Journal of Biomedical Informatics 43 (2010) 1–14

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

SoFoCles: Feature filtering for microarray classification based on Gene Ontology

Georgios Papachristoudis [a], Sotiris Diplaris [b,*], Pericles A. Mitkas [b]

[a] MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
[b] Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 25 June 2008
Available online 1 July 2009

Keywords:
Data mining
Semantic similarity
Ontologies
Bioinformatics
Feature filtering
Microarray classification

ABSTRACT

Marker gene selection has been an important research topic in the classification analysis of gene expression data. Current methods try to reduce the "curse of dimensionality" by using statistical intra-feature set calculations, or classifiers that are based on the given dataset. In this paper, we present SoFoCles, an interactive tool that enables semantic feature filtering in microarray classification problems with the use of external, well-defined knowledge retrieved from the Gene Ontology. The notion of semantic similarity is used to derive genes that are involved in the same biological path during the microarray experiment, by enriching a feature set that has been initially produced with legacy methods. Among its other functionalities, SoFoCles offers a large repository of semantic similarity methods that are used in order to derive feature sets and marker genes. The structure and functionality of the tool are discussed in detail, as well as its ability to improve classification accuracy. Through experimental evaluation, SoFoCles is shown to outperform other classification schemes in terms of classification accuracy in two real datasets using different semantic similarity computation approaches.

1. Introduction

A challenging problem in bioinformatics is the analysis of microarray experiments. Microarrays can monitor the expression of thousands of genes or gene products simultaneously, under varying conditions. Different types of analysis can be applied to the datasets produced, including clustering, classification or density estimation, among others. Recent studies have widely used classification techniques in problems related to cancer-gene expression studies, where a typical problem is the differentiation between tumors based on a set of tissue-specific gene expression profiles [1–3], or between patients and normal controls [4]. Despite the advantages of such techniques, the sheer data volume coupled with the great asymmetry between the number of genes monitored and the number of different conditions renders the extraction of useful information a formidable task. To tackle this problem, several methods have been exploited, most of them stemming from the field of statistical analysis. The isolation of the most informative genes is achieved by detecting statistical similarities among the features/genes based on their values across the whole dataset. The application of feature filtering methods in microarray classification aims at improving classification accuracy [5], mainly through dimensionality reduction achieved by keeping the most informative features while rejecting irrelevant and noisy ones [6]. The "curse of dimensionality", which can lead to overfitting problems, can be minimized by feature filtering. On the other hand, the presence of features that are irrelevant to the problem may affect the discrimination ability of models induced from the data, as well as the extraction of correlations among them [7].

In microarray analysis problems, the selected features that represent genes are known as marker genes. By selecting groups of features, instead of single features, the dataset is enriched with information regarding the interaction between genes, since the regulation of genes in the same biological path is correlated [8]. Thus, feature filtering can possibly provide a better understanding of the process that has produced the data. Finally, the functionality of new genes can be defined through the study of structural and functional behavior of known marker genes. The information gain [9], the chi-square approach [10], and the relief method [11] are among the legacy methods that have been used in feature selection for microarray analysis so far [12].

Classic feature filtering algorithms have proved quite useful in microarray classification. However, they also exhibit some drawbacks that can be overcome by semantic feature filtering. One drawback is the risk to select features/genes that are highly correlated [7]. This usually happens because genes that belong to common biological paths exhibit similar regulation, thus having similar expression profiles. The result is that when using classic feature filtering methods, statistically correlated genes receive the highest scores [13]. However, the use of many highly correlated genes does

* Corresponding author. Address: Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece. Fax: +30 2310 996398.
E-mail address: diplaris@issel.ee.auth.gr (S. Diplaris).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.06.002

not guarantee better classification accuracy, since such a selection may lead to information redundancy.

Moreover, if a certain threshold is set in the gene selection process, there is a high probability of selecting genes that belong only to the biological path with the main influence. As a result, complementary genes that reveal other class labels may not be included in the feature set, since they would receive lower scores. Finally, the number of selected genes always requires human intuition, thus making it difficult to select an optimal feature set. In this way, it is possible to exclude genes that participate in biological paths that might equally contribute in a robust data representation of the problem. Our approach aims to overcome this problem, by using existing biological knowledge in order to allow the reselection of genes that participate in crucial biological paths, ultimately aiming at the improvement of classification accuracy.

Statistical methods cannot incorporate the *a priori* semantic knowledge that lies behind the feature set and has been produced by other bioinformatics tools. The need to standardize the functions and activity regions of gene products has led to the Gene Ontology Annotation (GOA) project [14], which is supported by all major Sequence/Genome Databases, such as SwissProt [15], InterPro [16], Gramene [17], MGI [18], FlyBase [19], DictyBase [20], and TAIR [21]. So far, Gene Ontology analysis tools for microarray data have been developed in order to look up for existing annotations, cluster genes into categories and perform a significance analysis over them. There is a plethora of such tools, all implementing the same idea. The interested reader can refer to a relative review by Khatri and Draghici [22] for a performance comparison. However, the gene product characterization in conjunction with *semantic similarity* discovery in a structured vocabulary of genetic concepts can also provide quantitative knowledge that can be exploited in order to produce better datasets suitable for microarray classification. Semantic similarity has been longtime used in the Information Retrieval and Natural Language Processing fields in order to measure the likeness of the semantic content of documents or terms [23–27]. The same principle has also been used in order to derive geospatial content maps [28].

However, in microarray supervised data mining the existing semantic approaches, which have been recently reviewed by Bellazi and Zupan [29], utilize semantic knowledge either for defining class labels [30], or for statistically defining *a priori* sets of gene groups (e.g. genes found in the same metabolic pathways, located in the same cytogenetic band, or sharing the same Gene Ontology category) that are used for the classification task instead of selecting individual problem-specific genes [31]. In another study by Qi and Tang [32], gene expression data were classified with a methodology, which for the feature selection task used statistical methods and the Gene Ontology in order to remove redundant features from an already arbitrarily reduced initial feature set. In that study similarity tests involved the computation of a metric called *conjunctive similarity*, which was used in order to measure gene similarities within their feature set. The metric they introduced combines statistical correlation with semantic similarity using weights. In this sense, and given that the weight they usually assigned for the semantic similarity result was small, the use of semantic similarity aimed rather at verifying the results of the statistical feature selection methods or in the best case just to support them. Their aim to use semantic similarity only as a means of verification of the statistical methods results, and not as a self-contained means of feature set enrichment, is further cleared in their follow-up paper [33], where the authors explicitly claim that they use the discriminative values of GO terms to verify the statistical discriminative values of genes. Moreover, in these studies the feature set consisted of a small fraction of the whole set of genes (it typically comprised about 100 genes), while the rest of them were ignored and not used at all, claiming that the remaining genes were traditionally enough to describe the classification problem, and that there were even more redundant genes in this feature set that needed to be removed.

In this paper a novel approach for microarray classification is presented that integrates semantic knowledge with legacy statistical methods with the help of semantic similarity, aiming to reinforce the feature set used for classification by enriching it with new marker genes. Our approach has been implemented as a software platform called SoFoCles. The quality of microarray classification can be enhanced by exploiting the knowledge available in a structured hierarchy of genetic concepts, the Gene Ontology (GO), instead of merely applying feature selection methods. In this way, scientifically proven information with biological meaning is incorporated into the feature filtering procedure, in order to improve the classification accuracy. To achieve this, a repository of semantic similarity methods has been built into SoFoCles. The user may choose one or more of these methods to better identify feature similarities based on the terms of the Gene Ontology that each gene or gene product is tied to. The initial whole feature set (*W*-set) is first filtered using legacy statistical methods producing a refined small feature set (*R*-set). Then, the GO-based similarities are used to enrich the *R*-set using the SoFoCles algorithm. Information concerning genes or gene products of the refined dataset is preprocessed in order to identify the related GO terms and infer semantically similar genes that are involved in the same biological process. These genes enrich the *R*-set, yielding the semantically aware *S*-set, which describes better the biological paths regulated in the conditions tested, thus improving classification accuracy.

Compared to the approach in [31] that also utilizes the Gene Ontology, SoFoCles is able of selecting features that are problem-specific, constructing feature sets on demand and not *a priori*. With respect to the feature selection approaches in [32,33], SoFoCles is substantially different and innovative in terms of aim, methodology and metrics, firstly since it aims at *enriching* a small initial feature set (*R*-set), while the other approaches aim at *removing* redundant features from a larger initial feature set. In this context, SoFoCles involves the calculation of *semantic* similarities between two feature sets (*R*-set and the rest of the genes) in order to derive an enriched, semantically-aware final feature set (*S*-set), and not the calculation of *conjunctive* similarities in one single feature set. Within SoFoCles, semantic similarity is used solely as the gene discriminative method between the two feature sets for discovering new marker genes, instead of only verifying the statistical correlations between genes as in the other two approaches. Most importantly, a major innovation of SoFoCles is that it uses and semantically compares *all* genes of the initial feature set (*W*-set) and not only a small fraction of them (as in [32,33]), implying that prior knowledge (in the form of semantic similarities in the Gene Ontology) can be used in order to enrich the *R*-set with features that would not be normally included in the feature set if the traditional statistical techniques were used. Thus, using SoFoCles, the discovered genes could be found anywhere in the initial feature space, even between the large multitude of genes that are arbitrarily discarded by the other two methodologies.

The rest of the paper is organized as follows. In Section 2 the Gene Ontology aspects are presented and reliability issues in GO annotations are discussed. Section 3 covers the discovery of semantic similarity in the Gene Ontology using the different methods implemented within SoFoCles. Section 4 describes our methodology for semantic feature selection, while Section 5 presents the structure and functionality of the SoFoCles platform. Several experiments in microarray classification have been conducted with SoFoCles. Some of them are described in Section 6 together with representative results indicating the improvement in terms of classification accuracy, along with the relevant discussion regarding the added value of SoFoCles. Finally, Section 7 concludes and presents the future outlooks of SoFoCles.

