataset in two subsets: *training data* and *test data*, with 80% and 20% of the original dataset, respectively, selected randomly without reposition [62] and apply STEPS 2 and 3 to the training data; 2) for each gene in the test set we use its GO annotation to calculate its membership to the set of substructures identified in 1) using (13) and select the substructure with the highest membership value as the best prediction; and 3) we test the accuracy of the inference process by: (3.1) identifying the expression class explained by the selected substructure; (3.2) calculating its centroid as a weighted average of the expression values of its members [34]; and (3.3) computing the Pearson correlation (PC) [63] between the expression of the predicted gene and the centroid of (3.2).

We illustrate this process by: 1) evaluating the gene `203107_x_at` from the test set [Fig. 19(a)]; 2) calculating its

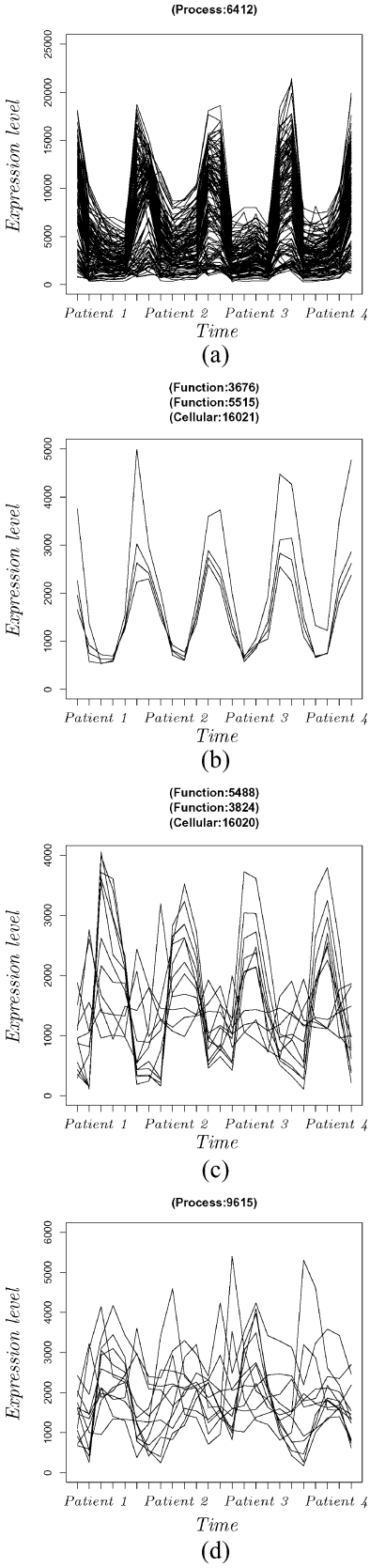

Fig. 19. The EMO-CC inference process. The new observation classified by EMO-CC is color-coded in red within the inferred substructure, while the centroid of the substructure is color-coded in blue. Expression of the substructures that classify (a) gene `203107_x_at`, (b) gene `208982_at`, (c) gene `216316_x_at`, and (d) gene `211676_x_at`.

membership to the set of the previously identified substructures. Since the obtained substructures are not disjointed, a given observation may belong to more than one substructure [e.g., probe set `203107_x_at` has a membership degree greater than zero in substructure #2 (0.24), #8 (0.25), #16 (0.63), #28 (0.68), #33 (0.70), #34 (0.76), and #127 (0.91)]. Therefore, we select the maximum value among the different memberships, classifying the target probe set into substructure #127. Then, we test the accuracy of the predictions by (3.2) calculating the centroid corresponding to substructure #127 [Fig. 19(a)], which is a cohesive profile with very similar expression pattern of its members. Afterwards, in (3.3), we calculate the correlation between the gene `203107_x_at` and the former centroid ($PC > 0.6$) and evaluate the prediction as a positive matching. Similar results are observed with other genes in the test set [Fig. 19(b)–(d)].

