BIOGRAM: Identifying Similar Proteins by Functional Annotation

Judith D. Cohn[1], Susan M. Mniszewski[2], Jennifer F. Harris[3], Hong Cai[4] and Ruy M. Rubeiro[5]

***Short Abstract* — In the study of complex biological systems, it is often useful to identify proteins with a similar functional signature. BIOGRAM (BIOlogical GRAphical Measurement) is a software system in development, which uses the mathematics of partially ordered sets to measure protein similarity based on annotations assigned to nodes in the three branches of the Gene Ontology. We present preliminary results from the use of BIOGRAM to guide the selection of experimental targets in the context of modeling host-pathogen interactions (pathomics) in avian influenza.**

***Keywords* — functional signature, protein similarity, Gene Ontology, pathomics, avian influenza.**

I. Introduction

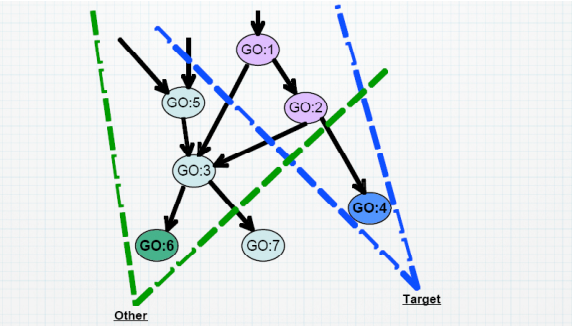
In the study of complex biological systems, it is often useful to identify proteins which are similar to a target protein. For example, it may be necessary to reduce the number of target proteins for either experimental or computational analysis by constructing a "non-redundant" set of proteins. Quite commonly non-redundant sets are selected using sequence homology, structural similarity, or both (e.g. the Astral Compendium for Sequence and Structure Analysis [1]). However, in many cases, it may be more useful to assemble a reduced dataset on the basis of functional signature rather than sequence or structure. Grouping proteins by function may also be helpful in substituting for a target protein which is found to be unsuitable for a variety of reasons.

II. Methods

BIOGRAM (BIOlogical Graphical Measurement) is a software system under development, which uses the mathematical structure underlying the Gene Ontology (GO) [2] to measure protein similarity based on annotations assigned to nodes in the three branches of the GO: Biological Process, Molecular Function, and Cellular Component. This system builds upon POSOC (POSet Ontology Categorizer), a software tool for automating functional annotation [3]. In particular, BIOGRAM incorporates the measures of hierarchical recall (HR) and hierarchical precision (HP), which are used in POSOC to evaluate annotation performance. Figure 1 illustrates the calculation of HR, HP and HF (hierarchical F-score: the geometric mean of HR and HP) in BIOGRAM when comparing the functional annotations of a target protein with those of a second protein (other) in the simplest case where each protein is annotated to a single GO node (target=blue, other=green).

Initially, BIOGRAM is being developed as a tool to guide the selection of experimental targets for the study of host-pathogen interactions (pathomics) in avian influenza.

Figure 1



III. Results

We will present preliminary results from running BIOGRAM to compare functional signatures of human target proteins selected for their response to influenza infection versus all other human proteins with GO annotations. We will also discuss future plans for BIOGRAM, including additional measures of similarity and an appropriate GUI interface.

Acknowledgements: This work was funded by the US Department of Energy through contract DE-AC52-06NA25396.

[1]Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545. E-mail: jcohn@lanl.gov

[2]Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545. E-mail: smm@lanl.gov

[3] Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545. E-mail: jfharris@lanl.gov

[4]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545. E-mail: cai_hong@lanl.gov

[5]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545. E-mail: ruy@lanl.gov

References

[1] Chandonia JM et al (2004). The ASTRAL compendium in 2004. *Nucleic Acids Research* **32,** D189-192.

[2] Gene Ontology website, http://www.geneontology.org

[3] Verspoor et al (2006). A categorization approach to automated ontological function annotation. *Protein Science* **15**, 1544-1549.