Algorithmen zur labelfreien quantitativen Proteomanalyse auf Basis datenunabhängig-akquirierter LC-MS-Daten

Dissertation

zur Erlangung des Grades

Doktor der Naturwissenschaften

am Fachbereich Biologie

der Johannes Gutenberg-Universität Mainz

Jörg Kuharev

geboren am 01. April 1980

in Woskresenowka (Kasachstan)

Mainz, April 2015

Dekan:

1. Gutachter:

2. Gutachter:

Tag der Promotion: 10.12.2015

Inhaltsverzeichnis

1 Einleitung **1**

1.1 Das Proteom und die Proteomik 1

1.2 Massenspektrometrie in der Proteomik 2

1.3 Massenspektrometer 2

1.3.1 Elektrospray-Ionisation 5

1.3.2 Flugzeitmassenspektrometrie 7

1.3.3 Tandem-Massenspektrometrie 8

1.3.4 Ionenmobilitätsspektrometrie 9

1.3.5 Ionenmobilitätsspektrometrie-Massenspektrometrie 9

1.4 Reduktion der Probenkomplexität 11

1.4.1 Fraktionierung 11

1.4.2 Online-Kopplung 11

1.4.3 Hochleistungsflüssigkeitschromatographie 12

1.5 Top-down und Bottom-up Analysen 13

1.6 Datenakquisition 15

1.6.1 Datenabhängige Akquisition 15

1.6.2 Gezielte Akquisition 16

1.6.3 Datenunabhängige Akquisition 16

1.7 Datenanalyse 18

1.7.1 Rohdaten 18

1.7.2 Signalverarbeitung 18

1.7.3 Peptid- und Proteinidentifikation 20

1.7.4 Von der Proteinidentifikation zur Quantifizierung 21

1.7.5 Proteinquantifizierung mit Markierungstechniken 22

1.7.6 Labelfreie Proteinquantifizierung 23

1.7.7 Absolute Quantifizierung 24

1.8 Spezifische Analyse der MS^E/$HDMS^E$/$UDMS^E$-Daten 26

1.8.1 Herausforderungen 26

1.8.2 Stand der Technik 27

1.9 Zielsetzung 30

2 Material und Methoden **31**

2.1 Chemikalien 31

2.2 Testdatensätze 31

2.2.1 HeLa-Proteom . . . 31
2.2.2 Metaproteom . . . 31
2.3 Filter-gestützte Probenvorbereitung . . . 33
2.4 LC-MS-Analyse . . . 35
2.5 Datenanalyse . . . 37
2.5.1 Proteindatenbanken . . . 37
2.5.2 PLGS . . . 37
2.5.3 ISOQuant . . . 38
2.5.4 synapter . . . 39
2.5.5 Progenesis QIP . . . 40
2.6 Geräte und Software . . . 42

3 Ergebnisse **44**
3.1 Analyseworkflow . . . 45
3.2 Datenzugriff und Zusammenführung der PLGS-Ergebnisse . . . 47
3.3 Retentionszeitalignment . . . 49
3.3.1 Paarweises Retentionszeitalignment . . . 50
3.3.2 Dynamic Retention Time Warping . . . 51
3.3.3 Fast Dynamic Retention Time Warping . . . 54
3.3.4 Linear Dynamic Retention Time Warping . . . 57
3.3.5 Fast Linear Dynamic Retention Time Warping . . . 60
3.3.6 Effizienzvergleich der Retentionszeitalignmentalgorithmen . . . 60
3.3.7 Multiples Retentionszeitalignment . . . 63
3.4 Feature-Clustering . . . 65
3.4.1 Preclustering . . . 65
3.4.2 Raumtransformation . . . 66
3.4.3 Dichtebasiertes Clustering . . . 67
3.5 Normalisierung der Feature-Intensitäten . . . 68
3.6 Filterung der Peptididentifikationen . . . 71
3.7 Annotation der Feature-Cluster . . . 71
3.8 Filterung der Peptid-FDR . . . 74
3.9 Protein-Homologie-Filter . . . 75
3.10 Verteilung der Peptidintensitäten . . . 76
3.11 Absolute Proteinquantifizierung . . . 77
3.12 Filterung der Protein-FDR . . . 77
3.13 Implementierung des Analyseworkflows - ISOQuant . . . 78

3.14 Vergleich mit PLGS 79
3.15 Vergleich mit Progenesis QI for Proteomics und synapter 84
3.15.1 HeLa-Hefe-*E.coli*-Metaproteom 84
3.15.2 Analyseworkflows 86
3.15.3 Filterung der Analyseergebnisse 87
3.15.4 Proteinidentifikation 88
3.15.5 Präzision der relativen labelfreien Quantifizierung 91
3.15.6 Richtigkeit der relativen labelfreien Quantifizierung 93

4 Diskussion **96**
4.1 Datenzugriff und Zusammenführung der PLGS-Ergebnisse 97
4.2 Retentionszeitalignment 97
4.3 Feature-Clustering 99
4.4 Normalisierung der Feature-Intensitäten 101
4.5 Filterung der Peptididentifikationen und Annotation der Feature-Cluster . . 103
4.6 Filterung der False Discovery Rate 104
4.7 Protein-Homologie-Filter 104
4.8 Verteilung der Peptidintensitäten 105
4.9 Absolute Proteinquantifizierung 106
4.10 Analyseworkflow und Implementierung 106
4.11 Vergleich mit PLGS 109
4.12 Vergleich mit synapter und Progenesis 110

Zusammenfassung **115**

Eigene Publikationen **116**

Literaturverzeichnis **117**

Abkürzungsverzeichnis **125**

Abbildungsverzeichnis **129**

Tabellenverzeichnis **130**

1 Einleitung

1.1 Das Proteom und die Proteomik

Proteine bilden die Grundlage unseres Lebens. Erst durch sie wird der genetische Bauplan verwirklicht. Sie übernehmen im Organismus vielfältige Funktionen. Proteine bilden Strukturen und Komplexe, sie transportieren Stoffe, sie katalysieren Reaktionen, sie interagieren miteinander und mit anderen Molekülen und sie übertragen Signale. Proteine weisen eine hohe Variabilität in ihrer Struktur und ihren chemischen Eigenschaften auf. Im Gegensatz zum vergleichsweise statischen Genom unterliegen Proteine einer ständigen Dynamik. Ihre Expressionsmuster ändern sich mit jedem Zelltyp und Entwicklungsstadium. Alternatives Spleißen und posttranslationale Proteinmodifikationen sorgen in vielen Organismen für eine erheblich höhere Zahl an Proteinen als die Anzahl der entsprechenden Gene erwarten ließe[1]. So liefert die Analyse des Genoms in den meisten Fällen lediglich begrenzte Informationen über die exprimierten Proteine. Auch Rückschlüsse vom Transkriptom auf Proteine sind nur eingeschränkt möglich, da keine direkte Korrelation zwischen den Mengen der Boten-RNA (mRNA, engl. messenger RNA) und den Mengen entsprechender Proteine nachgewiesen werden konnte[2]. Deshalb widmet sich die Proteomforschung oder Proteomik (engl. proteomics) der Untersuchung von Proteinen auf der Ebene des sogenannten Proteoms – der Gesamtheit aller Proteine in einem Organismus, einem Gewebe oder einer Zelle, die zu einer bestimmten Zeit und unter bestimmten Bedingungen exprimiert werden[3]. Die Proteomforschung ergänzt die Genom- und Transkriptomforschung, indem sie Zustände erfasst, die auf Ebene des Genoms nicht ermittelt werden können, z.B. posttranslationale Modifikationen wie Phosphorylierungen, Glykosylierungen, Acetylierungen, Methylierungen, etc. aber auch proteolytische Abbaureaktionen. Die Proteomforschung zielt auf ein besseres Verständnis der Zellmechanismen sowie auf die Entwicklung neuer Therapieansätze und Wirkstoffe gegen verschiedene Krankheiten, wie z.B. Krebs, Infektionen oder Erkrankungen des zentralen Nervensystems ab[4]. Auf dem Weg zur Klärung der Frage danach, welche Proteine zu welcher Zeit in welchen Zellen welche Wirkung erzielen[5], besteht eine wichtige Aufgabe der Proteomforschung in der qualitativen und quantitativen Charakterisierung der Proteome.

1.2 Massenspektrometrie in der Proteomik

Seit Beginn des 20. Jahrhunderts wurden Massenspektrometer mehrere Jahrzehnte lang von Forschern für Untersuchungen chemischer Elemente, ihrer Isotope und einfacher Moleküle genutzt. Die systematische Analyse von Proteinen mit Hilfe der Massenspektrometrie (MS) wurde jedoch erst in den 1980er Jahren mit der Entwicklung schonender Ionisationsverfahren ermöglicht. Seitdem hat sich die MS durch technologische Fortschritte bei der Entwicklung sensitiver, hochauflösender Massenspektrometer zu einer zentralen Analysemethode der Proteomik entwickelt. Ihre Dominanz in der Proteomik wird durch die Fähigkeit begründet, umfangreiche qualitative und quantitative Information über biologische Proben enormer Komplexität abzubilden[6]. Mit Hilfe der MS können heutzutage komplexe Proteingemische wie subzelluläre Fraktionen oder gesamte Proteome verschiedener Zelltypen untersucht werden. So ermöglichte die MS-basierte Proteomik bereits entscheidende Einblicke in die Zusammensetzung, Regulation und Funktion molekularer Komplexe und Signalwege.

1.3 Massenspektrometer

Ein Massenspektrometer besteht im Wesentlichen aus drei nacheinander geschalteten Bauteilen: einer Ionenquelle, einem Analysator und einem Detektor. Der Analyt wird in der Ionenquelle ionisiert. Der Analysator oder Massenselektor trennt Ionen nach ihrem Masse-zu-Ladung-Verhältnis (*m/z*) bevor diese im Detektor erfasst werden. Diese drei Komponenten können nach unterschiedlichen technischen Prinzipien ausgeführt sein und werden innerhalb eines Instrumentes oft in bestimmten Kombinationen verwendet.

Zur Ionisation des Analyts wurden im Laufe der Jahre verschiedene Methoden entwickelt, wie z.B. Fast Atom Bombardment (FAB)[7], chemische Ionisation (CI)[8] oder chemische Ionisation bei Atmosphärendruck (APCI, engl. atmospheric-pressure chemical ionization)[9]. Am weitesten verbreitet sind in der Proteomik schonende Ionisationsverfahren die Matrix-unterstützte Laser-Desorption/Ionisation (MALDI, engl. matrix-assisted laser desorption/ionization)[10] und die Elektrospray-Ionisation (ESI)[11]. Die prinzipielle Funktionsweise der ESI wird im Kapitel 1.3.1 näher beschrieben.

Verschiedene Analysatoren verwenden für die Massentrennung unterschiedliche physikalische Prinzipien. Im Sektorfeld-Massenspektrometer werden die Ionen in statischen elektromagnetischen Feldern abgelenkt und beschreiben in Kreisbahnen unterschiedliche Radien abhängig von ihrer Masse. Im Quadrupol-Massenspektrometer (QMS) wird ein

elektrischer Quadrupol als Analysator eingesetzt. Dabei durchfliegen die durch ein statisches Feld beschleunigten Ionen zwischen den parallel angeordneten Stabelektroden des Quadrupols ein Wechselfeld, in welchem eine Selektion der Ionen nach *m/z* stattfindet[12]. Eine ganze Familie von Masseanalysatoren umfassen die sogenannten Ionenfallen-Massenspektrometer. Ionenfallen halten Ionen mit Hilfe elektromagnetischer Felder in einem definierten Bereich, wo sie analysiert und manipuliert werden können. Folgende Ionenfallentypen werden in der MS verwendet: Quadrupol-Ionenfalle (Paul-Falle)[12], Linear trap[13], Fouriertransformations-Ionenzyklotronresonanz (FT-ICR)[14] und Orbitrap[15]. In einem Flugzeitmassenspektrometer (TOF, engl. time of flight) werden Ionen im definierten elektrischen Feld beschleunigt und durchlaufen daraufhin eine Flugstrecke bekannter Länge in unterschiedlicher Zeit abhängig von ihrem *m/z*[16]. Die Flugzeitmassenspektrometrie (TOF-MS, engl. time of flight mass spectrometry) wird im Kapitel 1.3.2 näher erläutert. Mehrere Analysatoren können in einem einzigen mehrstufigen oder hybriden Massenspektrometer kombiniert werden, wo ihnen abhängig vom Experiment unterschiedliche Rollen zugeteilt werden können, z.B. als m/z-Filter, Kollisionszelle, Ionenfalle. In Abbildung 1 wird der prinzipielle Aufbau eines einstufigen Massenspektrometers mit dem eines hybriden Massenspektrometers verglichen. Ein hybrides Massenspektrometer ermöglich die so genannte Tandem-Massenspektrometrie (MS/MS), bei der zwischen den Messvorgängen die Analytionen fragmentiert werden[17]. Die Funktionsweise der MS/MS wird im Kapitel 1.3.3 näher erläutert.

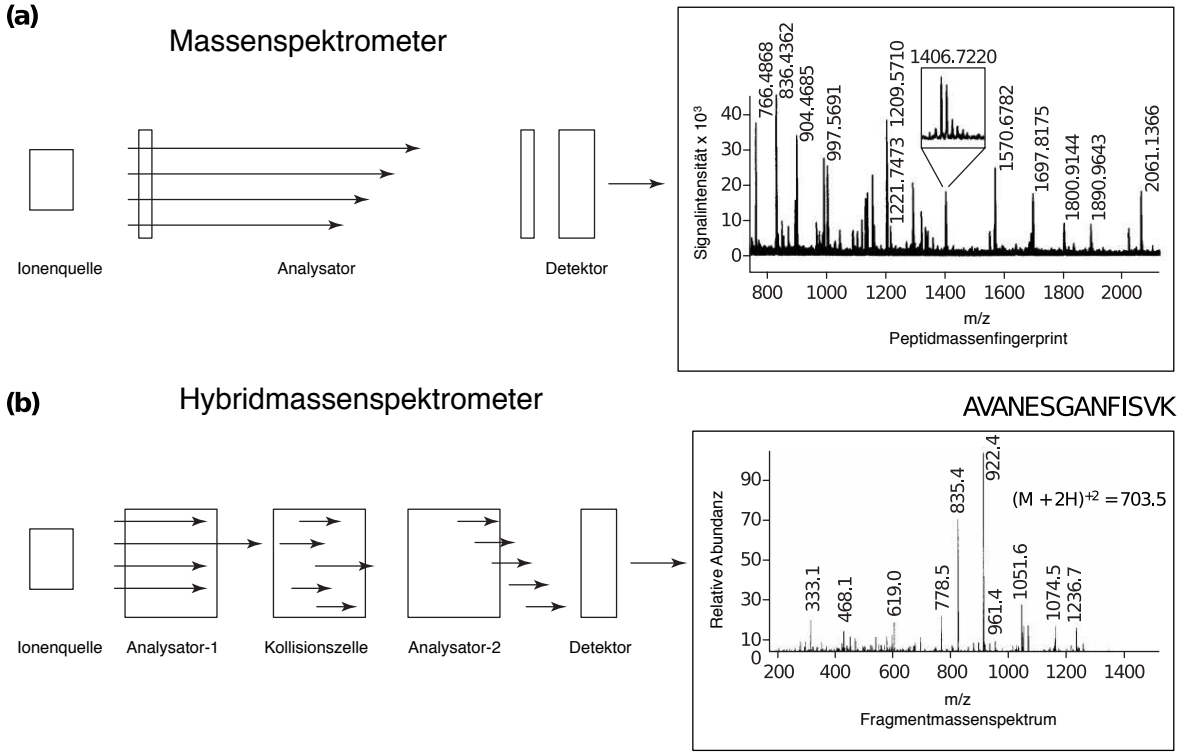

Abb. 1: Aufbau eines einstufigen und eines hybriden Massenspektrometers.
(a) Einstufiges Massenspektrometer. Ein einstufiges Massenspektrometer besteht aus einer Ionenquelle, einem Analysator (gezeigt Flugzeitanalysator, TOF) und einem Detektor. Durch einstufige MS können Proteine durch Erfassung der absoluten Massen ihrer tryptischen Peptide mit der sogenannten Peptidmassenfingerprint-Methode identifiziert werden[18–22]. An einem Beispiel wird rechts das Massenspektrum eines Proteinverdaus gezeigt.
(b) Hybrides Massenspektrometer. Ein hybrides Massenspektrometer besteht aus einer Ionenquelle, mehreren hintereinander geschalteten Analysatoren und einem Detektor. Für den tandemmassenspektrometrischen Betrieb (s. Kapitel 1.3.3) wird der erste Analysator als *m/z*-Filter zur Isolation bestimmter Ionenspezies, der zweite als Kollisionszelle und der dritte zur *m/z*-Analyse verwendet. So können Fragmentionenmuster einzelner oder mehrerer Analytionenspezies erzeugt werden. Die entstehenden Fragmentierungsmuster können zur Proteinidentifikation genutzt werden. An einem Beispiel wird rechts das Fragmentmassenspektrum eines tryptischen Peptides gezeigt.[23]
Das Bild wurde adaptiert von Yates et al., 2000.[23]

1.3.1 Elektrospray-Ionisation

ESI ist ein schonendes Ionisationsverfahren, das Ionen aus einer Lösung mit Hilfe eines sogenannten Elektrosprays erzeugt. Wie in Abbildung 2 gezeigt, wird der in einem flüchtigen Lösungsmittel gelöste Analyt durch eine dünne Metallkapillare geleitet, an dessen Ende ein starkes elektrisches Feld angelegt wird.

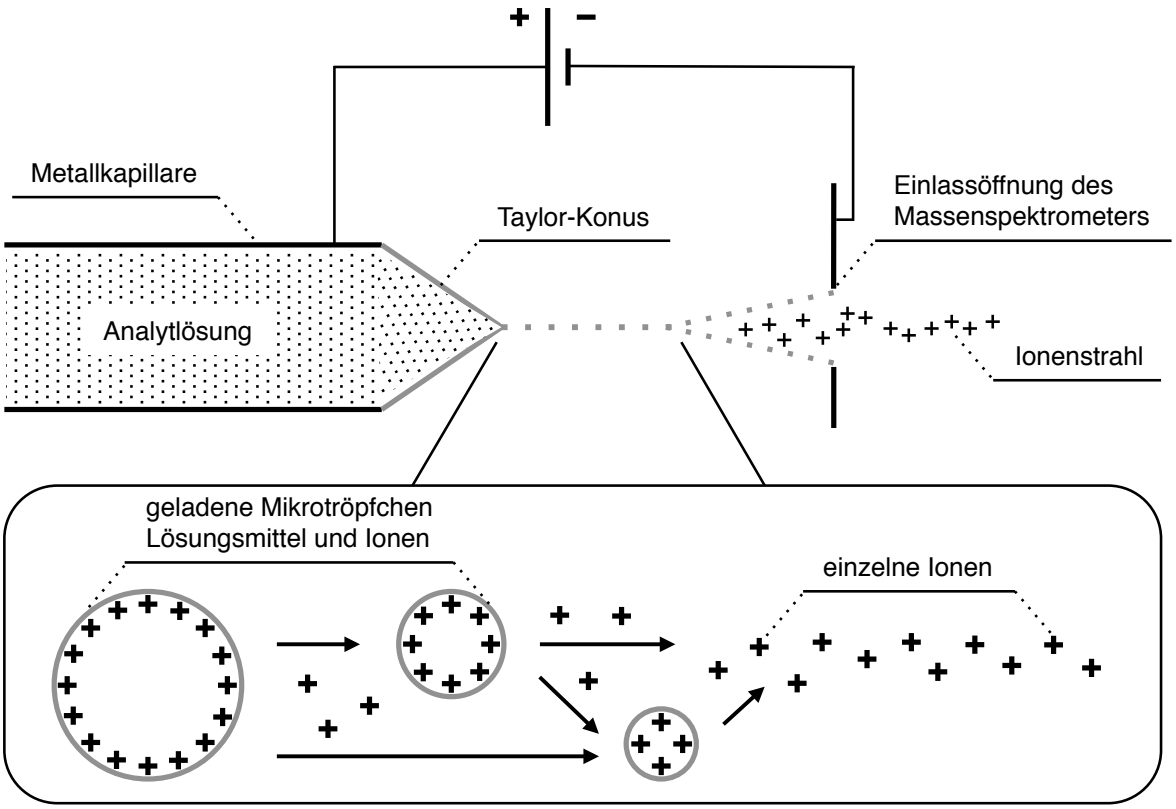

Abb. 2: Funktionsweise der Elektrospray-Ionisation.
Die Analytlösung wird durch eine dünne Metallkapillare geleitet, an der ein starkes elektrisches Potential gegenüber einer Gegenelektrode am Einlass des Massenspektrometers anliegt. Das Lösungsmittel verdunstet und an der Spitze der Kapillare bildet sich ein Taylor-Konus[24], von dem sich geladene Mikrotröpfchen ablösen, die unter Verdunstung immer kleiner werden und nach Überschreiten des Rayleigh-Limits[25] mit sogenannten Coulomb-Explosionen einzelne gasförmige Ionen freisetzen und einen Ionenstrahl erzeugen.

