**CAFA Training Data**

This archive contains training data that can be used for the development of prediction models of the Gene Ontology (GO) terms. The data was extracted from four different databases (UniProt-GOA, GO, Swiss-Prot, and BioCyc). Each of these data sets contains a set of protein sequences (in the FASTA format) as well as experimentally characterized GO terms (using Molecular Function, Biological Process, and Cellular Component ontologies) corresponding to these sequences. Note that the GO terms are not propagated towards the root of the respective ontology.

The proteins contained in these data sets overlap but the sequence IDs are not necessarily the same across databases. In some cases, the same protein may have the same functional annotation in different databases and in other cases the same protein may have exactly the same annotation. This is because the databases are maintained by different biocuration teams. It is important to mention that the information about the organism (given by NIH taxon ID) from which the protein was taken is important: two identical protein sequences taken from two different organisms do not necessarily have the same function (e.g. their binding partners may be present in one but missing in the other species).

Other types of data (protein-protein interactions, gene expression, structure, etc.) can be parsed from a variety of databases. These resources are discussed in the document

Radivojac P. A (not so) quick introduction to protein function prediction,

that is posted on the CAFA web site. This document also contains a brief introduction to proteins, formulation of the protein function prediction problem and a discussion of the Critical Assessment of Functional Annotation (CAFA) experiment.