Entwicklung einer flexiblen bioinformatischen Plattform zur Analyse von Massenspektrometriedaten

DISSERTATION

zur Erlangung des akademischen Grades

Dr. med.

an der
Medizinischen Fakultät der Universität Leipzig

eingereicht von

Sebastian Gibb
geboren am 16.03.1987 in Greifswald

angefertigt am
Institut für Medizinische Informatik, Statistik und Epidemiologie der Universität Leipzig

Betreut von
Prof. Dr. Korbinian Strimmer

Beschluss über die Verleihung des Doktorgrades vom 09.09.2015

Inhaltsverzeichnis

Bibliographische Beschreibung

Sebastian Gibb

Entwicklung einer flexiblen bioinformatischen Plattform zur Analyse von Massenspektrometriedaten

Universität Leipzig, Dissertation

76 Seiten, 104 Literaturangaben, 15 Abbildungen, 5 Tabellen, 3 Anhänge

Sowohl in der Klinischen Labormedizin, der Klinischen Mikrobiologie als auch in der Pathologie ist die Massenspektrometrie (MS) ein bedeutender Bestandteil der Diagnostik geworden. Der Fortschritt in der Gerätetechnik ermöglicht in kurzer Zeit viele, hochaufgelöste Spektren zu generieren. Diese Informationsvielfalt macht die manuelle Auswertung durch den Anwender sehr kompliziert bis unmöglich. Aus diesem Grund ist die Unterstützung durch bioinformatische Programme notwendig. Für die Reproduzierbarkeit der Ergebnisse und die Qualitätskontrolle ist es essentiell, dass die verwendeten Algorithmen transparent und die Programme als *Open Source Software* (OSS) frei verfügbar sind (Aebersold and Mann, 2003).

Das Ziel dieser Arbeit war die Entwicklung von `MALDIquant`, einer unter der *GNU General Public License* (GPL) stehenden, flexiblen OSS, die für die o.g. Anwendungsbereiche modernste Algorithmen für die komplette Analyse bietet und in der freien Programmiersprache *R* (R Core Team, 2014) geschrieben ist. Im Zusammenspiel mit dem dazugehörigen Paket

`MALDIquantForeign` ist `MALDIquant` in der Lage die üblichen Dateiformate der verschiedenen MS-Geräte zu verarbeiten. Dadurch ist `MALDIquant` hersteller- und geräteunabhängig und eignet sich nicht nur für MALDI/TOF, sondern für alle zweidimensionalen MS-Daten.

Angefangen vom Datenimport über die Prozessierung bis hin zur Analyse der Spektren bietet `MALDIquant` eine komplette Analyse-Pipeline und implementiert state-of-the-art Methoden. Neben weit verbreiteten Verfahren zur *Baseline Correction* und *Peak Detection* zeichnet sich `MALDIquant` besonders durch ein hervorragendes *Peak Alignment* aus. Dieses ist sehr genau und aufgrund des Fokus auf die Peaks schneller als die meisten anderen Verfahren und weitestgehend unabhängig von der Qualität der Intensitätenkalibrierung. Eine weitere Stärke von `MALDIquant` ist die Möglichkeit, eigene Algorithmen zu integrieren, sowie den Ablauf der Analyse den individuellen Bedürfnissen anzupassen.

In der beispielhaften Analyse der Daten von Fiedler et al. (2009) konnten durch `MALDIquant` Peaks gefunden werden, die Patienten mit Pankreaskarzinom von nicht erkrankten Probanden unterscheiden. Einige dieser Peaks wurden bereits in anderen Publikationen beschrieben. Neben diesem Beispiel hat `MALDIquant` seine Nützlichkeit bereits in verschiedenen Anwendungsbereichen und Publikationen bewiesen, wie etwa in Ouedraogo et al. (2013) oder Jung et al. (2014).

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ASCII *American Standard Code for Information Interchange*, Standard zur Zeichenkodierung für Textdateien.

CSV *Comma-Separated Values*, textbasiertes Dateiformat für die Speicherung von Tabellen.

CRAN *Comprehensive R Archive Network*, Verbund von öffentlichen Servern zur Verteilung von *R*-Paketen.

COW *Correlation Optimized Warping*, Verschiebung, Streckung oder Stauchung eines Spektrums zur Anpassung an ein Referenzspektrum mit dem Ziel die Korrelation beider zu maximieren.

DDA Diagonale Diskriminanzanalyse, lineare Diskriminanzanalyse mit diagonaler Kovarianzmatrix.

DTW *Dynamic Time Warping*, Streckung oder Stauchung eines Spektrums zur Anpassung an ein Referenzspektrum mittels Dynamischer Programmierung.

GC-MS *Gas Chromatography - Mass Spectrometry*, Kombination der Gaschromatographie mit der Massenspektrometrie.

GPL *GNU General Public License*, freie Softwarelizenz der *Free Software Foundation*, einer von Richard Stallman 1985 gegründeten, gemeinnützigen Organisation zur Förderung freier Software.

HCCA α-Cyano-4-hydroxyzimtsäure.

LC-MS *Liquid Chromatography - Mass Spectrometry*, Kombination der Flüssigchromatographie mit der Massenspektrometrie.

LOWESS *LOcally WEighted Scatterplot Smoothing*, lokal gewichtete, polynomiale Regression.

MA *Moving Average*, gleitender Mittelwert.

MAD *Median Absolute Deviation*, Median der absoluten Abweichungen.

MALDI *Matrix-Assisted Laser Desorption/Ionisation*, Verfahren zur Ionisation von Molekülen, basierend auf einer Kokristallisation von Matrix und Analyt. Die von der Matrix absorbierte Laserenergie ermöglicht das Herauslösen und die Ionisation der Moleküle des Analyten.

MS Massenspektrometrie oder Massenspektrometer.

m/z *mass-to-charge ratio*, Verhältnis von Masse zu Ladung.

NetCDF *Network Common Data Format*, binäres Datenformat für multidimensionale Daten.

OSS *Open Source Software*, quelloffene Software.

PF4 *Platelet Factor 4*, Protein, das an der Blutgerinnung beteiligt ist.

PQN *Probabilistic Quotient Normalization*, Verfahren zur Intensitätenkalibrierung (Dieterle et al., 2006).

PTW *Parametic Time Warping*, Streckung oder Stauchung eines Spektrums zur Anpassung an ein Referenzspektrum mittels polynomialer Funktionen (Bloemberg et al., 2010).

SNIP *Statistics-sensitive Non-linear Iterative Peak-clipping*, Verfahren zur Grundlinienkorrektur (Ryan et al., 1988).

SNR *Signal-to-Noise-Ratio*, Verhältnis von Signal zu Rauschen.

TIC *Total Ion Current*, entspricht der Fläche unter der Spektrumkurve.

TOF *Time-Of-Flight*, Massenanalysator, in dem das Masse-zu-Ladungs-Verhältnis durch die Messung der Flugzeit der geladenen Teilchen ermittelt wird.

1 Einleitung

1.1 Intention

Seit jeher ist die Therapie von Krankheiten ein Wettlauf gegen die Zeit. Eine zeitige Diagnose verbessert in der Regel das Therapieergebnis. Während die Ärzte bei der Diagnostik früher auf ihre fünf Sinne und ihre Intuition angewiesen waren, werden sie heutzutage durch laborchemische und bildgebende Verfahren unterstützt. Diese ermöglichen es bei einigen Krankheiten eine Diagnose zu stellen, bevor der Patient eindeutige, wahrnehmbare Symptome zeigt.

Seit einigen Jahren gewinnen Genomik und Proteomik zunehmend an Bedeutung. Sie nähren die Hoffnung, Diagnosen noch frühzeitiger und kostengünstiger zu stellen. Da die meisten Funktionen des Organismus von Proteinen gesteuert werden, kann die Proteomik neue Einblicke in die Entstehung von Erkrankungen geben. Durch die hohe Dynamik und die Abhängigkeit von Umgebungsbedingungen eignet sich das Proteom hervorragend für die Suche nach neuen, spezifischen Biomarkern

Nach wie vor stellt es eine große und bislang ungelöste Herausforderung dar, mit proteomischen Methoden entdeckte Biomarker in die klinische Routine zu überführen (Diamandis, 2010). Ein Problem ist die häufig fehlende Reproduzierbarkeit der publizierten Ergebnisse. Mögliche Ursachen liegen in fehlenden Standards in der Präanalytik, der Analyse und in der Auswertung der Daten eines MS-Experiments (Aebersold and Mann, 2003; Leichtle et al., 2013).

Für Letzteres wird häufig gerätespezifische, proprietäre Software verwendet, deren Algorithmen unbekannt sind und die zwischen verschiedenen Geräten und Laboren nicht vergleichbar ist. Um die Auswertung der MS-Daten zu standardisieren, um Vergleiche zwischen verschiedenen Geräten und Laboren zu ermöglichen und eine höhere Reproduzierbarkeit zu erreichen, ist offene und transparente Software notwendig (Aebersold and Mann, 2003; Gentleman et al., 2004).

Das Ziel dieser Arbeit war die Entwicklung von `MALDIquant`, einer freien und offenen Software, die moderne und anerkannte Algorithmen für die Analyse von MS-Daten bietet und so ihren Teil zur Integration der Proteomik in den klinischen Alltag beiträgt.

1.2 Eigene Beiträge

Im Rahmen dieser Dissertation wurde `MALDIquant` entwickelt. `MALDIquant` ist eine freie und flexible Plattform zur Analyse zweidimensionaler MS-Daten wie z.B MALDI/TOF MS-Daten. Im Gegensatz zu der üblichen Hersteller- und Gerätesoftware ist `MALDIquant` in der Lage individuelle Arbeitsabläufe abzubilden, Spektren unterschiedlicher Auflösung und technische sowie biologische Replikate zu verarbeiten. Des Weiteren enthält es moderne, anerkannte und gut dokumentierte Methoden zur Analyse von MS-Daten. Besonders hervorzuheben ist das nicht-lineare *Peak Alignment* (siehe Abschnitt 3.7). Aufgrund seines offenen Charakters kann `MALDIquant` bei Bedarf einfach erweitert werden.

Software

Zusätzlich zu `MALDIquant` entstanden weitere, mit `MALDIquant` assoziierte, freie Software-Pakete in der freien Programmiersprache *R* (R Core Team, 2014). Alle Pakete sind auf den Servern des *Comprehensive R Archive Network* (CRAN) verfügbar und unterstehen der GPL. Zusammen umfassen sie mehr als 6000 Zeilen *R*- und *C*-Code sowie mehr als 4000 Zeilen Dokumentation (siehe auch Tabellen B.1-B.4).

- `MALDIquant`: Ein Paket zur Analyse von MS-Daten (siehe Tabelle B.1 sowie `http://github.com/sgibb/MALDIquant`).
- `MALDIquantForeign`: Ein Paket zum Im- und Export verschiedener MS-Dateiformate (siehe Tabelle B.2 sowie `http://github.com/sgibb/MALDIquantForeign`).
- `readBrukerFlexData`: Ein Paket zum Einlesen von Rohdaten von *Bruker Daltonics *flex*-MS-Geräten (siehe Tabelle B.3 sowie `http://github.com/sgibb/readBrukerFlexData`).
- `readMzXmlData`: Ein Paket zum Einlesen von Daten im *mzXML* Format (siehe Tabelle B.4 sowie `http://github.com/sgibb/readMzXmlData`).

Publikation

Eine Beschreibung von `MALDIquant` wurde 2012 in der Fachzeitschrift *Bioinformatics* publiziert (siehe auch Anhang A):

- Gibb, S. and Strimmer, K. (2012). MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, 28:2270–2271.

1.3 Übersicht

Die nachfolgenden Kapitel erläutern die in `MALDIquant` implementierten Methoden und ihre Anwendung. Zum besseren Verständnis werden zu Beginn einige Hintergründe erklärt. Der Abschnitt 2.1 informiert über den Wissenschaftszweig Proteomik. In Abschnitt 2.2 wird der Aufbau und die Funktionsweise eines MS anhand eines MALDI/TOF MS verdeutlicht. Der folgende Abschnitt 2.3 zeigt, dass freie Software zur Auswertung der MS-Daten notwendig ist.

Eine detailierte Beschreibung der einzelnen in `MALDIquant` genutzten Algorithmen ist in Kapitel 3 in der Abfolge einer typischen Analyse zu finden. Ihre Nutzung wird in Kapitel 4 durch eine beispielhaften Analyse eines realen Datensatzes aus Fiedler et al. (2009) veranschaulicht. Schließlich vergleicht Kapitel 5 unsere Ergebnisse mit denen aus Fiedler et al. (2009) und weist auf verschiedene Anwendungsgebiete von `MALDIquant` hin.

2 Hintergrund

2.1 Proteomik

Der Begriff des Proteoms wurde erstmalig von Marc Wilkins während eines Kongresses in Siena 1994 verwendet. Er definierte das Proteom als Gesamtheit aller Proteine, die von einem Genom einer Zelle oder eines Gewebes exprimiert werden. Des Weiteren stellte er heraus, dass das Proteom zwar ein Produkt des Genoms, jedoch im Gegensatz zu diesem nicht einzigartig für einen Organismus ist. Denn je nach Umgebungsbedingungen, nach Zell- oder Gewebetyp kann sich das Proteom im selben Organismus quantitativ als auch qualitativ völlig unterschiedlich darstellen (Wilkins et al., 1996).

Das Proteom ist abhängig von wechselnden Umgebungsfaktoren, wie etwa dem pH-Wert, der Temperatur, dem Vorhandensein von Sauerstoff oder von Medikamenten. Dadurch bildet das Proteom, als funktionelles Endprodukt, den aktuellen Zustand des biologischen Systems bzw. des Organismus sowie dessen Reaktion auf die Umwelt ab (Banks et al., 2000).

Den Wissenschaftszweig, der sich mit der Erforschung der Zusammensetzung des Proteoms des jeweiligen biologischen Systems und der unterschiedlichen Eigenschaften der Proteine, ihrer Funktionen sowie ihren Wechselwirkungen untereinander widmet, bezeichnet man als Proteomik (Patterson and Aebersold, 2003).

Neben der Struktur- und Funktionsaufklärung ist ein wichtiges Teilgebiet der Proteomik die vergleichende Analyse verschiedener biologischer Systeme, z.B. verschiedener Spezies (Bernardo et al., 2002; Feltens et al., 2010), sowie verschiedener Zustände gleicher biologischer Systeme, z.B. krank und gesund (Wulfkuhle et al., 2003; Fiedler et al., 2009). Man hofft dabei vor allem auf die Entdeckung von Mustern, respektive Markern, die eine Identifikation der Systeme oder der Zustände ermöglichen. Diese objektiv messbaren Indikatoren für physiologische oder pathologische Prozesse werden als Biomarker bezeichnet (Diamandis, 2010).

In den letzten Jahren wurden eine Vielzahl proteomischer Biomarker für die unterschiedlichsten Krebsarten beschrieben, z.B. für das Ovarialkarzinom (Petricoin et al., 2002), das Nicht-Kleinzellige-Lungenkarzinom (Yanagisawa et al., 2003), das Hepatozelluläre Karzinom (Schwegler et al., 2005), das Kolorektale Karzinom (de Noo et al., 2006; Zhu et al., 2013), das Pankreaskarzinom (Fiedler et al., 2009) und viele andere.

Das Finden solcher Biomarker ist ein aufwendiger Prozess, der von der Probenentnahme, deren Vorverarbeitung im Labor, der Analyse mittels MS und der bioinformatischen sowie der statistischen Auswertung abhängt.

2.2 Massenspektrometrie

Die Massenspektrometrie (MS) hat in den letzten Jahren als Methode zur Gewinnung von proteomischen Informationen dramatisch an Bedeutung zugenommen (Aebersold and Mann, 2003).

Ein Massenspektrometer (MS) ist ein Gerät, das aus dem Analyten Ionen erzeugt, diese nach den *mass-to-charge ratios* (m/z) auftrennt und schließlich deren Art und Anzahl detektiert (Gross, 2004, S. 2–3).

Es existiert eine Vielzahl verschiedener MS-Typen. Im folgenden werden die allgemeinen Prinzipien eines MS am Beispiel eines *Matrix-Assisted Laser Desorption/Ionisation* (MALDI)/*Time-Of-Flight* (TOF) MS erläutert. Zum einen weil das MALDI/TOF MS einem anschaulichen Aufbau folgt und zum anderen, weil alle uns vorliegenden Daten auf eben so einem Gerät erzeugt worden sind. Das MALDI/TOF MS zeichnet sich außerdem durch eine hohe Massengenauigkeit, hohe Auflösung und hohe Sensitivität aus (Aebersold and Mann, 2003). Des Weiteren wird MALDI als sanfte Ionisationsmethode betrachtet, die die Moleküle kaum fragmentiert und sich somit hervorragend für die Analyse intakter Peptide eignet (Aebersold and Mann, 2003).

Jedes MS besteht aus drei Modulen: der Ionenquelle, dem Massenanalysator und dem Detektor (siehe auch Abb. 2.1 sowie Gross (2004, S. 3)).

Die zu analysierende Probe wird bei der MALDI Methode mit einer lichtabsorbierenden, organischen Matrix wie z.B. α-Cyano-4-hydroxyzimtsäure. (HCCA) kokristallisiert. Nach dieser Vorbereitung wird die Probe in die Ionenquelle platziert und mit einem gepulsten UV-Laser beschossen. Die von der Matrix absorbierte Laserenergie ermöglicht das Herauslösen sowie die Ionisation der Atome oder Moleküle des Analyten. Durch ein elektromagnetisches Feld werden die Ionen gebündelt, beschleunigt und in den Massenanalysator überführt (siehe auch Abb. 2.1 sowie Gross (2004, S. 411–440)).

Im Falle des MALDI/TOF MS ist der Massenanalysator eine einfache, feldfreie Vakuumröhre. In dieser trennen sich die zuvor gleichmäßig beschleunigten Ionen entsprechend ihrer Massen auf: leichte Ionen durchfliegen die Röhre schneller als schwere und erreichen den Detektor früher (siehe auch Abb. 2.1, sowie Gross (2004, S. 113–124)).

Der Detektor zählt die Anzahl der eintreffenden Ionen und misst die vergangene Zeit vom Laserbeschuss bzw. vom Anlegen der Beschleunigungsspannung bis zum Eintreffen der Ionen (siehe auch Abb. 2.1, sowie Gross (2004, S. 175–180)).

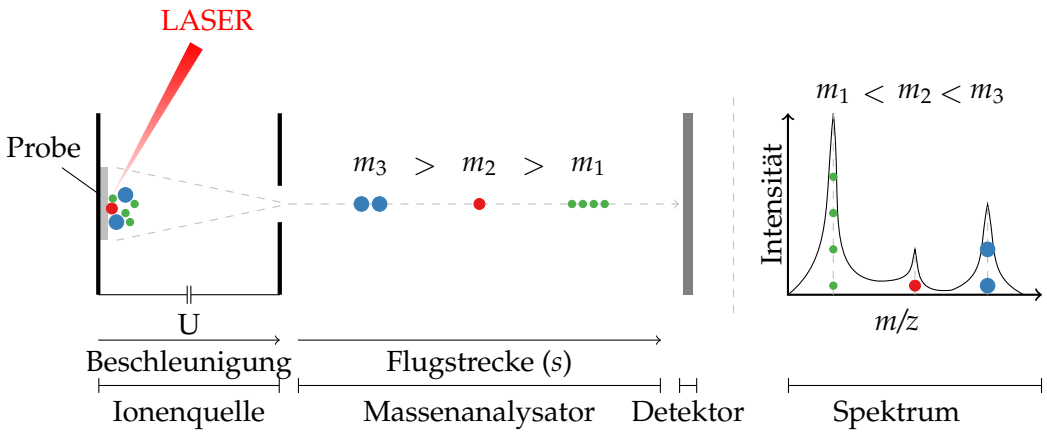


Abbildung 2.1: Schematischer Aufbau eines MALDI/TOF MS nach Gross (2004, Fig. 4.4, S. 117). Die vom gepulsten Laser aus der Probe herausgelösten Ionen werden beschleunigt (U) und fliegen durch eine feldfreie Vakuumröhre, den Massenanalysator (s), zum Detektor. Je leichter die Ionen sind (m_1-m_3) desto früher erreichen sie diesen. Anschließend werden die Flugzeiten (t) in m/z umgerechnet und gegen die Anzahl der Ionen (Intensitäten) als Spektrum aufgetragen. Siehe auch Gleichung 2.1.

