Chapter V

Proteomics with Mass Spectrometry

Simon Lin, Northwestern University, USA
Salvatore Mungal, Duke University Medical Center, USA
Richard Haney, Duke University Medical Center, USA
Edward F. Patz, Jr., Duke University Medical Center, USA
Patrick McConnell, Duke University Medical Center, USA

Abstract

This chapter provides a rudimentary review of the field of proteomics as it applies to mass spectrometry, data handling, and analysis. It points out the potential significance of the field suggesting that the study of nucleic acids has its limitations and that the progressive field of proteomics with spectrometry in tandem with transcription studies could potentially elucidate the link between RNA transcription and concomitant protein expression. Furthermore, we describe the fundamentals of proteomics with mass spectrometry and expound the methodology necessary to manage the vast amounts of data generated in order to facilitate statistical analysis. We explore the underlying technologies with the intention to demystify the complexities of the nascent field and to fuel interest by readers in the near future.

Copyright © 2006, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.
Introduction

The science of proteomics seeks to identify and characterize protein expression in biological systems. It leverages technologies that have been around for decades, such as mass spectrometry, liquid chromatography, and electrophoresis. Though the basic science is mature, proteomics technology has made much progress over the last decade (Aebersold & Mann, 2003). Proteomics has been gaining momentum as the limitations of studying DNA and RNA alone have been documented (Greenbaum, Colangelo, Williams, & Gerstein, 2003). Gene sequences themselves yield little information about how much of its transcribed protein will be expressed and in what cellular states. Furthermore, the nature and prevalence of alternative splicing has shown that studying gene expression at the protein level can yield complementary knowledge at the nucleic acid level.

Proteomics is a vast and complex field consisting of a variety of platforms. In this chapter, we focus on mass spectrometry technology and data because of its wide use in the field for protein identification and profiling experiments. The goal of identification is to find out the amino acid sequence of extracted proteins; whereas the goal of profiling is to quantify the expression level of proteins.

Proteomics data is measured in mass-to-charge ratios, termed m/z values. Peaks can be identified when we plot measured intensities against m/z values. The location of a peak corresponds to the chemical composition of the protein and thus can be used for protein identification, and the amplitude of a peak carries information on the protein abundance, which is the basis of profiling. Such data provides a unique challenge both to encode and to analyze. Firstly, even modest-sized experiments provide gigabytes worth of data. Encoding and exchanging data of this size is quite a technical challenge without mention of the problem of analyzing such a large dataset. Secondly, the data itself has intrinsic properties that make it particularly challenging to analyze and interpret. For example, one common task in analyzing spectra data is to locate data peaks. However, what characteristics of the peaks are most important: height, area under the peak, or some ratio thereof? Furthermore, noise in a single spectrum and aggregating multiple spectra provide unique statistical challenges.

There has been some work towards standardization of proteomics data. The Proteomics Standards Initiative (PSI) has been created under the auspices of the Human Proteome Organization (HUPO). PSI has working groups to develop data standards for protein-protein interaction and mass spectrometry (Orchard et al., 2004). Though good intentioned, the PSI mass spectrometry data standards are still in an abstract form. The Minimum Information About Proteomics Experiments is a document describing the types of data items needed to fully described proteomics experiments, but no data standard exists. The PSI Object Model (PSI-OM) and PSI Markup Language (PSI-ML) have defined a set of concepts and the relationships of those concepts, but not concrete data formats. Finally, the PSI Ontology (PSI-Ont) is being developed as an extension to the Microarray and Gene Expression Data Ontology (MGED Ontology) with the ‘PSI’ namespace, but is not yet complete. Proteomics data standardization is far from complete, and there is much work to be done to define data formats and ontologies.
A Transfer Learning Approach and Selective Integration of Multiple Types of Assays for Biological Network Inference
Tsuyoshi Kato, Kinya Okada, Hisashi Kashima and Masashi Sugiyama (2012). Computational Knowledge Discovery for Bioinformatics Research (pp. 188-202). www.igi-global.com/chapter/transfer-learning-approach-selective-integration/66711?camid=4v1a