2. GO aspects and annotation evidence codes

Gene Ontology constitutes a controlled and structured vocabulary, which supports querying in different layers. Curators, who annotate its terms to genes or gene products, are allowed to select the term of a layer that corresponds to the existing knowledge for a gene or a gene product. GO has three different aspects/ontologies that contain information regarding all species: (a) biological process, (b) molecular function, and (c) cellular component. The main characteristic of these ontologies is that they are orthogonal to each other; each term exists only within one of the three aspects, thus guaranteeing the uniqueness of a feature annotated to a gene. The biological process (BP) aspect comprises one or more function stages; nevertheless it cannot describe full biological paths. The molecular function (MF) aspect describes the biochemical activity of a gene product without defining the location or the timing of the activity, thus describing activities rather than gene products. The cellular component (CC) aspect describes locations within the cell, where gene products are active. Therefore, MF and CC aspects answer the question "what a gene product does and where it is active inside a cell", while the BP aspect explains the biological effect of a gene product [34]. Fig. 1 depicts an instance of the biological process aspect.

GO supports two types of relations, the "is_a" and the "part_of" relation. In general "is_a" relations are not interconnected with "part_of" relations in order to avoid abstractness [35]. Table 1 depicts the terms and relations distributions in Gene Ontology.

Table 1
Distribution of GO terms and relations in the ontology (The Gene Ontology, May 2007 release).

Ontology aspect	# Terms	# Relations (# edges)		
		# is_a	# part_of	Total
Cellular component	1957	2925	831	3756
Molecular function	7592	8859	1	8860
Biological process	13,509	20,425	3899	24,324
Obsolete set	1101	–	–	–
Total	24,159	32,209	4731	36,940

GO terms are annotated to gene products through the GOA project. Gene annotation follows two principles: (a) the source of the annotation should be mentioned, and (b) the type of evidence that supports the annotation must be provided. In general, annotations apply to gene products and not to genes themselves, since a gene may encode the formation of many different products with distinct properties.

The reliability of a GO annotation is directly dependent on the type of evidence that supports it. In general, electronic and manual evidence is available within the Gene Ontology terms and gene products. Manual evidence is based on published information, either in scientific bibliography, experiments, or biological knowledge, always with respect to the trustworthiness level of the GO curator. Therefore, a certain level of subjectivity cannot be avoided. Automated processes, on the other hand, exploit corresponding

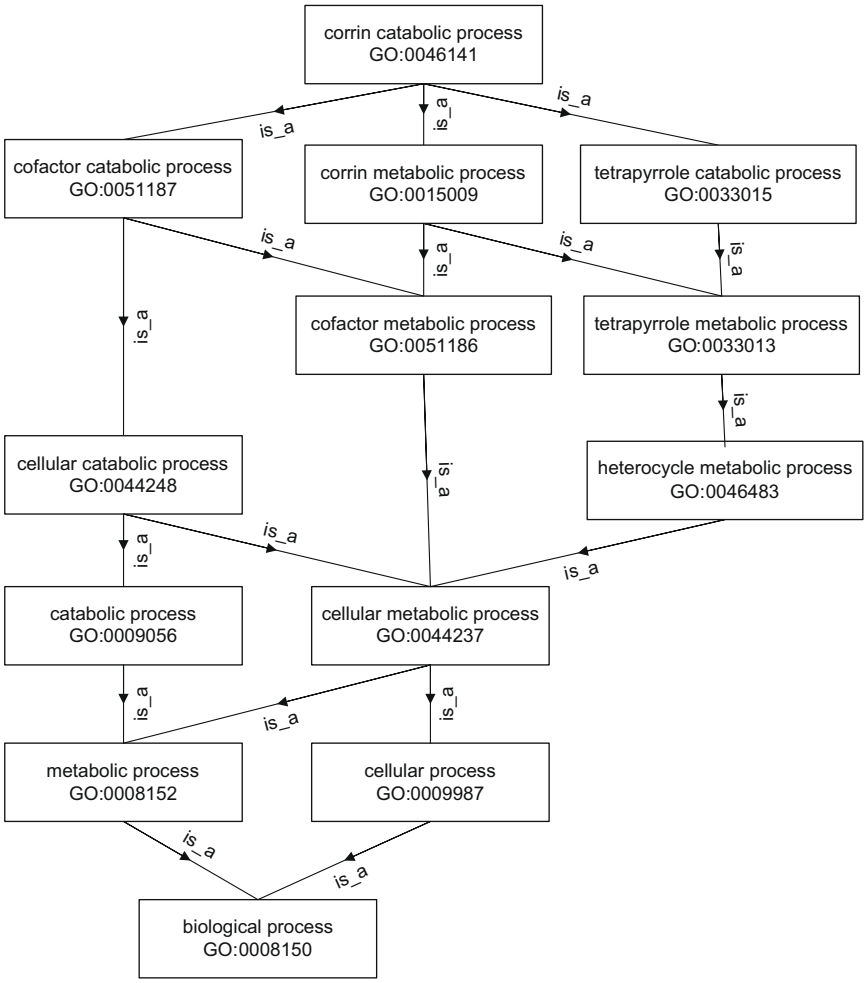

Fig. 1. An instance of the biological process aspect of Gene Ontology in the form of a Directed Acyclic Graph. It can be observed that there are nodes having more than one parent. The root of the ontology aspect is the term: GO:0008150, biological_process.

data and cross-references from other databases. These annotations are usually products of bioinformatics methods, like sequence alignments or scientific text mining. Such methods prove much faster and consistent, compared to supervised methods, since they rely on solid rules; nevertheless, they lack credibility compared to manual annotations, since automated methods tend to reproduce human errors that occur during the recording procedure in biological databases [36], and also because they tend to transfer annotation that is correct for one gene onto a gene where it is inappropriate.

In Gene Ontology there exist 14 different evidence codes that represent different types of evidence, though only one of them is a completely automatic approach (IEA: Inferred from Electronic Annotation). The rest of them, including some electronic methods, require the results to be reviewed manually by a curator. Specifically, these methods are: IC (Inferred by Curator), IDA (Inferred from Direct Assay), IEP (Inferred from Expression Pattern), IGC (Inferred from Genomic Context), IGI (Inferred from Genetic Interaction), IMP (Inferred from Mutant Phenotype), IPI (Inferred from Physical Interaction), ISS (Inferred from Sequence or Structural Similarity), NAS (Non-traceable Author Statement), ND (No biological Data available), RCA (inferred from Reviewed Computational Analysis), TAS (Traceable Author Statement) and NR (Not Recorded). Fig. 2 depicts a ranking of the evidence codes with respect to their reliability.

3. Semantic similarity in Gene Ontology

A major innovation of SoFoCles is the exploitation of existing biological knowledge for the microarray classification task. This can be realized by drawing knowledge from the Gene Ontology. It is of great importance, though, to employ a mechanism for knowledge quantification in order to allow the application of concepts such as correlation and similarity between GO terms.

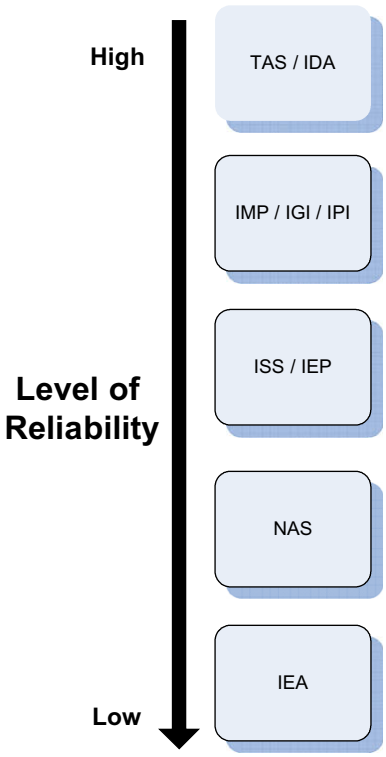

Fig. 2. Reliability index of GO evidence codes. In general, TAS and IDA evidence codes are considered more reliable, since they are inferred from publications. IEA is the least reliable code, being inferred by automatic sequence alignments that are not validated by curators.

Quantifying such information is enabled via the concept of *semantic similarity*, by which the integration of ontology's structured information becomes plausible. By definition, semantic similarity is a concept whereby a set of documents or terms within term lists are assigned a metric based on the likeness of their meaning/semantic content. In general, the terms that are labeled as ontology nodes are placed farther apart in the ontology graph, as their semantic similarity decreases.

Apart from the obvious metric of the distance between two ontology nodes, i.e. the number of edges of the minimum path that connects the nodes, there are also other quantification metrics. A well-established metric that is hereby used in semantic similarity techniques is the concept of information content of an ontology node.

The concept of information content in biological terms depicts an index of generality of a biological term. Specifically, the Gene Ontology comprises a structured vocabulary with hierarchical relations among its terms. Some are more generic, whereas others are more specific. Intuitively, the deepest layers of the ontology contain the more specific terms. An obvious metric to measure the relation between two nodes or a node's generality is the measurement of the distance among the nodes or between a node and the graph's root. Although the conceptual approach of such a methodology is quite simple, it features the important disadvantage of ignoring the local density in each node, thus failing in non-uniform density graphs or trees. Moreover, the fact that some ontology branches are longer than others is not taken into account, thus normalization is needed. For example, the GO terms GO:0008218 "*bioluminescence*" (depth 3) and GO:0045994 "*positive regulation of translational initiation by iron*" (depth 8) are both ontology leafs. Leafs in this example describe a certain whole biological process in the specific time. Thus, these two terms should include the same information content [37].

The information content *IC* of a term t is defined by Eq. (1):

$$IC(t) = -\log(p(t)) = -\log\left(\frac{n_t}{n_r}\right) \tag{1}$$

where:

$p(t)$: probability of t,
n_t: the frequency of the term or any of its descendants,
n_r: the frequency of the root or any of its descendants.

