The iProXpress Knowledge System for Proteomic Data Analysis

Zhang-Zhi Hu[1]; Hongzhan Huang[1]; Peter McGarvey[1]; An Chi[2]; Julio Valencia[3]; Cathy H. Wu[1]

[1]Georgetown University Medical Center, Washington, DC, USA; [2]University of Virginia, Charlottesville, VA 22904, USA; [3]National Cancer Institute, Bethesda, MD 20892

Large-scale proteomic profiling of biological samples such as cells, organelles or biological fluids has led to discovery of numerous key and novel proteins involved in many biological/disease processes including cancers, as well as to the identification of novel disease biomarkers and potential therapeutic targets. Bioinformatics infrastructure and systems are instrumental in analyses of functional involvement of the identified proteins in metabolic and signaling pathways, cell cycles, apoptosis, and other cellular functions and processes. iProXpress (integrated Protein eXpression) is an integrated protein expression analysis system, which is designed to help analyze proteomic and genomic data such as protein/peptide and gene profiles from IP, 2D and MS proteomic and microarray gene expression experiments. The iProXpress knowledge system consists of three major components, the PIR data warehouse with integrated protein information, analytical tools for sequence analysis and functional annotation, and a graphical user interface for categorization and visualization of expression data. The system includes the following functionalities: 1) Gene/Peptide to Protein Mapping. Gene or protein lists are mapped to corresponding entries in UniProtKB of all known proteins based on gene/protein IDs, names or sequences. 2) Protein information matrix. Protein family, domain, and functional site features for each protein are identified by BLAST, HMM, signal peptide, transmembrane helix predictions and other automated searches. For direct comparison of expressed genes/proteins, a comprehensive protein information matrix is generated, summarizing salient features retrieved from the underlying PIR data warehouse or inferred based on sequence similarity. 3) Protein Data Analysis for Pathway and Network Discovery. Users can conduct iterative categorization and sorting of proteins in the information matrix and correlate expression and interaction patterns to salient protein properties for pathway and network discovery. Proteins are clustered based on functions, pathways, and/or other attributes in the information matrix to identify hidden relationships not apparent in the data on expression patterns and interacting proteins, and to recognize candidate proteins of unknown identity that warrant further investigation. This global bioinformatics analysis provides a composite view of functional changes to help identify critical nodes and hidden relationships in the biological pathways and networks. The iterative categorization steps in the process are currently conducted manually; however, many of them can be automated and rules developed to flag significant clusters. The system has been applied to several studies including the expression profile analysis of hormone-induced changes in endocrine tumor cells, and is currently being adopted for analyses of pathogen/host genomic and proteomic data produced from the NIAID Biodefense Proteomic Program. Here we presented a case study where organellar proteomes of various stages of melanosomes from human melanoma cell lines were analyzed using the iProXpress knowledge systems, which included: 1) Mapping to known mouse coat color genes led to identification of 17 essential human melanosome proteins; 2) Identification of possible stage-specific melanosome proteins for validation; 3) Comparison of melanosome proteome with those of several other organelles permitted a proposed list of proteins characteristic of melanosome. This study has greatly facilitated a better understanding of melanin synthesis and melanosome biogenesis