



A Rule-Based Characterization of Clusters of Genes

Aleksandra Gruca

Silesian University of Technology,
ul. Akademicka 16, 44-100 Gliwice, Poland
Aleksandra.Gruca@polsl.pl

Abstract. This paper presents a method of using the attribute analysis and the rough sets theory for describing and analyzing the Gene Ontology (GO) composition of clusters of genes obtained in DNA microarray experiments. GO terms are understood as attributes of genes and gene clusters are characterized by decision rules related to attributes. A modification of the known algorithm for computing the decision rules for the information system is proposed, which makes it suitable for large-size problems encountered in the analysis of DNA microarray data. Additionally methods are developed for the assessment of the statistical significance of the computed rules. Presented approach is used for DNA microarray data obtained in experiments of measuring transcriptome response of cells to ionizing radiation. The results of computations are compared with those obtained with the use of GO browsers.

1 Introduction

The computational techniques that accompany the development in DNA microarray experiments [1, 2] are most often focused on tasks of (i) identifying genes or groups of genes expressed differentially between different experimental conditions, or (ii) identifying groups (clusters) of genes coexpressed in sequences of experiments [3, 4]. Information on differential expression or coexpression of genes is helpful in predicting outcomes of further experiments or in classifying biological samples on the basis of their gene expression profiles.

Another element of the interpretation of DNA microarray data includes confronting information on levels of gene expressions with the existing biological knowledge on genes, their classes and functions. Including that knowledge to analysis leads to robustifying the conclusions of the study against statistical artefacts, and also supports extracting biological knowledge from performed experiments.

The most widely used source of data on functions of genes is the Gene Ontology (GO) database [5] which provides standardized and structured vocabulary that describes genes and their products. The GO database is represented as three disjoint directed acyclic graphs describing biological process, molecular function and cellular component. Each has a single root and a thousands of depended nodes. Upper nodes represent more general concepts and as the DAG is

traversed towards deeper levels, the definitions are more and more precise. GO browsers, such as [6] allow one to characterize the clusters of genes obtained in some experiment by computing and comparing frequencies of GO terms. Many other studies [7–12] develop various systematic methods to use GO terms in conjunction with various information processing methods to obtain biological interpretation of experimental results.

Method proposed in this paper allows characterizing the composition of clusters of genes on the basis of the decision rules expressed as logical functions of the GO terms. The algorithm for computing logical decision rules on the basis of the rough sets theory [13] is elaborated. A modification of the published method [14] for computing the decision rules in the information system is introduced that makes it possible to use the algorithm for large-size problems. There are also methods developed for the assessment of the statistical significance of the computed rules. Presented approach is used for DNA microarray data obtained in the experiment of measuring transcriptome response of cells to ionizing radiation. The results of computations are compared to those obtained by using of one the popular GO browser.

2 The proposed approach

Let there be defined the set U which is a set of all genes whose probes are placed on the DNA microarray chips used in the analyzed experiment,

$$U = \{x_1, x_2, \dots, x_N\}. \quad (1)$$

In the above x_1, \dots, x_N denote genes, distinguished by numbers or labels, and N is the number of genes. GO terms of genes, such as “regulation of transcription”, “ribosome” or “tRNA processing” are interpreted as the binary attributes, in other words functions which assign to each of the genes, the value 0 or 1,

$$a: x \rightarrow a(x) \in \{0, 1\}. \quad (2)$$

In the above x denotes one of the genes and a denotes one of the attributes. So, for example if a corresponds to “tRNA processing”, then $a(x) = 1$ holds for all genes x which contain “tRNA processing” among their GO terms. For all other genes $a(x) = 0$.

Considering a cluster of genes, $G \subset U$, $G = x_1, x_2, \dots, x_P$, the problem is whether and how G can be described by the attributes of the genes x_1, x_2, \dots, x_P ?

GO browsers, for example [6], typically describe the gene composition of the cluster G or compare the gene composition of two clusters of genes G_1 and G_2 by computing frequencies of GO terms (attributes). Using terminology proposed in this paper GO browsers characterize G by using single-attribute rules of the following form:

$$\text{IF } a(x) = 1 \text{ THEN } x \in G, \quad (3)$$

which are either true or false. By frequency of GO term corresponding to a in G one can understand how often (3) is true for genes $x \in G$.