Normalizing the quantity in Eq. (1), it is possible to acquire the information gain of a leaf. Ontology leaves are considered to describe in full a certain process, by also incorporating the maximum information content. The probability of such a leaf, would be:

$$p(leaf) = \frac{1}{n_r} \tag{2}$$

while its information content would be:

$$IC(leaf) = -\log(p(leaf)) = -\log\left(\frac{1}{n_r}\right) \tag{3}$$

Thus, the information content of a term t could be normalized by dividing with the information content of an ontology leaf:

$$IC_{normal}(t) = \frac{\log(p(t))}{\log(p(leaf))} = \frac{\log\left(\frac{n_t}{n_r}\right)}{\log\left(\frac{1}{n_r}\right)} = 1 - \frac{\log(n_t)}{\log(n_r)} \tag{4}$$

The above quantity takes values in $[0, 1]$ and it is minimized when the term is the ontology root, whereas it takes its maximum value of 1 when the term is a leaf.

A term in higher levels of the ontology is considered to refer to a gene product, even if any descendant of the specific term is assigned to this gene product. Given the above, the position of a term within the ontology is directly related to the number of annotations of this term to gene products. In general, more generic terms occur in higher layers of the ontology. However, this in not always the case; in Table 2 it can be observed that, for some terms that lie deeper in the ontology, there may be others above them that are more generic having lower information content. Normalization attempts to override these disadvantages, however the disambiguation of the ontology leafs sense is the key to improve the descriptive capability of the ontology.

The convention to equally assign a gene product to the ascendants of a term is based on the *true path rule*, which is valid in Gene Ontology, and stipulates that if the child term describes the gene product then all its parent terms must also apply to that gene product. For example, the term "*alkali metal ion binding*" can be considered as "*metal ion binding*" (one of its direct parents) or also as "*ion binding*" (one of its ancestors). Thus, a term reserves all of its ancestors' features and it can be regarded as any one of them.

In this context, the employed semantic similarity methods are able to compute pairwise similarities between GO terms. The various approaches involve either the use of edges or the use of nodes of the ontology, or even a combination of both. In general, a node-based approach takes into account the information lying inside ontology nodes, by measuring how common the information that lies inside an ontology node is with that of another node.

The edge-based methods rely on the measurement of distance (edge length) between the nodes considered. Both approaches aim at computing semantic similarity, but from different aspects. The edge-based approach lies closer to human intuition, while the node-based approach is theoretically more robust [38].

In terms of SoFoCles, the following semantic similarity methods were employed:

(a) Resnik [39]:

This is the simplest node-based technique. It utilizes the minimal subsumer of two terms, defined as the common ancestor concept between t_1 and t_2 that has the minimal number of descendants, in order to derive the similarity measure between terms t_1 and t_2:

$$\text{R-sim}_{norm}(t_1,t_2)=\frac{\max_{t\in S(t_1,t_2)}[IC(t)]}{IC(leaf)} \tag{5}$$

where $S(t_1,t_2)$ is the set of common ancestors of t_1 and t_2.

Table 2
Some GO terms of the Biological Process Ontology in increasing order of their information content. It can be observed that the information content of a term is mostly dependent on its density rather than its depth in the ontology (The Gene Ontology, May 2007 release).

Term		Information content	Depth
id	Name		
GO:0008150	Biological_process	0.000	0
GO:0009987	Cellular process	0.042	1
GO:0008152	Metabolic process	0.099	1
GO:0044237	Cellular metabolic process	0.105	2
GO:0044238	Primary metabolic process	0.128	2
GO:0032502	Developmental process	0.143	1
GO:0007275	Multicellular organismal development	0.177	2
GO:0044248	Cellular catabolic process	0.285	3
GO:0015979	Photosynthesis	0.623	2
GO:0040011	Locomotion	0.700	1
GO:0008218	Bioluminescence	1.000	3

(b) Lin [40]:

Lin's method is another node-based technique that also considers the information content of the terms in order to derive their similarity:

$$\text{L-sim}_{norm}(t_1,t_2)=\frac{2\cdot\max_{t\in S(t1,t2)}[IC(t)]}{IC(t_1)+IC(t_2)} \tag{6}$$

(c) Jiang and Conrath [38]:

A method that combines node-based and edge-based approaches, by also considering network densities, node depth and types of connection along with the information content of the terms t_1 and t_2 and their minimal subsumer. The similarity measure is given by:

$$\text{JC-sim}_{norm}(t_1,t_2)=1-\frac{IC(t_1)+IC(t_2)-2\cdot IC(\text{min\_subsumer}(t_1,t_2))}{2\cdot\log(n_r)} \tag{7}$$

(d) Cao [41]:

A node-based method by which the similarity between two terms t_1 and t_2 is computed as the sum of the information content of all their common ancestors versus the sum of the information content for all the ancestors of each node:

$$\text{C-sim}_{norm}(t_1,t_2)=\frac{\sum_{t\in S(t_1,t_2)}IC(t)}{\sum_{t\in S(t_1,t_2)}IC(t)+\sum_{t'\notin S(t_1,t_2)}IC(t)} \tag{8}$$

where $t\notin S(t_1,t_2)$ is considered equivalent to $t\in((S(t_1)\cup S(t_2))-S(t_1,t_2))$.

(e) OldZZL [42]:

An edge-based method that computes the maximum common path among the longest paths from each term t_1, t_2 to the root:

$$\text{OZZL-sim}_{norm}(t_1,t_2)=1-\left(\frac{1}{2^{l_{common}}}-\frac{1}{2^{l_1+1}}-\frac{1}{2^{l_2+1}}\right) \tag{9}$$

where:

$l_1=\max[length(paths(t_1))]$,
$l_2=\max[length(paths(t_2))]$,
$l_{common}=\max[length(T_1\cap T_2)],\ T_1\in\max\_paths(t_1)$ and $T_2\in\max\_paths(t_2)$,

$paths(t_1)$: the set of paths from t_1 to the root,
$paths(t_2)$: the set of paths from t_2 to the root,
$\max\_paths(t_1)\subseteq paths(t_1)$ and $length(\max\_paths(t_1))=l_1$,
$\max\_paths(t_2)\subseteq paths(t_2)$ and $length(\max\_paths(t_2))=l_2$,
l_{common}: the maximum sub-path that is common between some path of t_1 maximum paths ($\max\_paths(t_1)$) and some path of t_2 maximum paths ($\max\_paths(t_2)$).

(f) ZZL [43]:

An improvement of the oldZZL method by which the similarity between two terms t_1 and t_2 takes into consideration the common path from the terms' minimal subsumer *ms* to the root:

$$\text{ZZL-sim}_{norm}(t_1,t_2)=1-\left(\frac{1}{2^{l_{ms}}}-\frac{1}{2^{l_1+1}}-\frac{1}{2^{l_2+1}}\right) \tag{10}$$

where in this case:

$$l_1 = \max[length(paths'(t_1))],$$
$$l_2 = \max[length(paths'(t_2))],$$
$$l_{ms} = length(path(ms)),$$

$$paths'(t_1) = paths(t_1, ms) \cup path(ms),$$
$$paths'(t_2) = paths(t_2, ms) \cup path(ms),$$

$$\begin{aligned} length(path(ms)) &= \max[length(paths'(t_1))] \\ &\quad - \max[length(paths(t_1, ms))] \\ &= \max[length(paths'(t_2))] \\ &\quad - \max[length(paths(t_2, ms))]. \end{aligned}$$

(g) Leacock and Chodorow [44]:

An edge-based method that considers the length of the minimum path between two terms and the maximum depth of the ontology. The similarity measure is:

$$\text{LC-sim}_{norm}(t_1, t_2) = \begin{cases} 1, & length(\min\_path(t_1, t_2)) = 0 \\ \frac{-\log \frac{length(\min\_path(t_1, t_2)) + 0.1}{2D}}{-\log \frac{1}{2D}}, & length(\min\_path(t_1, t_2)) \neq 0 \end{cases} \quad (11)$$

where

$\min\_path(t_1, t_2)$: the minimum path between t_1 and t_2,
D: the maximum depth of the ontology, i.e. the length from the farthest leaf to the root.

(h) Wu and Palmer [45]:

Semantic similarity is computed as follows:

$$\text{WP-sim}(t_1, t_2) = \frac{2N_3}{N_1 + N_2 + 2N_3} \quad (12)$$

where:

$$N_1 = length(\min\_path(t_1, ms)),$$
$$N_2 = length(\min\_path(t_2, ms)),$$
$$N_3 = length(\max\_path(ms, root)),$$

and ms is the minimal subsumer.

Finally, SoFoCles also implements the GraSM [46] versions of Resnik, Lin and Jiang–Conrath techniques, using the common disjunctive ancestors instead of the minimal subsumer.

4. Feature filtering using GO semantic similarities

4.1. The SoFoCles algorithm

Using the semantic similarity measure it is now possible to define an algorithm that can combine classic feature selecting techniques and the biological knowledge from the Gene Ontology in order to derive new feature sets that can better describe the microarray dataset. A flowchart of the SoFoCles algorithm, which comprises five steps, is depicted in Fig. 3.

