

# Abschlußbericht zum Karl-Steinbuch-Stipendium: Software zur robusten, quantitativen Vorverarbeitung experimenteller Proteomikdaten

**Bernhard Renard**

Heidelberg Collaboratory for Image Processing (HCI),  
Interdisciplinary Center for Scientific Computing (IWR),  
Universität Heidelberg

## Zusammenfassung

Im Fokus des Projektes für das Karl-Steinbuch-Stipendium stand die Entwicklung von Software für die Analyse von experimentellen Proteomikdaten. Durch die zentrale Stellung der Proteomik für die Systembiologie und die Weiterentwicklung der Aufnahmetechnik stehen im zunehmenden Masse komplexe Datensätze zur Verfügung, die nicht mehr manuell ausgewertet werden können, sondern zuverlässig automatisiert vorverarbeitet werden müssen. Hierfür wurde im Rahmen des Projektes ein theoretischer Ansatz der  $L_1$  regularisierten Regression adaptiert und mit der exakten Modellierung der Daten verbunden. Besonderer Augenmerk dabei wurde auf die Robustheit des Verfahrens und die exakte Quantifizierung gelegt. Die Ergebnisse zeigen signifikante Verbesserungen gegenüber momentan verwendeten Verfahren. Die Ergebnisse, die zusammen mit Kollegen am HCI in Heidelberg und dem Children's Hospital in Boston erreicht wurden, sind in mehreren Artikeln in wissenschaftlichen, peer-reviewed Zeitschriften und Konferenzbeiträgen veröffentlicht worden (siehe Kapitel 6), die in den Anhängen 1-6 diesem Bericht beiliegen. Um die Lesbarkeit zu vereinfachen, gibt dieser Abschlußbericht einen Überblick über die Problemstellung (Kapitel 1), die entwickelte Methodik (Kapitel 2), die Implementierung (Kapitel 3) und die Ergebnisse (Kapitel 4) sowie einen Ausblick (Kapitel 5), während für technische Details, Herleitungen, Literaturangaben und Hintergründe auf die entsprechenden, beiliegenden Publikationen verwiesen wird.

## 1 Problemstellung

Die Systembiologie mit ihrem Ansatz, biologisch Komplexe in ihrer Gesamtheit und unter Anwendung quantitativer Methoden zu beobachten, hat in den letzten Jahren zunehmend an Bedeutung gewonnen. Ein Hauptaugenmerk hierbei kommt - insbesondere nachdem die Genomik nicht die in ihr gesetzten Erwartungen im vollen Maße erfüllen konnte - der Proteomik zu. Mit dem Wissen über ausgedrückte Proteine besteht die Möglichkeit die Entstehung von Krankheiten wie bspw. Krebs besser zu verstehen und neuartige Diagnose- und Therapiemethoden zu entwickeln.

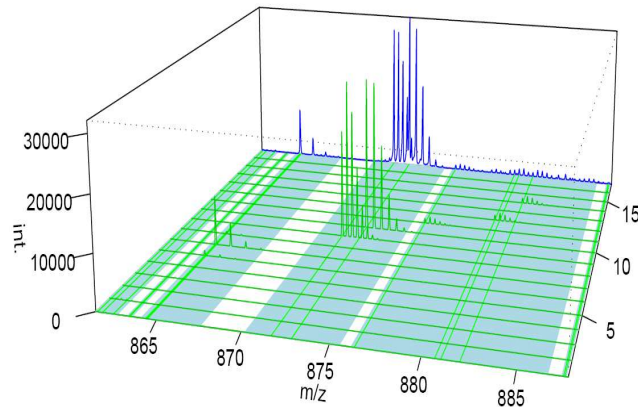


Abbildung 1: Beispiel eines Ausschnitt eines Massenspektrums (blau) und der automatisierten Zerlegung in die darin enthaltenen Peptidmodelle (grün).