This paper proposes a method that allows characterizing clusters of genes by rules of the form more complex than (3), such as the one below:

$$\text{IF } a_1(x) = v_1 \text{ and } a_2(x) = v_2 \dots \text{ and } a_R(x) = v_R \text{ THEN } x \in G. \quad (4)$$

In the above v_1, \dots, v_R are values from the set $\{0, 1\}$.

3 The algorithm for computing decision rules

In this section some necessary terminology from the rough sets theory is introduced that is essential to understand the idea of the presented method. An information system S is a pair $S = (U, A)$, where U is called a universe and A is a set of attributes. Objects from U represent the investigated cases (genes in the experiment) and attributes are features describing these objects. The attribute $a \in A$ is a map $a : U \rightarrow V_a$, where V_a is the value set of the attribute a .

Decision table $DT = (U, A \cup \{d\})$ is another system of objects including a universe U , a set of attributes A called conditional attributes and a distinguished attribute $d \notin A$, called the decision attribute. The decision attribute determines the partition $\{X_{d1}, \dots, X_{dk}\}$ of the U with respect to the value of the decision attribute d_i . With any subset $B \subseteq A$ one can associate the equivalence relation called B-indiscernibility relation, defined as:

$$IND(B) = \{x, y \in U \times U : \forall a \in B \quad a(x) = a(y)\}. \quad (5)$$

The objects from U satisfying the relation $IND(B)$ are indiscernible from each other with respect to attributes from B .

A decision rule is a logical expression of the form:

$$\text{IF } a_1 \in V_{a1} \text{ and } a_2 \in V_{a2} \text{ and } \dots \text{ and } a_n \in V_{an} \text{ THEN } d = v. \quad (6)$$

where $v \in D_d$, $\{a_1, a_2, \dots, a_n\} \subseteq A$ and $V_{ai} \subseteq D_{ai}$, $i = 1, 2, \dots, n$. The interpretation of the decision rule is intuitive – values of the attributes on the left-hand side of the rule should imply the value of the decision attribute.

Given object recognizes the rule if its attributes values satisfy the premise of the rule. The object supports the rule if it recognizes the rule and the decision given to the object is the same as a decision from the right side of the rule.

3.1 Feature reduction

Usually among all attributes that describe objects there are attributes irrelevant or redundant with respect to knowledge represented by the whole data set. The process of removing these attributes is called feature selection or reduction. Apart from reduction of the time of further computations and reduction of the amount of achieved data the most important benefit of feature selection is that obtained knowledge is both more structured and more compact and therefore easier to understand. In the rough sets theory, the process of feature selection

is considered as a computation of the subset of attributes called the reduct. For a given information system $S = (U, A)$ and its indiscernibility relation $IND(A)$ with respect to the whole attribute set A , the reduct is a minimal subset $B \subseteq A$ such that $IND(A) = IND(B)$. In other words, the reduct is a minimal subset of attributes that preserves the same abilities of discerning objects as the whole attribute set A . More than one reduct may exist for an information system.

Concerning the decision table $DT = (U, A \cup \{d\})$ there is a notion of the relative reduct, that is a minimal subset $B \subseteq A$ such that $IND(B) \subseteq IND(d)$. The set of all relative reducts in A is denoted by $RED_{DT}(A, d)$. The problem of finding a minimal reduct has been proven to be NP-hard [15], hence, the heuristic algorithm is always employed to find an approximate reduct.

