Mining Structural Databases: An Evolutionary Multi-Objetive Conceptual Clustering Methodology

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

[1] Dept. Computer Science and Artificial Intelligence, University of Granada, E-18071, Spain
{rocio, crubio, igor, ocordon, delval}@decsai.ugr.es
[2] Howard Hughes Medical Institute, Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110-1093, USA
zwir@borcim.wustl.edu

Abstract. The increased availability of biological databases containing representations of complex objects permits access to vast amounts of data. In spite of the recent renewed interest in knowledge-discovery techniques (or data mining), there is a dearth of data analysis methods intended to facilitate understanding of the represented objects and related systems by their most representative features and those relationship derived from these features (i.e., structural data). In this paper we propose a conceptual clustering methodology termed *EMO-CC* for *Evolutionary Multi-Objective Conceptual Clustering* that uses multi-objective and multi-modal optimization techniques based on Evolutionary Algorithms that uncover representative substructures from structural databases. Besides, EMO-CC provides annotations of the uncovered substructures, and based on them, applies an unsupervised classification approach to retrieve new members of previously discovered substructures. We apply EMO-CC to the Gene Ontology database to recover interesting substructures that describes problems from different points of view and use them to explain inmuno-inflammatory responses measured in terms of gene expression profiles derived from the analysis of longitudinal blood expression profiles of human volunteers treated with intravenous endotoxin compared to placebo.

1 Introduction

The increased availability of biological databases containing representations of complex objects such as microarray time series, regulatory networks or metabolic pathways permits access to vast amounts of data where these objects may be found, observed, or developed [1, 2, 3]. In spite of the recent renewed interest in knowledge-discovery techniques (or data mining), there is a dearth of data analysis methods intended to facilitate understanding of the represented objects

F. Rothlauf et al. (Eds.): EvoWorkshops 2006, LNCS 3907, pp. 159–171, 2006.

and related systems by their most representative features and those relationship derived from these features (i.e., structural data).

Structural data can be viewed as a graph containing nodes representing objects, which have features linked to other nodes by edges corresponding to their relationships. Interesting objects in structural data are represented as substructures, which consists of subgraph partitions of the datasets [4]. Conceptual clustering techniques have been successfully applied to structural data to uncover objects or concepts that relates objects, by searching through a predefined space of potential hypothesis (i.e., subgraphs that represent associations of features) for the hypothesis that best fits the training examples [5]. However, the formulation of the search problem in a graph-based structure would result in the generation of many substructures with small extent as it is easier to explain or model match 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 size of the substructure being explained, the number of retrieved substructures, and their diversity [4, 6]. The former are conflicting criteria that can be approached as an optimization problem. Multi-objective optimization techniques can evaluate concepts or substructures based on the conflicting criteria, and thus, to retrieve meaningful substructures from structural databases.

In this paper we propose a conceptual clustering methodology termed *EMO-CC* for *Evolutionary Multi-Objective Conceptual Clustering* that uses multi-objective and multi-modal optimization techniques. The EMO-CC methodology uses an efficient search process based on Evolutionary Algorithms [7, 8, 9], which inspects large data spaces that otherwise would be intractable. Besides, EMO-CC provides annotations of the uncovered substructures, and based on them, applies an unsupervised classification approach to retrieve new members of previously discovered substructures. We apply EMO-CC to the Gene Ontology database (i.e., the GO Project [3]) to recover interesting substructures containing genes sharing a common set of terms, which are defined at different levels of specificity and correspond to different ontologies, producing novel annotations based on them. Particularly, we use these substructures to explain inmuno-inflammatory responses measured in terms of gene expression profiles derived from the analysis of longitudinal blood expression profiles of human volunteers treated with intravenous endotoxin compared to placebo [10].

This work is organized as follows. Section 2 reviews the conceptual clustering problem. Section 3 describes the EMO-CC methodology. Section 4 shows the customization and results of applying EMO-CC to the GO database to explain gene expression profiles from the inflammatory problem. Section 5 introduces the discussion.

