GOODIES: GO Based Data Mining Tool for Characteristic Attribute Interpretation on a Group of Biological Entities

Sung Geun Lee[1] **Wan Seon Lee**[1] **Yang Seok Kim**[1,2]
sglee@istech21.com konan@istech21.com yskim@iste21.com

[1] Bioinformatics Unit. ISTECH Inc. #704, Hyundai Town Vill, 848-1 Janghang-dong, Ilsan-gu, Goyang city, Gyeonggi-do, 411-380, Korea
[2] Cancer Metastasis Research Center, Yonsei University College of Medicine, 134 Shinchon-dong, Seodaemun-gu, Seoul, 120-752, Korea

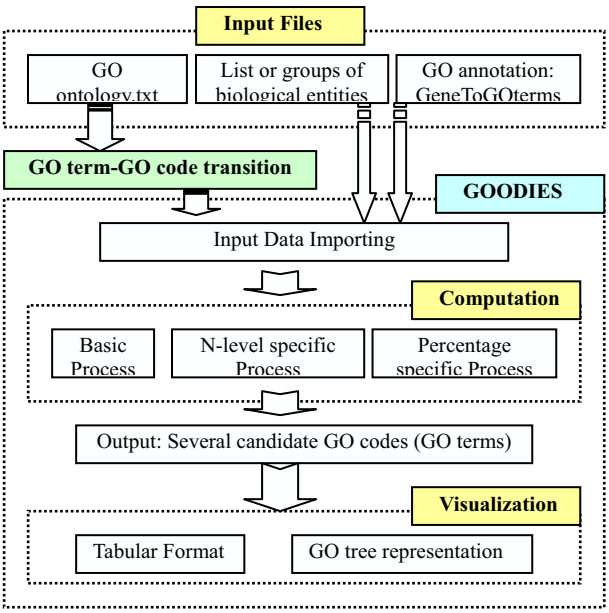
Keywords: Gene Ontology, GO tree, MaxPd, AverPd

1 Introduction

GOODIES^{TM} is a Gene Ontology (GO) based data-mining tool with intuitive visualization on a GO tree. Its algorithm uses the graph structure of GO to interpret and classify aggregates of biological entities [3]. Given gene or protein lists, e.g. gene clusters obtained from DNA chip experiments, GOODIES takes the multiple functionalities of genes into account and computationally selects the optimal GO candidate terms for the most suitable biological interpretation of given lists.

2 Method and Results

The main usage of GOODIES can be as follows: biologically-oriented cluster analysis of DNA microarray data, automated functional annotation via clustering, and functional categorization of biological objects. First, in cluster analysis of DNA microarrays, biologists primarily want to know how well clusters of expression profiles are associated with known functional categories and cellular processes. GOODIES can perform such tasks in terms of GO that it can be complementary to statistical clustering methods. Secondly, the unknown function of genes can be putatively predicted through the clustering interpretation of GOODIES.



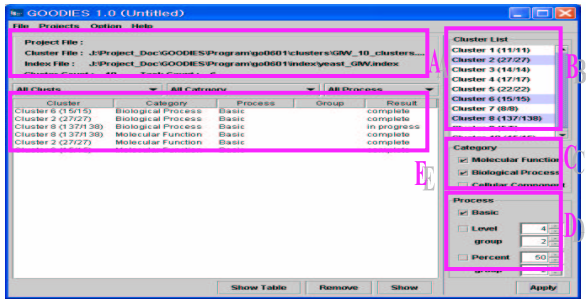


Figure 1: Schematic block diagram of GOODIES (left) and Sample analysis (right). Once GOODIES completes the matching process between the input groups and corresponding information from the GO annotation file, main window in A and B will be filled. After executing selected clusters(B), categories(C), and processes(D), GOODIES displays the results in E.

After the biological relationship among genes in each cluster is quantitatively estimated by *AverPd*, the clusters whose *AverPd* score is sufficiently low can be used for functional assignment of unknown

genes in those clusters. Thirdly, GOODIES can accomplish a large-scale functional categorization of biological entities - e.g. ESTs, genes, and proteins - according to the GO annotations of each entity that are extracted from reliable, curated databases [1, 2, 4].