In [14] Nguyen and Nguyen proposed an algorithm for computing an approximate reduct of the information system, based on the equivalence relation, suitable for large data-size problems. Let $S = (U, A)$ be an information system and $X \subseteq U$. For any attribute $a \in A$ there is an indiscernibility relation $IND^X(a)$ over objects from X determined by an attribute a . This relation determines the partition of the set X into equivalence classes denoted by $[IND^X(a)]$. Assuming that there is a partition of the X such that $[IND^X(a)] = \{X_1, \dots, X_m\}$ and cardinalities of the sets X, X_1, \dots, X_m are respectively x, x_1, \dots, x_m . The number of pairs of objects from X discerned by an attribute a may be computed by the following formula:

$$W^X(a) = \frac{\sum_{i \neq j} x_i x_j}{2} = \frac{x^2 - \sum_{i=1}^m x_i^2}{2}. \quad (7)$$

The number of pairs of objects discerned by any attribute can be computed using the following pseudocode [14]:

Require: an information system $S = (U, A)$
Ensure: a minimal reduct $R, L = [IND^U(R)]$
 $R = \emptyset, L = \{U\}$
repeat
 for all $a \in A$ **do**
 for all $X_i \in L$ **do**
 search for $[IND^{X_i}(a)]$
 count $W^{X_i}(a)$
 end for
 $W^U(a) = W^{X_1}(a) + \dots + W^{X_m}(a)$
 end for
 choose attribute a with max value of $W^U(a)$
 $A = A \setminus \{a\}, R = R \cup \{a\}$
 $L = [IND^{X_1}(a)] \cup \dots \cup [IND^{X_m}(a)]$
until $A = \{\emptyset\}$ or $W^U(a) = 0$

For the decision table $DT = (U, A \cup \{d\})$, the important information from the feature selection perspective is the information about attributes that discern

objects only from a different decision classes. In other words there is no need to discern objects with the same value of the decision attribute. Hence, this paper introduces modification to the original algorithm that allows computing the reduct relative to the given decision table. The algorithm searches for the attribute that discerns the maximal number of pairs of objects, but only from the different decision classes.

Assuming there is a subset $X \subseteq U$, two decision classes that split the X into two subsets X_{d_i} and X_{d_j} and numbers of objects belonging to these classes are respectively x_{d_i} and x_{d_j} , all pairs of objects discerned by these decision classes can be simply computed by multiplication of x_{d_i} by x_{d_j} . Thus the attribute discerning the most objects from different decision classes can be computed as follows:

$$W^X(a) = \frac{\sum_{i \neq j} x_i x_j}{2} - \frac{\sum_{i \neq j, d(x_i) \neq d(x_j)} x_i x_j}{2}. \quad (8)$$

Assuming that $card(U) = n$ and $card(A) = k$ the time complexity of the whole algorithm is $O(k^2 n \log n)$. The space complexity of the algorithm is $O(n)$.

3.2 Sequential covering algorithm for rules generation based on the relative reduct

After the relative reduct is computed, the next step of the algorithm involves generating decision rules with the use of the sequential covering method. The method is based on a simple idea: learn one rule for a given object, search for the objects recognized by that rule are remove them from the processed data. The decision rules are induced using the properties of minimal relative reduct such that for any $R \in RED_{DT}(A, d)$ the set of reduct attributes determines a decision value d .

Let $DT = (U, A \cup \{d\})$ be a decision table. The coverage of the decision system denoted as $[RUL(DT)]$ is a set of decision rules that for each object $u \in U$ there is at least one rule supporting that object. Computing the set of decision rules can be described by the following pseudocode:

Require: $DT = (U, A \cup \{d\})$, $R \in RED_{DT}(A, d)$

Ensure: $RUL(DT)$ - set of decision rules for S

$RUL(S) = \emptyset$

while $U \neq \emptyset$ **do**

$r = \forall_{a_i \in R} a_i = a_i(u) \rightarrow d = d(u)$ {create a decision rule r }

if $r \notin RUL(DT)$ **then**

$RUL(DT) = RUL(DT) \cup \{r\}$

$U = U \setminus [RUL(DT)]$

end if

end while

4 Statistical significance of the decision rules

Having the clustering of genes obtained in DNA microarray experiment, and the decision rules, the next task is to assess the statistical significance of the rules. Decision rules are statistically significant if the null hypothesis of purely random composition of clusters of genes can be rejected. Statistical significance of decision rules confirms a non-random composition of gene clusters and supports drawing further biological conclusions.

