STUDIA INFORMATICA 2014
Volume 35 Number 2 (116)

Łukasz STYPKA
Silesian University of Technology, Institute of Computer Science
Future Processing Sp. z o.o.
Michał KOZIELSKI
Silesian University of Technology, Institute of Electronics

METHODS OF NORMALIZATION THE RESULTS OF GENE ONTOLOGY TERM SIMILARITY

Summary. The article addresses the issue of improvement of the results quality when Gene Ontology (GO) term similarity is calculated. Several GO similarity measures produce results out of the range [0; 1]. Whereas, in order to compare different similarity measures or apply further processing, it is needed to normalise the results to this range. The most popular and well-known method of normalization is the *min-max* normalization. The article introduces seven normalization functions of different characteristics that can improve the results of the analysis. The comparison of the analysed methods on three different gene datasets and their evaluation is presented in this paper.

Keywords: Gene Ontology, Gene Ontology term similarity, normalization, normalization function

METODY NORMALIZACJI PODOBIEŃSTWA WYZNACZONEGO DLA TERMINÓW ONTOLOGII GENE ONTOLOGY

Streszczenie. Artykuł porusza problem normalizacji podobieństwa wyznaczonego dla terminów ontologii Gene Ontology (GO). Wiele metod pozwalających wyznaczyć podobieństwo terminów GO daje wyniki spoza przedziału [0; 1], podczas gdy przedział ten jest wymagany w celu porównania wybranych metod oraz dalszych analiz. W niniejszej pracy zaprezentowano siedem różnych funkcji normalizacyjnych oraz ich porównanie w odniesieniu do metody normalizacji *min-max*. Badania zostały przeprowadzone na trzech zbiorach genów o różnej charakterystyce.

Słowa kluczowe: ontologia genowa, podobieństwo terminów ontologii Gene Ontology, normalizacja, funkcje normalizujące

1. Introduction

Gene Ontology (GO) [2] is a popular knowledge base developed by GO Consortium [7]. The Gene Ontology consists of the terms and relations between them. Terms can belong to one of the sub-ontologies such as: biological process, molecular function or biological component. Relations can be of different types such as e.g., *is a*, *part of* or *regulates*.

Gene Ontology is represented as a directed acyclic graph where terms and relations are represented by nodes and edges respectively.

Gene Ontology is an important source of knowledge utilized in several research projects and analysis [7] as it enables to annotate gene products to Gene Ontology terms. One of the important issues that can be approached in Gene Ontology analysis is evaluation of GO terms similarity. It is often a preliminary step of the comparison and analysis of the annotated gene products.

There are several measures enabling GO term similarity calculation, what was analysed in a survey presented by Pesquita et al., [14]. Some of them, such as measures introduced by Resnik [15], Jiang and Conrath [10] and Lin [12] can be named as classical approaches as most newer methods refer to them and several methods further develop them, e.g., [5, 18]. This area of research is being further developed, as recently, new approaches introducing more complex ideas to term similarity analysis were presented [1, 9].

In this work, we focus on a single measure that refers directly to the graph representation of Gene Ontology. This approach applies shortest path analysis [13] in order to determine the similarity of GO terms. The results of the method have to be normalized in order to be utilized in further analysis, e.g. clustering. Therefore, the goal of this work is to present and analyse several normalization functions improving the quality of term similarity calculated. The analysis are performed on three gene datasets and the best normalization functions are selected.

The results quality evaluation is based on a comparative study performed in two data (gene) representations.

The first representation is Gene Ontology, where term and gene product similarity is calculated and where normalization functions are applied. The resulting similarity is compared with the results obtained in the second representation, which is gene expression identified in a microarray experiment.

Pearson correlation between the gene similarities in both representations verifies the final quality of the applied methods. Correlation as a verification method was used in such applications, e.g. [11, 16], what motivated the current application of this method.

The structure of this work is as follows. Section 2 presents the utilized GO term similarity measure and the whole process of gene similarity calculation. The normalization solutions that were introduced and implemented are presented in section 3. The results of the quality analysis are presented and discussed in section 4. Conclusions of the work are presented in section 5.

