

Modeling Gene Ontology terms using Finite State Automata

Christos Gkekas*, Fotis E. Psomopoulos† and Pericles A. Mitkas‡

Intelligent Systems and Software Engineering Lab, Dept. of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece 54124, <http://issel.ee.auth.gr>

**email: chggr@egnatia.ee.auth.gr †email: fpsom@danae.ee.auth.gr (Corresponding author) ‡email: mitkas@eng.auth.gr*

Protein classification is one of the most commonly discussed problems in bioinformatics. Due to the fact that proteins perform numerous and diverse functions in a cell, there exist quite a few metrics for describing their functionality. One of the latest tools for protein function annotation is the Gene Ontology¹ (GO) project. It provides a controlled vocabulary to describe gene and gene product attributes in organisms.

However, GO can be thought of as a database of expert-based terms. Although there are several cases where a term is derived through automated inference, the bulk of the annotation process is performed by human curators. Any new protein sequence must be either processed directly in a lab, or characterized through similarity to an already annotated sequence. In this paper a novel methodology is presented, which utilizes the motifs that are present in annotated protein sequences, in order to model the corresponding GO terms.

The first step in the methodology is the creation of the protein training sets. Using the UNIPROT code of each protein and the InterProScan² tool, we extract the motifs present in its sequence. This process allows us to take under consideration every available sequence database (such as PRODOM, PRINTS, PFAM, PROFILE etc) together with any GO annotation correlated to the specific sequence. For each GO term that appears in the original protein set, a new training set is created, which contains all the protein sequences that have been annotated with the specific GO term.

In the next step, each of the produced training sets is processed independently in parallel; first a PrefixTree Acceptor(PTA) is constructed using the motif sequence of the proteins in the training set. This PTA is consequently transformed into a more generalized Stochastic Finite State Automaton (FSA), which in turn can be used to model the corresponding GO term. In order to predict the annotation of an unknown protein, its motif sequence is run through each GO model thus producing similarity scores for every term.

The methodology has been implemented so it can be used both as a standalone or as a grid-based application. Although the process itself is efficient, the Grid provides for the seamless integration of the training process and the actual model evaluation, by allowing the concurrent retraining of GO models from different input sources or experts and the use of the existing ones. This methodology can be easily generalized to produce models for different protein classification schemes, such as SCOPfamilies etc.

Submitted to: Hellenic Bioinformatics and Medical Informatics Meeting, 4-5 October, 2007

1<http://www.geneontology.org/>

2<http://www.ebi.ac.uk/InterProScan/>