STUDIA INFORMATICA 2012
Volume 33 Number 2A (105)

Michał KOZIELSKI, Aleksandra GRUCA
Politechnika Śląska, Instytut Informatyki

APPLICATION OF BINARY SIMILARITY MEASURES TO ANALYSIS OF GENES REPRESENTED IN GENE ONTOLOGY DOMAIN

Summary. The work presents application of four binary similarity measures to analysis of Gene Ontology data. The measures are analysed and compared with semantic measure calculating term and gene similarity. Two kinds of experiments performed on two gene datasets show that binary similarity measures are valuable and interesting methods for the considered application.

Keywords: similarity measures, binary similarity measures, gene similarity, Gene Ontology clustering

ZASTOSOWANIE BINARNYCH MIAR PODOBIEŃSTWA DO ANALIZAY GENÓW REPREZENTOWANYCH W DZIEDZINIE ONTOLOGII GENOWYCH

Streszczenie. Artykuł przedstawia zastosowanie czterech binarnych miar podobieństwa do analizy danych z ontologii Gene Ontology. Miary są analizowane i porównywane z semantyczną miarą wyznaczającą podobieństwo genów na podstawie podobieństwa terminów ontologii. Przeprowadzone zostały dwa typy eksperymentów na dwóch zbiorach danych o różnej charakterystyce. Eksperymenty te pokazują, że binarne miary podobieństwa są wartościowymi i interesującymi metodami analizy dla opisywanego zastosowania.

Słowa kluczowe: miary podobieństwa, binarne miary podobieństwa, podobieństwo genów, grupowanie ontologii genowych, Gene Ontology

1. Introduction

Bioinformatics is currently one of the most intensively developing areas. A significant position within this area is occupied by gene function analysis. Years of research conducted on genes and gene products resulted in a knowledge represented among others by Gene Ontology (GO) database. GO provides an ontology of defined terms where each term represents gene product properties in three separate domains: biological process, molecular function and cellular component. When a new function of gene or gene product is discovered it may be annotated by the Gene Ontology terms describing that function, which makes GO a valuable source of knowledge. The knowledge represented by Gene Ontology may be also applied as an input to further analysis. It can be used e.g., as an additional expert knowledge gathered throughout the years and supporting an analysis of a new dataset [14, 15].

Clustering is one of the data mining methods that can be applied when gene analysis is performed. Clustering is particularly useful for analysis of gene expression values received from a microarray experiment [5, 9]. Such analysis can be supported by an expert knowledge in the form of Gene Ontology. In this case it is needed to combine the clustering of genes in two domains (expression values and Gene Ontology) [7].

The concept of similarity is fundamental for clustering process. Similarity can influence significantly the results of cluster analysis and its quality. While a clustering algorithm can be a general approach, the similarity measure applied should fit the characteristics of data very well. Gene Ontology is a specific type of data where similarity measures possible to apply were analysed [1, 6, 11, 12, 13] but still not exhaustively enough verified.

This work is a continuation of the research conducted in the field of Gene Ontology similarity measures [6, 11, 12]. The goal of this work is comparison of the methods determining the similarity of genes when they are described in Gene Ontology domain.

Similarity is a basic notion utilised by clustering algorithms. There are several similarity measures that can be applied to GO clustering [10, 13]. Some attempts to compare different similarity measures were performed in previous years [1, 6, 11, 12, 13]. However, in the opinion of the authors additional analysis and research is needed. The analysis presented in the work [1] is not clear to draw the convincing final conclusions. The results of the authors' research [6, 11, 12] show that traditional semantic similarity measures perform poorly when their correlation with gene similarity in gene expression domain is considered and when clustering quality based on these measures is considered. Therefore, the other than semantic measures are worth verifying.

The data that is analysed within the experiments is an ontology of defined terms where gene products can be annotated to the terms. Annotations are represented as a table containing binary values: 1 representing annotation of a gene to a term, 0 representing lack of the annotation. This type of data can be analysed by means of the methods dedicated to binary data such as binary similarity measures.

The contribution of this work is evaluation of binary similarity measures in application to Gene Ontology data analysis. Two aspects of such analysis are considered:

- correlation of GO based (binary) similarity with gene similarity in gene expression domain and
- quality of cluster analysis based on binary similarity measures.