2. GO-based similarity

Several GO term similarity measures, e.g., Resnik [15] or shortest path [13], can produce results in a range [0, ∞]. Whereas, in order to compare different similarity measures or apply further processing, it is needed to normalise the results to [0, 1] range [5].

A typical approach to this issue is to perform *min-max* normalization [5] having the following form:

$$v' = \frac{v-m}{M-m} v' = \frac{v-m}{M-m} \qquad (1)$$

where v' is a normalised value of v, M and m are respectively: maximal and minimal values calculated for a given dataset.

In order to thoroughly analyse impact of normalization on the resulting values we focus in the given work, on one approach to measure GO term similarity, which is based on analysis of the shortest paths in a Gene Ontology graph.

The distance between two terms a_i and a_j is defined as a length $l(a_i, a_j)$ of the shortest path between them in this method.

Calculating shortest paths in Gene Ontology it has to be taken into consideration that the ontology graph is a directed one. Therefore, the length of a path between two ontology terms that are not connected by a parent-child relation can be set as infinity or it can be calculated as a sum of path lengths leading to the nearest common ancestor a_{ca}. The latter approach was chosen in the work presented. Therefore, the distance of the two GO terms a_i and a_j having the nearest common ancestor a_{ca} can be defined as:

$$d_A(a_i, a_j) = \lambda\ (\ l(a_i, a_{ca}) + l(a_j, a_{ca})) \qquad (2)$$

where λ is a parameter setting the weight of the relations between GO terms.

The example results of distance calculated by means of measure (2), where λ=0.2 is presented in Table 1.

Table 1

The example results of distance matrix

	GO:0000027	GO:0002181	GO:0006412
GO:0008152	1.4	1.0	0.8
GO:0016310	2.0	1.4	1.2
GO:0006096	1.8	1.2	1.0

Next, the similarity of the GO terms can be determined as:

$$s_A = 1 - d_A \tag{3}$$

Once the term similarity is known it is possible to calculate gene similarity based on the similarity of terms describing the genes. There are many methods of calculating similarity between pair of genes [3, 18]. In this paper the similarity $s_G(g_k, g_p)$ between genes g_k and g_p was calculated by means of the following formula ([3]):

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

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.

3. Normalization functions

The method chosen to normalise the GO term similarity values can have significant impact on the results of the analysis. Therefore, we introduce in this section seven normalization functions (f, g, h, i, j, k, l: $[0, \infty] \rightarrow [0,1]$) listed below. These functions can be applied to the gene similarity calculation process presented in section 2.

$$f(x) = 2\frac{1}{1+e^{-x}} - 1, f(x) = 2\ \frac{1}{1+e^{-x}} - 1 \tag{5}$$

$$g(x) = 2\frac{1}{1+e^{-0.5x}} - 1, g(x) = 2\ \frac{1}{1+e^{-0.5x}} - 1 \tag{6}$$

$$h(x) = 2\frac{1}{1+e^{-0.33x}} - 1, h(x) = 2\ \frac{1}{1+e^{-0.33x}} - 1 \tag{7}$$

$$i(x) = \frac{\arctan(x)}{\frac{\pi}{2}}, i(x) = \frac{\arctan(x)}{\frac{\pi}{2}} \tag{8}$$

$$j(x) = \frac{x}{\sqrt{1+x^2}}, j(x) = \frac{x}{\sqrt{1+x^2}} \tag{9}$$

$$k(x) = \frac{x}{1+|x|}, k(x) = \frac{x}{1+|x|} \tag{10}$$

$$l(x) = \tanh(x). l(x) = \tanh(x) \tag{11}$$

The characteristics of the introduced normalization functions are presented in Fig. 1.

