Biocybernetics and Biomedical Engineering
2008, Volume 28, Number 4, pp. 21–32

Analysis of GO Composition of Gene Clusters by Using Multiattribute Decision Rules

ALEKSANDRA GRUCA*

Institute of Informatics, Silesian University of Technology, Gliwice, Poland

In this paper, a novel method for characterizing the Gene Ontology (GO) composition of the gene clusters on basis of the decision rules is presented. The rules are expressed as logical functions of the Gene Ontology terms which are interpreted as binary attributes. A new method for evaluating the quality of decision rules based on statistical significance is developed. The presented approach is applied to the well-known data set and the results are compared with the results obtained by other authors.

K e y w o r d s: Gene Ontologies, decision rules, attributes, rules quality evaluation, gene clusters, DNA microarrays, bioinformatics, rough sets theory

1. Introduction

The advent of the DNA microarray technology provided a great opportunity to better learn and understand complicated biological fundaments that rule the world of living organisms. Gene expression profiles obtained with the use of the DNA microarray technology allows analyzing simultaneously thousands of genes in a single experiment [1, 2]. Experiments in the laboratory provide the abundance of the data on biological and molecular process that brings both chances and challenges. Nowadays, without specialized mathematical and informatics tools, interpretation of the microarray data is impossible.

Development of the DNA microarrays entailed development of many computational techniques and numerical algorithms — especially various data mining techniques appeared to be very useful and efficient in the field of analysis of specific biological data. Supervised classification allows to identify groups (clusters) of genes expressed differentially among different experimental conditions, while clustering

* Correspondence to: Aleksandra Gruca, Institute of Informatics, Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, e-mail: Aleksandra.Gruca@polsl.pl
Received 02 January 2008; Accepted 01 August 2008

algorithms are used to identify groups of genes coexpressed in sequences of experiments or in repeated experiments [3, 4]. Information on differential expression or coexpression of genes is helpful in predicting outcomes of further experiments or classifying biological samples on the basis of their gene expression profiles. Numerous literature items report various results of this kind [5–10].

Classification or clustering the results of microarray experiment is only one element of the DNA microarray analysis. Most of the algorithms that are commonly used for classification or clustering do not consider any biological knowledge that lies behind the processed data. Thus the obtained results need to be confronted with existing biological knowledge on genes, their classes and functions. Including the data on genes and their functions to the analysis paves the way towards extracting biological knowledge from the performed experiments. This aspect of the data analysis is often done by an expert in the field, frequently manually, which is time consuming for large-scale data. There is a strong motivation for supporting the experts work by developing computer systems for storing, organizing and presenting the relevant information. Even more desirable is developing data processing algorithms, including knowledge discovery, artificial intelligence and automatic reasoning for incorporating the biological knowledge of genes to the results of data clustering.

In several past years the Gene Ontology (GO) database [11] became one of the most popular and widely used source of information on the genes and their products. There are many tools, methodologies and algorithms for using the GO terms that are available for the researchers. GO browsers, such as [12] allow one to characterize the clusters of genes obtained in some experiments by computing and comparing frequencies of the GO terms. More advanced methods, such as those presented in [13–18], allow combining the GO terms with various information processing methods to explain experimental results. Papers [13, 14] are devoted to methods for building sets of genes sharing common biological functions and for using these sets for classification and clusterization. Arguments are given that using gene-set approach leads to improving the quality of classification of samples on the basis of their expression profiles. In [15] methods and a computer program for forming the clusters of genes on the basis of their GO terms are presented. In [16] a graph-oriented approach for extracting the most important terms for the description of clusters of genes is proposed. In the papers [17, 18] computational intelligence methods including rough sets theory are used to predict biological functions of genes on the basis of levels of their expressions observed in experiments.

