Int. J. Appl. Math. Comput. Sci., 2010, Vol. 20, No. 3, 555–570
DOI: 10.2478/v10006-010-0041-3

QUALITY IMPROVEMENT OF RULE–BASED GENE GROUP DESCRIPTIONS USING INFORMATION ABOUT GO TERMS IMPORTANCE OCCURRING IN PREMISES OF DETERMINED RULES

MAREK SIKORA [*,**], ALEKSANDRA GRUCA [*]

[*] Institute of Informatics
Silesian University of Technology, Akademicka 16, 44–100 Gliwice, Poland
e-mail: {marek.sikora,aleksandra.gruca}@polsl.pl

[**] Institute of Innovative Technologies EMAG, Leopolda 31, 40–189 Katowice, Poland

In this paper we present a method for evaluating the importance of GO terms which compose multi-attribute rules. The rules are generated for the purpose of biological interpretation of gene groups. Each multi-attribute rule is a combination of GO terms and, based on relationships among them, one can obtain a functional description of gene groups. We present a method which allows evaluating the influence of a given GO term on the quality of a rule and the quality of a whole set of rules. For each GO term, we compute how big its influence on the quality of generated set of rules and therefore the quality of the obtained description is. Based on the computed quality of GO terms, we propose a new algorithm of rule induction in order to obtain a more synthetic and more accurate description of gene groups than the description obtained by initially determined rules. The obtained GO terms ranking and newly obtained rules provide additional information about the biological function of genes that compose the analyzed group of genes.

Keywords: decision rules, importance of rules premises, measures of rules interestingness, gene ontology, descriptions of gene groups.

1. Introduction

In the last decade, DNA microarray chips have proved to be a powerful tool used in biological and medical laboratories in genome scale experiments (Baldi and Hatfield, 2002). The analysis of data obtained in a DNA microarray experiment usually consists of four main steps: data normalization, the identification of differentially expressed genes, the application of algorithms grouping (clustering) together genes with similar expression patterns, and the interpretation of biological functions of genes co-expressed together.

Biological interpretation of the obtained gene groups is a very important part of the whole experiment and this aspect of data analysis is often carried out by an expert in the field of experimental design, frequently manually (which is time consuming for large data sets). However, to support such analysis, specialized systems are designed to store, organize and extract information on genes, their functions and products. The most popular and widely used are gene ontology (GO) terms (Ashburner *et al*., 2000) that are sources of information about biological processes and genes involved in these processes. The GO database is organized into three disjoint directed-acyclic graphs (DAGs) describing the biological process (BP), the molecular function (MF) and the cellular component (CC). The dependences among GO terms are hierarchical—nodes close to the root describe general concepts and, as the DAG is traversed from the root into its leaves, the description is more and more specific. Figure 1 presents a part of the GO directed-acyclic graph structure.

The popularity of the GO database results in developing the number of GO processing tools based on the idea of annotating the analyzed group of genes with GO terms and then performing a statistical test to extract over- or underrepresented GO terms in the analyzed set of genes (Maere *et al*., 2005; Al-Shahrour *et al*., 2005; Khatri and Drăghici, 2005).

Recently, research on rules induction which combines gene expression data and biological information has been performed (Hvidsten *et al.*, 2003; Midelfart, 2005a; 2005b). In the paper by Hvidstein *et al.* (2003), conditio-

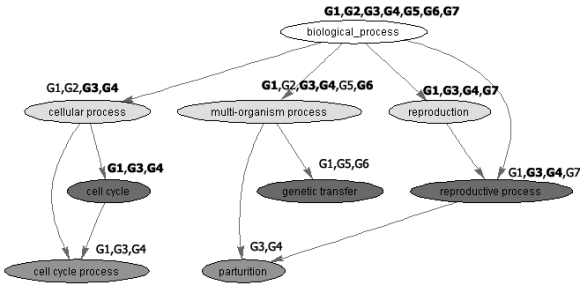

Fig. 1. Part of the gene ontology graph. Regular symbols denote gene annotations that were assigned to particular GO terms by curators. Gene assignments resulting from the hierarchy of the ontology graph are represented in bold.

nal rules of the form *"IF conjunction of conditions describing time series of gene expression profile THEN ontological term"* were proposed. The authors wanted to assign genes with specified expression profiles to a specific gene ontology term. Conclusions of rules with the same conditional parts are joined, and thus rules describing a group of genes with similar expression profiles are obtained. In rule conclusions, a set of gene ontology terms describing the group is included.

To discover the co-appearance of some ontology terms, algorithms of association rule induction have been applied so far. The method proposed by Carmona-Saez *et al.* (2006) combines expression data and biological information. In the paper Carmona-Saez *et al.* (2007), the Genecodis web-based tool for integrated analysis of annotations from different sources was introduced. The method uses the Apriori algorithm (Agrawal and Srikant, 1994) to discover sets of annotations that frequently co-occur in the analyzed group of genes. A similar tool that enables finding combinations of annotations in many different fields such as functional categories, gene regulation, sequence properties, evolution, conservation, etc., was presented by Hackenberg and Matthiesen (2008).

Rule induction techniques mentioned above have drawbacks which can make the obtained rules difficult or even impossible to interpret. Firstly, known rules induction methods do not take into consideration the hierarchy of GO terms, which may result in replacing a conjunction of attributes with one GO term being the lowest in the hierarchy. Secondly, all the methods mentioned above lead to the generation of a huge number of rules without providing more advanced (apart from the p-value and the rule coverage) methods of rules evaluation and selection.

In order to avoid the above inconveniences, a method of rule induction that considers the location of GO terms in the ontology graph was presented by Gruca and Sikora (2009). An interpretation of determined rules is as follows:

> ***If*** *a gene is described by a conjunction of gene ontology terms appearing in a rule premise,* (1)
> ***then*** *it belongs to a specific group of genes.*

The method of their induction guarantees that terms lying on the same path in the ontology graph do not appear in a rule premise.

Having induced a set of the rules that create a description of a gene group, one can be interested in evaluating the importance of each single GO term appearing in the rules. In this paper we present a method for evaluating the importance of GO terms that compose the obtained set of rules. Based on the induced rules set, we can determine which GO terms occurring in rule premises are characterized by the highest significance. We do not consider statistical significance but the influence of a given GO term on the quality of rules that include this term only.

The problem of evaluating the importance of conditions appearing in rules premises was considered by Greco *et al.* (2007), who applied indexes used in game theory for the evaluation of coalition quality. In conducted research, the Banzhaf index (Banzhaf, 1965) and a modified version of the approach presented by Greco *et al.* (2007) were applied to evaluate the importance of GO terms occurring in rule premises. An algorithm of rule induction that considers the obtained GO terms ranking is also presented in this paper. The algorithm determines rules which describe gene groups in even a more synthetic way.

The paper is organized as follows. In the next section, results obtained so far by the authors in the field of rule induction used for gene group description are presented briefly; especially methods of rules induction, their evaluation and filtration are described. In Section 3 a proposition of the assessment of the importance of GO terms appearing in the obtained rule premises is presented. Moreover, a new algorithm of rule induction which considers the obtained GO terms ranking is introduced. Section 4 contains results of performed experiments. Summary and directions of further research are given Section 5.

2. Description of gene groups by multiattribute logical rules

2.1. Data and rules representation. Consider a set of genes G and a set of descriptions of genes and gene products A. Formally, gene ontology is a directed acyclic graph $GO = (A, \leq)$, where A is a set of GO terms describing genes and its products and $\leq$ is a binary relation on A such that the genes described by GO term a_j are a subset of the genes described by GO term a_i, denoted by $a_j \leq a_i$, if and only if there is a path $(a_i, a_{i+1}, \ldots, a_{j-1}, a_j)$ such that $a_m \leq a_{m-1}$ for $m = i+1, i+2, \ldots, j-1, j$. The

relation $\leq$ is an order relation (i.e., it is reflexive, antisymmetric and transitive).

The root of the DAG is the largest element and we assume that the root is at a zero level in the ontology. Each node from the DAG is represented by a single GO term from the set A. Each level of the graph is defined in the following way: the i-th level of the graph is formed by all GO terms $a \in A$ for which there is a path $(root, a_1, \ldots, a_{i-1}, a_i)$ such that $a_1 \leq root$, $a_m \leq a_{m-1}$ for $m = 2, 3, \ldots, i-1$ and $a_i \leq a_{i-1}$. GO terms at higher levels (closer to the root) describe a more general function or process while terms at lower levels are more specific. Each annotation is an association between a gene and a GO term describing it. Thus, for simplicity, we can assume that each node of gene ontology is also annotated by genes from the set G. A gene can be annotated to zero or more nodes for each ontology, at any level within each ontology. All GO terms that exist in the DAG must follow the true path rule: "the pathway from a child term all the way up to its top-level parent(s) must always be true". A consequence of such an approach is that annotating a gene to a GO term implies annotation to all its parents via any path.