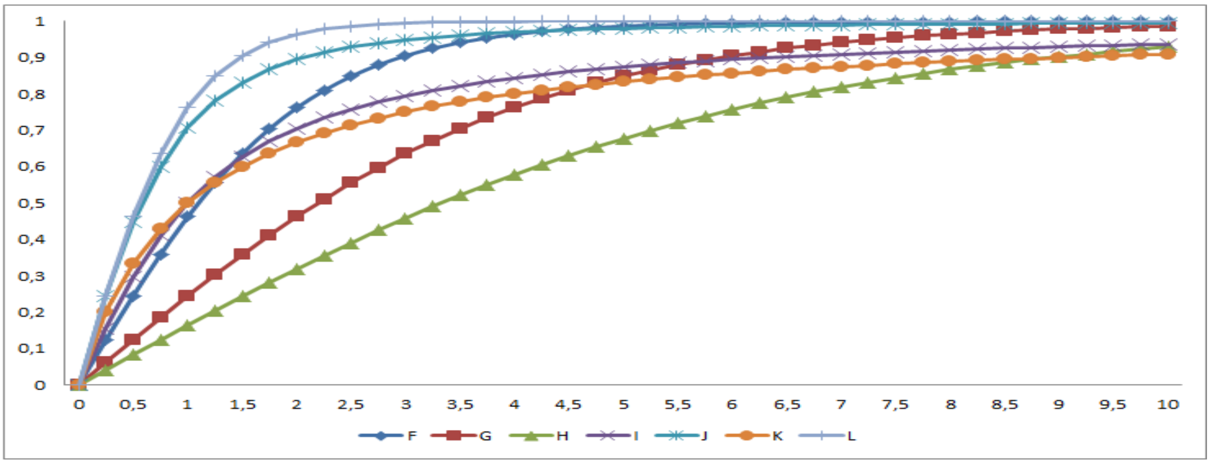

Fig. 1. The characteristics of the introduced normalization functions
Rys. 1. Charakterystyka przedstawionych funkcji normalizacyjnych

The results of the normalization by function *f* (5) of the example values from Table 1 are presented in Table 2.

Table 2

The results of the normalization of the example values from Table 1

	GO:0000027	GO:0002181	GO:0006412
GO:0008152	0.604	0.462	0.380
GO:0016310	0.762	0.604	0.537
GO:0006096	0.716	0.537	0.462

4. Experiments and results

Three gene databases were used during the experiments: Yeast1 [6], Human [8] and Yeast2 [4]. These datasets are characterized by the properties presented in Table 3.

Table 3

Gene datasets used during the experiments

Dataset	Number of genes	Number of terms used to describe the genes	Average number of terms used to describe one gene
Yeast1	274	248	3.72
Human	285	1413	10.18
Yeast2	1111	887	3.29

The datasets were analysed according to the procedure presented in section 2, where GO term similarity was calculated first, next the normalization was performed and finally gene similarity was calculated. The shortest path method utilized several values of the λ parameter, where $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. In the normalization step both *min-max* method (1) and introduced functions (5 - 11) were applied.

The results calculated (GO based gene similarity) were referred to gene similarity calculated in gene expression representation. Such approach has been used already [11, 16] and a correlation coefficient between gene similarity values is then applied to compare different methods. The more similar are the similarities in both representations (the higher correlation coefficient is received), the better is the outcome of the given method.

The gene similarity in gene expression representation was calculated as Pearson correlation coefficient. This method is usually applied as the data (expression values) consist of a set of time series and Pearson correlation coefficient is more suitable in this application then Euclidean distance [6].

The results of the analysis of the three datasets are presented in tables 4, 5 and 6 and visualized on figures 2, 3 and 4 respectively.

Table 4

Quality of results, expressed by correlation coefficient values, for Yeast1 dataset

λ	min-max	f	G	h	i	j	K	l
0.1	0.305	0.564	0.555	0.596	0.554	0.584	0.607	0.589
0.2	0.305	0.589	0.564	0.616	0.558	0.631	0.626	0.630
0.3	0.305	0.613	0.576	0.632	0.564	0.645	0.636	0.647
0.4	0.305	0.630	0.589	0.640	0.572	0.650	0.641	0.653
0.5	0.305	0.641	0.602	0.644	0.580	0.652	0.644	0.654
0.6	0.305	0.647	0.613	0.647	0.589	0.652	0.646	0.653
0.7	0.305	0.651	0.622	0.649	0.597	0.651	0.647	0.651
0.8	0.305	0.653	0.630	0.605	0.649	0.651	0.648	0.650
0.9	0.305	0.653	0.636	0.612	0.650	0.650	0.649	0.648
1	0.305	0.654	0.641	0.619	0.650	0.649	0.649	0.646