2 Conceptual Clustering

Cluster analysis –or simply clustering– is a data mining technique often used to identify various groupings or taxonomies in real-world databases [11]. Most ex-

isting methods for clustering are designed for linear feature-value data. However, sometimes we need to represent structural data that do not only contains descriptions of individual observations in databases, but also relationships among these observations. Therefore, mining into structural databases entails addressing both the uncertainty of which observations should be placed together, and also which distinct relationships among features best characterize different sets of observations, having in mind that, a priori, we do not know which feature is meaningful for a given relationship.

Conceptual clustering, in contrast to most typical clustering techniques [12], have been successfully applied to structural databases to uncover concepts that are embedded in subsets of structural data or substructures [4]. While most machine learning techniques applied directly or indirectly to structural databases exhibit methodological differences, they do share the same framework even though they employ distinct metrics, heuristics or probability interpretations [13, 4]: (1) *Database representation.* Structural data can be viewed as a graph containing nodes representing objects, which have features linked to other nodes by edges corresponding to their relations. A substructure consists of a subgraph of structural data [4]; (2) *Structure Learning.* This process consists of searching through the space for potential substructures, and either returning the best one found or an optimal sample of them; (3)*Cluster evaluation.* The substructure quality is measured by optimizing several criteria, including specificity, where harboring more features always increases the inferential power; sensitivity, where a large coverage of the dataset produces good generality; and diversity, where minimally overlapping between clusters generates more distinct clusters and descriptions from different angles; (4) *Database compression.* The database compression provides simpler representations of the objects in a database; and (5) *Inference.* New observations can be predicted from previously learned substructures by using classifiers that optimize their matching based on distance [14] or probabilistic metrics [5]).

3 An Evolutionary Multi-Objective Conceptual Clustering Methodology (EMO-CC)

We explicitly propose a method for each of the conceptual clustering steps mentioned:

(1) **Database representation** by using structures as graphs, where nodes correspond to database features and edges to the relationships among these features.
(2) **Structure learning** by searching in the feature space to obtain optimal substructures using an efficient multi-objective evolutionary algorithm, as well as appropriate objective definitions to guide the search relying on the NSGA-II algorithm [15]. Basic configuration of this algorithm is explained below:
Chromosome representation. EMO-CC encodes feasible substructures in the chromosomes of the algorithm population. Each chromosome is implemented

as a tree, where this representation in GAs is known as Genetic Programming (GP) [16]. This chromosome representation encodes each node and edge of the tree with a label, describing the type of feature, and an associated tag that indicates the value of such feature. The initial population consists of a set of chromosomes, each one built by choosing a random observation from the input database and extracting a subtree from its tree representation. The set of all non-dominated chromosomes of the final population represents a clustering of the given data.

Genetic operators. EMO-CC applies crossover and mutation operators with a given probability over the chromosomes composing the population of the GP. 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: (1) *Delete a leaf,* where a random leaf of the tree is selected and deleted along with the edge that connects it to the tree; (2) *Change a node,* where a random node is selected and replaced by another node belonging to the set of nodes constrained to have the same tag; and, (3) *Add a leaf,* where a random leaf is created and connected to the tree by a new edge.

Selection. EMO-CC uses a classical binary tournament selection method [17], which chooses two parent chromosomes and selects the one with the higher fitness value.

Multi-objective optimization. We consider that good substructures are those ones that maximize the *specificity* and *sensitivity* objectives. On the one hand, the specificity of a substructure is associated with its size (i.e., the number of objects and features that compose the substructure), which corresponds to the size of the tree represented in the chromosome. On the other hand, the sensitivity of a substructure is calculated as the number of instances that occur in the substructure, where an instance occur in a substructure if its tree representation is a subtree of the substructure tree. These are opposing objectives since the more specific the substructure, the less sensitive it becomes to detect new instances.

