Bioinformatics for mass spectrometry. Novel statistical algorithms
Dijkstra, Martijn

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2008

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Biomarkers are molecules that correlate to pathological states and physiological processes. Therefore, they have many applications in life sciences research areas, as is illustrated in Figure 1.1 for medicine. Mass spectrometry (MS) is a technology to separate complex mixtures of molecules, which can be useful for the discovery of new biomarkers.

The main contribution of this thesis is the development of novel statistical methods for the optimal analysis of complex MS data. This thesis also develops an approach to analyze MS time series data, that can be helpful in predicting patient recovery after surgery.
1. Introduction

1.1 Mass spectrometry

A mass spectrometer is a tool which can be used to determine the composition of complex mixtures of molecules. The mass spectrometer analyzes the solid or liquid mixture in three steps: ionization, separation and detection. In the first step, the molecules are ionized (i.e. charged) and brought into gas phase. The most commonly used ionization techniques are Electrospray ionization techniques (ESI) for liquid samples, and Matrix Assisted Laser Desorption/Ionization (MALDI) for solid samples (Aebersold and Mann 2003). A detailed overview of these and other ionization techniques is presented in (Vestal 2001). In the second step, the ions are separated in electric and/or magnetic fields according to their mass-to-charge ratios ($m/z$). This step can be repeated with some form of ion fragmentation in between, such as occurs in so-called Tandem-MS methods. A significant match between a measured fragmentation pattern and a known fragmentation pattern of a known molecule, can lead to the identification of analyzed molecules. In the third and final step, the separated ions are detected. An ion-to-electron converter is an example of a detector that converts the number of ions to a (proportional) electric current (Dubois et al. 1999), which in turn can be quantified by an ammeter. The resulting mass spectrum can be visualized by plotting the measured ion intensity ($y$-axis) versus the $m/z$ of the ions ($x$-axis). The location and the size of the peaks in a mass spectrum correspond to the mass and abundance of the detected ions, respectively. Thus, ions with similar $m/z$ give rise to overlapping peaks. This can complicate analysis of the mass spectrum.

The number of overlapping peaks can be reduced by means of an extra biophysical or chemical fractionation step. The main result of this additional step is that different ions with similar $m/z$ end up in separate fractions, which can be analyzed separately with the mass spectrometer. Surface Enhanced Laser Desorption/Ionization (SELDI) is a well-known example of such a separation technique (Hutchens and Yip 1993). SELDI extends MALDI chips with different (bio-)chemical coatings which bind different sets of molecules (Merchant and Weinberger 2000). Each set of mole-
1.2 Bioinformatics challenges

Molecules can be analyzed in a separate, and therefore less complex, mass spectrum. Another example of a separation technique is Liquid Chromatography (LC), where molecules with different characteristics need a different amount of time to pass through a chromatographic column. Examples of LC methods are strong cation exchange and reversed phase chromatography (Khalsa-Moyers and McDonald 2006). Other separation techniques are, isoelectric focusing, 1-D and 2-D gel electrophoresis, and capillary electrophoresis (Bischoff and Luider 2004). Successful separation reduces the number of overlapping peaks in the mass spectrum, and therefore facilitating its analysis.

Easy to operate, high throughput, high sensitivity, high resolution, high dynamic range and high accuracy are desirable properties of a mass spectrometer. MALDI and SELDI are high throughput methods that are easy to operate. MALDI and SELDI are commonly combined with Time-Of-Flight (TOF) mass analyzers. For TOF analyzers, there is a trade-off between sensitivity and resolution. Many SELDI-TOF mass spectrometers have a high sensitivity and a low resolution. ESI is most commonly combined with other types of mass analyzers, including multipole mass filters. The quadrupole is the most common example of multipole mass filters. Besides producing a normal mass spectrum, such filters can be used to make a narrow selection of ions for each given m/z. In this way, ions of a particular m/z can be selected for further analysis. Mass spectrometers can make use of ion-traps to increase signal-to-noise ratios by trapping and accumulating ions. Fourier transform ion cyclotron resonance detectors (Marshall et al. 1998, Marshall et al. 2007) and the more recently invented orbitrap (Scigelova and Makarov 2006), for example, make use of these ion-traps. These methods are the most difficult to operate, but have the highest dynamic range, resolution and precision.