In this paper, a novel method for characterizing the gene clusters on the basis of the decision rules expressed as logical functions of the GO terms, is presented. Decision rules are logical expressions that can be easily interpreted and understood by a human. To derive the decision rules the rough sets theory [19] is used which is a mathematical tool that allows to induce decision rules with good descriptive abilities. A method for assessment of the statistical significance of the decision rules that allows to evaluate the quality of obtained rules is also developed. The proposed approach is applied to the data set including expression profiles of the budding

yeast *Saccharomyces cerevisiae* published by Eisen et al. in [20]. For 10 clusters of the genes defined in [20] the GO characterization which follows from the proposed method is compared with the Eisen et al. description and the description obtained by Lee et al. [16] by the graph-oriented algorithm.

2. Methods

2.1. Formulation of the Problem of Characterization of Gene Clusters by GO Terms

Let there be defined a set U of genes whose probes are placed on the DNA microarray chip used in some experiment:

$$U = \{x_1, x_2, \ldots, x_N\}. \tag{1}$$

The set U consists of the genes, denoted as $x_1, \ldots, x_n$, which are distinguished by labels. Each gene is described by a set of the GO terms, such as "tRNA processing", "DNA binding", "translation" that can be interpreted as binary attributes assuming values from a set $\{0,1\}$.

In a more formal manner, the following mapping for each attribute may be defined:

$$a : x = a(x) \in \{0,1\}, \tag{2}$$

where $x \in U$ and denotes any gene, and a denotes an attribute from the whole set of attributes describing all genes belonging to the set U. The above formula is interpreted in the following way: if e.g., a corresponds to the GO term "DNA binding", then $a(x) = 1$ holds for all genes x which contain "DNA binding" among their GO terms. For other genes $a(x) = 0$.

Let us consider a cluster of genes $G \subset U$, $G = \{x_1, x_2, \ldots, x_P\}$ where each gene is annotated with the GO terms. The goal is to find the common biological meanings shared by genes composing the cluster and provide a method that allows describing the cluster G by the attributes (GO terms) of the genes $x_1, x_2, \ldots, x_P$.

The GO browsers, such as [12] allow comparing the composition of the GO terms between the analyzed gene clusters. The results of the analysis are presented in the form of the list of GO terms that are most significant in the investigated set of genes. Assuming the terminology used in this paper, the GO terms obtained by using a GO browser characterize the cluster G by single-attribute rules of the following form:

$$\textbf{IF } a(x) = 1 \textbf{ THEN } x \in U, \tag{3}$$

which are either true or false. Using the frequency of the GO terms corresponding to a in G one may determine if the gene function or product corresponding to that GO

term is representative for the analyzed cluster G. The more genes the rule is true for, the more important is the GO term that appears in the left side of the rule. Apart from counting the frequencies of the GO terms, the important part of the computations is to evaluate the quality of the created rules. This is most frequently done by assessing the statistical significances of the rules by applying the hypergeometric test, Fisher exact test or the chi-square test for independence [12].

In this paper, a new method for characterizing clusters of genes by GO terms, by using rules of the form more complex than (3), such as the one below:

$$\textbf{IF } a_1(x)=v_1 \text{ and } a_2(x)=v_2 \ \ldots \text{ and } a_R(x)=v_R \textbf{ THEN } x \in G\,, \tag{4}$$

where $v_1,\ldots,v_R$ are values from the set $\{0,1\}$is proposed. Some aspects of deriving rules of the form (4), evaluating their significance and using them for real data are described in the sequel.

2.2. Rough Sets

To compute the decision rules the rough sets theory [19] – a mathematical tool that allows to compute the decision rules of good, descriptive features is used. In the rough sets theory the data are represented in the form of the information system S which is a pair $S = (U, A)$, where U is non-empty set of objects called universe, and A is a set of attributes. The attribute $a \in A$ is a map $a: U \rightarrow V_a$, where V_a is the value set of the attribute a. Objects from the set U represent the investigated cases (genes in the experiment) and attributes are features describing these objects.