It stems from the above definitions that each gene annotated with a GO term $a_j \in A$ is also annotated with a GO term $a_i \in A$ such that $a_j \leq a_i$. Assuming that G_a is a set of genes annotated with a GO term corresponding to the node a, for each node a_i such that $a \leq a_i$, $G_a \subseteq G_{a_i}$ is satisfied. In other words, the higher level of a GO term, the more genes are annotated to that term.

To preserve the clarity of the ontology, annotation files that are available at the Gene Ontology Consortium website include only "original" annotations, that is, annotations that were assigned to a particular GO terms by curators. In Fig. 1, regular symbols of genes denote "original" genes assigned to a given GO term, whereas gene assignments resulting from a hierarchy of the ontology graph are represented in bold. Various datasets are obtained depending on whether relationships among terms are taken into consideration or not. In our research we assign to each node not only genes that were directly extracted from the gene ontology annotation database but also genes that are annotated to all descendant terms of that node. We call such a graph "GO-Inc" (Inclusive Analysis).

To summarize, there is a set G of genes, a set A of GO terms that create the GO-Inc ontology graph and n gene groups with similar expression profiles $\{G(1), G(2), \ldots, G(n)\}$. It is possible to create a decision table ***DT-Inc***$= (G, A \cup d)$, where for all $a \in A$, $a : G \rightarrow \{0, 1\}$, and $d(g) \in \{G(1), G(2), \ldots, G(n)\}$ for all $g \in G$. Thus, rows in the table ***DT-Inc*** contain descriptions of single genes belonging to the set G created by means of GO terms from A. The notation $a(g) = 1$ (in short, $a(g)$) denotes that a gene g is assigned to the term a in the GO-Inc graph. The value $a(g) = 0$ (in short, not $a(g)$) means that a gene g is not assigned to the term a. Each gene is also characterized by membership to a specific group of genes (value $d(g)$). An example of ***DT-Inc*** formed based on Fig. 1 is presented in Table 1.

Table 1. Example of the ***DT-Inc*** decision table formed based on Fig.1.

annot. /gene	bp	cp	mp	r	cc	gt	rp	ccp	p
$G1$	**1**	1	**1**	**1**	**1**	1	1	1	0
$G2$	**1**	1	1	0	0	0	0	0	0
$G3$	**1**	**1**	**1**	**1**	**1**	0	**1**	1	1
$G4$	**1**	**1**	**1**	**1**	**1**	1	**1**	1	1
$G5$	**1**	0	1	0	0	1	0	0	0
$G6$	**1**	0	**1**	0	0	0	0	0	0
$G7$	**1**	0	0	**1**	0	0	1	0	0

For simplification, a column informing about gene assignment to a group was omitted. Moreover, abbreviations of GO term names were used in the columns' titles. Annotations following from the hierarchy are in bold, in a simple ***DT***, which does not consider the hierarchy, in these places 0 is set. In the table ***DT-Inc*** we try to find all statistically significant rules of the form (2):

$$\textbf{IF} \quad a_{i1} \quad \text{and} \quad a_{i2} \quad \text{and} \quad \ldots \quad \text{and} \quad a_{ik} \quad \textbf{THEN} \quad d = G(l), \quad (2)$$

where

$$\{a_{i1}, a_{i2}, \ldots, a_{ik}\} \subseteq A, \qquad G(l) \in \{G(1), G(2), \ldots, G(n)\}.$$

The interpretation of the rule (2) is consistent with the expression (1). We denote by $RUL_{G(l)}$ a set of rules with identical conclusions and call the description of the gene group $G(l)$.

2.2. Rule induction, evaluation and filtration. The induction of decision rules can be classification-or discovery-oriented. The purpose of classification-oriented induction is to find, on the basis of a set of learning examples, a set of decision rules that will be used to classify new unknown examples (Michalski *et al.*, 1998; Fürnkranz, 1999; Grzymała-Busse and Ziarko, 2003). The purpose of discovery-oriented induction is to discover rules which are interesting and useful for different kinds of users (Fayyad *et al.*, 1996).

Our aim is to find rules describing gene groups by means of co-occurred GO terms. We are not interested in the classification of unknown objects because the selection of all genes that create gene groups was already done during the formation of groups. This means that the concept of test object has no application in our case, and thus we will not use these rules for classification.

For description purposes, all rules satisfying some requirements are usually determined. This approach is implemented, among others, in the Apriori association rule induction algorithm (Agrawal and Srikant, 1994), and the Explore decision rule induction algorithm (Stefanowski and Vanderpooten, 2001). Another possibility is to induce all rules and then filter them to find the most interesting ones (Brzezinska *et al*., 2007; Sikora, 2010).

We generate rules with p-values less than or equal to a threshold established by the user. Therefore, to describe a given gene group, we must determine all possible combinations of all possible subsets of GO terms. Since we consider only premises with descriptors $a(g) = 1$ (at the current stage of the analysis we are not interested in descriptors $a(g) = 0$), in a pessimistic case we would have to determine $\sum_{k=1}^{|A|} \binom{|A|}{k} = 2^{|A|}-1$ rules, which is impossible in the case of a big number of GO terms considered.

A modified version of the Explore algorithm which enables generating iteratively (from one-condition rules) all possible conjunctions of GO terms for each gene group (Stefanowski and Vanderpooten, 2001) was used for rules induction.

For our purpose, the Explore algorithm was modified. The main part of the algorithm generates premises with increasing size, beginning from premises containing one GO term. When a rule created satisfies a p-value criterion established by the user, it is added to an output rule set and a conjunction is extended (assuming that other statistically significant rules can be determined from the conjunction). If all GO terms for a premise being currently created were already considered, then a new GO term is selected (not chosen yet) and a new rule creation begins. In order to limit the search space and shorten the algorithm operating time, during rule induction no terms lying on any path (from the ontology leaf to the root) that leads to a term a are added to the premise since the conjunction of such terms will always be reduced to a term lying lower in the ontology graph (Gruca and Sikora, 2009). However, even after applying the above modifications to the Explore algorithm, the number of rules determined can be huge and then difficult to interpret. Hence we need to provide a method of rule quality evaluation and mutual similarity assessment.

Rules of the form (2) are a special case of the so-called decision rules, and several measures that reflect a quality of a decision rule can be computed (An and Cercone, 2001; Sikora, 2006; Fürnkranz and Flach, 2005; Guillet and Hamilton, 2007).

If the rule r is shortly written as $\varphi \rightarrow \psi$, then $n_\varphi = n_{\varphi\psi}+n_{\varphi\neg\psi} = |G_\varphi|$ is the number of genes that recognize the rule $\varphi \rightarrow \psi$; $n_{\neg\varphi} = n_{\neg\varphi\psi} + n_{\neg\varphi\neg\psi} = |G_{\neg\varphi}|$ is the number of genes that do not recognize the rule $\varphi \rightarrow \psi$; $n_\psi = n_{\varphi\psi} + n_{\neg\varphi\psi} = |G_\psi|$ is the number of genes that belong to the gene group described by the rule $\varphi \rightarrow \psi$; $n_{\neg\psi} = n_{\varphi\neg\psi} + n_{\neg\varphi\neg\psi} = |G_{\neg\psi}|$ is the number of genes that do not belong to the gene group described by the rule $\varphi \rightarrow \psi$; $n_{\varphi\psi} = |G_\varphi \cap G_\psi|$ is the number of genes that support the rule $\varphi \rightarrow \psi$; $n_{\neg\varphi\neg\psi} = |(G-G_\varphi)\cap(G-G_\psi)|$ is the number of genes that do not belong to the gene group described by the rule $\varphi \rightarrow \psi$ and do not recognize it. Values $n_{\varphi\neg\psi}, n_{\neg\varphi\psi}$ are calculated similarly as $n_{\varphi\psi}$ and $n_{\neg\varphi\neg\psi}$.