Die Massenspektrometrie (MS) ist zum Hauptwerkzeug der modernen Proteomik geworden, weil sie unübertroffene Spezitivität im Vergleich zu anderen Analyseverfahren bietet. Die MS basiert hierbei auf der Idee, daß die Proteine normalerweise erst in Peptide, kürzere Ketten von Aminosäuren, zerlegt und anschließend ionisiert gemessen werden. Die Messung erfolgt hierbei als Zählung der Anzahl der Ionen, die bei einer bestimmten Masse-zu-Ladung-Verhältnis auftreten. Hieraus ergibt sich direkt ein Problem der Analyse von MS-Daten, da mehrere unterschiedliche Peptide in einem Massenbereich auftreten können und verbunden mit Detektorrauschen und chemischen Rauschen die Komplexität der Analyse deutlich erhöhen.

Die Vorverarbeitung proteomischer Massenspektren ist ein nicht-triviales Problem. Dies hat mehrere Ursachen. Aufgrund der schweren Isotope von den in Aminosäuren vorliegenden Molekülen (also bspw. dem Auftreten von  $^{13}\text{C}$  anstelle von  $^{12}\text{C}$  in ca. 1% aller Kohlenstoffatome) wird bei der Bestimmung der Anzahl der geladenen Ionen für ein Peptid, dieses mehrfach für die verschiedenen Massen der Isotope der Atome auftreten. Gleichzeitig können zwei verschiedene Peptide leicht eine ähnliche Masse aufweisen, so daß sich die gegebenen Isotopenmuster überlappen. Ein weiteres Problem ergibt sich daraus, daß Rauschen vorhanden ist; dieses kann entweder durch den Detektor bedingt sein oder dadurch entstehen, daß Lösungsmittel oder Verunreinigungen im Spektrum auftreten oder sich kleinere Moleküle an ein Peptid im geringen Maße ankoppeln. Ein Beispiel ist in Abbildung 1 gegeben.

## 2 Methodik

### 2.1 Peptidmodellierung

Ein grundlegender Punkt für die zuverlässige Analyse von Peptiden ist ein zuverlässiges Peptidmodell. Aufbauend auf einem Ansatz von Senko [Senko 1995], in dem die chemische Zusammensetzung eines Durchschnittspeptids ermittelt wurde, wurde dieser Ansatz von uns erweitert, um durch das theoretische Zulas-

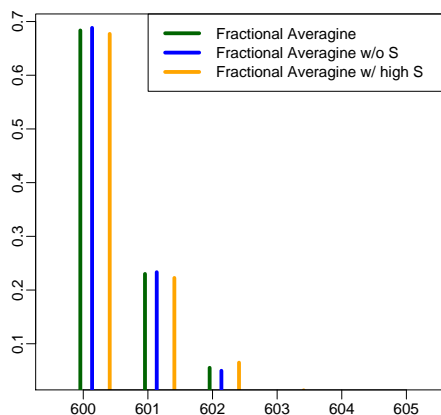


Abbildung 2: Beispiel eines Fractional Averagine Peptidmodells (grün) sowie der beiden Modelle von Extended Fractional Averagine für ein einfach ionisiertes Peptid mit einer Masse von 600 Dalton.

sen von nicht ganzzahligen Zusammensetzungen ein besseres Modell zu erhalten. Dieser Ansatz, Fractional Averagine, läßt sich in der selben Laufzeit berechnen wie der Ansatz von Senko, zeigt aber gerade für leichte Peptide deutlich bessere Modellierungseigenschaften. Abbildung 2 zeigt ein Beispiel für eine solche Peptidmodellierung. Für Details wird auf Anhang 1 und Anhang 3 verwiesen.