The structure of this work is as follows. Section 2 presents the similarity measures that are compared in the analysis. Section 3 presents the datasets, the experiments and their results, and a discussion of the obtained outcome. Conclusions of the work are presented in section 4.

2. Similarity measures

The goal of the analysis is to evaluate binary similarity measures applied to GO analysis. Therefore, a classical semantic similarity measure, Jiang-Conrath measure, was applied as a reference method enabling a proper comparison of the methods.

2.1. Gene similarity based on semantic term similarity

Semantic term similarity measures utilise the concept of Information Content $\tau(a_i)$ of an ontology term $a_i \in A$ (where A is a set of all GO terms) given by the following formula:

$$\tau(a_i) = -\ln(P(a_i)), \tag{1}$$

where $P(a)$ is a ratio of a number of gene annotations to a term a, to a number of analysed genes.

Jiang-Conrath [8] method calculates distance between ontology terms according to the following formula:

$$d_A^{(JC)}(a_i, a_j) = \tau(a_i) + \tau(a_j) - 2\tau_{ca}(a_i, a_j), \tag{2}$$

where $\tau_{ca}(a_i,a_j)$ is the Information Content of the most informative common ancestor of the compared terms a_i and a_j.

Similarity based on formula (2) can be derived as:

$$s_A^{(JC)}(a_i, a_j) = \left(d_A^{(JC)}(a_i, a_j) + 1\right)^{-1}. \tag{3}$$

When the term similarity calculated by means of semantic measure is known, it is possible to calculate gene similarity based on the similarity of terms describing the genes. The similarity $s_G(g_k,g_p)$ between genes g_k and g_p can be calculated according to one of the approaches presented in literature.

The first approach [3], which will be further referred to as Avg-max, is defined as:

$$s_G\left(g_k,g_p\right)=\left(m_k+m_p\right)^{-1}\left(\sum_i \max_j\left(s_A\left(a_i,a_j\right)\right)+\sum_j \max_i\left(s_A\left(a_i,a_j\right)\right)\right), \tag{4}$$

where m_k and m_p are the number of annotations of genes g_k and g_p respectively, a_i and a_j belong to the term sets describing genes g_k and g_p respectively.

Another approach, which will be further referred to as Avg-sum, was applied in [16]:

$$s_G\left(g_k,g_p\right)=\left(m_k m_p\right)^{-1}\sum s_A\left(a_i,a_j\right), \tag{5}$$

where m_k and m_p are the number of annotations of genes g_k and g_p respectively, a_i and a_j belong to the term sets describing genes g_k and g_p respectively.

2.2. Binary similarity

The most popular binary similarity measure is Jaccard coefficient [4, 8]. However, there are plenty of other methods that can be applied to the task of binary data analysis. The measures that were analysed in the research described are presented below.

When two binary vectors x_k and x_m are compared the result of the comparison of each bit i can be encoded by means of the symbols presented in Table 1.

Table 1

Encoding of binary data comparison

		x_{mi}	
		1	0
x_{ki}	1	a	b
	0	c	d

The examples of categorical data similarity formulas (based on the symbols defined in Table 1) are mentioned below [4, 8].

Jaccard coefficient:

$$s^{(J)}=\frac{a}{a+b+c}, \tag{6}$$

Czekanowski similarity

$$s^{(C)}=\frac{2a}{2a+b+c}. \tag{7}$$

The total similarity of calculated bit vectors, e.g. using a Jaccard coefficient, can be defined as:

$$s^{(J)}(\mathbf{x}_k, \mathbf{x}_m) = \frac{|and(\mathbf{x}_k, \mathbf{x}_m)|}{|or(\mathbf{x}_k, \mathbf{x}_m)|}, \tag{8}$$

where $|and(\mathbf{x}_k, \mathbf{x}_m)|$ denotes the number of features with a value 1 for both feature vectors $\mathbf{x}_k$ and $\mathbf{x}_m$, $|or(\mathbf{x}_k, \mathbf{x}_m)|$ denotes the number of features with a value 1 for any of the feature vectors $\mathbf{x}_k$ or $\mathbf{x}_m$, n is the number of features creating the feature vector.

