Bioinformatics for Proteomics Plays a Crucial Role in Human Proteome Project

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
expertise, was responsible for this critical task. Based on the reported proteins from the search engine, we first used PeptideShaker, an existing tool, to exclude proteins/peptides with FDR higher than 1%, which is a required criterion in C-HPP. Second, since proteins reported by search engines are based on inference from peptides identified from MS/MS spectra, ambiguity in protein inference is inevitable. To avoid ambiguity in protein inference, we further excluded identified proteins without any identified peptide that belongs to only a single protein in the entire human protein database. Up to this stage, we have confidently identified 7702 proteins, with 66% being membrane proteins. Third, these proteins were compared with the missing protein list provided by the project consortium to find if any missing proteins may have been detected. To confirm whether missing proteins were indeed detected, we performed two critical investigations for further filtering: (1) checking whether the identified peptides used to infer these proteins had been previously detected and deposited in PeptideAtlas, a huge MS/MS spectra repository and recommended for use in this project; (2) checking whether these identified peptides could be derived from a single amino acid variation of a peptide in another protein. By performing the above confirmations, we successfully detected 178 missing proteins, including 74 membrane proteins. Dr. Chen's lab used the multiple reaction monitoring MS technique on eight synthetic peptides to confirm our rigorous workflow to determine detected missing proteins. This work was published in the 2015 special issue of C-HPP in the Journal of Proteome Research (JPR).

Furthermore, from the perspective of bioinformatics, we attempted to address one of the priority questions for the C-HPP: which of the missing proteins are unlikely to be detectable by even an “ideal” shotgun MS/MS analysis of the human proteome. Thus we performed three in silico digestions by commonly used trypsin, Lys-C and both on the human proteins to generate all in silico fully digested peptides. With these presumed peptides, we found 145 proteins, including 77 missing proteins, containing no unique peptide, consisting of only shared peptides. These missing proteins were hard to confirm in shotgun proteomics experiments. In addition, missing proteins with high sequence similarity, even up to 100% similarity, are also hard to identify. We also noted that among all missing proteins with evidence at the transcript level, G protein-coupled receptors and olfactory receptors, based on InterPro classification, were the largest families of proteins that exhibited more frequent variants, and thus are hard to identify. In order to identify the above-mentioned types of missing proteins, new MS experiment designs and improved identification methods are needed. This work was published in JPR in 2015.

Proteomics research is gaining ever-increasing attention. Computing and analyzing big data acquired from mass spectrometry requires more IT expertise. IT experts are welcome to join the fascinating area of bioinformatics for proteomics.