1.2 Bioinformatics challenges

Nowadays high throughput and high resolution MS produces increasing amounts of data that are way too large for comprehensible manual analysis. This challenges the bioinformatics field to develop novel methods and
1. Introduction

models to efficiently process these data with statistical accuracy and objective criteria. In addition, the field is challenged to develop statistical methods for interpretation and analysis of MS (protein) time-series data.

Phenomena like isotopes, the formation of intermolecular complexes and/or multiple charges, generate complex mass spectra with many more peaks than the underlying number of molecule species in the sample. Models with strong interrelationships between peaks have to be developed to reduce a complex mass spectrum with many peaks to only a few molecule species, thus improving the estimates of the mass and the abundance of the molecules in the sample. Moreover, it will reduce the statistical test multiplicity in the biomarker discovery phase and therefore increase the power, and ultimately the chance on finding real biomarkers.

Different peaks in a given spectrum can take various shapes, e.g., symmetric or skewed, sharp or broad. This requires suitable peak models. Also, one has to take into account the shifts between locations of interrelated peaks, which are observed both within and between spectra. Furthermore, overlap of two or more close peaks can severely bias estimations of peak locations and sizes. Noise, and small ‘shoulder peaks’ on tails of peaks, are additional complicating factors. This is particularly difficult for low resolution spectra where many peaks show strong overlap, as occurs in many SELDI-TOF mass spectra.

This thesis discusses the most important sources of variation that affect a sample’s spectrum, and defines advanced parsimonious models for optimal and reliable predictions of sample content, thus enabling high power in biomarker discovery. Therefore, this thesis improves mass spectrometry data analysis by tackling its most striking bioinformatics challenges.

1.3 Outline of thesis

Chapter 2 reviews the various sources of variation in SELDI-TOF MS. A simple example with two molecular species (myoglobin and matrix molecules) illustrates physical and chemical sources of variation which can lead to complex mass spectra with many interrelated peaks. These sources include the
formation of matrix adducts. Chapter 2 shows that skewed peaks in low-resolution spectra can be explained by and decomposed in different adduct peaks. It also introduces a statistical finite mixture which makes use of normal distributions to model the peaks in the mass spectrum. Mixture models can be used to adequately locate and quantify individual peaks, even in very difficult cases with many overlapping peaks.

Chapter 3 presents a mixture model which uses log-normal distributions to fit skewed peaks. This model is illustrated on a large data set, and evidence is presented that the model is superior to the standard (Ciphergen) approach, in cases with overlapping peaks. Moreover, the new model is shown to be insensitive to disturbances in the data due to noise and shoulder peaks.

Chapter 4 uses the theory from Chapter 2 to link the locations of related peaks and to link the shapes of all the peaks in the mixture models. Linking peaks (i) improves parameter estimates, (ii) reduces the total number of observed peaks in a spectrum to a much smaller number of underlying molecule species, which in turn (iii) reduces the statistical test multiplicity in the biomarker discovery phase and therefore increases the power, and ultimately the chance on finding real biomarkers. In Section 6.2.1 (Peak detection in complex mixtures) it is explained in detail that linking peaks (iv) enables the detection of small peaks hidden in complex mixtures. This is important because each peak is a potential biomarker. Chapter 4 also introduces a new method, called “self-calibration”. Self-calibration can be used to locate multiply charged peaks at the correct position in the spectrum. Moreover, self-calibration reduces the alignment of different spectra to just a proportional scaling of the x-axis.

Chapter 5 develops an approach to analyze MS time series data, that can be useful to predict patient recovery after surgery and to evaluate nutritional, and pharmacological interventions as well. The developed methods are used to analyze the intricate protein kinetics in serum induced by surgery, and divide the detected proteins in four main clusters based on their kinetic behavior. Two protein species with pronounced recovery related temporal behavior are isolated and identified by tandem mass spectrometry. In addition, a prediction is made about the function of other clustered proteins.
Chapter 6 contains the general discussion of this thesis, including a literature discussion. It anticipates that models developed in this thesis can be very useful for the analysis of data acquired with a diverse number of technologies, \textit{i.e.} with different kinds of ionization, separation and detectors, as well as nuclear magnetic resonance data. In addition Chapter 6 discusses numerous applications of the developed models, including the analysis of molecular transformations, and the identification of molecules and of their atomic composition. Furthermore, it points out that the developed models improve biomarker discovery.