We evaluate the complete test set by considering substructures with at least n GO terms, where n ranges between 1 and 4. Our results indicate that 70% of the successful predictions can be achieved by using four GO terms [Fig. 20(a)], showing that the performance increases as the number of GO terms increases. However, this monotonic process is not conserved when the specificity of a given substructure is improved. For example, by increasing the specificity values of the former substructures from 0.5 to 0.9, we cannot observe an improvement in the prediction performance [Fig. 20(b)]. These results suggest that approaches that widely explore GO database in the complete feature space (i.e. all GO terms from biological process, molecular function and cellular component) can be appropriate for describing and predicting gene expression patterns.

The proposed testing process indicates a strategy to predict gene expression patterns based on an independent source of data such as GO terms. However, several classification errors result from ambiguous annotation terms or too general categories, as well as missing information in the GO database rather than misclassifications [64]. Many of these problems will be solved when the GO database becomes more accurately curated.

VI. Conclusion

Unlike typical clustering techniques, conceptual clustering methods have been successfully applied to structural information in order to reveal hidden concepts by searching through a predefined space of potential hypothesis. However, the formulation of the search problem in a structural database would often result in a conflicting paradigm. On the one hand, generating a large number of substructures, each containing a very small number of instances that share many features, makes it hard to find commonalities among similar observations. On the other hand, generating a small number of substructures, where their members share a limited number of features, would fail to discriminate between similar members. Therefore, any successful methodology should also consider more adequate tradeoffs among different criteria to evaluate substructures that uncover meaningful concepts always hidden in large datasets.

A. EMO-CC Methodology

Several characteristics distinguish EMO-CC from other conceptual clustering methods.

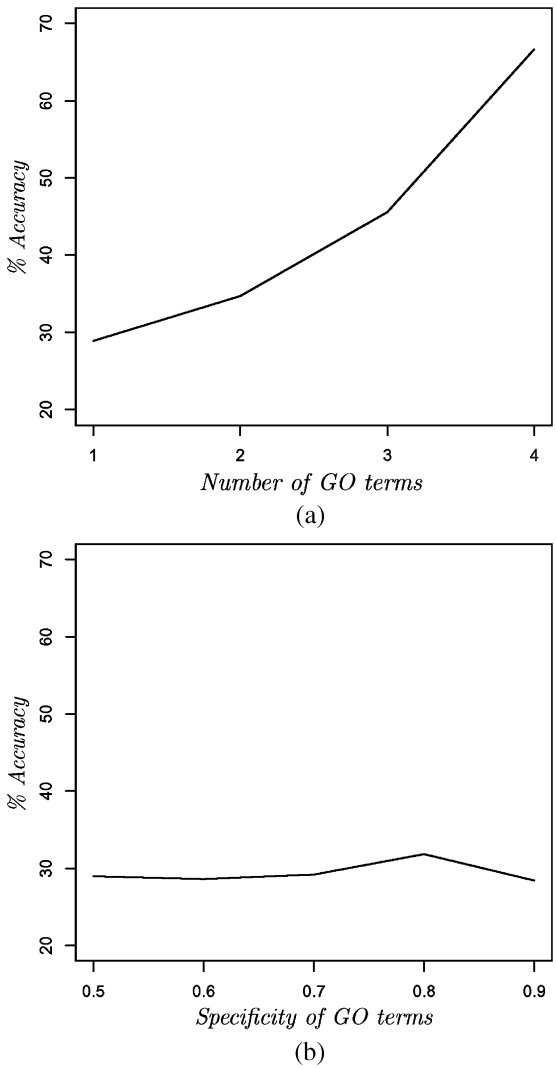

Fig. 20. Performance of the EMO-CC inference process evaluated by considering substructures with different number of terms defined at distinct specificity levels. (a) Accuracy of the inference process evaluating the test set, where substructures contain 1–4 GO terms. (b) Accuracy of the EMO-CC inference process evaluating the test set, where substructures contain only one term with specificity levels from 0.5 to 0.9.