The concept of the information system may be extended to the concept of a decision table which is also a pair $DT = (U, A \cup \{d\})$ with one distinguished attribute $d \notin A$ called decision attribute with range D_d. One can see that the decision attribute determines a partition of the universe $U = \{X_{d1}, X_{d2}, \ldots, X_{dk}\}$ with respect to the value of decision attribute d_i. Each i-th set from this partition is called the i-th decision class.

A decision rule is a logical expression of the form:

$$\textbf{\textit{IF }} a_1 \in V_{a1} \textbf{\textit{ and }} a_2 \in V_{a2} \textbf{\textit{ and }} \ldots \textbf{\textit{ and }} a_n \in V_{an} \textbf{\textit{ THEN }} d = v\,, \tag{5}$$

where $v \in D_{\mathrm{d}}$, $\{a_1, a_2, \ldots, a_n\} \subseteq A$ and $V_{ai} \subseteq D_{ai}$, $i = 1,2,\ldots,n$. The left side of the rule is called the conditional part while the right side is called the decision part. An expression $a_n \in V_{an}$ is called a descriptor. The interpretation of the rule is intuitive – values of the attributes on the left-hand side of the rule should imply the value of the decision attribute. The object is recognized by the decision rule if its attributes values are concordant with the conditional part of the rule. The object supports the decision rule if it is recognized by the rule and the decision assigned to the object is the same as the decision pointed by the rule.

In the case of using the decision table *DT* for characterizing the results of the DNA microarray experiment, the set of conditional attributes A is represented by all GO terms, that describe the whole set of genes, and $V_a = \{0,1\}$. The universe U is a set of genes, whose probes were placed in the analyzed DNA microarray chip, and the value of the decision attribute d is an index of cluster, to which a given gene was assigned by some clustering algorithm. Using the above assumptions and the rough sets theory as a tool to induce decision rules, the rules are obtained with the conditional part, that consists of the conjunction of GO terms, and the decision part pointing to a cluster, which is considered to be best described by these GO terms.

2.3. Computing Decision Rules

Before inducing decision rules, the attributes that are irrelevant or redundant with respect to the knowledge represented by the whole data set should be eliminated from the information system. In the rough sets theory a minimal subset of attributes that preserves the same abilities of discerning the objects with respect to different decision classes, as the whole set of attributes, is called a minimal relative reduct [19]. Computation of the relative reduct is a very important step of preprocessing the data, because it allows to obtain the decision table, which contains the relevant information. Additionally, the induced decision rules are both more structured and compact and therefore easier to understand and interpret.

The problem of finding a minimal (relative) reduct has been proven to be NP-hard [21], thus a heuristic algorithm is always employed to find it. In [22] Nguyen and Nguyen proposed the algorithm for computing the approximate reduct of an information system for a large-scale data sets. In this paper the modified version of the algorithm that makes it possible to compute approximate relative reduct for adecision system [23] was used.

Having the decision table reduced, the sequential covering algorithm [24] was applied to induce decision rules. The sequential covering method involves learning one rule for a given object. After the rule induction, the whole set of the objects is searched and all the objects covered by that rule are removed from the processed data. The *learn-one-rule* method is based on the idea that for any object the set of attributes from the relative reduct determines its decision class. Assuming that a relative reduct for a decision table is given, the algorithm for computing the set of decision rules using the sequential covering method is given by the following pseudocode:

```
input: DT = (U,A∪{d}) , R ∈ RED_DT(A, d) – set of relative reducts
output: RUL(DT) – set of decision rules for DT
begin
    RUL(DT) = ∅
    while U ≠ ∅
        r := ⋀_{a_i∈R} a_i = a_i(u)→d = d(u)      /* create a decision rule r  */
```

```
        if ({r} ∉ RUL(DT) ) then RUL(DT) = RUL(DT)∪{r}
        U=U\[RUL(DT)]
    end while
end
```

where [RUL(*DT*)] denotes the coverage of the decision system which is a set of decision rules such that for each object $u \in U$ there is at least one rule supporting that object.

