Evaluation of Semantic Term and Gene Similarity Measures

Michal Kozielski and Aleksandra Gruca

Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
{michal.kozielski,aleksandra.gruca}@polsl.pl

Abstract. In this paper we present the results of the research verifying how the functional description of genes contained in Gene Ontology database is related to genes expression values recorded during biological experiments. We compare several different gene similarity measures and semantic term similarity measures, and evaluate how the similarity of genes based on Gene Ontology terms is correlated with similarity of genes based on expression profiles. The analysis are preformed on three different datasets and we show that there is no single term similarity measure that always gives the best correlation results. The choice of the best term similarity measure depends on dataset characteristic.

Keywords: genes similarity, semantic term similarity, Gene Ontology database, experssion analysis.

1 Introduction

Gene Ontology (GO) is a database created throughout the years where the scientists introduced the knowledge resulting from the biological experiments in the form of gene-term annotations. Gene Ontology is also often used in the gene analysis where it can be regarded as an expert knowledge that helps to interpret results of the biological or medical experiments. In that way Gene Ontology can be also utilised in revealing the information hidden in gene expression data [7]. It is therefore needed to verify how the information contained in Gene Ontology is related to the experiments performed and which methods of the Gene Ontology analysis are best fitted to the gene expression analysis.

The semantic gene similarity evaluation consists of two steps: (1) term similarity calculation, (2) gene similarity based on term similarity calculation.

Several studies of Gene Ontology analysis were performed [5]. These studies, among others, analysed different similarity measures that can be applied to GO analysis and analysed how the Gene Ontology similarity is correlated with the similarity of different biological domains, e.g. protein sequence [5].

However, there are few approaches to analysis of the correlation of Gene Ontology based and gene expression based similarity. The work [8] presented interesting dependencies between both similarities but no conclusions on the quality of the similarity measures were drawn. The work [6] pointed Resnik similarity measure as giving the best correlation with the similarity of gene expression values.

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 406–411, 2011.

There are however two issues that are left ambiguous in this work: which method was used in order to calculate gene similarity in case of Lin and Jiang-Conrath term similarity methods; how was the gene expression similarity aggregated in order to improve the GO-expression correlation.

In our opinion additional research and discussion on the topic presented is needed. The contribution of our work covers analysis and comparison of three semantic term similarity measures and three pairwise, term based gene similarity measures. The analysis was performed on three datasets having different characteristics.

The paper is organised as follows. Sections 2 and 3 present semantic term similarity measures and term based gene similarity measures respectively. In section 4 the datasets, experiments and their results are presented. The final conclusions are drawn in section 5.

2 Semantic Term Similarity Measures

Semantic similarity of the Gene Ontology terms can be calculated applying the concept of *Information Content* $\tau(a)$ of an ontology term $a \in A$ defined as $\tau(a) = -ln(P(a))$, where $P(a)$ is a ratio of a number of annotations to a term a, to a number of analysed genes.

The simplest similarity measure proposed by Resnik [5,6] takes under consideration only the *Information Content* of the common ancestor $\tau_{ca}(a_i, a_j)$ of the compared terms a_i and a_j:

$$s_A^{(R)}(a_i, a_j) = \tau_{ca}(a_i, a_j). \tag{1}$$

More complex approach was proposed by Jiang-Conrath [5,6], where term similarity is defined as:

$$s_A^{(JC)}(a_i, a_j) = (d_A^{(JC)}(a_i, a_j) + 1)^{-1}, \tag{2}$$

where $d_A^{(JC)}(a_i, a_j)$ is a term distance defined as:

$$d_A^{(JC)}(a_i, a_j) = \tau(a_i) + \tau(a_j) - 2\tau_{ca}(a_i, a_j). \tag{3}$$

Another approach was presented by Lin [5,6]:

$$s_A^{(L)}(a_i, a_j) = \frac{2\tau_{ca}(a_i, a_j)}{\tau(a_i) + \tau(a_j)}. \tag{4}$$

3 Gene Similarity Measures

When the term similarity 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 very simple approach ([6]) may be to take the maximal similarity value of the terms annotated to the analysed genes

$$s_G(g_k, g_p) = max(s_A(a_i, a_j)), \tag{5}$$

where a_i and a_j belong to the term sets describing genes g_k and g_p respectively. This approach is referred further as Max method.