It can be noticed that for any rule $\varphi \rightarrow \psi$ the inequalities $1 \leq n_{\varphi\psi} \leq |G_\psi|, 0 \leq n_{\varphi\neg\psi} \leq |G_{\neg\psi}|$ hold. Hence, a quality measure is a function of two variables $n_{\varphi\psi}$ and $n_{\varphi\neg\psi}$, $q(\varphi \rightarrow \psi) : \{1,\ldots,|G_\psi|\} \times \{0,\ldots,|G_{\neg\psi|}\} \rightarrow \mathbb{R}$ (Sikora, 2006). Two basic quality measures are accuracy $acc(r) = n_{\varphi\psi}/n_\varphi$ and coverage $cov(r) = n_{\varphi\psi}/n_\psi$.

For knowledge discovery purposes we search for rules characterized simultaneously by high accuracy and high coverage. Unfortunately, a trade-off between the values of the measures exists, and therefore there are attempts to define quality measures (attractiveness or interestingness measures) that combine, among others, rule accuracy and coverage. The most important feature of created rules in biological and medical applications is their statistical significance. A p-value of a rule is calculated based on a suitably chosen statistical test and specifies how much the accuracy of a determined rule differs from the accuracy resulting from example distribution in analyzed data (the so-called *a priori* distribution), while rule coverage is fixed. The probability of obtaining a specific configuration of the number of examples recognized and covered by a rule is described by the hypergeometric distribution (3) (Agresti, 2002).

For the assessment of a statistical significance of a rule we consider the following null hypothesis: *the assignment of examples recognized by the rule to the decision class indicated by the rule is equivalent to random assignment of the examples to the class*.

A one-side (right-side) test is used to verify the hypothesis, because we are only interested in rules which assign examples to the class considered better than the tested rule. A p-value of the test is calculated by summing probabilities obtained for all possible rules recognizing as many examples as the analyzed rule but characterized by higher accuracy (3):

$$p(n_{\varphi\psi}, n_{\varphi\neg\psi}) = \frac{\binom{n_{\varphi\psi}+n_{\varphi\neg\psi}}{n_{\varphi\psi}}\binom{n-n_{\varphi\psi}-n_{\varphi\neg\psi}}{n_\psi-n_{\varphi\psi}}}{\binom{n}{n_\psi}}, \quad (3)$$

$$p_{val}(n_{\varphi\psi}, n_{\varphi\neg\psi}) = \sum_{k=0}^{\min\{n_\psi-n_{\varphi\psi}, n_{\varphi\neg\psi}\}} p(n_{\varphi\psi}+k, n_{\varphi\neg\psi}-k). \quad (4)$$

The formula (4) is a cost criterion which means that rules with lower p_{val} values are considered better than rules with higher p_{val} values. The p_{val} criterion gives us objective assessment of the determined rule quality, thus sorting

rules increasingly with respect to the value p_{val} we can obtain a ranking reflecting the rule quality. As we induce many rules and all of them need to be tested against the null hypothesis, this creates a multiple-testing problem. Thus, for each rule, we compute its corrected p-value using the standard FDR (false discovery rate) Benjamini and Hochberg correction method (Benjamini and Hochberg, 1995). The corrected p-value is computed only for informational purposes and can be further used as an additional indicator of the rule quality.

Apart from the objective criterion, subjective ones which reflect specific user preferences are also applied for rule interestingness assessment. The first subjective criterion of rules attractiveness is the number of GO terms included in a rule premise. We assume that the larger number of terms in a rule premise, the more information represented by the rule (we recall that terms occurring in a premise do not lie on a common path in the ontology graph). The second subjective criterion of rule evaluation is the level of GO terms occurring in the rule premise (5). As far as a description is concerned, we should prefer rules with premises including terms from as low a level of the GO graph as possible. The criterion

$$depth(r) = \frac{\sum_{i=1}^{NoGOterms(r)} level(a_i)}{\sum_{i=1}^{NoGOterms(r)} \max\_path(a_i)} \tag{5}$$

enables verifying how much specific knowledge, from the genes biological functions description point of view, the evaluated rule presents.

In the formula (5) $level(a_i)$ is the level of a GO term a_i that occurs in the rule premise, $\max\_path(a_i)$ is the longest path leading form the root to a leaf of GO-Inc that passes trough the node a_i.

The subjective measures are normalized (they take values from the interval $(0, 1]$) and monotone (adding a new GO term to a rule premise or moving a GO term to a lowest level of the ontology increases the values of both measures). Finally, the measure that incorporates all aspects (objective and subjective measures) of rule quality evaluation presented above is the product of all component measures:

$$Q(r) = \frac{1}{p_{val}(r)} length(r)\, depth(r). \tag{6}$$

The rules ranking established by the Q measure is the basis for selecting the most interesting rules from the determined rules set. We are interested in the best rules that cover as many genes from the described gene group as possible. Additionally, if a given subgroup of genes is described by a rule with a better quality, then we are not interested in rules with a lower quality covering the same genes. This leads to the filtration algorithm, which significantly limits the number of rules. The filtration algorithm creates a coverage of gene groups. Rules are added to a result (filtered) set of rules starting from the best rules describing each group. After adding a rule to a result set of rules, all genes that cover the rule are removed from the group. The filtration is performed until all genes covered by the unfiltered set of rules are covered by the filtered set of rules. However, if the rule rr covers (supports) the same objects as the reference rule but contains other biological knowledge (it is built of GO terms different from the GO terms included in the reference rule), then the rule rr is not removed but remains in the result set of rules. Dissimilarity of rules is determined by

$$diss(r_i, r_j) = 1 - \frac{uGOterms(r_i, r_j) + uGOterms(r_j, r_i)}{NoGOterms(r_i) + NoGOterms(r_j)}. \tag{7}$$

If rules r_i and r_j are dissimilar to a degree greater than the threshold ε defined by the user, both of them remain in the final set of determined rules. Here $uGOterms(r_i, r_j)$ is the number of unique GO terms occurring in the rule r_i and not occurring in the rule r_j; a GO term a from the rule r_i is unique if it does not occur directly in the rule r_j and there is no path in the GO-Inc graph that includes both term a and any other term from the rule r_j premise; $NoGOterms(r)$ is the number of GO terms in the rule r premise.

3. Evaluation of the importance of GO terms appearing in induced rule and renewed rules induction

Let us consider a given rule r of the form (2). A set of GO terms occurring in the rule premise is denoted by W, that is, $W = \{t_1, t_2, \ldots, t_n\}$. In the standard form presented by Greco *et al.* (2007), the Banzhaf measure that evaluates the contribution of elementary condition (a single GO term t_i here) to rule r accuracy is calculated according to

$$\phi_B(t_i, r) = \frac{1}{2^{n-1}} \sum_{Y \subseteq W \setminus \{t_i\}} \Big[acc(Y \cup \{t_i\}, r) - acc(Y, r) \Big] \tag{8}$$

where $acc(Y, r)$ denotes the accuracy of r in the premise part of which only GO terms included in the set Y occur; $acc(\emptyset, r) = 0$; $acc(W, r) = acc(r)$.

To evaluate the importance of a GO term in whole set RUL_G, it is necessary to compute its importance in each rule from the description (in each rule that includes the GO term considered) and verify whether, by any chance, the GO term also occurs in rules from other gene groups descriptions. We can represent the above requirements as a

formula which allows evaluating the GO term importance for the description of a given gene group:

$$G(t_i, RUL_G) = \sum_{r \in RUL_G} \phi_B(t_i, r) cov(r) - \sum_{r \notin RUL_G} \phi_B(t_i, r) cov(r), \quad (9)$$

where $cov(r)$ is the coverage of the rule r, and RUL_G is the set of rules that create the description of the gene group G.

In the described method of GO term importance evaluation, the contribution of each term to rule accuracy is evaluated, while its contribution to rule coverage is not (the coverage is considered for the whole rule only). For the purpose of importance evaluation of GO terms describing a gene group, it would be better to evaluate the contribution of the analyzed GO term to both accuracy and coverage. The measure proposed in the paper is empirical and enables evaluating the accuracy and coverage of a rule simultaneously, taking into consideration example distribution among the decision class indicated by the rule and the other decision classes. Making an analysis of the formula

$$Rss(\varphi \rightarrow \psi) = \frac{n_{\varphi\psi}}{n_{\varphi\psi} + n_{\neg\varphi\psi}} + \frac{n_{\neg\varphi\neg\psi}}{n_{\neg\varphi\neg\psi} + n_{\varphi\neg\psi}} - 1, \quad (10)$$

it can be noticed that the measure proposes the method of rule evaluation analogous to the method of classifier sensitivity (the first component of the sum) and specificity (the second component of the sum) evaluation (Bairagi and Suchindran, 1989; Grzymała-Busse *et al.*, 2005; Fürnkranz and Flach, 2005). The first component of the measure characterizes the conditional probability of an event in which an example covering the rule r belongs to a group of genes described by the rule. The second component specifies the conditional probability of an event in which examples not covering the rule r belong to gene groups different from the one the rule r points at. The third component has a normalization role and guarantees that the measure Rss has the property of confirmation (Greco *et al.*, 2004).