2.4. Evaluating the Quality of the Decision Rules

After computing the set of decision rules, the next step is to evaluate the quality of the obtained decision rules. There are many well-known measures that can be used to asses the significance of the rules [25]. A decision rule is statistically significant if the null hypothesis of purely random composition of the sets of genes recognized and supported by the rule may be rejected. Statistical significance of the rules confirms a non-random composition of gene clusters and encourages to extract biological conclusions from the obtained results. Below, a novel method of evaluating statistical significance of the decision rules, based on the conditional hypergeometic distribution, is derived.

A common method to verify the statistical significance of the decision rule D involves comparison of the attributes (decision rules) in gene sets G_i and $U \setminus G_i$. For every decision rule D one can form a contingency table:

Table 1. Contingency table describing the application of the decision rule D to partition of the universe U into clusters G_i and $U \setminus G_i$

Decision rule D	G_i	$U \setminus G_i$
True	N_{GT}	N_{UT}
False	N_{GF}	N_{UF}

where N_{GT} is the number of the genes that support the decision rule, N_{UT} is the number of the genes that recognize the rule but do not support it, and N_{GF}, N_{UT} are the numbers of the genes that do not recognize decision rules in the decision class and in its complement respectively.

Assuming genes in G_i and $U \setminus G_i$ as having two different "colors", the null hypothesis is stated as "the decision rule is color blind". Under the null hypothesis the probability of obtaining specific configuration of N_{GT}, N_{UT}, N_{GF}, N_{UF} follows the hypergeometric distribution:

$$p(N_{GT}, N_{UT}, N_{GF}, N_{UF}) = \frac{\binom{N_{GT}+N_{GF}}{N_{GT}}\binom{N_{UT}+N_{UF}}{N_{UT}}}{\binom{N_{GT}+N_{UT}+N_{GF}+N_{UF}}{N_{GT}+N_{UT}}}, \tag{6}$$

which, in the case of the testing for the overrepresentation of N_{GT}, leads to the following *p*-value of the hypergeometric test [26]:

$$p_H(N_{GT}, N_{UT}, N_{GF}, N_{UF}) = \sum_{k=1}^{N_{UT}} p(N_{GT} + k, N_{UT} - k, N_{GF}, N_{UF}) . \quad (7)$$

However, when multiattribute rules are applied to real data, it often happens, that the number of genes recognizing the rule (N_{GT} + N_{UT}) is very small. The extreme case of only one gene supporting and recognizing the rule (N_{UT} = 0 and N_{GT} = 1) is often encountered in the data. If the formula (7) is applied to such example, the very low *p*-value may be obtained, yet there is nothing extraordinary in the random choice of the single gene.

For the purpose of obtaining the satisfactory statistical model for small number of genes recognizing the computed rules, the conditional probability that the decision rule *D* has already classified one gene as gene belonging to G_i is proposed here. This conditional probability is given by the expression:

$$p^c(N_{GT}, N_{UT}, N_{GF}, N_{UF}) = p(N_{GT} - 1, N_{UT}, N_{GF}, N_{UF}) , \quad (8)$$

where p(.) is given by (6) and the *p*-value of the corresponding conditional hypergeometric test is:

$$p_H^c(N_{GT}, N_{UT}, N_{GF}, N_{UF}) = \sum_{k=1}^{N_{UT}} p^c(N_{GT} + k, N_{UT} - k, N_{GF}, N_{UF}) . \quad (9)$$

When (8) − (9) is used to the case N_{UT} = 0 and N_{GT}= 1 one may obtain $p_H^c = 1$ which is in accordance with intuition.

3. Data and Results

The presented method was applied to the data set published in 1998 by Eisen et al. in [20]. The time−course expression profiles of the budding yeast *Saccharomyces cerevisiae* genes were measured during several biological experiments with use of the DNA microarrays. Eisen et al. applied the hierarchical clustering algorithm to the obtained DNA microarray results and successfully revealed the existence of the clusters including genes of the very similar biological functions.