Table 5

Quality of results, expressed by correlation coefficient values, for Human dataset

λ	min-max	f	g	h	I	j	k	L
0.1	0.004	0.008	0.003	0.000	0.012	0.017	0.018	0.015
0.2	0.004	0.015	0.008	0.005	0.023	0.032	0.026	0.032
0.3	0.004	0.023	0.011	0.008	0.031	0.043	0.032	0.047
0.4	0.004	0.032	0.015	0.010	0.037	0.051	0.036	0.057
0.5	0.004	0.040	0.019	0.012	0.042	0.055	0.039	0.062
0.6	0.004	0.047	0.023	0.015	0.045	0.057	0.042	0.063
0.7	0.004	0.053	0.027	0.017	0.048	0.058	0.044	0.062
0.8	0.004	0.057	0.032	0.020	0.049	0.058	0.046	0.061
0.9	0.004	0.060	0.036	0.023	0.051	0.058	0.047	0.060
1	0.004	0.062	0.040	0.026	0.052	0.058	0.048	0.058

In case of Yeast1 dataset, there can be observed a clear difference between the *min-max* normalization (1) and normalization based on the introduced functions (5, 6, 7, 8, 9, 10, 11). Correlation coefficient equal to 30% was obtained for the first type of normalization, whereas the second type achieved results in a range from 55% to 65%. It is worth noting, that increase of the weight of a path affects the value of the correlation coefficient. However, this affect is not significant.

There can be observed significantly greater impact of the weight of a path λ on the quality results for Human dataset. The quality of the results for the five normalization functions (f, g, h, i, k) increase along the increase of λ value. In case of *j* and *l* functions, the maximum value was achieved for the weight equal to 0.8 and 0.6 respectively, and afterwards the correlation coefficient began to decrease. The *min-max* normalization in this case also turned out to be the worst approach, where the correlation coefficient reached only 0.4%. The best results were achieved for the introduced function *l*, for which correlation coefficient reached above 6.2%.

Table 6

Quality of results, expressed by correlation coefficient values, for Yeast2 dataset

λ	min-max	f	g	h	i	j	k	L
0.1	0.009	0.018	0.018	0.017	0.020	0.021	0.021	0.020
0.2	0.009	0.020	0.018	0.018	0.022	0.024	0.023	0.024
0.3	0.009	0.022	0.019	0.018	0.023	0.024	0.023	0.025
0.4	0.009	0.024	0.020	0.019	0.024	0.024	0.023	0.024
0.5	0.009	0.024	0.021	0.019	0.024	0.024	0.023	0.024
0.6	0.009	0.025	0.022	0.020	0.024	0.023	0.023	0.023
0.7	0.009	0.025	0.023	0.021	0.024	0.023	0.023	0.022
0.8	0.009	0.024	0.024	0.022	0.023	0.023	0.023	0.022
0.9	0.009	0.024	0.024	0.022	0.023	0.022	0.023	0.022
1	0.009	0.024	0.024	0.023	0.023	0.022	0.023	0.021

Similar trends can be noticed in case of Yeast2 dataset. Very evident is the difference between *min-max* normalization and normalization resulting from the introduced functions. The value of the correlation coefficient for *min-max* normalization was below 0.9%, whereas for the other functions, these values were in the range of 1.7% - 2.46%.

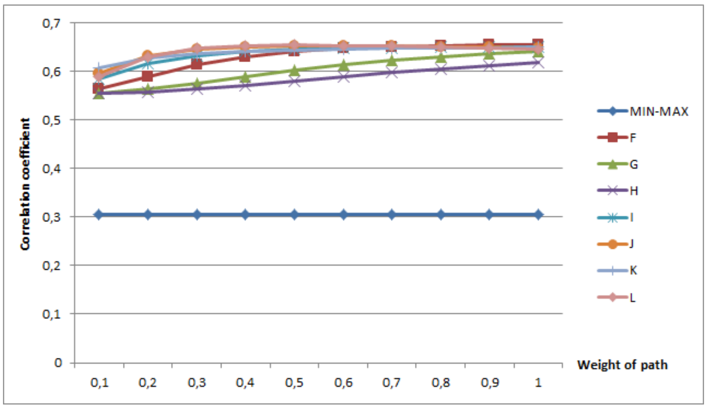

Fig. 2. Quality of results, expressed by correlation coefficient values, for Yeast1 dataset
Rys. 2. Jakość wyników, wyrażonych jako współczynnik korelacji, dla zbioru Yeast1