Schließlich wird die Anzahl der Ionen sowie deren Flugzeiten in Intensitäten und m/z-Werte umgerechnet. Dies erfolgt mit Hilfe der physikalischen Gleichung

$$\frac{m}{z} = \frac{2eUt^2}{s^2}, \tag{2.1}$$

die die Elementarladung e, die Beschleunigungsspannung U, die Flugzeit t sowie die Flugstrecke s miteinander verbindet.

2.3 Bioinformatik

Die Massenspektrometrie zählt zu den Hochdurchsatzverfahren. Darunter versteht man Verfahren, die in einem kurzen Zeitraum sehr viele Messungen ermöglichen und eine große Anzahl Daten produzieren. Aufgrund dieser Vielzahl an Daten ist die Auswertung durch einen erfahrenen Experimentator kaum zu bewältigen. Bioinformatische Softwarelösungen assistieren bei der Auswertung und bieten ein hohes Maß an Automatisierung.

Jeder Gerätehersteller liefert zu seinem Gerät eine entsprechende Software mit. Diese Software ist in der Regel proprietär und die verwendeten Algorithmen werden weder beschrieben, noch werden die Datenformate offen gelegt. Die Integration neuer oder das Evaluieren verschiedener Methoden durch den Wissenschaftler ist nicht vorgesehen. Vergleiche, die Verifizierung oder die Reproduktion der Ergebnisse, die mit diesen Analyseprogrammen erstellt wurden, sind kaum möglich und der wissenschaftliche Nutzen ist

daher gering. Auch die komplizierten und undokumentierten Datenformate erschweren bzw. verhindern eine geräte- und herstellerunabhängige Auswertung.

Durch die Entwicklung und Verwendung transparenter Programme und Datenformate kann die Reproduzierbarkeit und der Nutzen der Auswertung verbessert und der wissenschaftliche Fortschritt im Allgemeinen gefördert werden (Aebersold and Mann, 2003; Gentleman et al., 2004).

Transparente Software bedeutet OSS. Als OSS definiert man Software, die frei verfügbar und deren Quellcode öffentlich einsehbar, nachvollziehbar, erweiterungs- und anpassungsfähig ist (siehe auch Open Source Initiative (2014)). Neben der einfachen Offenlegung des Quellcodes ist aber auch die Beschreibung der implementierten Algorithmen sowie die Dokumentation der Verwendung der Software für die Reproduzierbarkeit essentiell. Leider ist festzustellen, dass die Entwicklung freier Software der Entwicklung der technischen Geräte häufig um Jahre hinterher hängt (Leichtle et al., 2013).

Seit einiger Zeit gibt es einige wegweisende OSS für MS-Daten, wie PROcess (Li, 2005), XCMS (Smith et al., 2006), OpenMS (Sturm et al., 2008), ProteoWizard (Kessner et al., 2008), mMass (Strohalm et al., 2008) oder MSnbase (Gatto and Lilley, 2012).

Jedoch sind all diese Programme entweder auf *Gas Chromatography - Mass Spectrometry* (GC-MS), *Liquid Chromatography - Mass Spectrometry* (LC-MS) oder Tandem-MS ausgelegt oder durch ihre Komplexität schwer erweiterungs- und anpassungfähig. Außerdem fehlte ein Programm, das speziell auf die klinische Diagnostik und klinische Fragestellungen ausgerichtet ist, für MALDI/TOF oder zweidimensionale MS-Daten entworfen wurde, technische bzw. biologische Replikate sowie verschiedene Auflösungen berücksichtigen kann.

Diese Lücke füllt nun `MALDIquant` (Gibb and Strimmer, 2012). `MALDIquant` ist eine OSS, die unter der GPL lizenziert und in der freien Programmiersprache *R* (R Core Team, 2014) geschrieben ist. *R* ist eine relativ einfach zu erlernende und im Bereich Statistik und Datenanalyse sehr populäre Programmiersprache. Sie ist einfach erweiterungsfähig und verfügt mittlerweile über 5000 Zusatzpakete, die hervorragende und sehr aktuelle Verfahren zur statistischen Auswertung bieten. Außerdem gibt es das Bioconductor-Projekt (Gentleman et al., 2004), das viele bioinformatische *R*-Pakete beinhaltet und reproduzierbare Wissenschaft mit *R* fördert.

Neben OSS wurden auch offene Datenformate entwickelt. So entstanden das *mzXML*-Format und sein Nachfolger *mzML* (Pedrioli et al., 2004; Martens et al., 2011), die mittlerweile von vielen Herstellern als Exportformat angeboten werden. `MALDIquant` kann dank des assoziierten Paketes `MALDIquantForeign` (Gibb, 2014) diese Datenformate ebenfalls verarbeiten.

3 Methoden

3.1 Überblick

Die von einem MS generierten Spektren können nicht direkt zur statistischen Analyse verwendet werden. Es sind einige Arbeitsschritte in einer bestimmten Reihenfolge notwendig um die Daten entsprechend vorzubereiten. In Abb. 3.1 ist ein typischer Arbeitsablauf dargestellt (siehe auch Norris et al. (2007); Morris et al. (2010)). Zuerst müssen die Rohdaten in die *R*-Umgebung eingelesen werden (*Data Import*). Als nächstes folgt die Varianzstabilisierung und das Glätten der Spektren (*Smoothing*). Zur Minimierung von Matrixeffekten sowie chemischer Verunreinigungen wird die Grundlinie des Spektrums korrigiert (*Baseline Correction*). Des Weiteren müssen die Intensitäten aller Spektren zur besseren Vergleichbarkeit angeglichen werden (*Intensity Calibration*). Anschließend werden zur Datenreduktion lokale Maxima (Peaks) gesucht und die restlichen Spektrendaten verworfen (*Peak Detection*). Da sich neben den Intensitäten auch die *m/z*-Werte über verschiedene Messungen ändern, müssen auch diese adjustiert werden (*Peak Alignment* und *Peak Binning*). Schließlich wird aus den Peaklisten eine sog. *Feature Matrix* generiert, die dann zur statistischen Auswertung, wie z.B. Klassifikation oder Variablenselektion, an entsprechende Algorithmen übergeben werden kann.

3.2 Import der Rohdaten

Der erste Schritt einer jeden Analyse ist der Datenimport. Dieser gestaltet sich in der Massenspektrometrie trotz der Existenz freier Datenformate wie *mzXML* (Pedrioli et al., 2004) und *mzML* (Martens et al., 2011) häufig schwierig. Das im Rahmen dieser Arbeit entwickelte *R*-Paket `MALDIquantForeign` bietet hier eine einfache Schnittstelle zwischen den Rohdaten und `MALDIquant`. Es unterstützt neben den freien Datenformaten *mzXML* (Pedrioli et al., 2004; Gibb, 2013b) und *mzML* (Martens et al., 2011) auch noch *Ciphergen XML*, *American Standard Code for Information Interchange* (ASCII), *Comma-Separated Values* (CSV), *Network Common Data Format* (NetCDF) sowie die Rohdaten der *Bruker Daltonics *flex Serie* (Gibb (2013a); siehe z.B. Abb. 3.2 und Anhang C, Abschnitt 3.2). Zusätzlich kann `MALDIquantForeign` noch die für MALDI-*Imaging* genutzten Datenformate *imzML* (Schramm et al., 2012) und *ANALYZE 7.5* (Robb et al., 1989) importieren. `MALDIquantForeign` liest

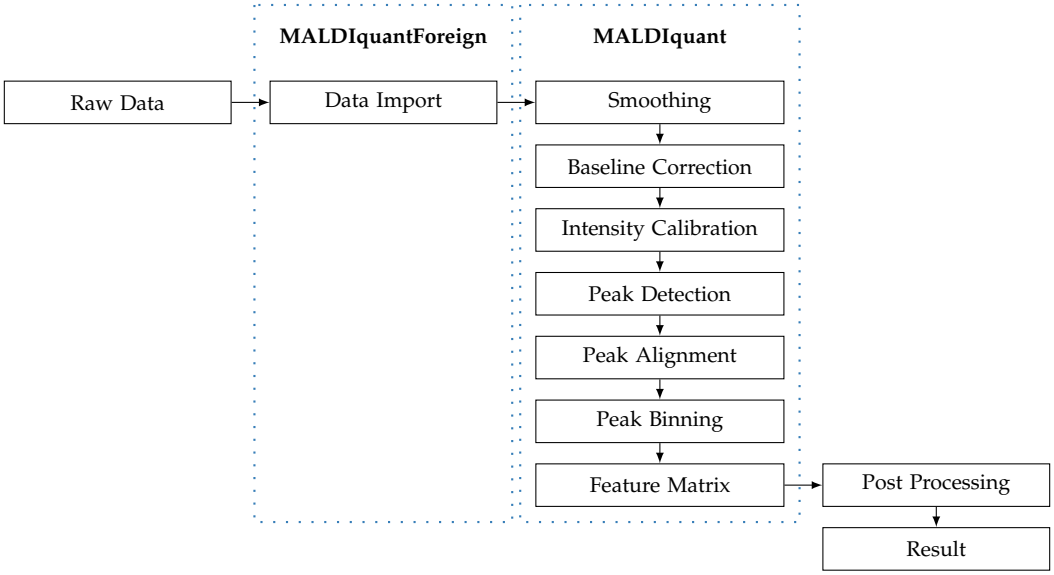


Abbildung 3.1: Schematischer Ablaufplan einer Analyse von MS-Daten mit `MALDIquantForeign` und `MALDIquant`.

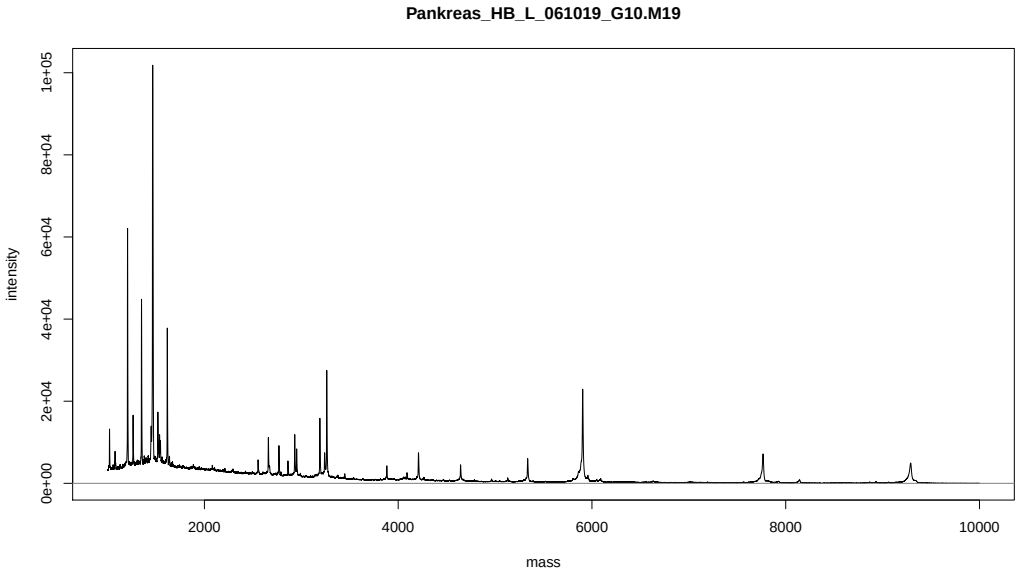


Abbildung 3.2: Beispiel für ein unbearbeitetes MALDI/TOF MS-Spektrum aus Fiedler et al. (2009).

nicht nur einzelne Dateien, sondern sucht ganze Verzeichnisbäume nach MS-Daten ab. Dabei erkennt es das entsprechende Datenformat in der Regel automatisch, auch wenn die Dateien komprimiert sind oder auf einem Server im Internet liegen.

3.3 Transformation der Intensitäten

Die eingelesenen Spektren werden als nächstes einer Transformation der Intensitäten unterzogen. Durch diese Datentransformation erleichtert man sich zum einen die grafische Darstellung zum anderen dient sie der Varianzstabilisierung.

Die Intensitäten eines Spektrums folgen näherungsweise der Poisson-Verteilung (Sköld et al., 2007; Du et al., 2008). Diese zeichnet sich dadurch aus, dass der Mittelwert und die Varianz identisch sind. Daraus folgt, dass sich die Varianz mit schwankenden Mittelwerten ebenfalls verändert. Nach einer Stabilisierung mit einer Wurzeltransformation ($f(x) = \sqrt{x}$) ist die Varianz annähernd konstant und damit die Grundvoraussetzung für eine Vielzahl statistischer Tests gegeben.

`MALDIquant` bietet für die Varianzstabilisierung eine eigene Funktion, die neben der für MS-Daten empfohlenen Wurzeltransformation (Purohit and Rocke, 2003) auch die häufig genutzten logarithmischen Transformationen (Tibshirani et al., 2004; Coombes et al., 2005) unterstützt (siehe auch Anhang C, Abschnitt 3.4).

Rauschen, also kleine, hochfrequente Schwankungen und Unebenheiten werden vor allen weiteren Schritten mit einem Glättungsverfahren herausgefiltert (*Smoothing*). Neben dem klassischen *Moving Average* (MA) bietet `MALDIquant` noch den *Savitzky-Golay*-Filter (Savitzky and Golay, 1964). Er basiert auf polynomialen Regressionen in einem sich über das Spektrum bewegenden Fenster. Im Gegensatz zum MA bewahrt der *Savitzky-Golay*-Filter lokale Maxima und Minima trotz Glättung der Grundlinie (siehe auch Anhang C, Abschnitt 3.4 und Abb. 3.3).

3.4 Korrektur der Grundlinie

Die Grundlinie (*Baseline*) ist durch sog. chemisches Rauschen, wie etwa Matrixeffekte und chemische Verunreinigungen, erhöht. Um die Einflüsse dieser Grundlinienartefakte in der nachfolgenden Analyse möglichst gering zu halten, wird eine sog. Korrektur der Grundlinie (*Baseline Correction*) durchgeführt. Über die Zeit wurde eine Vielzahl von Algorithmen entwickelt um dieses Problem zu lösen. Angefangen von relativ einfachen Methoden wie der Subtraktion des absoluten Minimums der Intensitäten (Gammerman et al., 2008), des gleitenden Minimums oder Medians (Liu et al., 2010), der Anpassung eines *LOcally WEighted Scatterplot Smoothing* (LOWESS), eines Spline oder einer Exponentialfunktion an die gleitenden Minima bzw. Mediane (Tibshirani et al., 2004; Williams et al., 2005; Li, 2005; Liu et al., 2009; He et al., 2011; House et al., 2011), über morphologische Filter wie z.B. den *TopHat* (Sauve and Speed, 2004), iterative Verfahren wie z.B. *Statistics-*

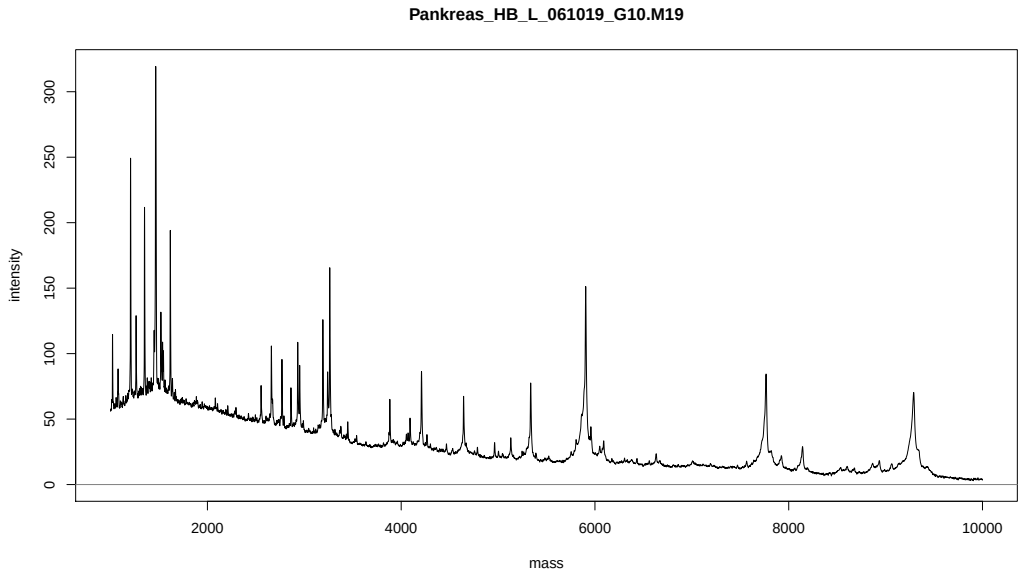


Abbildung 3.3: Beispiel für ein varianzstabilisiertes und geglättetes MALDI/TOF MS-Spektrum aus Fiedler et al. (2009).

sensitive Non-linear Iterative Peak-clipping (SNIP) (Ryan et al., 1988) bis hin zur Konstruktion einer konvexen Hülle (Liu et al., 2003).

Da nach wie vor kein geeignetes Verfahren existiert, um die Qualität einer *Baseline Correction* zu messen, bleibt nur die visuelle Inspektion (Williams et al., 2005). Idealerweise sollte die *Baseline Correction* die Peaks nicht beeinflussen. Dazu darf der jeweilige Algorithmus weder die Höhe noch die Form der Peaks verändern. Aus diesem Grund bietet `MALDIquant` außer dem gleitenden Median (siehe Abb. 3.4 A), der häufig als Referenz genutzt wird, nur Verfahren, die garantiert keine negativen Intensitäten verursachen und die Peaks schonen (je nach Einstellung der Parameter; siehe auch Gibb and Strimmer (2011)), namentlich SNIP (Implementierung nach Ryan et al. (1988); Morhác (2009)), *TopHat* (Implementierung nach van Herk (1992); Gil and Kimmel (2002)) und die konvexe Hülle (Implementierung nach Andrew (1979)).

Die konvexe Hülle ist aufgrund der nach oben konkaven Charakteristik der Matrixeffekte häufig nicht gut geeignet (siehe Abb. 3.4 B bei ca. 1500 Da).

Bei dem *TopHat*-Algorithmus handelt es sich um ein gleitendes Minimum (auch Erosion genannt), von dem im nachfolgenden Schritt ein gleitendes Maximum (auch Dilation genannt) berechnet wird (siehe Abb. 3.4 C).

