Applications of Probabilistic Model on Peptide MS/MS Spectra Identification and Protein Quantification

by

MA Chun Wai

Abstract

Shotgun proteomics, a bottom-up approach for complex protein mixture analysis using mass spectrometry coupled with liquid chromatography (LC-MS), can be viewed as a quasi-random model for its sampling of peptides for identification. In this process, a set of peptides eluting from the chromatographic column are selected for fragmentation in tandem mass spectrometry (MS/MS), and the measurement of the fragments enable the deduction of the peptide sequence. In this thesis, making use of probabilistic models, we attempt to improve various applications in mass spectrometry-based proteomics ranging from identification to quantification.

Post-translational modification (PTM) is ubiquitous in cellular processes such as cell differentiation, signaling pathways, regulation of enzymatic activities, and protein degradation. Reports estimated that up to 50% of proteins in the human body have undergone PTM. More than 200 types of PTM are found so far. These PTMs in protein largely increase the search space for peptide/protein identification in proteomics, and decrease the identification performance in term of accuracy and computational time. One strategy for post-translational modified protein identification is open modification search, which is based on a mass shifting strategy. In open modification search, the precursor mass difference between the candidate and query spectra is assumed to be the mass shift induced by the PTM. This mass shift is expected to be present in some, but not all, of the fragments produced in tandem mass spectrometry of the peptide. To allow for these mass-shifted fragments, the search and query spectra is matched several times with m/z ratio shifted values in different charge states, which greatly increases the chance of false matches. To reduce false positives and thus increase identification sensitivity, we introduced a tier-wise scoring strategy that filters peak matches by statistical significance calculated based on a probabilistic argument. The sensitivity of the identification is improved for both modified and unmodified peptides, compared with the state-of-the-art method previously published.

The peptides ionized by electrospray ionization (ESI) for mass spectrometric analysis is typically multiply charged. As the mass spectrometer can only measure the mass-to-charge (m/z) ratio of the analyte, the charge state have to be determined for mass calculation. Isotopic spacing and distribution is the dominating method for charge state determination. However, for complex mixtures, overlapping isotopic patterns are frequently observed, and can confound the charge determination process. For low-abundance peptides, sometimes the isotopic pattern was difficult to discern from the background noise, and algorithms that depend on the isotopic pattern will fail. We developed a novel method called ZDefine, for peptide charge determination, which makes use of the periodic pattern in mass
distribution found in short peptides. This method implements a Bayesian classifier based on the probabilistic distribution of mass defect values of all peptides. It only requires accurate mass measurement of one precursor peak, is much faster computationally than traditional methods, and offers comparable performance in charge determination to several alternative methods.

The same periodic pattern of mass distribution of peptides is also used to design optimal parameters for peptide spectral matching. We optimized the bin width and the intensity filter used in the calculation of the spectral dot product – a spectral similarity measure commonly used. The performance of our optimized scoring method compares favorably to the original algorithm.

Besides applications in qualitative proteomics, we also applied a probabilistic model to study the behavior of spectral counts as a protein quantification metric. Various indexes were proposed based on the spectral count (the number of spectra identified to peptides belonging to a protein) for absolute quantification purpose, but their use have been empirically driven and lack a theoretical basis. We propose a simple probabilistic model to describe the relation between those indexes and the protein abundance. Based on the theoretical result derived, we simplified a commonly used index called APEX, and still maintained similar performance. We also extended this index to the aggregate analysis of multiple experimental datasets, resulting in an improvement in the accuracy of quantification.

Date: 23 Jan 2014 (Thursday)
Time: 3:30pm
Venue: Room 5583 (Lift 29-30)

All are welcome!

Examination Committee:
Prof. Michael Zhang (Chair)
Prof. Henry Lam (Supervisor)
Prof. Gary Chan
Prof. Simon Chan
Prof. Albert Chung
Prof. Edmund Lam (HKU)