Diss. ETH No. 16844

Algorithms for Peptide Identification by Tandem Mass Spectrometry

A dissertation submitted to

ETH Zurich

for the degree of Doctor of Sciences

presented by

Felix Franz Roos

Dipl. Natw. ETH Zurich

born February 8, 1977

citizen of Zürich and Romoos LU, Switzerland

accepted on the recommendation of

Prof. Dr. Peter Widmayer, examiner
Prof. Dr. Wilhelm Gruissem, co-examiner
Prof. Dr. Joachim Buhmann, co-examiner

2006

Abstract

Bioinformatics is a hybrid science at the interface between biology and computer science. For nearly two centuries, data acquisition in the wet lab and in the field has limited the pace of progress in biology. With the advent of high-throughput technologies such as genome sequencing, data analysis has become an additional bottleneck, and simultaneously a rich source of computational and statistical problems.

In this thesis, we focus on two problems arising in the context of high-throughput protein identification via tandem mass spectrometry. The first problem is how to identify splice sites via protein analysis, and the second how to score the quality of tandem mass spectra. We present comprehensive, tailor-made solutions to these problems, thereby integrating theoretical analysis, algorithm design, machine learning and biological knowledge.

Just a decade ago, protein identification was a fairly painstaking process which took weeks or months per protein. In the meantime, hundreds and even thousands of proteins can be identified in a single day via tandem mass spectrometry and subsequent computational fragment pattern analysis. In a typical experiment, proteins in a complex sample are cut into shorter amino acid chains (peptides). Then, the intact mass and a fragment pattern of the peptide are recorded, yielding a tandem mass spectrum. Such a spectrum is usually sufficient to determine the amino acid sequence of the peptide if the genome of the organism is available. However, in the various production steps from the DNA to the protein, processes take place that increase the diversity of proteins in ways that are often difficult to predict from the genome. This may result in protein variants whose peptide fragment patterns are usually not identified by database searches simply because the counterpart is missing in the predicted pattern database.

In the first part of this thesis, we analyze the problem of how to search spliced variants of proteins. After estimating the size of the search space,

we find that a naive search for splice sites is computationally very expensive. We analyze the problem in the I/O model which pays special attention to the different sizes and speeds of storage levels in current computers. While applying a brute force approach, we exploit the computer architecture such as to maximize the computation speed. We present a lower bound for the problem of searching spectra against a peptide database. Furthermore, we develop and analyze several algorithms that are independent of scoring schemes and that match the lower bound. To apply those algorithms, we choose a hypergeometric model as the spectrum scoring scheme and adapt it slightly for our purposes. We implement the algorithm and the statistical framework in a C++ prototype and show in a timing experiment that with increasing datasets the calculation speed asymptotically approaches the limit given by the CPU and the processor cache. Finally, by applying the software to a biological dataset, we succeed in discovering splice variants on the protein level that could not be identified in a standard protein database search.

In the second part of the thesis, we present an approach to evaluate the quality of tandem mass spectra and demonstrate how it can be used to increase the spectrum identification rate at a moderate cost. In most proteomics experiments, only a fraction of a spectrum dataset can be identified by database searches. Depending on the measurement procedure, spectra may be produced even when the peptide concentration is so low that the resulting signal is not sufficient to identify the peptide. On the other hand, datasets may also contain high quality spectra of unexpected protein variants that are absent from the predicted protein database. We develop a method to detect unidentified high quality spectra by analyzing features of identified and unidentified spectra. Among the unidentified spectra, we search for atypical cases whose properties resemble more those of identified spectra than those of unidentified spectra. We then extract these spectra, yielding a relatively small subset, and submit this to more resource-hungry searches that take into account a greater number of potential protein variants.

Zusammenfassung

Bioinformatik ist eine fächerübergreifende Wissenschaft an der Schnittstelle von Biologie und Informatik. Während beinahe zwei Jahrhunderten wurde der Fortschritt in der Biologie durch die Datengewinnung im Labor und im Feld gebremst. Mit dem Aufkommen von automatisierten Technologien wie der Genom-Sequenzierung ist die Datenanalyse zu einem zusätzlichen Engpass geworden, im selben Zug aber auch zu einer reichhaltigen Quelle von Problemen für die Informatik und Statistik.

In dieser Doktorarbeit konzentrieren wir uns auf zwei Probleme, die im Zusammenhang mit der automatisierten Proteinidentifikation via Tandem-Massenspektrometrie auftreten. Das erste Problem besteht darin, wie Splice Sites auf der Ebene der Proteine identifiziert werden können. Beim zweiten Problem geht es darum, wie die Qualität von Tandem-Massenspektren beurteilt werden kann. Wir präsentieren umfassende, massgeschneiderte Lösungen für diese beiden Probleme und verknüpfen dabei theoretische Analyse, Algorithmen-Design, maschinelles Lernen und biologische Expertise.