Der SNIP-Algorithmus (Ryan et al., 1988), dargestellt durch die Gleichung

$$y_{i,k}(k) = \min \left\{ y_{i,k-1}, \frac{y_{i-k,k-1} + y_{i+k,k-1}}{2} \right\}, \tag{3.1}$$

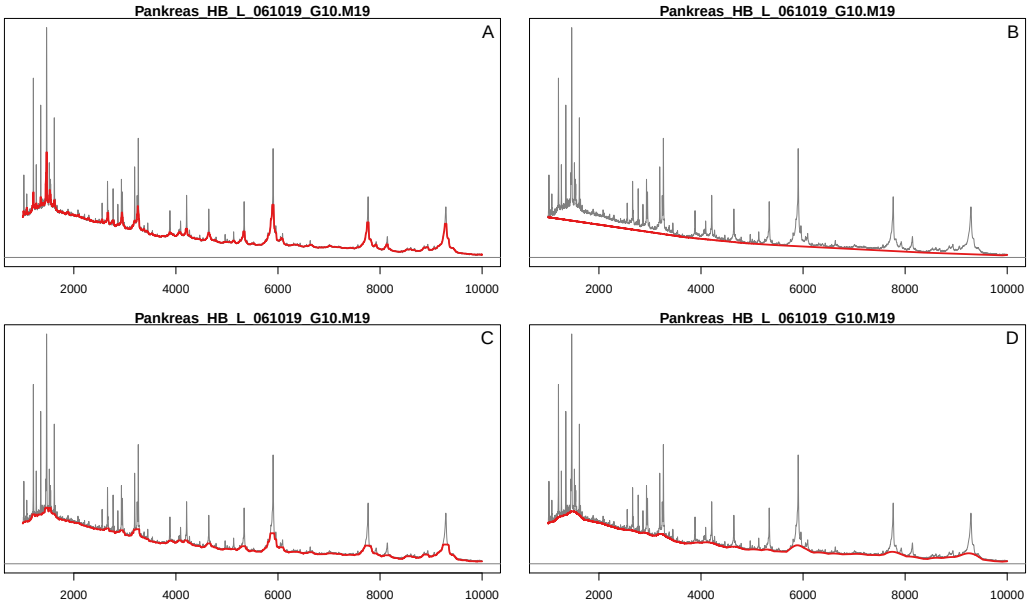


Abbildung 3.4: Beispiel für Grundlinien eines MALDI/TOF MS-Spektrum aus Fiedler et al. (2009) mittels gleitendem Median (A), konvexer Hülle (B), *TopHat* (C), SNIP (D).

mit $i = 1 \ldots n$ und $k = 1 \ldots w$, ersetzt jede Intensität (y) durch den Mittelwert ihrer k-ten Nachbarn, falls dieser kleiner ist. k wird bis zur gewünschten maximalen Fenstergröße (w), respektive Iterationsanzahl, erhöht und die Prozedur wiederholt. Als Resultat erhält man eine glatte und garantiert positive *Baseline* (siehe Abb. 3.4 D und Anhang C, Abschnitt 3.5).

3.5 Kalibrierung der Intensitäten

Die Intensitäten in einem MALDI/TOF MS-Spektrum repräsentieren die relativen Mengen eines beobachteten Proteins/Peptides. Jedoch ist diese Menge nicht proportional zur eigentlichen Konzentration und stark abhängig von präanalytischen Faktoren und Umgebungsfaktoren wie Probenentnahme, Probenaufbewahrung, Raumtemperatur, Luftfeuchtigkeit, Kristallisierung u.ä. (Baggerly et al., 2004; Leichtle et al., 2013). Ein weiteres Problem stellen sog. Batcheffekte dar. Dabei handelt es sich um systematische Veränderungen, die unter Umständen die wahre Aussage verschleiern (Hu et al., 2005; Gregori et al., 2012), z.B. durch ein anderes präanalytisches Vorgehen, Messungen an verschiedenen Tagen mit unterschiedlichen raumklimatischen Bedingungen, verschiedenen Experimentatoren, unterschiedlichen Geräten etc. (siehe Abb. 3.7 A). Diese Fehler sind durch bedachtes Vorgehen und gut geplante, standardisierte Studien zu minimieren.

Diese im weitesten Sinne als Batcheffekte bezeichneten Fehler verursachen sowohl Verschiebungen der m/z (x-Achse) als auch Variationen innerhalb der

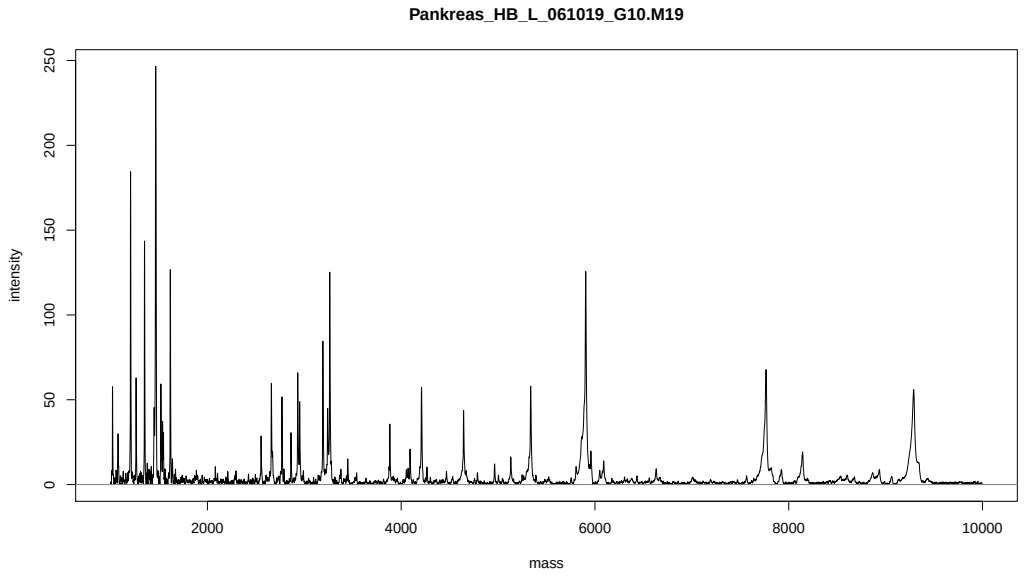


Abbildung 3.5: Beispiel für ein MALDI/TOF MS-Spektrum aus Fiedler et al. (2009) mit korrigierter Grundlinie nach Anwendung des SNIP-Algorithmus (vgl. auch Abb. 3.4D).

Intensitäten (y-Achse). Die Aussage und die Klassifizierung der MS-Daten hängt entscheidend von der Vergleichbarkeit der *m/z*-Werte als auch der Intensitäten ab. Somit stellt die Korrektur der Abweichungen der beiden Dimensionen den wichtigsten Schritt in einer Analyse dar.

Die Kalibrierung der *m/z*-Werte bezeichnet man als *Warping* bzw. *Peak Alignment* (siehe dazu Abschnitt 3.7). Während das *Warping* recht gut gelöst scheint, ist die Kalibrierung der Intensitäten (auch Normalisierung) nach wie vor sehr schwierig.

Man unterscheidet bei den Verfahren zu Intensitätenkalibrierung lokale und globale Verfahren (Meuleman et al., 2008).

Bei den lokalen Verfahren handelt es sich um einfache Rechenoperationen, die auf die einzelnen Spektren separat angewendet werden. In der Regel werden hier bestimmte Kennwerte der Spektren angeglichen, z.B. der *Total Ion Current* (TIC), der Median, der Mittelwert etc. (Callister et al., 2006; Meuleman et al., 2008; Borgaonkar et al., 2010).

Als globale Verfahren bezeichnet man Methoden, die spektrenübergreifend angewendet werden, wie z.B. Normalisierung mit Linearer Regression (Callister et al., 2006), Normalisierung der Quantile (Bolstad et al., 2003) oder die *Probabilistic Quotient Normalization* (PQN) (Dieterle et al., 2006). Diese setzen in der Regel bereits *m/z*-korrigierte Spektren voraus (siehe dazu Abschnitt 3.7).

In `MALDIquant` sind zwei lokale, TIC (siehe auch Anhang C, Abschnitt 3.6) und Median, und eine globale Kalibrierung, PQN, implementiert. Bei der PQN wird zuerst der TIC aller Spektren angeglichen und ein medianes

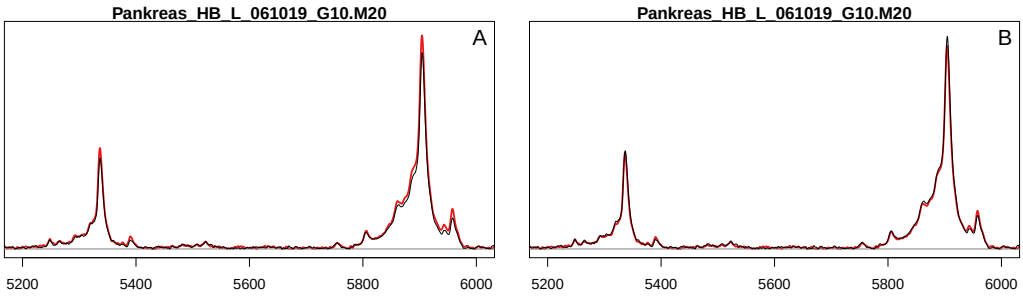


Abbildung 3.6: Beispiel für MALDI/TOF MS-Spektren (aus Fiedler et al. (2009)) technischer Replikate vor (A) und nach (B) Kalibrierung mittels TIC.

Spektrum als Referenzspektrum gebildet. Alle Spektren werden nun durch dieses Referenzspektrum dividiert. Aus den erhaltenen Quotienten wird der Median als spektrenspezifischer Skalierungsfaktor gewählt.

Wie bereits oben erwähnt, ist die Kalibrierung der Intensitäten schwierig und häufig nicht ausreichend um alle Batcheffekte zu neutralisieren. Dennoch ist erwiesen, dass die Anwendung einer Kalibrierung in der Regel bessere Ergebnisse erzielt als deren Verzicht (Meuleman et al., 2008). Ebenso hat sich gezeigt, dass die am häufigsten verwendete Methode, der TIC, trotz seiner Einfachheit recht robust ist (Shin and Markey, 2006; Meuleman et al., 2008). Er ist gut geeignet, um Ungleichheiten zwischen technischen Replikaten, die im selben Messvorgang generiert worden sind, auszugleichen (siehe Abb. 3.6). Jedoch reicht er zur Korrektur von Batcheffekten nicht aus. Das Problem wird deutlich, wenn man Spektren von unterschiedlichen Individuen oder Proben aus unterschiedlichen Kliniken vergleichen will, selbst wenn diese im selben Labor gemessen wurden (siehe Abb. 3.7).

Meiner Meinung nach existiert zur Zeit keine Methode, die Batcheffekte in MALDI/TOF MS-Daten im Nachhinein zufriedenstellend ausgleichen könnte. Umso wichtiger ist es die (prä-)analytischen Faktoren so konstant wie möglich zu halten (Baggerly et al., 2004).

3.6 Identifizierung von Merkmalen

Bei der *Peak Detection* handelt es sich um ein Verfahren zur Selektion lokaler Maxima. Dabei wird die Komplexität der Daten reduziert, was zum einen die Interpretation und zum anderen die weitere Verarbeitung der Spektren stark vereinfacht.

`MALDIquant` nutzt das am häufigsten eingesetzte Verfahren zur *Peak Detection* (Yasui et al., 2003a; Tibshirani et al., 2004; Li, 2005; Morris et al., 2005; Smith et al., 2006; Tracy et al., 2008). Zuerst werden in einem über das Spektrum gleitenden Fenster lokale Maxima gesucht. Diese Maxima werden als potentielle Peaks mit dem Grundrauschen, welches mittels der *Median*

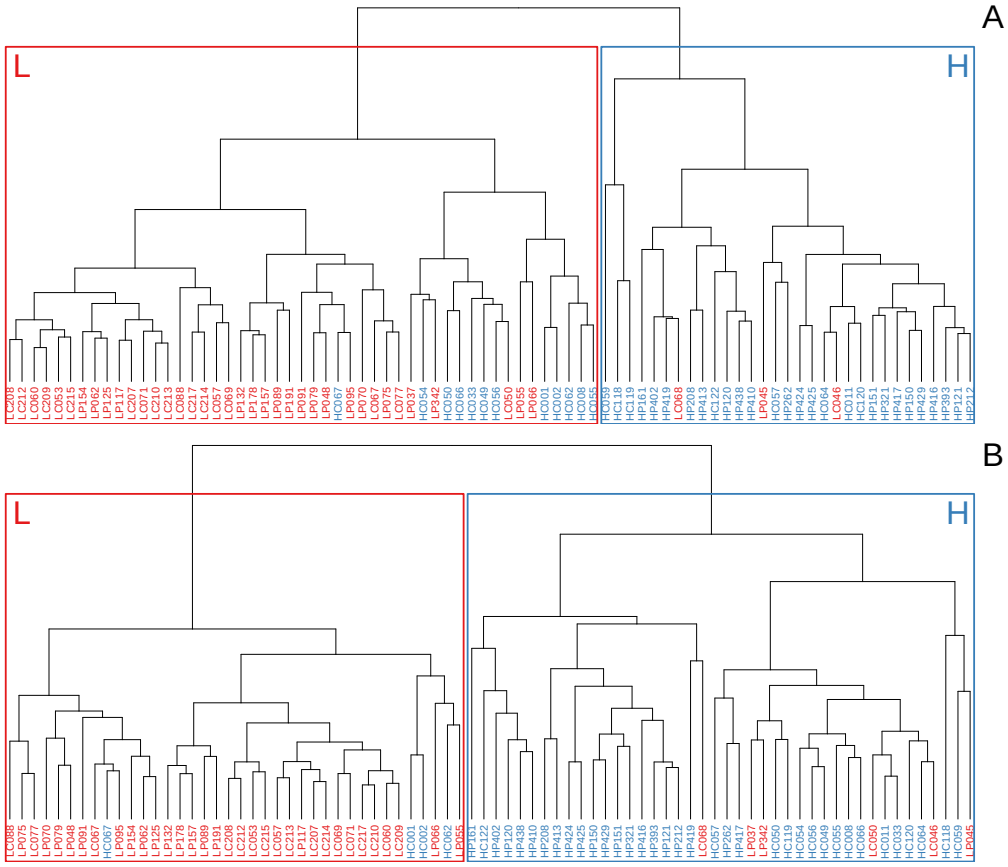


Abbildung 3.7: Hierarchische Clusteranalyse (*complete linkage* auf Grundlage einer euklidischen Distanzmatrix) von MALDI/TOF MS-Spektren aus Fiedler et al. (2009), deren Serumproben aus den Universitätskliniken Leipzig (L, rot) und Heidelberg (H, blau) stammen. Abbildung A zeigt die Clusteranalyse vor und Abbildung B nach der TIC-Kalibrierung. Deutlich erkennbar ist die fehlende Neutralisierung der Batcheffekte in B. Daraus resultiert eine unerwünschte Trennung nach Klinik und nicht nach Krankheitsstatus.

Absolute Deviation (MAD) bzw. Friedman's SuperSmoother (Friedman, 1984) geschätzt wird, verglichen. Wenn ein lokales Maximum ein gewisses *Signal-to-Noise-Ratio* (SNR) überschreitet, ist es per definitionem ein Peak. Lokale Maxima unter dem SNR werden verworfen (siehe Abb. 3.8 und Anhang C, Abschnitt 3.8).
Weitere populäre *Peak Detection*-Verfahren basieren auf *Wavelets* (Lange et al., 2006; Du et al., 2006) und sind bereits in dem *R*-Paket `MassSpecWavelet` (Du et al., 2006) implementiert.

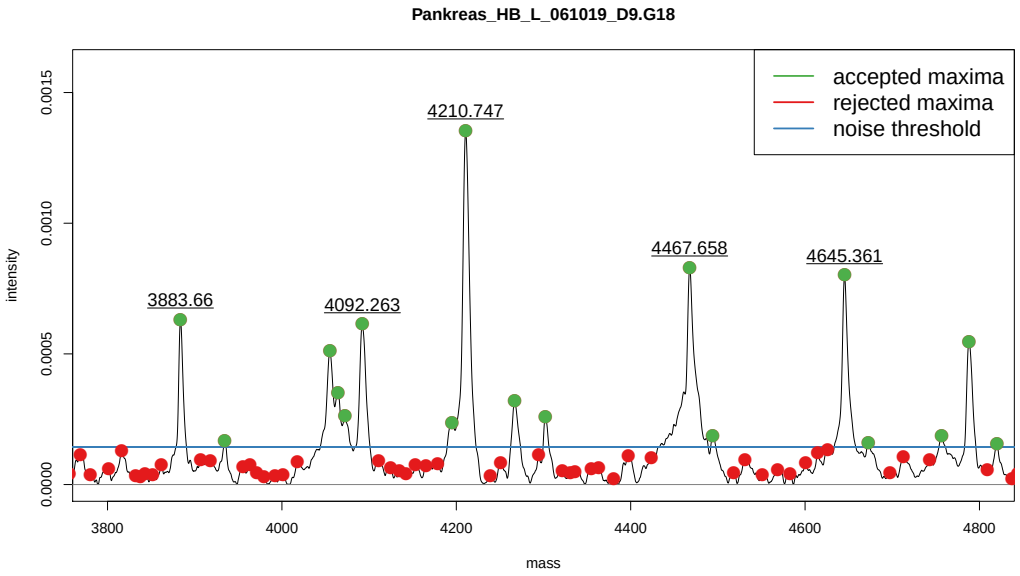


Abbildung 3.8: Ausschnitt eines MALDI/TOF MS-Spektrum (aus Fiedler et al. (2009)). Lokale Maxima sind mit Punkten markiert (rot: verworfene Maxima, grün: Peaks). Die blaue horizontale Linie zeigt das geschätzte Rauschlevel an (MAD).

3.7 Kalibrierung der *m/z*-Werte

Wie bereits unter 3.5 erwähnt, variieren nicht nur die Intensitäten (y-Achse) sondern auch die *m/z*-Werte (x-Achse) der Spektren untereinander (vgl. auch Abb. 3.10 A und D). Leider folgen diese Schwankungen keiner lineare Funktion (He et al., 2011). Die Ursachen ähneln denen der Abweichungen der Intensitäten. Hinzu kommt die fehlende oder wechselnde Kalibrierung der Instrumente durch verschiedene Experimentatoren bzw. in verschiedenen Laboren u.ä. (Baggerly et al., 2004; Smith et al., 2013).

Um die *m/z*-Werte der Spektren anzugleichen bzw. zu korrigieren, ist eine Rekalibrierung aller Spektren nötig, die als *Alignment* bezeichnet wird. Mittlerweile haben sich viele gute Methoden etabliert. Zu nennen wäre hier unter anderem das *Dynamic Time Warping* (DTW), das auf Dynamischer Programmierung beruht (Torgrip et al., 2003; Toppoo et al., 2008; Clifford et al., 2009; Kim et al., 2011). Bei diesem Ansatz werden die Spektren paarweise, jeder Datenpunkt gegen jeden, verglichen und mittels einer Distanzfunktion ein Punktwert vergeben. Danach wird die optimale Zuordnung der Punkte über die Minimierung/Maximierung der Summe der Punktwerte errechnet und die Spektren entsprechend gegeneinander verschoben. Der Vorteil liegt darin, dass dieser Algorithmus immer die optimale Lösung findet. Nachteilig ist allerdings, dass die Entwicklung eines Ähnlichkeitsbaumes notwendig wird, der die Reihenfolge der Spektren festlegt, die nacheinander paarweise kalibriert werden sollen. Oder es wird ein Referenzspektrum benötigt, ge-

gen das alle Spektren ausgerichtet werden. Außerdem ist der Vergleich aller Datenpunkte gegeneinander ineffizient, da viele unnötige Kombinationen berechnet werden, selbst nach Optimierung mit dem Sakoe-Chiba-Band o.ä. (Sakoe and Chiba, 1978). Des Weiteren benötigt das DTW viel Zeit und Arbeitsspeicher (so würden z.B. zwei Spektren mit je 50.000 Datenpunkten zu je 48 Bytes ohne Optimierung ca. 110 Gigabyte Arbeitspeicher belegen; bei Nutzung der Peak Informationen deutlich weniger).

Ein weiterer Ansatz ist das *Correlation Optimized Warping* (COW) (Veselkov et al., 2009; Morris et al., 2010; Wang et al., 2010). Beim COW werden die Spektren ebenfalls paarweise verglichen und so lange entlang der x-Achse (*m/z*-Werte) verschoben bis sie maximal korrelieren. Diese Methode ist viel schneller und viel weniger speicherintensiv als DTW. Jedoch braucht man wie beim DTW einen Ähnlichkeitsbaum oder ein Referenzspektrum. Zusätzlich lassen sich die nicht-linearen Schwankungen nur ausgleichen, wenn man das Spektrum zerstückelt und auf Höhe der Grundlinie Punkte entfernt oder hinzufügt.