Durch Potentialdifferenz zwischen der Kapillare und einer Elektrode am Eingang des Massenspektrometers bildet sich an der Spitze der Kapillare ein Überschuss gleichartig geladener Ionen aus der Analytlösung, die sich über die Bildung eines sogenannten Taylor-Konus (auch Taylor-Kegel)[24] als feine, geladene Tröpfchen von der Kapillare ablösen. Durch Verdampfung des Lösungsmittels sinkt die Größe der Tröpfchen und ihre Oberflächenladung steigt. Nach erreichen einer kritischen Größe, des Rayleigh-Limits[25], zerspringt das Tröpfchen in immer kleinere Mikrotröpfchen, bis diese wiederum das Rayleigh-Limit erreichen oder setzt mit einer sogenannten Coulomb-Explosion gasförmige, gleichgeladene

einzelne Ionen in der Gasphase frei. Die resultierenden Analytionen werden anschließend massenspektrometrisch untersucht.[11, 26, 27]

Im Jahr 1994 stellten M.S. Wilm und M. Mann eine Weiterentwicklung der ESI vor[28]. Bei dieser zunächst als Mikrospray, später als Nanospray oder nano-ESI bezeichneten Technik wird statt einer bis dahin üblichen ca. 100 µm weiten Metallkapillare eine dünne Glaskapillare mit einem Innendurchmesser im unteren µm-Bereich eingesetzt. Das Durchflussvolumen des Analyts und somit die Menge der Probe werden so weit reduziert, dass wenige µL des Analyts für eine massenspektrometrische Untersuchung ausreichen. Zusätzlich wird die Größe der emittierten Tröpfchen auf etwa 100 nm reduziert und so die Verdampfung des Lösungsmittels begünstigt. Nano-ESI ist in der Lage, Analytmoleküle aus einer wässrigen Lösung ohne Zugabe weiterer Lösungsmittel zu ionisieren und zeigt eine erhöhte Toleranz gegenüber Salzverunreinigungen.[29, 30]

Eine wichtige Voraussetzung für ein MS-Experiment ist eine gleichbleibende Signalstabilität über die Dauer eines Experiments. Die Elektrospraystabilität und die damit verbundene Signalstabilität hängen von einer ganzen Reihe verschiedener Faktoren ab[31–36]. Schwankungen der experimentellen Bedingungen einschließlich der angelegten elektrischen Spannung, Flussrate und Zusammensetzung der Analytlösung sowie die Emittergeometrie haben einen direkten Einfluss auf die Spraystabilität. Der Zusammenhang zwischen der Spraystabilität und der resultierenden analytischen Leistung des Systems wurde über Jahrzehnte hinreichend untersucht[34, 37, 38]. Die meisten Einflussfaktoren werden durch das spezifische Hardwaredesign, die Art der Analyse und die experimentellen Bedingungen, wie z.B. die Zusammensetzung des Analyts, vorgegeben. In der Praxis erfolgt eine individuelle Optimierung des Spraysystems durch manuelle Korrekturen der Emitterposition und der angelegten Spannung. Jedoch werden bereits Ansätze zur Optimierung der ESI durch automatische Korrektur der angelegten Spannung mit Hilfe von Rückkopplungssystemen[39]. Eine große Herausforderung hinsichtlich der Spraystabilität stellen Analytlösungen mit hohen Salz- und Detergenzienkonzentrationen dar[40]. Beim Betrieb des Elektrosprays zur Erzeugung von Anionen wird die Spraystabilität zusätzlich durch spontane Koronaentladungen beeinträchtigt[41]. Problematisch ist vor Allem eine Erhaltung der Spraystabilität über die Gesamtdauer eines Experiments, wenn die Zusammensetzung der Analytlösung sich im Laufe des Experiments ändert[40].

1.3.2 Flugzeitmassenspektrometrie

Bei der TOF-MS werden Ionen im definierten elektrischen Feld beschleunigt und durchlaufen daraufhin eine feldfreie Driftstrecke bekannter Länge in unterschiedlicher Zeit abhängig von ihrem *m/z*[16]. Die gemessene Flugzeit T_f ist proportional zur Quadratwurzel des Masse-zu-Ladung-Verhältnisses, $T_f \propto \sqrt{\frac{m}{z}}$. Werden alle Analytionen mit der gleichen Energie beschleunigt, so hängt der Zeitversatz zwischen den Ionen alleine von der Länge des Flugpfads ab. Die Massenauflösung kann durch Anwendung eines Ionenspiegels in Form eines Reflektrons verbessert werden. Ein Reflektron baut am Ende des Flugpfades ein der Beschleunigungsspannung entgegengesetztes elektrisches Feld mit einem Gradienten auf, so dass die Ionen abgebremst und in die entgegengesetzte Richtung umgelenkt werden[42]. Durch einen Reflektron kann die Flugbahn der Ionen V-förmig und durch mehrere Reflektrone etwa W-förmig mehrfach umgelenkt und fokussiert werden. Die Driftstrecke wird so entsprechend um ein Vielfaches verlängert. Neueste Generationen kommerzieller Flugzeitmassenspektrometer erlauben Massenauflösungen von über 60000 FWHM (engl. full width at half maximum), Genauigkeiten von wenigen ppm und decken dabei einen breiten Massenbereich ab. Sie eignen sich besonders gut für die schnelle Erfassung komplexer Massenspektren in Proteomik-Experimenten.

Die Geometrie eines üblicherweise aus einem Leichtmetall gefertigten Flugzeitanalyserohrs kann sich im Laufe eines Experiments durch Schwankungen der Umgebungstemperatur ändern. Dies führt zu Messfehlern. Eine gängige Lösung bietet die Nutzung eines so genannten Lockmassenstandards[43]. Als Lockmasse wird die Masse eines Ions mit einem bekannten *m/z* bezeichnet. Die Massenabweichung des Lockmassenstandards kann entweder für eine Echtzeit- oder eine nachträgliche Rekalibrierung, eine Korrektur systematischer Messfehler verwendet, die durch die Instrumentendrift verursacht wurden.

1.3.3 Tandem-Massenspektrometrie

Bei der MS/MS handelt es sich um eine mehrstufige MS, bei der zunächst intakte Analytionen, die sogenannten Vorläuferionen (auch Präkursorionen) analysiert werden, sie werden anschließend durch geeignete Techniken in Fragmente aufgeschlossen und die resultierenden Fragmentionen (auch Produktionen) wiederum massenspektrometrisch analysiert[17]. In der Proteomik bedeutet dies eine massenspektrometrische Erfassung der Peptidionen mit ihrer anschließenden hochenergetischen Fragmentierung sowie der massenspektrometrischen Erfassung der entstehenden Peptidfragmentionen, die eine oder mehrere wenige Aminosäuren enthalten können. Mehrere Technologien wurden beschrieben, bei denen durch unterschiedliche physikalische Effekte eine Fragmentierung der Vorläuferionen erreicht werden kann, z.B. “surface-induced dissociation” (SID)[44], “photon-induced dissociation” (PID)[45], “infrared multiphoton dissociation” (IRMPD)[46], “electron capture dissociation” (ECD)[47] und “electron-transfer dissociation” (ETD)[48]. Am weitesten verbreitet ist jedoch die sogenannte kollisionsinduzierte Dissoziation (CID oder CAD, engl. collision-induced dissociation bzw. collisionally activated dissociation), bei der eine Fragmentierung der Vorläuferionen durch Kollisionen mit neutralen Gasmolekülen wie Helium, Stickstoff oder Argon erreicht wird[49]. Bei Zusammenstößen der Kollisionsgas- und der Analytmoleküle wird von den Analytmolekülen ein Teil der Kollisionsenergie absorbiert. Der Anstieg der inneren Energie führt dann zu ihrer Fragmentierung. Neben der herkömmlichen CID kann in Orbitrap-Massenspektrometern die Fragmentierung der Vorläuferionen mit einer weiteren Methode der sogenannten kollisionsinduzierten Dissoziation höherer Energie (HCD, engl. higher-energy collisional dissociation bzw. higher-energy C-trap dissociation) durchgeführt werden[50].

Die einstufige MS erlaubt anhand der sogenannten Peptidmassenfingerprints die Identifikation einzelner Proteine[18–22]. Dabei werden die massenspektrometrisch erfassten Massen der Peptide mit theoretischen Peptidmassen aus einer Datenbank für bekannte oder hypothetische Proteinsequenzen verglichen. Eine Analyse von komplexen Proben wird erst durch die Anwendung der MS/MS möglich. Dabei wird gegenüber der einstufigen MS die Zuverlässigkeit der Proteinidentifikation durch die MS/MS gesteigert, indem Fragmentmassenspektren mit entsprechenden theoretisch berechneten Massenspektren verglichen werden[51]. Alternativ können die resultierenden Fragmentmassenspektren bei der sogenannten De-Novo-Peptidsequenzierung Aufschluss über die Aminosäuresequenz des analysierten Vorläuferpeptides liefern[52–56].

1.3.4 Ionenmobilitätsspektrometrie

Unter atmosphärischem Druck beschleunigen gasförmige Ionen in einem konstanten elektrischen Feld bis sie mit neutralen Gasmolekülen zusammenstoßen und beschleunigen wieder bis zur nächsten Kollision. Die Sequenz wiederholter Beschleunigungen und Kollisionen resultiert auf makroskopischer Ebene in konstanter Geschwindigkeit charakteristisch für eine Ionenspezies. Das Verhältnis der Geschwindigkeit einer Ionenspezies zur Stärke des elektrischen Feldes wird als Ionenmobilität und das Verfahren zur Analyse der Ionen anhand der Unterschiede ihrer Ionenmobilität wird als Ionenmobilitätsspektrometrie (IMS) bezeichnet[57]. Der Aufbau eines Ionenmobilitätsspektrometers ähnelt dem eines einfachen Massenspektrometers. Es besteht ebenfalls aus einer Ionenquelle, einem Analysator und einem Detektor. In einem Ionenmobilitätsanalysator werden die Ionen von einem elektrischen Feld gezielt durch ein Gas geleitet, in dem sie unterschiedliche, charakteristische Driftgeschwindigkeiten aufweisen. Nach der Kollisionsstrecke werden die Driftzeiten der Analytionen erfasst. Die erfassten Driftzeiten korrelieren mit der Ladung und Struktur der Moleküle und lassen auf die Identität der jeweiligen Analytionenspezies schließen.

Die IMS-Instrumente benötigen im Gegensatz zu Massenspektrometern kein Hochvakuum und die Umgebungsluft kann als Trägergas verwendet werden. Sie sind deshalb kompakter als Massenspektrometer. Typischerweise wird die IMS zur Detektion von chemischen Kampfstoffen, Sprengstoffen oder Drogen in Flughäfen und beim Militär aber auch für medizinische Atemgasanalysen eingesetzt.

1.3.5 Ionenmobilitätsspektrometrie-Massenspektrometrie

Bei der Ionenmobilitätsspektrometrie-Massenspektrometrie (IMS-MS) handelt es um eine Kombination der beiden Techniken um die Möglichkeiten der MS durch die Analyse der Ionenmobilität zu erweitern. Anfang der 1960er Jahre wurden die ersten Geräte beschrieben, die die IMS mit einem Sektorfeld-Massenspektrometer[58] bzw. mit einem Flugzeitmassenspektrometer[59] kombinieren. Einige Massenspektrometer neuerer Generationen integrieren zusätzlich zur *m/z*-Trennung die Ionenmobilitätsseparation[60]. Durch diese zusätzliche Trenndimension erlaubt die IMS-MS in vielen Fällen eine Unterscheidung isomerer Moleküle, welche die gleiche Masse und Ladung jedoch unterschiedliche Strukturen aufweisen.

Eine Möglichkeit die Ionen nach ihrer Mobilität zu trennen, ist die Technologie “travelling wave ion mobility mass spectrometry” (TWIMS)[61]. Vereinfacht dargestellt, werden bei dieser Technik die Analytionen über wellenförmig wandernde elektrische Hochfrequenzfelder durch eine gasgefüllte IMS-Zelle geleitet, erfahren proportional ihrer Größe und abhängig vom eingesetzten Driftgas einen Widerstand und werden quasi *m/z*-unabhängig aufgetrennt. Abbildung 3 zeigt am Beispiel des Massenspektrometers Waters Synapt G2-S den Aufbau eines modernen Massenspektrometers mit Ionenmobilitätsseparation.

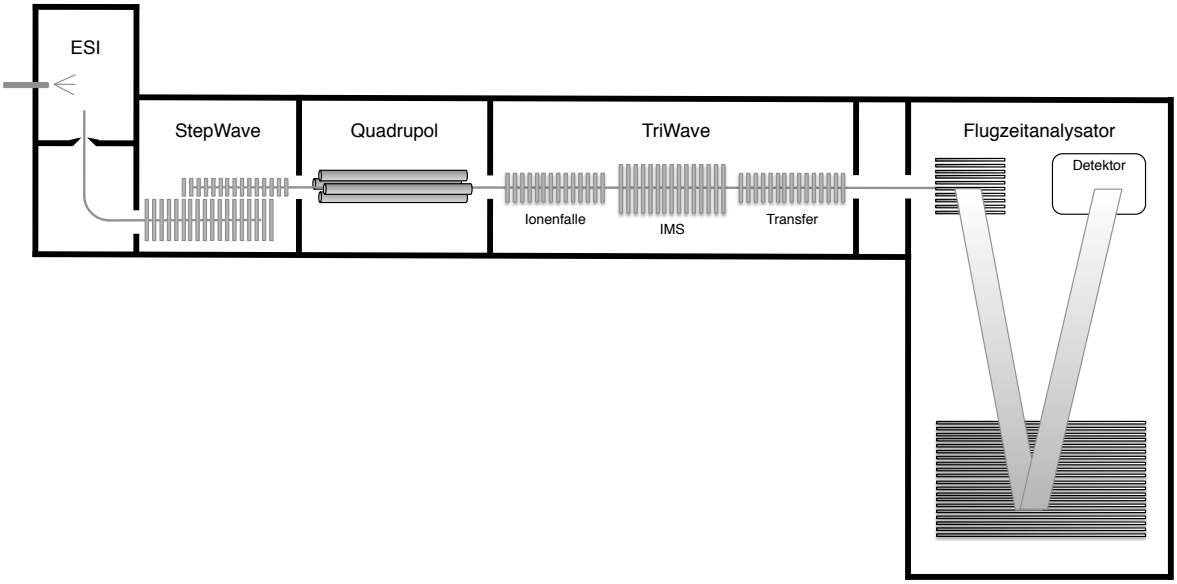

Abb. 3: Schematischer Aufbau des Waters Synapt G2-S Massenspektrometers.
Im Waters Synapt G2-S Massenspektrometer werden Quadupol, “travelling wave ion mobility mass spectrometry” (TWIMS) und Flugzeitmassenspektrometer kombiniert. Durch die Ionenquelle (hier ESI) gelangen die Analytionen in das Massenspektrometer und werden durch “StepWave” - eine Ionentransfereinheit zum Quadrupol geleitet, dem eine “TriWave”-Einheit nachgeschaltet ist. “TriWave” beinhaltet drei “Travelling-Wave-Ion-Guide” (TWIG) Geräte, die neben der Trennung nach Ionenmobilität als Ionenfalle und Kollisionszelle konfiguriert werden können. Schließlich werden die Ionen der Flugzeitanalyse zugeführt.[61]
Das Bild wurde adaptiert von Pringle et al. 2007.[61]

1.4 Reduktion der Probenkomplexität

Eine besondere Herausforderung für die MS stellt die Komplexität biologischer Proben dar. Die Ausgangsprobe kann durch Einsatz von Elektrophorese- oder Chromatographietechniken in weniger komplexe Fraktionen aufgetrennt werden, die sich auf Grund geringerer Komplexität für massenspektrometrische Untersuchungen besser eignen.

1.4.1 Fraktionierung

Für die Analyse von komplexen Proteingemischen eignen sich Methoden der 1D- und 2D-Gelelektrophorese mit der nachfolgenden Extraktion der entsprechenden Banden (bei 1D) oder Spots (bei 2D) und der anschließenden massenspektrometrischen Analyse. Überlagerungen von Gel-Banden oder -Spots sowie Detektionsgrenzen für niedrigabundante Proteine können den Einsatz dieser Techniken einschränken[62]. Weitere Einschränkungen ergeben aus der zeitaufwändigen, manuellen Extraktion der Banden bzw. Spots und deren Vorbereitung für die massenspektrometrische Analyse[63]. Deshalb werden diese Arbeitsschritte in modernen Laboren durch Einsatz spezieller Roboter automatisiert durchgeführt. Eine Alternative zum Einsatz moderner 2D-Gelelektrophorese stellen die Flüssigchromatographietechniken dar. So können Protein- oder Peptidgemische mit Hilfe der Flüssigchromatographie (LC) automatisiert in weniger komplexe Fraktionen aufgetrennt werden. Abhängig von der Komplexität der Proben werden ein- oder zweidimensionale LC-Techniken eingesetzt.

1.4.2 Online-Kopplung

Als Flüssigchromatographie mit Massenspektrometrie-Kopplung (LC-MS) wird eine Kombination der beiden Techniken bezeichnet. Insbesondere bietet die ESI eine einfache Möglichkeit der Online-Kopplung der LC an die MS[11, 64–66]. Das Ergebnis einer LC-ESI-MS-Messung sind kontinuierlich erfasste Massenspektren über die Gesamtdauer der LC-Methode. Jedes Massenspektrum korrespondiert dann mit einem Zeitpunkt der Chromatogramms. LC-ESI-MS ist heute die bevorzugte Methode der Proteomik zur Untersuchung komplexer Proteingemische oder ganzer Proteome in einer einzigen Probe.[1] Die Online-Kopplung bietet gegenüber der Fraktionierung den entscheidenden Vorteil, die eluierenden Moleküle nicht in einzelnen Fraktionen zu kumulieren, sondern führt diese direkt der massenspektrometrischen Analyse zu und reduziert dadurch die Komplexität der resultierenden Massenspektren. Zur

[1] In der Literatur wird häufig die Abkürzung LC-MS, wenn nicht näher spezifiziert, synonym zu LC-ESI-MS oder LC-ESI-MS/MS verwendet.

weiteren Reduktion der Probenkomplexität lassen sich Fraktionierungstechniken mit der Online-LC-MS kombinieren.

1.4.3 Hochleistungsflüssigkeitschromatographie

Bei LC-MS-Experimenten kommt meist die sogenannte Hochleistungsflüssigkeits-chromatographie[1] (HPLC, engl. high performance liquid chromatography) zum Einsatz. Durch evolutionäre Verbesserungen der Herstellungsprozesse und der verwendeten Materialien wurde die HPLC-Trennleistung in den letzten Jahrzehnten kontinuierlich gesteigert. Eine signifikante Steigerung der Trennleistung kann z.B. durch die Reduktion der verwendeten Partikelgröße erreicht werden. So betrug die gängige HPLC-Partikelgröße in den 1970er Jahren ca. 10 µm, in den 1980er Jahren ca. 5 µm und in den 1990er Jahren ca. 3,5 µm. Ein modernes HPLC-System arbeitet in Verbindung mit nano-ESI-MS mit niedrigen Analytmengen im Nanoliterbereich. In diesem Zusammenhang spricht man von nanoflow-HPLC. Die Reduktion der Analytmenge bei gleichbleibender, hoher analytischer Leistung wird durch eine weitere Miniaturisierung der Trennsäule ermöglicht. Eine weitere Reduktion der Partikelgröße auf 1,7 µm wurde mit der sogenannten "ultra-performance liquid chromatography" (UPLC)[2] vorgestellt[67]. Swartz et al. zeigten, dass die Verwendung von Partikelgrößen unterhalb von 2,5 µm nicht nur eine signifikante Steigerung der Trennleistung mit sich bringt, sondern dass die Trennleistung bei Steigerung der Durchflussrate und der linearen Geschwindigkeit des Analyts über weite Bereiche nicht abnimmt. Die Verwendung der UPLC-Technologie für LC-MS-Analysen bringt mit sich eine Reihe von weiteren Vorteilen gegenüber der herkömmlichen HPLC. So wird die Dauer der Experimente reduziert, gleichzeitig nimmt die Peakkapazität zu, d.h. die Zahl der über die Gradientenzeit aufgelösten Signale pro Zeiteinheit steigt.