Analogously, Czekanowski similarity can be represented as:

$$s^{(C)}(\mathbf{x}_k, \mathbf{x}_m) = \frac{2|and(\mathbf{x}_k, \mathbf{x}_m)|}{|and(\mathbf{x}_k, \mathbf{x}_m)| + |or(\mathbf{x}_k, \mathbf{x}_m)|}. \tag{9}$$

Another approach is to calculate distance between compared vectors by means of *xor* operation. The popular approach of this type is Hamming distance:

$$d^{(H)}(\mathbf{x}_k, \mathbf{x}_m) = \frac{|xor(\mathbf{x}_k, \mathbf{x}_m)|}{n}, \tag{10}$$

where $|xor(\mathbf{x}_k, \mathbf{x}_m)|$ denotes the number of features having different values for both feature vectors $\mathbf{x}_k$ and $\mathbf{x}_m$, n is the number of features creating the feature vector. The similarity values based on Hamming distance are calculated as:

$$s^{(H)} = 1 - d^{(H)}. \tag{11}$$

Applying encoding presented in Table 1 Hamming distance (11) can be expressed as:

$$d^{(H)} = \frac{b+c}{a+b+c+d}. \tag{12}$$

Hamming distance values are normalized by the number of all features, which is number of all terms in GO annotating any gene from the dataset. Feature vectors describing gene-term annotations are very sparse and contain only a few 1 values whereas there can be thousands terms creating a feature vector. Therefore, it can be a better idea to modify the similarity measure and normalize *xor* operation by a number of 1 values in a given feature vector calculated by means of *or* operation:

$$d^{(X)}(\mathbf{x}_k, \mathbf{x}_m) = \frac{|xor(\mathbf{x}_k, \mathbf{x}_m)|}{|or(\mathbf{x}_k, \mathbf{x}_m)|}. \tag{13}$$

The similarity values based on the above measure are again calculated as:

$$s^{(x)} = 1 - d^{(X)}. \tag{14}$$

Applying encoding presented in Table 1 the formula (14) can be expressed as:

$$d^{(X)} = \frac{b+c}{a+b+c}. \tag{15}$$

3. Analysis

Two types of analysis were performed in order to evaluate the similarity measures presented above. The first approach compares correlation of the values contained in the similarity matrix generated on gene expression data and the values contained in the similarity matrix generated on Gene Ontology data. Gene expression values are represented by means of real numbers. The similarity of genes in gene expression representation is calculated by means of Pearson correlation in most cases. Having two gene similarity matrices calculated, where one is based in gene expression data and the other is based on Gene Ontology data, it is possible to compare them by means of correlation analysis. The measure that can better express similarity of genes in GO representation should give the higher values of correlation.

The second approach is based on visual analysis of clustering results, similarly as it was presented in [12]. We focus here only on the plot produced by a density-based algorithm OPTICS [2] where the valleys representing clusters are separated by the hills representing the data objects distant in terms of density reachability from other data and not belonging to any cluster.

3.1. Datasets

The datasets analysed are represented in gene expression and Gene Ontology domains. The matrix containing annotations of genes to Gene Ontology terms can be created in two ways. One way is to represent each gene-term annotation as a single 1 value in the matrix. Another approach is to extend each gene-term annotation as an annotation of a gene to a set of terms creating a hierarchy leading from a root term to the given term. Such approach is allowed as each ontology term details the information represented by a root term.

Two datasets of different characteristics were used in the experiments performed.

Yeast dataset [5] that consists of 274 genes, 79 expression attributes and 248 GO terms (862 GO terms in a hierarchical representation). This dataset contains genes expression profiles from budding yeast *S. cerevisiae* that were measured during several different DNA microarray experiments. For analysis described in this paper we selected only 274 genes that composed 10 well defined functional groups described by the authors of the paper.

Human dataset [9] that consists of 285 genes, 18 expression attributes and 1413 GO terms (3385 GO terms in a hierarchical representation). This dataset contains expression values of human fibroblasts in response to serum. Similar as in the previous case, we selected only genes from functional groups described by the authors of the paper. However we would like to stress that genes composing these groups were not as functionally uniform as groups described in the case of Yeast dataset.

To annotate genes we used GO terms from Biological Process ontology only. In all cases we included into analysis only genes that were described by at least one GO term.