In **Step 1** genes are ranked with respect to their discriminative ability, by the use of legacy feature selection algorithms, such as chi-square, information gain, or relief-F. In this stage, genes with the most informative values with respect to the specified classification problem qualify. Specifically, when applying the information gain method, the measured metric by which the genes are ranked

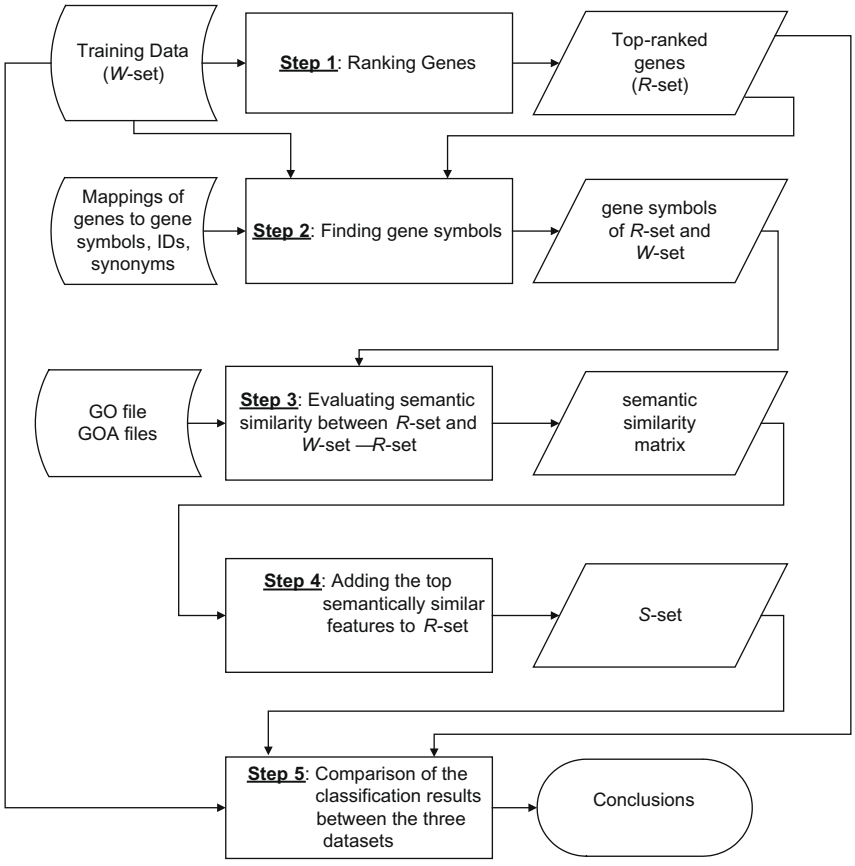

Fig. 3. Semantic feature filtering algorithm flowchart

is a correlation metric, whereas in the cases of relief-F and chi-square techniques the genes are ranked according to a distance metric. Correlation metrics in feature filtering measure an attribute's ability to correctly predict the value of a class. In particular, the information gain method tries to measure the reduction achieved in the entropy value of a class given that the value of one feature is known. Respectively, distance metrics measure an attribute's ability to discriminate among different classes. To this end, the chi-square method uses recursively the chi-square statistic, while relief-F tries to find a good estimate of the probability *RF(t)* to assign as the measure for each feature *t*. In the interest of brevity, Table 3 presents the metrics used for ranking a gene t_i given the problem classes c_i when using each one of these commonly known methods, without their mathematical justification.

In this step of the SoFoCles algorithm, values are calculated across all instances of the dataset in order to produce a ranking. The output is a new feature subset (*R*-set) that contains the most informative genes, which are also known as *marker genes*.

In **Step 2** of the algorithm, the gene symbols, identities, names or synonyms of the feature set genes are identified through mappings that transform the feature set genes to either of the above mentioned categories. In most microarray datasets, features are named after an inner microarray nomenclature system (e.g. the Affymetrix nomenclature). Thus, a preprocessing stage is required in order to translate the features to name categories that are recognizable forms by the Gene Ontology Annotation databases.

Step 3 of the algorithm covers the tasks of relating GO terms to genes and selecting the participating evidence codes as well as the aspects of the Gene Ontology that are to be used. The previously mentioned GOA system is used in order to discover the GO terms that are related to the feature set genes, by also setting the accepted evidence codes that must support the terms, as well as the desirable aspects.

After mapping each gene of the *W*-set to related GO terms, the algorithm computes semantic similarities in Gene Ontology between every pair of the two induced feature groups; (a) the refined group (*R*-set), that comprises the highly ranked genes, and (b) the rest of the genes (*W*-set – *R*-set) that have been temporarily discarded in Step 1 of the algorithm. Since a single gene may be represented by several GO terms in Gene Ontology, the semantic similarity between two genes that are represented by a series of GO terms has to be defined. This algorithm is described in Section 4.2.

Step 4 involves the semantic similarity based feature selection procedure. The temporarily discarded genes (*W*-set – *R*-set) are compared against each gene of the *R*-set. The genes that exhibit the highest similarity are selected to enrich the *R*-set, thus creating the semantically aware *S*-set. Eventually, *S*-set contains both genes/features filtered with classic feature selection methods as well as others that are semantically related to the former. The selection procedure is based on a two-level thresholding phase. In the first level, the top semantically relevant genes for each gene of the *R*-set qualify, creating a subset of candidate genes. In the second level, this subset is ranked and the top genes are finally selected to enrich the *R*-set. In this way, knowledge emerging from both the statistical similarity inside the dataset, and the semantic similarity from Gene Ontology is combined in order to produce the problem-specific, semantically aware dataset (*S*-set).

As a final step (**Step 5**), classification techniques are applied on the dataset, in order to validate the methodology. Various classifier models are trained and their parameters are estimated using the different feature sets induced in the previous steps. Different models are built using the *W*-set, the *R*-set and the *S*-set. The classification accuracy is then measured using the leave-one-out cross-validation technique in order to compare the models.

Table 3
Legacy feature filtering measures.

Description	Formula
Information gain	$IG(t_k,c_i)=\sum_{c\in\{c_i,\bar{c}_i\}}\sum_{t\in\{t_k,\bar{t}_k\}}P(t,c)\log_2\frac{P(t,c)}{P(t)P(c)}$
Chi-square	$CHI(t_k,c_i)=\frac{N[P(t_k,c_i)\cdot P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i)\cdot P(\bar{t}_k,c_i)]^2}{P(t_k)\cdot P(\bar{t}_k)\cdot P(c_i)\cdot P(\bar{c}_i)}$
Relief-F	$RF(t)=P(\text{different value of } t \mid \text{different class})$ $-P(\text{different value of } t \mid \text{same class})$

4.2. Semantic similarity at gene level

As mentioned above, in Step 3 of the SoFoCles algorithm we need to calculate the semantic similarity between two genes or gene products that may be represented by a series of GO terms.

Suppose that gene *A* is related with N_A GO terms and gene *B* is related with N_B GO terms. First, semantic similarities between all possible pairs of GO terms that are related with genes *A* and *B* are computed (sum of $N_A \cdot N_B$ combinations). Then a semantic similarity matrix (*SSM*) with dimensions $N_A \times N_B$ is created, whose elements $ss_{i,j}$ are the computed pairwise semantic similarities:

$$SSM=\underbrace{\left.\begin{bmatrix} ss_{1,1} & ss_{1,2} & ss_{1,3} & \dots & ss_{1,N_B} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ ss_{N_A,1} & ss_{N_A,2} & ss_{N_A,3} & \dots & ss_{N_A,N_B} \end{bmatrix}\right\}gene\ A}_{gene\ B} \tag{13}$$

The semantic similarity between the two genes can then be computed in three different ways.

MAX: The MAX approach involves the computation of the maximum similarity between any GO terms that are related to the genes. Formally, the semantic similarity between two genes using the MAX approach can be calculated as:

$$SSG_{MAX}=\max_{i,j}\{ss_{i,j}\},\quad (i,j)\in\{N_A\times N_B\} \tag{14}$$

Though, since a gene product usually encompasses all roles that its annotated GO terms describe, semantic similarity can be computed with respect to all annotated GO terms using the AVG process.

AVG: By this approach, semantic similarity is computed by finding the average similarity between any GO terms that refer to the genes [35]. Formally, semantic similarity between two genes with the AVG approach can be calculated as:

$$SSG_{AVG}=\frac{\sum_{i=1}^{N_A}\sum_{j=1}^{N_B}ss_{i,j}}{N_A\cdot N_B} \tag{15}$$

AVG_MAX: Finally, another way of computing semantic similarity between two genes that combines the MAX and AVG approaches is AVG_MAX. For each GO term that relates to a gene, its maximum similarity with respect to all terms of the other gene is calculated, a process that is repeated for both genes. As a result, two matrices $1\times N_A$ and $1\times N_B$ are formed, containing the similarities of gene *A* with respect to gene *B*, and the similarities of gene *B* with respect to gene *A*, respectively [46]. We denote as $sim(A,B)$ the matrix that contains the similarities of gene *A* with respect to gene *B*, while the matrix containing the similarities of gene *B* with respect to gene *A* is denoted as $sim(B,A)$.

The two matrices can be rewritten as:

$$sim(A,B)=\left[\max_{j\in\{1,2,\dots,N_B\}}\{ss_{1,j}\}\quad\max_{j\in\{1,2,\dots,N_B\}}\{ss_{2,j}\}\quad\cdots\quad\max_{j\in\{1,2,\dots,N_B\}}\{ss_{N_A,j}\}\right] \tag{16}$$

$$sim(B,A) = \left[\max_{i\in\{1,2,\ldots,N_A\}} \{ss_{i,1}\} \quad \max_{i\in\{1,2,\ldots,N_A\}} \{ss_{i,2}\} \quad \cdots \quad \max_{i\in\{1,2,\ldots,N_A\}} \{ss_{i,N_B}\} \right] \tag{17}$$

Finally, average similarities of gene A with respect to gene B and gene B with respect to gene A are computed:

$$\overline{sim(A,B)} = \frac{\sum_{i=1}^{N_A} \max_{j\in\{1,2,\ldots,N_B\}} \{ss_{i,j}\}}{N_A} \tag{18}$$

$$\overline{sim(B,A)} = \frac{\sum_{j=1}^{N_B} \max_{i\in\{1,2,\ldots,N_A\}} \{ss_{i,j}\}}{N_B} \tag{19}$$

The semantic similarity between the two genes according to the AVG_MAX approach can now be computed as the mean of the above average similarities:

$$SSG_{AVG\_MAX} = \frac{\overline{sim(A,B)} + \overline{sim(B,A)}}{2} \tag{20}$$

5. The SoFoCles platform

SoFoCles has been implemented as a Java platform that comprises four modules: (a) the user interface, (b) the data preprocessing module, (c) the semantic similarity unit and (d) the new feature set derivation subsystem. Fig. 4 depicts the general architecture of the platform and its environment.

The software uses and produces datafiles that are compatible with the WEKA data mining suite [47]. WEKA was selected as a data mining tool, because it provides a large variety of feature filtering and data mining algorithms, and its open source code facilitates its interconnection with SoFoCles.

