New Gene Ontology term similarity measure - comparison and performance evaluation based on DNA microarray data

Aleksandra Gruca [a], Michał Kozielski [b]
[a], *Institute of Informatics, Silesian University of Technology, Gliwice, Poland*
[b], *Institute of Electronics, Silesian University of Technology, Gliwice, Poland*
aleksandra.gruca@polsl.pl

Rapid evolution of high-throughput technologies provides us with more and more data and development of automated tools for data interpretation is necessary in order to process and understand results of such experiments. The main goal of the presented research is to analyze if and how information derived from the Gene Ontology (GO) database can be useful in the automated process of interpretation of gene groups obtained in expression level analysis.
A number of gene similarity measures based on Gene Ontology can be found in the literature but there is still a lack of complete studies that compare their performance. The first objective of this work is to propose new relatives-based GO terms similarity measure based on a granular approach and which allow comparing genes on a more general level. The second objective is to analyze existing similarity measures, compare them and evaluate in terms of clustering and correlation quality. We assume that good and efficient measure should reflect biological dependences among genes, therefore our conclusions are based on comparison with expression data from two different microarray experiments.

Following GO term similarity measures were analyzed and compared:

- Semantic term similarity
  - Information content
  - Jiang-Conrath
  - Lin
  - GraSM
  - G-SESAME
  - **Group and Group-soft relatives-based granular term similarity – new measures proposed**
- Path-based term similarity
- Binary Similarity
  - Jaccard measure
  - Czekanowski measure

Group and Group-soft are two new methods of Gene Ontology term similarity calculation based on the idea of granular analysis in order to compare ontology terms on more abstract and general level. In proposed approach, not a pair of terms is compared, but a pair of granules (sets) related to these terms is analyzed.

In the presented research we compare gene similarity in two representations: gene expression values and Gene Ontology graph. The rationale leading to such comparison is that genes that act in the same way (fact translating into similar expression patterns) should be similar in other representations, e.g., annotations to Biological Process Gene Ontology. Two types of analysis were performed:

- correlation of gene similarity in gene expression representation and GO representation,
- clusterability of the Gene Ontology data and comparison of clustering results in both representations,

From the clustering results perspective, gene similarity measures were used as a similarity/distance measures. Such analysis can show which similarity/distance measure gives the values making data objects more cohesive within a group and more easily separable between the groups, in other words, which measure gives a more clusterable data representation.

Two DNA microarray datasets were analysed: Eisen (Eisen et al, PNAS 1998) and Iyer (Iyer et al., Sciene 1999). The correlation and clustering quality results are presented in Table 1.

Table 1. Results of correlation and clustering analysis

	Correlation analysis		Clustering quality (NMI)	
	Eisen	*Iyer*	*Eisen*	*Iyer*
Binary Czekanowski	0.483	0.102	0.468	0.092
Binary Jaccard	0.475	0.119	0.468	0.092
Group	**0.571**	**0.124**	**0.569**	**0.103**
Group Soft	**0.572**	**0.136**	**0.705**	**0.109**
GSezame	0.522	0.088	0.526	0.072
Jiang-Conrath	0.412	0.104	0.518	0.109
Jiang-Conrath GraSM	0.427	0.112	0.597	0.121
Lin	0.36	0.088	0.444	0.123
Lin GraSM	0.385	0.103	0.526	0.103
Path	0.572	0.136	0.603	0.095
Resnik	0.458	0.085	0.45	0.092
Resnik GraSM	0.467	0.091	0.544	0.073
Weighted Czekanowski	0.477	0.11	0.592	0.094
Weighted Jaccard	0.461	0.125	0.592	0.094

Finally, to verify if the gene clusters obtained for the best measure (**Group Soft**) do have biological meaning, we analyzed their gene composition and compared the results with the reference partition for Eisen DNA microarray dataset. Analysis shows that our clusters have similar gene composition. In case of original cluster C (described by Eisen keyword Proteasome) and our group 7 we obtained identical partition. For other groups, differences were more visible, however typically it was not more than a few genes. In several cases we obtained group that consisted of reference groups merged together – for example gene composition of our group 1 is: CDC10, HTB2, HTB1, HHF1, HHF2, HTA2, HHT1, HHT2, HTA1, MCM7, DBF2, MCM4, MCM3 which mostly covers two Eisen groups: H which consist of genes: HTB2, HTB1, HHF1, HHF2, HTA2, HHT1, HHT2, HTA1 and J which consists of genes: MCM7, DBF2, MCM4, MCM3, MCM2. This result can be explained by the following facts. If we analyze the original dendrogram we can notice that genes composing clusters J and H are placed next to each other, therefore depending on selected cut-off value we can obtain one or two clusters. Another explanation of merging two clusters can be found by analyzing genes function. Original cluster H was described by Eisen by a keyword *chromatin structure* and includes, among others, genes HHF1, HHF2, HHT1, HHT2 that contribute to telomeric silencing. If we analyze biological function of MCM3 and MCM7 genes we can see that they also play a role in silencing and interact with the essential silencing chromatin factor, SIR2