The measure takes values from the interval $[-1, 1]$ and values equal to zero are achieved while a rule has the same accuracy as implied from the positive and negative example distribution in the training set. Rules with values greater than zero (the greater the value, the better the rule) should be recognized as good ones.

Using the Rss measure in the formula (8) allows considering the contribution of the GO term to the accuracy and coverage of all "subrules" which can be created from the rule r:

$$\phi_B(t_i, r, Rss) = \frac{1}{2^{n-1}} \sum_{Y \subseteq W \setminus \{t_i\}} \left[Rss(Y \cup \{t_i\}, r) - Rss(Y, r) \right], \quad (11)$$

where $Rss(Y, r)$ denotes the value of the Rss measure determined for the rule r, in the premise part of which only GO terms included in the Y occur; $Rss(\emptyset, r) = -1$; $Rss(W, r) = Rss(r)$. As a consequence of the introduced modification, the measure G that evaluates the quality (importance) of a GO term in the whole set of rules describing the analyzed gene group (12) was also modified:

$$G(t_i, RUL_G, Rss) = \sum_{r \in RUL_G} \phi_B(t_i, r, Rss(r)) - \sum_{r \notin RUL_G} \phi_B(t_i, r, Rss(r)) \quad (12)$$

Instead of the Rss measure, any rule quality evaluation measure can be used.

It is worth mentioning that $acc(Y, r)$ is a ‚monotonic measure which means that, if $Y \subseteq X \subseteq W$, then $acc(Y, r) \leq acc(X, r)$. It can be proved easily that, in the worst case, adding the next conjunction may not improve rule accuracy. Moreover, it is also easy to prove that the Rss measure is not monotonic, because removing the condition from the premise, though can decrease the accuracy of the rule, may also increase rule coverage, which can finally increase the value of the Rss measure. However, if the measure Rss is treated as a function of two variables $n_{\varphi\psi}, n_{\varphi\neg\psi}$, then the measure is monotone with respect to each variable (if the value of the second variable is fixed) (Fig. 2). Moreover, since Rss takes negative values (hence from the mathematical point of view it is not a measure, but we want to keep the convention used in the domain literature), GO terms with negative assessment of the importance can appear among GO terms composing the rule premise. The negative value of (12) means that "subrules" have, on average, a better quality without the GO term t_i than "subrules" with the term, since in such a case the term t_i should be considered unimportant (causing noise).

Based on the assessment of GO terms that occur in premises of rules describing individual gene groups, for each group we can obtain a ranking of GO terms which reflects the importance of a given term in the context of determined rules as well as in the context of a description of a given gene group. Obviously, terms from the top of the ranking will describe the biological function of genes from a given gene group better (stronger) than terms from the end of the ranking. In the rule induction algorithm presented in a previous section we searched for all rules with p_{val} less than a significance threshold fixed by the user. The obtained rules were evaluated and filtrated.

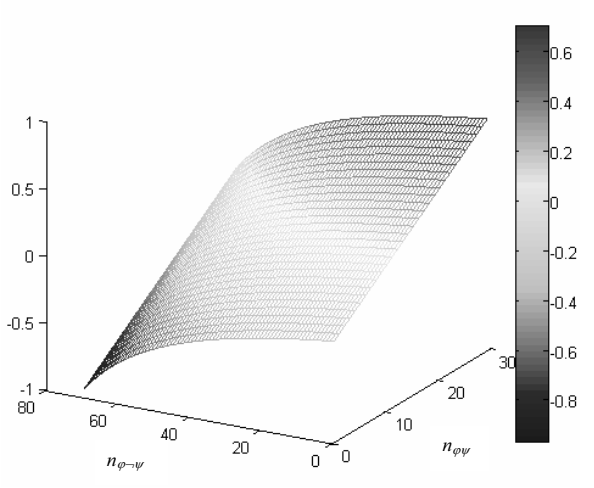

Fig. 2. Graph of the *Rss* measure for a classes distribution amounting to 30 positive and 70 negative examples.

Below, we present an algorithm that enables obtaining a coverage of a group of genes using the greedy search strategy for the creation of elementary conditions included in the rule premise. GO terms are added to the rule premise iteratively (Fürnkranz, 1999; Grzymała-Busse and Ziarko, 2003). However, only terms appearing in premises of previously determined statistically significant rules are used. Moreover, the procedure of rule creation also considers the importance of GO terms for previously determined rules. Including the expression *the term does not describe the given gene group* (in short, $not\ a(g)$) in rule premises is an additional unique feature of the presented algorithm.

For biological description of groups of genes, we are more interested in the information that genes from the group (or some part of them) have a specific biological function (they are annotated by a given GO term) than in the information that they do not have the function (they are not annotated by the term). However, if such a negative component is useful for making the description more precise, especially for describing these groups which normally have inaccurate or incomplete description (a big number of genes from the given group is covered by no statistically significant rules determined in a standard manner), then adding negative components to the rule premise will be purposeful. Moreover, for the sake of relations appearing in the GO-Inc graph, a negative component disqualifies all GO terms lying at levels lower than this component in GO-Inc.

Negative components of premises of determined rules are created based on the assumption that the GO terms important for the description of a particular gene group should be unimportant for another gene group. Therefore, it is worth trying to apply negations of terms describing that particular group to the description of the other group.

The idea of rule induction presented in the paper can be introduced by Algorithm 1.

Algorithm 1 Rules induction algorithm

Input: ***DT-Inc***$=(G, A\cup\{d\})$, $RssDesc_j = \{t_j^1, \ldots, t_j^{m_j}\}$ ranking of GO terms occurring in the description of gene group G_j established by the evaluation measure Rss.
Output: RUL_{out} set of rules describing all gene groups

$RUL_{out} = \emptyset$
for each gene group $G_j, j \in \{1, \ldots, n\}$ **do**
  $G_g := G_j$
  while ($G_g \neq \emptyset$ **or** any GO term from $RssDesc_j$ covers any gene belonging to G_g) **do**
    $r := \emptyset$ {*start from empty premise*}
    $Rss(r \rightarrow G) = -1$ {*put the minimal value of evaluation measure*}
    for $i = 1, \ldots, m_j$ **do** {*add positive descriptors to the rule premise*}
      if $Rss(r \wedge t_j^i \rightarrow G_j) > Rss(r \rightarrow G_j)$ **then**
        $r := r \wedge t_j^i$
      end if
    end for
    for $l = 1, \ldots, n$ **and** $l \neq j$ **do** {*add negative descriptors to the rule premise*}
      for $s = 1, \ldots, m_l$ **do**
        if $Rss(r \wedge \neg t_l^s \rightarrow G_j) > Rss(r \rightarrow G_j)$ **then**
          $r := r \wedge t_l^s$
        end if
      end for
    end for
    $Shorten(r \rightarrow G_j)$
    $RUL_{out} = RUL_{out} \cup \{r \rightarrow G_j\}$
    $G_g := G_g - [r]$ {$[r]$ *is the set of genes from the group* G_j *covered by* r}
  end while
end for

The function $Shorten(r \rightarrow G_j)$ is used to reduce the number of GO terms contained in a premise of a determined rule, because considering all GO terms that create a description of a given gene group and all negations of terms creating descriptions of the rest of the groups results in very long premises and specific rules.

During the shortening process, a hill climbing strategy that consists in successive removing a single GO term occurring in the premise and checking how it influences the rule quality (the measure Rss value) is applied. After removing all GO term trials, the one whose removal does not decrease the rule quality (the reference quality is the quality of a whole non-shortened rule) is removed permanently. If removing a GO term from a rule premise results in a rule quality increase, then this new (higher) quality becomes the reference quality. GO terms are being removed until the rule quality starts decreasing.

Since positive elementary conditions occurring in a rule premise are more desired, the algorithm tries to determine a rule from GO terms contained in rules describing a given gene group first and then tries to include negative elementary conditions in the premise in order to improve the rule quality by making it more accurate.