Implementing Step 1 of the algorithm, WEKA is used to derive an initially filtered dataset (R-set) from the original dataset (W-set). Then, through the user interface a researcher can upload the two datasets to SoFoCles, as well as any version of the Gene Ontology.

The same module is also used for setting the initial parameters for the two processes of data preprocessing (Step 2) and the discovery of semantic similarities (Step 3).

Microarray attribute names are transformed to gene symbols using the proper datafiles provided by the microarray manufacturer (e.g. Affymetrix). With respect to semantic similarity measurement, SoFoCles supports the setup of different experiments by exploiting the Gene Ontology in different ways. To this end, one or more aspects of the Gene Ontology and any subset of the evidence codes for GO annotations can be included for the semantic similarity computation task. Finally, a semantic similarity method can be selected from the SoFoCles repository along with its parameters (thresholds and gene-level semantic similarity method).

In the data preprocessing unit, gene symbols that correspond to the datasets are identified using the mapping files. The semantic similarity module then assumes the task of finding semantic similarities between all pairs of genes from the R-set and the (W-set – R-set) (Step 4). To accomplish this, a repository of semantic similarity methods has been developed, which contains the online collection of algorithms described in Section 3.

The outcome is the similarity matrix, as defined in Eq. (13), which is used by the feature set derivation module to detect the most semantically similar genes, and then construct the S-sets. The classification experiments are then performed in Step 5.

From a functional point of view, SoFoCles offers various services, which are briefly described below:

- *Data entry*. SoFoCles can receive input from multiple data sources, such as microarray datafiles in WEKA format, gene mapping files from the R package [48], Gene Ontology Annotation Files and the Gene Ontology itself. Furthermore, data can also be entered in a simple text format, thus avoiding the use of external packages.
- *Parameter setting environment*. SoFoCles enables experimenting with different scenarios by providing a flexible parameter setting environment. The user can select the evidence codes that will participate in the process of GO term identification for each gene and narrow the search within specific Gene Ontology terms, by selecting the search fields to be used. Such fields can be a gene's identity, symbol, name or synonym. Furthermore,

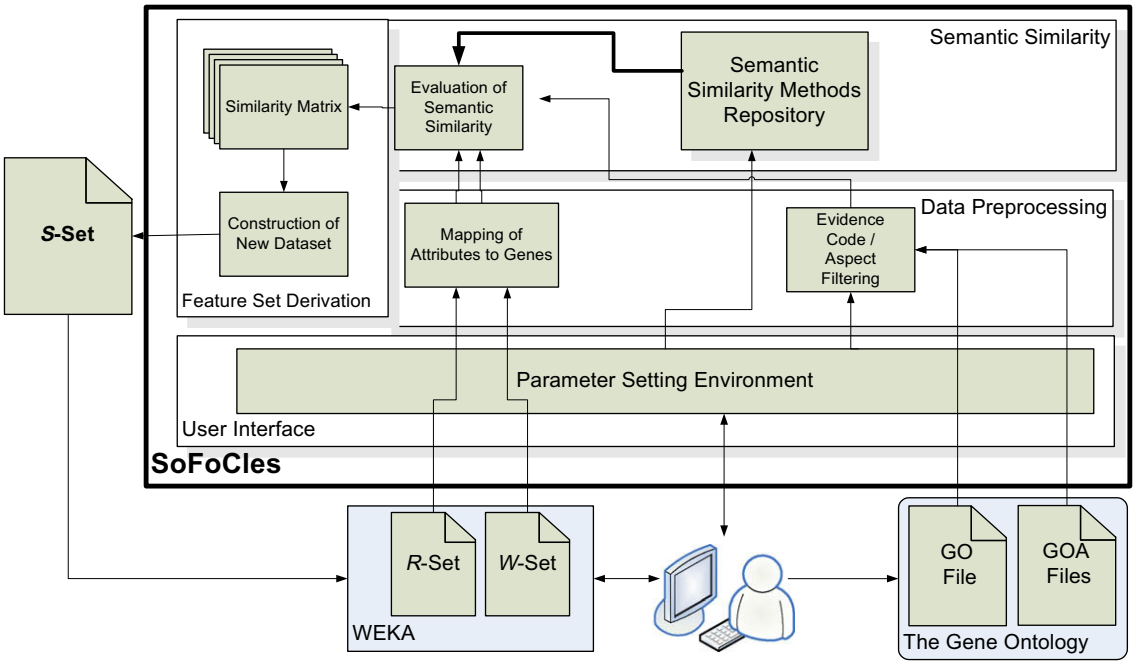

Fig. 4. General architecture of the SoFoCles platform.

the user can choose between an exact or partial match and exclude ontology aspects that are not considered important for the classification problem. A similarity threshold can be set, above which the semantically similar to the *R*-set genes qualify. Finally, the number of the derived genes can also be set manually. Fig. 5 depicts the data entry and parameter setting interface of SoFoCles.

- *Dataflow information*. While experimenting with SoFoCles, the user is directed with proper messages in order to form complete and sound queries.
- *Dataset production*. SoFoCles is able to output datasets suitable for processing with the WEKA environment that contain both statistically and semantically derived features. These datasets can be validated using various classification techniques.

6. Experiments

6.1. Experimental process

The SoFoCles platform was evaluated using two cases of microarray data. In both cases the general notion of the conducted experiments was to compare classification accuracy in each of the three resulting datasets: (a) the whole dataset *W*-set, (b) the refined dataset *R*-set (obtained by legacy feature selection techniques) and (c) the enriched dataset *S*-set generated by SoFoCles.

The main goal was the increase of classification accuracy. Apart from that, the improvement of other measures for the evaluation of the results such as TP rate, FP rate, precision and F-Measure, which refer more to each class label separately, were also considered.

The Support Vector Machines (SVM) algorithm [47] was used for classification in the last step of the process. SVM is computationally efficient, robust in high dimensions, which is the case in microarray problems, and it is based on sound theoretical foundations. Furthermore, no strong hypothesis is needed on the data generation process and the objects are classified with large confidence thus avoiding overfitting. In order to empower this decision, preliminary tests were conducted using the legacy feature selection algorithms, in which SVM outperformed other classification algorithms such as decision tables, rule learning algorithms (JRip and PART), decision trees (J48), instance based methods (IBk, K^*), neural networks (Perceptron), Bayesian and radial basis function networks. The datasets were evaluated using leave-one-out cross-validation.

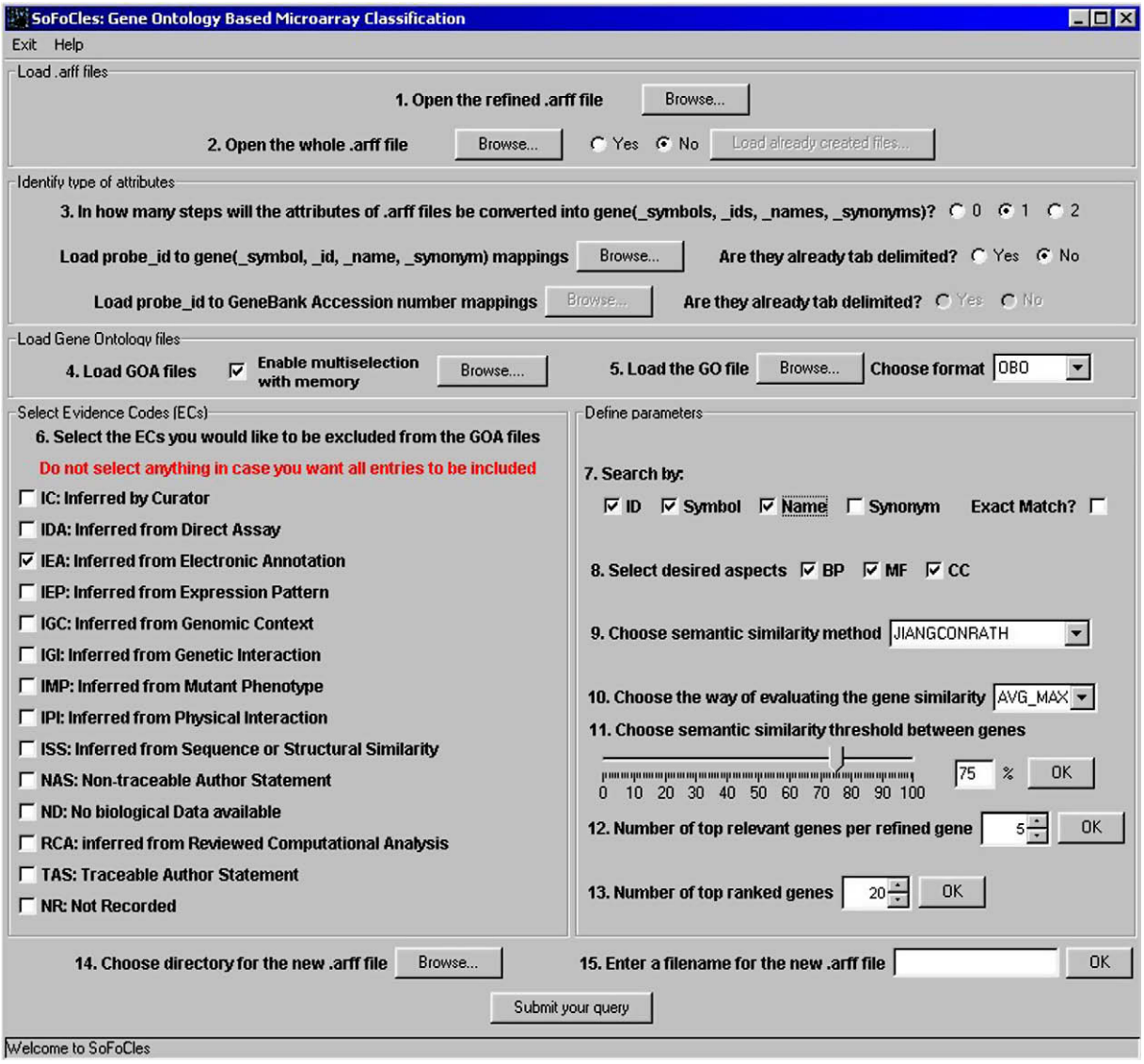

Fig. 5. SoFoCles data entry and parameter setting interface.