[1] auch Hochdruckflüssigchromatographie (engl. high pressure liquid chromatography)
[2] UPLC® ist ein eingetragenes Warenzeichen der Waters Corporation

1.5 Top-down und Bottom-up Analysen

Mit Hilfe der MS können intakte Proteine mit sogenannten Top-down-Ansätzen untersucht werden. So konnten Schey et al. im Jahr 2013 in einer Imaging-Studie 50 bis 100 intakte Proteine in einzelnen Gewebeproben aus Augenlinsen, Hirn und Nieren massenspektrometrisch nachweisen. Dabei konnten mehrere modifizierte Proteine z.B. verkürzte Formen der Augenlinsenproteine identifiziert werden[68]. MS-gestützte Top-down-Ansätze eignen sich besonders für den Nachweis verkürzter Proteinformen, integraler Membranproteine oder Proteine mit posttranslationalen Modifikationen[68–70]. Aufgrund der unterschiedlichen biochemischen Eigenschaften intakter Proteine können diese nur eingeschränkt durch flüssigchromatographische Methoden separiert werden[71]. Die molekularen Massen der Proteine variieren stark und verteilen sich über mehrere Größenordnungen. Massenspektrometrische Untersuchung von Proteinen mit einem Molekulargewicht über 5000 Da wird durch mehrere technische Faktoren erschwert. Diese Proteine lassen sich schlecht ionisieren und weisen komplizierte Isotopenverteilungen sowie mehrere Ladungszustände nach der Ionisation auf. Zusätzlich wird ihre Erfassung durch den eingeschränkten dynamischen Bereich der verwendeten Massenspektrometer limitiert. Alternativ können Proteine unabhängig ihrer molekularen Masse und ihrer biochemischen Eigenschaften mit Hilfe der sogenannten Bottom-up-Methoden untersucht werden. Die Bottom-up-Techniken beschäftigen sich mit den massenspektrometrischen Untersuchungen von Proteinfragmenten, den Peptiden, welche meist mit Hilfe spezifischer Endopeptidasen erzeugt werden. Die proteolytisch generierten Peptide eignen sich auf Grund ihrer biochemischen Eigenschaften und Molekulargewichte für massenspektrometrische Analysen besser als intakte Proteine. Eine Identifikation dieser Peptide erlaubt Rückschlüsse auf die Identität der analysierten Proteine. Die am weitesten verbreitete Bottom-up-Vorgehensweise zur Proteinanalyse aus einem Gemisch ist die sogenannte Schrotschuss-Proteomik[1] (engl. shotgun proteomics). Die Schrotschuss-Proteomik fasst analytische Vorgehensweisen zusammen, bei denen ein Proteingemisch enzymatisch verdaut und die resultierenden Peptide der LC-MS zugeführt werden. Die Schrotschuss-Proteomik ermöglicht eine globale Identifikation der Proteine sowie systematische Analysen dynamischer Proteome[73]. Bereits 1999 ermöglichte die Schrotschuss-Proteomik eine Identifikation von

[1] Der Begriff der Schrotschuss-Proteomik entstand in Anlehnung an die Schrotschuss-Sequenzierung, eine molekularbiologische Methode zur Sequenzierung langer DNA-Stränge, bei der aus den untersuchten DNA-Strängen zufällig kleine Fragmente von mehreren hundert Basenpaaren erzeugt und die Fragmente sequenziert werden[72].

mehr als 100 Proteinen aus einer einzigen Probe[66]. Die Weiterentwicklung der Technik ermöglichte nach nur wenigen Jahren die Identifikation von 270 Proteinen aus einer Probe sowie Untersuchungen von ko- und posttranslationalen Proteinmodifkationen wie Serin-, Threonin- und Tyrosin-Phosphorylierung, Arginin- und Lysin-Methylierung, Lysin-Acetylierung sowie Methionin-, Tyrosin- und Tryptophan-Oxidation[74]. Heutzutage lassen sich mit Hilfe der Schrotschuss-Proteomik über 4000 Proteine in einem einzigen Experiment identifizieren[75].

1.6 Datenakquisition

1.6.1 Datenabhängige Akquisition

Bei der MS/MS werden Vorläuferionen anhand ihres *m/z* ausgewählt und gezielt fragmentiert[17, 76]. Das Prinzip der gezielten Auswahl der Vorläuferionen findet bei der sogenannten datenabhängigen Akquisition (DDA, engl. data dependant acquisition) Anwendung. Dabei wird in einem MS1-Scan für alle Analytionen ein Vorläufermassenspektrum (auch MS1-Massenspektrum) erfasst und EDV gestützt ausgewertet. Im nächsten Schritt, dem MS2-Scan, wird anhand der Auswertung des vorangegangenen MS1-Scans nach vordefinierten Regeln eine Ionenspezies ausgewählt, der Fragmentierung zugeführt und das dazugehörige Fragmentmassenspektrum (auch Produkt- oder MS2-Massenspektrum) aufgenommen. Auf diese Weise kann die erfasste Fragmentioneninformation der ausgewählten Vorläuferionenspezies eindeutig zugeordnet werden. Instrumentenabhängig werden nach einem MS1-Scan in diskreten Zeitabständen mehrere MS2-Zyklen in Serie durchgeführt. Die Kandidaten für die Fragmentierung werden in der Regel anhand ihrer Intensität ausgewählt, d.h. nur die Analytionen mit der höchsten Signalintensität werden fragmentiert. Bei der typischen Anwendung in Verbindung mit der Schrotschuss-Proteomik steht zu jedem Zeitpunkt der LC-Separation nur eine begrenzte Zeit für die datenabhängige Fragmentierung der Vorläuferionen zur Verfügung. Dies führt dazu, dass in der Praxis nach jedem MS1-Scan eine begrenzte Zahl der Fragmentierungszyklen durchgeführt werden. Üblich sind die “top 10” bzw. “top 20” Strategien der DDA, bei welchen nach jedem MS1-Scan die Fragmentierung der 10 bzw. 20 Vorläuferionen mit den höchsten Intensitäten untersucht wird. Die DDA in Verbindung mit Schrotschuss-Proteomik ermöglichte die ersten, großangelegten Charakterisierungen ganzer Proteome. So wurden im Jahr 2001 in einer Studie von Washburn et al. insgesamt 1484 Hefeproteine, davon 131 Proteine mit drei und mehr Transmembrandomänen und im Jahr 2003 in einer weiteren Studie von Peng et al. insgesamt 1504 Hefeproteine identifiziert[77, 78]. Massenspektrometer mit höherer Auflösung, Sensitivität und Scan-Geschwindigkeit erlauben mittlerweile die Detektion von mehreren tausend Proteinen mit nur einer LC-MS-Messung. So identifizierten Nagaraj et al. mit Hilfe der DDA in einer Studie über 4000 Hefeproteine[75]. Dies entspricht der Annotation von etwa 63% der offenen Leseraster des Hefegenoms und damit der möglichen Proteine des Hefeproteoms.

DDA ist die meist verwendete massenspektrometrische Akquisitionsmethode in der Proteomik. Für komplexe Proben weist die datenabhängige Auswahl der Vorläuferionen

jedoch einen stochastischen Charakter sowie eine schlechte Reproduzierbarkeit auf[79, 80]. Zusätzlich begrenzen die langen, instrumentenabhängigen Zykluszeiten die Anzahl wählbarer Vorläuferionen für die Fragmentierung und führen zur Unterabtastung (engl. undersampling) des Analyts, so dass lediglich ein Bruchteil der Ausgangsprobe analysiert wird[81]. Da bei den gängigen DDA-Anwendungen die Auswahl der Vorläuferionenspezies auf Grund ihrer Intensität stattfindet, werden vorwiegend hochabundante Ionenspezies untersucht[82]. Infolgedessen wird der dynamische Bereich detektierbarer Peptide eingeschränkt.

1.6.2 Gezielte Akquisition

Im Gegensatz zur hohen Proteomabdeckung durch DDA wird in vielen Fällen lediglich der Nachweis von wenigen bekannten Proteinen oder Peptiden in einer Probe benötigt. Sind dann die Massen der Zielionen bekannt, können die Vorläuferionen und auch die zu erfassenden Fragmentionen, die in diesem Zusammenhang als Transitionen bezeichnet werden, nach Benutzervorgaben gefiltert werden. Diese Strategie wird als gezielte Analyse oder “targeted proteomics” bezeichnet. So werden bei der Methode “single reaction monitoring” oder “selected reaction monitoring” (SRM) jeweils nur ein Präkursorion und ein Fragmention nach Benutzervorgaben erfasst. In abgewandelter Form findet bei dem sogenannten “multiple reaction monitoring” (MRM) die Selektion mehrerer Zielfragmente und bei dem “parallel reaction monitoring” (PRM) aller Fragmente Anwendung[83–85]. Für die Analyse eines Peptides ist dabei seine Detektion in MS1-Spektrum nicht notwendig, da die Kriterien für die Vorläuferionenwahl bereits vor der Messung definiert werden. Auf diese Weise können auch niedrigabundante Peptide gezielt untersucht werden. Die Vorzüge der gezielten Akquisition liegen in ihrer hohen Spezifität und der Qualität der gewonnenen Daten[86]. Da jedoch nur wenige tausend Transitionen in einer LC-MS/MS-Messung erfasst werden können[87], eignen sich die Techniken der gezielten Akquisition nicht für explorative Untersuchungen, bei denen möglichst viele Peptide eines Proteoms charakterisiert werden sollen.

1.6.3 Datenunabhängige Akquisition

Die Einschränkungen der datenabhängigen und gezielten Akquisitionsmethoden führten zur Entwicklung von Methoden, die weder die Detektion der Vorläuferionen noch eine Vorkenntnis über ihre *m/z* voraussetzen, um die Fragmentierung der Ionen einzuleiten. Diese datenunabhängigen Akquisitionsmethoden (DIA, engl. data-independent acquisition) erfassen zyklisch zu jedem Zeitpunkt der vorgeschalteten Flüssigchromatographie abwechselnd zuerst

Übersichtsmassenspektren der Vorläuferionen und daraufhin Fragmentionenmassenspektren für alle eluierenden Analyten[88]. Um die Fragmentierung aller Vorläuferionen zu erreichen, gehen DIA-Implementierungen wie DIA[89], PAcIFIC[82, 90], SWATH-MS[88] ähnlich vor, indem sie nach jeweils einem Übersichtsmassenspektrum fortlaufend mehrere Fragmentmassenspektren für Vorläuferionen erfassen, die aus verschiedenen m/z-Bereichsfenstern ausgewählt werden. Eine weitere Gruppe von DIA-Methoden wie MS^E[91] und AIF[92] erfassen zyklisch jeweils ein Übersichtsmassenspektrum gefolgt von einem Fragmentmassenspektrum für alle eluierenden Analyten. Die simultane Fragmentierung vieler oder gar aller Vorläuferionen führt zur gesteigerten Komplexität der Fragmentmassenspektren sowie zum Verlust der Zuordnung der Fragmentionen zu ihren Vorläufern[88]. Bis zu einer gewissen Komplexität können die erzeugten Fragmentmassenspektren mit vorhandenen, ursprünglich für DDA-Daten entwickelten Softwarewerkzeugen zur Peptid- und Proteinidentifikation direkt verarbeitet werden[89, 93]. Für komplexere Daten kann zunächst eine Fragment-zu-Vorläufer-Zuordnung algorithmisch auf Basis ihrer Elutionsprofile erfolgen. Diese Zuordnung erlaubt eine Konstruktion von Pseudofragmentspektren für die jeweiligen Vorläuferionen, die dann zur Peptid- und Proteinidentifikation mit vorhandenen Softwarewerkzeugen genutzt werden können[80, 94–97].

Bei der IMS-MS Variante von MS^E, der sogenannten $HDMS^E$ (engl. high definition MS^E) werden die Vorläuferionen vor der Erfassung des Übersichtsmassenspektrums nach ihrer Ionenmobilität getrennt. Im MS2-Scan werden alle Vorläuferionen ebenfalls nach ihrer Ionenmobilität getrennt, in der Kollisionszelle fragmentiert und die entsprechenden Fragmentmassenspektren anschließend erfasst. Die erfasste Information über die Ionenmobilität in MS1- und MS2-Scans vereinfacht die Zuordnung der Fragmente zu ihren Präkursoren. $UDMS^E$ (ultra definition MS^E) erweitert die $HDMS^E$-Methode um driftzeitabhängige Energieprofile für die Fragmentierung der Peptidionen. Die erhöhte Fragmentierungseffizienz ermöglicht eine signifikante Steigerung der Proteomabdeckung in Bereiche, die bisher ausschließlich mit datenabhängigen Methoden erreicht werden konnten[98].

1.7 Datenanalyse

1.7.1 Rohdaten

Messwerte eines LC-MS-Experiments werden zunächst als Rohdaten in proprietären instrumentenspezifischen Datenformaten erfasst. Neben den eigentlichen Messwerten können die Rohdaten für die spätere Datenverarbeitung wichtige Informationen enthalten: Experimentbedingungen, Instrumenteneinstellungen, Kalibrierungsdaten des Lockmassenstandards sowie weitere Metainformationen. Beispielsweise umfassen die Rohdaten eines Waters Synapt G2S Massenspektrometers für ein HDMS^E-Experiment die Retentionszeit-, m/z- und Ionenmobilitätswerte der MS1- und MS2-Scans, die m/z-Werte des Lockmassenstandards sowie Metainformationen über das Experiment und Instrumenteneinstellungen. Rohdaten-Dateiformate verschiedener Instrumentenhersteller sind nicht untereinander kompatibel und setzen häufig auf die Speicherplatz schonende binäre Datenspeicherung. Dies erschwert jedoch den Zugriff auf die Daten außerhalb herstellerspezifischer Software. Um den Zugriff auf massenspektrometrische Daten zu vereinfachen, wurden offene Dateiformate entwickelt. Für die Speicherung und den Austausch massenspektrometrischer Proteomik-Rohdaten entwickelte die Proteomics Standards Initiative (PSI) der Human Proteome Organization (HUPO) das Format mzData[99]. Am Seattle Proteome Center wurde das Format mzXML entwickelt[100]. Beide Organisationen versuchten die Entwicklung der eigenen Dateiformate zu harmonisieren und entwickelten gemeinsam ein weiteres Format - mzML[101].

1.7.2 Signalverarbeitung

Enthalten die erfassten Rohdaten Messwerte des Lockmassenstandards, wird zunächst der systematische Fehler der Ionenmassen korrigiert, der durch Instrumentendrift hervorgerufen wird. Welche weiteren Einzelschritte die Verarbeitung der Rohdaten benötigt, hängt von den verwendeten Algorithmen ab.

LC-MS-Daten können ein hohes Maß an chemischem und technisch bedingtem Hintergrundrauschen enthalten[102]. Um das Rauschen effektiv zu unterdrücken, werden unterschiedliche Ansätze verfolgt[103–105]. Am häufigsten findet jedoch die Methode der Hintergrundsubtraktion (engl. background-subtraction) Anwendung[106]. Im einfachsten Fall entfernt die Hintergrundsubtraktion alle Signale, deren Intensität unterhalb eines vordefinierten Schwellenwertes liegt. Dadurch wird jedoch nicht das mit den verbleibenden

Signalen gefaltete Rauschen beseitigt. So kann es abhängig vom später verwendeten Algorithmus für die Signaldetektion notwendig sein, die LC-MS Signale entlang der Retentionszeit zu glätten, beispielsweise durch den Einsatz des Savitzky-Golay-Filters[107].

Die verbleibenden, geglätteten Signale werden im Rahmen der so genannten Feature-Detektion nach typischen Signalmustern der Peptide durchsucht. Als Features werden dabei die zusammenhängenden MS1-Signalgruppen bezeichnet, die durch Isotopenmuster der Peptidionen verursacht werden. Abbildung 4 zeigt ein typisches Isotopenmuster eines tryptischen Peptides. Da Teile der Signalgruppen unterschiedlicher Features überlappen können, werden Bestandteile der Ausgangssignale durch spezielle Dekonvolutionsalgorithmen wiederhergestellt[102, 108, 109]. Bei MS/MS-Daten werden den einzelnen Features entsprechende Fragmentmassenspektren zugeordnet. Für quantitative Untersuchungen können die detektierten Features anhand ihrer Eigenschaften, z.B. durch die Integration der Intensitäten der jeweiligen Signalgruppe, quantifiziert werden.

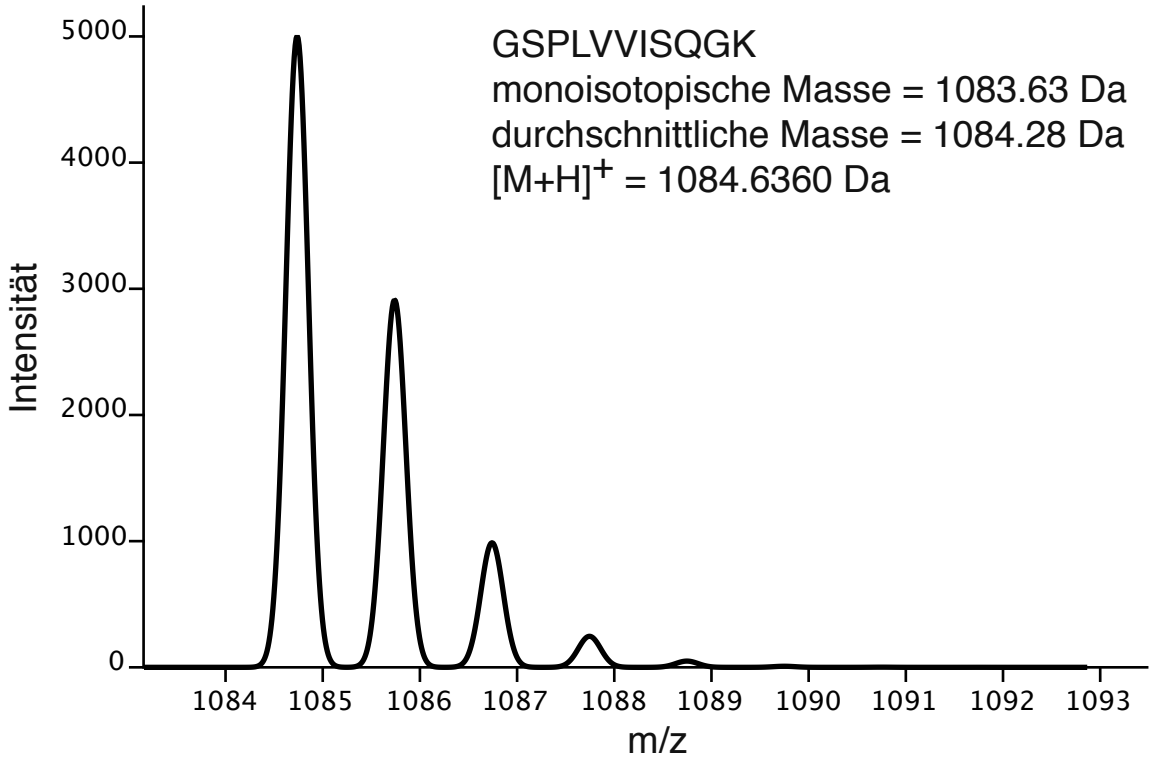

Abb. 4: Peptid-Isotopenmuster.
Dargestellt wird das Massenspektrum des theoretisch-berechneten Isotopenmusters eines einfach geladenen, tryptischen Peptides.
Das Bild wurde mit Envelope[110] erstellt.

1.7.3 Peptid- und Proteinidentifikation

Im weiteren Verlauf der Datenverarbeitung lassen sich die detektierten Features anhand ihrer Eigenschaften wie der Retentionszeit, der Masse, der Ladung, der Ionenmobilität, des Fragmentierungsmusters und der eventuell vorhandenen De-Novo-Sequenzinformation als bestimmte Peptide identifizieren.

Typischerweise setzt man bei der Proteinidentifikation Datenbanken mit bekannten Proteinsequenzen ein, wie z.B. “universal protein database” (Uniprot, http://www.uniprot.org)[111] oder “international protein index” (IPI, http://www.ebi.ac.uk/IPI)[112]. Die Identifikation erfolgt dann durch spezielle Datenbanksuchalgorithmen, die zunächst die in der jeweiligen Suchdatenbank enthaltenen Proteinsequenzen abhängig vom Experiment *in silico* verdauen und anschließend unterschiedliche Eigenschaften der Features mit den theoretischen Eigenschaften der *in silico*-Peptide vergleichen. Die besten statistischen Übereinstimmungen werden als Identifikationen der Features verwendet. Die identifizierten Peptide lassen nun auf die Identität der Ausgangsproteine schließen. Die bekanntesten Datenbanksuchalgorithmen sind Sequest[113], MASCOT[114], OMSSA[115], X!Tandem[116], MS-GF+[117] sowie der speziell für MS^E/$HDMS^E$-Daten angepasste Algorithmus PLGS Identity^E[96, 118].

Die Qualitätssicherung der datenbankbasierten Peptid- und Proteinidentifikation kann anhand der geschätzten “false discovery rate” (FDR) erfolgen[119]. Für die Schätzung der FDR verwenden die Datenbanksuchalgorithmen als Eingabe die sogenannten Target-Decoy-Datenbanken, die neben den gesuchten Target-Proteinsequenzen einen Anteil an Decoy-Proteinsequenzen enthalten, die nicht in der Natur vorkommen[120]. Der Decoy-Anteil kann dabei durch eine Umkehrung oder eine zufallsbehaftete Durchmischung der Aminosäuresequenzen der vorliegenden Zielproteine generiert werden. Die Ergebnisse der Target-Decoy-Datenbanksuche werden auf einen bestimmten, benutzerdefinierten FDR-Schwellenwert, üblicherweise 1% bis 5% beschränkt.

