Qualitätsbewertung von LC-MS-Maps[1]

Exposé zur Diplomarbeit

September 2006

Autor
Alexander Haupt
Institut für Informatik
Humboldt-Universität zu Berlin
haupt@informatik.hu-berlin.de

Betreuer
Prof. Ulf Leser
Institut für Informatik
Humboldt-Universität zu Berlin

Prof. Knut Reinert
Institut für Informatik
Freie Universität Berlin

[1] *Liquid Chromatography Mass Spectrometry*

1 Motivation

Die automatische Analyse von LC-MS-Massenspektren (*Maps*) zur Identifizierung und Quantifizierung von Proteinen spielt heutzutage eine grundlegende Rolle in der Proteomforschung. Auch in der klinischen Diagnostik werden große Mengen biologischer Proben mittels LC-MS-Massenspektrometrie analysiert. Die Auswertung dieser mehrdimensionalen Spektren (Intensität, Masse/Ladung, Zeit) ist aufgrund der großen Datenmengen jedoch ein zeitaufwändiger Prozess – die Verarbeitung einer Map kann bis zu mehreren Stunden dauern.

Wie alle experimentellen Messdaten unterliegen auch LC-MS-Maps starken qualitativen Schwankungen. Ihre Qualität bestimmt maßgeblich den Erfolg der Identifizierung und Quantifizierung von Proteinen. Speziell in klinischen Studien möchte man ungenaue Ergebnisse durch schlechte Datenqualität vermeiden, kann jedoch nicht hunderte von Maps manuell überprüfen, um sie je nach Qualität gesondert zu verarbeiten oder zu verwerfen.

Eine Bestimmung von Qualitätsmerkmalen und eine entsprechende Untersuchung und Aussortierung von Maps mit zu geringer Qualität könnte die zeitintensive vollständige Analyse des Spektrums ersparen.

2 Zielstellung

2.1 Simulation

Ziel ist die Erstellung von Maps zum Testen der im zweiten Teil zu erstellenden Software. Die Zusammensetzung der virtuellen Probe, einige Parameter der Chromatographie und des Massenspektrometers sowie die Qualität der Maps sollen einstellbar sein.

2.2 Qualitätsprüfung

Der im zweiten Teil zu entwickelnde Algorithmus soll als Teil des OpenMS-Frameworks[1] die Aufgabe übernehmen, qualitativ schlechte Maps vor der weiteren Analyse auszusortieren. Es sollen möglichst alle qualitativ schlechten und möglichst keine qualitativ guten Spektren entfernt werden.

[1] http://www.openms.de/

3 Vorgehensweise

Die Arbeit unterteilt sich in drei Abschnitte. In den ersten anderthalb Monaten erfolgt die Implementation der Simulationssoftware, in den folgenden zweieinhalb Monaten die der Qualitätsüberprüfung (Klassifizierung). Im folgenden fünften Monat soll der Einfluss der Spektrumsqualität auf die Leistungsfähigkeit der Klassifizierung genauer untersucht werden.

3.1 Simulation

Notwendig für eine Simulation von LC-MS-Maps ist sowohl ein gutes Modell für die Vorhersage der Laufzeiten durch die chromatographischen Säulen (*retention time)* sowie eine exakte Berechnung des Massenspektrums anhand der Proteinsequenzen. Neuere Vorhersagemodelle für die Laufzeit, welche die Peptidsequenz statt nur die Aminosäurezusammensetzung auswerten, versprechen Fehlerraten von nur etwa 1,5%, teilweise allerdings erst nach aufwändigem Training der Algorithmen (Petritis et al., 2006; Krokhin et al., 2006). Ein geeigneter Algorithmus muss ausgesucht und eventuell angepasst werden. Für die Berechnung der einzelnen Massenspektren kann auf entsprechende Software zum Beispiel von Rusconi (2006) zurückgegriffen werden.