A comparison of the efficiency of the discussed algorithms is presented in the next section.

4. Data analysis

Experiments were conducted on two freely available data sets: YEAST and HUMAN, which can be treated as benchmark sets designed for the comparison of genes annotation methods. The YEAST data set contains values of expression levels of budding yeast *Saccharomyces cerevisiae* measured in several DNA microarray experiments (Eisen *et al*., 1998). Our analyses were performed on 274 genes from 10 top clusters presented by Eisen *et al.*(1998). The HUMAN data set contains values of expression levels of human fibroblasts in response to serum (Iyer *et al*., 1999). In the paper by Iyer *etal.* (1999), 517 EST sequences were reported and divided into ten clusters. After translating the sequences for unique gene names and removing sequences that are duplicated or that are currently considered to be invalid, we obtained a set of 309 genes. Then, each gene from the YEAST and HUMAN data sets were described by GO terms from biological process (BP) ontology.

Table 2. Number of genes and GO terms obtained for each decision table.

Decision table	No. of GO terms	No. of genes
YEAST BP	249	274
HUMAN BP	390	309

We used GO terms from at least the second ontology level, describing at least five genes from our data sets, and we finally obtained four ***DT-Inc*** (the numbers of genes and GO terms for each decision table are presented in Table 2). For each decision table we computed decision rules with statistical significance lower than or equal to 0.05 and at most five GO terms in induced rules premises.

Based on the rules determined in such a manner, the importance of GO terms occurring in these rules was assessed, and then repeated rule induction using the obtained GO terms ranking was executed. Since the algorithm considering terms importance creates the coverage of a group of genes, a 100% coverage of each gene group was obtained every time, althoug statistically unimportant rules were among the determined ones. Thus, in the next step, all statistically unimportant rules ($p_{val} > 0.05$) were removed from the determined set of rules. During the first run of the algorithm, we did not consider GO terms negations—this part of experiments is described in the columns *Rules induction without negative GO terms*. Negations of GO terms are considered in the second experiment (*Rules induction with negative GO terms*).

A summary statement of the total coverage and the number of rules obtained by various induction methods are presented in the last rows of Tables 3 and 4.

The next comparison concerns the number of GO terms occurring in the determined rules; summary statements are presented in Tables 5 and 6. In the case of rules including negations of GO terms, the number of terms that had the form of negation is presented in parentheses.

The results presented in Tables 3 and 4 show that an exhaustive strategy for generating all statistically significant rules results in determining a huge number of rules, although applying the compound rule quality measure and the filtration algorithm enables limiting the description to the most interesting rules. The number of rules describing each of the analyzed gene groups is not big, and thus they can be used for biological interpretation of the analyzed groups. A set of rules describing an analyzed gene group enables covering all or the majority of genes contained in the group, which is also an important feature of the presented method.

Moreover, it is worth noticing that the coverage of a gene group, the number of rules after filtration and the quality of the rules strongly depend on the group analyzed. For example, it is commonly known that groups from the YEAST dataset are better defined than groups from the HUMAN dataset. This can be also observed in results presented in the paper, since the coverage of decision classes and the quality of rules obtained are better for the YEAST dataset than for HUMAN.

The application of the rule induction algorithm using a ranking of GO terms importance enables reducing the number of rules describing genes, simultaneously increasing the number of genes covered by the rules. Rules without negative condition terms occurring in premises do not lie on the same path of the ontology graph, like rules that are the basis of GO term ranking creation. Among rules determined in such a way there are those belonging to the input rules set (which is the basis of GO terms ranking creation) as well as those which are not contained in the set. This means that the rules were previously removed during filtration, because they did not satisfy the filtration conditions given by the compound measure Q. The algorithm considering the importance of GO terms does not use the measure Q. Moreover, it puts greater emphasis on the conciseness of the obtained description. Thus such rules can appear in the description now. Applying negative conditions significantly increases the coverage of gene groups descriptions but, unfortunately, makes the interpretation of the obtained rules more difficult. From the definition of the relation $\leq$ in the ontology graph it follows that if the gene g is not annotated by the term t, then it is annotated by no term s satisfying the relation $s \leq t$, either. Therefore, rules that include negative annotations can help in the interpretation of biological information contained in a gene group by excluding negatively annotated terms together with all annotations placed below these terms in the ontology graph (at lower levels). For example, the rule pre-

Table 3. Number of decision rules and gene group coverage obtained for the YEAST data set.

Gene group	Examples	Significant rule induction			Rules induction without negative GO terms		Rules induction with negative GO terms	
		No of rules	No of rules after filtration	Coverage	No of rules	Coverage	No of rules	Coverage
1	11	377	6	100%	3	100%	3	100%
2	27	1447	4	100%	1	100%	1	100%
3	14	4308	12	93%	3	93%	3	93%
4	17	43083	22	100%	3	100%	3	100%
5	22	307	4	41%	2	64%	2	77%
6	15	20225	11	93%	3	100%	3	93%
7	8	6645	12	100%	1	100%	1	100%
8	139	3842	12	100%	2	100%	1	99%
9	5	54426	12	100%	2	100%	1	80%
10	16	810	8	94%	5	100%	4	94%
$\sum$	**274**	**-**	**103**	avg **94%**	**25**	avg **97%**	**22**	avg **96%**

Table 4. Number of decision rules and gene group coverage obtained for the HUMAN data set.

Gene group	Examples	Significant rule induction			Rules induction without negative GO terms		Rules induction with negative GO terms	
		No of rules	No of rules after filtration	Coverage	No of rules	Coverage	No of rules	Coverage
1	62	5	5	27%	4	15%	5	100%
2	83	1253	15	53%	2	58%	5	90%
3	27	16	4	44%	4	48%	6	100%
4	31	2336	20	88%	6	84%	10	77%
5	5	7134	8	80%	3	80%	3	80%
6	21	37432	44	90%	9	71%	7	62%
7	11	841	10	46%	4	64%	4	45%
8	40	23	3	26%	5	35%	8	78%
9	14	147	6	43%	4	43%	4	50%
10	15	12988	30	81%	5	47%	5	47%
$\sum$	**309**	**-**	**145**	avg **47%**	**46**	avg **48%**	**57**	avg **83%**

sented below excludes from the description 3101 terms, which makes up 17% of the whole ontology graph:

IF regulation of nitrogen compound metabolic process **and** ($p_{val}=0.00277$)
gene expression **and** ($p_{val}=0.00143$)
not (negative regulation of metabolic process **or** ($p_{val}=0.62721$)
cell division **or** ($p_{val}=0.05836$)
behavior **or** ($p_{val}=0.13548$)
multicellular organismal homeostasis **or** ($p_{val}=0.13548$)
wound healing **or** ($p_{val}=0.28381$)
regulation of catalytic activity **or** ($p_{val}=0.05172$)
chromosome organization **or** ($p_{val}=0.22460$)
cellular catabolic process) ($p_{val}=0.65636$)
THEN gene group is 2
($p_{val} = 3.35e-07, FDR = 8.55e-06$). (13)

Rules with negative annotations can be especially useful for negative verification of hypotheses concerning biological functions of genes that create a given group.

Irrespective of the induction algorithm applied, premises of rules are composed of statistically significant as well as of insignificant GO terms. The following two rules present this property:

IF aerobic respiration **and** ($p_{val} = 0.05496$)
biopolymer metabolic process **and** ($p_{val} = 0.26048$)
cellular macromolecule metabolic process ($p_{val} = 0.32327$)
THEN gene group is 3
($p_{val} = 0.00011, FDR = 0.00016$), (14)

IF generation of precursor metabolites and energy **and** ($p_{val}=2.88e-11$)
primary metabolic process **and** ($p_{val} = 0.39566$)
THEN gene group is 6
($p_{val} = 9.69e-12, FDR = 2.10e-11$). (15)

In the case of the first rule, a conjunction of three statistically insignificant GO terms gives as a result a sta-

Table 5. Number of rules and descriptors obtained for descriptions of each gene group (YEAST dataset).

Gene group	Significant rules indiction		Rules induction without negative GO terms		Rules induction with negative GO terms	
	Rules	GO rules	Rules	GO rules	Rules	GO rules
1	6	17	3	4	3	4 (0)
2	4	12	1	1	1	1 (1)
3	12	25	3	5	3	8 (3)
4	22	31	3	5	3	6 (1)
5	4	7	2	3	2	8 (6)
6	11	11	3	5	3	8 (6)
7	12	19	1	1	1	1 (0)
8	12	22	2	5	1	5 (3)
9	12	17	2	4	1	2 (0)
10	8	12	5	6	4	8 (3)

Table 6. Number of rules and descriptors obtained for descriptions of each gene group (HUMAN dataset).