Again, a stronger impact of λ parameter on the quality of results can be noticed. The functions *f*, *j* and *l* reached their maximum quality for the weight λ=0.6, 0.3 and 0.3 respectively.

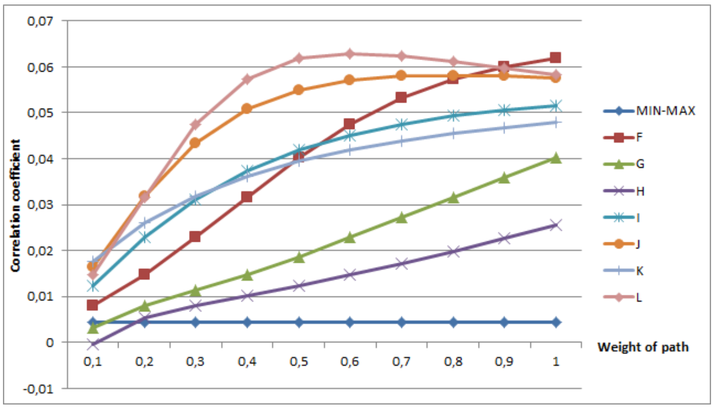

Fig. 3. Quality of results, expressed by correlation coefficient values, for Human dataset
Rys. 3. Jakość wyników, wyrażonych jako współczynnik korelacji, dla zbioru Human

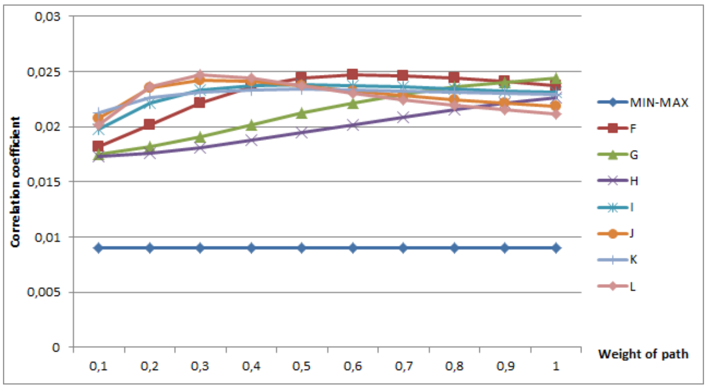

Fig. 4. Quality of results, expressed by correlation coefficient values, for Yeast2 dataset
Rys. 4. Jakość wyników, wyrażonych jako współczynnik korelacji, dla zbioru Yeast2

It is also possible to analyse the results presented in Fig. 2, 3 and 4 in the context of the introduced normalization function characteristics that can be recognized in Fig. 1.

Functions *j* and *l* are the steepest and they have similar characteristics of the quality of results. It can be noticed on Fig. 3 and 4 and it was also mentioned in the analysis above that they achieve maximal quality for a chosen value of parameter λ and then their quality decreases with the increase of λ value. Therefore it can be said that their maximal quality is λ dependent, as most of the functions achieve their maximal quality for $\lambda = 1$, although they are able to achieve very good results (especially *l* function).

The *h* and *g* functions are the less steep in turn. It results in a slow increase of their quality value along with the increase of λ values. It also causes a poor performance in terms of quality results of these functions (an exception is the quality reached by *g* function for $\lambda = 1$ in Yeast2 dataset).

The characteristic feature of *i* and *k* functions is that they converge to 1 the most slowly what makes their results of poor quality in general. However, it enables function *k* to achieve the best results independently of a dataset for $\lambda = 0.1$.

