

## BIOGRAM: Algorithms for Identifying Similar Proteins by Functional Annotation

Judith D. Cohn, Susan M. Mniszewski, CCS-3; Hong Cai, Jennifer F. Harris, B-7; Ruy M. Ribeiro, T-6; Cliff A. Joslyn, PNNL; Karin M. Verspoor, University of Colorado Health Sciences Center

In the study of complex biological systems, it is often useful to identify proteins that exhibit a similar functional signature within a cell or organism but do not necessarily have homologous sequences or any single biological feature in common. BIOlogical GRAPHical Measurement (BIOGRAM) is a software system under development that uses the mathematical structure underlying the controlled vocabulary of the Gene Ontology (GO) [1] to measure protein similarity based on annotations assigned to nodes in the three branches of the GO: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). This system builds upon the Partially Ordered Set Ontology Categorizer (POSOC), a software tool for automating protein functional annotation [2,3].

BIOGRAM uses a hierarchical adaptation of information retrieval (IR) performance statistics (recall, precision, and f-score) to measure protein similarity based on annotations assigned to the nodes of the GO. The network structure of the GO along with the use of hierarchical statistics allows us to make fuzzy comparisons of functional profiles, giving partial credit for matches to nearby nodes. Figure 1 illustrates the calculation of hierarchical recall (HR), hierarchical precision (HP), and hierarchical f-score (HF, the geometric mean of HR and HP) within a partially ordered set comparing the functional annotations of a target protein with those of a second protein (other). BIOGRAM is being developed to guide selection of experimental targets in the context of modeling host-pathogen interactions (pathomics) in avian influenza.

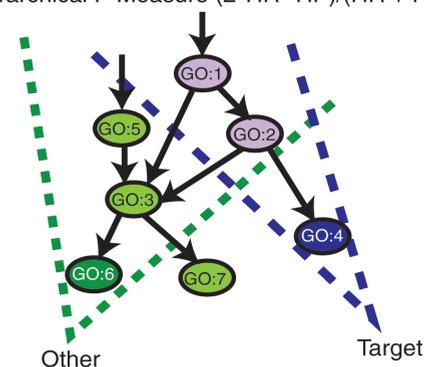
GO annotations for all human proteins in the Swiss-Prot database were obtained from the European Bioinformatics Institute (EBI). Our analysis was performed using

*Fig. 1. Hierarchical recall (HR), hierarchical precision (HP), and hierarchical f-score (HF) calculations are shown in context of a partially ordered set structure, such as the Gene Ontology (GO). The use of overlapping filters allows for a fuzzy measure of similarity.*

annotations drawn from all evidence codes (13,841 proteins) or a subset obtained only from curated evidence codes (9,090 proteins). In most cases, there were multiple annotations for each protein in each GO branch. Thus far, we have analyzed a number of proteins known to be involved in the response to influenza infection in humans by comparing annotations in the BP branch of the GO for each target protein (for example, tumor necrosis factor alpha or [TNFA]) against those of all other human proteins. A partial list of TNFA annotations includes response to virus, leukocyte adhesion, and anti-apoptosis. Only a small set of proteins show a high level of similarity to TNFA, whether using all or the curated subset of annotations (see Fig. 2). Protein similarity can be determined by rankings based on HR, HP, or HF. The top ten human proteins similar to TNFA based on HR using curated annotations are shown in Fig. 3. A detailed exposition of the specific annotations that underlie the hierarchical statistics can be obtained for any pair of proteins (e.g., TNFA and BCL3).

Future plans include the addition of reaction pathway and protein-protein interaction data.