Eine Alternative zur Proteinidentifikation auf Basis der Datenbanksuche bietet der Abgleich von erfassten Massenspektren mit speziellen Spektrenbibliotheken (engl. spectral libraries), in welchen die zuvor experimentell beobachteten Fragment-zu-Vorläufer-Zuordnungen samt Peptididentifikation gespeichert sind[88]. Die erfassten Massenspektren korrellieren mit den entsprechenden Massenspektren der Bibliothek und liefern dadurch die Identifikation der Peptide und Proteine. Als Konsequenz wird jedoch die Identifikation auf den Umfang und Stand der verwendeten Spektrenbibliothek eingeschränkt.

1.7.4 Von der Proteinidentifikation zur Quantifizierung

Massenspektrometrie galt in der Proteomik lange als ein Werkzeug zur reinen Proteinidentifikation. Zur Klärung vieler biologischen Fragestellungen wird jedoch neben dem qualitativen Nachweis der Proteine oft zusätzlich die quantitative Information benötigt. So bedarf die Erforschung biologischer Funktionen der Proteine, der Rollen ihrer Modifikationen für Signalmechanismen in Zellen und der spezifischen Biomarker für Krankheiten der Ermittlung quantitativer Informationen unter unterschiedlichen experimentellen Bedingungen[121]. Abhängig von der Fragestellung unterscheidet sich die quantitative Analyse. So lässt sich entweder die absolute Menge eines Proteins in der Probe oder das relative Verhältnis der Proteinmengen zwischen zwei Proben bestimmen[122]. Die absolute Quantifizierung von Proteinen erlaubt Aussagen über konkrete Proteinmengen, etwa die Menge eines Biomarkers im Serum in ng/ml oder die Zahl der Kopien eines Proteins pro Zelle. Die relative Proteinquantifizierung beschreibt das Expressionsniveau einzelner Proteine als ein vergleichendes Verhältnis zwischen den Proteomen aus unterschiedlichen, experimentellen oder biologischen Konditionen. Dabei wird die Proteinmenge in einer Probe als Vielfaches der Menge des korrespondierenden Proteins in einer anderen Probe bestimmt. Sind absolute Mengen der Proteine bekannt, kann ihr relatives Verhältnis einfach berechnet werden.

Die MS-gestützte Proteomik verfügt über zwei grundlegende Vorgehensweisen zur Quantifizierung von Proteinen[122]. Bei der markierungsfreien oder labelfreien Quantifizierung ermöglicht ein Vergleich der Signale korrespondierender Peptide unter unterschiedlichen Bedingungen eine Schätzung der relativen Mengenverhältnisse eines Proteins zwischen mehreren Proteomen. Alternativ kann durch eine Molekülmarkierung der untersuchten Proteome mit stabilen Isotopen eine genaue Erfassung der relativen Mengenverhältnisse erfolgen. Diese Vorgehensweisen werden in den Kapiteln 1.7.5 und 1.7.6 näher erläutert.

Die Proteinquantifizierung stellt eine der größten Herausforderungen der modernen Analytik dar. Lediglich ein Teil des Proteoms ist für die MS-gestützte Proteinidentifikation zugänglich, zudem können nicht alle identifizierten Proteine zuverlässig quantifiziert werden[123]. Dieser Zusammenhang wird in Abbildung 5 gezeigt.

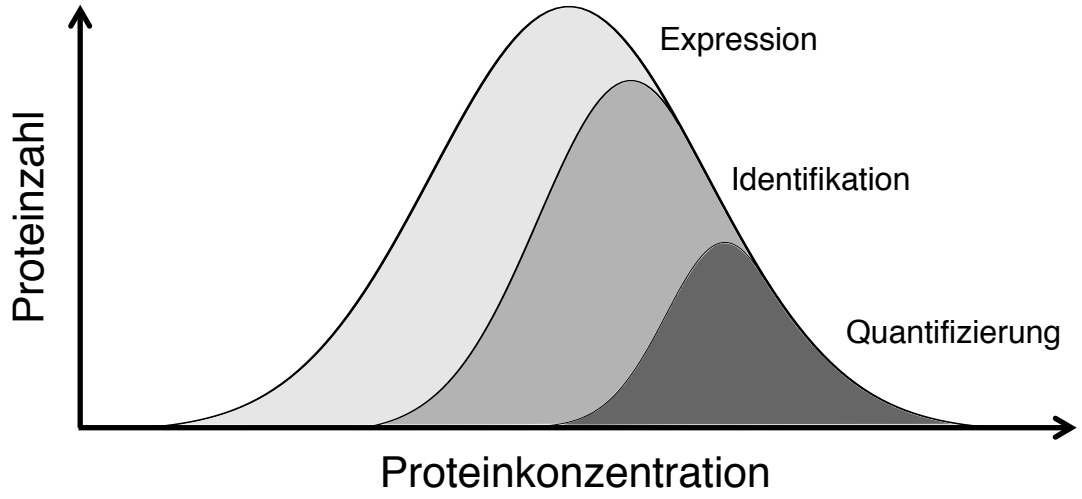

Abb. 5: Identifizierbare und Quantifizierbare Anteile eines Proteoms.
Anteile eines Proteoms, die durch MS-gestützte Proteomik identifiziert oder quantifiziert werden können, werden schematisch dargestellt. Nur ein Teil der exprimierten Proteine kann durch massenspektrometrische Methoden identifiziert werden. Durch Einschränkungen der Datenqualität kann nur ein Teil der identifizierten Proteine zuverlässig quantifiziert werden.[123]
Das Bild wurde adaptiert von Bantscheff et al., 2007.[123]

1.7.5 Proteinquantifizierung mit Markierungstechniken

Die gängigen Markierungstechniken zur massenspektrometrischen Proteinquantifizierung wie "isotope-coded affinity tag" (ICAT)[124] und "stable isotope labeling with amino acids in cell culture" (SILAC)[125] gewährleisten eine hohe Genauigkeit und finden routinemässigen Einsatz. Bei diesen Techniken werden bekannte Massenunterschiede der Markierung durch stabile Isotope mit identischen chemischen und physikalischen Eigenschaften ausgenutzt, um eine berechenbare Massendifferenz zwischen korrespondierenden Peptiden untersuchter Proteome zu induzieren[122]. Hierfür werden die Proteinproben entweder wie bei ICAT chemisch oder wie bei SILAC metabolisch durch stabile Isotopen markiert. Markierte Proteome werden zusammengeführt und gemeinsam massenspektrometrisch analysiert. Die bekannte Massendifferenz der verwendeten Isotopenmarkierungen erlaubt eine Unterscheidung zwischen den Peptidsignalen unterschiedlicher Proteome und eine genaue Berechnung der relativen Mengenverhältnisse direkt aus dem erfassten Massenspektrum. Neben der relativen Quantifizierung kann die Markierung mit stabilen Isotopen für die absolute Quantifizierung genutzt werden, z.B. durch Verwendung synthetischer mit ^{13}C-Kohlenstoffisotopen markierten Standardpeptide[126]. Bei der Methode "isobaric tags for relative and absolute quantitation" (iTRAQ) werden die tryptischen Peptide der untersuchten Proben mit speziellen Molekülen (Tags) markiert[127]. Ein iTRAQ-Tag besteht aus einer

Reporter- und einer Ausgleichsgruppe. Die Molekularmasse der Reportergruppe wird zur Markierung einzelner Proben unterschiedlich gewählt und durch Anpassung der Ausgleichsgruppe auf eine gemeinsame Molekularmasse von 145 Da gebracht. Die so markierten Peptide der bis zu acht Proben werden gemischt und gemeinsam mittels MS/MS analysiert. Während der Fragmentierung werden die Reporter-Ionen abgespalten. Da ihre Massen bekannt sind (113 Da bis 119 Da und 121 Da), erlaubt der Vergleich ihrer Signalintensitäten eine relative Quantifizierung der entsprechenden Peptide.

Massenspektrometrische Proteinquantifizierung mittels Markierungstechniken kann bei der Ermittlung präziser Proteinfunktionen und der Verfolgung zeitlicher Änderungen von Proteinmengen eingesetzt werden[122]. Die Markierungsmethoden ermöglichen eine genaue Quantifizierung der Proteine. Durch die begrenzte Anzahl der für die jeweilige Methode verfügbaren unterschiedlichen Markierungen wird die Anzahl der Proben limitiert, die in einem Experiment quantitativ untersucht werden können. Die Komplexität der massenspektrometrisch erfassten Daten steigt bei einem Experiment mit der Anzahl der markierten Proben, da sie gemeinsam untersucht werden. Die Datenanalyse bedarf speziell auf die verwendete Markierung abgestimmter Auswertungssoftware.

1.7.6 Labelfreie Proteinquantifizierung

Einschränkungen der Isotopenmarkierung bewirken ein wachsendes Interesse an den markierungsfreien Methoden. Bei der labelfreien Proteinquantifizierung (LFQ, engl. label free quantification) werden mehrere Proteome in jeweils eigenständigen massenspektrometrischen Experimenten unabhängig voneinander untersucht. Von den erfassten Signalen wird bei der Datenauswertung auf statistischer Basis quantitative Information über die identifizierten Peptide und Proteine abgeleitet.

Allet et al. postulierten, dass die bei der Proteinidentifikation üblichen Identifikationsscores mit der Proteinabundanz korrelieren können[128]. Weitere Zusammenhänge zwischen der Proteinabundanz und solchen statistischen Merkmalen, wie der erreichten Sequenzabdeckung, der Anzahl identifizierter Peptide (engl. peptide count) oder erfasster Fragmentspektren (engl. spectral count) wurden aufgezeigt, die sich im Rahmen der labelfreien Schrotschuss-Proteomik zur Ermittlung quantitativer Informationen über differentiell exprimierte oder koexprimierte Proteine eignen[129]. Sowohl die Größe eines Proteins als auch seine Abundanz beeinflussen die Zahl der identifizierten Peptide. Die Proteinabundanz kann durch Proteinabundanzindices (PAI) geschätzt werden, z.B. als Verhältnis der Zahl identifizierter Peptide zu der Zahl

theoretisch erfassbarer tryptischer Peptide eines Proteins[130]. Die logarithmische Natur der Beziehung zwischen der Zahl identifizierter Peptide und der Proteinmenge führte zur nachfolgenden Definition eines exponentiell modifizierten Proteinabundanzindizes (emPAI)[131].

In der LC-ESI-MS besteht ein Zusammenhang zwischen der Analytkonzentration in der Probe und der massenspektrometrisch erfassten Ionenintensität[32]. Dabei wird die Menge einer massenspektrometrisch analysierten Molekülspezies als Signalintensität erfasst. Die erfassten Signalintensitäten können zur Quantifizierung der Proteine herangezogen werden. Bei einer LC-MS-Messung kann die MS1-Signalintensität eines chromatographisch eluierenden Peptides als Funktion der Retentionszeit dargestellt werden. Die Fläche unter einer solchen Kurve, die als “extracted ion chromatogram” oder “extracted ion current” (XIC) bezeichnet wird, hat jeweils für das gleiche Peptid und unter gleichen experimentellen Bedingungen eine lineare Beziehung zur Menge des Peptides. Durch den Vergleich dieser integrierten Signalintensitäten zwischen den Messungen kann auf das relative Mengenverhältnis der Ausgangsproteine geschlossen werden[122].

1.7.7 Absolute Quantifizierung

Neben dem relativen Vergleich der Proteinmengen zwischen unterschiedlichen experimentellen Bedingungen kann die Ermittlung von absoluten Proteinmengen für bestimmte biologische Fragestellungen von besonderem Interesse sein. Wie bereits im Kapitel 1.7.5 erwähnt wurde, ergibt sich eine Möglichkeit der absoluten Proteinquantifizierung aus der Verwendung von synthetischen Peptiden, die mit ^{13}C-Isotopen (alternativ mit ^{15}N- oder ^{2}H-Isotopen) markiert wurden[126].

Durch die Zugabe definierter Mengen synthetischer, schwerer Peptide zum Analyt, können absolute Mengen der entsprechenden, nicht-markierten Peptide durch das Verhältnis der MS1-Intensitäten direkt berechnet werden. Die auf diese Weise berechneten, absoluten Mengen der Peptide erlauben Rückschlüsse auf die absoluten Mengen der entsprechenden Proteine. In Verbindung mit der gezielten Akquisition bietet diese Methode eine hohe Genauigkeit und Sensitivität der quantitativen Analyse[132], wird jedoch durch den zeitlichen und finanziellen Aufwand bei der Herstellung synthetischer Peptide eingeschränkt. Sie eignet sich deshalb besonders für Experimente, bei denen nur wenige Proteine quantifiziert werden. Eine Steigerung der Anzahl der Proteine, die in einem Experiment quantifiziert werden können, und bei der Qualität der Quantifizierung wird mit der Methode QconCAT (Abgeleitet von

engl. quantification concatamer) erreicht[133, 134]. Bei dieser Methode wird ein synthetisches Gen erzeugt, das proteotypische Peptide aller zu quantifizierenden Proteine kodiert. Das Gen wird synthetisiert und mit einem Expressionsvektor in ein Bakterium eingebracht. In einem Medium mit Aminosäuren, die mit stabilen Isotopen markiert wurden, wird das künstliche Protein exprimiert. Nach einem enzymatischem Verdau erhält man synthetische Peptide, die als Standard für eine gezielte Quantifizierung der Zielproteine eingesetzt werden. Diese Methode erlaubt eine zuverlässige absolute Quantifizierung von bis etwa einhundert Proteinen. Alternativ zur Zugabe synthetischer Standardpeptide können intakte Proteine als Standard genutzt werden. Die absoluten Proteinmengen werden dann durch die Relation der mit einer geeigneten relativen Quantifizierungsmethode ermittelten Proteinmengen zum eingebrachten Standard ermittelt. Besonders Quantifizierungsmethoden auf Basis der Signalintensitäten tryptischer Peptide eignen sich für die absolute Proteinquantifizierung. Im einfachsten Fall kann die Proteinmenge durch die Summe der Intensitäten oder die durchschnittliche Intensität aller identifizierten Peptide eines Proteins angegeben werden. Alternativ kann mit einem sogenannten "extracted ion intensity-based protein abundance index" (xPAI)[135], auch als Top3-Methode bekannt[136], die Proteinabundanz auf Basis der durchschnittlichen Signalintensität der drei meist abundanten Peptide des jeweiligen Proteins geschätzt werden. Für die xPAI bzw. Top3-Werte wurde in unabhängigen Studien hohe Korrelation mit den entsprechenden Proteinkonzentrationen in den Proben nachgewiesen[135–137]. Mit iBAQ (intensity based absolute quantification) stellten Schwanhaußer et al. eine weitere vielversprechende Methode für die absolute Proteinquantifizierung vor. iBAQ beruht auf dem Verhältnis der Summe der Intensitäten aller identifizierten Peptide zu der Zahl aller theoretisch detektierbaren tryptischen Peptide eines Proteins[138]. Das Verhältnis des ermittelten Top3 oder iBAQ Wertes eines beliebigen identifizierten Proteins zu dem entsprechend ermittelten Top3 oder iBAQ Wert des Standardproteins ermöglicht die Berechnung der absoluten Konzentration bzw. der absoluten Menge des jeweiligen Proteins in der Probe. Die beschriebenen Methoden weisen eine vergleichbare, hohe Eignung für die massenspektrometrische Ermittlung absoluter Proteinmengen auf[139].

1.8 Spezifische Analyse der MS^E/HDMS^E/UDMS^E-Daten

Bei einem typischen LFQ-LC-MS-Experiment wird die differentielle Proteinexpression zwischen mehreren Proteomen untersucht. Dazu werden die Proteome unabhängig voneinander massenspektrometrisch analysiert. Zur statistischen Absicherung wird jede LC-MS-Messung mehrfach wiederholt. Üblich ist die Erfassung gleicher Anzahl, etwa drei bis fünf, technischer Replikate von jeder Probe eines Experiments. Nach der Durchführung aller LC-MS-Messungen sowie den ersten Datenverarbeitungsschritten (s. Kapitel 1.7.2), z.B. durch den Einsatz der proprietären Software ProteinLynx Global Server^{TM} (PLGS), stehen bei einem solchen Experiment zunächst mehrere unabhängige Datensätze mit Informationen über die MS1- und MS2-Signale, die identifizierten Peptide und Proteine zur Verfügung. Für die Ermittlung der differentiellen Proteinexpression müssen die bis dahin unabhängigen Datensätze bei einer weiterführenden Datenanalyse gemeinsam betrachtet werden, mit dem Ziel relative oder absolute Mengen der in unterschiedlichen Messungen identifizierten Proteine gegenüberzustellen.

1.8.1 Herausforderungen

Über die Gesamtdauer eines LFQ-LC-MS-Experiments können experimentelle Bedingungen innerhalb einzelner Messvorgänge und zwischen ihnen variieren. Durch unterschiedliche Einflüsse können die erfassten Daten lückenhaft sein und systematische Fehler enthalten. Beispielsweise kann durch Luftdruck- und Temperaturunterschiede, durch Unterschiede in der chemischen Zusammensetzung des Analyts sowie durch Altern der Trennsäule die Stabilität der chromatographischen Methode beeinträchtigt werden und nicht-lineare Verschiebungen der Retentionszeit hervorrufen[140]. Die ESI hat einen direkten Einfluss auf die massenspektrometrisch erfassten Signale (s. Kapitel 1.3.1). Instabilitäten des Elektrosprays äussern sich in komplexen systematischen Fehlern der gemessenen Signalintensitäten. Zusätzlich werden diese durch die chemische Zusammensetzung der Probe und die Instrumentendrift beeinflusst. Komplexe Zusammenhänge und die Varianz in den erfassten Daten führen eine ganze Reihe von spezifischen Problemen ein. So können folgende Effekte beobachtet werden: a) Gleiche Peptidspezies können in unterschiedlichen Messungen zu unterschiedlichen Retentionszeiten eluieren. b) Gleiche Analytmengen können zu unterschiedlichen Zeitpunkten in unterschiedlichen Signalintensitäten resultieren. c) Gleiche Peptidspezies können in unterschiedlichen Messungen unterschiedliche Signalausprägungen

hervorrufen und werden in Grenzfällen als mehrere unabhängige Signale erkannt oder in einzelnen Messungen gar nicht detektiert[141]. Dies führt auch dazu, dass abweichende Fragmentierungsinformationen für gleiche Vorläuferionenspezies erfasst werden. In Folge dessen können die Daten unterschiedlicher LC-MS-Messungen bei der Peptid- und Proteinidentifikation durch Datenbanksuchalgorithmen entsprechend unterschiedlich interpretiert werden. Der Nachweis einzelner Peptide kann in einzelnen Messungen somit gänzlich fehlen. So kann ein Protein in unterschiedlichen Messungen des Experiments durch unterschiedliche Peptidsequenzen identifiziert werden. In einzelnen Messungen kann der Nachweis eines Proteins dann nur unzureichend sein, oder auch fehlen. Die Schrotschuss-Proteomik erfasst das Proteom auf Peptidebene. Der Nachweis der Proteine erfolgt über Rückschlüsse von den erfassten Peptidsequenzen. Einem Protein können dabei mehrere Peptide zugeordnet sein. Ein Peptid kann gleichzeitig auch verschiedenen Proteinen zugeordnet werden. Diese uneindeutige Peptid-Protein-Beziehung wird als das Proteininferenz-Problem bezeichnet[142]. Durch das Proteininferenz-Problem können komplexe Netzwerke an Peptid-Protein-Beziehungen entstehen. Wenn in einem solchen Netzwerk eindeutige Hinweise auf einzelne Proteine fehlen, spricht man sinngemäß von der Identifikation einer Proteingruppe. Dies ist v.a. bei homologen Proteinen oft der Fall. Das Problem der lückenhaften Identifikation der Peptide und Proteine wird durch das Proteininferenz-Problem in seiner Wirkung verstärkt und erschwert zusätzlich die Proteinquantifizierung.

1.8.2 Stand der Technik

DIA mit den Methoden MS^E sowie deren Ionenmobilitätsvarianten $HDMS^E$ und $UDMS^E$ ermöglicht bereits Proteomabdeckungen, die bis dato nur mit Einsatz von DDA-Methoden möglich waren. Im Gegensatz zu DDA behalten dabei die DIA-Methoden ihren quantitativen Charakter gleichermaßen für hoch- und niedrig-abundante Proteine. Auf Grund der hohen Komplexität der DIA-Daten spielen die Algorithmen für die Datenanalyse eine entscheidende Rolle. Für die noch junge MS^E/$HDMS^E$-Akquisitionstechnik existieren bislang nur wenige Softwarelösungen, die diese Art von Daten verarbeiten können. Die Software des Instrumentenherstellers Waters PLGS integriert Algorithmen für den grundlegenden Umgang mit MS^E/$HDMS^E$-Daten: Apex3D/Apex4D für die Signalverarbeitung, Peptide3D für die Feature-Detektion sowie $Identity^E$ für die Peptid- und Proteinidentifikation[118].

