Report of project

Incorporating Semantic Similarity Information Into Functional Genomics

LHNBC - Visiting Faculty Program, August 8th, 2003.

By Francisco Azuaje.
Advisor: Olivier Bodenreider.

Summary

This research explored the feasibility of semantic similarity approaches to supporting predictive and validation tasks in functional genomics. It aimed to establish potential relationships between the semantic similarity of gene products and important functional properties, such as gene expression correlation and protein complex membership. Semantic similarity measures based on the information content of the Gene Ontology (GO) were analyzed. Models were implemented using data obtained from well-known studies in *S. cerevisiae*. Results suggested that there may exist a significant and non-linear relationship between gene expression correlation and semantic similarity. Analyses of protein complex data showed that in general there is a significant correlation between the semantic similarity exhibited by a pair of genes and the probability of finding them in the same complex. Moreover, for this type of data, this study illustrates a relationship between the three GO aspects: molecular function, biological process and cellular function in terms of semantic similarity. This research will be reported in a paper, which will be submitted to the journal *Comparative and Functional Genomics*.

Semantic similarity and expression data

The integration of gene expression correlation and semantic similarity is based on data that characterize mRNA transcript levels during the cell cycle of *S. cerevisiae*. Semantic similarity analyses were performed on 225 genes that show significant and periodic transcriptional fluctuations during five cell cycle phases. 25200 pairs of semantic similarity values and 25200 expression correlation values were calculated. Graphical analyses and ANOVA were implemented to visualize potential relationships between semantic similarity, expression correlation and cell cycle phases.

Semantic similarity and protein complex membership

Associations between semantic similarity and protein complex membership were studied using two datasets. The first dataset consisted of 83 pairs of proteins. Each pair is labeled on the basis of their membership (or non-membership) in the same protein complex in *S. cerevisiae*. There are 46 pairs categorized as belonging to the same complex (true positives), using the MIPS complexes catalogue as the gold standard. Graphical analyses and ANOVA were performed to visualize potential correlations between semantic similarity of a pair of genes and the probability of finding them in the same complex. A second analysis was conducted using a larger protein complex dataset. It processes 4190 pairs of proteins that form distinct multi-protein complexes in yeast. Semantic similarity values were obtained for each pair, and their frequency distribution and dependencies were studied on the basis of the three GO taxonomies.

Results and conclusions

This research indicates that in general one may expect a strong connection between the degree of GO-based similarity and the expression correlation of two gene products. This relationship is perhaps more clearly illustrated in the case of pairs of genes showing a low expression correlation and weak semantic similarity. It also illustrated how semantic similarity may be used to describe differences between expression clusters. Based on the analysis of a relatively small number of protein pairs, this study suggests that pairs assigned to the same complex may exhibit strong semantic similarity. Such similarity levels may be significantly stronger than the similarity shown by protein pairs belonging to different complexes.

The existence of a strong relationship between semantic similarity and gene expression correlation may be applied to support functional prediction applications together with other genomic resources or predictive models. It may be used, for example, to evaluate or validate clusters of genes or similarity-based predictions using different types of data. Semantic similarity models may be used to automatically label clusters of genes in terms of their compactness or intra-cluster similarity attributes. They could also be applied to support annotation tasks. For instance, groups of gene products could be annotated using their lowest common ancestor rather than multiple annotations. These models may also contribute to assess differences in annotations across genes, within a database or across multiple model organisms.

This research will be reported in a paper, which will be submitted to the journal *Comparative and Functional Genomics*. We expect to expand this collaboration. Future research should provide stronger evidence to support these claims. It should investigate semantic similarity applications such as expression cluster validation and functional classification of gene products. It will be necessary to incorporate larger sets of gene expression data in yeast and other model organisms. We also aim to analyze larger protein complex datasets, consisting of thousands of pairs of proteins.