Gene group	Significant rules indiction		Rules induction without negative GO terms		Rules induction with negative GO terms	
	Rules	GO rules	Rules	GO rules	Rules	GO rules
1	6	11	4	9	5	11 (2)
2	15	20	2	2	5	55 (50)
3	11	21	4	7	6	47 (40)
4	26	40	6	9	10	19 (5)
5	8	18	3	5	3	5 (0)
6	42	63	9	16	7	15 (1)
7	8	18	4	8	4	11 (2)
8	4	8	5	8	8	42 (32)
9	6	19	4	6	4	8 (2)
10	31	56	5	12	5	12 (0)

tistically significant rule. In the case of the second rule, adding a statistically insignificant term to a statistically significant one allows obtaining a better rule.

Considering quantitative results of the conducted experiments and the purpose of rule induction, the following methodology of determining rules describing gene groups can be proposed:

- Set an acceptable level of rule statistical significance, set the maximal number of GO terms occurring in the premises of rules determined by the significant rule induction algorithm.
- Perform rule induction and apply the filtration algorithm using the measure (6).
- If, for a given gene group, the number of rules is too high or the coverage is too low, perform for these groups rules induction using the algorithm that applies GO term importance (the option without negative annotations).
- If the obtained descriptions are still unsatisfactory, apply the algorithm that evaluates GO term importance once more (the option with negative annotations).

To improve the quality of the output description of a gene group, descriptions obtained by means of particular algorithms can be joined. For rules included in the joined sets of rules, a measure evaluating the quality of rules can be defined and the description of a gene group can be created by means of a filtration algorithm similar to the one presented in Section 2.2. Defining such a measure and a filtration algorithm will be the subject of future research.

Decision rules are generated mainly for description purposes to support drawing biological conclusions from DNA microarray experiments. Thus, real verification of the rule quality is its ability to provide biological interpretation of the genes composing analyzed groups. Do the determined rules and information about the reduced set of GO terms have any interesting biological interpretation? Below we present an example decision rule from the YEAST data set generated for the gene group No. 6:

IF	oxidation reduction	**and**
	oxidative phosphorylation	**and**
	hydrogen transport	**and**
	monovalent inorganic cation transport	**and**
	ion transmembrane transport	
THEN	gene group is 6	
Accuracy: 1.0	**Coverage**: 0.6	
	($p_{val} = 2.383e-13$, $FDR = 3.93e-13$).	

(16)

The group indicated in the conclusion of the above rule was described as ATP synthesis in the original paper of Eisen *et al.* (1998), and it is a group which includes genes related to ATP synthesis. The ATP synthase enzyme is an enzyme which catalyses the reaction of ATP synthesis in mitochondria using transmembrane electrochemical proton potential difference which forces to move the protons from the inter-membrane space into the matrix in mitochondria during oxidative phosphorylation. The difference between the concentration of protons (H+) between the matrix and the inter-membrane space drives a flux of ions across the membrane down the proton gradient. The structure of this enzyme is complex—in yeast organisms, the ATP synthase complex consists of the two distinct multisubunit portions: F1 and FO. Group 6 includes 15 genes—among them there are nine genes that encode components of the ATP synthase. The rule above is supported and recognized by nine genes (ATP1, AT2, ATP3, ATP4, ATP5, ATP7, ATP14, ATP16, ATP17). This means all of the genes from Group 6 that encode components of ATP synthase.

Another exemple decision rule is the rule obtained for the HUMAN dataset for the gene group No. 6:

IF	organ development	**and**
	regulation of locomotion	**and**
	positive regulation of cell motion	**and**
	muscle cell proliferation	
THEN	gene group is 6	
Accuracy: 0.75	**Coverage**: 0.14	
	($p_{val} = 0.001$, $FDR = 0.05$).	

(17)

The above rule is supported by three genes which encode proteins involved in the process of angiogenesis—the development of a vascular supply system which is a fundamental requirement for organ development. VEGFA is one of the most important proteins from the VEGF (vascular endothelial growth factor) subfamily of growth factors. VEGFA stimulates cellular responses by binding to tyrosine kinase receptors Flt-1 and KDF/Flk-1. VEGFA and the fibroblast growth factor 2 (FGF2) both are well-investigated proangiogenic molecules (Kano *et al.*, 2005). There are evidences that VEGFA regulates the expression of FLT1 (Mata-Greenwood *et al.*, 2003) and FGF2 (Seghezzi *et al.*, 1998).

In Tables 7 and 8 we present the most and the least important GO terms obtained for each gene group for both datasets. In the tables we also included the description given by the authors of the original papers. As can be seen (in the case of the YEAST dataset) the most important GO terms obtained for each gene group are consistent with the description given by the authors of the original paper. For example, in yeast organisms, the spindle pole body controls the assembly of all microtubules in the cell, glycolysis is a metabolic pathway that converts glucose (hexose molecule) into pyruvate, and one of the functions of chromatin is to package DNA. In the case of the HUMAN dataset, we cannot find any similarity by comparing the most important GO terms and the description given by the authors of the paper. This is due to the fact that the authors described only part of the genes composing the clusters, for example, 24 genes from Cluster 1 were described as signal transduction genes while 28 genes were described as cell cycle and proliferation genes and 48 remaining genes were no function assigned. However, we would like to stress that the description of the gene group is created by all important GO terms obtained during the analysis, not only by one, most important of them. The analysis of the last columns of Tables 7 and 8 shows that some terms occur to be worst for several gene groups (i.e., *biopolymer biosynthetic process* in Table 7 and *response to stress* in Table 8). If these terms appear in rules describing other gene groups, we can treat them as noise. We can also notice that some of the terms that are worst for one of the groups are best for other groups (i.e., *biopolymer biosynthetic process* in Table 7 and *protein complex assembly* in Table 8), which is consistent with the intuition.

The results presented in Tables 7 and 8 show that terms which are most significant for the rule description of a given gene group are also characterized by high statistical significance. However, it does not mean that the term with the best assessment of the importance is also statistically the most significant term describing the given gene group. In the obtained data sets, many examples that confirm the above statement can be found, which means that the GO terms assessment (hence the whole rules, too) obtained using the measure Rss differs from the assessment obtained by p_{val}.

We also compared results of our analysis with those obtained from the Genecodis service (Carmona-Saez *et al.*, 2007). The results of the comparison for the YEAST dataset are presented in Table 9.

The rule

IF	transcription	**and**
	regulation of transcription, DNA dependent	**and**
	DNA replication	**and**
	protein complex assembly	**and**
	pre-replicative complex assembly	**and**
	DNA replication initiation	**and**
	S phase of mitotic cell cycle	**and**
	DNA strand elongation during DNA replication	
THEN	gene group is 9	

(18)

is an example rule obtained from Genecodis. This rule

Table 7. Results of the evaluation of GO terms significance for the YEAST dataset.

Gene group	Desription from the original paper	Best GO terms Name (importance; p_{val})	Worst GO term Name (importance)
1	spindle body assembly and function	microtubule-based process (0.48; 1.568112e-11)	modification-dependent macromolecule catabolic process (–1.17)
2	proteasome	modification-dependent macromolecule catabolic process (1.18;17 1.415534e-13)	sexual reproduction (–0.07)
3	mRNA splicing	mRNA metabolic process (0.33; 1.803550e-08)	mitochondrion organization (–1.32)
4	glycolysis	hexose catabolic process (0.89; 5.160317e-13)	gene expression (–0.51)
5	mitochondrial ribosome	mitochondrion organization (1.32; 4.087841e-13)	biopolymer biosynthetic process (–0.49)
6	ATP synthesis	ribonucleoside triphosphate biosynthetic process (0.81; 3.449463e-13)	cell death (–0.08)
7	chromatin structure	DNA packaging (1.66; 0)	biopolymer biosynthetic process (–0.54)
8	ribosome and translation	biopolymer biosynthetic process (0.49; 1.040279e-13)	ribosome assembly (–0.24)
9	DNA replication	DNA replication initiation (1.60; 1.084954e-07)	negative regulation of biosynthetic process (–0.52)
10	tricarboxylic acid cycle and respiration	electron transport chain (0.35; 3.504974e-13)	ion transport (–0.23)

Table 8. Results of the evaluation of GO terms significance for the HUMAN dataset.