To verify the method described in the previous sections, the data set including top 10 gene clusters from Eisen et al. experiment was created. This data set included 274 genes described by the 142 different GO terms from the biological process ontology. The GO terms used for annotations were extracted from the *Saccharomyces cerevisiae* (SGD) GO annotations file that is available on the Gene

Ontology Consortium website (www.geneontology.org). In this paper the version from 11/03/2007, revision 1.1365 was used. The file included 6476 genes annotated with 3054 different GO terms, including 1326 different GO terms from biological process ontology.

A decision table of 274 objects and 142 attributes where for each gene its attributes values were located at the intersections of the row corresponding to that gene and columns representing the attributes was created. Each attribute assumed the value "1" if the gene was described by the GO term corresponding to that attribute and "0" otherwise. Average number of the GO terms describing one gene was about four, so the created decision table was a sparse matrix. Then, the relative reduct which led to the reduction of the decision table to 62 attributes was computed. Using the sequential covering algorithm based on relative reduct, the 96 decision rules were obtained. For each decision rule its statistical significance was determined by computing conditional p-value using the formula (9). The obtained statistical significances of the rules were examined to reject these with the p-value greater than 0.1. Finally the set of 15 statistically significant decision rules covered 166 genes from the Eisen data set.

We compared the results obtained by our method to the results from two papers: the paper by Eisen et al. [20] where ten clusters were described by an expert and to the paper by Lee et al. [16] where significant biological features of gene clusters were obtained from the GO terms on the basis of the graph modeling algorithm.

The results of comparison are presented in Table 2. The method proposed in [16] by Lee et al., besides discovering GO terms for clusters, also leads to computing values that can be interpreted as measures of the quality of the obtained information (the *AverPd* measure). For the case of applying the proposed method the p-value of the statistical test based on conditional hypergeometric distribution (9) is reported. Table 2 also includes the level of the obtained GO terms given by the longest path from the root to that term – in other words, the level is the maximal number of edges between the root and that GO term plus one (+1). This parameter is related to the specificity of the discovered biological knowledge. The "biological process" term – the root of the biological process ontology assumes the level one.

4. Conclusions

The first conclusion is that the results obtained by the presented method are generally consistent with the results obtained by Lee et al. [16] and Eisen et al. [20]. The GO terms indicated by the computed rules are either exactly the same or close to the terms presented by other authors. The GO terms are close to each other if their distance on the GO DAG is small.

Table 2. Comparison of the GO terms obtained by our method with the description given by Eisen to its clusters and the best GO terms obtained by Lee presented in his paper

Cluster No.	Eisen description	Lee GO terms (*AverPd*) (level)	Go terms obtained by our method (conditional *p*-value) (level)
1	spindle body assembly and function	microtubule nucleation (72.0) (lev.8)	– microtubule nucleation, (0.0042) (lev.8) – axial cellular bud site selection, (0.0042)(lev.10) – protein ubiquitination (0.079)(lev.9) – regulation of cyclin-dependent protein kinase activity *AND* S phase of mitotic cell cycle (0.079) (lev.8)
2	preoteasome	ubiquitin-dependent protein catabolic process (0.0)(lev.10)	– ubiquitin-dependent protein catabolic process (4e-34) (lev.10)
3	mRNA splicing	mRNA splicing (88.0)(lev.9)	– translation *AND* aerobic respiration (0.0996)(lev.6) – mRNA cleavage (0.0996)(lev.8)
4	glycolysis	glycolysis (47.0)(lev.10)	– glycolysis (2.9e-10) (lev.10)
5	mitochondrial ribosome	protein biosynthesis(43.0)	–
6	ATP synthesis	ATP synthesis coupled proton transport (54.0)(lev.11)	– ATP synthesis coupled proton transport(7.9e-11) (lev.11)
7	chromatin structure	chromatin assembly or disassembly (0.0)(lev.8)	– chromatin assembly or disassembly (2.9e-12) (lev.8)
8	ribosome and translation	protein biosynthesis (12.0) (lev.6)	– translation (3.9e-17)(lev.6) – telomere maintenance(lev.8) *AND* translation(lev.6) (0.0336) – ribosomal large subunit assembly and maintenance(lev.9) *AND* translation(lev.6) (0.0934)
9	DNA replication	DNA replication initiation (22.0)(lev.8)	– S phase of mitotic cell cycle (1e-5)(lev.8)
10	tricarboxylic acid cycle and respiration	metabolism (72.0)(lev.2)	– tricarboxylic acid cycle (4e-5) (lev.8)