3.2. Experiments and results

The experiments were implemented and performed in Matlab computing environment installed on a desktop PC computer. Before the experiments were started the similarity matrices by means of each measure were calculated. It showed the expected significant difference in complexity between binary measures and semantic measure Jiang-Conrath what can be expressed by the following exemplary comparison:

- similarity calculation of Human genes by means of Jaccard coefficient took about 4 sec.,
- similarity calculation of terms describing Human genes by means of Jiang-Conrath measure took about 3 days and 0.5 hour.

The first experiment covered analyses of all the similarity measures presented in section 2. Jiang-Conrath term similarity measure was combined with two gene similarity measures avg-sum and avg-max. This measure can be calculated only on the basis of annotation matrix where hierarchy is introduced. Four binary similarity measures (Czekanowski, Jaccard, Hamming and Xor) were analysed. Binary similarity measures can be calculated on the basis of both types of annotation matrix. The results of the analysis of correlation between similarity matrix calculated for gene expression values and the similarity matrices calculated for GO data for the two datasets are presented in Table 2.

Ranking of the results presented in Table 2 reveals several interesting observations:

- The best results were obtained when binary measures were applied. Aggregating the results for both datasets Jaccard similarity seems to perform best and it is followed by Czekanowski and Xor measure.
- Analysis of non-hierarchical annotation matrix by means of binary measures gives better results.
- Jiang-Conrath similarity measure gives generally worse results then most of binary measures.
- The similarity measure based on Hamming distance performs very poorly.

Table 2 shows also significant difference in quality of results between Yeast and Human datasets. Analysis of Yeast dataset gives the results of higher quality what is a consequence of the gene pre-selection mentioned in point 3.1.

The second experiment covered visual analysis of OPTICS density reachability plot. The parameters of OPTICS algorithm were set to $\varepsilon=1$ and $m=15$. Each similarity matrix calculated by means of the analysed measures was transformed to distance matrix which is an expected input parameter of OPTICS algorithm. The examplary plots calculated for Jinag-

Conrath (avg-max) and Czekanowski methods applied to Yeast dataset are presented in Fig. 1. a) and b) respectively. The same analysis performed on Human data set is presented in Fig. 2.

Table 2

Ranking of measures according to correlation between similarity matrix calculated for gene expression values and the similarity matrices calculated for GO data for the dataset (a) Yest and (b) Human. Hierarchical annotation matrix was used in the analysis where a measure is described as (h)

a)

No.	Measure	Correlation
1	Czekanowski	0.553
2	Jaccard	0.519
3	Xor	0.519
4	Czekanowski (h)	0.483
5	Jaccard (h)	0.475
6	Xor (h)	0.475
7	Jiang-Conrath (avg-max)	0.412
8	Jiang-Conrath (avg-sum)	0.370
9	Hamming (h)	0.017
10	Hamming	-0.055

b)

No.	Measure	Correlation
1	Jaccard	0.136
2	Xor	0.136
3	Jiang-Conrath (avg-sum)	0.133
4	Czekanowski	0.132
5	Jaccard (h)	0.119
6	Xor (h)	0.119
7	Jiang-Conrath (avg-max)	0.104
8	Czekanowski (h)	0.102
9	Hamming (h)	-0.102
10	Hamming	-0.120

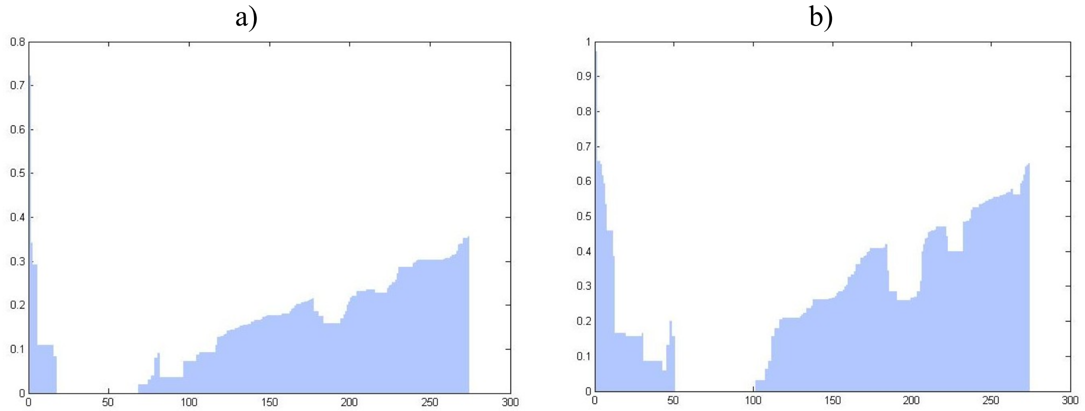

Fig. 1. OPTICS plots of Yeast data processed by means of: a) Jinag-Conrath (avg-max) and b) Czekanowski methods