- EMO-CC searches for all optimal solutions among multiple criteria (i.e., Pareto optimality) [17], which avoids the biases that might result from using any specific weighting scheme [10]. This allows the detection of cohesive substructures even those comprising a small number of instances that would remain undetected by methods that emphasize the support of a substructure [23]. We showed that the EMO-CC algorithm, by using a multiobjective approach, obtains more diverse Pareto optimal sets and dominates most of the substructures provided by the other methods.
- EMO-CC has a multimodal nature that allows alternative descriptions of a system by providing several adequate solutions [7], [17], thus recovering locally optimal targets that could be meaningful [15], [65]. This differentiates EMO-CC from methods that are focused on a single optimum [24].
- EMO-CC performs a local feature selection for each substructure, because not every feature is relevant for every substructure [28], and *a priori*, we do not know which features are meaningful for a given set of instances. This is in contrast to approaches that filter or reduce features for all possible clusters [29].
- EMO-CC allows gene membership to more than one substructure by using a flexible classifier [34], [57], [66], thus explicitly treating the substructures as hypotheses that can be tested and refined [6]. This distinguishes EMO-CC from other approaches that prematurely force instances into disjoint clusters [67].

Finally, EMO-CC is applicable to a wide set of domains, being easy to customize to particular problem, and may be an appropriate technique to uncover rare and unknown patterns in structural databases. Particularly, this guideline can be easily extended to more complex networks comprising protein–protein or different regulatory interactions [1], [2]. Indeed, EMO-CC efficiently searches the feature space in acceptable run times, while computational times of exhaustive search algorithms are intractable.

B. GO and Structural Database Domains

Again, several characteristics distinguish EMO-CC from other approaches typically used in GO databases.

- EMO-CC uses a multivariate and multilevel approach, where substructures are discovered based on several types of hierarchical features. For example, substructures identified in the GO database include features or terms derived from different information sources (e.g., cellular components or biological processes). Moreover, each feature is defined at a different specificity level in a graph-based structure. This guideline distinguishes our methodology from other approaches, like FatiGO [24] and OE [25], where each type of feature is individually treated, and the specificity level is selected *a priori*.
- EMO-CC considers gene expression as one independent feature, thereby allowing classification of genes even in the absence of its expression. This approach differs from supervised learning methods that group features and instances based on a explicitly defined dependent class. Instead, EMO-CC uses an unsupervised strategy that compresses similar groups of substructures based in their ability to describe independent classes derived from different experimental conditions (e.g., microarray expression, or chip-on-chip binding occupancy). This approach changes according to the experimental class, thus differing from fixed approaches that use an irreversible database compression [22]. Sometimes, expression-dependent characterizations only allows a relatively crude classification of genes into a limited number of classes, which can conceal rather than reveal novel interesting profiles [15].
- EMO-CC uses a simple classification procedure that allows to predict gene expression patterns based on an independent set of features like GO terms. Although the prediction accuracy increases when substructures with more features (i.e., GO terms) are considered, it does not change when the complexity of these features is improved. Curiously, the specificity of the ontologies is the center of the current debates about their applicability [68]. Therefore, approaches

like EMO-CC, that emphasize wide searches in the feature space can be appropriate for describing and predicting gene expression patterns. However, methods that focus on one type of feature cannot improve their results even if they add more levels to the analysis.

C. Biological Domain

We investigated the inflammatory response problem by means of a study performed on human volunteers treated with intravenous endotoxin compared to a placebo. Understanding this problem is critical because the majority of deaths are caused due to inflammatory diseases [21]. The response to inflammation is a complex problem that considers gene profiles that reflect differences in gene expression over time, treatment, and patient. Most state-of-the-art methods can normally recover the most obvious relations, but fail to discover the less frequent but more informative underlying data associations.

We showed that EMO-CC produces annotations that explain coexpressed genes and can be used to make predictions by using an independent source of information. We demonstrated that these annotations are biologically meaningful because they represent inflammatory regulatory networks.

ACKNOWLEDGMENT

The authors thank C. del Val, C. Previti, H. Huang, and A. Polpitiya for useful suggestions.

