Mass Informatics in Differential Proteomics

Xiang Zhang
University of Louisville, USA

Seza Orcun
Purdue University, USA

Mourad Ouzzani
Purdue University, USA

Cheolhwan Oh
Purdue University, USA

INTRODUCTION

Systems biology aims to understand biological systems on a comprehensive scale, such that the components that make up the whole are connected to one another and work in harmony. As a major component of systems biology, differential proteomics studies the differences between distinct but related proteomes such as normal versus diseased cells and diseased versus treated cells. High throughput mass spectrometry (MS) based analytical platforms are widely used in differential proteomics (Domon, 2006; Fenselau, 2007). As a common practice, the proteome is usually digested into peptides first. The peptide mixture is then separated using multidimensional liquid chromatography (MDLC) and is finally subjected to MS for further analysis. Thousands of mass spectra are generated in a single experiment. Discovering the significantly changed proteins from millions of peaks involves mass informatics. This paper introduces data mining steps used in mass informatics, and concludes with a descriptive examination of concepts, trends and challenges in this rapidly expanding field.

BACKGROUND

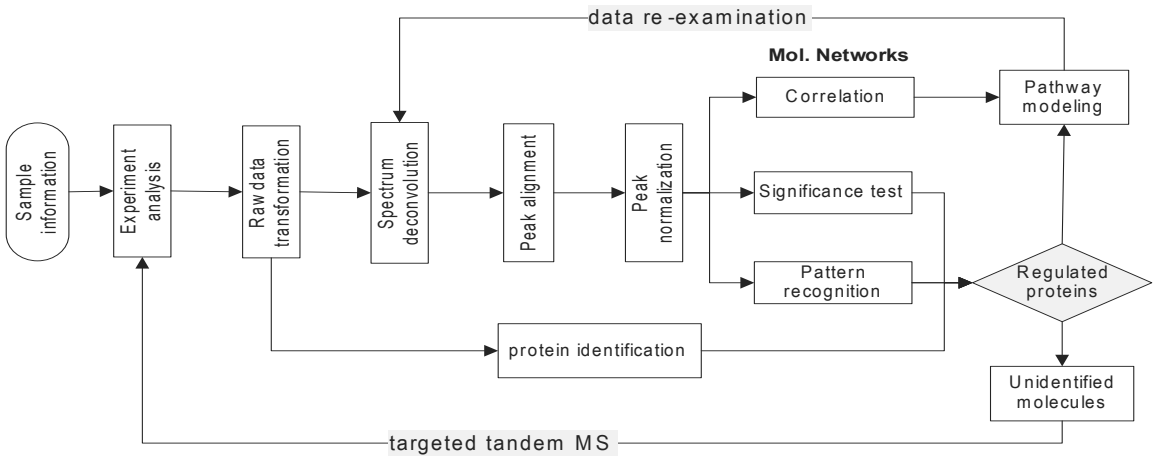
Proteomics was initially envisioned as a technique to globally and simultaneously characterize all components in a proteome. In recent years, a rapidly emerging set of key technologies is making it possible to identify large numbers of proteins in a mixture or complex, to map their interactions in a cellular context, and to analyze their biological activities. Several MS based technologies have emerged for identifying large numbers of proteins expressed in cells and globally detecting the differences in levels of proteins in different cell states (Asara, 2006).

Due to the complexity of the proteome, a major effort in proteomics research is devoted to fractionation of proteins and peptides prior to MS. One way to fractionate peptide mixtures to the extent required for MS analysis of the component peptides is to couple multiple chromatography columns in tandem (Hattan, 2005; Qui, 2007). The fractionated peptide mixture is then subjected to MS for further separation and mass analysis. Thousands of mass spectra will be generated in a single differential proteomics experiment. Mass informatics is involved in identifying the significantly changed proteins from millions of peptide peaks.

DATA MINING FRAMEWORK

A variety of mass spectrometers are commercially available. Each of these mass spectrometers stores raw mass spectra in a proprietary format. The raw spectra have to be transformed into common data format first. As in any study of biological phenomena, it is crucial that only relevant observations are identified and related to each other. The interpretation and comprehension of the collection of mass spectra presents major challenges and involve several data mining steps. The aim of mass informatics is to reduce data dimensionality and to extract relevant knowledge from thousands of mass spectra (Arneberg, 2007). Figure 1 shows an overall framework for mass informatics in differential proteomics. Most of the components of this framework

Figure 1. Information flow chart in differential proteomics



will be discussed in this paper with the exception of the pathway modeling.

Spectra Deconvolution

The purpose of spectra deconvolution is to differentiate signals arising from the real analyte as opposed to signals arising from contaminants or instrumental noise, and to reduce data dimensionality which will benefit down stream statistical analysis. Therefore, spectra deconvolution extracts peak information from thousands of raw mass spectra. The peak information is reported in a simple peak table. As an example, GISTool (Zhang, 2005a) is a software package using chemical noise filtering, charge state fitting, and de-isotoping for the analysis of complex peptide samples. Overlapping peptide signals in mass spectra are deconvoluted by correlating the observed spectrum with modeled peptide isotopic peak profiles. Isotopic peak profiles for peptides were generated *in silico* from a protein database producing reference model distributions. Several other spectra deconvolution packages such as RelEx (MacCoss, 2003), MSQuant (http://msquant.sourceforge.net/), and ASAPRatio (Li, 2003) have been developed to find quantitative information about proteins and peptides.

Protein Identification

Two methods are currently used for protein identification: database searching and *de novo* sequencing. Database searching correlates the spectra with protein sequences. The database-searching algorithm starts with spectrum reduction to remove chemical noise. A list of peptides is generated *in silico* from the protein database using enzyme specificity. After applying potential chemical modifications, *in silico* peptides that have similar molecular weight to the precursor ions of tandem mass spectrum (MS/MS) are selected as candidate peptides. A theoretical spectrum is then created for each candidate peptide and these theoretical spectra are compared with the experimental spectrum. A final ranking list is generated using different scoring functions. Disadvantages of the database searching strategy are very well understood: the protein of interest might not be present in the sequence database, prediction errors are present in gene-finding programs, the protein database may not be available in some cases, genes might undergo alternative splicing resulting in novel proteins, and amino acids may mutate and undergo unknown modifications. SEQUEST (Eng, 1994) and MASCOT (Perkins, 1999) are two database searching software packages used most frequently.

de novo sequencing derives the peptide sequence directly from the tandem mass spectrum (Kanazawa, 2007). Lutefisk (Taylor, 1997) is a popular *de novo*