## 2.2 Erweiterung des Peptidmodells zur Detektion schwefelhaltiger Moleküle

Eine Erweiterung von Fractional Averagine als Peptidmodell, Extended Fractional Averagine, versucht, stark schwefelhaltige Peptide zuverlässig zu modellieren. Dafür wurde der Ansatz erweitert, um stark schwefelhaltige Peptide im Verlauf der Analyse zunächst zu identifizieren und mit einem eigenen Ansatz zu modellieren, um so eine verbesserte Identifizierung von Massenpositionen von Peptiden zu erlauben. Abbildung 2 verdeutlicht den Ansatz. Neben einem nicht-schwefelhaltigen wird stets ein schwefelhaltiges Peptid modelliert und korrelationsgetrieben das passendere Modell für die weitere Verarbeitung gewählt. Anhang 3 beschreibt diesen Ansatz detailliert.

## 2.3 Regularisierte Regression

Zum Erkennen der Peaks im Massenspektrum verwenden wir einen regularisierten Regressionsansatz. Der Gedanke ist, daß mit Hilfe von Fractional Averagine bzw. Extended Fractional Averagine ein Modell für das Aussehen des Signals eines Peptides geschätzt werden kann. Für jede mögliche Masse in einem Spektrum und für jede mögliche Ladung kann dabei ein solches Modell konstruiert und dann mit Hilfe eines regularisierten, multivariaten Regressionsansatzes [Efron 2003] entschieden werde, welche der konstruierten Modelle tatsächlich im Signal erkannt werden können. Hierfür werden Verfahren der modernen, rechnergestützten Statistik auf das Problem hin adaptiert und optimiert. Durch diesen

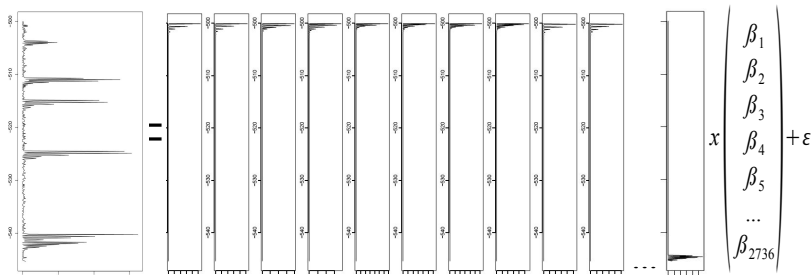


Abbildung 3: Ein Spektrum (links) kann als lineare Mischung aus, oft mehr als 1000, verschiedenen Peptidmodellen (Mitte), die mit einem Faktor  $\beta$  gewichtet werden, interpretiert werden, wobei ein Fehler  $\epsilon$  unerklärt bleibt. Peptidmodelle werden für jede Massenposition und mögliche Ladung konstruiert, der Faktor  $\beta$  wird dann über eine Nicht-negative,  $L_1$  regularisierte Regression bestimmt (zum genauen Vorgehen, siehe Anhang 1).

Ansatz ist es möglich, auch komplexe Überlappungen von Peptiden mit einer hohen Zuverlässigkeit zu identifizieren und zu trennen. Gleichzeitig wird durch die Regularisierung ein Overfitting verhindert und über ein Informationskriterium das richtige Maß an Regularisierung automatisiert gewählt. Abbildung 3 verdeutlicht das Verfahren und in Anhang 1 wird zudem die detaillierte mathematische Herleitung dargestellt.

## 2.4 Erweiterung des algorithmischen Modells auf zweidimensionale Daten

Das generelle theoretische Framework war zunächst auf das Identifizieren der Massenpositionen von Peptiden in eindimensionalen MS-Spektren ausgelegt. Allerdings wurde das Verfahren weiterentwickelt, um auch auf zweidimensionalen Spektren Anwendung finden zu können. Zweidimensionale Spektren entstehen mit Hilfe von Liquid-Chromatography-Massenspektrometrie (LCMS), bei denen vor der Aufnahme der Spektren, die Peptide mit Hilfe einer Flüssigchromatographie getrennt werden. Hierbei wurden zuerst Verfahren aus der Bildverarbeitung genutzt, um zusammenhängende Signalfächen in den zweidimensionalen Spektren zu identifizieren. Die Signale innerhalb dieser Flächen werden dann bezüglich der die Signale im Hinblick auf die Dimension der Flüssigchromatographie aufaddiert. Daraus entsteht wiederum ein eindimensionales Spektrum, was wiederum mit der ursprünglichen Methodik analysiert werden kann. Abbildung 4 zeigt die Identifizierung von Signalfächen und Anhang 5 beschreibt detailliert das Vorgehen.

