

Possibilities of the automatic classification of protein functions from the literature

Christian Blaschke and Alfonso Valencia

Protein Design Group, National Center for Biotechnology, CNB-CSIC
Cantoblanco, Madrid E-28049, Spain
valencia@cnb.uam.es <http://www.pdg.cnb.uam.es>

Detailed classifications, control vocabulary and organized terminology are widely used in different areas of science and technology. Their relatively recent introduction in molecular biology has been crucial for the progress in the analysis of genomics and proteomics massive experiments. Unfortunately the construction of the ontologies, including terminology, classification and entity relations requires considerable effort including the analysis of massive amount of literature. I will introduce here our recent efforts for automatic generation of classifications of gene-product functions using bibliographic information.

The initial results show a good correspondence of the classification structures with the ones constructed by human experts. The analysis of a large structure built for yeast gene-products, and the detailed inspection of various examples, show encouraging properties. In particular, the comparison with the well accepted GO ontology points to various situations in which the automatically derived classification can be useful for assisting human experts in the annotation of ontologies.

Biographical notes (Alfonso Valencia):

Born in Sevilla (Spain), bachelor degree in Biology by the U. Complutense (Madrid) and PhD degree in Biochemistry by the U. Autonoma (Madrid). Dr. Valencia main scientific interest is associated to the use of genomics and proteomics for the study of molecular evolution and for the development of new biotechnological resources. His scientific interest requires the development of Bioinformatics and Computational Biology methods, to which he has dedicated his scientific activity.

His specific training in Bioinformatics was acquired during a postdoctoral work in the group of Dr. Chris Sander, EMBL- Heidelberg (1988 – 1994). Since his incorporation to the Spanish National Center for Biotechnology he has created a multidisciplinary group of 20 researches, including biologist and computer scientist. The Protein Design Group has contributed to the development of various bioinformatics systems, and collaborates actively with experimental biologists in the study of different molecular systems, such as small GTPases and bacterial cell division proteins. The scientific activity of his group includes the analysis and comparison of genomes, prediction of protein structure and function, analysis of protein interactions, and more recently the extraction of information from scientific texts. His group at the CNB-CSIC has published more than 50 articles in relevant international journals, actively participated in the organization of meetings (i.e. system biology, Granada June 2002; Protein interactions, Verona, July 2002; Bioinformatics and proteomics, El Escorial August 2002; Training in Bioinformatics, Madrid July 2002), and frequent lecture in European and American institutions. The group is financed with National and European agencies and by collaborations with different companies in the biotechnology and pharmaceutical sectors.

Dr. Valencia is senior scientist of the Spanish research council (CSIC), belongs to the editorial board of the journal Bioinformatics, is the Coordinator of the Spanish Network of Bioinformatics, chair of the workshop committee of the Functional Genomics program of the ESF, and former Vicepresident of the International Association for Computational Biology (ISCB).