Mittels *Parametic Time Warping* (PTW) (Jeffries, 2005; Lin et al., 2005; Bloemberg et al., 2010; He et al., 2011) wird versucht eine polynomiale Funktion zu finden, die zwei Spektren so streckt oder staucht, dass sie ähnlicher werden. Diese Methode ist schnell und erzeugt oder entfernt keine Datenpunkte. Genau wie beim DTW und COW ist eine bestimmte Reihenfolge oder Referenz notwendig.

Alle bisher berichteten Methoden hängen u.a. von einer korrekten Kalibrierung der Intensitäten ab (Smith et al. (2013), siehe auch Abschnitt 3.5).

Die letzte große Gruppe der *Alignment*-Verfahren basiert auf verschiedenen Clustering-Methoden bzw. der Erzeugung von sog. *Bins*, also *m/z*-Bereichen, in denen Peaks als identisch betrachtet werden (Yasui et al., 2003b; Tibshirani et al., 2004; Tracy et al., 2008). Sie sind sehr einfach zu implementieren und bieten die Möglichkeit, im Gegensatz zu DTW, COW und PTW, alle Spektren auf einmal auszurichten. Methodenbedingt können jedoch nur kleine Schwankungen der Peaks um ihre reale Position ausgeglichen werden.

In `MALDIquant` wird ein zweiteiliger Ansatz verfolgt. Zuerst werden die *m/z*-Positionen der Peaks mit PTW angeglichen und danach in *Bins* zusammengefasst. Im Gegensatz zu den meisten PTW-basierten Methoden nutzt `MALDIquant` keinen spektrum- sondern einen peak-basierten Ansatz (He et al., 2011).

Dazu sucht `MALDIquant` eine Funktion (basierend auf LOWESS oder eine polynomialen Funktion), die die Peakpositionen eines Spektrums an die Referenz angleicht (siehe Abb. 3.9 und Anhang C, Abschnitt 3.7). Diese Referenz wird aus stabilen Peaks, die in den meisten Spektren vorkommen, gebildet (Wang et al., 2010). Dazu werden die größten Peaks in bestimmten, gröberen *m/z*-Bereichen herausgesucht und deren *m/z*-Position gemittelt.
Durch die Beschränkung auf Peaks ist die Methode aufgrund der geringeren

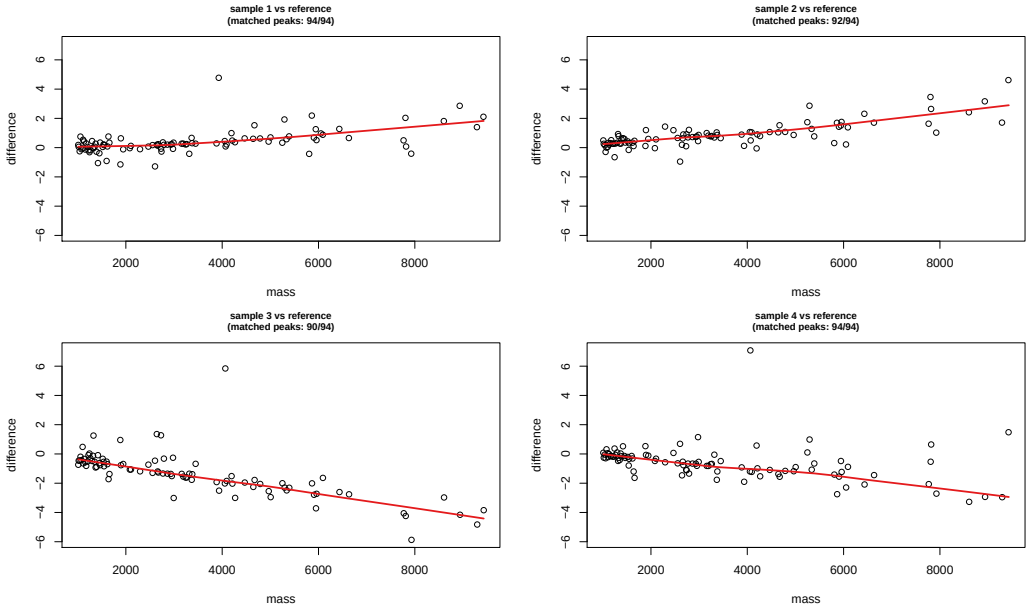


Abbildung 3.9: Warping-Funktionen für vier verschiedene Peaklisten. Dargestellt sind jeweils die Abweichungen der *m/z*-Positionen von der Referenzpeakliste (y-Achse) für die jeweilige Peakpositionen (x-Achse). Die rote Linie zeigt die ermittelte Warping-Funktion (Daten aus Fiedler et al. (2009)).

Datenmenge viel schneller als spektrenbasierte PTW-Varianten bei ähnlich guten Ergebnissen (siehe Abb. 3.10). Außerdem reduziert sie, im Vergleich zu dem spektrenbasierten PTW (als auch zu DTW und COW) die Abhängigkeit der *Warping*-Resultate von der Kalibrierung der Intensitäten.
Nach dem *Warping* sind die Peakpositionen ähnlich, jedoch numerisch nicht identisch (siehe Abb. 3.11 A und B). Deshalb erfolgt im zweiten Schritt die Gruppierung in *Bins* und die Zuordnung einer gemeinsamen *m/z*-Position (siehe auch Anhang C, Abschnitt 3.9). Dazu sortiert `MALDIquant` alle *m/z*-Werte und spaltet diese Liste solange an der größten Lücke zwischen zwei benachbarten *m/z*-Werten bis keine zwei Peaks aus einem Spektrum im selben *Bin* liegen und die Abweichung der *m/z*-Positionen von ihrem Mittelwert eine vom Nutzer erstellte Grenze unterschreitet. Schließlich erhalten alle Peaks in einem *Bin* die gleiche, gemittelte *m/z*-Position (siehe Abb. 3.11 C und D).

3.8 Nachbearbeitung

Nachdem die Peaks nun in *Bins* gruppiert sind, folgt die Umwandlung in eine sog. *Feature Matrix*. Bevor diese Matrix erstellt wird, können zur Erhöhung der Spezifität bei Bedarf noch seltene Peaks herausgefiltert und technische Replikate gemittelt werden. Zur Erhöhung der Sensitivität ist es ebenso möglich, Peaks, die in einigen Spektren nicht erkannt worden sind,

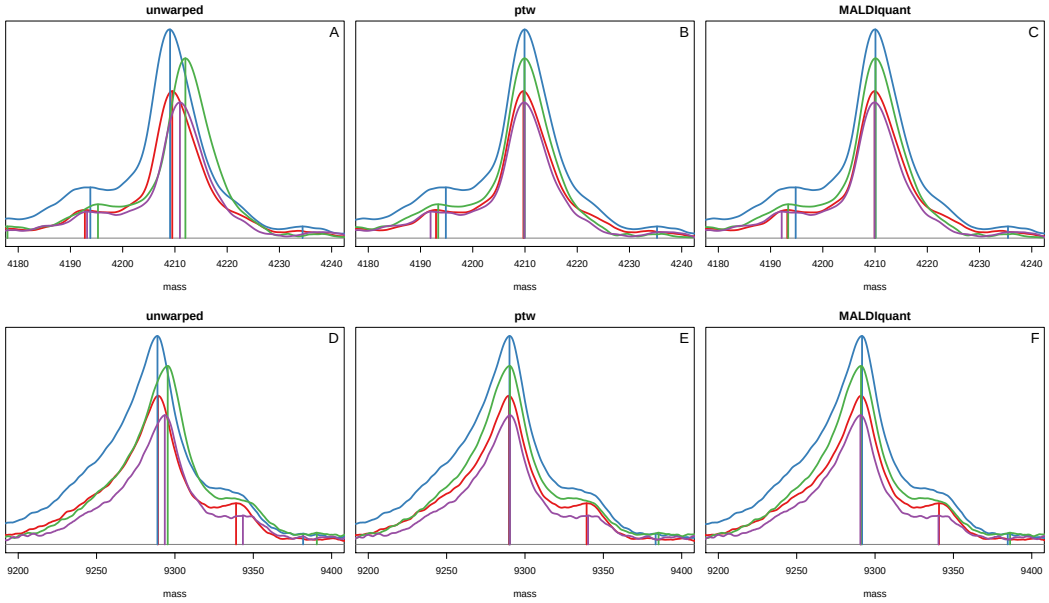


Abbildung 3.10: Vergleich zweier Peakpositionen in vier MALDI/TOF MS-Spektren (aus Fiedler et al. (2009)). (A, D) unkorrigiert; (B, E) mit dem *R*-Paket `ptw` (Bloemberg et al., 2010) korrigiert; (C, F) mit `MALDIquant` korrigiert.

nachträglich aus diesen Spektren in die *Feature Matrix* zu integrieren (siehe auch Anhang C, Abschnitt 3.9). Diese *Feature Matrix* kann nun mit beliebigen statistischen Methoden weiter analysiert werden (siehe dazu Abschnitt 4.3).

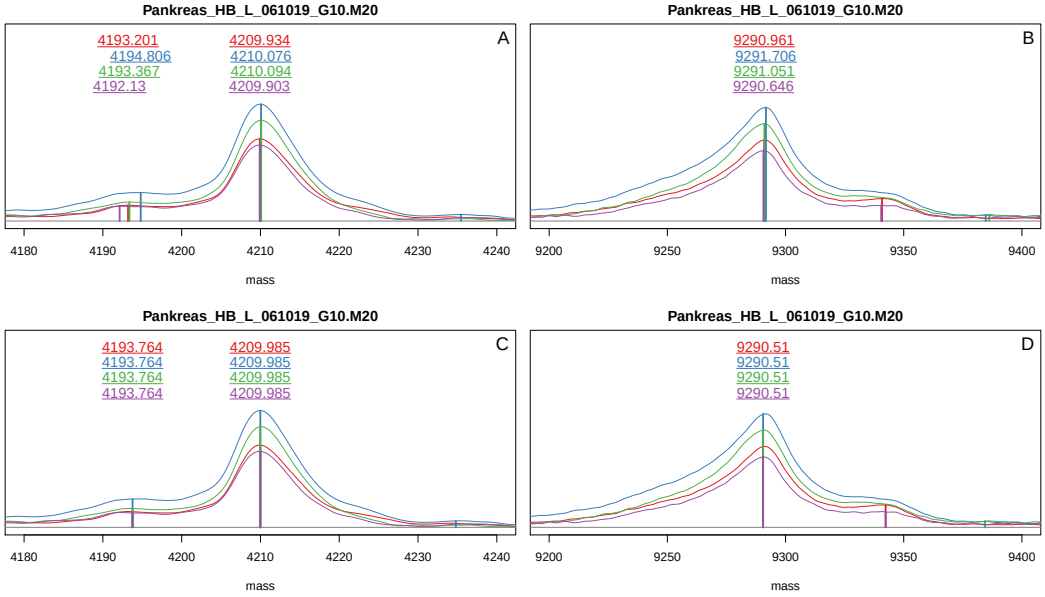


Abbildung 3.11: Peakpositionen aus 3.10 vor (A, B) und nach (C, D) der Gruppierung in *Bins*.

4 Ergebnisse

4.1 Implementierung

`MALDIquant` und `MALDIquantForeign` sind OSS im Sinne der GPL Version 3. Beide *R*-Pakete sind auf CRAN unter `http://cran.r-project.org/web/packages/MALDIquant/` bzw. `http://cran.r-project.org/web/packages/MALDIquantForeign/` zu finden. Zusammen mit `readBrukerFlexData` und `readMzXmlData` umfassen sie mehr als 6000 Zeilen *R*-Code (siehe auch Tabellen B.1-B.4). Zeitkritische Komponenten wie die *Baseline Correction* und die *Peak Detection* in `MALDIquant` sind in *C* geschrieben. Außerdem sind alle Pakete umfangreich dokumentiert und enthalten jeweils viele Demonstrationsskripte sowie eine Einsteiger-Vignette. Weiterführende Literatur und Beispiele zur Nutzung von `MALDIquant` sind unter `http://strimmerlab.org/software/maldiquant/` aufgeführt.

4.2 Anwendungsbeispiel Fiedler et al. 2009

Zur Veranschaulichung der Analyse eines *profiling*-MS-Experiments mit `MALDIquant` verwenden wir die Daten aus Fiedler et al. (2009). Im Sinne reproduzierbarer Forschung ist die komplette Analyse im Anhang C als kommentiertes *R*-Skript zu finden.

In der in Fiedler et al. (2009) beschriebenen Studie wurde ein Trainingsdatensatz mit 320 MALDI/TOF MS-Spektren aus Serumproben von jeweils 40 Patienten mit Pankreaskarzinom und 40 gesunden Probanden zu je vier technischen Replikaten erstellt. Die eine Hälfte der Studienteilnehmer wurde am Universitätsklinikum Heidelberg, die andere im Universitätsklinikum Leipzig rekrutiert. Aufgrund der starken Batcheffekte zwischen den beiden Kliniken (vgl. Abb. 3.7) beschränken wir uns in der nachfolgende Analyse auf den Datensatz aus Heidelberg.

Das Ziel der Studie war es, anhand der MS-Daten Patienten, die am Pankreaskarzinom erkrankt waren, von der Kontrollgruppe zu unterscheiden. So wurde in Fiedler et al. (2009) ein Peak, *m/z* 3884, als diskriminierender Marker gefunden. Dieser Peak ist das doppelt geladene Äquivalent zu *m/z* 7767 und wurde als Platelet Factor 4 (PF4) identifiziert. Er soll im Falle eines Patienten mit Pankreaskarzinom erniedrigt sein. Für weitere Details siehe Fiedler et al. (2009).

4.3 Vorbehandlung der Daten aus Fiedler et al. 2009 mit `MALDIquant`

In der Verarbeitung der MS-Daten folgen wir dem im Kapitel 3 beschriebenen Arbeitsablauf (siehe auch Abb. 3.1). Nach dem Import der Rohdaten durch `MALDIquantForeign` verwerfen wir aufgrund der o.g. Batcheffekte die Leipziger und verwenden für die weitere Analyse ausschließlich die Heidelberger Daten (Anhang C, Abschnitt 3.2). Anschließend überprüfen wir, ob die Spektren die grundlegenden Voraussetzungen für eine Analyse mit `MALDIquant` erfüllen (Anhang C, Abschnitt 3.3). Dazu sollten die Spektren weder leer sein, noch unregelmäßige Abstände zwischen den einzelnen *m/z*-Werten aufweisen. Außerdem möchten wir, dass die Spektren den gleichen *m/z*-Bereich abdecken. Dem weiteren Ablauf nach Abb. 3.1 folgend, führen wir dann die Wurzeltransformation zur Varianzstabilisierung durch und glätten die Spektren mit Hilfe des *Savitzky-Golay*-Filters (Savitzky and Golay, 1964) (Anhang C, Abschnitt 3.4). Mit dem SNIP Algorithmus (Ryan et al., 1988) korrigieren wir die Grundlinie bzw. entfernen den matrix-bedingten Hintergrund (Anhang C, Abschnitt 3.5). Im nächsten Schritt kalibrieren wir die Intensitäten der Spektren indem wir ihren TIC angleichen (Anhang C, Abschnitt 3.6).

Wie in Abschnitt 3.7 beschrieben, sind in `MALDIquant` die Peaks die Grundvoraussetzung für die Kalibrierung der *m/z*-Achse. `MALDIquant` bietet mit `alignSpectra` eine einfache Funktion, die die *Peak Detection* und das *Warping* vereint (Anhang C, Abschnitt 3.7). Die kalibrierten technischen Replikate fassen wir danach zu Mittelwertspektren zusammen. Jede Patientenprobe wird nun von einem einzigen Spektrum repräsentiert (Anhang C, Abschnitt 3.7). Die Mittelwertspektren werden anschließend mittels *Peak Detection* auf Peaks reduziert (Anhang C, Abschnitt 3.8). Dann gruppieren wir die Peaks in *Bins* und entfernen alle, die nicht in mindestens 50 % aller Spektren einer Gruppe vorkommen (Anhang C, Abschnitt 3.9). Zum Schluss wandeln wir die verbliebenen Peaks in eine *Feature Matrix* um (Anhang C, Abschnitt 3.9). Falls ein Peak in einem Spektrum nicht erkannt wurde, hinterlässt er in der *Feature Matrix* eine Lücke. Diese füllen wir mit den Intensitäten an den entsprechenden *m/z*-Werten aus dem jeweiligen Spektrum wieder auf. Die resultierende *Feature Matrix* enthält 166 Peaks je Spektrum.

4.4 Multivariate Analyse

Nachfolgend führen wir eine multivariate Analyse der *Feature Matrix* mittels Diagonaler Diskriminanzanalyse (DDA) durch. Bei der Diagonale Diskriminanzanalyse (DDA) handelt es sich um ein Klassifizierungsverfahren, das auch eine Rangfolge der Variablen, in unserem Falle der *m/z*-Werte, erstellt. Die Variablen werden entsprechend ihrer *t*-Statistik nach ihrem Diskriminie-

pos	mass	score	t.cancer	t.control
1	8936.97	90.69	9.52	-9.52
2	4468.07	80.80	8.99	-8.99
3	8868.27	80.06	8.95	-8.95
4	4494.80	67.00	8.19	-8.19
5	8989.20	66.19	8.14	-8.14
6	5864.49	37.56	-6.13	6.13
7	5906.17	34.43	-5.87	5.87
8	2022.94	33.30	5.77	-5.77
9	5945.57	32.66	-5.71	5.71
10	1866.17	32.12	5.67	-5.67
...	...	...	...	...
145	3884.03	0.39	-0.62	0.62
...	...	...	...	...
164	7768.08	0.00	-0.05	0.05
165	2660.93	0.00	0.02	-0.02
166	1741.19	0.00	0.01	-0.01

Tabelle 4.1: Rangliste der *m/z* aus Fiedler et al. (2009): Die Rangliste der Heidelberger *m/z*-Werte (Fiedler et al., 2009) basiert auf den Ergebnissen der DDA und wurde mit dem *R*-Paket `sda` erstellt (Ahdesmäki and Strimmer (2010)). Peaks mit einem positiven *t.cancer*-Wert weisen in Patientenproben eine höhere und in Kontrollen eine niedrigere Intensität auf. Ein negativer *t.cancer*-Wert bedeutet, dass der entsprechende Peak in den Patientenproben eine geringere Intensität als in den Kontrollen zeigt.

rungspotential geordnet. Wir verwenden zur Durchführung der DDA das *R*-Paket `sda` (Ahdesmäki and Strimmer, 2010) und erhalten als Resultat die Rangliste in Tabelle 4.1 (siehe auch Anhang C, Abschnitt 3.10).

Es ist bemerkenswert, dass wir unter den besten fünf Peaks zwei mögliche Pärchen aus einfach und doppelt geladenen Peaks finden: *m/z* 8936.97 und 4468.07 sowie *m/z* 8989.20 und 4494.80. Auffällig ist auch, dass die besten sieben Peaks in drei *m/z*-Bereichen eng beieinander liegen (vgl. Abb. 4.1). Jedoch lassen sie sich aufgrund der geringen Auflösung nicht scharf voneinander abgrenzen (insbesondere um *m/z* 8936, vgl. Abb. 4.1 C).

Während die komplette *Feature Matrix* in der hierarchischen Clusteranalyse keine klare Trennung der beiden Gruppen erlaubt (siehe Abb. 4.2 A und Anhang C, Abschnitt 3.11), erreichen wir mit den Peaks *m/z* 4468 und 8936 eine nahezu perfekte Diskriminierung von Kontroll- und Patientengruppe (siehe Abb. 4.2 B und Anhang C, Abschnitt 3.11). In der 10-fachen Kreuzvalidierung ergibt sich für diese beiden Peaks eine Sensitivität von 90 % und eine Spezifität von 100 % (siehe auch Anhang C, Abschnitt 3.12; ermittelt mit dem *R*-Paket `crossval` (Strimmer, 2014)).