HR = Hierarchical Recall  
 HP = Hierarchical Precision  
 HF = Hierarchical F-Measure  $(2 \cdot HR \cdot HP) / (HR + HP)$



$G(\text{Target}) = \{\text{GO:4}\}$ ,  $\text{filter}(\text{GO:4}) = \{\text{GO:1}, \text{GO:2}, \text{GO:4}\}$   
 $G(\text{Other}) = \{\text{GO:6}\}$ ,  $\text{filter}(\text{GO:6}) = \{\text{GO:1}, \text{GO:2}, \text{GO:3}, \text{GO:5}, \text{GO:6}\}$   
 $HR = 2/3$ ,  $HP = 2/5$ ,  $HF = 1/2$

For more information contact Susan Mniszewski at [smm@lanl.gov](mailto:smm@lanl.gov).

- [1] The Gene Ontology website, <http://www.geneontology.org>.
- [2] C.A Joslyn et al., *Bioinformatics* **20** Suppl 1, i169-77 (2004).
- [3] K.M. Verspoor et al., *Prot. Sci.* **15**(6), 1544-9 (2006).

TNFA: Cumulative Fraction Plot of Hierarchical Recall for Biological Process

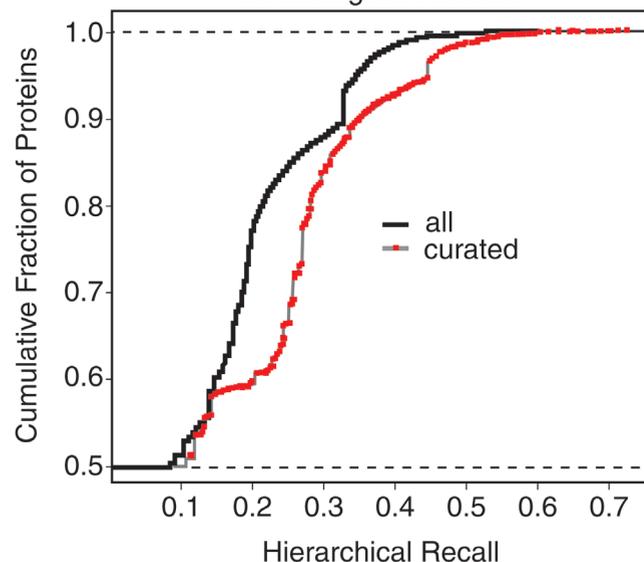


Fig. 2. Cumulative fraction plot of hierarchical recall (HR) for TNFA against all human proteins: comparison between results obtained from annotations of all evidence types (black) vs curated only (red).

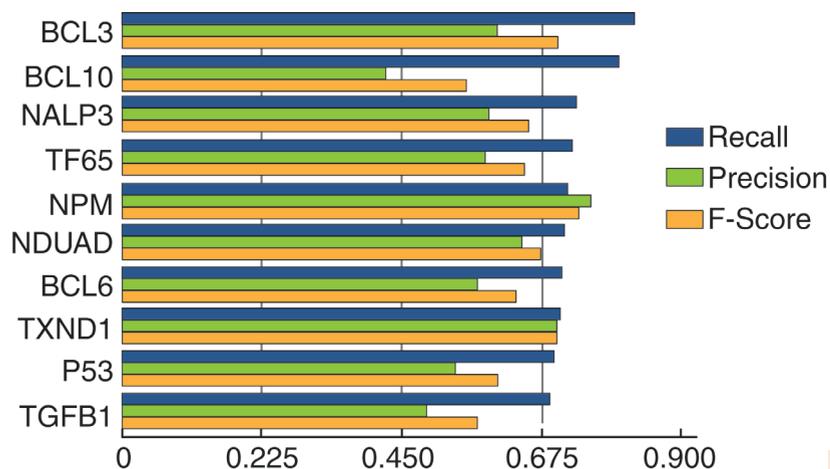


Fig. 3. The top ten human proteins that were determined to be similar to TNFA based on the BIOGRAM hierarchical recall (HR) are shown, along with their corresponding hierarchical precision (HP) and hierarchical f-score (HF) values.

**Funding**

**Acknowledgments**

LANL Directed Research and Development Program