As already mentioned, the *AverPd* is a measure that allows to assess the clustering quality – if the value of *AverPd* is relatively small, the cluster can be regarded as biologically well-clustered in the GO space [16]. If the *p*-value of the obtained rules is compared with the *AverPd* measure proposed by Lee et al. in [16] the results are also similar.

One can notice that the presented method has some features that make the obtained results more specific than those presented in [16] and [20]. This is seen by contemplating the GO levels in different columns of Table 2. The method proposed in this paper leads to the biggest values of GO levels, corresponding to the most specific term.

However, the most important property of the proposed method is that apart from selecting of the significant GO terms, which are included in the conditional part of the rule, the decision rule also indicates small groups of genes close related to each other. The genes that support the decision rules are the small sets of the genes that have very similar biological function. These small gene sets are easy to analyze. For example the cluster one consists of eleven genes related to the microtubules. One of the obtained decision rules:

IF “axial cellular bud site selection” = 1 THEN class is 1

is supported by three genes: BNR1, CDC10, CDC3 and from all genes composing the cluster only these three are involved into a process of bud neck emergence [27].

The presented method may be used to indicate small groups of genes and characterize more precisely the biological features of the gene cluster. These small groups of genes supporting the statistically significant decision rules can be presented to biologists as a specific genes that probably serve some important biological function and are more interesting than other genes in the analyzed cluster.

References

1. Baldi P., Hatfield G.W.: DNA Microarrays and Gene Expression. Cambridge University Press, Cambridge 2002.
2. Speed T. (ed.): Statistical Analysis of Gene Expression Micoroarray Data. Interdisciplinary Statistics. Chapman & Hall/CRC, Boca Ration 2003.
3. Allison D.B., Gadbury G., Heo M., Fernandez J., Lee C.K., Prolla T.A., Weindruch R.: A Mixture Model Approach for the Analysis of Microarray Gene Expression Data. Comput. Statist. Data Anal., 2002, 39, 1–20.
4. Storey J.D., Tibshirani R.: Statistical Significance for Genomewide Studies. Proc. Natl. Acad. Sci. USA, 2003, 100, 9440–9445.
5. Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., Boldrick J.C., Sabet H., Tran T., Yu X. et al.: Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling. Nature, 2000, 403, 503–511.
6. Beer D.G., Kardia S.L.R., Huang C.C., Giordano T.J., Levin A.M., Misek D.E., Lin L., Chen G.A., Gharib T.G. et al.: Gene Expression Profiles Predict Survival of Patients with Lung Adenocarcinoma. Nat. Med., 2002, 8, 816–824.