2005 stellten Silva et al. Expression Informatics vor - den ersten Workflow zur quantitativen Analyse labelfreier MS^E-Daten[91]. Der Workflow beinhaltet eine Clustering-Analyse der "accurate mass retention time pairs" (AMRT, auch "exact mass retention time pairs", EMRT)[1] mit der nachfolgenden globalen Normalisierung der Intensitäten und der statistischen Analyse. Dabei beinhaltet die Clustering-Analyse das Retentionszeitalignment und Feature-Clustering zwischen mehreren LC-MS-Messungen eines Experiments. Die Normalisierung führt globale Korrekturen der Feature-Intensitäten zum Ausgleich der Fluktuationen von Injektionsvolumina durch und normalisiert die Intensitäten aller EMRTs anhand injizierter interner Standardpeptide. Expression Informatics wurde als ein kommerziell verfügbares Modul mit dem Namen $Expression^E$ in die Software PLGS integriert. Die Integration von Expression Informatics in PLGS sorgte für eine weite Verbreitung dieses Analyseworkflows. Als Teil von PLGS vereinfachte Expression Informatics die Datenanalyse für den Benutzer und reduzierte den Bedarf an Fremdsoftware für die Post-PLGS-Analyse der MS^E-Daten. Durch den kommerziellen Charakter der Software wurden nicht alle Details der verwendeten Algorithmen und ihrer konkreten Implementierung offengelegt. Nur wenige Parameter der Analyse können durch den Benutzer geändert werden. Dadurch wird die Nachvollziehbarkeit der Analyseergebnisse eingeschränkt. Expression Informatics behandelt nicht das Problem der Proteininferenz. Die modernen Akquisitionsmethoden auf Basis der IMS-MS $HDMS^E$ und $UDMS^E$ werden von Expression Informatics nicht vollständig unterstützt. Hohe Anforderungen an die Ausführungsumgebung schränken den Einsatz von Expression Informatics auf Datenanalysen von maximal etwa 15 bis 20 LC-MS-Messungen in einem Experiment ein.

Eine weitere kommerzielle Lösung wird von Nonlinear Dynamics (http://www.nonlinear.com/) einer Tochtergesellschaft der Waters Corporation in Form der Software Progenesis QI for Proteomics (Progenesis QIP) angeboten. Progenesis QIP stellt eine vollintegrierte Lösung für die Analyse labelfreier MS^E/$HDMS^E$-Proteomikdaten dar. Die Software integriert Schritte der Signalprozessierung auf Basis der aus PLGS bekannten Algorithmen. Die weiterführende Analyse der Daten beinhaltet nicht näher spezifizierte Algorithmen für das Retentionszeitalignment, eine messungsübergreifende Feature-Detektion und Featurequantifizierung, die Peptid- und Proteinidentifikation, mehrere Methoden der Proteinquantifizierung, Werkzeuge zur Qualitätssicherung der Analyse in Form

[1] Ein AMRT/EMRT wird durch die monoisotopische Masse eines auf eine einfache Ladung reduzierten Peptidions und durch die Retentionszeit definiert, zu der dieses Peptid eluiert[143].

unterschiedlicher Filter und Visualisierungen sowie Möglichkeiten zum Export der Ergebnisse. Bei der Proteinquantifizierung umgeht Progenesis QIP das Problem der Proteininferenz durch ausschließliche Verwendung der Peptide, die einzelnen Proteinen oder Proteinfamilien eindeutig zugeordnet wurden. Die Software ist in der Lage massenspektrometrische Daten verschiedener Instrumentenhersteller zu analysieren. Es werden mehrere sowohl datenabhängige als auch datenunabhängige Akquisitionsstrategien unterstützt.

Mit dem R-Paket synapter wurde eine freie Lösung speziell für die Analyse labelfreier MS^E/$HDMS^E$-Proteomikdaten von Bond et al. entwickelt und im Sommer 2013 vorgestellt[144]. Der Workflow von synapter ermöglicht eine post-PLGS Datenanalyse und beinhaltet Algorithmen für die benutzerdefinierte Filterung der Daten, das Retentionszeitalignment sowie den Transfer von Identifikationen zwischen einzelnen LC-MS-Messungen des Experiments mit dem Ziel die Anzahl der Peptid- und Proteinidentifikationen im Experiment bei Einhaltung einer benutzerdefinierten FDR zu maximieren. Als R-Paket kann synapter für benutzerdefinierte Datenanalysen innerhalb der R-Umgebung flexibel eingesetzt werden. Innerhalb eines Projektes kann synapter mehrere mit unterschiedlichen Methoden akquirierte LC-MS-Messungen kombinieren. Der Entwickler empfiehlt eine Kombination von MS^E und $HDMS^E$-Messungen so, dass die $HDMS^E$-Daten für die Maximierung der Identifikationen genutzt werden, während die MS^E-Daten für die Proteinquantifizierung mit der Top3-Methode verwendet werden. Durch die Maximierung von Identifikationen eignet sich synapter besonders gut für qualitative Experimente. Die synapter-Workflow beinhaltet keine Normalisierungsalgorithmen und behandelt nicht das Problem der Proteininferenz. Eine weitere Hürde für den Endanwender stellt die Notwendigkeit der spezifischen R-Programmierung für jedes Experiment.

Neben den oben beschriebenen spezialisierten Lösungen ist eine Analyse der labelfreien MS^E-Daten mit Hilfe der Analysepipelines SuperHirn[145] oder OpenMS/TOPP[146–148] zumindest theoretisch möglich. Diese Lösungen erfordern eine Konvertierung der Rohdaten in spezifische Datenformate wie mzXML und mzML. Erschwerend kommt hinzu, dass bis dato die komplexen MS^E/$HDMS^E$/$UDMS^E$-Daten sich nur mit Hilfe der Waters-Algorithmen für die Peptid- und Proteinidentifikation, d.h. durch PLGS oder Progenesis QIP vollständig auswerten lassen[149].

1.9 Zielsetzung

Die Core-Facility für Massenspektrometrie am Institut für Immunologie an der Universitätsmedizin der Johannes Gutenberg-Universität Mainz setzt mehrere Massenspektrometer des Herstellers Waters Corporation, Modellreihen Premier Q-TOF, Xevo G2 und Synapt G2-S ein. Diese werden hauptsächlich zur Erfassung labelfreier Proteomikdaten unter Einsatz der MS^E/$HDMS^E$/$UDMS^E$-Akquisitionmethoden und der Software PLGS verwendet.

Mit dieser Arbeit sollte die differentielle, quantitative Analyse der labelfreien MS^E/$HDMS^E$/$UDMS^E$-Daten durch Entwicklung einer geeigneten Analysestrategie verbessert werden, die die im Kapitel 1.8.1 beschriebenen, datenspezifischen Probleme löst. Hierfür sollten PLGS-Ergebnisse mehrerer LC-MS-Messungen eines Experiments in einer geeigneter Datenstruktur zusammengeführt werden, die einen schnellen, messungsübergreifenden Datenzugriff ermöglicht. Die weiterführende Datenanalyse sollte unter Einhaltung einheitlicher Qualitätskriterien messungsübergreifend eine Korrektur der potentiellen Fluktuationen von Retentionszeiten vornehmen und korrespondierende Features aus unterschiedlichen Messungen einander zuordnen sowie systematische Fehler der Signalintensitäten korrigieren, Identifikationslücken schließen und potentielle Fehlidentifikationen reduzieren, uneindeutige Peptid-Protein-Zuordnungen auflösen und eine absolute Quantifizierung der Proteine durchführen. Einzelschritte der Datenanalyse sollten in Form einer Analysepipeline implementiert werden. Die Gesamteffektivität der entwickelten Analysestrategie sollte evaluiert und mit Analyseworkflows von synapter und Progenesis QIP verglichen werden.

2 Material und Methoden

2.1 Chemikalien

Soweit nicht anders vermerkt, wurden alle Chemikalien von der Firma Sigma-Aldrich Chemie GmbH (Taufkirchen bei München) und alle Lösungsmittel (Ultra LC-MS Grade) von der Firma Carl Roth GmbH und Co. KG (Karlsruhe) bezogen.

2.2 Testdatensätze

2.2.1 HeLa-Proteom

Epithelzellen eines Zervixkarzinoms (HeLa-Zellen) wurden vom Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (Braunschweig) bezogen. Die HeLa-Zellen wurden mit der Methode der filter-gestützten Probenvorbereitung (FASP, engl. filter-aided sample preparation)[150] lysiert und verdaut (s. Kapitel 2.3). Der HeLa-Proteinverdau wurde mit 0,1%iger Ameisensäure (v/v in H_2O) zu einer Peptidkonzentration von 500 ng/µL verdünnt. Zusätzlich wurde dem HeLa-Proteom zu 20 fmol/µL der Proteinverdaustandard Enolase 1 (*Saccharomyces cerevisiae*, MassPREP™, Waters Corporation, Eschborn) hinzugegeben. Jeweils 200 ng HeLa-Proteinverdau wurden einer LC-MS-Analyse (s. Kapitel 2.4) mit 90 min Gradientenzeit und einer Akquisition mit MS^E-, $HDMS^E$- und $UDMS^E$-Methoden unterzogen (HeLa-Datensätze MS^E, $HDMS^E$ und $UDMS^E$). Zusätzlich wurden 300 ng HeLa-Proteinverdau mit 180 min Gradientenzeit mit der Methode $UDMS^E$ akquiriert (HeLa-Datensatz $UDMS^E$-L). Jede LC-MS-Messung wurde dreimal wiederholt. Rohdaten der aus Triplikatmessungen bestehenden HeLa-Datensätze MS^E, $HDMS^E$, $UDMS^E$, $UDMS^E$-L wurden mit PLGS (s. Kapitel 2.5.2) prozessiert. Die PLGS-Ergebnisse wurden einer quantitativen Datenanalyse mit ISOQuant (s. Kapitel 2.5.3) unterzogen.

2.2.2 Metaproteom

Zur Herstellung der Metaproteomproben wurden HeLa-Zellen (Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, Braunschweig) und Reinkulturzellen der Bäcker- bzw. Weinhefe (*Saccharomyces cerevisiae bayanus* Stamm Lalvin® EC 1118, Institut Oenologique de Champagne, Epernay, Frankreich) nach dem oben beschriebenen FASP-Protokoll (s. Kapitel 2.3) lysiert und verdaut. Der *E.coli*-Proteomverdau

(MassPREP^{TM}-Standard) wurde von Waters Corporation (Eschborn) bezogen. Die verdauten Proteome wurden zu zwei Metaproteomen mit unterschiedlichen Gewichtsverhältnissen, wie folgt, kombiniert:

- Metaproteom A: HeLa 65%, Hefe 30% und *E.coli* 5%,
- Metaproteom B: HeLa 65%, Hefe 15% und *E.coli* 20%.

Die resultierenden Metaproteomproben A und B wurden anschließend mit 0,1%iger Ameisensäure (v/v in H_2O) zu einer Peptidkonzentration von 500 ng/µL verdünnt. Jeweils 300 ng Proteinverdau der Metaproteomproben wurden einer LC-MS-Analyse (s. Kapitel 2.4) mit 180 min Gradientenzeit in Akquisitionsmodi MS^E und UDMS^E unterzogen. Jede LC-MS-Messung wurde fünfmal wiederholt. Die erfassten Rohdaten der Metaproteomproben wurden nach Akquisitionsart zu den Datensätzen MS^E und UDMS^E (je 10 Messungen) kombiniert. Zusätzlich wurden alle MS^E- und UDMS^E-Rohdaten zu einem kombinierten Datensatz (20 Messungen) zusammengefasst. Die drei Metaproteomdatensätze wurden quantitativer Datenanalyse mit Progenesis QIP (s. Kapitel 2.5.5), synapter (s. Kapitel 2.5.4) und ISOQuant (s. Kapitel 2.5.3) unterzogen.

2.3 Filter-gestützte Probenvorbereitung

FASP ist eine integrierte Methode zur Vorbereitung biologischer Proben wie Zellen oder Gewebe für massenspektrometrische Proteomanalysen[150]. FASP beinhaltet eine Proteinaufreinigung durch Entfernung unerwünschter Substanzen aus der Probe mittels Ultrazentrifugation und einen anschließenden enzymatischen Verdau der aufgereinigten Proteine. Der Proteinverdau erfolgt mittels der Endopeptidase Trypsin, welche denaturierte Proteine mit einer hohen Spezifität am C-Terminus der Aminosäuren Lysin und Arginin spaltet[151]. Die in dieser Arbeit verwendeten Proben wurden mit Hilfe eines modifizierten FASP-Protokolls[150], wie im Folgenden beschrieben, für die massenspektrometrischen Analysen vorbereitet.

Material

- Filtereinheit: Vivacon® 500, 30000 MWCO (Sartorius Stedim Biotech, Göttingen)
- Harnstoffpuffer: 48,05 g Urea / 100 mL + 1,21 g TRIS / 100 mL, pH 8,5
- NH_4HCO_3-Puffer: 0,05 g NH_4HCO_3 in 10 mL Wasser
- IAA-Lösung: 50 mM Iodacetamid in 8 M Harnstoffpuffer
- DTT-Lösung: 8 mM Dithiothreitol in 8 M Harnstoffpuffer
- Trypsinlösung: 1 µL Trypsin (Promega GmbH, Mannheim) in 99 µL NH_4HCO_3-Puffer
- TFA-Lösung: 10% Trifluoressigsäure in H_2O
- FASP-Lysepuffer: 7 M Urea, 2 M Thiourea, 5 mM DTT, 2% CHAPS

Ablauf

Die Zellproben wurden zunächst in 200 µL FASP-Lysepuffer im Ultraschallbad (10 min, 4°C) lysiert und die Proteinmengen im Zelllysat mit dem Pierce 660 nm Protein Assay (Thermo Scientific) bestimmt. Filtereinheiten wurden mit 100 µL 1%iger Ameisensäure gereinigt (Zentrifugation, 12500 rpm, 15 min). 20 µg Protein wurden auf die Filtereinheiten gegeben und zentrifugiert (12500 rpm, 15 min). Die Filtereinheiten wurden zur Entfernung von Verunreinigungen mit niedrigem Molekulargewicht wie Salzen und Detergenzien mit 200 µL 8 M Harnstoffpuffer gewaschen (Zentrifugation, 12500 rpm, 15 min). Zur Reduktion von Disulfidbrücken wurden 100 µL DTT-Lösung auf die Filtereinheiten gegeben, die Proben wurden 15 min bei 56°C inkubiert und zentrifugiert (12500 rpm, 10 min). Filtereinheiten wurden zweimal mit 100 µL Harnstoffpuffer gewaschen (Zentrifugation, 12500 rpm, 15 min). Zur Alkylierung der Proteine wurden die Proben mit 100 µL IAA-Lösung 20 min bei Raumtemperatur im Dunkeln inkubiert und anschließend zentrifugiert (12500 rpm, 10 min). Filtereinheiten wurden erneut zweimal mit 100 µL Harnstoffpuffer gewaschen (Zentrifugation, 12500 rpm, 15 min). Nach einer Inkubation von 15 min mit 100 µL DTT-Lösung bei 56°C wurden die Proben zentrifugiert (12500 rpm, 10 min) und anschließend erneut zweimal mit 100 µL Harnstoffpuffer gewaschen (Zentrifugation, 12500 rpm, 15 min). Zur Vorbereitung der Proben für den tryptischen Verdau wurden die Filtereinheiten dreimal mit 100 µL NH_4HCO_3-Puffer gewaschen (Zentrifugation, 12500 rpm, 15 min). Proteine wurden mit 40 µL Trypsinlösung mit einem Enzym-zu-Protein-Verhältnis von 1:50 über Nacht bei 37°C verdaut. Die Trypsinlösung wurde entfernt (Zentrifugation, 12500 rpm, 10 min). Um die tryptischen Peptide aufzufangen, wurden die Filtereinheiten mit 40 µL NH_4HCO_3-Puffer gewaschen (Zentrifugation, 12500 rpm, 10 min). Der Durchfluss wurde mit 3 µL TFA-Lösung angesäuert und das Probenvolumen durch Sublimationstrocknung auf 20 µL reduziert.

2.4 LC-MS-Analyse

Die chromatographische Auftrennung der tryptischen Peptide erfolgte mittels eines nanoAcquity UPLC Systems (Waters Corporation). Zur Trennung wurde eine analytische Umkehrphasen-Trennsäule (engl. reversed-phase) HSS-T3 C18 1,8 µm, 75 µm x 250 mm verwendet. Die Proben wurden direkt auf die Säule aufgetragen. Die UPLC war online an das mit einer T-Wave-IMS-Zelle ausgestattete Massenspektrometer Synapt G2-S HDMS (Waters Corporation) gekoppelt.

Als mobile Phase A wurde Wasser mit 0,1% v/v Ameisensäure und als mobile Phase B Acetonitril mit 0,1% v/v Ameisensäure eingesetzt. Die Trennsäule wurde initial mit 1% mobiler Phase B gespült. Die chromatographische Auftrennung der Peptide erfolgte über einen Gradienten, bei dem die Konzentration der mobilen Phase B von 5% auf 40% über 90 min bzw. 180 min mit einer Flussrate von 300 nl/min stieg. Anschließend wurde die Trennsäule mit 90% mobiler Phase B bei einer Flussrate von 600 nl/min 5 min gespült und bei initialen Bedingungen reäquilibriert, so dass sich eine Gesamtlaufzeit von 110 min bzw. 200 min ergab. Die analytische Trennsäule wurde über die Dauer des Experiments auf 55°C temperiert. Der Lockmassenstandard Glu-1-Fibrinopeptid B wurde während der Messungen in Abständen von 30 s mit einer Konzentration von 100 fmol/µL durch die Hilfspumpe des LC-Systems mit einer Flussrate von 500 nl/min über den Referenzsprayer NanoLockSpray in das Massenspektrometer injiziert und mit jeweils einem MS1-Scan erfasst.

Das Massenspektrometer wurde im V-Modus mit einer typischen Auflösung von 25000 FWHM betrieben. Die Messungen erfolgten im Positiv-Ionen-Modus. Der TOF-Analysator des Massenspektrometers wurde mit einer Natriumformiatmischung von *m/z* 50 bis 1990 extern kalibriert. Die Lockmassenkorrektur der Daten erfolgte nachträglich während der Rohdatenprozessierung in PLGS anhand der zweifach geladenen Ionen des Lockmassenstandards.

Die MS/MS-Akquisition erfolgte mit den DIA-Methoden MS^{E}[91], $HDMS^{E}$[143] und $UDMS^{E}$[98]. Für die IMS wurden eine Wellenamplitude von 40 V und ein Gradient der Wellengeschwindigkeit über einen vollen IMS-Zyklus von 800 m/s bis 500 m/s verwendet. Für die Ionenfalle wurden eine Wellengeschwindigkeit von 313 m/s und eine Wellenhöhe von 8 V eingestellt, für die Transferzelle betrugen diese Werte entsprechend 190 m/s und 4 V. Jeder MS1- oder MS2-Scan dauerte 0,6 s mit einer Verzögerung von 0,05 s zwischen den Scans, so dass ein Gesamtzyklus aus einem MS1- und einem MS2-Scan 1,3 s dauerte. Für MS1-Scans wurde in allen Akquisitionsmodi eine konstante Kollisionsenergie von 4 eV verwendet. Für

die MS^E-Analysen wurde die Kollisionsenergie innerhalb eines vollen MS2-Scans linear von 16 eV bis 38 eV und für $HDMS^E$-Analysen von 25 eV bis 55 eV gesteigert[1]. Innerhalb eines IMS-Zyklus[2] wurde bei der $HDMS^E$-Akquisition die Kollisionsenergie konstant gehalten. Bei Messungen im $UDMS^E$-Modus wurden innerhalb der IMS-Zyklen unterschiedliche, ionenmobilitätsabhängige Kollisionsenergien verwendet: 17 eV für die Bins 1 bis 20, 17 eV bis 45 eV für Bins 21 bis 110 und 45 eV bis 60 eV für die Bins 111 bis 200. In allen Messungen wurde das Quadrupol für eine effiziente Transmission von Ionen mit einem *m/z* von 350 bis 2000 eingestellt. Auf diese Weise wurde sichergestellt, dass alle erfassten Ionen mit einem *m/z* unterhalb von 350 der Dissoziation von Vorläuferionen in der Kollisionszelle entstammen.

[1] Im IMS-Modus wird im Massenspektrometer Synapt G2-S die IMS-Zelle mit Stickstoff (1 mbar) geflutet. Aus der IMS-Zelle diffundiert ein Teil des Stickstoffgases in die Transferzelle, in der Argon als Kollisionsgas eingesetzt wird. Da Stickstoff eine niedrigere Molekularmasse als Argon besitzt wird im IMS-Modus eine höhere Kollisionsenergie benötigt.

[2] Ein voller IMS-Zyklus dauert 6,6 ms, innerhalb dieses Zyklus wird die Ionenmobilität 200-mal in diskreten Einheiten (Bins) erfasst.