The more complex approach, which will be further referred to as Avg-max, may be found in [1]:

$$s_G(g_k, g_p) = (m_k + m_p)^{-1}(\sum_i \max_j(s_A(a_i, a_j)) + \sum_j \max_i(s_A(a_i, a_j))), \tag{6}$$

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 method, which is further referred to as Avg-sum, was applied in [8]:

$$s_G(g_k, g_p) = (m_k m_p)^{-1} \sum (s_A(a_i, a_j)), \tag{7}$$

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.

4 Analysis

4.1 Datasets

Three datasets of different characteristics were used in the experiments performed. Yeast1 dataset [3] consists of 274 genes, 79 expression attributes, 645 GO terms. Human dataset [4] consists of 296 genes, 18 expression attributes, 1711 GO terms. Yeast2 dataset [2] consists of 1099 genes, 17 expression attributes, and 1552 GO terms.

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. Analysing correlation of GO based similarity and gene expression based similarity it is needed to present the values and distribution of gene expression similarity within each dataset. As it was shown in the work [8] there is a nonlinear relation between the GO and gene expression based similarity. This observation can influence the results of the analysis in the present work because if the similarity of genes in a gene expression domain is little then there will be

Table 1. Average value of gene expression similarity

Yeast1	Human	Yeast2
0.257	0.045	0.057

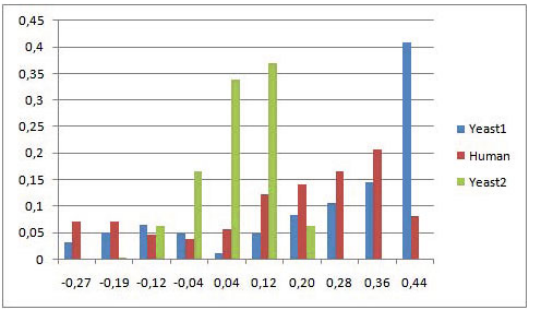

Fig. 1. Histogram showing the distribution of the similarity values in gene expression domain

little correlation of this similarity with the GO based one. It may be noticed in Table 1 that the average value of gene expression similarity, calculated as a Pearson correlation coefficient, is smaller for Human dataset then for Yeast2 dataset, which should result in its worse correlation with Gene Ontology based similarity. However, the distributions presented in Fig. 1 show that for the Human dataset the gene expression similarity has greater variance (the histogram is spread along the whole x axis) comparing to Yeast2 dataset. Thus, the differences in average values can be compensated and it is Human dataset which can give better results.

4.2 Experiments and Results

The following methods were applied in the experiments. Gene expression based similarity was calculated as Pearson correlation of the expression values.

Gene Ontology annotations were introduced to the analysis by means of binary array where "1" represented annotation of a gene to a term. The annotation table was constructed in such way that annotation of a gene to a term resulted in annotation of that gene to all the parents of a given term.

GO term similarity was calculated by means of Jiang-Conrath (2), Lin (4) and Resnik (1) methods. Gene similarity was calculated on the basis of term similarity by means of Max (5), Avg-max (6) and Avg-sum (7) methods.

The results of the experiments (correlation between gene expression based and GO based similarity matrices) are presented in the Table 2.