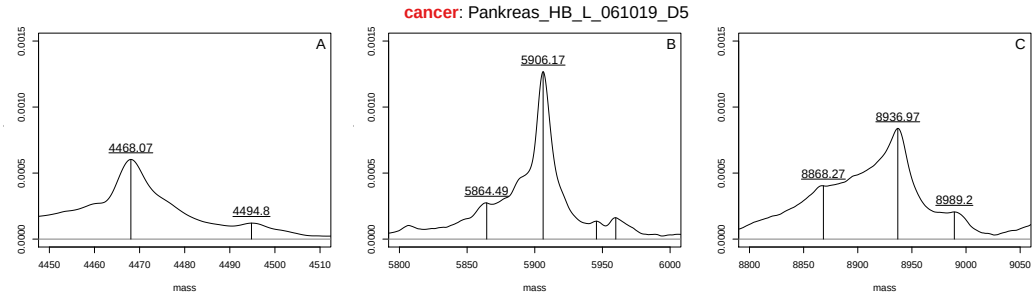


Abbildung 4.1: Detailübersicht der sieben Peaks mit dem größten Diskriminierungspotential (Gesamtansicht in Abb. 4.3 B).

4.5 Mögliche Biomarker

Eine Abfrage in der UniProtKB/Swiss-Prot-Datenbank (Magrane and Consortium, 2011) mit dem TagIdent-Tool (Gasteiger et al. (2005); Mw: 8936.97; Mw range: 0.05 %; Organism: *homo sapiens*) ergibt, dass der Peak *m/z* 8936.97 neben einigen anderen Kandidaten das *pancreatic progenitor cell differentiation and proliferation factor-like protein* (PDPFL_HUMAN) bzw. ein Fragment des *Complement C3* (*Acylation stimulating protein*, auch *C3adesArg*, CO3_HUMAN) repräsentieren könnte. Letzteres ist ein relativ unspezifisches, von der Leber produziertes, Akut-Phase-Protein, das auch schon in vielen anderen MS-basierten *profiling* Studien in Serumproben von Krebspatienten in erhöhter Konzentration gefunden wurde (Li et al., 2005; Lee et al., 2006; Miguet et al., 2006; Ward et al., 2006). Die genaue Identifikation des zugrundeliegenden Moleküls sowie die Klärung seiner Bedeutung für das Pankreaskarzinom liegt allerdings außerhalb des Fokus dieser Dissertation.

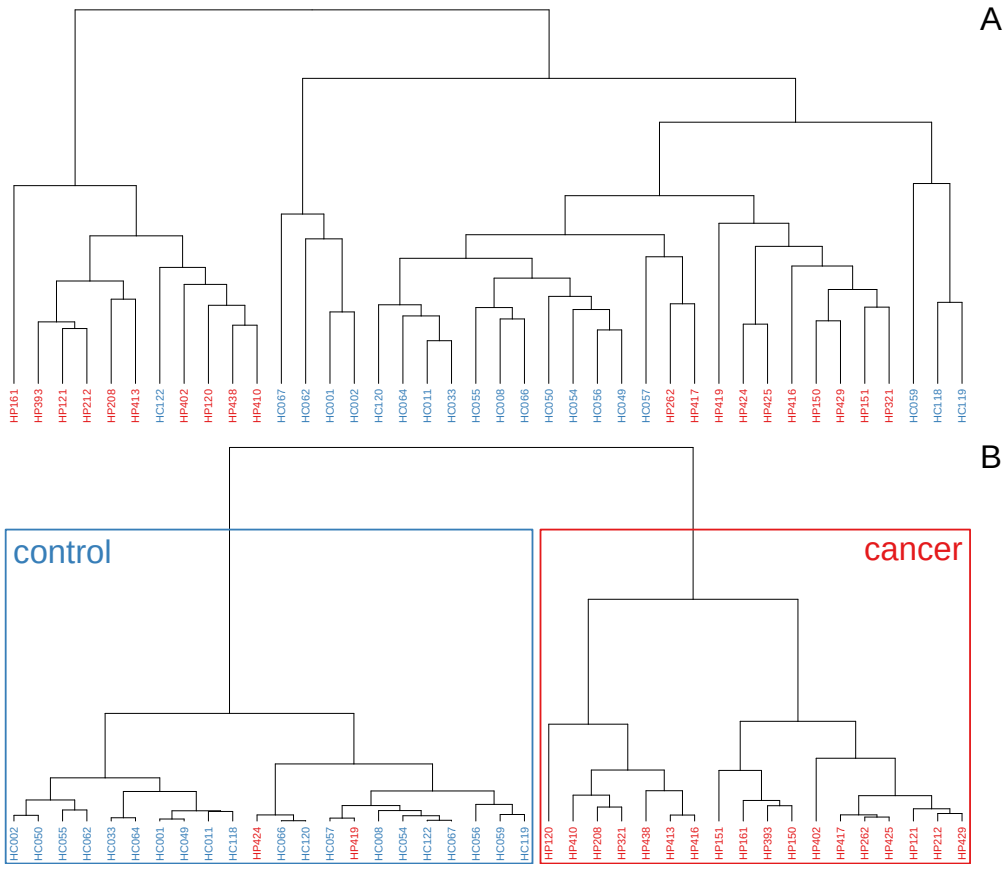


Abbildung 4.2: Hierarchische Clusteranalyse (*complete linkage* auf Grundlage einer euklidischen Distanzmatrix) von MALDI/TOF MS-Spektren aus Fiedler et al. (2009). Der obere Baum (A) ist mit allen Features erstellt und der untere (B) nur mit den, durch `sda` (Ahdesmäki and Strimmer, 2010) ausgewählten, am stärksten diskriminierenden Peaks *m/z* 4468 und 8936.

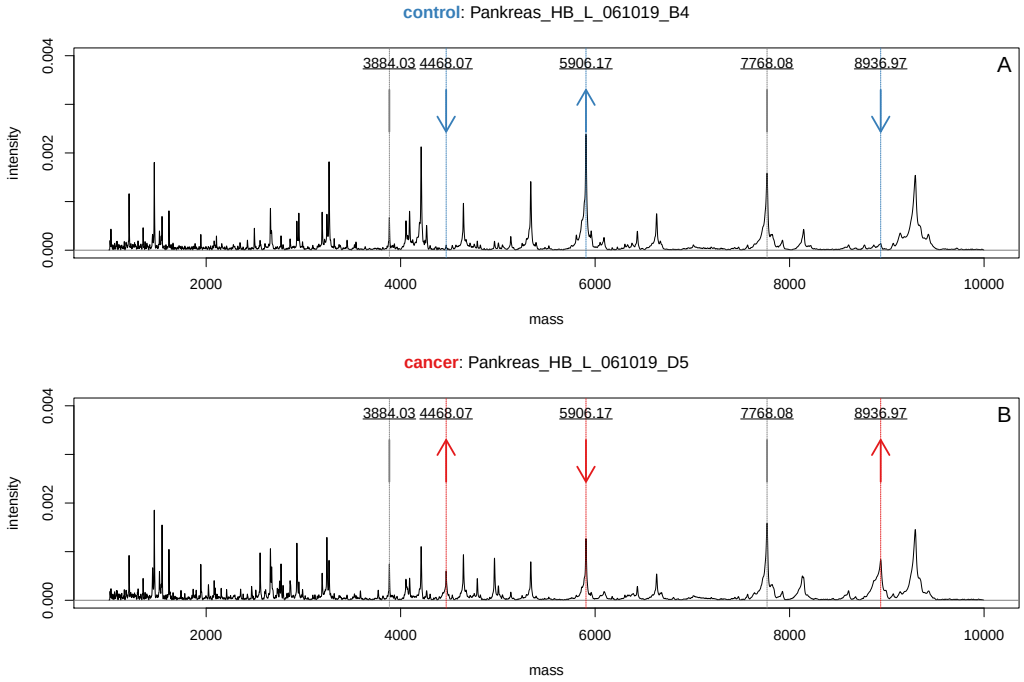


Abbildung 4.3: Vergleich zweier aus technischen Replikaten gemittelter Spektren aus Fiedler et al. (2009). Oben (A) ist ein Spektrum aus der Kontrollgruppe und unten (B) eines aus der Patientengruppe dargestellt. Markiert sind die beiden Peaks mit dem besten Diskriminierungspotential (*m/z* 8936 und 4468; niedrige Intensität in den Kontrollen, hohe Intensität in Patientenspektren), der Peak *m/z* 5906.17 (als Beispiel für hohe Intensität in den Kontrollen und niedrige Intensität in Patientenspektren) sowie die beiden Peaks aus Fiedler et al. (2009) (*m/z* 3884 und 7768; ohne nennenswerte Veränderung in der Intensität).

5 Diskussion

In der klinischen Diagnostik ist die Suche und Entdeckung neuer Biomarker in sog. MS-Profil-Vergleichsstudien von besonderem Interesse. Trotz einer Vielzahl an publizierten Markern (Petricoin et al., 2002; Yanagisawa et al., 2003; Schwegler et al., 2005; de Noo et al., 2006; Fiedler et al., 2009; Zhu et al., 2013) stellt es sich nach wie vor als Herausforderung dar, diese in die klinische Routine zu überführen (Diamandis, 2010). Ein häufiges Problem ist die Reproduzierbarkeit der Studienergebnisse. Die Ursachen dafür sind vielfältig und u.a. in schlechter Studienplanung, fehlender Standardisierung der Präanalytik, der Analyse und der Auswertung sowie in intransparenter Software zu suchen (Diamandis, 2010; Leichtle et al., 2013).

Das Ziel dieser Arbeit war die Entwicklung von `MALDIquant`, einer freien und offenen Software, die moderne und anerkannte Algorithmen bietet und reproduzierbare Analysen von MS-Daten ermöglicht. Im Abschnitt 4.3 haben wir zur Veranschaulichung eine beispielhafte Analyse eines *profiling*-Experiments mit den Daten aus Fiedler et al. (2009) durchgeführt und das komplette *R*-Skript im Anhang C beigelegt.

In unserer Analyse erreichen die Peaks *m/z* 8936 und 4468 das höchste Diskriminierungspotential. Möglicherweise repräsentieren diese Peaks ein unspezifisches Akut-Phase-Protein, ein Fragment des *Complement C3*, das bereits in mehreren MS-basierten Studien bei verschiedenen Krebsentitäten beschrieben wurde (Li et al., 2005; Lee et al., 2006; Miguet et al., 2006; Ward et al., 2006).

Beim Vergleich unserer Ergebnisse mit denen aus Fiedler et al. (2009) stellen wir fest, dass einige unserer am besten trennenden Peaks, wie *m/z* 4468 und 5906, auch in der Originalanalyse gefunden, aber wieder verworfen wurden. Die dort als signifikant beschriebenen Peaks *m/z* 3384 bzw. 7767 (kalibrierungsbedingt hier 7768) ordnen sich auf die Plätze 145 bzw. 164 ein und spielen somit für die Trennung der Kontroll- von der Patientengruppe in unserer Analyse keine Rolle (vgl. Abb. 4.3).

Auch Poruk et al. (2010) haben beobachtet, dass der PF4 evtl. nicht zur Unterscheidung von gesunden Probanden und Patienten mit Pankreaskarzinom geeignet ist, sondern vielmehr mit deren Überlebenszeit sowie mit dem Risiko tiefer Venenthrombosen assoziert sein könnte.

Unsere Analyse basiert aufgrund der starken Batcheffekte (vgl. Abb. 3.7) ausschließlich auf den Daten aus Heidelberg. Diese systematischen Fehler stellen ein sehr komplexes statistisches Problem dar, zu dessen Lösung noch kein Konsensus besteht und weitere Forschung notwendig ist.

Auch abseits von klinischen Studien spielt die MS in der Medizin eine immer größere Rolle. Die Suche nach erblichen Stoffwechselerkrankungen mittels Tandem-MS ist im Neugeborenenscreening seit Jahren etabliert (Chace et al., 1999; Rashed, 2001). Gleiches gilt für die Identifikation von Bakterien in der Klinischen Mikrobiologie, die dank MALDI/TOF MS effizienter, kostengünstiger und schneller durchgeführt werden kann als mit herkömmlichen Methoden (Carbonnelle et al., 2011; Croxatto et al., 2012; DeMarco and Ford, 2013). Um die Zeit bis zur korrekten Antibiotikatherapie weiter zu verringern, gibt es Ansätze mit MALDI/TOF MS nicht nur die Bakterienspezies sondern auch ihre Antibiotikaresistenzen zu identifizieren (Kostrzewa et al., 2013). Sowohl für die Identifikation (DeMarco and Ford, 2013) als auch für die Suche nach Resistenzen (Jung et al., 2013; Sparbier et al., 2013; Jung et al., 2014) wird `MALDIquant` bereits erfolgreich eingesetzt.

Eine weitere Möglichkeit der Diagnostik bakterieller Infektionen ist die Analyse der Zellen des Immunsystems sowie ihrer Reaktion auf Infektionen mittels MALDI/TOF MS. Wie in Ouedraogo et al. (2013) gezeigt, eignet sich `MALDIquant` auch für dieses Szenario.

Eine neue Methode, die seit einigen Jahren insbesondere die Histopathologie bereichert, ist das MALDI-*Imaging* (Cornett et al., 2007). Dabei werden histologische Bilder mit den molekularen Informationen aus den MS-Daten überlagert und damit möglicherweise die diagnostische Aussagekraft verbessert. Dank der Unterstützung üblicher Dateiformate (vgl. Abschnitt 3.2) durch `MALDIquantForeign` findet `MALDIquant` auch bei dieser Technologie Anwendung (Thomas et al., 2013).

6 Zusammenfassung

Das Proteom fasst die Gesamtheit aller Proteine zusammen, die von einem Genom einer Zelle oder eines Gewebes exprimiert werden (Wilkins et al., 1996). Das Besondere am Proteom ist seine hohe Dynamik und Abhängigkeit von Umgebungsfaktoren. Somit eignet es sich, die Reaktion eines biologischen Systems auf seine Umwelt abzuschätzen (Banks et al., 2000). Mittlerweile hat sich mit der Proteomik ein eigener Wissenschaftszweig zur Erforschung des Proteoms gebildet (Patterson and Aebersold, 2003). Eines der wichtigsten Analyseverfahren der Proteomik ist die Massenspektrometrie (MS). Mittels MS ist es möglich, Ionen zu erzeugen und diese nach ihren *mass-to-charge ratios* (*m/z*) aufzutrennen und Rückschlüsse auf ihre Art und Anzahl zu ziehen (Gross, 2004).

Im Alltag der Klinischen Labormedizin, der Klinischen Mikrobiologie wie auch in der Pathologie ist die MS bereits etabliert und ein essentieller Bestandteil der Diagnostik geworden. Trotz starker Verbreitung und technischen Fortschritts der Geräte stellt es sich nach wie vor als Herausforderung dar, publizierte krankheitsspezifische Marker (sog. Biomarker) in die klinische Routine zu überführen (Diamandis, 2010). Ein häufiges Problem ist die Reproduzierbarkeit der Ergebnisse. Die Ursachen sind vielfältig und u.a. in fehlender Standardisierung der Präanalytik, der Analyse und der Auswertung zu suchen (Aebersold and Mann, 2003; Leichtle et al., 2013).

Da es sich bei der MS um ein sog. Hochdurchsatzverfahren handelt, also sehr viele Daten in kurzer Zeit erzeugt werden, sind bioinformatische Programme zur Auswertung notwendig. Diese sind in der Regel proprietär und die Eigenschaften der verwendeten Algorithmen sind nicht bekannt. Ein Weg zur Verbesserung der Reproduzierbarkeit ist deshalb die Entwicklung und Nutzung transparenter Programme und Datenformate (Aebersold and Mann, 2003).

Ziel dieser Arbeit war die Entwicklung von `MALDIquant` (Gibb and Strimmer, 2012), einer freien und offenen Software, die eine einfache Automatisierung und verbesserte Reproduzierbarkeit der Analyse von MALDI/TOF und anderen zweidimensionalen MS-Daten ermöglicht. Dank des begleitenden Paketes `MALDIquantForeign` ist `MALDIquant` weitestgehend herstellerund geräteunabhängig. Sowohl `MALDIquant` als auch `MALDIquantForeign` sind freie und offene Software für die Programmiersprache *R* (R Core Team, 2014) und stehen unter der *GNU General Public License* (GPL).

In der Abfolge einer typischen Auswertung von MS-Daten werden die Funktionen und Algorithmen von `MALDIquant` bzw. `MALDIquantForeign`

erklärt. Nach einer kurzen Beschreibung von `MALDIquantForeign` und seinen Funktionen zum Einlesen der MS-Daten, folgt die Erläuterung der Bedeutung der Varianzstabilisierung sowie des Glättens von Spektren und deren Umsetzung in `MALDIquant`. Der nächste wichtige Schritt ist die Minimierung von Effekten der Matrix und chemischer Verunreinigung mittels Grundlinienkorrektur (*Baseline Correction*). Hier bietet `MALDIquant`, neben dem weit verbreiteten gleitenden Median, Algorithmen wie z.B. den SNIP-Algorithmus (Ryan et al., 1988), die die Form der lokalen Maxima (Peaks) schonen und garantiert keine negativen Intensitäten erzeugen. Um die Intensitäten und *m/z*-Werte von Spektren zu vergleichen, müssen beide kalibriert werden. In `MALDIquant` sind zur Kalibrierung der Intensitäten (oft auch als Normalisierung bezeichnet) der *Total Ion Current* (TIC), der Median und die *Probabilistic Quotient Normalization* (PQN) (Dieterle et al., 2006) implementiert. Da die Intensitätenkalibrierung in Gegenwart von Batcheffekten nicht immer zufriedenstellende Ergebnisse liefert, ist es besonders wichtig, dass die Kalibrierung der *m/z*-Werte möglichst unabhängig von dieser ist. `MALDIquant` konzentriert sich deswegen ausschließlich auf die *m/z*-Werte der Peaks, wird durch die Intensitätenkalibrierung kaum beeinflusst und ist schneller als die meisten anderen spektrenbasierten Methoden mit vergleichbar guten Resultaten.

Die *Peak Detection* dient der Selektion und Reduktion der Daten für die weitere Analyse. `MALDIquant` nutzt eines der am häufigsten eingesetzten Verfahren zur *Peak Detection* und betrachtet ein lokales Maximum als Peak, sobald es ein definiertes Verhältnis zum Hintergrundrauschen überschritten hat (*Signal-to-Noise-Ratio*).

Im letzten Schritt werden die Peaks in einer sog. *Feature Matrix* zusammengefasst. Diese kann nachfolgend mit beliebigen statistischen Methoden analysiert werden.

`MALDIquant` ist sehr flexibel und einfach an eigene Bedürfnisse anzupassen. So stellt der soeben skizzierte Arbeitsablauf nur einen Vorschlag dar, denn andere Vorgehensweisen lassen sich mit `MALDIquant` ebenfalls realisieren.

Wir demonstrieren die Verwendung von `MALDIquant`, indem wir die Analyse der Daten aus Fiedler et al. (2009) nachvollziehen. In dieser Studie wurde versucht, Patienten mit Pankreaskarzinom von der Kontrollgruppe anhand der MS-Profile aus Blutseren zu unterscheiden.

In unserer Analyse mit `MALDIquant` und `sda` (Ahdesmäki and Strimmer, 2010) finden wir unter anderem die Peaks *m/z* 4468 und 8936, mit denen eine nahezu perfekte Trennung der beiden Gruppen möglich ist. Dabei könnte es sich um ein Fragment des *Complement C3* handeln, welches bereits in anderen Studien in Serumproben von Krebspatienten beschrieben worden ist (Li et al., 2005; Lee et al., 2006; Miguet et al., 2006; Ward et al., 2006).

Dass `MALDIquant` sehr flexibel ist und neben dem eben beschriebenen Szenario weitere Anwendungsgebiete erschließt, zeigt seine breite Verwendung. So wurde `MALDIquant` bereits zur Analyse der Aktivierung von Immunzellen

(Ouedraogo et al., 2013) und beim MALDI-*Imaging* eingesetzt (Thomas et al., 2013). Auch in der Klinischen Mikrobiologie hat `MALDIquant` seine Nützlichkeit bei der Identifikation von Bakterienspezies (DeMarco and Ford, 2013) sowie der Analyse von Antibiotikaresistenzen (Jung et al., 2013; Sparbier et al., 2013; Jung et al., 2014) bereits bewiesen.