REFERENCES

[1] V. Siripurapu, J. Meth, N. Kobayashi, and M. Hamaguchi, "Dbc2 significantly influences cell-cycle, apoptosis, cytoskeleton and membrane-trafficking pathways," *J. Molecular Biol.*, vol. 346, no. 1, pp. 83–89, 2005.

[2] A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo, "Pathway studio—The analysis and navigation of molecular networks," *Bioinformatics*, vol. 19, no. 16, pp. 2155–2157, 2003.

[3] T. G. O. Consortium, "Gene ontology: Tool for the unification of biology," *Nature Genet.*, vol. 25, pp. 25–29, 2000.

[4] G. Schuler, J. Epstein, H. Ohkawa, and J. Kans, "Entrez: Molecular biology database and retrieval system," *Methods Enzymol.*, vol. 266, pp. 141–162, 1996.

[5] D. Cook, L. Holder, S. Su, R. Maglothin, and I. Jonyer, "Structural mining of molecular biology data," *IEEE Eng. Med. Biol., Special Issue on Advances in Genomics*, vol. 4, no. 20, pp. 67–74, 2001.

[6] T. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[7] E. Ruspini and I. Zwir, "Automated generation of qualitative representations of complex object by hybrid soft-computing methods," in *Pattern Recognition: From Classical to Modern Approaches*, S. Pal and A. Pal, Eds. Singapore: World Scientific Company, 2001, pp. 453–474.

[8] I. Zwir, R. Romero-Zaliz, and E. Ruspini, "Automated biological sequence description by genetic multiobjective generalized clustering," in *Techniques in Bioinformatics and Medical Informatics*, F. Valafar, Ed., 2002, vol. 980, pp. 65–82.

[9] R. Romero-Zaliz, I. Zwir, and E. Ruspini, "Generalized Analysis of Promoters (GAP): A method for DNA sequence description," in *Applications of Multi-Objective Evolutionary Algorithms*. Singapore: World Scientific, 2004, pp. 427–450.

[10] J. Rissanen, *Stochastic Complexity in Statistical Inquiry Theory*. Singapore: World Scientific, 1989.

[11] E. Ruspini and I. Zwir, "Automated qualitative description of measurements," in *Proc. 16th IEEE Instrum. Measure. Technol. Conf.*, Venice, Italy, 1999, vol. 2, pp. 1086–1091.

[12] I. Zwir, H. Huang, and E. Groisman, "Analysis of differentially-regulated genes within a regulatory network by GPS genome navigation," *Bioinformatics*, vol. 21, no. 22, pp. 4073–4083, Nov. 2005.

[13] L. McCue, W. Thompson, C. Carmack, M. P. Ryan, J. S. Liu, V. Derbyshire, and C. E. Lawrence, "Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes," *Nucleic Acids Res*, vol. 29, pp. 774–782, 2001.

[14] A. Martinez-Antonio and J. Collado-Vides, "Identifying global regulators in transcriptional regulatory networks in bacteria," *Curr. Opin. Microbiol.*, vol. 6, pp. 482–489, 2003.

[15] I. Zwir, D. Shin, D. A. Kato, K. Nishino, T. Kunihiko, F. Solomon, J. Hare, H. Huang, and E. Groisman, "Dissecting the PhoP regulatory network of Escherichia coli and Salmonella enterica," *PNAS*, vol. 102, no. 8, pp. 2862–2867, 2005.

[16] , T. Back, D. Fogel, and Z. Michalewicz, Eds., *Handbook of Evolutionary Computation*. Bristol, U.K.: IOP Publishing, 1997.

[17] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. New York: Wiley, 2001.

[18] C. Coello-Coello, D. V. Veldhuizen, and G. Lamont, *Evolutionary Algorithms for Solving Multi-Objective Problems*, ser. Genetic Algorithms and Evolutionary Computation. Norwell, MA: Kluwer, 2002.

[19] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, pp. 182–197, 2002.