Non-dominance relationship. We select substructures that satisfy a trade-off between their specificity and sensitivity by selecting a set of solutions that are non-dominated, in the sense that there is no other solution that is superior to them in all objectives (i.e., Pareto optimal front [8, 6]). Another objective that is indirectly considered is the substructure diversity, which consists of maintaining a distributed set of solutions in the Pareto front. Therefore, to address all of these objectives our approach applies the non-dominance relationship locally, that is, it identifies all non-dominated optimal substructures that have no better solution in a neighborhood [8, 6]. We consider that two substructures are in the same neighborhood if they have at least a 50% of instances occurring in both of them calculated based on the *Jaccard's coefficient* [18].

(3) **Clustering evaluation** applying the non-dominance relationship between conflicting criteria in a neighborhood to achieve cohesive, well supported, and diverse substructures.

(4) **Compression of substructures** based on an circumstantial query, thus allowing flexible and adaptive substructures to different contexts.
(5) **Inference** by using an unsupervised fuzzy k-nearest prototype classifier that characterizes new instances based on available knowledge. It calculates the membership of a query observation x_q in a set of I previously identified substructures.

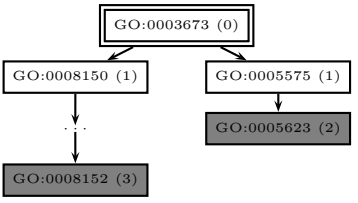
4 Application of the EMO-CC Methodology to the Gene Ontology Structural Database

Massive microarray experiments provide a wide view of the gene regulation problem; however, most of the biological knowledge extracted from these experiments include few relevant genes, some of which are difficult to be identified because of their low expression levels. Moreover, it is also difficult to distinguish among expressed genes that behave differentially between treatments, time, patients and other factors that are always hidden in typical microarray protocols (e.g., gender or age). Here we focus on the challenge of explaining these profiles and re-discover them based on independent biological information.

We therefore apply EMO-CC to discover interesting substructures in the Gene Ontology database that can explain classes composed of microarray gene profiles having similar behaviors of their expression over time, treatment, and patient. The Gene Ontology (GO) network stores one of the most powerful characterization of genes, containing three structured vocabularies (i.e., ontologies) that 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). Particularly, from the computational point of view, these networks are organized as structures called directed acyclic graphs (DAGs), which are one way routed graphs that can be represented as trees. Therefore, identifying which distinct relationships among features best characterize different sets of observations does not only have to consider the process of grouping distinct type of features, but also defining at which level of specificity they have to be represented.

4.1 EMO-CC Customization for the GO Domain

We used the GO database and compatibilized the terms with descriptions provided by Affymetrix, where each observation of the database has the following features: (1) *Name:* Affymetrix identifier for each gene in HG-U133A v2.0 set of arrays; (2) *Biological process:* List of the biological processes where a gene product is involved (e.g., mitosis or purine metabolism); (3) *Molecular function:* List of the biological functions of the gene product (e.g., carbohydrate binding and ATPase activity), which is indexed by a list of integer GO codes; and (3) *Cellular component:* List of the cellular components indicating location of gene



(a) Chromosome

207929_at
202442_at
213475_s_at
214467_at
219889_at
203508_at
201087_at
209864_at
200625_s_at

(b) Genes

Specificity = 0.6769
Sensitivity = 0.0051

(c) Objectives

Fig. 1. An example of a chromosome representing a cluster. (a) The tree representation, gray boxes represent the most specific GO terms of the concept of the cluster, the level of each term is shown between parenthesis. (b) The list of genes that correspond to the cluster. (c) The values corresponding to the sensitivity and specificity objective functions.

products (e.g., nucleus, telomere, and origin recognition complex), which are indexed by a list of integer GO codes.

An instance for the GO domain is redefined as the particular subset of values that constitutes a prefix tree[1] of a database observation in contrast to a subtree as in the general case. Then, an instance occurs in a substructure if a subgraph of the prefix tree that represent that instance matches with the substructure tree, where this tree contains tagged nodes with the type of feature (e.g., biological process), and the corresponding values (e.g., GO:0007165), and the edges represent relationship between features (i.e., tagged nodes).