The two gene expression datasets that were used in our evaluation experiments are described in Section 6.2. The first dataset, D_S, contains information on the effects of cigarette smoke on the human epithelial cell transcriptome and the other, D_{CNV}, examines the central nervous system embryonal tumor outcome based on gene expression.

The GO release of May 2007 was used and searching was limited to the criteria of ID, symbol and synonym partial matching. Only GO terms from the BP aspect were taken into consideration excluding those supported by the IEA evidence code as this type of annotation is not validated by a curator.

For annotation purposes, three GOA files were used:

(a) Human (gene_association.goa_human),
(b) Rat (gene_association.rat) and
(c) Mouse (gene_association.mgi).

The decision to use the last two files was taken because these two organisms are highly similar to Human.

In addition, the minimum semantic similarity threshold was set to 50%, while five was selected as the maximum number of top semantically similar genes for each gene of the *R*-set and the maximum number of finally chosen genes among the top semantically similar ones was set to 50% of the number of features of the *R*-set, i.e. 20 genes for the D_S dataset and 10 genes for the D_{CNV} dataset.

Lastly, five semantic similarity methods (Resnik, Lin, Jiang&Conrath, Cao, oldZZL) and three ways of evaluating the semantic similarity between genes (AVG, MAX, AVG_MAX) were used for the experimentation, thus leading to the creation of $5 \times 3 = 15$ sets of results.

After presenting the parameters used, a short discussion on the nature of the tested data is essential. First, the classification accuracy of the whole dataset (*W*-set) was estimated as a reference value for the comparison of the other datasets' performance.

Furthermore, a refined set (*R*-set) was created from *W*-set after the application of a legacy feature selection method. In order to achieve fair comparison, when creating the *R*-set preliminary experiments were conducted and the threshold for which the legacy methods yielded the optimal results was selected. Based on this *R*-set, 15 new *S*-sets were produced by SoFoCles using different semantic similarity techniques. The *S*-sets were enriched with a number of genes with respect to the *R*-set. Thus, for a more complete attempt of evaluating the strength of SoFoCles, a new *R*-set, comprised of as much features as the *S*-sets had in average, was created using the legacy feature selection algorithm. For completeness purposes, two additional *R*-sets were also created; one containing only randomly extracted features out of the initial *W*-set and another containing the features of the initial *R*-set plus some randomly extracted features from the *W*-set. The number of the randomly extracted features was selected so that these *R*-sets have the same number of features as the average number of features of the *S*-sets.

6.2. Experimental data

6.2.1. D_S dataset

This dataset was a collection of 74 samples profiling the gene expression of the human airway epithelial cells across a broad spectrum of individuals in order to discover the reversible and irreversible effects of cigarette smoking [49]. These profiles were obtained by hybridization on the Affymetrix HG-U133A Genechip containing probes for 22,215 genes. Thus, the dataset comprises 22,215 features plus one class divided into three labels: current-smoker, never-smoked and former-smoker. Out of the 74 samples 33 belong to the "current-smoker" label, 23 to the "never-smoked" label and the remaining to the "former-smoker" label.

The features which represent GenBank accession numbers were transformed into gene symbols by appropriate mappings in order to be processed by the GOA files.

Relief-F was used as the feature selection method for the creation of the *R*-set from the *W*-set. In more detail, the top 40 features, as selected by the relief-F algorithm, were isolated and produced the *R*-set. Based on these 40 features their top semantically similar genes were mined according to one of the semantic similarity methods implemented in SoFoCles and the *S*-sets were produced.

6.2.2. D_{CNV} dataset

The second group of data was a set of 60 samples examining the mechanism of central nervous system embryonal tumor based on gene expression [50]. The RNA amounts were hybridized to HuGeneFL arrays which, in turn, were scanned on Affymetrix scanners. Finally, the expression values for each gene were calculated using Genechip software. The dataset comprises 7129 genes plus one binary class. Out of the 60 samples 39 belong to "medulloblastoma survivors" and 21 to "treatment failures".

The features which represent probe IDs were transformed into gene symbols by appropriate mappings in order to be processed by the GOA files. The feature selection method that was used was the Information Gain. More specifically, the top 20 features coming out of the Information Gain algorithm constituted the *R*-set.

6.3. Results

6.3.1. D_S dataset

Accuracies for the performed classification experiments are depicted in Fig. 6. The bar chart is based on Table I (Supplementary information).

As it can be observed in the chart the two *R*-sets (those with 40 and 74 features) outperform the *W*-set roughly by 14%. This can be justified by the fact that many of the genes, which are described in [49] as cancer-related, are found in the *R*-sets. More explicitly, the *R*-sets include genes that code for xenobiotic functions, such as the anti-oxidants GPX2 and ALDH3A1, several putative tumor suppressor genes, such as SLIT1, SLIT2 and TU3A, the oncogene pirin (PIR), a transcription factor involved in the induction of oxygen regulated genes (EPAS), as well as the genes CLDN10, MMP10, LOC92689, ME1 and MT1F, which were also considered in the analysis in [49].

As expected, the sets with randomly selected genes yielded poorer results than the ones refined with legacy feature selection methods.

Fig. 7 depicts the percentile accuracy improvement of each method and gene semantic similarity approach with respect to the *R*-set for the D_S dataset. Out of the 15 sets produced by SoFoCles, all five that used the AVG metric for evaluating semantic similarity exhibited invariably worse performance. Thus, it can be inferred that this method produced no new knowledge and it was not suitable for the specific dataset.

Of the remaining 10 *S*-sets, six outperformed the *R*-set (of 40 attributes). In short, the combinations of Jiang&Conrath and OldZZL with MAX or AVG_MAX and the pairs Resnik/MAX and Cao/AVG_MAX produced better results.

Thus, Jiang&Conrath and oldZZL semantic methods seem to perform adequately for this specific dataset and derive the necessary information, particularly when MAX is chosen as the method for *evaluating semantic similarity*.

It must be noted that among the genes of the above *S*-sets that have semantically enriched the *R*-sets, there can be found genes that are believed to contribute to cancer development according to [49]. For instance, CX3CL1, a potential tumor suppressor gene, which is absent in the *R*-sets, is found in the Cao.AVG_MAX, Jiang-

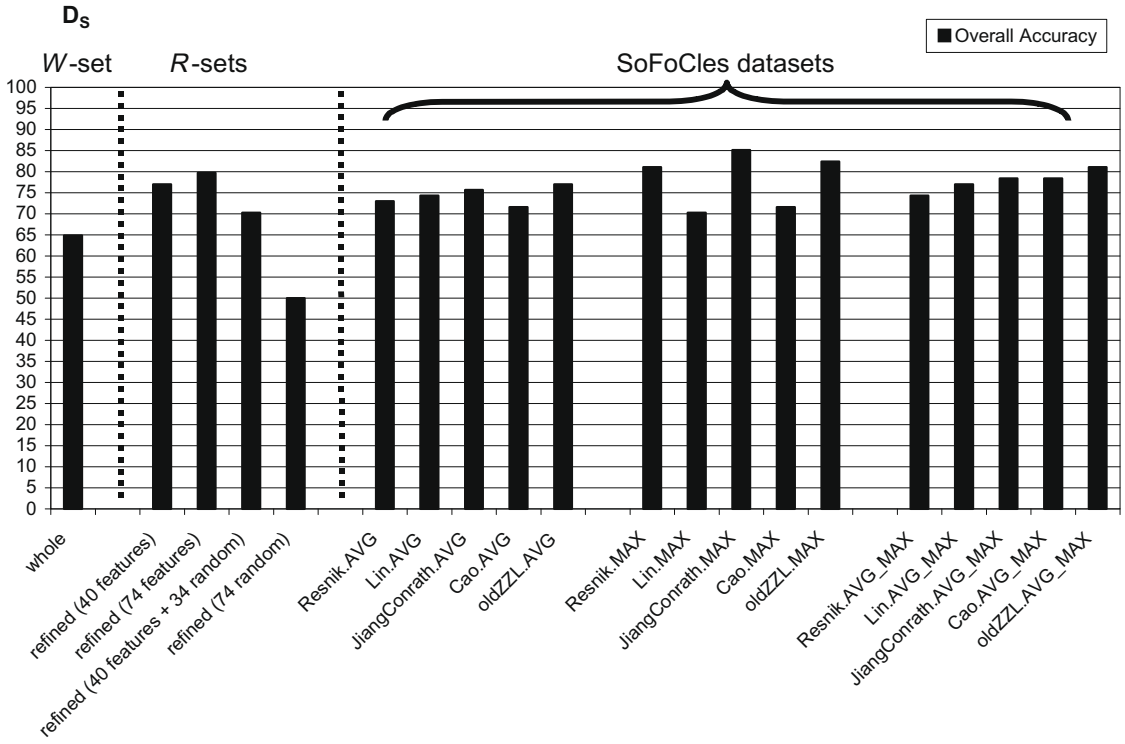

Fig. 6. Classification accuracy for the D_S dataset. Accuracies were measured for the whole dataset (*W*-set), four different datasets refined with legacy statistical methods (*R*-sets), and the SoFoCles datasets (*S*-sets) that are semantically enhanced using five semantic similarity methods and three different approaches for computing semantic similarity between genes.

Conrath.AVG_MAX and oldZZL.AVG_MAX sets. Moreover, the presence of tumor suppressor SLIT1 gene is reinforced in the Resnik.MAX set. Thus, if the legacy feature selection algorithms were solely used, the inclusion of the above genes into the feature set would also require the inclusion of more irrelevant genes, thus decreasing the classification accuracy. Specifically, the inclusion of the above genes using the legacy method would require the use of 13,547 attributes in total, resulting to a classification accuracy of 77.027%, a very low rate compared to the 85.14% of SoFoCles. The SoFoCles methodology can override this drawback using the semantic similarity notion.

