BIOINFORMATICS APPLICATIONS NOTE
Vol. 26 no. 10 2010, pages 1384–1385
doi:10.1093/bioinformatics/btq133

Data and text mining
Advance Access publication March 26, 2010

# SICAGO: Semi-supervised cluster analysis using semantic distance between gene pairs in Gene Ontology

Bo-Yeong Kang[1], Song Ko[2] and Dae-Won Kim[2,*]
[1]School of Mechanical Engineering, Kyungpook National University, 1370 Sankyuk-dong, Buk-gu, Daegu 702-701 and [2]School of Computer Science and Engineering, Chung-Ang University, 221 Heukseok-dong, Dongjak-gu, Seoul 156-756, Korea

Associate Editor: Limsoon Wong

## ABSTRACT

**Summary:** Despite the importance of using the semantic distance to improve the performance of conventional expression-based clustering, there are few freely available software that provides a clustering algorithm using the ontology-based semantic distances as prior knowledge. Here, we present the SICAGO (SemI-supervised Cluster Analysis using semantic distance between gene pairs in Gene Ontology) system that helps to discover the groups of genes more effectively using prior knowledge extracted from Gene Ontology.

**Availability:** http://ai.cau.ac.kr/sicago.html

**Contact:** dwkim@cau.ac.kr

Received on November 26, 2009; revised on February 9, 2010; accepted on March 23, 2010

## 1 INTRODUCTION

Based on the assumption that genes with similar expression patterns are more likely to have similar biological functions, many clustering algorithms have been used to predict the function of a gene from the known functions of other genes in the cluster. However, most existing methods are limited in improving the clustering performance since microarray expression data are noisy. Noise comes from the damage caused by the experimental device or observing system error. Thus, it has been observed that some genes with similar expression patterns do not have common biological characteristics. In some cases, coregulated genes did not exhibit similar expression patterns.

New clustering approaches have been proposed to handle the limitations of the conventional methods. They attempted to incorporate prior biological knowledge into clustering algorithms, exploiting known gene functions in the process of clustering. The machine learning community terms this semi-supervised clustering. It uses existing domain knowledge to guide the clustering process. It is observed that incorporating biological knowledge into cluster analysis is effective in obtaining reliable clustering performance and interpreting the analysis results of expression data with high noise levels.

The best known approach to incorporating biological knowledge is to measure the semantic distance between known genes in Gene Ontology (Azuaje and Bodenreider, 2004; Lord and Stevens, 2003; Sevilla *et al*., 2005). This GO-based distance between two genes is incorporated into the expression-based distance during the process of clustering (Cheng *et al*., 2004; Fang *et al*., 2006; Huang and Pan, 2006; Kustra and Zagdanski, 2006). Information content theory has been widely used to measure a semantic distance between genes in GO.

Despite the importance of utilizing the semantic distance into cluster analysis, few freely available application programs can perform clustering using the ontology-based semantic distances as prior biological knowledge. We have developed a software program, SICAGO, a tool for SemI-supervised Cluster Analysis using semantic distance between gene pairs in Gene Ontology to satisfy these needs. SICAGO supports (i) three well-known semantic distance measures based on information content. Thus, we can see which type of semantic distance is most effective to construct prior knowledge using GO. It also supports (ii) three Gene Ontologies, so that we can analyze how the difference of semantic distance between two genes, according to GO structures, influences clustering performance. Finally, it supports (iii) the most widely used *k*-means clustering algorithm, so that we can analyze the effectiveness of semi-supervised clustering when prior knowledge was incorporated.

## 2 SICAGO OVERVIEW

Figure 1 shows the schematic diagram of SICAGO software. It performs three tasks: semantic calculation, semantic correlation and cluster analysis. A system interface lets users to upload an input file, and then select the type of task in order. The source is freely available and modified under GNU GPL.

Task I of SICAGO, semantic calculation, provides the semantic distances between gene pairs in GO. Three well-known semantic distance measures (Jiang and David, 1997; Lin, 1998; Resnik, 1995) are supported to achieve this. These measures are based on information content theory, where the semantic similarity between genes is determined by the amount of information they share in common. This shared information is computed using the set of parent nodes (genes); the more parents two genes share, the more similar they are. The ontology selection in SICAGO provides three domains of GO; biological process (BP) that indicates the operations of molecular events with a defined beginning and end; cellular component (CC) that indicates the parts of a cell or its extracellular environment; and molecular function (MF) that indicates the elemental activities of a gene product at the molecular level. After users select the type of semantic measure and type of ontology, SICAGO calculates all the semantic distances between genes in the ontology.

*To whom correspondence should be addressed.