Good substructures are those ones that result in a trade-off between sensitivity and specificity. Although, the sensitivity can be calculated based on the number of instances in a substructure, the specificity of the substructure is not linearly dependent to its size, as it was previously defined based on the number of nodes and edges because of the level component included in the GO domain. Thus, we redefine the specificity as the distance among all most specific nodes of an instance i and the closest leaf-node in the substructure S:

$$Specificity(S) = \frac{\sum_i^K \sum_u^U \frac{dist(node_u, node_i)}{level(node_i)}}{K} \quad (1)$$

where the distance is calculated as the number of edges between two nodes, the level of a node is calculated as the length of the shortest path to the root node, U is the number of leaf-nodes in substructure S, and K is the number of instances occurring in substructure S. An example of a chormosome representing a cluster concept is shown in Figure 1.

4.2 Experiments and Analysis of Results

The structural database used for the GO domain is composed of 1770 instances of genes and their GO associated terms. The population of the evolutionary

[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.

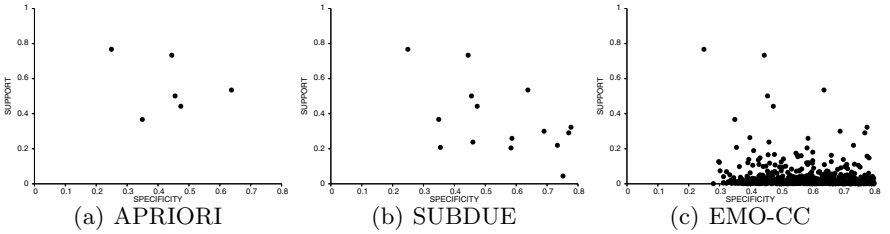
algorithm is initialized by 50% of randomly chosen subtrees of the database and by another 50% of random instances. The parameters of the algorithms used for this domain are shown in Table 1. The EMO-CC approach was run ten times with different seeds and the average of these runs is reported.

Table 1. Parameters for the GO domain

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

4.3 Computational Analysis

We compare EMO-CC with two other methods, APRIORI and SUBDUE, all of which satisfy in some extent those features shared by machine learning methods introduced in Section 3. Although APRIORI and SUBDUE are not MO algorithms, we illustrate the obtained Pareto fronts in Figure 2 to perform fair comparisons with EMO-CC. In addition, we verify the performance of the former methods by applying some multi-objective comparison metrics, namely $\mathcal{C}$ and $\mathcal{ND}$ [19, 20]. The metric $\mathcal{C}(X', X'')$ measures the dominance relationship between the set of non-dominated solutions X' over other set of non-dominated solutions X''. The value $\mathcal{C}(X', X'') = 1$ means that all points in X'' are dominated by points in X'. The opposite, $\mathcal{C}(X', X'') = 0$, represents the situation where none of the points in X'' are covered by the set X'. The metric $\mathcal{ND}(X', X'')$ compares two sets of non-dominated solutions and gives the number of solutions of X' not equal and not dominated by any member of X''. The values obtained by the methods are shown in Table 2,



(a) APRIORI (b) SUBDUE (c) EMO-CC

Fig. 2. Pareto fronts for the GO domain by using two conflicting objectives: specificity and sensitivity. (a) Non-dominated solutions reported by the APRIORI method. (b) Solutions recovered by the SUBDUE method. (c) Substructures recovered by the EMO-CC methodology, where more than one solution for the same specificity level indicates that they correspond to different neighborhoods.

