Bioinformatics for Proteomics Plays a Crucial Role in Human Proteome Project

Ting-Yi Sung

Project Coordinator

Proteins are final products of genes and perform functions in living organisms. In the area of biomedical research, proteins are prominent drug targets. Thus, in the post-genomics era, proteomics has been gaining ever-increasing attention. In "Mass spectrometry-based proteomics," Aebersold and Mann (*Nature* 2003) proclaimed mass spectrometry as an indispensable tool for proteomics. Later in 2005, Ong and Mann published "Mass spectrometry-based proteomics turns quantitative" in *Nature Chemical Biology*. Currently, liquid chromatography (LC) coupled with mass spectrometry (MS) technology has been widely used in proteomics research, especially in biomarker discovery and cancer research. In LC-MS experiments, proteins are digested into peptides, separated by LC, and then analyzed by MS. Proteins differentially expressed in different bio-samples can be determined by analyzing the acquired large-scale mass spectra. However, analysis of mass spectral data is challenging for a variety of reasons, including different fragmentation modes in mass spectrometry, coeluting, quality of samples and sample complexity, and noise in the mass spectra. In 2003, in collaboration with Dr. Yu-Ju Chen of Academia Sinica's Institute of Chemistry, we began to develop computation methods and automated tools for mass spectrometry-based quantitative proteomics. We have published three quantitation tools available for download: Multi-Q, MaXIC-Q, and IDEAL-Q.

At the advent of the proteomics age, the Human Proteome Organization initiated the Chromosome-centric Human Proteome Project (C-HPP), similar to Human Genome Project, in which an international collaborative effort has been organized, with 25 working groups — one per chromosome. Taiwan's team, led by Dr. Yu-Ju Chen, is responsible for chromosome 4. The project is aimed at discovering and characterizing all human proteins encoded from genes for the purpose of filling the gap between genomics and proteomics. Since 2013, the main theme of this project has been to experimentally discover missing proteins, which have not been detected by MS or antibody experiments, i.e., those that lack of experiment evidence at the protein level. These proteins are missing for various reasons, such as low abundance, expression in transient states or rare samples, and unfavorable cleavage sites for MS experiments. In order to detect missing proteins, Dr. Chen conducted LC-MS/MS experiments on 11 non-small cell lung cancer cell lines. By using existing database sequence search engines (e.g., Mascot), proteins can be confidently reported from acquired LC-MS/MS spectra. However, because most search engines do not report false discovery rate (FDR) at the protein level, those confidently reported proteins may not be really identified. Rigorous analysis of search engine results is essential to claim missing proteins being detected. Thus, our lab, which has bioinformatics