Selbst implementiert wird das Einlesen der vom Benutzer gewünschten Proteine und Parameter, die Steuerung und das Training der Algorithmen, die Ein- und Ausgabe und Umleitung der Daten sowie der Export der vollständigen LC-MS-Map in den Formaten *mzXML* und *mzData*. Da keine idealen, sondern rauschende und gestörte Spektren erzeugen werden sollen, muss der Algorithmus zur Spektrumsberechnung entsprechend angepasst werden.

Die Bewertung der Simulationsgüte erfolgt durch einen Vergleich mit realen, experimentell gewonnen Spektren.

3.2 Qualitätsprüfung

Um Spektren in *gute* und *schlechte* klassifizieren zu können, müssen vorher die qualitätsbestimmenden Merkmale von LC-MS-Maps identifiziert werden – manuell oder automatisch. Da auf eine große Menge bereits manuell klassifizierter Maps als Trainingsdaten zurückgegriffen werden kann, bietet sich der Einsatz automatischer Klassifizierungsverfahren, z.B. von *Support Vector Machines* an. Bern et al. (2004) klassifizieren MS/MS-Massenspektren mit Hilfe von *SVMs*, während Flikka et al. (2006) bayesianische Algorithmen verwenden. Als problematisch für derartige Algorithmen könnte sich jedoch die Größe der

einzelnen Maps erweisen, die häufig zwischen 300 und 500 Megabytes messen. Möglicherweise muss die Anzahl der betrachteten *Features* beschränkt werden. Auch eine starke Quantisierung der Daten ist denkbar, um die Datenmenge zu reduzieren. Es kann dann untersucht werden, inwieweit eine Klassifizierung zum Beispiel allein auf Basis der visuellen Darstellung der mehrdimensionalen Spektren möglich ist.

Zur weiteren Evaluierung des Klassifizierungsalgorithmus soll abschließend geklärt werden, welche Korrelation zwischen der Qualität der simulierten Spektren (Rausch- und Störungsanteil) und der Leistungsfähigkeit des Klassifizierers besteht.

Aus Gründen der Portierbarkeit, Schnelligkeit und möglichen Einbindung in das OpenMS-Framework erfolgt die Implementierung in C++.

Literatur

M. Bern, D. Goldberg, W. H. McDonald, and J. R. Yates. Automatic quality assessment of Peptide tandem mass spectra. *Bioinformatics*, 20 Suppl 1:I49–I54, Aug 2004. doi: 10.1093/bioinformatics/bth947. URL `http://dx.doi.org/10.1093/bioinformatics/bth947`.

K. Flikka, L. Martens, J. Vandekerckhove, K. Gevaert, and I. Eidhammer. Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics*, 6(7):2086–2094, Apr 2006. URL `http://dx.doi.org/10.1002/pmic.200500309`.

O. Krokhin, S. Ying, J. Cortens, D. Ghosh, V. Spicer, W. Ens, K. Standing, R. Beavis, and J. Wilkins. Use of Peptide Retention Time Prediction for Protein Identification by off-line Reversed-Phase HPLC-MALDI MS/MS. *Anal. Chem.*, 78(17):6265–6269, 2006. URL `http://dx.doi.org/10.1021/ac060251b`.

K. Petritis, L. J. Kangas, B. Yan, M. E. Monroe, E. F. Strittmatter, W.-J. Qian, J. N. Adkins, R. J. Moore, Y. Xu, M. S. Lipton, D. G. Camp, and R. D. Smith. Improved peptide elution time prediction for reversed-phase liquid chromatography-MS by incorporating peptide sequence information. *Anal Chem*, 78(14):5026–5039, Jul 2006. URL `http://dx.doi.org/10.1021/ac060143p`.

F. Rusconi. GNU polyxmass: a software framework for mass spectrometric simulations of linear (bio-)polymeric analytes. *BMC Bioinformatics*, 7:226, 2006. URL `http://dx.doi.org/10.1186/1471-2105-7-226`.