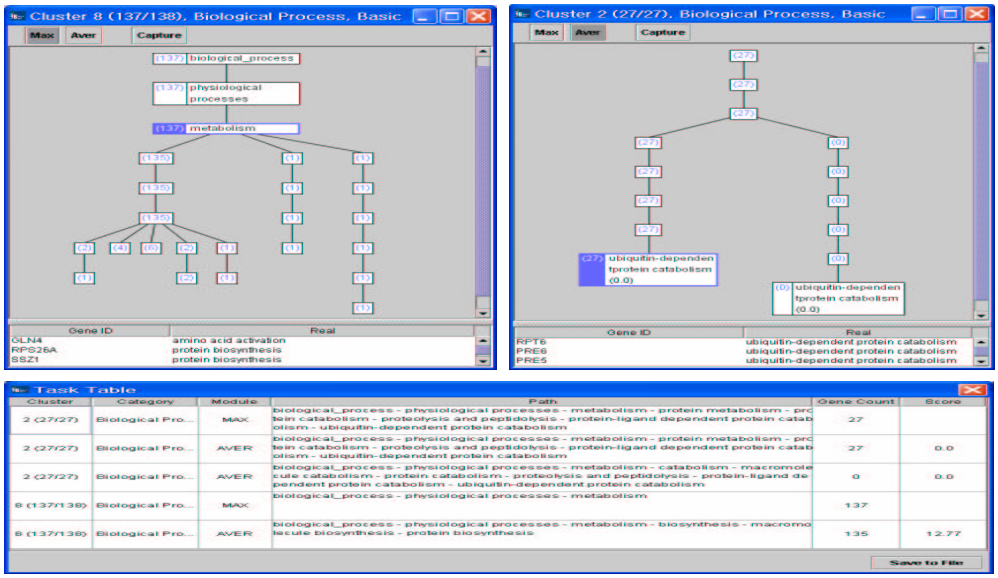


Figure 2: Graphical display of Basic Process results. *MaxPd* (upper left), *AverPd* (upper right) and table format summary (below).

3 Discussion

In a viewpoint of software engineering, choosing Gene Ontology [6], or any other ontology that is rapidly changing, as the core base for devising a computational algorithm may be regarded as risky in terms of robustness. Moreover, it may be suspected that the results can be fluctuating whenever GO gets updated. We have checked the result of [3] and other experiments periodically and the outcome was remarkably robust. The overall structure of GO seem to be undisturbed and consistent even though many new GO terms are being added to the old versions of GO.

Current GOODIES version entirely depends on the GO annotations. Although GO is widely accepted nowadays as a standard ontology protocol for functional annotation, in some cases it may be more sensible to employ species-specific or disease-centric terms that GO lacks. So we plan to construct customized GO categories and to use the information in OMIM. GO has been often used for functional prediction of unknown genes [5]. GOODIES can be another choice for that purpose. By a framework using *AverPd* score, functional prediction of genes can be automatically taken coupled with DNA microarray cluster analysis. In the near future, as more information is accumulating about the genomes, GOODIES will be a more helpful data-mining tool in functional genomics.

References

[1] Camon, E. *et al.,* The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro, *Genome Res.,* 13:662–672, 2003.

[2] Dwight, S.S. *et al.*, Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO), *Nucleic Acids Res.*, 30:69–72, 2002.

[3] Lee, S.G. *et al.*, A graph-theoretic modeling on GO space for biological interpretation of gene clusters, *Bioinformatics*, in press.

[4] Wheeler, D.L. *et al.*, Database resources of the national center for biotechnology, *Nucleic Acids Res.*, 31:28–33, 2003.

[5] Zhou, X. *et al.*, Transitive functional annotation by shortest path analysis of gene expression data, *Proc. Natl. Acad. Sci. USA*, 99:12783–12788, 2002.

[6] `http://www.geneontology.org/`