available at www.sciencedirect.com
ScienceDirect
www.elsevier.com/locate/jprot

Review Article

A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics

Alexey I. Nesvizhskii[*]

Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA
Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

ARTICLE INFO

Article history:
Received 2 July 2010
Accepted 25 August 2010

Keywords:
Proteomics
Bioinformatics
Mass spectrometry
Peptide identification
Protein inference
Statistical models
False discovery rates

ABSTRACT

This manuscript provides a comprehensive review of the peptide and protein identification process using tandem mass spectrometry (MS/MS) data generated in shotgun proteomic experiments. The commonly used methods for assigning peptide sequences to MS/MS spectra are critically discussed and compared, from basic strategies to advanced multi-stage approaches. A particular attention is paid to the problem of false-positive identifications. Existing statistical approaches for assessing the significance of peptide to spectrum matches are surveyed, ranging from single-spectrum approaches such as expectation values to global error rate estimation procedures such as false discovery rates and posterior probabilities. The importance of using auxiliary discriminant information (mass accuracy, peptide separation coordinates, digestion properties, and etc.) is discussed, and advanced computational approaches for joint modeling of multiple sources of information are presented. This review also includes a detailed analysis of the issues affecting the interpretation of data at the protein level, including the amplification of error rates when going from peptide to protein level, and the ambiguities in inferring the identifies of sample proteins in the presence of shared peptides. Commonly used methods for computing protein-level confidence scores are discussed in detail. The review concludes with a discussion of several outstanding computational issues.

Contents

[*] Department of Pathology, University of Michigan, 4237 Medical Science I, Ann Arbor, MI 48109, USA. Tel.: +1 734 764 3516.
E-mail address: nesvi@umich.edu.

1874-3919/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jprot.2010.08.009

1. Introduction

More than a decade after the beginning of rapid expansion in proteomic technologies and applications, proteomics remains a fast growing field. Generally defined, proteomics is an integrative study of proteins, and their biological functions and processes. An overarching goal of proteomics is to achieve complete and quantitative analysis of the proteome of a species, including the sub proteomes of various cells or tissue types in the case of multi-cellular organisms. This also includes the reconstruction of protein interaction networks and protein complexes and their dynamic changes, cellular localization analysis, delineation of kinase — substrate relationships, and many other biological applications [1].

While there exist a number of alternative proteomics strategies (e.g. protein array based methods [2]), mass spectrometry (MS)-based strategies have become the method of choice for both identification and quantification of proteins in most studies. In this regard, the last several years have been particularly exciting for the field. With the advent of new MS instrumentation, alternative fragmentation mechanisms, and advanced data acquisition strategies, the throughput and the depth of the proteomic analysis have improved by an order of magnitude compared to earlier applications. This has enabled many powerful proteomic applications, including global analysis of post-translational modifications [3], large-scale reconstruction of protein interaction networks [4], and deep quantitative proteome profiling of model organisms [5]. Significant efforts are being made to introduce proteomic technologies in clinical and translational research [6]. MS-based proteomics is now increasingly applied in the context of systems biology studies where it is used in parallel with other technologies such as gene expression analysis and metabolomics. MS-based findings are being increasingly annotated in knowledge repositories such as UniProt. MS-specific repositories are also quickly growing, with new resources for various domain applications such as phosphoproteomics being constantly created [7].

Proteomics, like all high-throughput technologies, is extremely dependent on the ability to quickly and reliably analyze large amounts of experimental data. In the absence of robust statistical and computational methods, proteomic datasets contain significant numbers of false positives [8–13], and statements referring to computational analysis of MS/MS data as e.g. "the Achilles heels of proteomics" are common in the literature [14]. The high rate of false positives in early proteomic publications was so alarming to the scientific community that it lead to the establishment of specific data analysis guidelines by the Editorial Boards of the leading proteomic journals [15]. In recent years, there has been a substantial progress in addressing the most immediate proteomic data analysis needs. Several commercial and open source data analysis pipelines became available and allowed faster and more transparent analysis of proteomic data than previously possible. At the same time, the dramatic change in the size, diversity, and context in which proteomic datasets of today are generated creates a need for a survey and detail discussion of the existing and emerging computational strategies. In this manuscript, we review the