Preface

Currently, a number of books cover the experimental side of proteomics and only briefly describe the theory and practice of data analysis. Additionally, the generation of mass spectrometry (MS) data already has become a high-throughput technique, which calls for efficient high-quality algorithms for data analysis. The intention with this volume is to support researchers in deciding which programs to use in various tasks related to analysis of MS data in proteomics. *Mass Spectrometry Data Analysis in Proteomics* gives a precise description of the theoretical background of each topic followed by accurate descriptions of programs and the parameters best suited for different cases. The focus has been on covering the most common steps in analyzing MS data.

First, different types of MS data and the data format are introduced, followed by a description of the best way to convert raw data into peak list, which can be input to various search engines. For searching databases with MS data, it is important to use databases that are as complete as possible. Sequences can be gathered from several resources, i.e., predicted genes from genomic data, expressed sequence tags (ESTs), and protein sequences. When searching predicted genes from genomic data it is important to consider the accuracy of the predicted exons, for ESTs possible frame-shift is a central issue, and for protein sequences potential signal peptides are worth considering. *Mass Spectrometry Data Analysis in Proteomics* not only gives a report of the available sequence databases, but also covers how to assemble ESTs into nonredundant databases and to further process the sequences into a format suitable for searching with MS data.

In the proteomics field there is a figure of speech, "100% sequence coverage is not enough." The proteomics field has to deal with more than 200 possible modifications of amino acids. When looking for modifications, it is important to have high mass accuracy and it is also an advantage to use other experimental techniques or consider information stated in the literature, which can help to limit the number of possible modifications.

Quantification is an important issue in proteome projects because the dynamic range of protein concentrations is thought to be around 105–106 for eukaryotic cells. Relative quantification of proteins has been available by densitometry of protein spots in two-dimensional electrophoresis (2-DE) gels for many years. Recently, relative quantification of stable isotope-labeled peptides analyzed by liquid chromatography (LC)-MS/MS has drawn great attention. This interest is mainly due to easy automation of the LC-MS/MS runs, whereas

the preparations of 2-DE gels are quite tedious. However, the data analysis is more complex especially if the isotopic peaks from the nonstable isotope-labeled and the stable isotope-labeled version of the peptides are overlapping, as will be discussed further.

Mass Spectrometry Data Analysis in Proteomics mainly describes publicly available programs. However, for computations where no publicly accessible programs are available, commercial programs have been described. The choice of programs in proteomics is, unfortunately, often limited by the data format. The Proteomics Standards Initiative has been established to define community standards for data representation in proteomics. The XML format has recently been suggested as a good tool for interchanging data between applications and is used already by several public and commercial applications.

There are some similarities between the proteomics field of today and the situation in the structural proteomics community in the late 1960s with the lack of public databases containing results and detailed experimental procedures. In the last chapter, strategies for creating such databases are discussed.

Rune Matthiesen