6.3.2. D_{CNV} dataset

Classification accuracy measurements for the experiments with the D_{CNV} dataset are given in Fig. 8. The bar chart is based on the Table II (Supplementary information).

In this dataset, the *R*-sets (except the totally random one) exceed the classification accuracy of the *W*-set by as much as 13%. One profound reason is that discriminatory genes, as described in [50], are included in the *R*-sets. More specifically, the aforementioned sets contain features that, according to the analysis in [50], are responsible for either the expression of Class1 (tumors with high ribosomal content), Class0 (tumors with low ribosomal

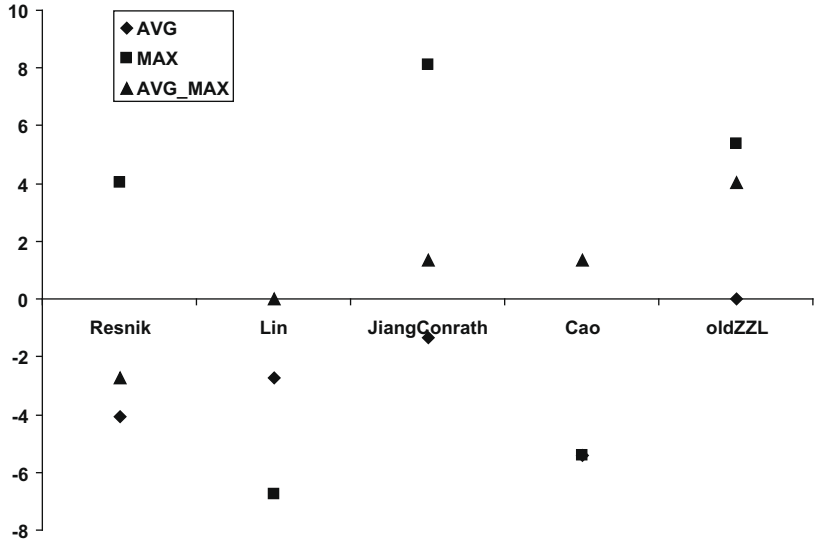

Fig. 7. Accuracy improvement (%) with respect to *R*-set per semantic similarity method and per Gene Similarity Approaches (AVG, MAX, AVG_MAX) for the D_S dataset.

content) or both. Such genes are the markers of survival in the study GPS2, AMT, ACADVL, AP3B2 (Beta-NAP), KIAA0220, NHLH1, and C5orf18 (human polyposis locus).

Besides, the *R*-set with mixed filtered and random features performs similarly with the other two *R*-sets, although the one with all of its features being random normally gave bad results.

Fig. 9 depicts the percentile accuracy improvement of each method and gene semantic similarity approach with respect to the *R*-set for the D_{CNV} dataset. For this experiment nine out of the 15 *S*-sets outperformed the *R*-sets and two others showed similar results. It seems that the biological knowledge entered by the tool really helps in deciphering useful information from *W*-set. Indeed, when the AVG_MAX metric was used, the classification accuracy increased by almost 10% compared to that of the *R*-sets. Therefore, the AVG_MAX parameter appears to have a critical role in the enrichment of biological pathways related to the problem.

As in the previous case, the Jiang&Conrath and oldZZL semantic methods appear to apply effectively on this dataset as well, especially in combination with AVG_MAX value.

However, three out of five *S*-sets with the *way of evaluating semantic similarity* parameter set to AVG rendered worse results as it happens in the set having the parameters *semantic method* and *way of evaluating gene semantic similarity* set to Resnik and AVG_MAX respectively. Again, when the *way of evaluating gene semantic similarity* equals AVG no improvement as far as the accuracy is concerned is achieved.

Last but not least, the four *S*-sets reaching levels of accuracy near 90% (Cao.AVG_MAX, JiangConrath.AVG_MAX, Lin.AVG_MAX and oldZZL.AVG_MAX) contain genes which are highly semantically similar to problem related ones such as DSCR1L1, FMR1 showing high similarity with NHLH1 and ACTR1A being similar to AP3B2. Specifically, DSCR1L1, a gene considered as critical for the Down syndrome (Ca interaction proteins), and FMR1, which may play a role in the development of synaptic connections between nerve cells in the brain (also known as "fragile X mental retardation protein") are found to be semantically, and also functionally, similar with NHLH1, which may serve as DNA-binding protein and may be involved in the control of cell-type determination, possibly within the developing nervous system. Likewise, the final *S*-set contains the gene ACTR1A, which is involved in a diverse array of cellular functions, including ER-to-Golgi transport, the centripetal movement of lysosomes and endosomes, spindle formation, chromosome movement, nuclear positioning, and axonogenesis, due to its semantic similarity with AP3B2, a gene that is thought to serve neuron-specific functions such as neurotransmitter release. As in the previous case, in order to achieve the presence of these specific genes using the legacy statistical feature selection methods, the inclusion of more irrelevant to the problem genes would be required, thus decreasing the classification efficiency. Specifically, using the legacy method, the inclusion of the above genes would require the use of 3996 attributes in total, resulting to a classification accuracy of 75%, much poorer than the 91.67% that SoFoCles yields.

6.4. Discussion

We have measured the effect of semantically enhancing gene expression feature sets in order to improve microarray classification accuracy. To this end, we have used several semantic similarity methods and three approaches for computing semantic similarity between genes.

To summarize, better results are obtained when only the terms with the maximum similarity are taken into consideration at a time (MAX approach). More specifically, in the D_S dataset, three out of five *S*-sets with the MAX parameter perform more efficiently than the *R*-sets, as is the case in the D_{CNV} dataset, as well. Even more satisfactory results were produced when the AVG_MAX value was used for evaluating gene semantic similarity, especially in the D_{CNV} set. This can be attributed to the fact that genes (or gene products) have multiple biological roles. Thus, for each term annotated to a gene, only its similarity to the most similar one of the other genes is taken into account. Consequently, all the characteristics of the genes are considered during the evaluation of their semantic similarity. The AVG *way of evaluating gene semantic similarity* was found to be inadequate.

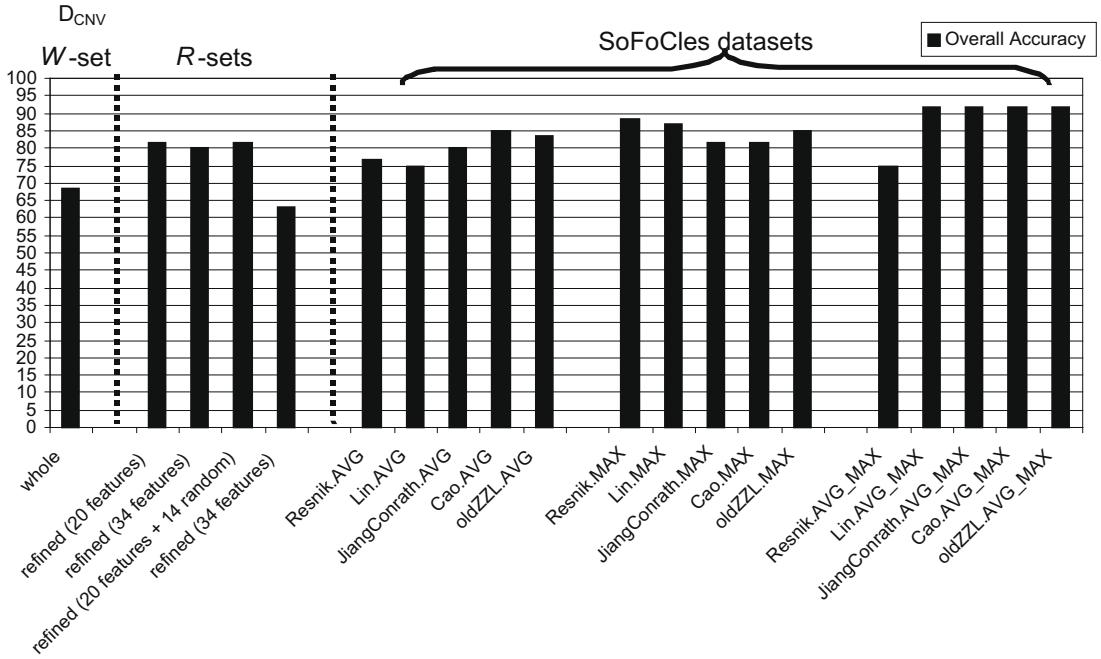

Fig. 8. Classification accuracy for the D_{CNV} dataset. Accuracies were measured for the whole dataset (*W*-set), four different datasets refined with legacy statistical methods (*R*-sets), and the SoFoCles datasets (*S*-sets) that are semantically enhanced using five semantic similarity methods and three different approaches for computing semantic similarity between genes.

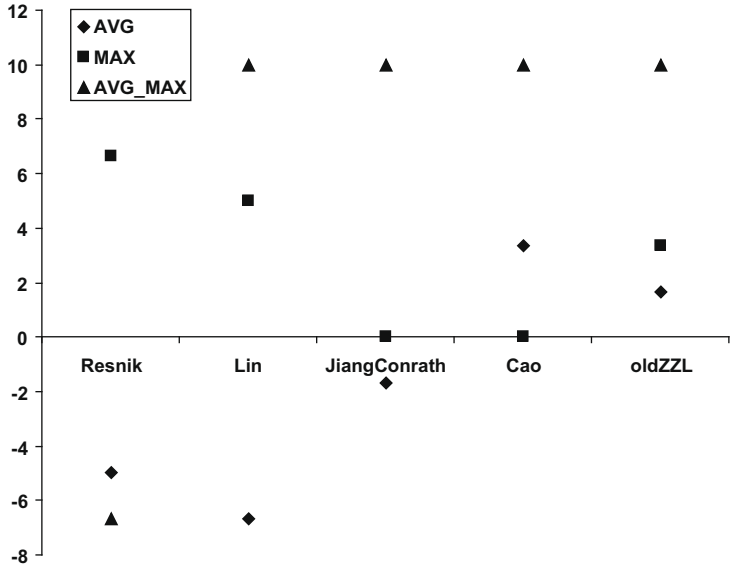

Fig. 9. Accuracy Improvement (%) with respect to *R*-set per semantic similarity method and per Gene Similarity Approaches (AVG, MAX, AVG_MAX) for the D_{CNV} dataset.