2.5 Datenanalyse

2.5.1 Proteindatenbanken

Für die datenbankbasierte Peptid- und Proteinidentifikation wurden benutzerdefinierte Proteindatenbanken erstellt, welche Proteinsequenzen der jeweils analysierten Spezies und zusätzlich Proteinsequenzen typischer Probenkontaminationen sowie einen Decoy-Anteil enthalten. Für die die Analyse der HeLa-Rohdaten wurde eine Suchdatenbank (HeLa-Suchdatenbank) anhand der UniProtKB/Swiss-Prot-Einträge des humanen Referenzproteoms (UniProtKB Version 07.2012, 20231 Proteine) generiert. Für die Analyse von Rohdaten der Metaproteomproben wurden UniProtKB/Swiss-Prot-Einträge der Referenzproteome des Menschen, der Hefe Lalvin EC 1118 und des Bakteriums *E.coli* (UniProtKB Version 14.02.2014, 20266 humane Proteine, 5982 Hefeproteine und 4303 *E.coli*-Proteine) kombiniert (Metaproteom-Suchdatenbank). Während der Probenvorbereitung können Fremdproteine in die Ausgangsprobe gelangen. Proteinsequenzen typischer Kontaminanten wie humaner Keratine, porzinen Trypsins, etc. wurden den beiden Suchdatenbanken hinzugefügt. Der HeLa-Suchdatenbank wurde die Sequenz des Standardproteins für die absolute Proteinquantifizierung (Hefe-Enolase 1, Uniprot-ID: ENO1_YEAST) hinzugefügt. Die Suchdatenbanken wurden mit Hilfe des in PLGS eingebauten Werkzeugs für das Datenbankmanagement um Decoy-Proteinsequenzen[120] erweitert. Die Decoy-Sequenzen wurden dabei durch Umkehrung der Proteinsequenzen generiert. Der Decoy-Anteil an den Suchdatenbanken betrug 50%.

2.5.2 PLGS

Die initiale Datenanalyse wurde mit PLGS (v3.0.1 bzw. v3.0.2) durchgeführt und beinhaltete eine Rohdatenanalyse und eine datenbankbasierte Peptid- und Proteinidentifikation. Im ersten Schritt der Rohdatenanalyse mit der Software PLGS wurde eine Lockmassenkorrektur anhand der zweifach geladenen Ionen des Lockmassenstandards (Glu-1-Fibrinopeptid B, *m/z* 785,8426 $[M+2H]^{2+}$) durchgeführt. Die Peakdetektion wurde mit einem Schwellenwert von 135 Counts für die Detektion von MS1-Peaks (low energy ion threshold), 25 Counts für die Detektion von MS2-Peaks (elevated energy ion threshold) und einer minimalen Gesamtintensität der Ionen von 750 Counts. Für die Peptid- und Proteinidentifikation wurden entsprechend benutzerdefinierte Suchdatenbanken (s. Kapitel 2.5.1) verwendet. Die Massentoleranzen für die Identifikation der Vorläufer- und

Fragmentionen wurden durch PLGS während der Datenbanksuche automatisch definiert und betrugen durchschnittlich weniger als 5 ppm für Vorläuferionen und weniger als 10 ppm für Fragmentionen. Die Datenbanksuche wurde auf eine Identifikation tryptischer Peptide mit maximal bis zu zwei verfehlten Schnittstellen (engl. missed cleavages) eingeschränkt. Weiterhin wurde eine Carbamidomethylierung aller Cysteine und eine Oxidation mancher Methionine angenommen sowie bei der Datenbanksuche als Modifikationen berücksichtigt. Zusätzlich wurde die Datenbanksuche auf Peptide mit mindestens drei identifizierten Fragmentionen und auf Proteine mit mindestens zwei identifizierten Peptiden limitiert. Für die FDR der Pept- und Proteinidentifikation, die in PLGS anhand der Decoy-Identifikationen automatisch ermittelt wird, wurde ein Schwellenwert von 1% verwendet.

2.5.3 ISOQuant

Für die quantitative Datenanalyse mit dem entwickelten Analyseworkflow wurden die LC-MS-Rohdaten zunächst mit PLGS (s. Kapitel 2.5.2) prozessiert. Die Ergebnisse der PLGS-Prozessierung wurden durch ISOQuant in eine MySQL-Datenbank importiert und einer quantitativen Datenanalyse unterzogen. Bei der ISOQuant-Analyse wurden Features mit einer Masse von mindestens 500 Da und einer Intensität von 1000 berücksichtigt. Das paarweise Retentionszeitalignment aller Messungen wurde gegenüber einer automatisch in jedem Experiment ausgewählten Referenzmessung mit dem Algorithmus FastLinDRTW durchgeführt. Der Warping-Pfad wurde mit Voralignments von 1000, 5000 und 20000 repräsentativen Features und einem konstanten Radius von 500 durchgeführt. Eine Übereinstimmung zweier Features wurde bei einer Abweichung ihrer Massen von bis zu 10 ppm und bei IMS-Daten bei einer Abweichung ihrer Driftzeiten von bis zu 2,0 Bins angenommen. Korrespondierende Features wurden mit der dichtebasierten Clustering-Analyse bestimmt. Zur Reduktion der Datenmenge wurde zunächst eine Einteilung von Features in Untergruppen mit dem Preclustering durchgeführt. Das Preclustering erfolgte auf Basis der Massen und Retentionszeiten der Features. Untergruppen von Features wurden bei einer Massendifferenz zwischen benachbarten Features von mehr als 6,0 ppm und einer Retentionszeitdifferenz von mehr als 0,2 min (0,1 min bei dem UDMS^{E}- und kombinierten Metaproteomdatensatz) abgespalten. Die Preclustering-Sequenz aus der Masse- und Retentionszeitauftrennung wurde dreimal wiederholt. Untergruppen der Features wurden separat mit DBSCAN auf Feature-Cluster untersucht. Eigenschaften der Features wurden für die Analyse mit DBSCAN mit einer Massenauflösung von 6,0 ppm, einer

Retentionszeitauflösung von 0,2 min und einer Driftzeitauflösung von 2,0 Bins in einen geometrischen Raum übertragen. Die DBSCAN-Analyse der HeLa-Daten erfolgte mit mindestens einem benachbarten Punkt für die Cluster-Expansion. Bei Metaproteomdaten wurden für MS^E- und $UDMS^E$-Datensätze mindestens zwei und für den kombinierten Datensatz mindestens vier Punkte für die Cluster-Expansion verwendet. Systematische Fehler der Signalintensitäten wurden mit der multidimensionalen Normalisierung untersucht und korrigiert. Dabei wurden Abhängigkeiten der systematischen Fehler der Intensitäten von der logarithmierten Intensität und der Retentionszeit nacheinander gegenüber der durchschnittlichen Signalintensität der Feature-Cluster analysiert. In beiden Dimensionen wurde die Fehlerfunktion durch LOWESS-Regression mit einer Bandbreite von 0,3 geschätzt. PLGS-Peptididentifikationen wurden gefiltert. Bei der Annotation der Feature-Cluster wurden ausschließlich Peptide von Typ "PEP_FRAG_1" und einer Länge von mindestens sechs Aminosäuren verwendet, deren PLGS-Score in einer der Messungen mindestens 5,5 erreichte und die in mindestens zwei Messungen identifiziert wurden. Die Annotation der Feature-Cluster erfolgte restriktiv, so dass Cluster-Annotation in Konfliktfällen verworfen wurden. Nach der Annotation der Feature-Cluster wurde die FDR auf Peptidebene auf 1% beschränkt. Anschließend wurden Peptid-Protein-Beziehungen mit dem Protein-Homologie-Filter untersucht und Proteine entfernt, die mit einer unzureichenden Wahrscheinlichkeit identifiziert wurden. Intensitäten der Features, die mehreren Proteinen zugeordnet wurden, wurden zwischen ihren Ursprungsproteinen verteilt. Proteine, für die mindestens zwei Peptididentifikationen vorlagen, wurden mit der TopX-Methode quantifiziert. Bei der TopX-Quantifizierung wurden in jeder Messung maximal drei Peptide mit der höchsten Signalintensität für jedes Protein berücksichtigt. Für HeLa-Proben wurden unter Einbeziehung des Standardproteins Hefe-Enolase 1 absolute Proteinmengen berechnet. Nach der Proteinquantifizierung wurde die FDR auf Proteinebene auf 1% beschränkt.

2.5.4 synapter

Um Anforderungen von synapter zu erfüllen, wurde die PLGS-Datenanalyse der Metaproteomdaten mit geänderten Einstellungen wiederholt. Die Peptid- und Proteinidentifikation wurde hierfür in PLGS mit 100% FDR durchgeführt. PLGS-Ergebnisse wurden während der Datenanalyse zusätzlich in CSV-Dateien ausgegeben. Die exportierten CSV-Dateien wurden mit synapter in R 3.1.0 nach Empfehlungen der Urheber der Software prozessiert[144]. Die Schritte der Prozessierung wurden spezifisch für jeden

Datensatz in R programmiert. Um die Anzahl der Identifikationen unter Einhaltung einer vorgegebenen FDR zu maximieren, werden in synapter Peptididentifikationen mehrerer Messungen zu einer sogenannten Masterdatei kombiniert. Zur Bestimmung der besten Messungskombination wertet synapter alle möglichen Kombinationen der Messungen aus. Um den Vorgang der Bestimmung der besten Messungskombination zu beschleunigen wurde der Quellcode von synapter adaptiert, indem die Anzahl der in einem Schritt zu kombinierenden Messungen beschränkt wurde. Die Masterdatei wurde so aus der besten Kombination der maximal drei der insgesamt 10 Messungen der MS^E- oder UDMSˆE-Metaproteomdatensätze bzw. 20 Messungen des kombinierten Metaproteomdatensatzes generiert. Die Peptididentifikationen der entsprechenden Masterdatei wurden dann auf alle Messungen des jeweiligen Datensatzes übertragen (synapter-Autoren sprechen von "synergized"). Dabei wurden ausschließlich Peptide mit einer Sequenzlänge von mindestens sechs Aminosäuren berücksichtigt. Die Parameter für die FDR und die Falsch-Positiv-Rate (FPR) betrugen 1%. Die Analyseergebnisse von synapter wurden automatisch für jede einzelne Messung des Datensatzes in ein separates Verzeichnis ausgegeben. Aus den Ergebnisdateien auf Basis einzelner Messungen wurden Proteinidentifikationen, ihre Top3-Mengen, ihre Identifikationsscores, ihre Sequenzabdeckung und die entsprechenden Peptididentifikationen manuell extrahiert für jeden Datensatz zu messungsübergreifenden Ergebnissen zusammengetragen. Für jedes Protein wurde die Anzahl der Peptide, durch die dieses Protein identifiziert wurde, durch Auszählen eindeutiger Peptidsequenz-Modifikation-Kombinationen in R berechnet.

2.5.5 Progenesis QIP

LC-MS-Rohdaten wurden in Progenesis QIP importiert. Während des Importvorgangs wurde automatisch die Lockmassenkorrektur anhand der zweifach geladenen Ionen des Lockmassenstandards (Glu-1-Fibrinopeptid B, *m/z* 785,8426 $[M+2H]^{2+}$) und eine initiale Signaldetektion durchgeführt. Die Signaldetektion wurde mit einem Schwellenwert von 135 Counts für die Detektion von MS1-Peaks, 25 Counts für die Detektion von MS2-Peaks und einer minimalen Gesamtintensität der Ionen von 750 Counts. Ein multiples Retentionszeitalignment der importierten Daten aller Messungen eines Datensatzes zu einer automatisch ausgewählten Referenzmessung wurde ohne weitere manuelle Nachkorrekturen durchgeführt. Für die Feature-Detektion wurde die höchste Einstellung (Wert 5) der automatischen Bestimmung der Sensitivität verwendet. Die minimale Breite

eines chromatographischen Peaks wurde auf 0,15 min und die maximale Ladung der Peptidionen auf 6 beschränkt. Die Feature-Detektion wurde für jeden Datensatz anhand aller Messungen durchgeführt. Intensitäten der detektierten Features wurden zu einer Referenzmessung normalisiert, welche durch die Software automatisch bestimmt wurde. Die Peptid- und Proteinidentifikation wurde anhand der im Kapitel 2.5.1 beschriebenen Metaproteom-Suchdatenbank durchgeführt. Die Datenbanksuche wurde auf eine Identifikation tryptischer Peptide mit maximal bis zu zwei verfehlten Schnittstellen eingeschränkt. Die maximale Proteinmasse wurde auf 2500 kDa beschränkt. Bei der Datenbanksuche wurden Proteinmodifikationen wie eine Carbamidomethylierung aller Cysteine, eine Desaminierung mancher Asparagine und Glutamine sowie eine Oxidation mancher Methionine berücksichtigt. Toleranzparameter der Datenbanksuche wurden automatisch durch die Software zur Laufzeit ermittelt. Die Datenbanksuche wurde auf Peptide mit mindestens zwei identifizierten Fragmentionen und Proteine mit insgesamt mindestens fünf Fragmentionen ihrer Peptide sowie eine FDR von bis zu 1% limitiert. Identifikationen der Peptide mit einer Sequenzlänge von weniger als sechs Aminosäuren oder einem Identifikationsscore von weniger als 4,0 wurden durch die Definition entsprechender Filterkriterien entfernt. Für die Proteinquantifizierung wurden Standardeinstellungen der Software verwendet, so dass ähnliche Proteine gruppiert wurden und die Quantifizierung ausschließlich anhand konfliktfreier Features durchgeführt wurde. Die Ergebnisse der quantitativen Analysen mit Progenesis QIP wurden durch die eingebaute Berichtsfunktion “export protein measurements” für jeden Datensatz in separate CSV-Dateien exportiert. Da Progenesis QIP die FDR der Peptid- und Proteinidentifikation auf Basis einzelner Messungen kontrolliert, wurde für die Analyseergebnisse einzelner Datensätze die FDR messungsübergreifend berechnet und die Einhaltung des 1% FDR-Niveau sichergestellt.

2.6 Geräte und Software

Für die Anfertigung dieser Arbeit wurden für Tätigkeiten wie Probenvorbereitung, Datenerfassung, Datenanalyse und sonstige EDV-gestützte Tätigkeiten unterschiedliche Geräte und Software verwendet. In Tabelle 1 werden die verwendeten Geräte und in Tabelle 2 die verwendeten Softwarepakete zusammengefasst.

Tab. 1: Die im Zusammenhang mit dieser Arbeit verwendeten Geräte.

Gerät	Typ	Hersteller
Analytische Trennsäule	HSS-T3 C18 1,8 µm, 75 µm x 250 mm	Waters Corporation
Computer	ThinkStation C30	Lenovo
Computer	Mac Pro 3.1	Apple Inc.
Computer	MacBook Pro 7.1	Apple Inc.
Computer	ThinkStation D20	Lenovo
Gefriertrocknungsanlage	Finn-Aqua® Lyovac GT2	Steris
Heizblock	BTD	Grant
Inkubator	Heraeus Function Line	Thermo Scientific
LC-System	nanoAcquity UPLC System	Waters Corporation
Massenspektrometer	Synapt G2-S HDMS	Waters Corporation
Pipette	Eppendorf Research® plus	Eppendorf
Pipettenspitze	MAXYMum Recovery TR-222-C-L-R	Axygen Scientific
Pipettenspitze	MAXYMum Recovery TF-1000-L-R-S	Axygen Scientific
Tiefkühlschrank	Heraeus HERAfreeze™ HFU 486 Basic	Thermo Scientific
Ultraschallbad	SONOREX	Bandelin
Vakuumzentrifuge	SpeedVac SVC 100	Savant
Zentrifugaleinheit	Vivacon® 500, 30000 MWCO	Sartorius AG
Zentrifuge	Heraeus Biofuge Pico	Thermo Scientific

Tab. 2: Die im Zusammenhang mit dieser Arbeit verwendeten Softwarepakete.

Software	Version	Hersteller/Autor
Adobe Illustrator	CS6	Adobe Systems Inc.
Adobe Reader	10 - 11	Adobe Systems Inc.
Draw.io	4.2.3.1	Jgraph Ltd.
Eclipse IDE	4.2 - 4.4	Eclipse Foundation
Envelope	0.9.20	Sykes et al.[110]
ISOQuant	1.5 - 1.6	Kuharev et al. (http://isoquant.org)
Java SE JDK	6 - 8	Oracle Corporation
Java SE JRE	5.0, 6 - 8	Oracle Corporation
Mac OS X	10.7 - 10.10	Apple Inc.
MacTeX	2014	TeX Users Group
MassLynx	4.1	Waters Corporation
Microsoft Excel	14	Microsoft Corporation
Microsoft Powerpoint	14	Microsoft Corporation
Microsoft Word	14	Microsoft Corporation
Mozilla Firefox	17 - 36	Mozilla Corporation
MySQL Server	5.1 - 5.6	Oracle Corporation
Notepad++	6.1 - 6.6	Don Ho (http://notepad-plus-plus.org)
pandoc	1.9 - 1.13	John MacFarlane (http://johnmacfarlane.net)
PLGS	3.0 - 3.02	Waters Corporation
GraphPad PRISM	5.0	GraphPad Software, Inc.
Progenesis QIP	1.0	Nonlinear Dynamics, Waters Corporation
R	3.1.0	R Development Core Team[152]
RStudio	0.98	RStudio, Inc.
synapter	1.7.0	Bond et al.[144]
TextWrangler	4.5	Bare Bones Software, Inc.
Windows	7	Microsoft Corporation
Zotero	4	Roy Rosenzweig Center for History and New Media

3 Ergebnisse

Im Rahmen dieser Arbeit wurde ein Workflow und eine Reihe von Algorithmen zur labelfreien, quantitativen Analyse von MS^E/$HDMS^E$/$UDMS^E$-Daten entwickelt. Die entwickelten Methoden sowie der Analyseworkflow wurden als Bestandteile der freien Software ISOQuant implementiert. Die Auswirkung des entwickelten Analyseworkflows auf die Proteinidentifikation und die Proteinquantifizierung wurde durch einen Vergleich von Daten vor und nach der ISOQuant-Analyse gezeigt. Der Analyseworkflow und die entwickelten Algorithmen sowie deren Auswirkung auf die Proteinidentifikation und Proteinquantifizierung wurden teilweise mit dem Artikel “Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics”[98] in Nature Methods publiziert. Die Effizienz des entwickelten Analyseworkflows, die Reproduzierbarkeit der Proteinidentifikation und der Proteinquantifizierung sowie die Richtigkeit der Proteinquantifizierung wurden im Vergleich von ISOQuant mit synapter und Progenesis QIP evaluiert. Die Ergebnisse dieser Evaluierung wurden mit dem Artikel “In-depth evaluation of software tools for data-independent acquisition based label-free quantification”[153] in Proteomics veröffentlicht.

3.1 Analyseworkflow

Zur Lösung der im Kapitel 1.8.1 beschriebenen, spezifischen Probleme der Analyse von MS^E/$HDMS^E$/$UDMS^E$-Daten wurde in der vorliegenden Arbeit ein Datenanalyseworkflow entwickelt, der einzelne Unterprobleme der Datenanalyse schrittweise mit hierfür entwickelten Algorithmen löst. Die initiale Prozessierung der Rohdaten einschließlich der Peptid- und Proteinidentifikation wird mit Hilfe von PLGS durchgeführt. Nach der Vorverarbeitung der MS^E/$HDMS^E$/$UDMS^E$-Daten mit PLGS werden die bis dahin unabhängigen PLGS-Ergebnisse einzelner LC-MS-Messungen zusammengeführt und schrittweise analysiert. Abbildung 6 zeigt schematisch den Ablauf der Datenanalyse nach der PLGS-Analyse. Die Einzelschritte des Analyseworkflows können prinzipiell in drei Blöcke zusammengefasst werden: die Signal-, die Identifikations- und die Quantifizierungsanalyse. Die Signalanalyse dient hauptsächlich der Zuordnung korrespondierender Features zwischen den Messungen des Experiments. Dabei werden mit einem Retentionszeitalignment die nichtlinearen Verschiebungen der Retentionszeiten zwischen den einzelnen LC-MS-Messungen des Experiments ausgeglichen und die korrespondierenden Features durch das Feature-Clustering gruppiert. Im Anschluss werden mit der Normalisierung der Signalintensitäten die systematischen Fehler der gemessenen Signalintensitäten korrigiert. Die Identifikationsanalyse stellt messungsübergreifend eine einheitliche Qualität der Peptididentifikation sicher, annotiert Feature-Cluster und löst das Proteininferenz-Problem. Bei der Quantifizierungsanalyse werden zunächst die Signalintensitäten der geteilten Peptide (engl. shared peptides), die aus mehreren Proteinen stammen, zwischen ihren Ursprungsproteinen verteilt. Unter Einhaltung zusätzlicher Qualitätskriterien für die Peptid- und Proteinidentifikation wird eine absolute Proteinquantifizierung durchgeführt. Für die Einzelschritte der Analyse wurden dedizierte Algorithmen entwickelt, deren Funktionsweisen in den folgenden Kapiteln (3.3 bis 3.12) beschrieben werden.

