



INTERNSHIP OPPORTUNITY

Project Description: Ontological Clustering

In the last few years we have seen the rise of ontologies in informatics, particularly in the realm of bioinformatics. We are now at the point where we should be able to move beyond the description of biological entities using ontological terms to statistical analysis of groups of biological entities according to their ontological attributes.

The project will be to create software, and associated database tables, that will cluster entities according to their attributes where the attributes are terms or nodes in one or more ontologies (the ontologies may be arranged as trees or as directed acyclic graphs). The algorithm should be able to find clusters in the space of all ontologies or in the space defined by arbitrary subsets of ontologies. The algorithm should return distance or probability measures that describe the significance of the clusters. The algorithm should calculate the weights given to probabilities or distances between nodes as roughly proportional to the distance from the root.

The student will have access to expert molecular biologists for the purposes of testing and validation of the software.

Languages: Java, Perl, or C

Project Description: Network Comparison

Public pathway databases are growing rapidly and this profusion of data is allowing researchers to develop analytical tools that can be used to generative comparative and predictive data. The project is to create an algorithm and associated software that could be used to find similar networks given one network where some fraction of the nodes are described with terms from one or more tree or DAG ontologies. The algorithm should be able to evaluate similarities using all ontologies or an arbitrary subset of ontologies. The algorithm should also be able find similarities and return meaningful statistical measures of similarity.

The student will have access to expert molecular biologists for the purposes of testing and validation of the software.

Languages: Java, C, or Perl

Project Description: Automated Classification using Motifs and Ontologies

Many public databases contain sequence information as well as reliable human annotation of protein sequence, and much of this annotation is in the form of ontology terms. One should be able to automatically match protein sequences to Hidden Markov Models constructed from public domain "motif" databases and correlate these matches to the corresponding annotations. The aim is to assign motifs, or groups of motifs, to positions on various ontological graphs and to use these assignments to automatically classify protein sequences.

Interesting problems should arise as motifs will be assigned to different positions on the ontological graphs – the solution may be through the use of measures of statistical significance.

Languages: Java or Perl

About Cognia

Cognia Corporation is a developer and distributor of information products and services that facilitate the use of biological and chemical information. Cognia's information solutions help pharmaceutical and biotechnology companies cost-effectively accelerate discovery and research processes.

Cognia's business/product model was developed through experience and use with over one hundred Pharmaceutical, Biotechnology and Research Centers. Customers like Merck & Co., GlaxoSmithKline, Aventis, Schering-Plough Research Institute, and research groups at academic institutes such as Harvard Medical School, The Rockefeller University and The Whitehead Institute.

Our core product, Cognia Molecular™, enables industry and academic users alike to integrate, manage, and utilize biological and chemical information from many formats and sources in support of drug discovery and basic research. Cognia Molecular can be augmented with Cognia's value-added content products and services, as well as most datasets and systems our customers may already be using.

Cognia also distributes content products such as the gene regulation database products of BIOBASE GmbH and the mass spectral and anti-microbial chemical database products of John Wiley & Sons.