7 Literaturverzeichnis

Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422:198–207.

Ahdesmäki, M. and Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *The Annals of Applied Statistics*, 4(1):503–519.

Andrew, M. A. (1979). Another Efficient Algorithm for Convex Hulls in Two Dimensions. In *Information Processing Letters 9*, pages 216–219. Elsevier.

Baggerly, K. A., Morris, J. S., and Coombes, K. R. (2004). Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, 20:777–785.

Banks, R. E., Dunn, M. J., Hochstrasser, D. F., Sanchez, J. C., Blackstock, W., Pappin, D. J., and Selby, P. J. (2000). Proteomics: new perspectives, new biomedical opportunities. *Lancet*, 356:1749–1756.

Bernardo, K., Pakulat, N., Macht, M., Krut, O., Seifert, H., Fleer, S., Hünger, F., and Krönke, M. (2002). Identification and discrimination of Staphylococcus aureus strains using matrix-assisted laser desorption/ionization-time of flight mass spectrometry. *PROTEOMICS*, 2:747–753.

Bloemberg, T. G., Gerretzen, J., Wouters, H. J. P., Gloerich, J., van Dael, M., Wessels, H. J. C. T., van den Heuvel, L. P., Eilers, P. H. C., Buydens, L. M. C., and Wehrens, R. (2010). Improved parametric time warping for proteomics. *Chemometrics and Intelligent Laboratory Systems*, 104:65–74.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185–193.

Borgaonkar, S. P., Hocker, H., Shin, H., and Markey, M. K. (2010). Comparison of normalization methods for the identification of biomarkers using MALDI-TOF and SELDI-TOF mass spectra. *OMICS*, 14:115–126.

Callister, S. J., Barry, R. C., Adkins, J. N., Johnson, E. T., Qian, W.-J., Webb-Robertson, B.-J. M., Smith, R. D., and Lipton, M. S. (2006). Normalization Approaches for Removing Systematic Biases Associated with Mass Spectrometry and Label-Free Proteomics. *Journal of Proteome Research*, 5:277–286.

Carbonnelle, E., Mesquita, C., Bille, E., Day, N., Dauphin, B., Beretti, J.-L., Ferroni, A., Gutmann, L., and Nassif, X. (2011). MALDI-TOF mass spectrometry tools for bacterial identification in clinical microbiology laboratory. *Clinical Biochemistry*, 44:104–109.

Chace, D. H., DiPerna, J. C., and Naylor, E. W. (1999). Laboratory integration and utilization of tandem mass spectrometry in neonatal screening: a model for clinical mass spectrometry in the next millennium. *Acta Paediatrica*, 88:45–47.

Clifford, D., Montoliu, G. S. I., Rezzi, S., Martin, F.-P., Guy, P., Bruce, S., and Kochhar, S. (2009). Alignment Using Variable Penalty Dynamic Time Warping. *Analytical Chemistry*, 81:1000–1007.

Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M.-C., and Kuerer, H. M. (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *PROTEOMICS*, 5:4107–4117.

Cornett, D. S., Reyzer, M. L., Chaurand, P., and Caprioli, R. M. (2007). MALDI imaging mass spectrometry: molecular snapshots of biochemical systems. *Nature Methods*, 4:828–833.

Croxatto, A., Prod'hom, G., and Greub, G. (2012). Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS Microbiology Reviews*, 36:380–407.

Danial, A. (2012). *CLOC: Count Lines of Code - version 1.56*.

de Noo, M. E., Mertens, B. J. A., Ozalp, A., Bladergroen, M. R., van der Werff, M. P. J., van de Velde, C. J. H., Deelder, A. M., and Tollenaar, R. A. E. M. (2006). Detection of colorectal cancer using MALDI-TOF serum protein profiling. *European Journal of Cancer*, 42:1068–1076.

DeMarco, M. L. and Ford, B. A. (2013). Beyond identification: emerging and future uses for MALDI-TOF mass spectrometry in the clinical microbiology laboratory. *Clinics in Laboratory Medicine*, 33:611–628.

Diamandis, E. P. (2010). Cancer biomarkers: can we turn recent failures into success? *Journal of the National Cancer Institute*, 102:1462–1467.

Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Analytical Chemistry*, 78:4281–4290.

Du, P., Kibbe, W. A., and Lin, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22:2059–2065.

Du, P., Stolovitzky, G., Horvatovich, P., Bischoff, R., Lim, J., and Suits, F. (2008). A noise model for mass spectrometry based proteomics. *Bioinformatics*, 24:1070–1077.

Feltens, R., Görner, R., Kalkhof, S., Gröger-Arndt, H., and von Bergen, M. (2010). Discrimination of different species from the genus Drosophila by intact protein profiling using matrix-assisted laser desorption ionization mass spectrometry. *BMC Evolutionary Biology*, 10:95.

Fiedler, G. M., Leichtle, A. B., Kase, J., Baumann, S., Ceglarek, U., Felix, K., Conrad, T., Witzigmann, H., Weimann, A., Schütte, C., Hauss, J., Büchler, M., and Thiery, J. (2009). Serum peptidome profiling revealed platelet factor 4 as a potential discriminating peptide associated with pancreatic cancer. *Clinical Cancer Research*, 15:3812–3819.

Friedman, J. H. (1984). A variable span smoother. Technical report, DTIC Document.

Gammerman, A., Nouretdinov, I., Burford, B., Chervonenkis, A., Vovk, V., and Luo, Z. (2008). Clinical mass spectrometry proteomic diagnosis by conformal predictors. *Statistical Applications in Genetics and Molecular Biology*, 7.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., and Bairoch, A. (2005). *The Proteomics Protocols Handbook/Protein Identification and Analysis Tools on the ExPASy Server*. Humana Press.

Gatto, L. and Lilley, K. S. (2012). MSnbase - an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, 28:288–289.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

Gibb, S. (2013a). *readBrukerFlexData: Reads mass spectrometry data in Bruker *flex format*. R package version 1.7.

Gibb, S. (2013b). *readMzXmlData: Reads mass spectrometry data in mzXML format*. R package version 2.7.

Gibb, S. (2014). *MALDIquantForeign: Import/Export routines for MALDIquant*. R package version 0.7.

Gibb, S. and Strimmer, K. (2011). Analysis of proteomic data using MALDIquant. In *Proceedings of the 8th International Workshop on Computational Systems Biology, WCSB 2011 (June 6-8, 2011, Zurich, Switzerland)*, pages 49–52.

Gibb, S. and Strimmer, K. (2012). MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, 28:2270–2271.

Gil, J. Y. and Kimmel, R. (2002). Efficient dilation, erosion, opening, and closing algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24:1606–1617.

Gregori, J., Villarreal, L., Méndez, O., Sánchez, A., Baselga, J., and Villanueva, J. (2012). Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics. *Journal of Proteomics*, 75(13):3938–3951.

Gross, J. H. (2004). *Mass Spectrometry: A Textbook*. Springer.

He, Q. P., Wang, J., Mobley, J. A., Richman, J., and Grizzle, W. E. (2011). Self-calibrated warping for mass spectra alignment. *Cancer Informatics*, 10:65–82.

House, L. L., Clyde, M. A., and Wolpert, R. L. (2011). Bayesian nonparametric models for peak identification in MALDI-TOF mass spectroscopy. *The Annals of Applied Statistics*, 5:1488–1511.

Hu, J., Coombes, K. R., Morris, J. S., and Baggerly, K. A. (2005). The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Briefings in Functional Genomics and Proteomics*, 3:322–331.

Jeffries, N. (2005). Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21:3066–3073.

Jung, J. S., Eberl, T., Sparbier, K., Lange, C., Kostrzewa, M., Schubert, S., and Wieser, A. (2013). Rapid detection of antibiotic resistance based on mass spectrometry and stable isotopes. *European Journal of Clinical Microbiology & Infectious Diseases*.

Jung, J. S., Popp, C., Sparbier, K., Lange, C., Kostrzewa, M., and Schubert, S. (2014). Evaluation of MALDI-TOF MS for rapid detection of β-lactam resistance in enterobacteriaceae derived from blood cultures. *Journal of Clinical Microbiology*.

Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008). ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24:2534–2536.

Kim, S., Koo, I., Fang, A., and Zhang, X. (2011). Smith-Waterman peak alignment for comprehensive two-dimensional gas chromatography-mass spectrometry. *BMC Bioinformatics*, 12:235.

Kostrzewa, M., Sparbier, K., Maier, T., and Schubert, S. (2013). MALDI-TOF MS: an upcoming tool for rapid detection of antibiotic resistance in microorganisms. *Proteomics - Clinical Applications*, 7:767–778.

Lange, E., Gröpl, C., Reinert, K., Kohlbacher, O., and Hildebrandt, A. (2006). High-Accuracy Peak Picking of Proteomics Data Using Wavelet Techniques. In *Pacific Symposium on Biocomputing*, volume 11, pages 243–254.

Lee, I. N., Chen, C.-H., Sheu, J.-C., Lee, H.-S., Huang, G.-T., Chen, D.-S., Yu, C.-Y., Wen, C.-L., Lu, F.-J., and Chow, L.-P. (2006). Identification of complement C3a as a candidate biomarker in human chronic hepatitis C and HCV-related hepatocellular carcinoma using a proteomics approach. *PROTEOMICS*, 6:2865–2873.

Leichtle, A. B., Dufour, J.-F., and Fiedler, G. M. (2013). Potentials and pitfalls of clinical peptidomics and metabolomics. *Swiss Medical Weekly*, 143:w13801.

Li, J., Orlandi, R., White, C. N., Rosenzweig, J., Zhao, J., Seregni, E., Morelli, D., Yu, Y., Meng, X.-Y., Zhang, Z., Davidson, N. E., Fung, E. T., and Chan, D. W. (2005). Independent validation of candidate breast cancer serum biomarkers identified by mass spectrometry. *Clinical Chemistry*, 51:2229–2235.

Li, X. (2005). *PROcess: Ciphergen SELDI-TOF Processing*. R package version 1.40.0.

Lin, S. M., Haney, R. P., Campa, M. J., Fitzgerald, M. C., and Patz, E. F. (2005). Characterising phase variations in MALDI-TOF data and correcting them by peak alignment. *Cancer Informatics*, 1:32–40.

Liu, L. H., Shan, B. E., Tian, Z. Q., Sang, M. X., Ai, J., Zhang, Z. F., Meng, J., Zhu, H., and Wang, S. J. (2010). Potential biomarkers for esophageal carcinoma detected by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Clinical Chemistry and Laboratory Medicine*, pages 855–861.

Liu, Q., Krishnapuram, B., Pratapa, P., Liao, X., Hartemink, A., and Carin, L. (2003). Identification of Differentially Expressed Proteins Using MALDI-TOF Mass Spectra. *Signals, Systems and Computers, 2003. Conference Record*, 2:1323–1327.

Liu, Q., Sung, A. H., Qiao, M., Chen, Z., Yang, J. Y., Yang, M. Q., Huang, X., and Deng, Y. (2009). Comparison of feature selection and classification for MALDI-MS data. *BMC Genomics*, 10(Suppl 1):S3.

Magrane, M. and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database*, 2011:bar009.

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., and Deutsch, E. W. (2011). mzML–a community standard for mass spectrometry data. *Molecular & Cellular Proteomics*, 10:R110.000133.

Meuleman, W., Engwegen, J. Y., Gast, M.-C. W., Beijnen, J. H., Reinders, M. J., and Wessels, L. F. (2008). Comparison of normalisation methods for surface-enhanced laser desorption and ionisation (SELDI) time-of-flight (TOF) mass spectrometry data. *BMC Bioinformatics*, 9:88.

Miguet, L., Bogumil, R., Decloquement, P., Herbrecht, R., Potier, N., Mauvieux, L., and van Dorsselaer, A. (2006). Discovery and Identification of Potential Biomarkers in a Prospective Study of Chronic Lymphoid Malignancies Using SELDI-TOF-MS. *Journal of Proteome Research*, 5:2258–2269.

Morhác, M. (2009). An algorithm for determination of peak regions and baseline elimination in spectroscopic data. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 600:478–487.

Morris, J. S., Baggerly, K. A., Gutstein, H. B., and Coombes, K. R. (2010). Statistical contributions to proteomic research. *Methods in Molecular Biology*, 641:143–166.

Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A., and Kobayashi, R. (2005). Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21:1764–1775.

Norris, J. L., Cornett, D. S., Mobley, J. A., Andersson, M., Seeley, E. H., Chaurand, P., and Caprioli, R. M. (2007). Processing MALDI Mass Spectra to Improve Mass Spectral Direct Tissue Analysis. *International Journal of Mass Spectrometry*, 260:212–221.

Open Source Initiative (2014). Open Source Definition. `http://opensource.org/definition`. Accessed: 2014-01-21.

Ouedraogo, R., Textoris, J., Daumas, A., Capo, C., and Mege, J.-L. (2013). Whole-cell MALDI-TOF mass spectrometry: a tool for immune cell analysis and characterization. *Methods in Molecular Biology*, 1061:197–209.

Patterson, S. D. and Aebersold, R. H. (2003). Proteomics: the first decade and beyond. *Nature Genetics*, 33:311–323.

Pedrioli, P. G. A., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., and Aebersold, R. (2004). A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology*, 22:1459–1466.

Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359:572–577.

Poruk, K. E., Firpo, M. A., Huerter, L. M., Scaife, C. L., Emerson, L. L., Boucher, K. M., Jones, K. A., and Mulvihill, S. J. (2010). Serum platelet factor 4 is an independent predictor of survival and venous thromboembolism in patients with pancreatic adenocarcinoma. *Cancer Epidemiology, Biomarkers & Prevention*, 19:2605–2610.

Purohit, P. V. and Rocke, D. M. (2003). Discriminant models for high-throughput proteomics mass spectrometer data. *PROTEOMICS*, 3:1699–1703.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rashed, M. S. (2001). Clinical applications of tandem mass spectrometry: ten years of diagnosis and screening for inherited metabolic diseases. *Journal of Chromatography B: Biomedical Sciences and Applications*, 758:27–48.

Robb, R. A., Hanson, D. P., Karwoski, R. A., Larson, A. G., Workman, E. L., and Stacy, M. C. (1989). Analyze: a comprehensive, operator-interactive software package for multidimensional medical image display and analysis. *Computerized Medical Imaging and Graphics*, 13:433–454.

Ryan, C. G., Clayton, E., Griffin, W. L., Sie, S. H., and Cousens, D. R. (1988). SNIP, a statistics-sensitive background treatment for the quantitative

analysis of PIXE spectra in geoscience applications. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 34:396–402.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26:43–49.

Sauve, A. C. and Speed, T. P. (2004). Normalization, baseline correction and alignment of high-throughput mass spectrometry data. *Proceedings Gensips*.

Savitzky, A. and Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36:1627–1639.

Schramm, T., Hester, A., Klinkert, I., Both, J.-P., Heeren, R. M. A., Brunelle, A., Laprévote, O., Desbenoit, N., Robbe, M.-F., Stoeckli, M., Spengler, B., and Römpp, A. (2012). imzML–a common data format for the flexible exchange and processing of mass spectrometry imaging data. *Journal of Proteomics*, 75:5106–5110.

Schwegler, E. E., Cazares, L., Steel, L. F., Adam, B.-L., Johnson, D. A., Semmes, O. J., Block, T. M., Marrero, J. A., and Drake, R. R. (2005). SELDI-TOF MS profiling of serum for detection of the progression of chronic hepatitis C to hepatocellular carcinoma. *Hepatology*, 41:634–642.

Shin, H. and Markey, M. K. (2006). A machine learning perspective on the development of clinical decision support systems utilizing mass spectra of blood samples. *Journal of Biomedical Informatics*, 39:227–248.

Sköld, M., Rydén, T., Samuelsson, V., Bratt, C., Ekblad, L., Olsson, H., and Baldetorp, B. (2007). Regression analysis and modelling of data acquisition for SELDI-TOF mass spectrometry. *Bioinformatics*, 23:1401–1409.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78:779–787.

Smith, R., Ventura, D., and Prince, J. T. (2013). LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Briefings in Bioinformatics*.

Sparbier, K., Lange, C., Jung, J., Wieser, A., Schubert, S., and Kostrzewa, M. (2013). MALDI biotyper-based rapid resistance detection by stable-isotope labeling. *Journal of Clinical Microbiology*, 51:3741–3748.

Strimmer, K. (2014). *crossval: Generic Functions for Cross Validation*. R package version 1.0.1.

Strohalm, M., Hassman, M., Košata, B., and Kodíček, M. (2008). mMass data miner: an open source alternative for mass spectrometric data analysis. *Rapid Communications in Mass Spectrometry*, 22:905–908.

Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., and Kohlbacher, O. (2008). OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics*, 9:163.

Thomas, A., Patterson, N. H., Marcinkiewicz, M. M., Lazaris, A., Metrakos, P., and Chaurand, P. (2013). Histology-driven data mining of lipid signatures from multiple imaging mass spectrometry analyses: application to human colorectal cancer liver metastasis biopsies. *Analytical Chemistry*, 85:2860–2866.

Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q.-T. (2004). Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics*, 20:3034–3044.

Toppoo, S., Roveri, A., Vitale, M. P., Zaccarin, M., Serain, E., Apostolidis, E., Gion, M., Maiorino, M., and Ursini, F. (2008). MPA: A multiple peak alignment algorithm to perform multiple comparisons of liquid-phase proteomic profiles. *PROTEOMICS*, 8:250–253.

Torgrip, R. J. O., Åberg, M., Karlberg, B., and Jacobsson, S. P. (2003). Peak alignment using reduced set mapping. *Journal of Chemometrics*, 17:573–582.

Tracy, M. B., Chen, H., Weaver, D. M., Malyarenko, D. I., Sasinowski, M., Cazares, L. H., Drake, R. R., Semmes, O. J., Tracy, E. R., and Cooke, W. E. (2008). Precision enhancement of MALDI-TOF MS using high resolution peak detection and label-free alignment. *PROTEOMICS*, 8:1530–1538.

van Herk, M. (1992). A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels. *Pattern Recognition Letters*, 13:517–521.

Veselkov, K. A., Lindon, J. C., Ebbels, T. M. D., Crockford, D., Volynkin, V. V., Holmes, E., Davies, D. B., and Nicholson, J. K. (2009). Recursive segment-wise peak alignment of biological (1)h NMR spectra for improved metabolic biomarker recovery. *Analytical Chemistry*, 81:56–66.

Wang, B., Fang, A., Heim, J., Bogdanov, B., Pugh, S., Libardoni, M., and Zhang, X. (2010). DISCO: distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. *Analytical Chemistry*, 82:5069–5081.

Ward, D. G., Suggett, N., Cheng, Y., Wei, W., Johnson, H., Billingham, L. J., Ismail, T., Wakelam, M. J. O., Johnson, P. J., and Martin, A. (2006). Identification of serum biomarkers for colon cancer by proteomic analysis. *British Journal of Cancer*, 94:1898–1905.

Wilkins, M. R., Sanchez, J.-C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F., and Williams, K. L. (1996). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnology and genetic engineering reviews*, 13:19–50.

Williams, B., Cornett, S., Dawant, B., Crecelius, A., Bodenheimer, B., and Caprioli, R. (2005). An Algorithm for Baseline Correction of MALDI Mass Spectra. In *Proceedings of the 43rd Annual Southeast Regional Conference - Volume 1*, ACM-SE 43, pages 137–142, New York, NY, USA. ACM.

Wulfkuhle, J. D., Liotta, L. A., and Petricoin, E. F. (2003). Proteomic applications for the early detection of cancer. *Nature Reviews Cancer*, 3:267–275.