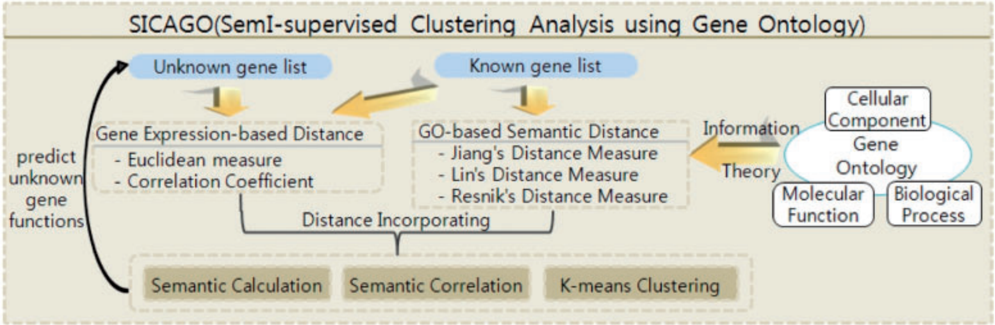

**Fig. 1.** Schematic diagram of SICAGO software that incorporates GO-based semantic distance into expression-based distance in the process of clustering.

Task II, semantic correlation, provides the semantic distance results between two genes obtained from Task I. SICAGO shows the comparative results for the two genes picked by users using three distance measures: expression-based Euclidean distance, GO-based semantic distance and correlation coefficient. Users can see the difference between expression and ontology-based distance results. We showed the distance between two genes in the ontology using a tree view to understand the semantic distance better.

Task III, cluster analysis, provides the semi-supervised $k$-means clustering algorithm. We amended the distance between a gene and a group calculation method to exploit the prior knowledge generated from the selected ontology and semantic distance. We modified the distance measure (dist) between a gene ($g$) and a group ($G$) to be a linear combination of their expression-based Euclidean/correlation distance and their semantic distance to select a cluster where a gene is assigned. Thus, dist$(g,G)$ is defined by:

$$\text{dist}(g,G)=\alpha\times E(g,G)+(1-\alpha)\times S(g,G) \qquad (1)$$

where

$$E(g,G)=\|g-\text{center}(G)\|^2 \qquad (2)$$

is an expression-based Euclidean distance,

$$S(g,G)=\frac{1}{n}\sum_{g_i\in G}\frac{\text{dist}_{\text{GO}}(g,g_i)}{M} \qquad (3)$$

is an ontology-based semantic distance. $n$ is the number of genes that have semantic distances to $g$, and $M$ is the maximum value of the semantic distances between $g$ and other genes. The weight coefficient $0\le\alpha\le 1$ controls the weights of the two distances.

## 3 RESULTS AND DISCUSSION

We validated the presented system on three well-known yeast and human gene datasets (Cho *et al.*, 1998; Eisen *et al.*, 1998; Spellman, 1998; Whitfield *et al.*, 2002). From the series of experimental results, we could see that the performance of the presented semi-supervised clustering method was surprisingly increased compared with that of the only expression-based clustering:

- The composition of the Euclidean distance with Lin/Resnik's measures showed the best performance in $k$-means clustering.
- The best $\alpha$ values in Equation (1) are $\alpha=0.1$ and 0.3 for Jiang's measure incorporated; $\alpha=0.3$ and 0.5 for Lin/Resnik's measures.
- The cluster obtained by SICAGO shows a high enrichment of GO functional categories for predicting the function of a gene.

Please refer to our website (http://ai.cau.ac.kr/sicago.html) for more information.

*Funding*: This work were supported by Basic Science Research Programs through the National Research Foundation of Korea (NRF) grand funded by the Korea government (MEST) (No.2009-0064861 and No.2009-0068718).

*Conflict of Interest*: none declared.

## REFERENCES

Azuaje,F. and Bodenreider,O. (2004) Incorporating ontology-driven similarity knowledge into functional genomics: an exploratory study. In *Proceedings of the IEEE Fourth Symposium Bioinformatics and Bioengineering*, Taichung, Taiwan.

Cheng,J. *et al.* (2004) A Knowledge-based clustering algorithm driven by Gene Ontology. *J. Biopharm. Stat.*, **14**, 687–700.

Cho,R.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Eisen,M.B. *et al.* (1998) Cluster analysis display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fang,Z. *et al.* (2006) Knowledge guided analysis of microarray data. *J. Biomed. Inform.*, **39**, 401–411.

Huang,D. and Pan,W. (2006) Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics*, **22**, 1259–1268.

Jiang,J. and David,W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics*, Tapei, Taiwan.

Kustra,R. and Zagdanski,A. (2006) Incorporating gene ontology in clustering gene exresspion data. In *Proceedings of the 20th IEEE International Symposium on Computer-Based Medical Systems*, Maribor, Slovenia.

Lin,D. (1998) An information-theoretic definition of similarity. In *Proceeding of the 15th International Conference on Machine Learning*, Madison, Wisconsin, USA.

Lord,P.W. and Stevens,R.D. (2003) Investigating semantic semiliarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada.

Spellman,P.T. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccaromyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Sevilla,J.L. *et al.* (2005) Correlation between gene expression and GO semantic similarity. In *IEEE Trans. Comput. Biol. Bioinform.*, **2**, 330–338.

Whitfield,M.L. *et al.* (2002) Indentification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.