A common method to verify the statistical significance of the decision rule D , given the gene cluster G_i involves comparison of the attributes (decision rules) in gene sets G_i and $U \setminus G_i$. For every rule D there is 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} and N_{UT} denote numbers of genes, respectively in G_i and $U \setminus G_i$, $N_{GT} + N_{UT}$ is the number of genes recognizing the rule D and N_{GT} is the number of genes supporting the rule D . N_{GF} and N_{UF} denote numbers of genes, for which the rule is false.

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

$$p(N_{GT}, N_{UT}, N_{GF}, N_{UF}) = \frac{\binom{N_{GT} + N_{GF}}{N_{GT}} \binom{N_{UT} + N_{UF}}{N_{UT}}}{\binom{N_{GT} + N_{UT} + N_{GF} + N_{UF}}{N_{GT} + N_{UT}}}. \quad (9)$$

which, in the case of testing for the overrepresentation of N_{GT} , leads to the following p-value of the Fisher exact test [16]:

$$p_{FET}(N_{GT}, N_{UT}, N_{GF}, N_{UF}) = \sum_{k=1}^{N_{UT}} p(N_{GT} + k, N_{UT} - k, N_{GF}, N_{UF}). \quad (10)$$

Fisher exact tests like the above, or their approximations, chi-square tests for independence, are typically applied in the GO browsers. Since tests are performed for large number of rules, corrections for multiple testing are applied [3, 4].

The construction of the contingency table, in Table 1, is the same for the single attribute rules (3) and for the multi-attribute rules (4). However, there is a quantitative difference between single attribute rules (3) and multi-attribute

rules (4). For multi-attribute rules the number of genes recognizing the rule is typically very small. The extreme case $N_{UT} = 0$ and $N_{GT} = 1$ is often encountered in the data. In such case the calculated p-value of the test in (10) can be very low, yet there is nothing improbable in the random choice of a single gene.

In order to obtain a statistical model for the small number of genes recognizing the computed rules, more realistic than (9)–(10), there is considered conditional probability of obtaining the configuration N_{GT}, N_{UT}, N_{GF} and N_{UF} given that the decision rule D has already classified one gene as belonging to G_i , given by the expression:

$$p^c(N_{GT}, N_{UT}, N_{GF}, N_{UF}) = p(N_{GT} - 1, N_{UT}, N_{GF}, N_{UF}), \quad (11)$$

where $p(\cdot)$ is given by (10) and the p-value of the corresponding variant of the Fisher exact test is:

$$p_{FET}^c(N_{GT}, N_{UT}, N_{GF}, N_{UF}) = \sum_{k=1}^{N_{UT}} p^c(N_{GT} + k, N_{UT} - k, N_{GF}, N_{UF}). \quad (12)$$

When (11)–(12) is used to the case $N_{UT} = 0$ and $N_{GT} = 1$, $p_{FET}^c = 1$ is obtained which is in accordance with intuition.

5 Application to the DNA microarray data set

The elaborated methodology was applied to the data set obtained in the DNA microarray experiment related to observing the transcriptome dynamic response of the human melanoma cell line ME45 to ionizing radiation [17]. Gene expression profiles of the cell line were measured with the use of the Affymetrix chip U133A [18] containing approximately 22000 spots corresponding to known human genes. Based on the expression profiles of the cell line measured at several time instants genes on the U133A Affy chip were divided into 48 gene clusters [19].

5.1 Computing the decision rules

The annotations of Affymetrix DNA microarray probe sets with GO terms are provided and updated on regular basis [20] and are available on the NetAffx web site: <http://www.affymetrix.com>. For the computations the annotation file from March, 2007 was used. GO terms belonging to three ontologies: biological process, molecular function and cellular component were used to describe each gene. Some of the genes had no GO term assigned so they were removed from further investigations. Having removed these genes the decision table was generated including 12050 objects (genes) described by the 5171 attributes (GO terms). For each gene its attributes values were set in the following way: if the gene had a GO term assigned the attribute value corresponding to that GO term was set to 1 otherwise the value was set to 0. Then the minimal relative reduct was computed that decreased the number of attributes to 627.