Rys. 1. Wykresy algorytmu OPTICS wygenerowane dla danych Yeast przy zastosowaniu miar podobieństwa: a) Jiang-Conrath (avg-max) oraz b) Czekanowskiego

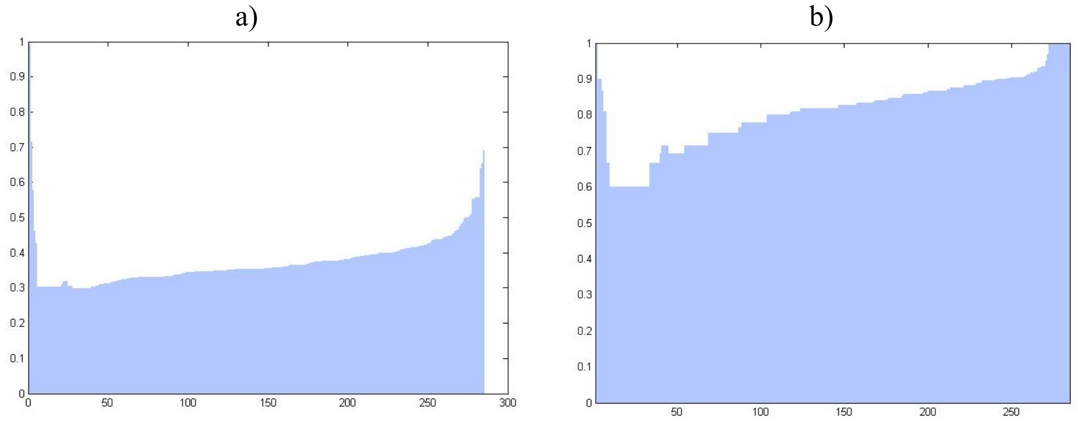

Fig. 2. OPTICS plots of Human data processed by means of: a) Jinag-Conrath (avg-max) and b) Czekanowski methods

Rys. 2. Wykresy algorytmu OPTICS wygenerowane dla danych Human przy zastosowaniu miar podobieństwa: a) Jiang-Conrath (avg-max) oraz b) Czekanowskiego

The plots presented in Fig. 1 and Fig. 2 are representative for the two classes of measures (binary and semantic) that were analysed. The results of the analysis show that there is no significant qualitative difference between the obtained plots. Czekanowski measure produces distance values enabling easier separation of clusters (valleys are deeper) comparing to Jiang-Conrath measure. However, both measures enabled the algorithm to reveal very similar structures in data. It is also visible that the analysis of Human data set is much more difficult and the results does not allow to point several dens clusters in data.

4. Conclusions

Application of binary similarity measures to the analysis of Gene Ontology data is presented in the article. The measures are presented, analysed and compared with Jiang-Conrath semantic similarity measure in two types of experiments on two datasets representing different characteristics.

The results, that were presented, show that binary similarity measures applied to Gene Ontology data analysis can produce similarity matrices that are better correlated with gene expression based similarity matrix then semantic similarity measures (Jiang-Conrath). Additionally, the results show also that binary similarity measures produce the results that are at least as informative (discriminative) for clustering algorithm as semantic similarity measures (Jiang-Conrath). Moreover, application of binary similarity measures is significantly less costly concerning processing power then application of semantic similarity measures (Jiang-Conrath).

These characteristics of binary similarity measures applied to the analysis of Gene Ontology data will be taken into consideration in further research connected with cluster analysis of such data.