Focusing on the semantic methods used, the Jiang&Conrath and the oldZZL seem more appropriate at least for the two datasets used (D_S and D_{CNV}). In particular, these methods achieved optimum performance in both datasets when combined with MAX and AVG_MAX values of the *way of evaluating gene semantic similarity* parameter.

SoFoCles can be a valuable tool in microarray data analysis by contributing to the process of improving classification results in microarray data, as well as in providing them with more biological insight. SoFoCles is achieving that by reinforcing the discriminative power of marker genes not only by just adding highly semantically similar genes which are likely to participate in common biological pathways, but also by exploiting knowledge from one of the most respected biological ontologies, the Gene Ontology.

7. Conclusions and future work

SoFoCles is a methodology for marker gene selection in microarray experiments that combines semantic and statistical knowledge in a unified environment. The methodology has been implemented as a software platform which provides the user with an experimenting suite comprising a series of semantic similarity algorithms and an easy way of handling and parameterizing the Gene Ontology. SoFoCles reinforces already spotted biological process pathways by finding the most semantically similar genes of the *R*-sets. The user of the suite can experiment with different semantic similarity algorithms and fine-tuning parameters in order to achieve optimal performance. It is also worth noting that the interface environment of the tool is user-friendly, guiding the user step by step and presenting information messages, where necessary, in order to ensure its proper functionality.

Microarray classification experiments have demonstrated the ability of the proposed technique to improve classification accuracy, by enhancing the dataset with strong marker genes that might otherwise have been omitted. The performance of SoFoCles proved better not only compared to the *W*-set but to the *R*-sets, as well.

The ability of SoFoCles to reveal semantically similar marker genes is directly dependent on the quality of the Gene Ontology itself. Thus, further improvement or enrichment of the Gene Ontology may yield even better results with SoFoCles. At present, only a small fraction of the existing genes is annotated to GO terms. From this percentage, the majority of the annotations is based on electronic techniques, which contain high levels of unreliability especially when not supported by a GO curator. Although SoFoCles is able to exclude such annotations, an improvement in GO annotations quality will automatically enhance the quality of SoFoCles results.

Future work on SoFoCles aims at uncovering more of the "shadowy" pathways that are not revealed after the application of legacy feature selection methods. Such pathways may hide important information and have considerable discriminative power of uncovering the separate labels of the problem. One possible solution is the clustering of genes of the *R*-set. Clustering these genes into groups that are likely to participate in common pathways might take more paths into consideration and prevent excessive reinforcement of main paths as well. Moreover, statistical measures can be applied to the semantically similar genes retrieved in order to isolate representative genes of a group rather than all genes. In that way, genes having same functionalities with a referenced one and do not serve further in the amplification of a pathway will be ignored.

In conclusion, the first version of SoFoCles seems to reinforce the synergy between statistical and semantic methods for the improvement of classification results and the attribution of more biological insight on them.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2009.06.002.

References

[1] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7:673–9.
[2] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999;96:6745–50.
[3] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
[4] Bomprezzi R, Ringner M, Kim S, Bittner ML, Khan J, Chen Y, et al. Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. Hum Mol Genet 2003;12:2191–9.

[5] Wang Y, Makedon F, Ford J, Pearlman J. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics 2005;21(8):1530–7.
[6] Xiong F, Huang H, Ford J, Makedon F, Pearlman J. A new test system for stability measurement of marker gene selection in DNA microarray data analysis. In: Proceedings of the 10th Panhellenic conference on informatics; 2005. p. 437–47.
[7] Li J, Su H, Chen H. Identification of marker genes from high-dimensional microarray data for cancer classification. Knowledge discovery in bioinformatics. New York: Wiley; 2007.
[8] Lu Y, Han J. Cancer classification using gene expression data. Data Manag Bioinform 2003;28(4):243–78 [special issue].
[9] Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. Belmont, CA: Wadsworth International Group; 1984.
[10] Liu H, Setiono R. Chi2: feature selection and discretization of numeric attributes. In: Proceedings of the 7th international conference on tools with artificial intelligence; 1995. p. 388–91.
[11] Kononenko I. Estimating attributes: analysis and extensions of relief. In: Proceedings of the European conference on machine learning; 1994. p. 171–82.
[12] Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23(19):D2507–17.
[13] Jaeger J, Sengupta R, Ruzzo WL. Improved gene selection for classification of microarrays. Proc Pacific Symp Biocomput 2003;8:53–64.
[14] Camon E, Magrane M, Barrel D, Lee V, Dimmer E, Maslen J, et al. The Gene Ontology Annotation (GOA) database: sharing knowledge in UniProt with Gene Ontology. Nucl Acids Res 2004;32(1):D262–6.
[15] The UniProt Consortium. The universal protein resource (UniProt). Nucl Acids Res (Database issue) 2007;35:D193–7.
[16] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. New developments in the InterPro database. Nucl Acids Res 2007;35:D224–8.
[17] Jaiswal P, Ni J, Yap I, Ware D, Spooner W, Youens-Clark K, et al. Gramene: a bird's eye view of cereal genomes. Nucl Acids Res 2006;34:D717–23.
[18] Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, the members of the Mouse Genome Database Group. The mouse genome database (MGD): from genes to mice—a community resource for mouse biology. Nucl Acids Res 2005;33:D471–5.
[19] Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM, the FlyBase Consortium. FlyBase: genomes by the dozen. Nucl Acids Res 2007;35:D486–91.
[20] Chisholm RL, Gaudet P, Just EM, Pilcher KE, Fey P, Merchant SN, et al. DictyBase, the model organism database for *Dictyostelium discoideum*. Nucl Acids Res (Database issue) 2006;34:D423–7.
[21] Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, et al. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucl Acids Res 2003;31(1):224–8.
[22] Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005;21(18):3587–95.
[23] Goldstone RL, Son J. Similarity. In: Holyoak K, Morrison R, editors. Cambridge handbook of thinking and reasoning. Cambridge: Cambridge University Press; 2005. p. 13–36.
[24] Larkey LB, Markman AB. Processes of similarity judgment. Cogn Sci 2005;29(6):1061–76.
[25] Tversky A. Features of similarity. Psychol Rev 1977;84(4):327–52.
[26] Goldstone RL. The role of similarity in categorization: providing a groundwork. Cognition 1994;52:125–57.
[27] Rips LJ, Shoben EJ, Smith EE. Semantic distance and the verification of semantic relations. J Verbal Learn Verbal Behav 1973;12:1–20.
[28] Schwering A, Raubal M. Measuring semantic similarity between geospatial conceptual regions. In: Rodriguez A, Cruz I, Egenhofer M, Levashkin S, editors. GeoSpatial Semantics—1st international conference. Lecture notes in computer science 3799; 2005. p. 90–106.
[29] Bellazi R, Zupan B. Towards knowledge-based gene expression data mining. J Biomed Inform 2007;40(6):787–802.
[30] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 2000;25:25–9.
[31] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50.
[32] Qi J, Tang J. Gene Ontology driven feature selection from microarray gene expression data. In: Proceedings IEEE symposium on computational intelligence and bioinformatics and computational biology (CIBCB); 2006. p. 1–7.
[33] Qi J, Tang J. Integrating gene ontology into discriminative powers of genes for feature selection in microarray data. In: Proceedings ACM symposium on applied computing; 2007. p. 430–4.
[34] The Gene Ontology Consortium. Creating the Gene Ontology resource. design and implementation. Genome Res 2001;11:1425–33.
[35] Lord P, Stevens R, Brass A, Goble C. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003;19(10):1275–83.
[36] Schlicker A. A global approach to comparative genomics: comparison of functional annotation over the taxonomic tree. Master thesis. Saarbrücken, Germany: Center for Bioinformatics of Saarland University; 2005. p. 1–82.
[37] Bérard S, Tichit L, Herrmann C. ClusterInspector: a tool to visualize ontology-based relationships between biological entities. In: Proceedings Journees Ouvertes Biologie Informatique Mathematiques (JOBIM); 2005. p. 1–12.
[38] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of international conference on research in computational linguistics; 1997. p. 1–15.
[39] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on artificial intelligence; 1995. p. 448–53.
[40] Lin D. An information—theoretic definition of similarity. In: Proceedings of 15th international conference on machine learning; 1998. p. 296–304.
[41] Cao S, Qin L, He W, Zhong Y, Zhu Y, Li Y. Semantic search among heterogeneous biological databases based on Gene Ontology. Acta Biochim Biophys Sinica 2004;36(5):365–70.
[42] Zhu H, Zhong J, Li J, Yu Y. An approach for semantic search by matching RDF graphs. In: Proceedings of the 15th international Florida artificial intelligence research society conference; 2002. p. 450–4.
[43] Zhong J, Zhu H, Li J, Yu Y. Conceptual graph matching for semantic search. In: Proceedings of the 10th international conference on conceptual structures: integration and interfaces; 2002. p. 92–106.
[44] Patwardhan S, Banerjee S, Pedersen T. Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the 4th international conference on intelligent text processing and computational linguistics; 2003. p. 1–17.
[45] Wu Z, Palmer M. Verb semantics and lexical selection. In: Proceedings of the 32nd annual meeting of the associations for computational linguistics; 1994. p. 133–8.
[46] Couto F, Silva M, Coutinho P. Measuring semantic similarity between Gene Ontology terms. Data Knowledge Eng 2007;61(1):137–52.
[47] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufman; 2005.
[48] Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat 1996;5(3):299–314.
[49] Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, et al. Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc Natl Acad Sci USA 2004;101(27):10143–8.
[50] Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 2002;415(6870):436–42.