Gene group	Desription from the original paper – partially	Best GO terms Name (importance; p_{val})	Worst GO term Name (importance)
1	signal transduction / cell cycle and proliferation	primary metabolic process (0.04; 0.03889)	cell communication (–0.15)
2	immediate-early transcription factors/ coagulation and hemostasis	cellular biopolymer metabolic process (0.18; 0.00441)	regulation of RNA metabolic process (–0.12)
3	other transcription factors/ inflammation	lipid metabolic process (0.07; 0.00073)	adaptive immune response based on somatic recombination of immune receptors built from immunoglob. superfamily domains (–0.40)
4	angiogenesis	cell cycle checkpoint (0.40; 0.00004)	protein complex assembly (–0.22)
5	tissue remodeling	response to toxin (1.25; 0.00206)	response to stress (–0.19)
6	cytoskeletal reorganization	positive regulation of cell division (1.21; 0.00008)	positive regulation of RNA metabolic process (–0.39)
7	re-epithelialization	protein complex assembly (0.22; 0.15404)	positive regulation of cellular biosynthetic process (–0.32)
8	unidentified role in wound healing	blood circulation (0.10; 0.00282)	cell communication (–0.12)
9	cholesterol biosynthesis	developmental growth (0.36; 0.00234)	regulation of angiogenesis (–0.41)
10	no description	skeletal system development (0.61; 0.01570)	response to stress (–0.17)

Table 9. Comparison of rules obtained from Genecodis and Explore for the YEAST dataset.

Gene group	Method of analysis	Coverage	Number of rules
1	Genecodis	45%	4
	Significant rules induction	100%	6
2	Genecodis	100%	6
	Significant rules induction	100%	4
3	Genecodis	0%	0
	Significant rules induction	93%	12
4	Genecodis	88%	23
	Significant rules induction	100%	22
5	Genecodis	0%	0
	Significant rules induction	41%	4
6	Genecodis	60%	3
	Significant rules induction	93%	11
7	Genecodis	100%	5
	Significant rules induction	100%	12
8	Genecodis	25%	2
	Significant rules induction	100%	12
9	Genecodis	80%	4
	Significant rules induction	100%	12
10	Genecodis	75%	12
	Significant rules induction	94%	8

was generated for objects from the gene group No. 9, from the YEAST dataset.

The analysis of the structure of the GO graph for BP ontology revealed that there are the following relations among GO terms composing the above rule:

regulation of transcription, DNA dependent $\leq$ *transcription,*
DNA replication initiation $\leq$ *DNA replication,*
DNA strand elongation during DNA replication $\leq$ *DNA replication,*
pre-replicative complex assembly $\leq$ *DNA replication.*

Genecodis does not perform any initial selection of the attributes that are added to the premise of the created rule—it simply generates all possible combinations of GO terms. As a result of such an approach, one may obtain rules that include redundant information in their premises, i.e., GO terms that are in relation $\leq$ with other GO terms composing the rule. With the Explore method, the rules obtained include smaller number of GO terms, but each term describes a different biological process.

Below we present the rule obtained by our version of the Explore algorithm for the same gene group:

IF	transcription, DNA dependent	**and**	(19)
	regulation of transcription	**and**	
	protein-DNA complex assembly	**and**	
	interphase of mitotic cell cycle	**and**	
	DNA replication initiation		
THEN	gene group is 9.		

The rule (19) covers exactly the same genes as the previous rule (18). As can be easily noticed, the rule obtained with the modified Explore method does not include GO terms *DNA replication* and *transcription*, which are parent terms to the terms: *regulation of transcription*, *DNA dependent and DNA replication initiation*; *pre-replicative complex assembly*; *DNA strand elongation during DNA replication, respectively*. The term *S phase of mitotic cell cycle* from the rule (18) is replaced by the term *interphase of mitotic cell cycle,* which is the immediate parent of the term *S phase of mitotic cell cycle*. Also the term *regulation of transcription*, *DNA dependent* was replaced by its two immediate parent terms: *transcription*, *DNA dependent and regulation of transcription*. There is also a close relation among the terms protein complex assembly and *the protein DNA complex assembly* from the Genecodis rule and the term *protein-DNA complex assembly* from the Explore rule. The absence of the term *DNA strand elongation during the DNA replication* can be explained by the fact that during the generation of the rules by our version of the Explore algorithm we limited the number of rule descriptors to five.

It is very difficult to compare the results obtained with the use of both methods. Due to the number of possible combinations, similarities among different GO terms describing the same biological processes and differences between the two methods, the obtained descriptions differ in many aspects. However, as can be easily noticed, the rules obtained by our version of the Explore algorithm are easier to interpret than Genecodis rules, since we do not include terms lying on the common ontology path to a rule premise.

5. Conclusion

Issues of the description of gene groups by means of GO terms were presented in the paper. Logical rules were used

as a language of description. Work related to gene groups description by means of rules with various representations was presented. A novel method of the induction, evaluation and filtration of multiattribute rules describing gene groups was outlined in the main part of the paper. A method of importance evaluation of a single GO term occurring in the premises of the obtained rules and an algorithm of rule induction that considers the obtained GO terms ranking were also introduced in the paper.

The proposed method of significant rule induction, evaluation and filtration as algorithms that consider ranking of GO terms appeared to be very effective. We are able to obtain small rule sets having better average statistical significance than unfiltered rule sets.

The presented method of significant rules induction guarantees that all statistically significant rules are determined. The proposed approach differs from other methods in the following features: the method of the evaluation and filtration of the rules and the fact that terms lying on the same path in the gene ontology graph do not occur in the rule premises simultaneously (like, for example, in Genecodis). Both objective factors (statistical significance of rules) and subjective factors (premises composed of many GO terms assigned to the lowest possible level in the ontology graph) are involved in rule evaluation.

The analysis of the obtained results shows that establishing a ranking of GO terms that describe the given gene group provides additional knowledge about the group. The analysis of the rules and the results presented in Tables 7 and 8 leads to the conclusion that the best GO terms describing the given gene group usually occur in the best rules describing the group and do not occur in rules describing other groups (even if they do, they are recognized as the worst or almost the worst terms and are placed at the bottom of the other group rankings). Among the worst GO terms there are those that describe more than one group of genes—it could be interesting to verify the similarity of the groups (or at least part of genes belonging to these groups) which are described by the same, least significant, GO terms. The least significant terms usually have negative values of the coefficient (12); this result is consistent with intuition and can be justified by the fact that such terms introduce noise in gene groups description.

The application of the rule induction algorithm using a ranking of important GO terms enables reducing the number of rules describing genes while simultaneously increasing the number of genes covered by the rules. Rules with negative annotations can be especially useful for negative verification of hypotheses concerning biological functions of genes that create a given group. The presented algorithms may be useful tools that help biologists to understand and interpret results of DNA microarray experiments. Results of experiments demonstrate that the proposed method of rule induction and postprocessing is efficient. In particular, the method enables discovering automatically the dependences which were found during research published by biologists (Bruckmann *et al*., 2007; Mata-Greenwood *et al*., 2003; Kano *et al*., 2005; Seghezzi *et al*., 1998).

The algorithms of significant rule induction and rule filtration are available through the RuleGO Internet service (`www.rulego.polsl.pl`) (Gruca *et al*., 2009). In the future, we plan to extend the service by adding the possibility of evaluating GO terms importance. We also plan to add a new functionality to the service to allow the user to semi-automatically generate a description of groups of genes. The user will be able to generate three different sets of rules for an analyzed group of genes using each of the rules generation methods described in this paper. By comparing the obtained sets of rules, the user will be able to choose the best description. They will also have the possibility to create their own description of the gene group. By adding rules from the all three sets of rules to a new description and analyzing the coverage of the newly created set of rules, the user will be able to create the best possible description of the analyzed group of genes.

Acknowledgment

This work was partially supported by the European Community through the European Social Fund. The first version of this paper was presented during the 15th National Conference on *Application of Mathematics to Biology and Medicine* (Szczyrk, Poland, 2009) and published in a shortened form in the conference proceedings. We would like to thank the anonymous reviewers for helpful feedback and comments on drafts of this paper.

References

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules, *VLDB'94, Proceedings of the 20th International Conference on Very Large Data Bases, Santiago de Chile, Chile*, pp. 487–499.

Agresti, A. (2002). *Categorical Data Analysis*, Wiley Interscience, Hoboken, NJ.