[20] J. Dróo, A. Pétrowski, E. Taillard, and A. Chatterjee, *Metaheuristics for Hard Optimization: Methods and Case Studies*. New York: Springer, 2005.

[21] S. Calvano, W. Xiao, D. Richards, R. Felciano, H. Baker, R. Cho, R. Chen, B. Brownstein, J. Cobb, T. S.K., C. Miller-Graziano, L. Moldawer, M. Mindrinos, R. Davis, R. Tompkins, and S. Lowry, L. S. C. R. ProgramInflamm, and H. R. to Injury, "A network-based analysis of systemic inflammation in humans," *Nature*, vol. 437, no. 7061, pp. 1032–1037, 2005.

[22] I. Jonyer, D. J. Cook, and L. B. Holder, "Discovery and evaluation of graph-based hierarchical conceptual clusters," *J. Mach. Learn. Res.*, vol. 2, pp. 19–43, 2001.

[23] R. Agrawal and J. Shafer, "Parallel mining of association rules," *IEEE Trans. Knowl. Data Eng.*, vol. 8, pp. 962–969, 1996.

[24] F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo, "Fatigo: A web tool for finding significant associations of gene ontology terms with groups of genes," *Bioinformatics*, vol. 20, pp. 578–580, 2004.

[25] P. Khatri, P. Bhavsar, G. Bawa, and S. Draghici, "Onto-tools: An ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments," *Nucleic Acids Research*, vol. 32, pp. 449–456, 2004.

[26] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2 ed. New York: Wiley, 2000.

[27] G. Der and B. Everitt, *A Handbook of Statistical Analyses Using SAS*. London, U.K.: Chapman-Hall, 1996.

[28] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Dordrecht: Kluwer, 1988.

[29] K. Yeung and W. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, pp. 763–774, 2001.

[30] P. Cheeseman and R. W. Oldfors, *Selecting Models From Data*. Berlin, Germany: Springer-Verlag, 1994.

[31] A. Aho, J. Hopcroft, E. John, and J. Ullman, *Data Structures and Algorithms*, ser. Addison-Wesley Series in Computer Science and Information Processing. Reading, MA: Addison-Wesley, 1983.

[32] D. M. Chickering, "Optimal structure identification with greedy search," *J. Mach. Learn. Res.*, vol. 3, pp. 507–554, 2003.

[33] P. Larranaga, M. Poza, Y. Yurramendi, R. H. Murga, and C. M. H. Kuijpers, "Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters," *IEEE J. Pattern Anal. Mach. Intell.*, vol. 18, pp. 912–926, 1996.

[34] J. Bezdek, "Fuzzy clustering," in *Handbook of Fuzzy Computation*, E. Ruspini, P. Bonissone, and W. Pedrycz, Eds. College Park, MD: Institute of Physics Press, 1998, pp. f6.1:1–f6.6:19.

[35] G. Grothaus, A. Mufti, and T. Murali, "Automatic layout and visualization of biclusters," *Algorithms Molecular Biology*, vol. 1, no. 15, 2006.

[36] L. Hall, I. Ozyurt, and J. Bezdek, "Clustering with a genetically optimized approach," *IEEE Trans. Evol. Comput.*, vol. 3, no. 2, pp. 103–112, 1999.

[37] S. Pan and K. Cheng, "Evolution-based tabu search approach to automatic clustering," *IEEE Trans. Syst., Man, Cybern. Part C: Applications and Reviews*, vol. 37, no. 5, pp. 827–838, 2007.

[38] J. Handl and J. Knowles, "An evolutionary approach to multiobjective clustering," *IEEE Trans. Evol. Comput.*, vol. 11, no. 1, pp. 56–76, Feb. 2007.

[39] M. Morita, R. Sabourin, F. Bortolozzi, and C. Y. Suen, "Unsupervised feature selection using multi-objective genetic algorithms for handwritten word recognition," *ICDAR*, vol. 02, p. 666, 2003.

[40] P. Delima and G. Yen, "Multiple objective evolutionary algorithm for temporal linguistic rule extraction," *ISA Trans.*, vol. 44, no. 2, pp. 315–327, 2005.

