YPED: A Proteomics Database for Protein Expression Analysis

Mark A. Shifman, MD, PhD[1,2], Kexin Sun, MD, MS[1,2], Christopher M. Colangelo, PhD[3], Kei-Hoi Cheung, PhD[1,2,4], Perry Miller, MD, PhD[1,2,5], Kenneth Williams, PhD[3,6]

[1]Center for Medical Informatics, [2]Department of Anesthesiology,
[3]Keck Biotechnology Resource Laboratory, [4]Department of Genetics
[5]Department of Molecular, Cellular and Developmental Biology
[6]Department of Molecular Biophysics and Biochemistry
Yale University, New Haven, CT

Abstract: We have developed the Yale Protein Expression Database (YPED) to address the storage, retrieval, and integrated analysis of proteomics data generated by Yale's Keck Protein Chemistry and Mass Spectrometry Facility. YPED is Web-accessible and currently handles sample requisition, result reporting and sample comparison for ICAT, DIGE and MUDPIT samples. Sample descriptions are compatible with the evolving MIAPE standards. Peptides and proteins identified using Sequest or Mascot are validated with the Trans-Proteomic Pipeline developed at the Institute of Systems Biology and data from the resulting XML file are stored in the database. Researchers can view, subset and download their data through a secure Web interface.

Recent advances in analytical chemistry and mass spectroscopy have made large-scale analysis of the human proteome feasible. As new bioinformatics tools are facilitating the identification and quantitation of proteins and peptides in complex mixtures, vast quantities of proteomic data are being generated. To handle the growing needs of researchers at Yale, we have developed the Yale Protein Expression Database (YPED).

YPED (http://info.med.yale.edu/proteome) was developed using an Oracle database for data storage and retrieval. The Web interface was built using Java, Tomcat and Struts.

The system currently deals with three types of proteomic samples: MUDPIT (Multidimensional Protein Identification Technology), DIGE (Two dimensional difference gel electrophoresis with mass spectroscopic identification of selected protein spots), and ICAT (Isotope Coded Affinity Tag profiling). Samples may be requisitioned by researchers via the Web. The sample descriptions include, experiment type, organism, tissue, etc, which are recommended by the MIAME and MIAPE standards.

After processing of the samples, the mass spectra are analyzed for the identification of peptides and proteins using the commercial programs Sequest or Mascot. In the future, the open source program, X!Tandem, will also be utilized. These programs compare and score the experimental spectra with spectra generated from protein databases allowing identification of peptides in the samples.

The results of the database search are validated using the Trans-Proteomic Pipeline suite of programs developed by the Institute of Systems Biology. This suite includes programs for computing the probability that the peptide assignments are correct, for computing the probability that the protein assignments are correct, and for quantitating the relative abundance of proteins obtained from ICAT experiments. The output of the suite is an XML file which is transformed and entered into YPED.

The validated results can be viewed via the Web interface. The results consist of two parts: a statistical overview summarizing the number of proteins, peptides and quantitation, and a table of results. A protein probability cutoff can be selected to restrict the number of proteins viewed. The result table contains hyperlinks for obtaining more information on the identified proteins and for obtaining peptide and quantitation details. The results can also be downloaded as Excel spreadsheets for local analysis.

Multiple samples can also be compared and viewed via the Web interface. Proteins which are common to several samples are easily visualized as well as proteins which are distinct. An overview summary is also presented.

In the future, we plan to extend YPED to incorporate new experimental techniques as they are implemented in our laboratory. We also are planning to implement an interface with the Yale Microarray Database which would facilitate comparison of gene expression results with protein expression results.