Finally, there is *f* function, which can be characterized as medium steep and quickly converging to 1. The important feature of this function is that it enabled to achieve nearly or exactly the best results for each dataset and it was quite stable in terms of λ parameter value. It means that setting λ to 0.9 or 1 and applying *f* normalization function during the analysis assures the results of at least good quality. This observation is additionally visualized in Fig. 5, 6 and 7.

5. Conclusions

The article presented two different approaches to normalization issue. The first approach was based on a well-known *min-max* function. The second approach used special normalization functions. Seven different normalization functions were introduced in this paper, where each function had different characteristics. Three gene datasets of different characteristics were analyzed in order to verify the quality of the new approaches.

The experiments performed proved that the functions that were introduced achieve significantly better results. The impact of the parameter (path weight λ) of GO term similarity measure was also verified in the article. It was shown that λ can have significant impact on the results. Increasing the weight of paths was usually followed by the improvement of the quality of the results in case of the most functions. The best and the most stable results were obtained by applying normalization function denoted as *f*.

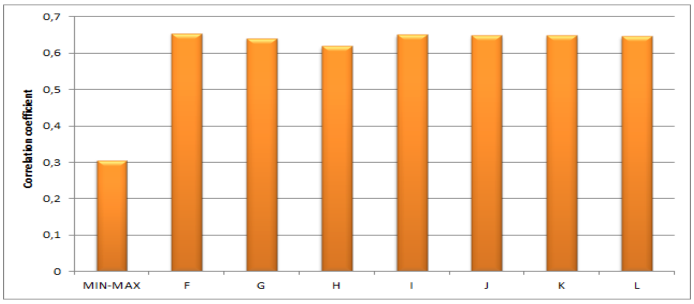

Fig. 5. Quality of results, expressed by correlation coefficient values, for Yeast1 dataset, for weight $\lambda = 1$

Rys. 5. Jakość wyników, wyrażonych jako współczynnik korelacji, dla zbioru Yeast1, dla wagi $\lambda = 1$

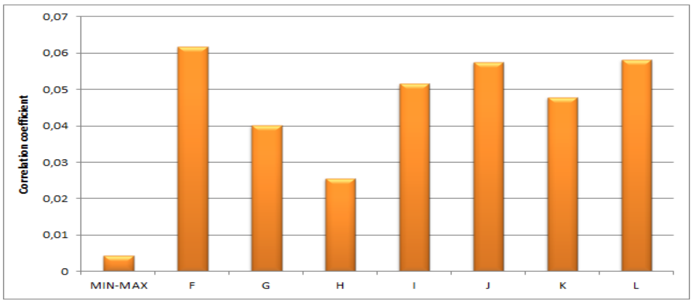

Fig. 6. Quality of results, expressed by correlation coefficient values, for Human dataset, for weight $\lambda = 1$

Rys. 6. Jakość wyników, wyrażonych jako współczynnik korelacji, dla zbioru Human, dla wagi $\lambda = 1$

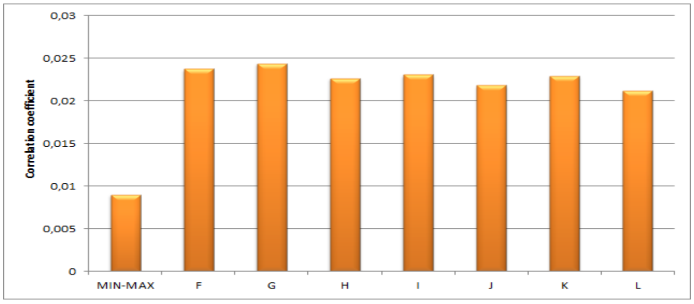

Fig. 7. Quality of results, expressed by correlation coefficient values, for Yeast2 dataset, for weight $\lambda = 1$

Rys. 7. Jakość wyników, wyrażonych jako współczynnik korelacji, dla zbioru Yeast2, dla wagi $\lambda = 1$

The work was supported by National Science Centre (decision DEC-2011/01/D/ST6/-07007) and by Ministry of Science and Higher Education as a Statutory Research Project (decision 8686/E-367/S/2013).