[41] B. Alatas, E. Akin, and A. Karci, "Modenar: Multi-objective differential evolution algorithm for mining numeric association rules," *Appl. Soft Comput. J.*, vol. 8, no. 1, pp. 646–656, 2008.
[42] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006.
[43] S. Mitra and H. Banka, "Multi-objective evolutionary biclustering of gene expression data," *Pattern Recogn.*, vol. 39, no. 12, pp. 2464–2477, 2006.
[44] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach," *IEEE Trans. Evol. Comput.*, vol. 3, no. 4, pp. 257–271, Nov. 1999.
[45] E. Zitzler, K. Deb, and L. Thiele, "Comparison of multiobjective evolutionary algorithms: Empirical results," *Evol. Comput.*, vol. 8, no. 2, pp. 173–195, 2000.
[46] K. Deb and A. R. Reddy, "Reliable classification of two-class cancer data using evolutionary algorithms," *BioSystems*, vol. 72, no. 1–2, pp. 111–129, Nov. 2003.
[47] J. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: MIT Press, 1992.
[48] M. Oltean and C. Grosan, "Using traceless genetic programming for solving multi-objective optimization problems," *J. Experimental and Theoretical Artif. Intell.*, vol. 19, no. 3, pp. 227–248, 2007.
[49] Y. Zhang and P. Rockett, "Feature extraction using multi-objective genetic programming," *Studies in Computational Intelligence*, vol. 16, pp. 75–99, 2006.
[50] K. Rodriguez-Vazquez, C. Fonseca, and P. Fleming, "Multiobjective genetic programming: A nonlinear system identification application," in *Proc. Late Breaking Papers at the 1997 Genetic Programming Conf.*, J. Koza, Ed., Stanford, CA, Jul. 13–16, 1997, pp. 207–212, Stanford Univ..
[51] O. Cordón, E. Herrera-Viedma, and M. Luque, "Evolutionary learning of Boolean queries by multiobjective genetic programming," in *Proc. 7th International Conference on Parallel Problem Solving from Nature (PPSN-VII)*, Granada, Spain, 2002, vol. 2439, Lecture Notes in Computer Science, pp. 710–719.
[52] O. Cordón, E. Herrera-Viedma, and M. Luque, "Improving the learning of Boolean queries by means of a multiobjective IQBE evolutionary algorithm," *Inf. Process. Manage.*, vol. 42, no. 3, pp. 615–632, 2006.
[53] M. Wong and K. Leung, *Data Mining Using Grammar-Based Genetic Programming and Applications*. Norwell, MA: Kluwer, 2000.
[54] J. Horn and N. Nafpliotis, Multiobjective optimization using the niched Pareto genetic algorithm Urbana, IL, Tech. Rep. IlliGAl Rep. 93005, 1993.
[55] P. Jaccard, "The distribution of flora in the alpine zone," *The New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
[56] S. Tavazoie, J. Hughes, M. Campbell, R. Cho, and G. Church, "Systematic determination of genetic network architecture," *Nature Genet.*, vol. 22, no. 3, pp. 281–285, 1999.
[57] A. Gasch and M. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biology*, vol. 3, p. RESEARCH0059, 2002.
[58] C. Rubio-Escudero, R. Romero-Zaliz, O. Cordón, O. Harari, C. del Val, and I. Zwir, "Optimal selection of microarray analysis methods using a conceptual clustering algorithm," in *Proc. Applications of Evolutionary Computing: EvoWorkshops 2006: EvoBIO, EvoCOMNET, EvoHOT, EvoIASP, EvoINTERACTION, EvoMUSART, and EvoSTOC*, F. Rothlauf *et al.*, Ed., Apr. 10–12, 2006, EvoBIO Contributions, pp. 172–183.
[59] Affymetrix, Microarray Platform Comparisons Affymetrix White Paper, 2006.
[60] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. 1993 ACM SIGMOD Int. Conf. Manage. Data*, P. Buneman and S. Jajodia, Eds., Washington, D.C., 1993, pp. 207–216. [Online]. Available: citeseer.ist.psu.edu/agrawal93mining.html
[61] I. Systems, Ingenuity Pathways Analysis. [Online]. Available: http://www.ingenuity.com
[62] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. Cambridge, MA: MIT Press, 2001, hAN d3 01:1 1.Ex.
[63] K. E. Applegate and P. E. Crewson, "An introduction to biostatistics," *Radiology*, pp. 318–322, 2002.
[64] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir, "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data," *Genetics*, vol. 101, no. 9, pp. 2981–2986, 2004.
[65] T. Azevedo, R. Lohaus, V. Braun, M. Gumbel, M. Umamaheshwar, P. Agapow, W. Houthoofd, U. Platzer, G. Borgonie, H. Meinzer, and A. Leroi, "The simplicity of metazoan cell lineages," *Nature*, vol. 433, pp. 152–156, 2005.
[66] O. Cordón, F. Herrera, F. Hoffmann, and L. Magdalena, *Genetic Fuzzy Systems. Evolutionary Tuning and Learning of Fuzzy Knowledge Bases*, ser. Advances in Fuzzy Systems—Applications and Theory. Singapore: World Scientific, 2001, vol. 19.
[67] Z. Qin, L. McCue, W. Thompson, L. Mayerhofer, C. Lawrence, and J. Liu, "Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites," *Nature Biotechnol.*, vol. 21, pp. 435–439, 2003.
[68] J. Bard and S. Rhee, "Ontologies in biology: Design, applications and future challenges," *Nature Reviews Genetics*, vol. 5, pp. 213–222, 2004.