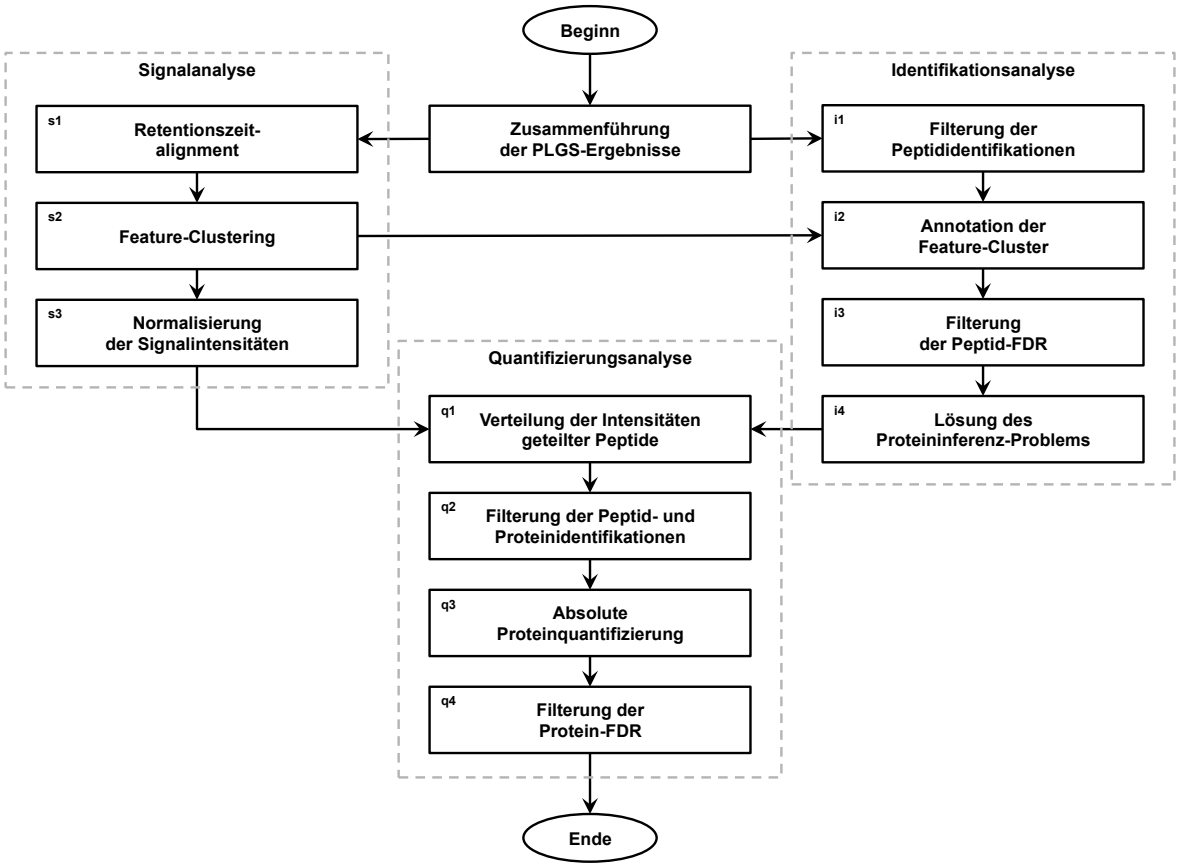

Abb. 6: Ablauf der Post-PLGS-Analyse von MS^E/$HDMS^E$/$UDMS^E$-Daten.
Nach der Rohdatenprozessierung sowie der Peptid- und Proteinidentifikation werden die PLGS-Ergebnisse der einzelnen LC-MS-Messungen des Experiments zusammengeführt und in mehreren Schritten analysiert. (**s1**) Mit dem Retentionszeitalignment werden die nichtlinearen Retentionszeitverschiebungen zwischen den Messungen analysiert und ausgeglichen. (**s2**) Mit dem Feature-Clustering werden korrespondierende Features aus unterschiedlichen LC-MS-Messungen gruppiert. (**s3**) Die Signalintensitäten der Features werden normalisiert. (**i1**) Die identifizierten Peptide werden nach Benutzervorgaben gefiltert und (**i2**) die Feature-Cluster mit Peptididentifikationen annotiert. (**i3**) Die FDR der Peptididentifikationen der annotierten Feature-Cluster wird ermittelt und beschränkt. (**i4**) Das Proteininferenz-Problem wird gelöst. (**q1**) Signalintensitäten von Peptiden, die mehreren Proteinen zugeordnet sind, werden zwischen ihren Ursprungsproteinen verteilt. (**q2**) Die Peptid- und Proteinidentifikationen werden für den Einsatz in der Proteinquantifizierung gefiltert. (**q3**) Proteine werden quantifiziert sowie (**q4**) ihre FDR ermittelt und beschränkt.

3.2 Datenzugriff und Zusammenführung der PLGS-Ergebnisse

Analyseergebnisse der Messungen eines LFQ-LC-MS-Experiments werden in PLGS in Form eines Projektes zusammengefasst und in einer Verzeichnisstruktur im lokalen Dateisystem gespeichert. Die vorliegende Datenstruktur wurde vom Hersteller nicht öffentlich spezifiziert. Teile der verfügbaren Daten liegen im XML-Format vor, so dass die zugrunde liegenden Datenstrukturen nachvollzogen werden konnten. Diese Datenstrukturen umfassen Ergebnisse der Feature-Detektion als EMRT[1] sowie der Peptid- und Proteinidentifikation. Zusätzlich enthalten die PLGS-Datenstrukturen Metainformationen über das jeweilige Projekt, Instrumenten- und Softwareparameter, die biologischen Proben und ihre technischen Replikate. Zur Zusammenführung der PLGS-Ergebnisse der einzelnen LC-MS-Messungen eines Experiments sowie zur Vereinfachung des Datenzugriffs wurde eine relationale Datenbank entworfen, deren vereinfachte Struktur in Abbildung 7 dargestellt wird. Die relationale Datenbank umfasst das Design des Experiments und die Ergebnisse der Feature-Detektion sowie der Peptid- und Proteinidentifikation. Die Zusammenführung der Daten in der relationalen Datenbank erlaubt eine einfache Analyse messungsübergreifender Zusammenhänge.

[1] In der vereinfachten Darstellung als EMRT werden die detektierten Features durch ihre integrierte Signalintensität, die monoisotopische Peptidmasse, Retentionszeit und im Falle der IMS-MS zusätzlich durch die Driftzeit beschrieben.

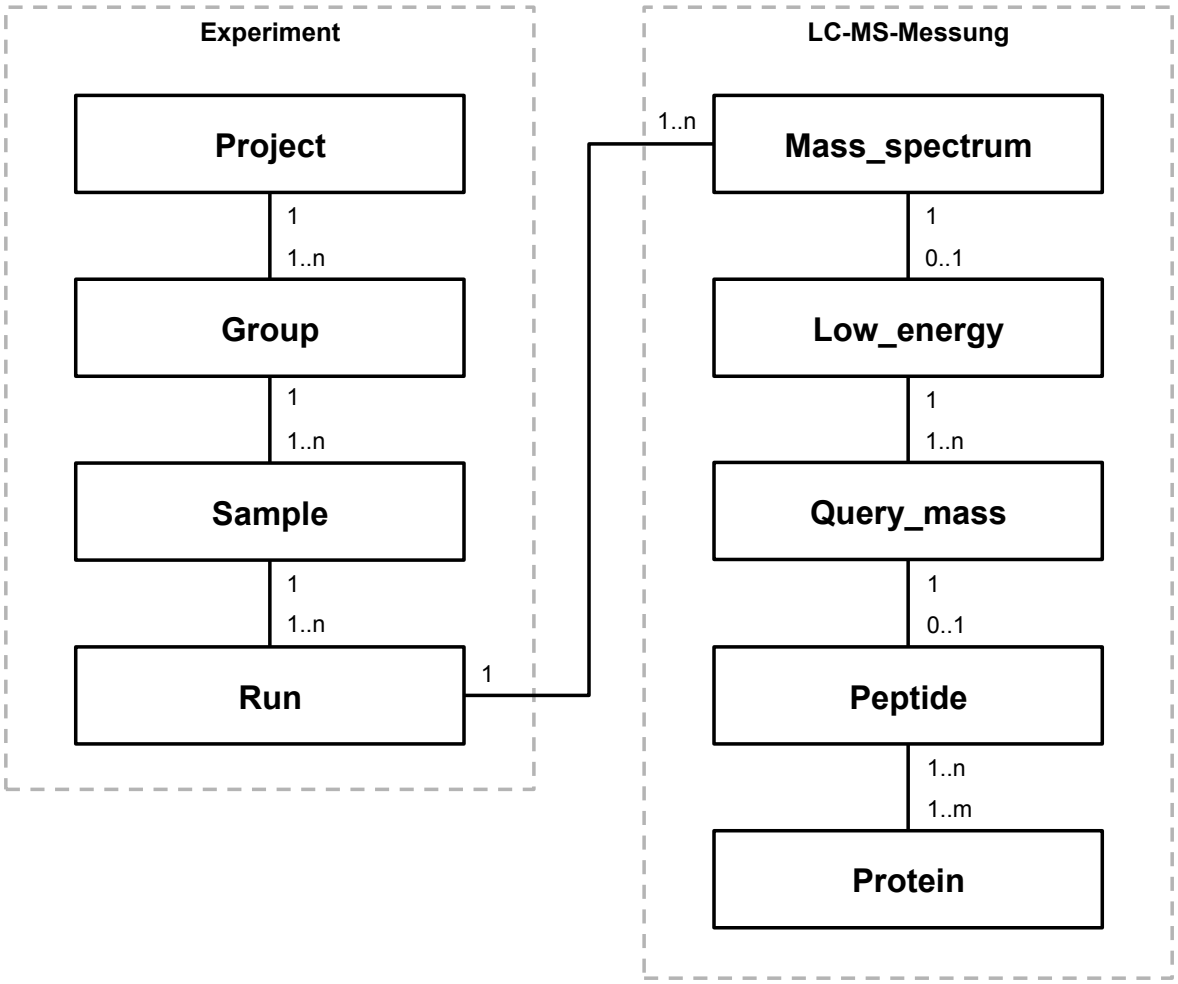

Abb. 7: Relationale Datenbank zur Speicherung von PLGS-Ergebnissen.
Die Struktur der relationalen Datenbank zur Speicherung von PLGS-Ergebnissen orientiert sich an PLGS-Datenstrukturen. Die Datenbank bildet das Design eines LFQ Experimentes und die Ergebnisse der LC-MS-Messungen ab. Dabei können in einem Experiment (Project) mehrere Probengruppen (Group), in einer Gruppe mehrere Proben (Sample) und in einer Probe mehrfache Messungen (Run) vorkommen. Jeder LC-MS-Messung werden die PLGS-Ergebnisse der Signalprozessierung als Profile der detektierten Signale der MS1-Massenspektren (Mass_spectrum), der Feature-Detektion mit Features als EMRTs (Low_energy) sowie der Peptid- und Proteinidentifikation (Query_mass, Peptide und Protein) zugeordnet. Dargestellt wird eine vereinfachte Struktur der relationalen Datenbank. Eigenschaften der Entitäten werden hier nicht aufgeführt.

3.3 Retentionszeitalignment

Instabilitäten des flüssigchromatographischen Trennverfahrens können zu nichtlinearen Verzerrungen der Retentionszeiten von Peptiden führen (s. Kapitel 1.8.1). Als Folge eluieren gleiche Peptidspezies in unterschiedlichen LC-Läufen zu unterschiedlichen Retentionszeiten. Der zeitliche Versatz zwischen den nacheinander eluierenden Peptiden kann variieren. Die Ermittlung dieser Zeitverzerrungen findet mit einem sogenannten Retentionszeitalignment statt. Das Retentionszeitalignment zwischen zwei Messungen wird als paarweises und zwischen drei oder mehr Messungen als multiples Retentionszeitalignment bezeichnet. Im Folgenden wird zunächst das Problem des paarweisen Retentionszeitalignments erläutert. Danach werden mehrere aufeinander aufbauende Algorithmen zur Lösung des paarweisen Retentionszeitalignmentproblems beschrieben und ihre Effizienz verglichen. Anschließend wird eine Strategie für das multiple Retentionszeitalignment aller Messungen des Experiments und entsprechende Korrektur der Retentionszeitverschiebungen vorgestellt.

3.3.1 Paarweises Retentionszeitalignment

Das paarweise Retentionszeitalignment zweier LC-MS-Messungen untereinander kann, wie folgt, formalisiert werden:
Es gibt zwei Sequenzen T und L, die entsprechend n und m Features enthalten.

$$T = t_1, t_2, ..., \quad t_{i-1}, t_i, t_{i+1}, ..., \quad t_n$$
$$L = l_1, l_2, ..., \quad l_{j-1}, l_j, l_{j+1}, ..., \quad l_m$$

Jedes Feature wird in abstrakter Form als EMRT mit seinen Eigenschaften: der Retentionszeit (rt), der monoisotopischen Masse (mass) und der Signalintensität (intensity) als Mengenausprägung notiert. Für IMS-MS Daten wird ein EMRT zusätzlich um die Angabe der Driftzeit (dt) erweitert. Die Sequenzen T und L enthalten Features in chronologischer Reihenfolge.

$$rt_{(t_{i-1})} \leq rt_{(t_i)} \leq rt_{(t_{i+1})}$$
$$rt_{(l_{j-1})} \leq rt_{(l_j)} \leq rt_{(l_{j+1})}$$

Die Features-Sequenzen werden mit der Annahme aligniert, dass sie Features gleicher Peptide enthalten, die zu unterschiedlichen Zeiten mit unbekannten nichtlinearen Zeitverzerrungen eluieren und dass die Mehrheit der korrespondierenden Peptide in beiden Sequenzen in gleicher Reihenfolge auftreten. Dann ist das Retentionszeitalignment $\mathbb{A}$ für T und L die Superposition der Retentionszeiten der Untermengen $\overline{T}$ und $\overline{L}$.

$$\mathbb{A} = \begin{pmatrix} rt(\overline{T}) \\ rt(\overline{L}) \end{pmatrix} = \left\{ \quad \begin{pmatrix} rt(\bar{t}_1) \\ rt(\bar{l}_1) \end{pmatrix}, \ldots, \begin{pmatrix} rt(\bar{t}_{\bar{i}}) \\ rt(\bar{l}_{\bar{j}}) \end{pmatrix}, \ldots, \begin{pmatrix} rt(\bar{t}_{\overline{m}}) \\ rt(\bar{l}_{\overline{n}}) \end{pmatrix} \quad \right\}$$

mit

$$\overline{T} \in T, \qquad \overline{T} = \bar{t}_1, ..., \bar{t}_{\overline{n}}, \quad 1 \leq \overline{n} \leq n, \qquad rt(\bar{t}_{\bar{i}-1}) \leq rt(\bar{t}_{\bar{i}}) \leq rt(\bar{t}_{\bar{i}+1})$$
$$\overline{L} \in L, \qquad \overline{L} = \bar{l}_1, ..., \bar{l}_{\overline{m}}, \quad 1 \leq \overline{m} \leq m, \qquad rt(\bar{l}_{\bar{j}-1}) \leq rt(\bar{l}_{\bar{j}}) \leq rt(\bar{l}_{\bar{j}+1})$$

Das Ergebnis des paarweisen Alignments erlaubt eine Projektion der Retentionszeiten übereinstimmender Features einer Sequenz auf die Zeitachse der zweiten Sequenz.

3.3.2 Dynamic Retention Time Warping

Mit dem Dynamic Time Warping (DTW) wurde ein generischer Algorithmus zur Messung von Distanzen zwischen zwei zeitaufgelösten Signalsequenzen beschrieben[154, 155]. Der DTW-Algorithmus nutzt das Prinzip der dynamischen Programmierung, um ein nichtlineares Mapping der Signale einer Sequenz zu den korrespondierenden Signalen der anderen Sequenz durch die Minimierung der Unähnlichkeitsdistanz (engl. dissimilarity distance) zwischen den Signalreihen zu berechnen. Auf der Basis von DTW wird im Folgenden ein neuer Algorithmus das Dynamic Retention Time Warping (DRTW) definiert, der die Anwendung des DTW-Algorithmus auf das Problem des paarweisen Retentionszeitalignments von EMRT-Sequenzen ermöglicht. Der Ablauf des DRTW-Algorithmus wird in der Abbildung 8 gezeigt.

Beim DRTW erfolgt die Berechnung des Alignments in zwei Stadien. Zunächst wird durch dynamische Programmierung eine Distanzmatrix ($\mathbb{M}$) erstellt, deren Größe $(n+1) \times (m+1)$ durch die beiden Zeitreihen vorgegeben wird. Die Distanzmatrix wird initialisiert und mit kumulativen Unähnlichkeitsdistanzen der Sequenzpräfixe an jeder Position der verglichenen Zeitreihen gefüllt. Ist die Distanzmatrix berechnet, folgt ein Rückverfolgungsalgorithmus der minimalen Unähnlichkeit einem möglichen, optimalen Warping-Pfad durch die Distanzmatrix und konstruiert das entsprechende Alignment der Eingabesequenzen. Als eine Basis zur Berechnung der Distanzmatrix dient die problemspezifische Bewertungsfunktion $w(t, l)$, die die Unähnlichkeit zwischen einzelnen Elementen der Eingabesequenzen mit einem Score beziffert. Die Bewertung der Unähnlichkeit zweier Features $w(t, l)$, bzw. eines Features und der leeren Menge ε erfolgt auf Basis konstanter Scores $SCORE_{gap} = 1$, $SCORE_{match} = -1$ und $SCORE_{mismatch} = 3$, die auf Grund der Übereinstimmung der ausgewerteten Features vergeben werden. Dazu wird der relative Massenunterschied zweier EMRTs (in ppm) berechnet und mit dem benutzerdefinierten Schwellenwert Δ_{mass}^{max} verglichen. Für IMS-MS-Daten wird zusätzlich die absolute Differenz der Driftzeit (in instrumentenspezifischen Einheiten) mit dem benutzerdefinierten Schwellenwert Δ_{dt}^{max} verglichen. Übereinstimmung zweier Features wird angenommen, solange die Differenzen ihrer Massen und ihrer Driftzeiten die jeweiligen Schwellenwerte nicht übersteigen. Nach der Initialisierung der ersten Zeile und der ersten Spalte der Distanzmatrix $\mathbb{M}(0, 0..m)$ und $\mathbb{M}(0..n, 0)$ werden weitere Matrix-Einträge auf Basis der Vorläufer-Scores und der Unähnlichkeit an der jeweiligen Position rekursiv berechnet. Auf diese Weise werden während der rekurrenten Berechnung der Distanzmatrix die Unähnlichkeitsdistanzen durch Addition der Vorläufer-Scores kumuliert.