The work was supported by National Science Centre (decision DEC-2011/01/D/ST6/07007)

BIBLIOGRAPHY

1. Al Mubaid H., Nagar A.: Comparison of four similarity measures based on GO annotations for Gene Clustering. IEEE Symposium on Computers and Communications, ISCC 2008, p. 531÷536.
2. Ankerst M., Breunig M., Kriegel H. P., Sander J.: OPTICS: ordering points to identify the clustering structure. SIGMOD Rec., Vol. 28, No. 2, 1999, p. 49÷60.
3. Azuaje F., Wang H., Bodenreider O.: Ontology-driven similarity approaches to supporting gene functional assessment. Proc. Of The Eighth Annual Bio-Ontologies Meeting, 2005.
4. Doreian P., Batagelj V., Ferligoj A.: Generalized blockmodeling. Cambridge University Press, Cambridge 2005.
5. 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, Vol. 95, 1998, p. 14863÷14868.
6. Gruca A., Kozielski M.: Correlation of genes similarity measures based on GO terms similarity and gene expression values. Advances in Intelligent and Soft Computing, Vol. 103, Springer, 2011, p. 137÷144.
7. Gruca A., Kozielski M., Sikora M.: Fuzzy Clustering and Gene Ontology Based Decision Rules for Identification and Description of Gene Groups. AISC, Vol. 59, 2009, p. 141÷149.
8. Han J., Kamber M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Academic Press, San Francisco 2001.
9. Iyer V. R., Eisen M. B., Ross D. T., Schuler G., Moore T., Lee J.C., Trent J. M., Staudt L. M., Hudson J., Boguski M. S., Lashkari D., Shalon D., Botstein D., Brown P. O.: The transcriptional program in the response of human fibroblasts to serum. Science, Vol. 283, 1999, p. 83÷87.

10. Jiang J. J., Conrath D. W.: Semantic similarity based on corpus statistics and lexical ontology. Proc. on Int. Conference on Research in Computational Linguistics, 1997, p. 19÷33.
11. Kozielski M., Gruca A.: Evaluation of Semantic Term and Gene Similarity Measures. LNCS, Vol. 6744, Springer, 2011, p. 406÷412.
12. Kozielski M., Gruca A.: Visual comparison of clustering Gene Ontology data when different similarity measures are applied. Studia Informatica, Vol. 32, No 2A(96), Wydawnictwo Politechniki Śląskiej, Gliwice 2011, p. 169÷180.
13. Pesquita C., Faria D., Falca A. O., Lord P., Couto F. M.: Semantic Similarity in Biomedical Ontologies. PLoS Comput. Biol., Vol. 5(7), 2009, p. 1÷12.
14. Sikora M., Gruca A.: Induction and selection of the most interesting Gene Ontology based multiattribute rules for descriptions of gene groups. Pattern Recogn. Letters, Vol. 32, 2011, p. 258÷269.
15. Sikora M., Gruca A.: Quality improvement of rules based gene groups descriptions using information about GO terms importance occurring in premises of determined rules. Int. Journal of Applied Mathematics & Computer Science, Vol. 20, No. 3, 2010, p. 555÷570.
16. 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 CIBCB ‘04, 2004, p. 25÷31.

Wpłynęło do Redakcji 31 stycznia 2012 r.

Omówienie

Celem niniejszego artykułu jest ocena możliwości zastosowania binarnych miar podobieństwa do analizy podobieństwa genów reprezentowanych w dziedzinie ontologii genowych, takich jak baza Gene Ontology (GO). Ocena oparta jest na dwóch eksperymentach, dla których porównano wyniki uzyskane przy zastosowaniu semantycznej miary Jiang-Conrath (2) oraz czterech miar binarnych: Jaccarda (6), Czekanowskiego (7), Hamminga (12) oraz Xor (15).

Przeprowadzone eksperymenty obejmują: (a) wyznaczenie rankingu miar opartego na wartościach korelacji macierzy podobieństwa genów, wyznaczonej za pomocą analizowanej miary w dziedzinie GO z macierzą podobieństwa genów opisanych w dziedzinie ekspresji oraz (b) porównanie jakości miar przez wizualną analizę wykresów uzyskanych w procesie grupowania genów.

Wyniki analizy wskazują, że zastosowanie binarnych miar podobieństwa pozwala na uzyskanie lepszych rezultatów. Wyniki te zostaną uwzględnione w dalszych badaniach związanych z grupowaniem genów opisanych w dziedzinie GO.

Addresses

Michał KOZIELSKI: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16,
44-100 Gliwice, Polska, michal.kozielski@polsl.pl.
Aleksandra GRUCA: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16,
44-100 Gliwice, Polska, aleksandra.gruca@polsl.pl.