BIBLIOGRAPHY

1. Alvarez M. A., Qi X., Yan C.: A shortest-path graph kernel for estimating gene product semantic similarity. J. Biomedical Semantics, 2, 3, 2011.
2. Ashburner M. et al.: Gene Ontology: tool for the unification of biology. Nature genetics 25.1, 2000, p. 25÷29.
3. Azuaje F., Wang H., Bodenreider O.: Ontology-driven similarity approaches to supporting gene functional assessment. Proceedings Of The Eighth Annual Bio-Ontologies Meeting, Michigan 2005, p. 9÷10.
4. 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 2, 1998, p. 65÷73.
5. Couto F. M., Silva M. J., Coutinho, P. M.: Measuring semantic similarity between Gene Ontology terms. Data & knowledge engineering, 61(1), 2007, p. 137÷152.

6. 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 95, 1998, p. 14863÷14868.
7. GO-Consortium: The Gene Ontology (GO) database and informatics resource, Nucleic Acids Research, 32, 2004 (http://www.geneontology.org).
8. 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., Lashkari D., Shalon D., Botstein D., Brown P.: The transcriptional program in the response of human fibroblasts to serum. Science, 283, 1999, p. 83÷87.
9. Jain S., Bader G.: An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. BMC Bioinformatics, 11(1), 2010, 562.
10. Jiang J. J., Conrath D. W.: Semantic similarity based on corpus statistics and lexical ontology. Proc. on International Conference on Research in Computational Linguistics, 1997, p. 19÷33.
11. Kozielski M., Gruca A.: Evaluation of semantic term and gene similarity measures. Pattern Recognition and Machine Intelligence, 2011, p. 406÷411.
12. Lin D:. An information-theoretic definition of similarity. Proc. of the 15th Int'l Conference on Machine Learning, 1998, p. 296÷304.
13. Al Mubaid H., Nagar A.: Comparison of four similarity measures based on GO annotations for gene clustering. Computers and Communications. ISCC 2008. IEEE Symposium, 2008, p. 531÷536.
14. Pesquita C., Faria D., Falcão A. O., Lord P., Couto F. M.: Semantic Similarity in Biomedical Ontologies. PLoS Comput Biol 5(7), 2009, p. 1÷12.
15. Resnik P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. J. Artif. Intell. Res. (JAIR), Vol. 11, 1999, p. 95÷130.
16. Sevilla J. L., Segura V., Podhorski A., Guruceaga E., Mato J. M., Martinez-Cruz L. A., Corrales F. J., Rubio A.: Correlation between gene expression and GO semantic similarity. IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2(4), 2005, p. 330÷338.
17. Wang H., Azuaje F., Bodenreider O., Dopazo J.: Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB '04, 2004, p. 25÷31
18. Yang H., Nepusz T., Paccanaro A.: Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. Bioinformatics, 28(10), 2012, p. 1383÷1389.

Wpłynęło do Redakcji 7 lutego 2014 r.

Omówienie

Wiele metod pozwalających wyznaczyć podobieństwo terminów ontologii Gene Ontology (GO) daje wyniki spoza przedziału [0; 1], podczas gdy przedział ten jest wymagany w celu porównania wybranych metod oraz dalszych analiz. Wybór metody normalizacji może mieć istotne znaczenie, jeśli chodzi o jakość otrzymanych wyników podobieństwa. W artykule przedstawiono siedem różnych funkcji normalizujących, które zostały porównane ze sobą oraz z typową metodą, jaką jest normalizacja *min-max*. Analizy zostały przeprowadzone na trzech zbiorach genów, które mają różną charakterystykę.

Analizy pokazały, że każda z zaproponowanych funkcji pozwala na uzyskanie lepszych wyników od metody *min-max*. Badania ujawniły również charakterystykę analizowanych funkcji, co pozwoliło na wskazanie najbardziej interesującego podejścia.

Addresses

Łukasz STYPKA: Silesian University of Technology, Institute of Computer Science,
ul. Akademicka 16, 44-100 Gliwice, Polska, lukasz.stypka@polsl.pl.
Future Processing Sp. z o. o. ul. Bojkowska 37, 44-100 Gliwice, Polska
Michał KOZIELSKI: Silesian University of Technology, Institute of Electronics,
ul. Akademicka 16, 44-100 Gliwice, Polska, michal.kozielski@polsl.pl.