Organizing, Validating, and Reusing Massive Amount of Proteomics Data

by

Long WU

Abstract
Tandem mass spectrometry plays a vital role in proteomics research. With the rapid advances in high-resolution, high-throughput mass spectrometers (MS), the volume of MS data from proteomics experiments has increased exponentially in the past few years. Nowadays, proteomics data-sets are organized in the public space in three forms: data repositories, spectral archives or spectral libraries, but there is a pressing need for methods to better organize such data for effective reuse. In this thesis, we focus on building a new technological platform for organizing, validating and reusing massive amounts of proteomics data. The platform is called a searchable spectral archive, which organizes tandem mass spectra of peptides by their similarities to each other. Data can be added to the archive incrementally with linear time, and the nearest neighbors (most similar spectra) can be retrieved quickly for any query spectrum, with the help of a powerful indexing technique and parallel computing. This allows the identification of any newly acquired spectrum by referencing all existing spectra, a new paradigm in proteomic data analysis. To further verify the reliability of retrieved nearest neighbors, several validation tools are developed. First, a context-free validator is trained with a machine learning framework to evaluate peptide-spectrum matches. It does not rely on traditional sequence search engine scores or any data-set-specific information, and achieves similar performance compared to existing tools. Second, a Benjamini-Hochberg procedure-based method for error control was developed and applied to spectral archive search results, to demonstrate the validity of this new approach in comparison to traditional search engines. Third, a web interface was developed to visualize and navigate a spectral archive as a force-directed graph in real time, giving the user a direct and intuitive view of the archive search process. The whole platform is lightweight and can be deployed even in small research groups with limited computing resources.

Date: 22 May 2020 (Friday)
Time: 3:00pm
Venue: Online via Zoom

Examination Committee:
Prof. Sam Garg (Chair)
Prof. Henry Lam (Supervisor)
Prof. Weichuan Yu
Prof. Jiguang Wang
Prof. Terence Wong
Prof. Terence Poon (University of Macau)

All are welcome!

Room 4572, Academic Building, Clear Water Bay, Kowloon. http://bien.ust.hk Tel: (852) 23588483