Using the sequential covering algorithm based on relative reduct 11522 decision rules were obtained. For each rule the statistical significance was calculated and only 46 rules with p-value lower than 0.05 were considered as rules with good descriptive features that could be used for further description of gene clusters.

5.2 Comparison of the statistical significance of decision rules obtained by using two different methods

The statistical significance for 48 gene clusters G_1, G_2, \dots, G_{48} of the partitions $G_i, U \setminus G_i$ assessed by using the GO internet tool Fatigo+ [6] was compared with the statistical significance of multi-attribute decision rules presented in the previous subsection. For the GO term the GO level defines the distance from that term to the most upper node of a DAG. For three GO levels: 3, 5, and 7 the lowest (most significant) value of all statistical tests performed was recorded. Statistical tests performed by the applied internet tool were based on expressions (9)–(10). Multiple testing corrections were not taken into account. For multi-attribute rules (4), expressions (11)–(12) were used for computing p-values of the statistical test and, as previously, for each of the partitions $G_i, U \setminus G_i$ the lowest p-value was recorded. Figure 1 presents comparison of the statistical significances of the decision rules analyzed by the Fatigo+ internet tool and the decision rules obtained by presented method.

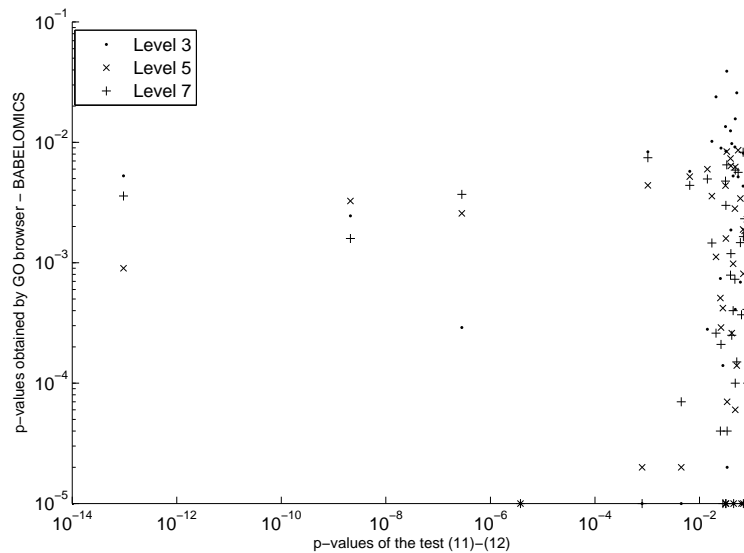


Fig. 1. Comparison of statistical significances of decision rules analyzed by the Fatigo+ internet tool on 3rd, 5th, 7th GO level and decision rules obtained by presented method.

6 Conclusions

The analysis of the contents of gene clusters performed by GO browsers [6] have been interpreted by defining a single-attribute rule (3) and assessing its statistical significance by (9)–(10). A characterization of clusters of genes by multi-attribute rules of the form (4) have been considered. An example of a multi-attribute rule computed on the basis of the analyzed data set [17] is:

```
a1304 = ‘‘binding’’ AND
a1312 = ‘‘iron ion binding’’ AND
a1316 = ‘‘protein binding’’ AND
a1950 = ‘‘transport’’ AND
a3485 = ‘‘oxygen binding’’ AND
a4710 = ‘‘metal ion binding’’ => class is 23 ,
```

(14)

recognized by 4 and supported by 3 genes.

Multi-attribute rules, such as (14), are typically recognized by only small number of genes. In order to assess their statistical significance a variant of the Fisher exact test based on probability (11) was employed conditioned on the event of one gene already supporting the rule.

Comparing statistical significances of single- and multi-attribute rules for the same gene clustering, leads to some conclusions. For one group of clusters, roughly half of the data, p-values of both tests are well correlated. For another group there is no correlation. This may suggest different mechanisms behind forming of the detected gene clusters.

Statistical significance of some of the obtained multi-attribute rules suggests the existence of the biological meaning behind them. Information properties of multi-attribute rules, small number of recognizing genes and many GO terms in one rule, provide a new tool for the characterization of gene clusters by their GO terms, which can be useful for biologists.