Yanagisawa, K., Shyr, Y., Xu, B. J., Massion, P. P., Larsen, P. H., White, B. C., Roberts, J. R., Edgerton, M., Gonzalez, A., Nadaf, S., Moore, J. H., Caprioli, R. M., and Carbone, D. P. (2003). Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet*, 362:433–439.

Yasui, Y., McLerran, D., Adam, B., Winget, M., Thornquist, M., and Feng, Z. (2003a). An automated peak-identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *Journal of Biomedicine and Biotechnology*, 4:242–248.

Yasui, Y., Pepe, M., Thompson, M. L., Adam, B.-L., Wright, G. L., Qu, Y., Potter, J. D., Winget, M., Thornquist, M., and Feng, Z. (2003b). A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, 4:449–463.

Zhu, D., Wang, J., Ren, L., Li, Y., Xu, B., Wei, Y., Zhong, Y., Yu, X., Zhai, S., Xu, J., and Qin, X. (2013). Serum proteomic profiling for the early diagnosis of colorectal cancer. *Journal of Cellular Biochemistry*, 114:448–455.

A Publikation

BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 17 2012, pages 2270–2271
doi:10.1093/bioinformatics/bts447

Genome analysis

Advance Access publication July 12, 2012

MALDIquant: a versatile R package for the analysis of mass spectrometry data

Sebastian Gibb* and Korbinian Strimmer

Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany

Associate Editor: Alex Bateman

ABSTRACT

Summary: `MALDIquant` is an R package providing a complete and modular analysis pipeline for quantitative analysis of mass spectrometry data. `MALDIquant` is specifically designed with application in clinical diagnostics in mind and implements sophisticated routines for importing raw data, preprocessing, non-linear peak alignment and calibration. It also handles technical replicates as well as spectra with unequal resolution.

Availability: `MALDIquant` and its associated R packages `readBrukerFlexData` and `readMzXmlData` are freely available from the R archive CRAN (http://cran.r-project.org). The software is distributed under the GNU General Public License (version 3 or later) and is accompanied by example files and data. Additional documentation is available from http://strimmerlab.org/software/maldiquant/.

Contact: mail@sebastiangibb.de

Received on June 6, 2012; revised on July 5, 2012; accepted on July 9, 2012

1 INTRODUCTION

Mass spectrometry profiling is increasingly becoming an important tool in clinical diagnostics, for example to identify biomarkers for cancer (e.g. Fiedler *et al.*, 2009). Similarly as with other high-throughput technologies, sophisticated statistical algorithms are essential in the analysis of spectrometry data (Morris *et al.*, 2010).

We have developed `MALDIquant` to provide a complete open-source analysis pipeline on the R platform (R Development Core Team, 2012) comprising all steps from importing of raw data, preprocessing (e.g. baseline removal), peak detection and non-linear peak alignment to calibration of mass spectra. `MALDIquant` is written as a standalone package using S4 object-oriented programming to facilitate further extension.

`MALDIquant` was initially developed for clinical proteomics using Matrix-Assisted Laser Desorption/Ionization (MALDI) technology. However, the algorithms implemented in `MALDIquant` are generic and may be equally applied to other 2D mass spectrometry data.

2 DISTINCTIVE FEATURES

In comparison with related R packages for mass spectrometry analysis, `MALDIquant` features a number of unique capabilities. In particular, it implements a sophisticated non-linear peak alignment algorithm (He *et al.*, 2011; Wang *et al.*, 2010) as well as a calibration procedure for normalization of peak intensities across spectra that are modeled on a related method for sequence count data (Anders and Huber, 2010). In addition, `MALDIquant` allows to analyze technical replicates and spectra with unequal resolution, a crucial feature in clinical mass spectrometry where spectra from multiple sources need to be compared.

3 DETAILS ON ALGORITHMS

An example workflow for mass spectrometry analysis using `MALDIquant` is depicted in Fig. 1, starting with a raw unprocessed MALDI spectrum (A), followed by smoothing, baseline correction and peak detection (B), local alignment of peaks across spectra by warping (C–E) and merging and visualization (F). In the following, we briefly provide some background on the respective algorithms.

3.1 Data import

`MALDIquant` is carefully designed to be independent of any specific mass spectrometry hardware. Nonetheless, native input of binary data files (as well as complete folder hierarchies) from Bruker flex series instruments and input of the `mzXML` data format is supported through the associated R packages `readBrukerFlexData` and `readMzXmlData`.

3.2 Data preprocessing

For preprocessing spectral data, `MALDIquant` offers a complete set of routines for smoothing, variance stabilization, baseline correction and peak detection. `MALDIquant` implements several approaches to adjust the baseline and uses per default the SNIP algorithm (Ryan *et al.*, 1988) that returns a smooth baseline and leads to positive corrected intensities (Fig. 1B).

3.3 Peak alignment

For comparison of peaks across different spectra, it is essential to conduct alignment. In order to match peaks belonging to the same mass, `MALDIquant` uses a statistical regression-based approach combining the algorithms of He *et al.* (2011) and Wang *et al.* (2010). Specifically, first landmark peaks are identified that occur in most spectra. Subsequently, a non-linear warping function is computed for each spectrum by fitting a local regression

*To whom correspondence should be addressed.

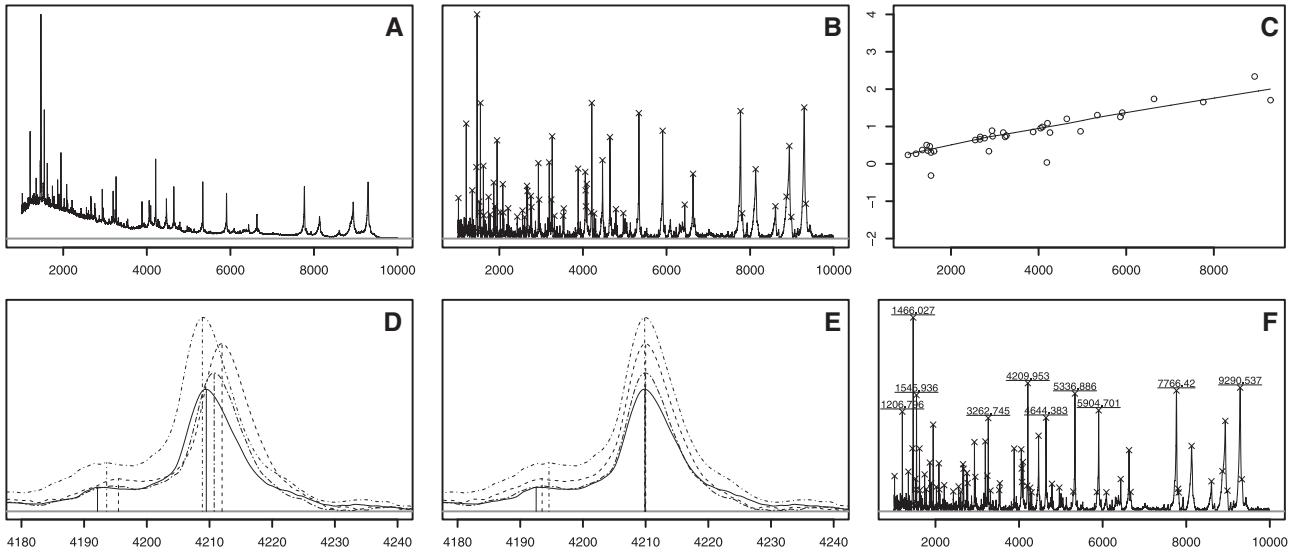


Fig. 1. Example of MALDIquant output: **(A)** raw spectrum; **(B)** variance-stabilized, smoothed and baseline-corrected spectrum with detected peaks; **(C)** fitted warping function for peak alignment; **(D)** four unaligned peaks; **(E)** four aligned peaks; and (**F**) merged spectrum with detected and labeled peaks.

to the matched reference peaks (Fig. 1C–E). This also allows to merge aligned spectra from technical replicates. An example of a merged spectrum with identified and labeled peaks is shown in Fig. 1F.

3.4 Calibration

Quantitative analysis of multiple spectra, e.g. to detect differentially expressed peaks, requires calibration. In order to render peak intensities comparable across spectra, a suitable scale factor for each individual spectrum needs to be determined. Experimentally, quantification of intensities is performed by reference to spike-in samples. In absence of spike-ins, MALDIquant offers a way of calibrating relative intensities by adapting an algorithm for calibrating next generation sequencing data (Anders and Huber, 2010). In this procedure first a reference spectrum is created using the median intensity of aligned peaks from all spectra. Subsequently, a scale factor is computed for each spectrum by using a robust estimator of the overall ratio of the peak intensities of the uncalibrated spectrum versus the reference spectrum. Additionally, calibration based on total ion current is available.

3.5 Classification and feature selection

Finally, the resulting calibrated peak intensity matrix may be exported for further use in high-level statistical analysis, for instance classification and feature selection using shrinkage discriminant analysis (Ahdesmäki and Strimmer, 2010).

4 CONCLUSION

`MALDIquant` is a versatile R package providing a flexible analysis pipeline for MALDI-TOF and other mass spectrometry data. It offers a number of distinctive features, in particular for alignment by non-linear warping and simultaneous calibration of peak intensities.

An overview of its capabilities is given by running the included demo script

```
library(''MALDIquant'')
demo(''MALDIquant'')
```

ACKNOWLEDGMENTS

We thank Alexander Leichtle for many valuable and helpful suggestions and Fiedler *et al.* (2009) for their kind permission to use their data in MALDIquant.

Funding: S.G. received funding from the German National Academic Foundation.

Conflict of Interest: none declared.

REFERENCES

Ahdesmäki,M. and Strimmer,K. (2010) Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Statist.*, **4**, 503–519.

Anders,S. and Huber,W. (2010) Differential expression for sequence count data. *Genome Biol.*, **11**, R106.

Fiedler,G.M. *et al.* (2009) Serum peptidome profiling revealed platelet factor 4 as a potential discriminating peptide associated with pancreatic cancer. *Clin. Cancer Res.*, **15**, 3812–3819.

He,Q.P. *et al.* (2011) Self-calibrated warping for mass spectra alignment. *Cancer Inform.*, **10**, 65–82.

Morris,J.S. *et al.* (2010) Statistical contributions to proteomic research. *Methods Mol. Biol.*, **641**, 143–166.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*.. R Foundation for Statistical Computing Vienna, Austria.

Ryan,C.G. *et al.* (1988) SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nucl. Instrument. Meth. B*, **34**, 396–402.

Wang,B. *et al.* (2010) DISCO: distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. *Anal. Chem.*, **82**, 5069–5081.

B Übersicht Codeumfang

Sprache	Dateien	Leerzeilen	Kommentare	Code
R	123	908	2267	3043
tex	40	306	13	2736
C	7	96	229	313
Gesamt	170	1310	2509	6092

Tabelle B.1: Codeübersicht MALDIquant 1.10: Übersicht erstellt mit *cloc* (Danial, 2012).

Sprache	Dateien	Leerzeilen	Kommentare	Code
R	48	482	1485	1986
tex	24	151	5	1117
Gesamt	72	633	1490	3103

Tabelle B.2: Codeübersicht MALDIquantForeign 0.7: Übersicht erstellt mit *cloc* (Danial, 2012).

Sprache	Dateien	Leerzeilen	Kommentare	Code
tex	15	75	0	710
R	12	108	995	352
Gesamt	27	183	995	1062

Tabelle B.3: Codeübersicht readBrukerFlexData 1.7: Übersicht erstellt mit *cloc* (Danial, 2012).

Sprache	Dateien	Leerzeilen	Kommentare	Code
R	9	135	543	530
tex	14	62	0	466
Gesamt	23	197	543	996

Tabelle B.4: Codeübersicht readMzXmlData 2.7: Übersicht erstellt mit *cloc* (Danial, 2012).

C Analyse Fiedler et al. 2009

Dieser Anhang stellt die komplette Analyse der Heidelberger Spektren aus Fiedler et al. (2009) entsprechend den Abschnitten 4.3 und 4.4 dar.

Auf `https://github.com/sgibb/MALDIquantExamples/` ist die jeweils aktuellste Version zu finden.

Analysis of Fiedler et al. 2009 using MALDIquant

Sebastian Gibb*

February 1, 2015

Abstract

This vignette describes the analysis of the MALDI-TOF spectra described in Fiedler et al. (2009) using MALDIquant

Contents

*mail@sebastiangibb.de

1 Foreword

`MALDIquant` is free and open source software for the `R` (R Core Team, 2014) environment and under active development. If you use it, please support the project by citing it in publications:

> Gibb, S. and Strimmer, K. (2012). MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, 28(17):2270–2271

If you have any questions, bugs, or suggestions do not hesitate to contact me (`mail@sebastiangibb.de`).
Please visit `http://strimmerlab.org/software/maldiquant/`.

2 Dataset

In this vignette we use the dataset described in Fiedler et al. (2009). Please contact the authors directly if you want to use the dataset in your own analysis.

This dataset contains 480 MALDI-TOF mass spectra from blood sera of 60 patients and 60 healthy controls (each sample has four technical replicates).

It is divided in three set:

1. *Discovery Set A*: 20 patients with pancreatic cancer and 20 healthy patients from the University Hospital Leipzig.
2. *Discovery Set B*: 20 patients with pancreatic cancer and 20 healthy patients from the University Hospital Heidelberg.
3. *Discovery Set C*: 20 patients with pancreatic cancer and 20 healthy patients from the University Hospital Leipzig (half resolution).

Both discovery sets *A* and *B* were measured on the same target (batch). The validation set *C* was measured a few months later.

Please see Fiedler et al. (2009) for details.

3 Analysis

3.1 Setup

First we need to install the necessary packages (you can skip this part if you have already done this). You can install `MALDIquant` (Gibb and Strimmer, 2012), `MALDIquantForeign` (Gibb, 2014), `sda` (Ahdesmäki and Strimmer, 2010) and `crossval` (Strimmer, 2014) directly from `CRAN`. To install this data package from `http://github.com/sgibb/MALDIquantExamples` you need the `devtools` (Wickham and Chang, 2014) package.

```
install.packages(c("MALDIquant", "MALDIquantForeign",
                   "sda", "crossval", "devtools"))
library("devtools")
install_github("sgibb/MALDIquantExamples")
```

Next we load the packages.

```
library("MALDIquant")
library("MALDIquantForeign")
library("sda")
library("crossval")

library("MALDIquantExamples")
```

3.2 Import Raw Data

We use the `getPathFiedler2009` function to get the correct file path to the spectra and the metadata file respectively.

```
## import the spectra
spectra <- import(getPathFiedler2009()["spectra"],
                  verbose=FALSE)

## import metadata
spectra.info <- read.table(getPathFiedler2009()["info"],
                           sep=",", header=TRUE)
```

Because of heavy batch effects between the two hospitals we consider only the data collected in the University Hospital Heidelberg.

```
isHeidelberg <- spectra.info$location == "heidelberg"

spectra <- spectra[isHeidelberg]
spectra.info <- spectra.info[isHeidelberg,]
```

We do a basic quality control and test whether all spectra contain the same number of data points and are not empty.

3.3 Quality Control

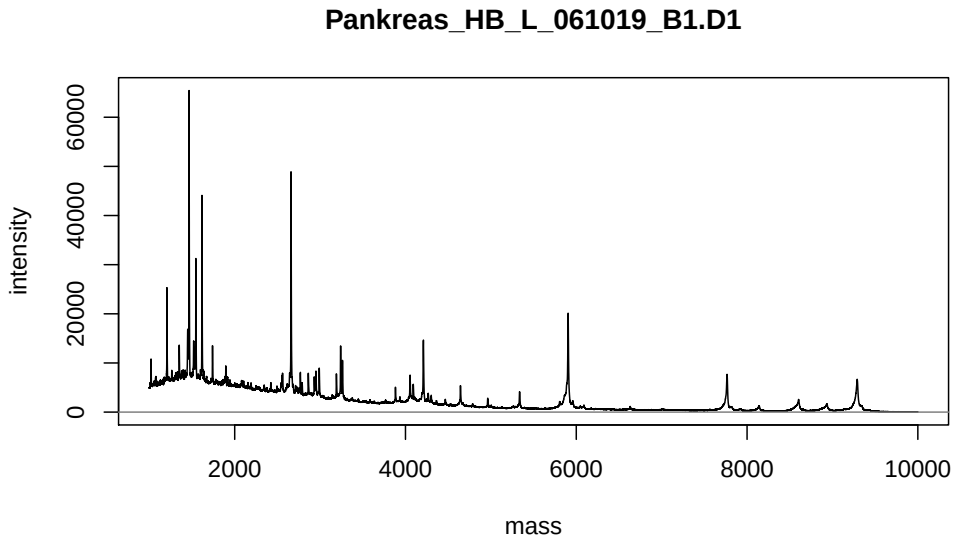
```
table(sapply(spectra, length))


42388
  160

any(sapply(spectra, isEmpty))

[1] FALSE

all(sapply(spectra, isRegular))

[1] TRUE
```

Subsequently we ensure that all spectra have the same mass range.

```
spectra <- trim(spectra)
```

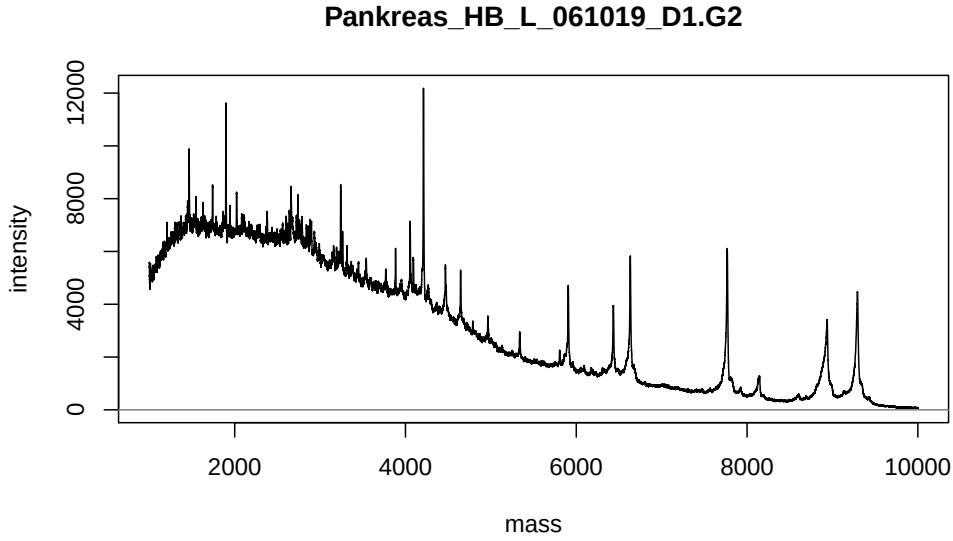
Finally we draw some plots and inspect the spectra visually.

```
idx <- sample(length(spectra), size=2)
plot(spectra[[idx[1]]])
```



LDIquantForeign_uncompress/spectra_6ab6119eb509/fiedler_et_al_2009/set B − discovery heidelberg/control/Pankreas_HB_

```
plot(spectra[[idx[2]]])
```



LDIquantForeign_uncompress/spectra_6ab6119eb509/fiedler_et_al_2009/set B − discovery heidelberg/tumor/Pankreas_HB_L

3.4 Transformation and Smoothing

We apply the square root transformation to simplify graphical visualization and to overcome the potential dependency of the variance from the mean.

```
spectra <- transformIntensity(spectra, method="sqrt")
```