## 2.5 Erweiterung auf die Identifikation künstlich markierter Moleküle

Es ist in der Massenspektrometrie oft von Interesse, einige Peptide künstlich Änderungen zu unterziehen, bspw. durch das Austauschen von einzelnen Atomen gegen ihre schweren Isotope (bspw. Wasserstoff gegen Deuterium), um so

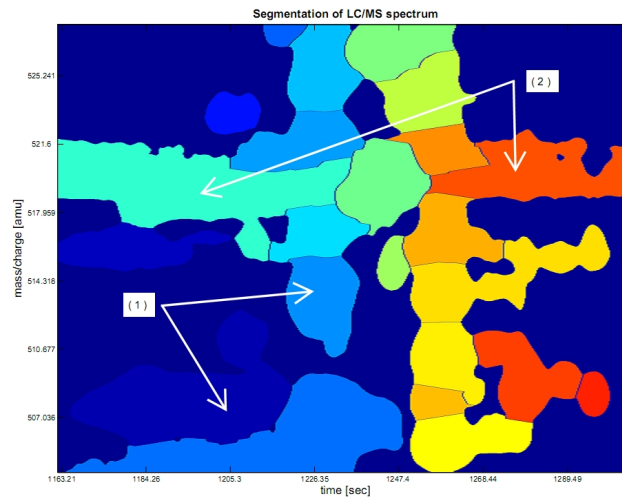


Abbildung 4: Segmentierung eines LC/MS-Spektrums, in dem die resultierenden Signalfächen farblich markiert sind. (1) hebt Flächen hervor, in denen Überlappungen von Peptiden mit dem eindimensionalen Ansatz noch getrennt werden müssen, (2) Regionen, in denen chemische Lösungsmittel das Signal überlagern.

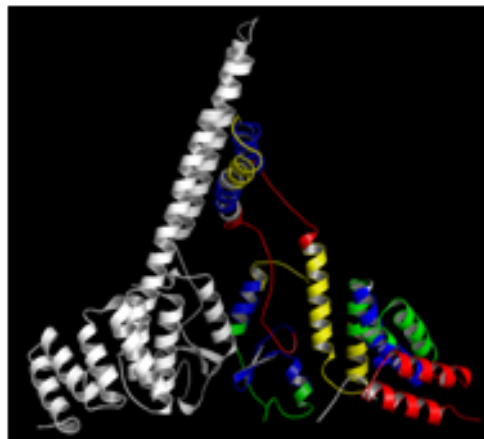


Abbildung 5: Rekonstruktion der Struktur eines Proteins. Die farblichen Hervorhebungen geben die Erreichbarkeit des Moleküls für Wasserstoff-Deuterium-Austauschprozesse, wie sie mit dem vorgeschlagenen Ansatz errechnet werden können, wieder.

spezielle Eigenschaften von Peptiden erkennbar zu machen. Bei der anschließenden Aufnahme im Massenspektrum erscheinen dann die markierten und nicht markierten Moleküle als überlappend. Die vorgeschlagene Methode wurde dahin adaptiert, daß künstliche Änderungen im Peptidmodel berücksichtigt wurden und die entstehenden Überlappungen soweit als möglich eindeutig getrennt werden. Aus den entsprechenden Messungen können Rückschlüsse über die Raumstruktur von Proteinen gewonnen werden (siehe Abbildung 5). Für Details wird hierbei auf Anhang 6 verwiesen.