Rocío C. Romero-Zaliz was born in 1977 in Buenos Aires, Argentina. She received the M.Sc. degree in computer science from the University of Buenos Aires, Buenos Aires, Argentina, in 2001 and the Ph.D. degree on computer science from the University of Granada, Granada, Spain, in 2005.

She is currently an Associate Professor with the Department of Computer Science and Artificial Intelligence, University of Granada, where she is a member of M4M Bioinformatics Lab, part of the Soft Computing and Intelligent Information Systems Research Group. She has worked in several research projects supported by the Argentinian and Spanish Government. Her research interests include bioinformatic tools, machine learning, and evolutionary algorithms.

Cristina Rubio-Escudero received the B.Sc. and M.Sc. graduate degrees in computer science from the University of Granada, Granada, Spain, in 2003 and 2005, respectively.

She is an Assistant Teacher with the Department of Computing Systems and Languages, University of Sevilla, in the Soft Computing and Intelligent Information Systems Research Group. She has worked in several research projects supported by the Spanish Government and the European Union. She is coauthor of journal papers, book chapters, and conference papers. Her research interests include gene expression analysis, knowledge discovery, and systems biology.

J. Perren Cobb graduated (*cum laude*) with a Degree in medicine from the University of Louisville, School of Medicine, Louisville, KY, in 1986.

From 1986 to 1994, he completed his residency in General Surgery at the University of California, San Francisco. From 1989 to 1992, he was a Fellow in the Critical Care Medicine Department, NIH, and from 1994 to 1995, he completed a fellowship in the Multidisciplinary (Surgical) Critical Care Training Program at the University of Pittsburgh. He is currently Professor of Surgery and Associate Professor of Genetics at the Washington University School of Medicine, St. Louis, MO. He specializes in surgical critical care with research interests in the pathophysiology of sepsis and injury. He is also the Director of the Cellular Injury and Adaptation Laboratory and the University's Center for Critical Illness and Health Engineering. His investigative work has been supported by the National Institutes of Health, the American Association for the Surgery of Trauma, the Society of Critical Care Medicine, and the Barnes-Jewish Hospital Foundation.

Dr. Cobb serves on the Steering Committee of the NIGMS Inflammation and the Host Response to Injury Program ("Trauma Glue Grant"). An active member of the Society of Critical Care Medicine (SCMM) and the American Thoracic Society. He was the Co-Chair of the SCCM Surgical Section Research Committee, as well as a member of the SCCM Membership Committee. He is a Fellow in the American College of Surgeons and Past-President of the Association for Academic Surgery.

