Diss. ETH No. 17212 2007

New Strategies in Proteomics Data Analysis

A dissertation submitted to the

ETH Zurich

for the degree of Doctor of Science

presented by

Jonas Tobias Grossmann

Dipl. Natw. ETH Zurich
born April 20, 1978
citizen of Fischbach (LU), Switzerland

accepted on the recommendation of

Prof. Dr. Wilhelm Gruissem, examiner
Prof. Dr. Ruedi Aebersold, co-examiner
Dr. Sacha Baginsky, co-examiner

Abstract

The advent of complete genome sequence information permits a comprehensive analysis of biological systems. In conjunction with high throughput technologies, novel and complex scientific questions can thus be addressed. Mass spectrometry is increasingly the technology of choice for proteome characterization. Thanks to remarkable technological progress, thousands of proteins can now be identified on a daily basis. This leads to a rapid accumulation of enormous amounts of data which cannot be interpreted manually anymore. Instead, custom-tailored statistical methods and computational tools are required.

In this thesis we present methods which are able to extract more information from the data than the currently established methods alone. Our results show that such an extended analysis of proteomics data can be advantageous and novel proteins are identified.

In the first paper we show that while standard methods such as database searches are able to identify many proteins, a significant amount of high quality data is not recognized as such and is not exploited in further analysis. We present a method that successfully detects such data and extracts them for further more in-depth analysis.
The second paper deals with the problem of de novo sequencing of peptide mass spectra, where a sequence is derived exclusively from information that is contained in the spectrum itself. We show that spectrum preprocessing by heuristic filters for example, considerably influences the performance of such approaches.
In the third paper, we present a method for the identification of proteins in organisms for which no genome sequence is available yet. By combining several automated software tools and using cross-species information from error-tolerant sequence alignments, an increase in the identification rate of up to 30% is achieved.

The focus of the fourth paper is on the problem of relative quantitative proteomics without isotopic labeling. The main challenge is to align two analogous mass spectrometry experiments and detect common elements as well as differences. Machine learning methods provide an elegant solution to this problem.
In a final chapter, we show how a combination of these tools is able to extract substantially more information from mass spectrometry data than currently established methods alone. We circumvent drawbacks of protein database searches by the development of custom-tailored computational tools and alternative strategies. Thereby it is made possible to use these data to re-annotate the genome and improve genome annotation.

Zusammenfassung

Dank der Verfügbarkeit kompletter Genomsequenzen verschiedener Modellorganismen ist es möglich, biologische Systeme in einem globalen Ansatz zu betrachten. In Verbindung mit Hochdurchsatztechnologien knnen neuartige und komplexe Fragestellungen beantwortet werden.
Zur Charakterisierung von Proteomen hat sich die Massenspektrometrie als bevorzugte Technologie durchgesetzt. Die rasante Entwicklung dieses Verfahrens ermöglicht es, täglich tausende von Proteinen zu identifizieren. Dies führt zu einer Ansammlung grosser Datenmengen, welche nicht mehr manuell interpretiert werden können. Zur Analyse derselben sind massgeschneiderte statistische Methoden und Software notwendig.

In dieser Doktorarbeit präsentieren wir Methoden, die umfangreichere und bisher unzugängliche Informationen aus den Daten extrahieren können als die derzeit gebräuchlichen Verfahren. Unsere Resultate zeigen, dass solch eine erweiterte Analyse von Proteomikdaten durchaus lohnenswert sein kann und sogar neue Protein entdeckt werden können.

Im ersten Artikel wird gezeigt, dass Standardmethoden wie zum Beispiel eine Datenbanksuche, viele Proteine identifizieren kann. Andererseits beobachtet man aber auch, dass Daten von hoher Qualität zu keiner Identifikation führen und so für die Analyse typischerweise verloren gehen. Wir stellen eine Methode vor welche es erlaubt, solche Daten ausfindig zu machen, zu extrahieren und dann einer vertieften Analyse zu unterziehen.
Im zweiten Artikel wird das Problem der de novo Sequenzierung von Peptidspektren behandelt. Hierbei wird eine Sequenz ausschliesslich aufgrund von Informationen die im Spektrum selber enthalten sind hergeleitet. Wir zeigen dass dabei die Vorprozessierung des Spektrums mit zum Beispiel heuristischen Filter eine grosse Rolle spielt.
Der dritte Artikel behandelt eine Methode zur Proteinidentifikation in Organismen, für die noch keine Genomsequenz verfügbar ist. Durch die Kom-

bination von verschiedenen automatisierten Programmen und Information von sequenzierten Organismen ist es möglich die Identifikationsrate gegenüber einer gewöhnlichen Datenbanksuche um bis zu 30% zu steigern.
Im vierten Artikel wird auf ein Problem der relativen quantitativen Proteomanalytik ohne isotopische Markierung eingegangen. Die Herausforderung hier ist es, zwei analoge Massenspektrometrie-Experimente abzugleichen und sowohl Gemeinsamkeiten als auch Unterschiede zu quantifizieren. Dies kann mit Hilfe von Machine Learning-Methoden elegant gelöst werden.
In einem abschliessenden Kapitel zeigen wir, dass man durch die richtige Kombination dieser Methoden deutlich mehr Information aus denselben Daten gewinnen kann als mit den gängigen Verfahren alleine. Wir umgehen die Nachteile von Proteindatenbanksuchen mit der Entwicklung von massgeschneiderten Methoden und alternativen Strategien. Dadurch wird es ermöglicht mit Massenspektrometriedaten das Genom genauer zu beschreiben und Genomanotierung zu verbessern.