Data Mining in Protein Identification by Tandem Mass Spectrometry

Haipeng Wang
Institute of Computing Technology & Graduate University of Chinese Academy of Sciences, China

INTRODUCTION

Protein identification (sequencing) by tandem mass spectrometry is a fundamental technique for proteomics which studies structures and functions of proteins in large scale and acts as a complement to genomics. Analysis and interpretation of vast amounts of spectral data generated in proteomics experiments present unprecedented challenges and opportunities for data mining in areas such as data preprocessing, peptide-spectrum matching, results validation, peptide fragmentation pattern discovery and modeling, and post-translational modification (PTM) analysis. This article introduces the basic concepts and terms of protein identification and briefly reviews the state-of-the-art relevant data mining applications. It also outlines challenges and future potential hot spots in this field.

BACKGROUND

Amino Acids, Peptides, and Proteins

An amino acid is composed of an amino group (NH_2), a carboxylic acid group (COOH), and a differentiating side chain (*R*). Each amino acid is represented by a letter from the English alphabet except B, J, O, U, X, and Z. A peptide or a protein is a chain that consists of amino acids linked together by peptide bonds. In this context, we refer to the products of enzymatic digestion of proteins as peptides.

Tandem Mass Spectrometry

Mass spectrometry (MS) is an analytical technique used to separate molecular ions and measure their mass-to-charge ratios (*m*/*z*). Tandem mass spectrometry (MS/MS), which can additionally fragment ionized molecules into pieces in a collision cell and measure the *m*/*z* values and ion current intensities of the pieces, is becoming increasingly indispensable for identifying complex protein mixtures in high-throughput proteomics.

MS-Based Protein Identification

According to the MS or MS/MS instrument adopted, there are two strategies for protein identification: peptide mass fingerprinting (PMF) and peptide fragment fingerprinting (PFF). The PFF approach is the focus of this article.

In a typical high-throughput PFF experiment, protein mixtures are digested with a site-specific protease (often trypsin) into complex peptide mixtures (e.g., the protein [AFCEFIVKLEDSE] digested into peptides [AFCEFIVK] and [LEDSE]). The resulted peptides are separated typically by liquid chromatography (LC) and sequentially fed into a MS/MS instrument. In this instrument the separated peptides are ionized with one or more units of charges, selected according to their *m*/*z* values, and broken into pieces by low-energy collision-induced dissociation (CID) resulting in various types of fragment ions (Figure 1). The fragment ions are measured to obtain a bundle of spectral peaks each comprising an *m*/*z* and an intensity values. Peaks plus the *m*/*z* value and charge state of a peptide ion constitutes a peptide MS/MS spectrum (Figure 2). It is not necessary that every possible fragment ion appear in the spectrum.

Figure 1. The nomenclature for various types of fragment ions

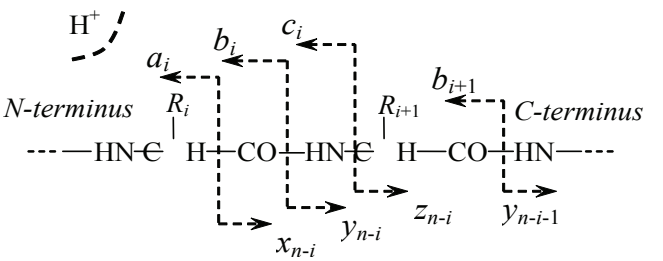
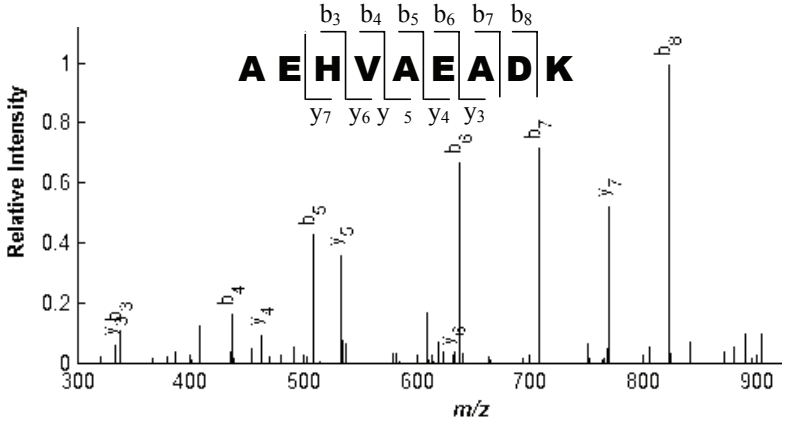


Figure 2. An example of a MS/MS spectrum and its annotation



Peptide identification is the core of protein sequencing in that a protein sequence is identified from its constituent peptides with the help of protein databases. There are three different approaches for PFF peptide identification: database searching, *de novo* sequencing, and sequence tagging.

Database searching is the most widely used method in high-throughput proteomics. It correlates an experimental MS/MS spectrum with theoretical ones generated from a list of peptides digested *in silicon* from proteins in a database. As annotated MS/MS spectra increase, a spectrum library searching method which compares an experimental spectrum to previously identified ones in the reference library has recently proposed (Craig et al., 2006; Frewen et al., 2006). *De novo* sequencing tries to infer the complete peptide sequence directly from the *m*/*z* differences between peaks in a MS/MS spectrum without any help of databases. Sequence tagging yields one or more partial sequences by *de novo* sequencing, then finds homologous sequences in the protein database and finally scores the homologous candidate sequences to obtain the true peptide.

Robust data preprocessing, scoring scheme, validation algorithm, and fragmentation model are the keys to all of the above three approaches.

MAIN FOCUS

Preprocessing of MS/MS Spectra

The purpose of data preprocessing is to retain the potentially contributing data while removing the noisy or misleading ones as many as possible to improve the reliability and efficiency of protein identification. In general, only a relatively small fraction (say, <30%) of a large MS/MS spectra set is identified confidently due to some reasons. First, there exist some chemical or electronical noises in a spectrum resulting in random matches to peptides. Second, the peptide ion charge state of a spectrum is usually ambiguous due to the inability of instruments. Third, there are considerable duplicate spectra most likely representing a unique peptide. Four, many poor-quality spectra often have no results which can be filtered in advance using classification or regression methods. Clustering algorithms (Tabb et al., 2003; Beer et al., 2004), linear discriminant analysis (LDA) (Nesvizhskii et al., 2006), quadratic discriminant function (Xu et al., 2005), SVM (Bern et al., 2004; Klammer et al., 2005), linear regression (Bern et al., 2004), Bayesian rules (Colinge et al., 2003a; Flikka et al., 2006), decision tree and random forest (Salmi et al., 2006) have been widely used to address the above problems.

Scoring Scheme for Peptide Candidates

The goal of scoring is to find the true peptide-spectrum match (PSM). In principle there are two implementation frameworks for this goal: descriptive or probabilistic.

In descriptive frameworks, an experimental and a theoretical MS/MS spectra are represented as vectors $\mathbf{S} = (s_1,s_2,...,s_n)$ and $\mathbf{T} = (t_1,t_2,...,t_n)$, respectively, where n denotes the number of predicted fragments, s_i and t_i are binary values or the observed and predicted intensity