The obtained results of applying the former metrics reveal that there is no solution obtained by EMO-CC that is dominated by APRIORI, and only one solution obtained by SUBDUE dominates solutions belonging to the Pareto front found by EMO-CC (Table 2(a)), as described by metric $\mathcal{C}$, while there is no solution of the latter method that dominates any solution from the other two approaches. Moreover, the EMO-CC method discovers more non-dominated solutions, as evaluated by metric $\mathcal{ND}$ (Table 2(b)), than both APRIORI and SUBDUE methods. The difference between the values reported by the $\mathcal{ND}$ metric from EMO- CC and those ones from APRIORI and SUBDUE (i.e., 181.89 and 171.80 vs. 1.20 and 1.60 from Table 2(b)) suggests that EMO-CC retrieves almost all solutions identified by the other methods and covers a wide set of all of optimal solutions that can be obtained in the GO domain. This is in contrast to the few solutions that are identified by the APRIORI and SUBDUE methods, but remain undetected by the EMO-CC method (i.e., 1.20 and 1.60 in average from Table 2(b)).

In addition, the EMO-CC method recovers most and more diverse solutions than those found by the APRIORI and SUBDUE methods. Particularly, our approach retrieves substructures of the Pareto optimal front containing few instances harboring several features (i.e., cohesive substructures), which were undetected by the other methods.

Table 2. Comparative evaluation of the solutions identified by APRIORI, SUBDUE and EMO-CC for the GO domain by using different metrics

(a) $\mathcal{C}$ metric

$\mathcal{C}(X', X'')$	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*)	-

(b) $\mathcal{ND}$ metric

$\mathcal{ND}(X', X'')$	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*)	-

Biological results analysis using gene expression profiles. We consider 24 independent classes containing gene expression profiles derived from the analysis of 48 GeneChips® HG-U133A v2.0 from Affymetrix Inc., corresponding to an inflammatory response study performed on human volunteers treated with intravenous endotoxin compared to placebo [10]. The data has been acquired from samples taken from human blood to eight patients over time at 0, 2, 4, 6, 9 and 24 hours, where four had been treated with intravenous endotoxin (i.e., patients 1 to 4) and four with placebo (i.e., patients 5 to 8). We will use these gene expression profiles for validating the substructures detected by EMO-CC, or, in other words, which are explained by these substructures.

Table 3. Clusters derived from the GO information by EMO-CC intersecting significantly with class #13 from the gene expression information. Solid lines separate groups of clusters which GO information is not related, while dashed lines separate clusters within these groups, as shown in Figure 3.

#Substr.	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)

For example class #13 is described by several substructures (Table 3). Significantly, these descriptions are based on different types of descriptions (e.g., process and cellular components) that belong to different levels of the GO structure (e.g., level 6 or level 4). These diverse substructures are optimal in the sense that belong to the Pareto optimal front (Figure 2) between specific and sensitive descriptions. The effect of the substructures on the explained class #13 can be visualized in (Figure 3).

EMO-CC, as a machine learning method (see Section 3 (4)), compresses those substructures that explain an expression profile from the same point of view to provide a summarized explanation of this phenomena (Table 3). For example, substructures #89 and #216 are compressed because they are indistinguishable for the class corresponding to the expression profile #13, while substructure #179 describes it from a very different point of view and is preserved as a diverse solution. This compression is dynamic because substructures are re-grouped in a context-dependent fashion, where the context corresponds to an explained class and a different classification can produce a distinct substructure association (e.g., substructures #89 and #216 are indistinguishable for class #13, while may be not the case for other class of microarray or clinical experiments). Notably, this