Francisco Herrera received the M.Sc. and Ph.D. degrees in mathematics from the University of Granada, Granada, Spain, in 1988 and 1991, respectively.

He is currently a Professor in the Department of Computer Science and Artificial Intelligence, University of Granada. He has published more than 100 papers in international journals. He is coauthor of the book *Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases* (World Scientific, 2001). As edited activities, he has coedited four international books and coedited 165 special issues in international journals on different soft computing topics. He currently serves as Area Editor for the *Journal on Soft Computing* (area of genetic algorithms and genetic fuzzy systems), and he serves as member of the Editorial Board of the journals *Fuzzy Sets and Systems*, the *International Journal of Hybrid Intelligent Systems*, the *International Journal of Computational Intelligence Research*, the *Mediterranean Journal of Artificial Intelligence*, the *International Journal of Information Technology and Intelligent and Computing*, *Evolutionary Intelligence*, and *Memetic Computation*. He acts as Associated Editor for the journals *Mathware and Soft Computing*, *Advances in Fuzzy Systems*, and *Advances in Computational Sciences and Technology*. His current research interests include computing with words and decision making, data mining and knowledge discovery, data preparation, fuzzy rule-based systems, genetic fuzzy systems, knowledge extraction based on evolutionary algorithms, memetic algorithms, and genetic algorithms.

Óscar Cordón received the M.S. and Ph.D. degrees in computer science from the University of Granada, Granada, Spain, in 1994 and 1997, respectively.

He is a Professor with the Department of Computer Science and Artificial Intelligence since 1995 and Associate Professor from 2001 to 2007. Since 2006, he is the Principal Researcher of the Applications of Fuzzy Logic and Evolutionary Algorithms Research Unit at the European Centre for Soft Computing, Spain. He has published more than 40 papers in international journals indexed at the JCR Science Citation Index and coauthored the book *Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases* (Singapore: World Scientific, 2001). He has coedited six special issues of *Information Sciences*, *Mathware & Soft Computing*, the *International Journal of Approximate Reasoning, Fuzzy Sets and Systems*, the IEEE TRANSACTIONS ON FUZZY SYSTEMS, and *Applied Soft Computing*, as well as three books. He became Area Editor of the *International Journal of Approximate Reasoning* in 2005. He has worked on 18 research projects (as Coordinator of seven of them) supported by the European Commission, the Spain's and Andalusian Governments, the University of Granada, and the Puleva Food S.A. business concerning several aspects of genetic algorithms, fuzzy systems, genetic fuzzy systems, ant colony optimization and other metaheuristics, and e-Learning. His current main research interests are in the fields of soft computing for forensic anthropology and medical imaging, genetic fuzzy systems, soft computing and information retrieval, and evolutionary computation, ant colony optimization, and other metaheuristics.

Dr. Cordón created and Chaired from 2004 to 2007 the Genetic Fuzzy Systems Task Force within the IEEE CIS Fuzzy Systems Technical Committee, and was Treasurer of the EUSFLAT Society between 2005 and 2007. He was General Co-Chairman of the First International Workshop on Genetic Fuzzy Systems (GFS205), Granada, in March 2005.

Igor Zwir received the M.Sc. degree in computer science from the University of Buenos Aires, Buenos Aires, Argentina, in 1997 and the Ph.D. degree in computer science from the University of Granada, Granada, Spain, in 2001.

He is currently a Senior Research Scientist at the University of Granada and at Howard Hughes Medical Institute, Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO. He has been applying all of his original background in computational intelligence to system biology and bioinformatics problems for seven years, resulting in several publications in high-profile interdisciplinary journals His research interests in the biological field include genetic networks, transcriptional regulation, and gene expression dynamics. He also develops computational methods based on knowledge discovery based on conceptual clustering and evolutionary algorithms, genetic fuzzy systems, and control systems.