Al-Shahrour, F., Minguez, P., Vaquerizas, J., Conde, L. and Dopazo, J. (2005). Babelomics: A suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments, *Nucleic Acids Research* **33**: W460–W464.

An, A. and Cercone, N. (2001). Rule quality measures for rule induction systems: Description and evaluation, *Computational Intelligence* **17**(3): 409–424.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G. and Sherlock, G. (2000). Gene ontology: Tool for the unification of biology, *Nature Genetics* **25**(1): 25–29.

Bairagi, R. and Suchindran, C. (1989). An estimator of the cutoff point maximizing sum of sensitivity and specificity, *Sankhya, Indian Journal of Statistics* **51**(B-2): 263–269.

Baldi, P. and Hatfield, G. (2002). *DNA Microarrays and Gene Expression*, Cambridge University Press, Cambridge.

Banzhaf, J. (1965). Weighted voting doesn't work: A mathematical analysis, *Rutgers Law Review* **19**(2): 317–343.

Benjamini, Y. and Hochberg, T. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B* **57**(1): 289–300.

Bruckmann, A., Hensbergen, P., Balog, C., Deelder, A., de Steensma, H. and van Heusden, G. (2007). Post-transcriptional control of the saccharomyces cerevisiae proteome by 14-3-3 proteins, *Journal of Proteome Research* **6**(5): 1689–1699.

Brzezinska, I., Greco, S. and Slowinski, R. (2007). Mining pareto-optimal rules with respect to support and confirmation or support and anti-support, *Engineering Applications of Artificial Intelligence* **20**(5): 587–600.

Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J. and Pascual-Montano, A. (2006). Integrated analysis of gene expression by association rules discovery, *BMC Bioinformatics* **7**(1): 54.

Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J. and Pascual-Montano, A. (2007). Genecodis: A web-based tool for finding significant concurrent annotations in gene lists, *Genome Biology* **8**(1): R3.

Eisen, M., Spellman, P., Brown, P. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns, *Proceedings of the National Academy of Sciences of the United States of America* **95**(25): 14863–14868.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery: An overview, *in* U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, American Association for Artificial Intelligence, Menlo Park, CA, pp. 1–34.

Fürnkranz, J. (1999). Separate-and-conquer rule learning, *Artificial Intelligence Review* **13**(1): 3–54.

Fürnkranz, J. and Flach, P. (2005). Roc'n' rule learning—Towards a better understanding of covering algorithms, *Machine Learning* **58**(1): 39–77.

Greco, S., Pawlak, Z. and Słowiński, R. (2004). Can Bayesian confirmation measures be useful for rough set decision rules?, *Engineering Applications of Artificial Intelligence* **17**(4): 345–361.

Greco, S., Słowiński, R. and Stefanowski, J. (2007). Evaluating importance of conditions in the set of discovered rules, *RSFDGrC '07: Proceedings of the 11th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Toronto, Ontario, Canada,* pp. 314–321.

Gruca, A. and Sikora, M. (2009). Ontological description of gene groups by the multiattribute statistically significant logical rules, *in* S. Safeeullah (Ed.), *Engineering the Computer Science and IT*, INTECH, Vukovar, pp. 277–303.

Gruca, A., Sikora, M., Chróst, Ł. and Polański, A. (2009). Rulego. Bioinformatical internet service system architecture, *Proceedings of the 16th Conference on Computer Networks. Communications in Computer and Information Sciences, Wisła, Poland*, pp. 160–167.

Grzymała-Busse, J., Stefanowski, J. and Wilk, S. (2005). A comparison of two approaches to data mining from imbalanced data, *Journal of Intelligent Manufacturing* **16**(6): 565–573.

Grzymała-Busse, J. and Ziarko, W. (2003). Data mining based on rough sets, *in* J. Wang (Ed.), *Data Mining: Opportunities and Challenges*, IGI Publishing, Hershey, PA, pp. 142–173.

Guillet, F. and Hamilton, H. (2007). *Quality Measures in Data Mining (Studies in Computational Intelligence)*, Springer-Verlag New York, Inc., Secaucus, NJ.

Hackenberg, M. and Matthiesen, R. (2008). Annotation-modules: A tool for finding significant combinations of multisource annotations for gene lists, *Bioinformatics* **24**(11): 1386–1393.

Hvidsten, T., Legreid, A. and Komorowski, H. (2003). Learning rule-based models of biological process from gene expression time profiles using gene ontology, *Bioinformatics* **19**(9): 1116–1123.

Iyer, V., Eisen, M., Ross, D., Schuler, G., Moore, T., Lee, J., Trent, J., Staudt, L., Hudson, J., Boguski, M., Lashkari, D., Shalon, D., Botstein, D. and Brown, P. (1999). The transcriptional program in the response of human fibroblasts to serum, *Science* **283**(5398): 83–87.

Kano, M., Morishita, Y., Iwata, C., Iwasaka, S., Watabe, T., Ouchi, Y., Miyazono, K. and Miyazawa, K. (2005). Vegf-a and fgf-2 synergistically promote neoangiogenesis through enhancement of endogenous pdgf-b-pdgfrbeta signaling, *Journal of Cell Science* **118**(Pt 16): 3759–3768.

Khatri, P. and Drăghici, S. (2005). Ontological analysis of gene expression data: Current tools, limitations, and open problems, *Bioinformatics* **21**(18): 3587–3595.

Maere, S., Heymans, K. and Kuiper, M. (2005). Bingo: A cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks, *Bioinformatics* **21**(16): 3448–3449.

Mata-Greenwood, E., Meyrick, B., Soifer, S., Fineman, J. and Black, S. (2003). Expression of vegf and its receptors flt-1 and flk-1/kdr is altered in lambs with increased pulmonary blood flow and pulmonary hypertension, *American Journal of Physiology: Lung Cellular and Molecular Physiology* **285**(1): L222–L231.

Michalski, R., Bratko, I. and Kubar, M. (1998). *Machine Learning and Data Mining: Methods and Applications*, John Wiley and Sons, New York, NY.

Midelfart, H. (2005a). Supervised learning in the gene ontology, Part I: A rough set framework, *in* J. Peters and A. Skowron (Eds.) *Transactions on Rough Sets IV*, Lecture Notes in Computer Science, Vol. 3700, Springer, Berlin/Heidelberg, pp. 69–97.

Midelfart, H. (2005b). Supervised learning in gene ontology, Part II: A bottom-up algorithm, *in* J. Peters and A. Skowron (Eds.) *Transactions on Rough Sets IV*, Lecture Notes in Computer Science, Vol. 3700, Springer, Berlin/Heidelberg, pp. 98–124.

Seghezzi, G., Patel, S., Ren, C., Gualandris, A., Pintucci, G., Robbins, E., Shapiro, R., Galloway, A., Rifkin, D. and Mignatti, P. (1998). Fibroblast growth factor-2 (fgf-2) induces vascular endothelial growth factor (vegf) expression in the endothelial cells of forming capillaries: an autocrine mechanism contributing to angiogenesis, *The Journal of Cell Biology* **141**(7): 1659–1673.

Sikora, M. (2006). *Rule Quality Measures in Creation and Reduction of Data Role Models*, Lecture Notes in Artificial Intelligence, Vol. 4259, Springer, Heidelberg, pp. 716–725.

Sikora, M. (2010). Decision rules-based data models using TRS and NetTRS—Methods and algorithms, *in* J., Peters and A. Skowron (Eds.), *Transactions on Rough Sets XI*, Lecture Notes on Computer Sciences, Vol. 5946, Springer, Berlin/Heidelberg, pp. 130–160.

Stefanowski, J. and Vanderpooten, D. (2001). Induction of decision rules in classification and discovery-oriented perspectives, *International Journal on Intelligent Systems* **16**(1): 13–27.

Marek Sikora was born in Poland in 1969. He received the M.Sc. degree in applied mathematics from Silesian University in 1993 and the Ph.D. degree in informatics from the Silesian University of Technology in 2002. His scientific interest is in rule induction and evaluation, machine learning, the application of intelligent systems in industry, biology and medicine. He is the author or a co-author of more than 50 scientific papers.

Aleksandra Gruca was born in Poland in 1980. She received the M.Sc. degree in computer science from the Silesian University of Technology in 2004 and the Ph.D. degree in informatics (bioinformatics focus) from the same university in 2009. Her scientific interest is in the application of machine learning algorithms and data mining techniques in the field of bioinformatics. She is the author or a co-author of more than 20 scientific papers.

Received: 10 January 2010
Revised: 1 June 2010