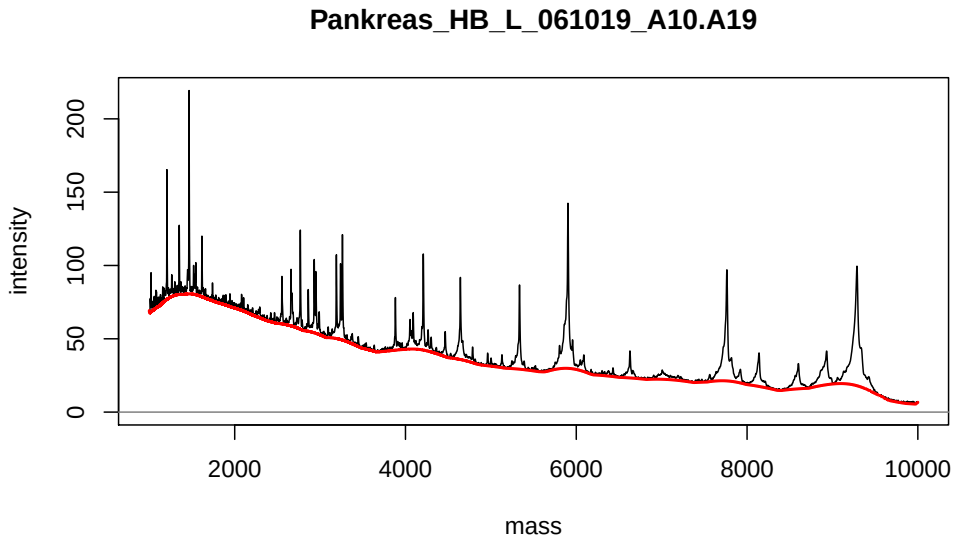
In the next step we use a 21 point *Savitzky-Golay*-Filter (Savitzky and Golay, 1964) to smooth the spectra.

```
spectra <- smoothIntensity(spectra, method="SavitzkyGolay",
                           halfWindowSize=20)
```

3.5 Baseline Correction

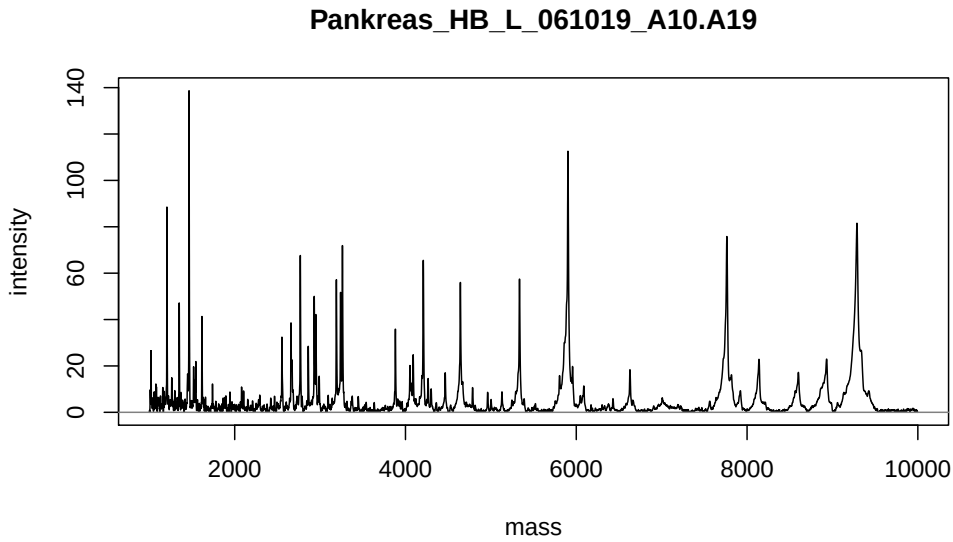
Matrix effects and chemical noise results in some background noise. That's why we have to apply a baseline correction. In this example we use the *SNIP* algorithm (Ryan et al., 1988) to correct the baseline.

```
baseline <- estimateBaseline(spectra[[1]], method="SNIP",
                             iterations=150)
plot(spectra[[1]])
lines(baseline, col="red", lwd=2)
```



.DIquantForeign_uncompress/spectra_6ab6119eb509/fiedler_et_al_2009/set B – discovery heidelberg/control/Pankreas_HB_L.

```
spectra <- removeBaseline(spectra, method="SNIP",
                          iterations=150)
plot(spectra[[1]])
```



.DIquantForeign_uncompress/spectra_6ab6119eb509/fiedler_et_al_2009/set B – discovery heidelberg/control/Pankreas_HB_L.

3.6 Intensity Calibration

We perform the *Total-Ion-Current*-calibration (TIC; often called normalization) to equalize the intensities across spectra.

```
spectra <- calibrateIntensity(spectra, method="TIC")
```

3.7 Alignment

Next we need to (re)calibrate the mass values. Our alignment procedure is a peak based warping algorithm. `MALDIquant` offers `alignSpectra` as a wrapper around more complicated functions. If you need a finer control or want to investigate the impact of different parameters please use `determineWarpingFunctions` instead (see `?determineWarpingFunctions` for details).

```
spectra <- alignSpectra(spectra)
```

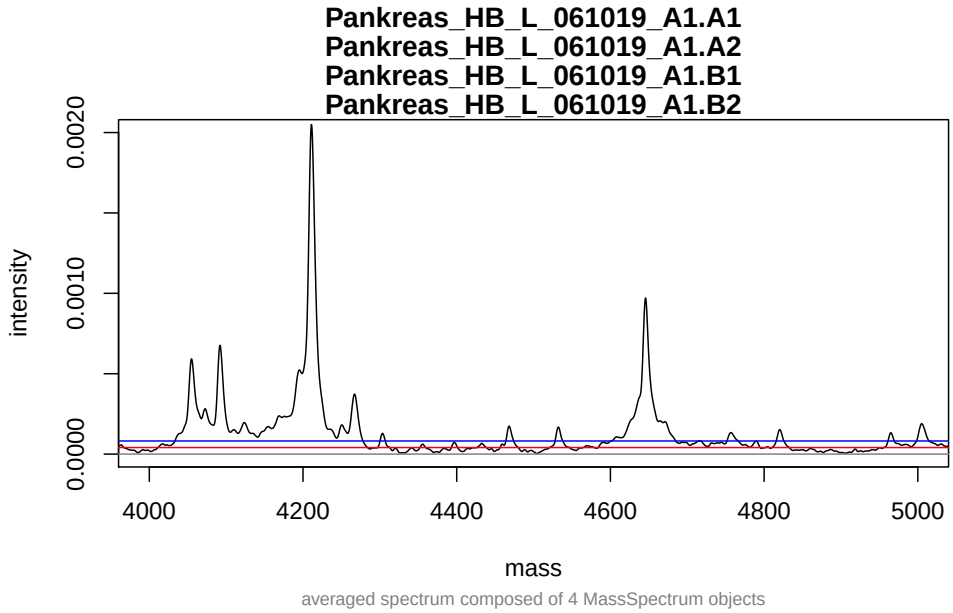
We average the technical replicates before we look for peaks and adjust our metadata table accordingly.

```
avgSpectra <-
  averageMassSpectra(spectra, labels=spectra.info$patientID)
avgSpectra.info <-
  spectra.info[!duplicated(spectra.info$patientID), ]
```

3.8 Peak Detection

The peak detection is the crucial feature reduction step. Before performing the peak detection we estimate the noise of some spectra to get a feeling for the *signal-to-noise ratio* (SNR).

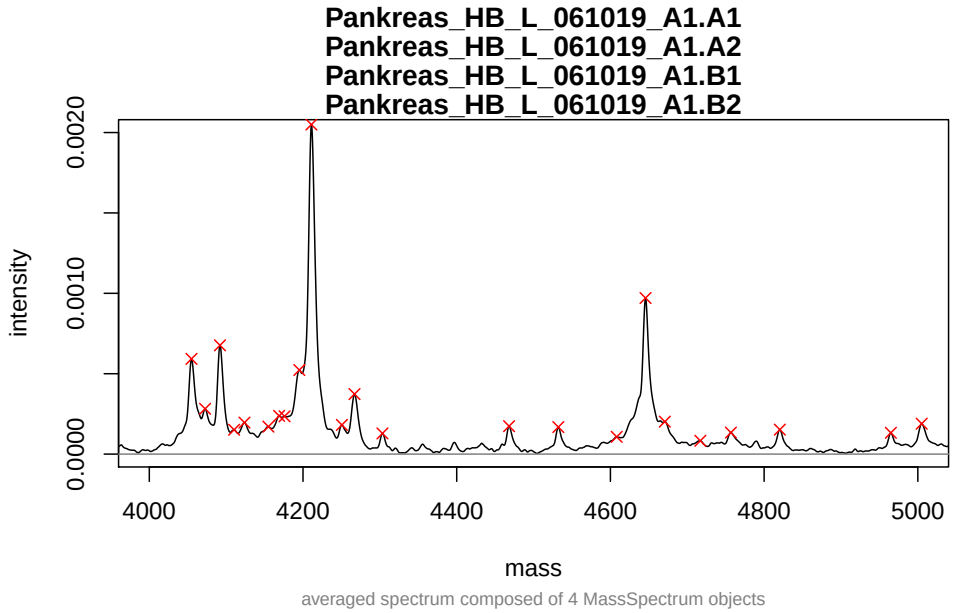
```
noise <- estimateNoise(avgSpectra[[1]])
plot(avgSpectra[[1]], xlim=c(4000, 5000), ylim=c(0, 0.002))
lines(noise, col="red")                      # SNR == 1
lines(noise[, 1], 2*noise[, 2], col="blue") # SNR == 2
```



In this case we decide to set a *SNR* of 2 (blue line).

```
peaks <- detectPeaks(avgSpectra, SNR=2, halfWindowSize=20)
```

```
plot(avgSpectra[[1]], xlim=c(4000, 5000), ylim=c(0, 0.002))
points(peaks[[1]], col="red", pch=4)
```



3.9 Post Processing

After the alignment the peak positions (mass) are very similar but not identical. The binning is needed to make similar peak mass values identical.

```
peaks <- binPeaks(peaks)
```

We choose a very low signal-to-noise ratio to keep as much features as possible. To remove some false positive peaks we remove peaks that appear in less than 50 % of all spectra in each group.

```
peaks <- filterPeaks(peaks, minFrequency=c(0.5, 0.5),
                     labels=avgSpectra.info$health,
                     mergeWhitelists=TRUE)
```

Finally we create the feature matrix and label the rows with the corresponding patient ID.

```
featureMatrix <- intensityMatrix(peaks, avgSpectra)
rownames(featureMatrix) <- avgSpectra.info$patientID
```

3.10 Diagonal Discriminant Analysis

We finish the `MALDIquant` preprocessing and use the *diagonal discriminant analysis* (DDA) function of `sda` (Ahdesmäki and Strimmer, 2010) to find the most important peaks.

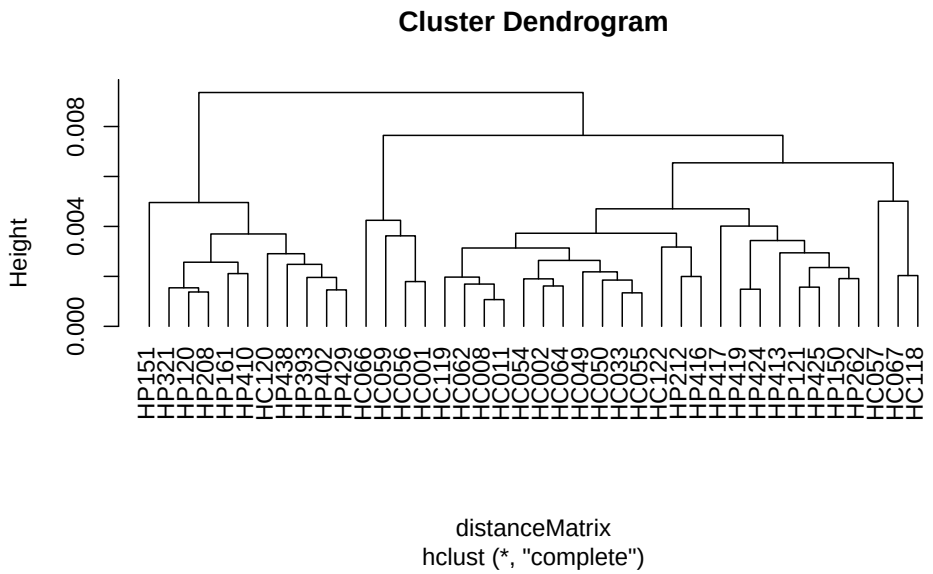
```
Xtrain <- featureMatrix
Ytrain <- avgSpectra.info$health
ddar <- sda.ranking(Xtrain=featureMatrix, L=Ytrain, fdr=FALSE,
                    diagonal=TRUE)

Computing t-scores (centroid vs. pooled mean) for feature ranking

Number of variables: 166
Number of observations: 40
Number of classes: 2

Estimating optimal shrinkage intensity lambda.freq (frequencies): 1
Estimating variances (pooled across classes)
Estimating optimal shrinkage intensity lambda.var (variance vector): 0.107
```

	idx	score	t.cancer	t.control
8936.97236585095	158.00	90.69	9.52	-9.52
4468.06600951353	116.00	80.80	8.99	-8.99
8868.2678310697	157.00	80.06	8.95	-8.95
4494.80267780907	117.00	67.00	8.19	-8.19
8989.20382965523	159.00	66.19	8.14	-8.14
5864.49053296298	135.00	37.56	-6.13	6.13
5906.17351903972	136.00	34.43	-5.87	5.87
2022.94475790442	49.00	33.30	5.77	-5.77
5945.5697657874	137.00	32.66	-5.71	5.71
1866.16591692444	44.00	32.12	5.67	-5.67

3.11 Hierarchical Clustering

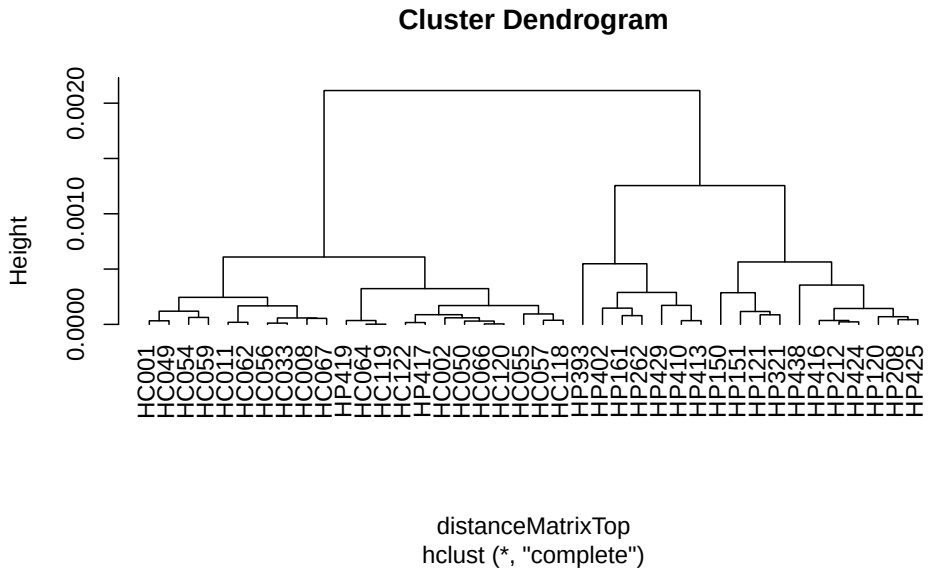
To visualize the results without any feature selection by *DDA* we apply a hierarchical cluster analysis based on the euclidean distance.

```
distanceMatrix <- dist(featureMatrix, method="euclidean")

hClust <- hclust(distanceMatrix, method="complete")

plot(hClust, hang=-1)
```



Next we use only the 2 top peaks selected in the *DDA* and we get a nearly perfect split between the cancer and control group.

```
top <- ddar[1:2, "idx"]

distanceMatrixTop <- dist(featureMatrix[, top],
                          method="euclidean")

hClustTop <- hclust(distanceMatrixTop, method="complete")

plot(hClustTop, hang=-1)
```



3.12 Cross Validation

Subsequently we use the `crossval` (Strimmer, 2014) package to perform a 10-fold cross validation of these two selected peaks.

```
# create a prediction function for the cross validation
predfun.dda <- function(Xtrain, Ytrain, Xtest, Ytest,
                        negative) {
  dda.fit <- sda(Xtrain, Ytrain, diagonal=TRUE, verbose=FALSE)
  ynew <- predict(dda.fit, Xtest, verbose=FALSE)$class
  return(confusionMatrix(Ytest, ynew, negative=negative))
}

# set seed to get reproducible results
set.seed(1234)

cv.out <- crossval(predfun.dda,
                   X=featureMatrix[, top],
                   Y=avgSpectra.info$health,
                   K=10, B=20,
```

```
                  negative="control",
                  verbose=FALSE)
diagnosticErrors(cv.out$stat)

      acc      sens      spec       ppv       npv       lor
0.9500000 0.9000000 1.0000000 1.0000000 0.9090909       Inf
```

3.13 Summary

We found the peaks *m/z* 8937 and 4467 as important features for the discrimination between the cancer and control group.

4 Session Information

- R version 3.1.2 (2014-10-31), `x86_64-pc-linux-gnu`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: colorout 1.0-2, corpcor 1.6.7, crossval 1.0.2, devtools 1.7.0, entropy 1.2.1, fdrtool 1.2.13, knitr 1.8, MALDIquant 1.11.1, MALDIquantExamples 0.1, MALDIquantForeign 0.9, pvclust 1.3-2, sda 1.3.5, setwidth 1.0-3, vimcom.plus 0.9-93, xtable 1.7-4
- Loaded via a namespace (and not attached): base64enc 0.1-2, digest 0.6.8, downloader 0.3, evaluate 0.5.5, formatR 1.0, highr 0.4, readBrukerFlexData 1.8.2, readMzXmlData 2.8, stringr 0.6.2, tools 3.1.2, XML 3.98-1.1

References

Ahdesmäki, M. and Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *The Annals of Applied Statistics*, 4(1):503–519.

Fiedler, G. M., Leichtle, A. B., Kase, J., Baumann, S., Ceglarek, U., Felix, K., Conrad, T., Witzigmann, H., Weimann, A., Schtte, C., Hauss, J., Büchler, M., and Thiery, J. (2009). Serum peptidome profiling revealed platelet factor 4 as a potential discriminating peptide associated with pancreatic cancer. *Clinical Cancer Research*, 15:3812–3819.

Gibb, S. (2014). *MALDIquantForeign: Import/Export routines for MALDIquant*. R package version 0.7.

Gibb, S. and Strimmer, K. (2012). MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, 28(17):2270–2271.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ryan, C. G., Clayton, E., Griffin, W. L., Sie, S. H., and Cousens, D. R. (1988). SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 34:396–402.

Savitzky, A. and Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36:1627–1639.

Strimmer, K. (2014). *crossval: Generic Functions for Cross Validation*. R package version 1.0.0.

Wickham, H. and Chang, W. (2014). *devtools: Tools to make developing R code easier*. R package version 1.5.

D Erklärung über die eigenständige Abfassung der Arbeit

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und ohne unzulässige Hilfe oder Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Ich versichere, dass Dritte von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen, und dass die vorgelegte Arbeit weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde zum Zweck einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt wurde. Alles aus anderen Quellen und von anderen Personen übernommene Material, das in der Arbeit verwendet wurde oder auf das direkt Bezug genommen wird, wurde als solches kenntlich gemacht. Insbesondere wurden alle Personen genannt, die direkt an der Entstehung der vorliegenden Arbeit beteiligt waren.

Leipzig, den 11. September 2015 _____

Sebastian Gibb

E Lebenslauf

F Danksagung

Ich möchte mich ganz herzlich bei Herrn Prof. Dr. Korbinian Strimmer für seine exzellente Betreuung bedanken.

Ich danke den Autoren von Fiedler et al. (2009) für die Bereitstellung der in dieser Dissertation genutzten Daten. Außerdem möchte ich Herrn Dr. med. Alexander B. Leichtle für die vielen hilfreichen Erklärungen und Hinweise rund um die Massenspektrometrie danken.

Für den einen oder anderen statistischen bzw. mathematischen Rat, die vielen Lebensweisheiten und das angenehme Arbeitsklima bedanke ich mich bei allen Kollegen des IMISE, insbesondere bei meinen Bürokollegen Dr. Bernd Klaus, Dipl.-Math. Katja Rösch, M.Sc. DIC Maryam Yahiaoui-Doktor und Dr. Verena Zuber.