

metaProt: A Proteomics Metaclassification

Anand Kumar^{ab}

^aLaboratory of Medical Informatics, Dept. of Computer Science, University of Pavia, Italy.

^bInstitute for Formal Ontology and Medical Science, University of Leipzig, Germany.

Abstract

There is a recognized need for a proteomics meta-classification which can serve as a foundation for more refined classifications and ontologies. metaProt is a manually curated proteomics metaclassification, comprising a large collection of relations integrating SCOP, GO and SWISS-PROT.

Keywords: Proteomics, Metaclassification, Ontology, Gene Ontology, SCOP, SWISS-PROT.

Introduction

A wide range of factors needs to be taken into account for a complete description of a protein's function, including cellular roles, molecular functions and the involvement of proteins in physiology and pathology [1].

metaProt is based on the notion that existing proteomics databases and related ontologies cannot be integrated semantically without certain changes being made within them and thus it includes both modification in the existing relations and the addition of new ones.

Structural Classification of Proteins

SCOP (<http://scop.mrc-lmb.cam.ac.uk/scop/>) assigns proteins to 1110 superfamilies.

SCOP asserts that it will classify proteins exclusively in terms of structure, however, it mixes topology with structure with its classification. Classes like *Membrane and cell surface proteins and peptides* have been modified within metaProt, since it suggests a classification based on location.

Gene Ontology

The three root classes in GO are *molecular function*, *biological process* and *cellular component* (<http://www.gene-ontology.org/>), each of the three constituent taxonomies being organized via **is-a** and **part-of** relations between the different classes.

Replacing is-a by is-part-of relations: There are situations when GO treats parthood relations between processes as subsumption relations. They have been corrected within metaProt.

Distinguishing functions and activities: Functions cannot be types or parts of processes as present in GO; instead, they should be types or parts of other functions, just as processes should be types or parts of other processes.

Problems with synonyms: There are many terms and their synonyms within GO's *molecular function* axis which actually refer to substances and are reassigned accordingly.

Aligning molecular activity with biological process: The difference between biological process and molecular activity is based on granularity. Instead of keeping them as separate root classes, they are aligned within the same hierarchy in metaProt.

SWISS-PROT

SWISS-PROT (<http://www.expasy.org/sprot/>) is a curated protein sequence database with various annotations. metaProt's reuration of SWISS-PROT is designed to interconnect different types of information related to proteins. New classes are introduced as necessary in order to make the ontology compatible with the modified GO terms. metaProt also specifies the locations of proteins mentioned in SWISS-PROT within GO more accurately.

References

[1] Liu J and Rost B. Comparing function and structure between entire proteomes. *Protein Sci.* Oct;10(10):1970-9, 2001.

Acknowledgments

Work on this paper was supported by the Wolfgang Paul Program of the Alexander von Humboldt Foundation.

Address for correspondence

Anand Kumar, Laboratory of Medical Informatics, Deptt. Computer Science, University of Pavia, Italy 27100. email: anand@aim.unipv.it