The Max gene similarity is not applicable to Jiang-Conrath and Lin similarity measures since it produces in these cases the similarity matrix containing only a value 1. It results from the fact that if two genes are annotated to the terms having common ancestor, then that ancestor annotates both genes in the calculated annotation table. Calculating gene similarity as a pairwise term similarity we have to calculate also self-similarity of such common ancestor. The self-similarity of a term in case of both Jiang-Conrath and Lin similarity measures equals 1 and this is maximal similarity value that can be calculated. In

Table 2. Correlation of gene expression based and GO based similarity

		Yeast1	Human	Yeast2
Max	Resnik	0.257	0.006	-0.013
Avg-max	Jiang-Conrath	0.420	0.089	0.015
	Lin	0.368	0.075	0.005
	Resnik	0.180	-0.024	0.001
Avg-sum	Jiang-Conrath	0.383	0.117	0.028
	Lin	0.352	0.089	0.011
	Resnik	0.039	-0.068	-0.038

case of Resnik measure the self-similarity of a term is equal to its information content and therefore it is possible to apply the Max method jointly only with Resnik measure.

The results show that the method Avg-sum gave the highest absolute correlation values in case of two out of three datasets analysed. Considering term similarity measures it is Jiang-Conrath that gives the best results in most cases.

The second experiment performed stems from the following two facts mentioned above: (1) the dependency between gene similarity matrices based on expression values and on Gene Ontology is not linear, (2) the best correlation between such similarity matrices is received when the genes that are highly similar in expression domain are taken under consideration.

Knowing these two facts we compared the correlation of the full similarity matrices with the correlation calculated for the matrices indexed by the values of expression based similarity s_e fulfilling the condition $s_e > \varepsilon$. The values of $\varepsilon = 0.5$ and $\varepsilon = 0.6$ were taken under consideration. The results of the experiment, when gene similarity was calculated by means of Avg-sum, are presented in Table 3.

Table 3. Correlation of gene expression based similarity and GO based similarity for the matrices reduced by ε condition (Avg-sum method); A - full data, B - data reduced for $\varepsilon = 0.5$, C - data reduced for $\varepsilon = 0.6$

	Yeast1			Human			Yeast2		
	A	*B*	C	A	*B*	C	A	*B*	C
Jiang-Conrath	0.383	0.402	0.334	0.117	0.232	0.304	0.028	0.176	0.298
Lin	0.352	0.392	0.323	0.089	0.133	0.192	0.011	0.149	0.256
Resnik	0.039	-0.181	-0.224	-0.068	-0.285	-0.355	-0.038	-0.250	-0.385

The results presented in Table 3 show that the dataset reduction applied improves in most cases the correlation between the similarity matrices compared. It can be also noticed that the results for the Resnik similarity measure gain the greater improvement (when absolute value of a correlation coefficient is taken under consideration) and become better then the results for Jiang-Conrath

measure. It can be however disputable if the negatively correlated similarity matrices can be applied successfully in all types of analysis.

5 Conclusions

The work presented analysis and comparison of three semantic term similarity measures and three pairwise, term based gene similarity measures. The analysis was performed on three datasets having different characteristics.

The results of the experiments showed that Jiang-Conrath and Resnik term similarity measures can give the best results of gene expression based similarity and GO based similarity. Jiang-Conrath term similarity measure however, performs better on raw data, gives always positively correlated results and seems to be more suitable for further applications. Considering gene similarity measures that were analysed it is Avg-sum method that gives the best results.

References

1. Azuaje, F., Wang, H., Bodenreider, O.: Ontology-driven similarity approaches to supporting gene functional assessment. In: Proceedings Of The Eighth Annual Bio-Ontologies Meeting, Michigan (2005)
2. 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, 65–73 (1998)
3. 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, 14863–14868 (1998)
4. 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 283, 83–87 (1999)
5. Pesquita, C., Faria, D., Falcao, A.O., Lord, P., Couto, F.M.: Semantic Similarity in Biomedical Ontologies. PLoS Comput. Biol. 5(7), 1–12 (2009)
6. 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 Transactions on Computational Biology and Bioinformaticsm 2(4), 330–338 (2005)
7. Sikora, M., Gruca, A.: Induction and selection of the most interesting Gene Ontology based multiattribute rules for descriptions of gene groups. Pattern Recogn. Letters 32, 258–269 (2011)
8. Wang, H., Azuaje, F., Bodenreider, O., Dopazo, J.: Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In: Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2004, pp. 25–31 (2004)