## 2.6 Anwendung auf räumlich aufgenommene Daten (MS Imaging)

Eine weiteres wichtiges Anwendungsfeld ergibt sich aus einer weiteren Fortentwicklung massenspektrometrischer Verfahren, dem MS Imaging. Hierbei wird nicht nur eine einzelne, isolierte, Probe der Massenspektrometrie unterzogen, sondern vielmehr werden ganze Gewebeschnitte rasterförmig mit einem massenspektrometrisch abgetastet. Hierfür erscheint für jeden Rasterpunkt ein entsprechendes Massenspektrum, indem wieder die Massenpositionen von Peptiden ermittelt werden müssen. Bisher wurde aufgrund der enormen Datenmengen und Laufzeitanforderungen allerdings nur ein vereinfachtes Verfahren eingesetzt (siehe Anhang 2 und 4).

## 3 Implementierung

Die Implementierung des Softwarepakets erfolgte im Frameworks der statistischen Sprache R und baute auf bisher in der Gruppen vorhandene Softwarepakete auf. Zudem erlaubte dies, bestehenden Code für die Regularisierung [Efron 2003] einzubinden und zu adaptieren. Es entstand ein R-Paket was kostenlos, frei und inklusive Offenlegung des Quellcodes unter <http://hci.iwr.uni-heidelberg.de/mip/proteomics/> veröffentlicht wurde. Auf das Softwarepaket wurde auch in den entsprechenden Publikationen hingewiesen (siehe Anhang 1 und 3). Es wurde auch in den ersten drei Monaten seit der Veröffentlichung schon von mehr als ein hundert interessierten Nutzern heruntergeladen.

## 4 Ergebnisse

Im Rahmen des Projektes wurden die Performanz des Algorithmus und der Software intensiv auf simulierten und realen Daten verschiedener Instrumententypen getestet. Zudem wurde die vorgeschlagene Methodik gegen bestehende Algorithmen aus dem Feld verglichen. Hierbei zeigte sich, daß die vorgeschlagene Methodik gegenüber dem momentanen de facto Standard im Feld der Peak Identifikation, THRASH [Horn 2000], einerseits mehr Massenpositionen von Peptiden identifiziert und andererseits auch, die Wahrscheinlichkeit, daß eine gefundene Massenposition korrekt ist, erhöht. Abbildung 6 zeigt das Ergebnis der Anwendung auf einem realen, eindimensionalen Datensatz eines BSA-Samples, während Abbildung 7 die Simulationsergebnisse und den Vergleich von THRASH und NITPICK darstellt. Für eine detailliertere Darstellung der Ergebnisse im Bezug auf Unterpunkte des Algorithmus bzw. im Hinblick auf

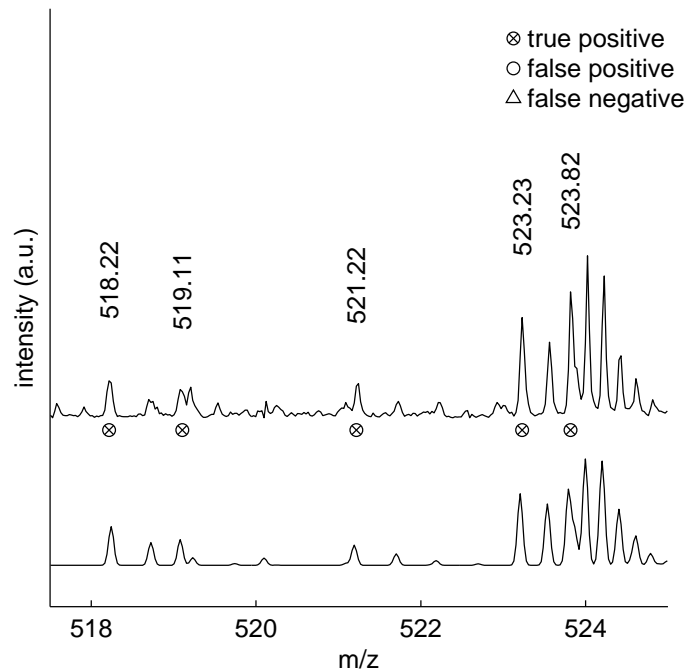


Abbildung 6: Ausschnitt eines BSA-Samples, das mit unserer NITPICK-Methodik analysiert wurde. Es wird deutlich, wie es dem Ansatz gelingt, zwischen Rausch- und Signalpeaks zu trennen und so bspw. das Peptid bei 521,22 Dalton zu identifizieren. Zudem gelingt die korrekte Trennung von zwei sich überlappenden Peptiden bei 523,32 und 523,82 Dalton.

die verschiedenen Anwendung wird auf die Anhänge 1-6 verwiesen, genauso wie für die detaillierte Beschreibung des experimentellen Vorgehens.

