The IU-IBM Protein Family Annotator Project Final Report

Background

In June 2002, IU and IBM signed a joint study agreement (W0125820) to develop protein family annotator software as part of a larger IU-IBM life sciences partnership. IBM provided $35,000 for the project. Matching funds were provided by IU's Indiana Genomics (INGEN) project and School of Informatics to procure hardware (two IBM Intel servers) and to be applied toward a programmer's salary. This document serves as the final report on the joint study agreement.

The Curation Alignment Tool for Protein Analysis (CATPA)

To address the deficiencies in protein family annotation and curation, a "Curation Alignment Tool for Protein Analysis" (CATPA) was developed at the IU School of Informatics by Dr. Mehmet Dalkilic. CATPA is a full-fledged information system for biologists that creates, stores, manages, and queries curated protein families. CATPA utilizes a GUI front-end that provides biologists with an environment they are accustomed to, while at the same time providing the integrity and power of a local, stand-alone database at the backend. CATPA incorporates the GO ontology in its curation vocabulary.

CATPA recognizes a number of well-known and widely used formats both for importing to and exporting from the system. CATPA utilizes a Java GUI front-end that allows biologists to interact with information in an environment to which they are accustomed. Protein families are aligned, conservation and curation are easily discernable via colors that users can change according to their preferences. Additionally, other kinds of information can be displayed, e.g, entropy, hydrophobicity.

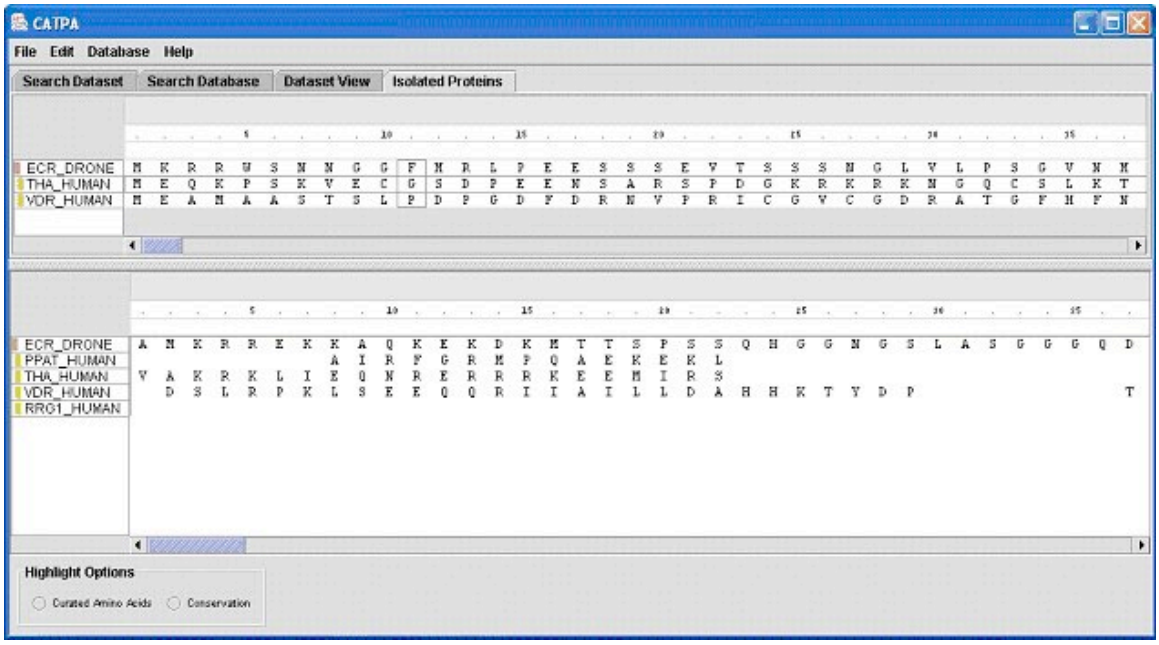

Figure 1. CATPA interface that allows segration of interesting proteins.

Activities

- A paper on CATPA titled "Design and Evaluation of CATPA: Curation and Analysis Tool for Protein Analysis" by Mehmet Dalkilic and Arijit Sengupta has recently been submitted to "Special Issue of the ACM Transactions on Information Systems: Genomic Information Retrieval." In the paper, IBM and other partners in the projects are duly acknowledged.
- A CATPA web site (http://www.protein.informatics.iu.edu/) will soon be available to provide the software freely for academic use.
- A NSF ITR proposal is being developed for submission in early 2004 to seek external funding to continue the project.
- A study to explore using DB2 as the database back-end and IU's IBM Regatta node for a production instance of CATPA is under way.

Future

The vision is to institute CATPA metaservers globally to allow groups of biologists who work with specific families, e.g., nuclear receptor, to curate and share information. The main metaserver will be housed at IU. The query mechanism includes both textual and graphical.