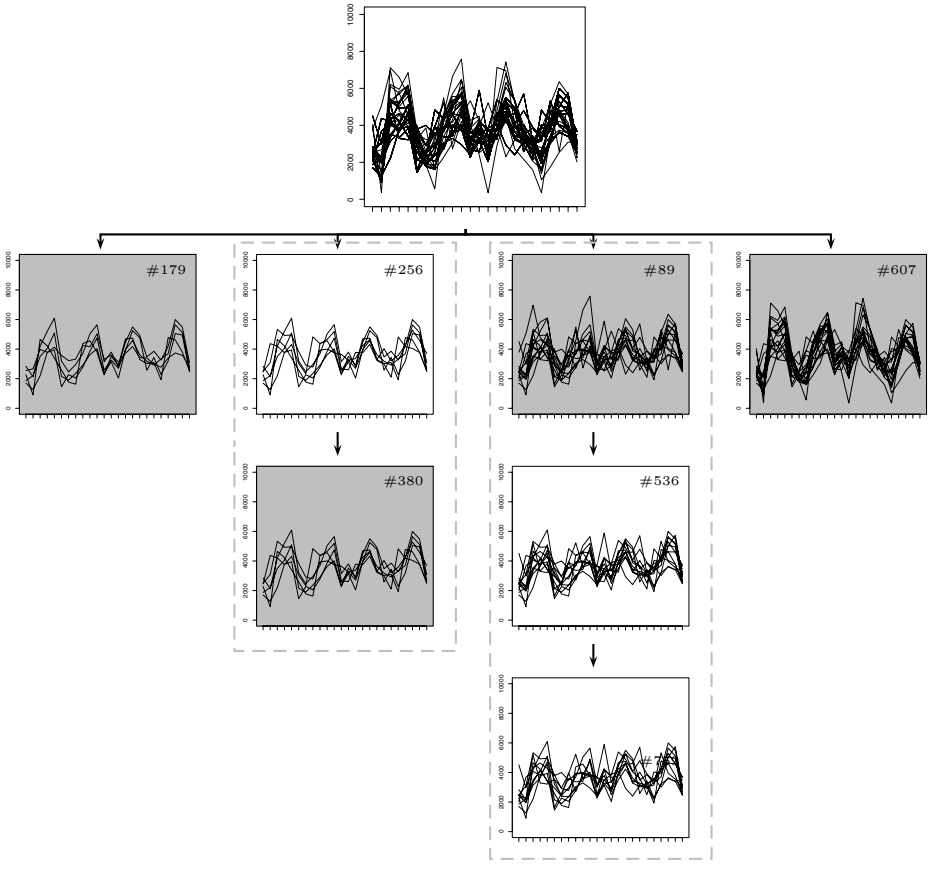


Fig. 3. The effects of the explanation of the expression class #13 based on the GO substructures identified by EMO-CC. The dashed rectangle illustrated the local application of the non-dominance relationship within a class, and the summarization of two indistinguishable substructures for this class. Grey filled graphs correspond to the compressed substructures of Table 3.

classification is performed based on completely external information provided by GO database, instead of the levels of expression.

In addition, EMO-CC applies an unsupervised inferential approach (see Section 3 (5)) which calculates the membership of a query observation x_q in a set of I previously identified substructures, to classify new instances. Since the obtained substructures are not disjoint, a given observation may belong to more than one cluster.

The unsupervised inferential mechanism of EMO-CC allows to identify new genes belonging to a particular expression profile. This is exemplified by the gene `212659_s_at`, which was recovered by its proximity to substructure #824 and shows a similar expression pattern to the genes of class #17 (Figure 4), but was ignored by the statistical methods used to recover differentially expressed genes

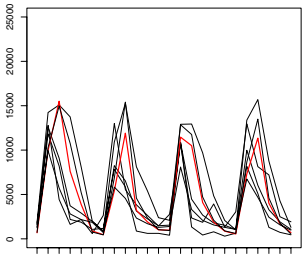


Fig. 4. Expression of Substructure B #824 where gene product `212659_s_at` is classified. The observation classified is highlighted.

[10]. It is noteworthy that this gene was not identified by its similarity with the centroid of the expression class #17, but from an independent substructure provided by EMO-CC.

5 Discussion

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 biological network would often result in a conflicting paradigm. On the one hand, generating a large number of substructures, each containing a very small number of genes that share all considered features, makes it hard to find commonalities among similarly regulated genes. On the other hand, generating a small number of groups in which their members share a limited number of features, would fail to discriminate between members of a molecular pathway.

In order to tackle these problems, we proposed the EMO-CC methodology that identifies conceptual clusters and classifies co-regulated genes based on multiple features that characterizes them, including functional descriptions, molecular processes and cellular components, at different levels of specificity.