## 5 Ausblick

Während im Rahmen dieses Projektes große Fortschritte für die robuste und quantitative Vorverarbeitung von Massenspektren erzielt werden konnte, bleibt es dennoch zu konstatieren, daß diese Problematik keineswegs umfassend gelöst ist. Aufgrund sich stetig verbessernder Aufnahmemöglichkeiten entstehen zunehmend größere Datensätze, die eine beschleunigte Verarbeitung der Daten erfordern. Hierfür wird momentan in unserer Gruppe unter meiner Mithilfe eine C++ Implementierung erstellt, die die Laufzeit des Ansatzes deutlich verringern wird. Die vorliegende R-Implementierung erweist sich hierfür als große Hilfe. Des weiteren arbeiten wir an Verfahren, die die bestehende Algorithmik erweitern, um räumliche Zusammenhänge bei der Analyse von MS Imaging Daten auszunutzen. Gleichzeitig ermöglicht die Vorverarbeitung der Daten nun, Probleme wie die Identifikation von Biomarkern auf Basis von zuverlässig extrahierten Peptiddaten durchzuführen.

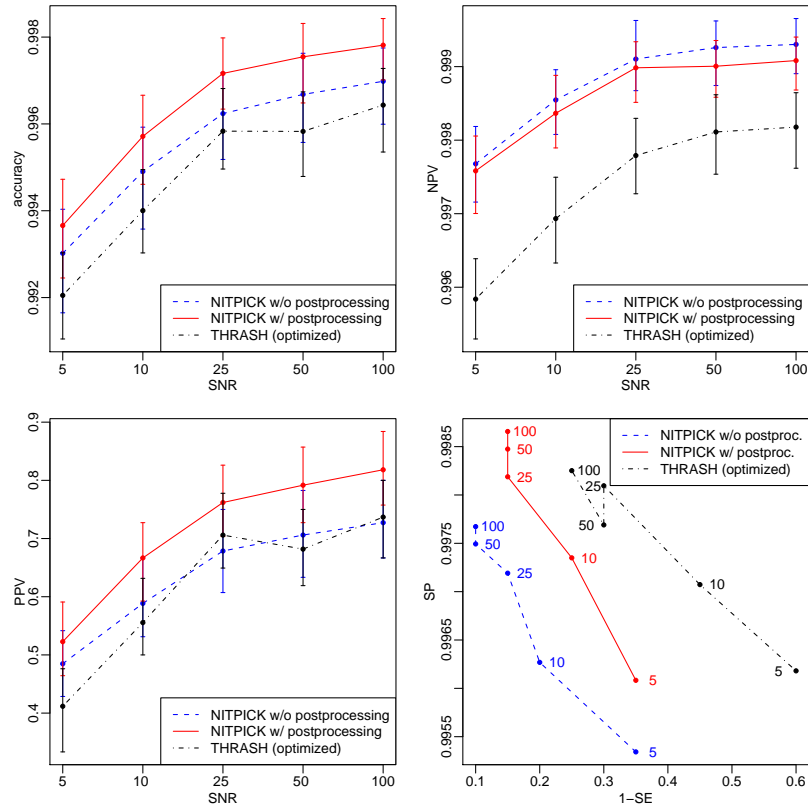


Abbildung 7: Vergleich der Ergebnisse von THRASH (schwarz) und NITPICK (blau bzw. rot mit einem nachgeschalteten Postprocessing) auf 500 simulierten Spektren für 5 verschiedene Signal-Rausch-Verhältnisse (SNR). Die Grafik oben links zeigt die Accuracy als generelles Maß der Performanz der beiden Verfahren für die höhere Werte vorteilhaft sind. Die Grafik oben rechts zeigt das Negative Predictive Value (NPV) als Maß für die Wahrscheinlichkeit, daß wenn an einer Massenposition kein Peptid detektiert wurde, dort auch tatsächlich keines vorliegt, während das Positive Predictive Values (PPV, unten links) die Wahrscheinlichkeit bemißt, daß eine detektierte Massenposition eines Peptides tatsächlich korrekt ist. Bei beiden Werten ist eine größere Wahrscheinlichkeit ein Indikator für einen zu bevorzugenden Ansatz. Das selbe gilt auch für die Darstellung von Sensitivität (SE) und Spezifität (SP) unten rechts. Für alle SNR und alle Maße ist ein bessere Performanz von NITPICK zu beobachten und mit einer Ausnahme (PPV für SNR 25) gilt dies auch, wenn NITPICK ohne Postprocessing angewandt wird.

