Semantic Similarity-based Validation of Human Protein-Protein Interactions

[1]Xiang Guo, [2]Craig D. Shriver, [1]Hai Hu, and [1]Michael N. Liebman
[1]*Windber Research Institute, Windber, PA 15963, {s.guo, h.hu, m.liebman}@wriwindber.org*
[2]*Walter Reed Army Medical Center, Washington, DC 20307, craig.shriver@na.amedd.army,mil*

Biological networks are fundamental to understanding the dynamics of human health and disease. They are built based on the identification of protein-protein interactions. Traditionally, information about protein interactions was collected from the small-scale screens. The accuracy of each interaction has often been validated with multiple experiments. With the development of high-throughput methods such as the two-hybrid assay and protein chip technology, the information within interaction databases has increased tremendously. However, large-scale protein interaction assays are notoriously noisy. Therefore, it is essential to develop strategies to validate the high-throughput data sets.

Different evidence has been used to validate the high-throughput protein-protein interaction data. Among them, functional association of protein pairs is used to verify the biological relevance of putative interacting proteins. It's a reliable measure for the validity of protein interactions [1]. Traditionally, functional association has been assessed by the shared annotation of proteins in a controlled vocabulary system [2]. However, those methods are restricted to protein pairs having the same annotation. Human proteins have a lower level of accurate annotation than proteins in other organisms such as yeast. The percentage of human proteins sharing the same annotation is low, and the shared annotation may be too general to verify the functional association of two proteins. Therefore, the current methods may not be applicable to human interactome analysis.

Gene Ontology (GO) is a controlled vocabulary of over 17,000 terms used to describe biological process, molecular function and cellular component of genes and gene products in a generic cell. GO terms and their relationships are represented in the form of directed acyclic graphs (DAGs). Given a pair of terms, the traditional method for measuring similarity is to calculate the path distance between two nodes associated with these terms. Edges are weighted according to the depth of GO. This approach assumes that nodes and links in ontology are uniformly distributed, which is not accurate in GO. An alternative approach is based on the amount of information two terms share. The more information two terms share, the more they are similar. The shared information is indicated by the information content of the terms that subsume them in the DAG. The information content *p(t)* is defined as the frequency of each term, or any of its children occur within a corpus. Less frequently occurring terms are "more informative". Since GO allows multiple parents for each term, the similarity score between two terms can be defined as

$$sim(t1,t2) = -\ln\left(\min_{t \in S(t1,t2)} \{p(t)\}\right)$$

where *S(t1, t2)* is the set of terms that subsume both t1 and t2. This measurement has been shown to be significantly correlated with sequence similarity [3].

In this study, we perform a quantitative assessment on the application of GO-based semantic similarity measures in human protein interaction analysis. The ability of different similarity measures to discriminate true positive and false positive protein interactions is assessed using receiver operating characteristic (ROC) graphs. Our results indicate that two **s**emantic similarity measures based on GO biological process and molecular function annotation can be used to stratify human protein interactions. However, the similarity measurement based on GO cellular component annotation may not have the ability to discriminate true protein interactions from false positives. The combination of two measures yields better performance as compared to using either GO category alone for the classification. In addition, GO-derived semantic similarity measures have been shown to be valuable for the characterization of biological pathways. We believe that the integration of these measures with other information such as gene expression and network topology would greatly reduce the error rate in high-throughput protein interaction data.

[1]. Sprinzak, E., Sattath, S. and Margalit H. (2003) How reliable are experimental protein-protein interaction data? *J. Mol. Biol.* **327**, 919-923.

[2]. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302,** 449-453.

[3]. Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**,1275-1283.