7. Cho R.J., Campbell M.J., Winzeler E.A., Steinmetz L., Conway A., Wodicka L., Wolfsberg T.G., Gabrielian A.E., Landsman D., Lockhart D.J., Davis R.W.: A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. Mol. Cell., 1998, 2, 65–73.
8. Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.: Molecular Classification of Cancer Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 1999, 286, 531–537.
9. Jarząb B., Wiench M., Fujarewicz K., Simek K., Jarząb M., Oczko-Wojciechowska M., Wloch J., Czarniecka A., Chmielik E., Lange D., Pawlaczek A., Szpak S., Gubala E., Siwerniak A.: Gene Expression Profile of Papillary Thyroid Cancer: Sources of Variability and Diagnostic Implications. Cancer Res., 2005, 65, 1587–1597.
10. Rhodes D.R., Yu J., Shanker K., Deshpande N., Varambally R., Ghosh D., Barrette T., Pandey A., Chinnaiyan A.M.: ONCOMINE, A Cancer Microarray Database and Integrated Data Mining Platform. Neoplasia, 2004, 6, 1–6.
11. Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M. et al.: Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium. Nature genetics, 2000, 25, 25–29.
12. Al-Shahrour F., Minguez P., Vaquerizas J.M., Conde L., Dopazo J.: BABELOMICS: A Suite of Web Tools for Functional Annotation and Analysis of Groups of Genes in High-Throughput Experiments. Nucleic Acid Research, 2005, 33, W460–W464.
13. Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A. et al.: Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. Proc. Natl. Acad. Sci. USA, 2005, 102, 15545–15550.
14. Svensson J.P., Stalpers L.J., Esveldt-van Lange R.E., Franken N.A., Haveman J., Klein B., Turesson I., Vrieling H. and Giphart-Gassler M.: Analysis of Gene Expression Using Gene Sets Discriminates Cancer Patients with and without Late Radiation Toxicity, PLOS Med., 2006, 3, 1905–1914.
15. Lee I.Y., Ho J.M., Chen M.S.: CLUGO: A Clustering Algorithm for Automated Functional Annotations Based on Gene Ontology. Proc. of the Fifth IEEE Int. Conf. on Data Mining, Houston, Texas, 2005, 705–708.
16. Lee S.G., Hur J.U., Kim Y.S.: A Graph-Theoretic Modeling on GO Space for Biological Interpretation of Gene Clusters. Bioinformatics, 2004, 20, 381–388.
17. Midelfart H., Komorowski H.J.: A Rough Set Framework for Learning in a Directed Acyclic Graph. Rough Sets and Current Trends in Computing, 2002, 144–155.
18. Wang H., Azuaje F., Bodenreider O., Dopazo J.: Gene Expression Correlation and Gene Ontology--Based Similarity: an Assessment of Quantitative Relationships. Proc of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, La Jolla, California, 2004, 25–31.
19. Pawlak Z.: Rough Sets: Theoretical aspects of reasoning about data. Kluwer Academic Publisher, Dordrecht 1991.
20. Eisen M.B., Spellman P.T., Brown P.O., Botstein D.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 1998, 95, 14863–14868.
21. Skowron A., Rauszer C.: The Discernibility Matrices and Functions in Information Systems, in: Słowinski, R. (Ed.), Intelligent Decision Support: Handbook of Applications and Advances to Rough Sets Theory, Kluwer Academic Publisher, Dordrecht 1992, 331–362.
22. Nguyen S.H., Nguyen H.S.: Some Efficient Algorithms for Rough Set Methods. Proc. of the Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems, Granada, Spain, 1996, 1451-145.
23. Gruca A.: A Rule-Based Characterization of Gene Clusters. Proc. of the International Multiconference on Computer Science and Information Technology, Wisla, Poland, 2007, 2, 93–102.
24. Mitchell T.: Machine Learning. McGraw-Hill, New York 1997.

25. Bruha I.: Quality of Decision Rules: Definitions and Classification Schemes for Multiple Rules, in: Nakhaeizadeh G., Taylor C.C. (Eds.), Machine Learning and Statistics, The Interface, John Wiley and Sons 1997.
26. Rice J.A.: Mathematical Statistics and Data Analysis. 2nd edn. Duxbury Press, Belmont 1995.
27. Kikyo M. et al.: An FH domain-containing Bnr1p is a multifunctional protein interacting with a variety of cytoskeletal proteins in Saccharomyces cerevisiae, Oncogene, 1999, 18, 7046–7054.