EMO-CC allows gene membership to more than one substructure by using a flexible classifier [14, 21], thus, explicitly treating the substructures as hypotheses, that can be tested and refined [5]. Moreover, these hypotheses can produce novel annotations among different types of features at multiple specificity levels, which explain co-regulation phenotypes and can be used to conduct gene-wide searches.

Also, EMO-CC considers gene expression as one independent feature, thereby allowing classification of genes even in the absence of its expression. Moreover, EMO-CC minimizes the number of substructures by using a flexible compression strategy that groups similar substructures based on their ability to describe gene profiles derived from different experimental conditions (e.g., microarray expression, or Chip-on-Chip binding occupancy).

Our proposed methodology is applicable to a wide set of domains, being easily to customize to particular problem, and may be an appropriate white-box technique to uncover rear 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].

References

1. Siripurapu, V., Meth, J., Kobayashi, N., Hamaguchi, M.: Dbc2 significantly influences cell-cycle, apoptosis, cytoskeleton and membrane-trafficking pathways. Journal of Molecular Biology **346** (2005) 83–89
2. Nikitin, A., Egorov, S., Daraselia, N., Mazo, I.: Pathway studio–the analysis and navigation of molecular networks. Bioinformatics **19** (2003) 2155–2157
3. Consortium, T.G.O.: Gene ontology: tool for the unification of biology. Nature Genet. **25** (2000) 25–29
4. Cook, D., Holder, L., Su, S., Maglothin, R., Jonyer, I.: Structural mining of molecular biology data. IEEE Engineering in Medicine and Biology, special issue on Advances in Genomics **4** (2001) 67–74
5. Mitchell, T.: Machine Learning. McGraw-Hill, New York (1997)
6. Ruspini, E., Zwir, I.: Automated generation of qualitative representations of complex object by hybrid soft-computing methods. In Pal, S., Pal, A., eds.: Pattern Recognition: From Classical to Modern Approaches, Singapore, World Scientific Company (2001) 453–474
7. Back, T., Fogel, D., Michalewicz, Z., eds.: Handbook of Evolutionary Computation. IOP Publishing Ltd., Bristol, UK (1997)
8. Deb, K.: Multi-Objective Optimization Using Evolutionary Algorithms. John Wiley & Sons, Inc. (2001)
9. Coello-Coello, C., Veldhuizen, D.V., Lamont, G.: Evolutionary Algorithms for Solving Multi-Objective Problems. Kluwer (2002)
10. Romero-Zaliz, R., Cordón, O., Rubio-Escudero, C., Zwir, I., Cobb, J.: (A multi-objective evolutionary conceptual clustering methodology for gene annotation from networking databases) Submited.
11. Duda, R., Hart, P., Stork, D.: Pattern Classification (2nd Edition). Wiley-Interscience (2000)
12. Der, G., Everitt, B.: A handbook of statistical analyses using SAS. CHAPMAN-HALL (1996)
13. Cheeseman, P., Oldfors, R.W.: Selecting models from data. Springer-Vlg (1994)
14. Bezdek, J.: Fuzzy clustering. In Ruspini, E., Bonissone, P., Pedrycz, W., eds.: Handbook of Fuzzy Computation, Institute of Physics Press (1998) f6.1:1–f6.6:19
15. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation **6** (2002) 182–197
16. Koza, J.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, MA, USA (1992)
17. Goldberg, D.: Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley (1989)
18. Jaccard, P.: The distribution of flora in the alpine zone. The New Phytologist **11** (1912) 37–50

19. Zitzler, E., Thiele, L.: Multiobjective evolutionary algorithms: A comparative case study and the Strength Pareto Approach. IEEE Transactions on Evolutionary Computation **3** (1999) 257–271
20. Romero-Zaliz, R., Zwir, I., Ruspini, E.: Generalized Analysis of Promoters (GAP): A method for DNA sequence description. In: Applications of Multi-Objective Evolutionary Algorithms. World Scientific (2004) 427–450
21. Gasch, A., Eisen, M.: Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. Genome Biology **3** (2002)