Noch vor einem Jahrzehnt war Proteinidentifikation ein langwieriger Prozess, der mehrere Wochen oder Monate Arbeit pro Protein erforderte. Inzwischen können täglich Hunderte oder sogar Tausende von Proteinen mittels automatisierter Tandem-Massenspektrometrie und anschliessender computergestützter Fragmentmusteranalyse identifiziert werden. In einem typischen Experiment werden Proteine in einer komplexen Probe in kürzere Aminosäurenketten (Peptide) zerlegt. Dann wird Peptid um Peptid die intakte Masse und ein Fragmentmuster aufgezeichnet, woraus jeweils ein Tandem-Massenspektrum resultiert. Ein solches Spektrum genügt im Normalfall, um die Aminosäuresequenz des Peptids zu bestimmen, sofern das Genom des Organismus bekannt ist. In den verschiedenen Produktionsschritten von der DNA zum Protein finden allerdings Prozesse statt, welche die Vielfalt der Proteine erhöhen - und zwar in einer Art und Weise, welche aufgrund des Genoms oft nur schwer vorhersagbar ist. Dar-

aus können Proteinvarianten hervorgehen, deren Peptid-Fragmentmuster bei Datenbanksuchen normalerweise nicht identifiziert werden, und zwar schlicht deshalb, weil ihr Gegenstück in der Datenbank mit den vorhergesagten Mustern fehlt.

Im ersten Teil dieser Arbeit analysieren wir das Problem, wie nach gesplicten Proteinvarianten gesucht werden kann. Aufgrund einer ungefähren Schätzung des Suchumfangs stellen wir fest, dass die naive Suche nach Splice Sites ziemlich aufwändig ist. Deshalb analysieren wir das Problem im Input/Output-Modell, welches ein besonderes Augenmerk auf die unterschiedlichen Grössen und Transfergeschwindigkeiten von Speicherebenen in gängigen Computern richtet. Dann wählen wir zwar einen Brute-Force-Ansatz, nützen aber gleichzeitig die Computerarchitektur aus, um die Rechengeschwindigkeit zu maximieren. Ausserdem präsentieren wir eine untere Schranke für die Suche von Spektren in Peptid-Datenbanken. Wir entwickeln und analysieren mehrere Algorithmen, die unabhängig von Spektren-Korrelationsmethoden funktionieren und welche die untere Schranke matchen. Um diese Algorithmen anzuwenden, benutzen wir ein hypergeometrisches Modell als Spektren-Korrelationsmodell und passen es etwas für unsere Zwecke an. Wir implementieren den Algorithmus und eine statistische Auswertung in einem C++-Prototyp und zeigen in einem Zeitmessungsexperiment, dass sich die Rechengeschwindigkeit mit wachsender Datensatzgrösse asymptotisch der Grenze annähert, die durch den Takt der CPU und des Prozessor-Caches gegeben ist. Indem wir die Software auf einen biologischen Datensatz anwenden, gelingt es uns schliesslich, auf der Proteinebene Splice-Varianten zu entdecken, welche bei einer gewöhnlichen Protein-Datenbanksuche nicht identifiziert werden können.

Im zweiten Teil dieser Arbeit präsentieren wir einen Ansatz, um die Qualität von Tandem-Massenspektren zu bewerten, und wir zeigen, dass dieser Ansatz dazu benützt werden kann, bei nur mässigem Mehraufwand die Anzahl der identifizierten Spektren zu erhöhen. In den meisten Proteomics-Experimenten kann nur ein Bruchteil eines Spektren-Datensatzes mittels einer gewöhnlichen Protein-Datenbanksuche identifiziert werden. Einer der Gründe dafür ist, dass je nach Messverfahren Spektren auch dann gemessen werden, wenn die momentane Peptid-Konzentration zu tief ist und das daraus resultierende Signal kaum ausreicht, um das Peptid zu identifizieren. Andererseits können Datensätze auch qualitativ hochwertige Spektren von nicht erwarteten Proteinvarianten enthalten, welche in der Datenbank mit den prognostizierten Spektren fehlen. Wir entwickeln eine Methode, um nicht identifizierte qualitativ hochwertige Spektren aufzuspüren, indem wir Eigenschaften von identifizierten und nicht identifizierten Spektren analysieren. Unter den nicht identifizierten Spek-

tren suchen wir nach untypischen Fällen, die von ihren Eigenschaften her eher den identifizierten Spektren gleichen als den nicht identifizierten Spektren. Danach extrahieren wir diese Spektren, welche meist nur einen kleinen Teil des gesamten Datensatzes ausmachen, und unterziehen diese Spektren aufwändigeren Suchen, die dafür aber eine grössere Anzahl von potenziellen Proteinvarianten berücksichtigen.