## 6 Publikation der Ergebnisse

Die Ergebnisse der Arbeit wurden bisher in 2 Journal-Artikeln und 4 Konferenzbeiträgen publiziert. Eine weitere Publikation ist momentan in Vorbereitung und sollte Anfang des kommenden Jahres submitiert werden. Die Publikationen liegen diesem Bericht bei und geben einen deutlich detaillierten Eindruck über das Projekt.

- (Anhang 1) **BY Renard\***, M Kirchner\*, H Steen, JAJ Steen, and FA Hamprecht. NITPICK: Peak Identification for Mass Spectrometry Data, *BMC Bioinformatics*, 2008, 9:355

(Diese Publikation wurde in den vergangengen 3 Monaten bereits über 3000 Mal heruntergeladen und von den Herausgebern von BMC Bioinformatics als 'highly accessed' ausgezeichnet.)

- (Anhang 2) M Hanselmann, **BY Renard\***, M Kirchner\*, ER Amstalden, RMA Heeren, FA Hamprecht. Concise Representation of MS Images by Probabilistic Latent Semantic Analysis, *Analytical Chemistry*, 2008, epub ahead of print

- (Anhang 3) **BY Renard\***, M Kirchner\*, JAJ Steen, H Steen and FA Hamprecht. STRAP2: Reliable, Hierarchical Feature Extraction from Multicomponent Mass Spectra, *Conference of the American Society for Mass Spectrometry (ASMS)*, 2008.

- (Anhang 4) K Hanselmann, **BY Renard\***, M Kirchner\*, A Kharchenko, L Klerk, U Koethe, RMA Heeren and FA Hamprecht. Concise representation of MS Images by Probabilistic Latent Semantic Analysis, *Conference of the American Society for Mass Spectrometry (ASMS)*, 2008.

- (Anhang 5) S Boppel, **BY Renard\***, M Kirchner\*, H Steen, U Koethe and FA Hamprecht. Sparse Profile Reconstruction for LC/MS Feature Extraction, *Conference of the American Society for Mass Spectrometry (ASMS)*, 2008.

- (Anhang 6) X Lou, **BY Renard\***, M Kirchner\*, U Koethe, H Steen, MA Mayer and FA Hamprecht. Fully Automated HX-MS Data Analysis with Complete Deuteration Distribution Estimation, *Conference of the American Society for Mass Spectrometry (ASMS)*, 2008.

**BY Renard\***, M Kirchner\*, H Steen, JAJ Steen, and FA Hamprecht. NITPICK<sup>H</sup>: Improved Peak Identification for Mass Spectrometry Data, *Manuscript in Preparation*

**BY Renard\***, M Kirchner\*, H Steen, JAJ Steen, and FA Hamprecht. NITPICK R-Package, <http://hci.iwr.uni-heidelberg.de/mip/proteomics/>