Acknowledgement. This paper was partially supported by the European FP6 grant, GENEPI, lowRT, Genetic Pathways for the Prediction of the Effects of Ionising Radiation: Low dose radiosensitivity and risk to normal tissue after Radiotherapy.

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 Microarray 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.* 39, 1–20 (2000).
4. Storey J. D., Tibshirani R.: *Statistical Significance for Genomewide Studies*, *Proc. Natl. Acad. Sci. USA.* 100, 9440–9445 (2003).

5. 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*. 25, 25–29 (2000).
6. Al-Shahrour F., Minguez P., Vaquerizas J. M., Conde L., Dopazo J.: *BABEL-LOMICS: A Suite of Web Tools for Functional Annotation and Analysis of Groups of Genes in High-Throughput Experiments*, *Nucleic Acid Research*. 33, W460–W464 (2005).
7. 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*. 102, 15545–15550 (2005).
8. Lee I. Y., Ho J. M., Chen M. S.: *vCLUGO: A Clustering Algorithm for Automated Functional Annotations Based on Gene Ontology*, In: *Proc. of the Fifth IEEE Int. Conf. on Data Mining*, pp 705–708. IEEE Computer Society, Washington, DC (2005).
9. Lee S. G., Hur J. U., Kim Y. S.: *A Graph-Theoretic Modeling on GO Space for Biological Interpretation of Gene Clusters*. *Bioinformatics*, 20, 381–388 (2004).
10. Midelfart H., Komorowski H. J.: *A Rough Set Framework for Learning in a Directed Acyclic Graph*, *Rough Sets and Current Trends In Computing*. 144–155 (2002).
11. Wang H., Azuaje F., Bodenreider O., Dopazo J.: *Gene Expression Correlation and Gene Ontology-Based Similarity: an Assessment of Quantitative Relationships*, In: *Proc. of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. 25–31 (2004).
12. Svensson J. P., Stalpers L. J., Esveldt-van Lange R. E., Franken N. A., Haveman J., et al.: *Analysis of Gene Expression Using Gene Sets Discriminates Cancer Patients with and without Late Radiation Toxicity*, *PLoS Med*. 3, e422 (2006).
13. Pawlak Z.: *Rough Sets: Theoretical aspects of reasoning about data*, Kluwer Academic Publisher, Dordrecht (1991).
14. Nguyen S. H., Nguyen H. S.: *Some Efficient Algorithms for Rough Set Methods*, In: *Proc. of the Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 1451–1456, Granada (1996).
15. Skowron A., Rauszer C.: *The Discernibility Matrices and Functions in Information Systems*, In: Slowinski R. (ed.): *Intelligent Decision Support: Handbook of Applications and Advances to Rough Sets Theory*, pp. 331–362. Kluwer Academic Publisher, Dordrecht (1992).
16. Rice J. A.: *Mathematical Statistics and Data Analysis*, 2nd edn. Duxbury Press, Belmont (1995).
17. Konopacka M., Rogolinski J., Herok R., Polanska J., Fujarewicz K., Hancock R., Rzeszowska-Wolny J.: *Radiation Induced Genetic Changes in Directly Exposed and Neighboring K562 Cells in Vitro*, Manuscript submitted for publication.
18. *Affymetrix, GeneChip Expression Analysis, Data Analysis Fundamentals*, Affymetrix Manual (2002).
19. Polanska J., Widlak P., Rzeszowska-Wolny J., Kimmel M., Polanski A.: *Gaussian Mixture Decomposition of Time-Course DNA Microarray Data*, In Deutsch A., Bruschi L., Byrne H., de Vries G., Herzog H. (eds.): *Cellular Biophysics, Development, Biomedicine, and Data Analysis*, series: Modeling and Simulation in Science, Engineering and Technology, pp. 351–359. Birkhäuser, Boston (2007).
20. Liu G., Loraine A. E., Shigeta R., Cline M., Cheng J., Valmeekam V., Sun S., Kulp D., Siani-Rose M. A.: *NetAffx: Affymetrix Probesets and Annotations*, *Nucleic Acids Res*. 31, pp. 82–86 (2003).