Eingabe:

$$\begin{array}{l} T = t_1, ..., t_n, \quad rt_{(t_{i-1})} \leq rt_{(t_i)} \leq rt_{(t_{i+1})} \\ L = l_1, ..., l_m, \quad rt_{(l_{j-1})} \leq rt_{(l_j)} \leq rt_{(l_{j+1})} \\ \Delta_{mass}^{max}, \quad \Delta_{dt}^{max} \end{array} \tag{a}$$

Ausgabe:

$$\mathbb{A}_{DRTW} = \left\{ \begin{matrix} \bar{T} \\ \bar{L} \end{matrix} \right\}, \qquad \bar{T} \in \{T \cup \varepsilon\} \wedge \bar{L} \in \{L \cup \varepsilon\} \tag{b}$$

Algorithmus:

$$match_{(t,l)} := \frac{\left| mass_{(t)} - mass_{(l)} \right|}{\max(mass_{(t)},\ mass_{(l)})} \leq \Delta_{mass}^{max} \quad \wedge \quad \left(dt_{(t)} = NA \quad \vee \quad dt_{(l)} = NA \quad \vee \quad \left| dt_{(t)} - dt_{(l)} \right| \leq \Delta_{dt}^{max} \right) \tag{c}$$

$$w_{(t,l)} := \begin{cases} SCORE_{gap} & \text{if} \quad (t = \varepsilon) \quad \vee \quad (l = \varepsilon) \\ SCORE_{match} & \text{if} \quad (t \neq \varepsilon) \quad \wedge \quad (l \neq \varepsilon) \quad \wedge \quad (match_{(t,l)} = true) \\ SCORE_{mismatch} & \text{if} \quad (t \neq \varepsilon) \quad \wedge \quad (l \neq \varepsilon) \quad \wedge \quad (match_{(t,l)} = false) \end{cases} \tag{d}$$

$$\begin{array}{l} \mathbb{M}_{0,0} := 0 \\ \mathbb{M}_{i,0} := \mathbb{M}_{i-1,0} + w_{(t_i,\varepsilon)}, \qquad 1 \leq i \leq n \\ \mathbb{M}_{0,j} := \mathbb{M}_{0,j-1} + w_{(\varepsilon,l_j)}, \qquad 1 \leq j \leq m \end{array} \tag{e}$$

$$\mathbb{M}_{i,j} := \min \left\{ \begin{matrix} \mathbb{M}_{i-1,j} & + w_{(\varepsilon,l_j)} \\ \mathbb{M}_{i,j-1} & + w_{(t_i,\varepsilon)} \\ \mathbb{M}_{i-1,j-1} & + w_{(t_i,l_j)} \end{matrix} \right\}, \quad 1 \leq i \leq n \quad \wedge \quad 1 \leq j \leq m \tag{f}$$

$$\mathbb{A}_{i,j} := \left\{ \begin{matrix} \left\{ \binom{t_i}{l_j} \bigcup \mathbb{A}_{i-1,j-1} \right\} & \text{if} \quad \big(i > 0\big) \wedge \big(j > 0\big) \wedge \big(\mathbb{M}_{i,j} = \mathbb{M}_{i-1,j-1} + w_{(t_i,l_j)}\big) \\ \left\{ \binom{t_i}{\varepsilon} \bigcup \mathbb{A}_{i-1,j} \right\} & \text{if} \quad \big(i > 0\big) \wedge \Big(\big(j = 0\big) \vee \big(\mathbb{M}_{i,j} = \mathbb{M}_{i-1,j} + w_{(\varepsilon,l_j)}\big)\Big) \\ \left\{ \binom{\varepsilon}{l_j} \bigcup \mathbb{A}_{i,j-1} \right\} & \text{if} \quad \big(j > 0\big) \wedge \Big(\big(i = 0\big) \vee \big(\mathbb{M}_{i,j} = \mathbb{M}_{i,j-1} + w_{(t_i,\varepsilon)}\big)\Big) \\ \left\{ \right\} & \text{if} \quad \big(i = 0\big) \wedge \big(j = 0\big) \end{matrix} \right. \tag{g}$$

$$\mathbb{A}_{DRTW} := \mathbb{A}_{n,m} \tag{h}$$

Abb. 8: Ablauf des Algorithmus Dynamic Retention Time Warping (DRTW).
(**a**) EMRT-Sequenzen T und L, Schwellenwerte für relative Massendifferenz Δ_{mass}^{max} und absolute Driftzeitdifferenz Δ_{dt}^{max}, (**b**) DRTW-Alignmentergebnis $\mathbb{A}_{DRTW}$, (**c**) Prüfung der Feature-Äquivalenz $match_{(t,l)}$, (**d**) Bewertung der Festure-Unähnlichkeit $w_{(t,l)}$, (**e**) Initialisierung der Distanzmatrix $\mathbb{M}$, (**f**) Distanzmatrix-Rekurrenz, (**g**) Rückverfolgung und Alignmentkonstruktion, (**h**) determinierende Rückgabe des Alignments $\mathbb{A}_{DRTW}$.
Die Eigenschaften eines Features x werden durch Funktionen dargestellt: Retentionszeit $rt_{(x)}$, Masse $mass_{(x)}$ und Driftzeit $dt_{(x)}$ (mit dem Wert NA bei nicht IMS-MS-Daten).
Für die Bewertung der Unähnlichkeit zweier Features werden konstante Scores verwendet: $SCORE_{gap} = 1$, $SCORE_{match} = -1$, $SCORE_{mismatch} = 3$. Die leere Menge wird dargestellt als $\varepsilon = \{\}$. Die Eingabeparameter Δ_{mass}^{max} und Δ_{dt}^{max} hängen von verwendeten Instrumenten und den Experimentenbedingungen ab.

In Abbildung 9 werden schematisch eine Distanzmatrix und ihre rekurrente Berechnung nach der Initialisierung verdeutlicht. Nachdem die gesamte Distanzmatrix berechnet wurde, findet der Rückverfolgungsalgorithmus ausgehend von der letzten Position $\mathbb{M}_{(n,m)}$ ein optimales Alignment der Features, indem er die möglichen Vorläuferpositionen berechnet und den entsprechenden Vorläufern zum Ursprung der Matrix $\mathbb{M}(0,0)$ folgt.

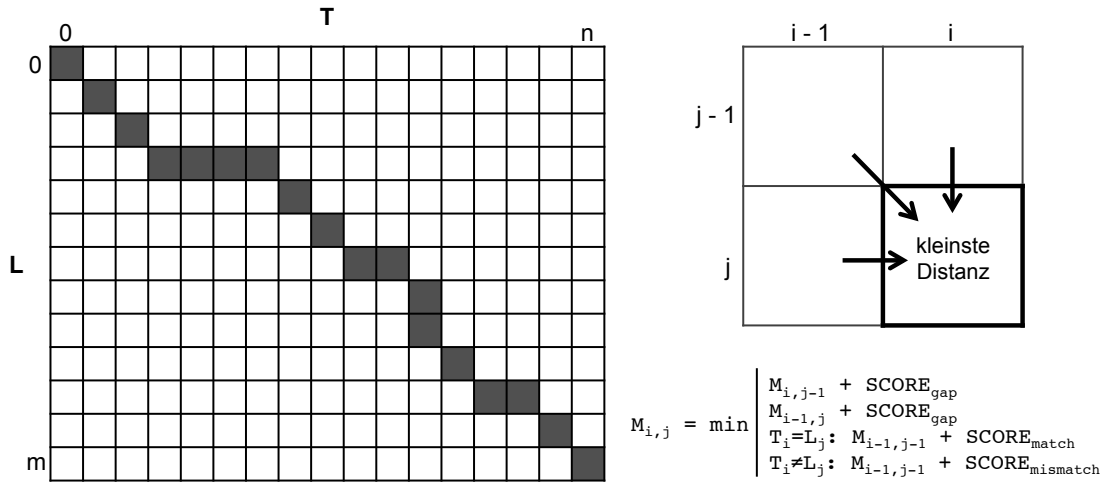

Abb. 9: DRTW-Distanzmatrix und rekurrente Distanzminimierung.
Nach der Initialisierung der jeweils ersten Zeile und Spalte werden alle weiteren Positionen der der DRTW-Distanzmatrix rekurrent anhand der minimalen Distanz zu den drei möglichen Vorläufern berechnet. Stimmen die Features an der aktuellen Zeilen- und Spaltenposition überein (T_i=L_j), wird zum diagonalen Vorläufer der $SCORE_{match}$ andernfalls der $SCORE_{mismatch}$ addiert. Zu der horizontalen sowie der vertikalen Vorläuferdistanz wird jeweils der $SCORE_{gap}$ addiert. Die Distanz an der aktuellen Position wird dann als das Minimum der drei berechneten Distanzen notiert. Nach der vollständigen Berechnung der Matrix wird von der letzten Position $\mathbb{M}(n,m)$ ausgehend zum ersten Position $\mathbb{M}(0,0)$ der so genannte Warping-Pfad (graue Positionen) konstruiert.

Der DRTW-Algorithmus hat eine quadratische Komplexität $\mathcal{O}(n^2)$. Die Berechnung der Distanzmatrix erfolgt in quadratischer Laufzeit (Zeitkomplexität $\mathcal{O}(n^2)$) und die Bereitstellung der entsprechenden Datenstrukturen hat einen quadratischen Speicherbedarf (Platzkomplexität $\mathcal{O}(n^2)$). DRTW liefert eine mathematisch optimale Lösung für das Retentionszeitalignment. Der Speicherbedarf für das Alignment von Sequenzen mit wenigen zehntausenden Features kann beim Einsatz moderner Hardware bereits die verfügbaren Ressourcen übersteigen.

3.3.3 Fast Dynamic Retention Time Warping

Mehrere Techniken zur Reduktion der Komplexität des DTW-Algorithmus wurden vorgestellt. So kann die Performance des DTW durch Techniken der globalen Restriktion des Suchraumes verbessert werden. Bei diesen Techniken wird auf die Berechnung der kompletten Distanzmatrix verzichtet. Die Repräsentation der Distanzmatrix wird dann auf einen fensterartigen zusammenhängenden Ausschnitt reduziert, in dem der Warping-Pfad mit hoher Wahrscheinlichkeit erwartet wird. Die Abbildung 10 zeigt die zwei bekanntesten Techniken der globalen Suchraumeinschränkung des DTW das Itakura-Parallelogramm[156] und die Sakoe-Chiba-Bande[157].

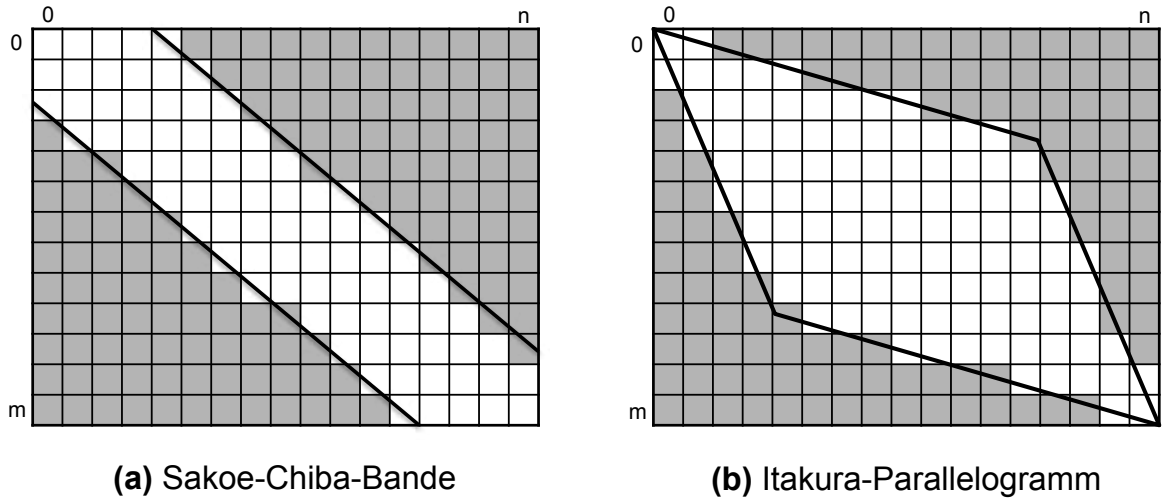

Abb. 10: Sakoe-Chiba-Bande und Itakura-Parallelogramm.
Die Performance von DTW kann durch globale Suchraumrestriktion in Form der Sakoe-Chiba-Bande[157] oder des Itakura-Parallelogramms[156] gesteigert werden. Beide Methoden definieren einen festen Bereich in der Distanzmatrix, der berechnet wird. Ausserhalb der Begrenzung (graue Bereiche) wird auf die Berechnung verzichtet.

Diese Techniken setzen hohe Ähnlichkeit sowie eine vergleichbare Länge der Eingabesequenzen voraus, so dass der Warping-Pfad diagonal etwa durch die Mitte der Distanzmatrix verläuft. Mit der steigenden Fensterbreite werden diese Bedingungen zu Lasten der Performance aufgeweicht, dabei nähert sich die resultierende Komplexität der quadratischen Komplexität des ursprünglichen DTW-Algorithmus. Zur Steigerung der Performance kann das Alignment auf einer reduzierten Repräsentation der Daten berechnet und auf die Ausgangsdaten übertragen werden. Beispielsweise werden beim Warp2D-Algorithmus LC-Daten durch Downsampling, d.h. Resampling des Signals mit einer niedrigeren Abtastrate, reduziert[158]. Die Reduktion der Datenmenge geht jedoch mit dem Verlust der Information über lokale Zeitverzerrungen einher. Sowohl die globale

Suchraumeinschränkung als auch das Downsampling der Daten reduzieren die effektive Komplexität des DTW durch potentiellen Verlust der Genauigkeit. Ohne Kenntnis über die Lage eines optimalen Warping-Pfads können statische Fenster falsche Warping-Pfade innerhalb des berechneten Bereichs erzwingen. Dies kann z.B. beim Alignment von LC-MS-Daten der Fall sein, die mit unterschiedlichen chromatographischen Methoden analysiert wurden, oder wenn die Komplexität der Datensätze stark variiert. Das Alignment auf reduzierten Daten findet zwar globale Zeitverzerrungen zwischen den Datensätzen, vernachlässigt jedoch kleine, lokale Zeitverschiebungen, die in chromatographischen Daten häufig auftreten.

Eine flexible Lösung des Problems liefert eine Kombination der Suchraumeinschränkung und der Datenreduktion. Wenngleich das Alignment auf reduzierten Daten lediglich eine Schätzung des Warping-Pfades liefert, kann es zur dynamischen Reduktion des Suchraumes für das Alignment der Ausgangsdaten verwendet werden. Dazu wird zunächst ein Voralignment auf einer reduzierten Repräsentation der Daten berechnet und auf die Distanzmatrix für das Alignment der Ausgangsdaten projiziert. Der Suchraum wird damit auf ein dynamisch berechnetes Fenster um die Projektion des Voralignments eingeschränkt. Algorithmen wie IDDTW[159] und FastDTW[160] zeigen, dass der Verlauf des Warping-Pfades iterativ verfeinert werden kann, indem das Alignment für unterschiedlich grobe Abstraktionsebenen der Ausgangsdaten berechnet wird.

Durch eine Adaption der Technik zur iterativen Verfeinerung des Warping-Pfades zur Lösung des paarweisen Retentionszeitalignments wird ein neuer Algorithmus das Fast Dynamic Retention Time Warping (FastDRTW) definiert, dessen Ablauf in Abbildung 11 schematisch dargestellt wird. In der Anwendung der iterativen Pfadverfeinerung für das Alignment von Feature-Sequenzen ergibt sich zunächst die Schwierigkeit der Datenabstraktion. Durch EMRTs repräsentierte Features können nicht durch Skalieren oder Resampling reduziert werden. FastDRTW erreicht eine Datenabstraktion durch Anwendung benutzerdefinierter Filter für die Auswahl von repräsentativen Features, z.B. Auswahl einer Zahl von Features mit der höchsten Intensität oder mit bestimmten Massen. Durch den Filter wird im ersten Schritt eine geringe Zahl repräsentativer Features, z.B. je 1000 Features, aus den jeweiligen Eingabedaten ausgewählt. Für die repräsentativen Features wird ein Alignment mit dem oben beschriebenen DRTW-Algorithmus berechnet. Im nächsten Schritt wird der Warping-Pfad des “groben” Voralignments für das Alignment der nächst größeren Datenrepräsentation, z.B. je 2000 Features, auf die Distanzmatrix projiziert und mit einem

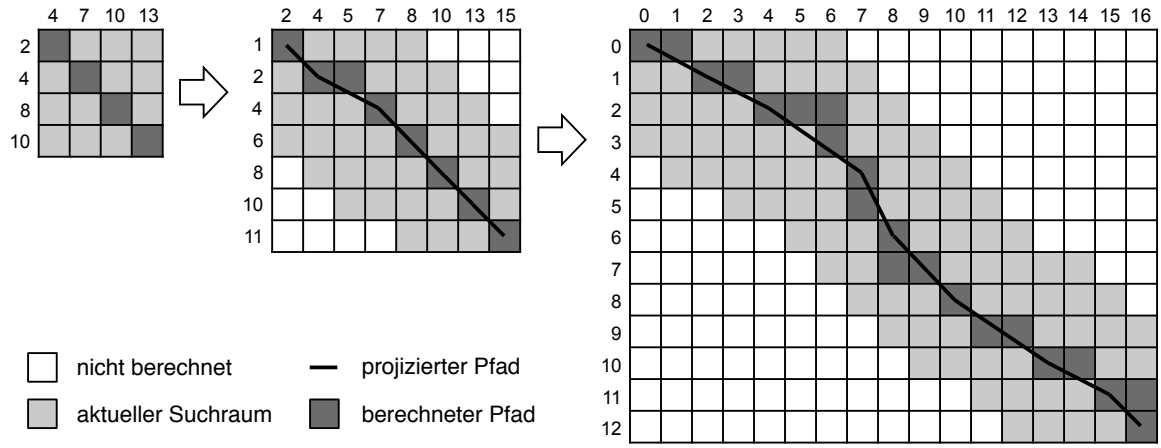

Abb. 11: FastDRTW: Iterative Verfeinerung des Warping-Pfads.
Anhand der groben Datenabstraktion durch eine Auswahl repräsentativer Features wird ein Voralignment berechnet. Für das Alignment der nächstgrößeren Datenabstraktion wird jeweils eine höhere Zahl der repräsentativen Features ausgewählt, für die das Alignment innerhalb eines Fensters in der dünnbesetzten Distanzmatrix berechnet wird. Dafür werden zunächst die Übereinstimmungen des Voralignments auf die Distanzmatrix projiziert und die Projektion um einen Radius erweitert. Schrittweise wird die Auswahl der repräsentativen Features für das Alignment erweitert und das Ergebnis des jeweiligen Voralignments zur Suchraumeinschränkung verwendet. Auf diese Weise wird der Warping-Pfad iterativ verfeinert, bis alle Features aligniert werden können.

benutzerdefinierten Radius, z.B. 200, zu einem Suchraumfenster vergrößert. Das Alignment wird dann innerhalb des eingeschränkten Suchraums berechnet. In folgenden Iterationen wird die Zahl der repräsentativen Features kontinuierlich erhöht, z.B. verdoppelt und der Warping-Pfad immer weiter verfeinert, bis die vollständigen Datensätze aligniert werden können.

Die iterative Pfadverfeinerung reduziert die Wahrscheinlichkeit eines fehlerhaften Warping-Pfades beim Alignment der Ausgangsdaten und steigert die Effizienz gegenüber dem Originalalgorithmus. Neben der Steigerung der Effizienz ermöglicht der FastDRTW-Algorithmus das Alignment von größeren Feature-Sequenzen durch die Reduktion des Speicherbedarfes, da die Distanzmatrix für die dynamische Programmierung nur partiell berechnet wird und nicht mehr vollständig im Speicher abgebildet werden muss. Die Genauigkeit der Methode hängt von der Wahl eines nicht zu kleinen Fensterradius ab. Gleichzeitig kann sich die Fenstergröße negativ auf die Effizienz des Algorithmus auswirken. Das Gleichgewicht zwischen einer akzeptablen Effizienz und einer akzeptablen Genauigkeit kann durch Optimierung des Radius-Parameters erreicht werden. Der Speicherbedarf für das FastDRTW-Alignment von Sequenzen mit einigen hunderttausenden Features und bei Verwendung großer Fensterradien kann beim Einsatz moderner Hardware die verfügbaren Ressourcen übersteigen.

3.3.4 Linear Dynamic Retention Time Warping

Einen neuen Weg zur Optimierung der dynamischen Programmierung auf der Basis des “teile und herrsche”-Prinzipes beschreibt der Hirschberg-Algorithmus, der ursprünglich zur Lösung des LCS-Problems (engl. longest common subsequence) entwickelt wurde[161]. Basierend auf dem Prinzip des Hirschberg-Algorithmus und der zuvor beschriebenen Distanzfunktion (s. Kapitel 3.3.2) zur Berechnung der Feature-Unähnlichkeiten wird ein neuer Algorithmus das Linear Dynamic Retention Time Warping (LinDRTW) definiert. Der LinDRTW-Algorithmus löst das Problem des paarweisen Retentionszeitalignments auf linearem Speicher.

Für die Berechnung einer beliebigen Position der Distanzmatrix werden beim DRTW nur die Distanzen der drei direkten Vorläufer an der vertikalen, horizontalen und diagonalen Nachbarpositionen benötigt. Die Distanzmatrix könnte in beide Richtungen berechnet werden. Für das Ergebnis spielt die Richtung der Berechnung jedoch keine Rolle. Die Vorwärtskalkulation, d.h. ausgehend von der ersten in Richtung der letzten Position der Distanzmatrix, bewertet auf dem Weg zum Gesamtergebnis die Unähnlichkeiten der Präfixe, also der Anfänge der Eingabesequenzen. Umgekehrt berechnet die Rückwärtskalkulation, d.h. ausgehend von der letzten in Richtung der ersten Position der Distanzmatrix, die Unähnlichkeiten der Suffixe, d.h. der Enden der Eingabesequenzen. In beiden Fällen wird die gesamte Unähnlichkeit der beiden Eingabesequenzen berechnet. Diese Tatsachen macht sich der LinDRTW-Algorithmus zunutze, indem für die dynamische Programmierung lediglich zwei Vektoren, für die zuletzt berechnete und die zu berechnende Zeile der Distanzmatrix im Speicher vorgehalten werden. Wie in Abbildung 12 gezeigt, wird die Berechnung der Distanzen in beide Richtungen gleichzeitig ausgeführt. Dazu wird eine der Eingabesequenzen in zwei etwa gleiche Teile, einen Präfix und einen Suffix getrennt. Für den Präfix und die zweite Eingabesequenz wird die Unähnlichkeitsdistanz vorwärts berechnet. Für den Suffix und die zweite Eingabesequenz wird die Unähnlichkeitsdistanz rückwärts berechnet. Das Ergebnis der einzelnen Subalgorithmen ist der jeweils zuletzt berechnete Distanzvektor und das gemeinsame Ergebnis ist die Summe dieser Distanzvektoren. Die Position des Minimums im Summenvektor wird zur Spaltung der zweiten Eingabesequenz in ein Präfix und ein Suffix genutzt. Im nächsten Schritt wird das Alignmentproblem in zwei Unterprobleme aufgeteilt. Das erste Unterproblem stellt das Alignment der Präfixe der beiden Eingabesequenzen dar und das zweite Unterproblem ist das Alignment ihrer Suffixe. Im trivialen Fall wird das jeweilige Alignment mit der Anwendung des ursprünglichen DRTW-Algorithmus gelöst, anderenfalls

wird der LinDRTW-Algorithmus rekursiv angewandt. Die Konkatenation der Einzellösungen jeweiliger Unterprobleme liefert dann die Gesamtlösung des Ausgangsproblems.

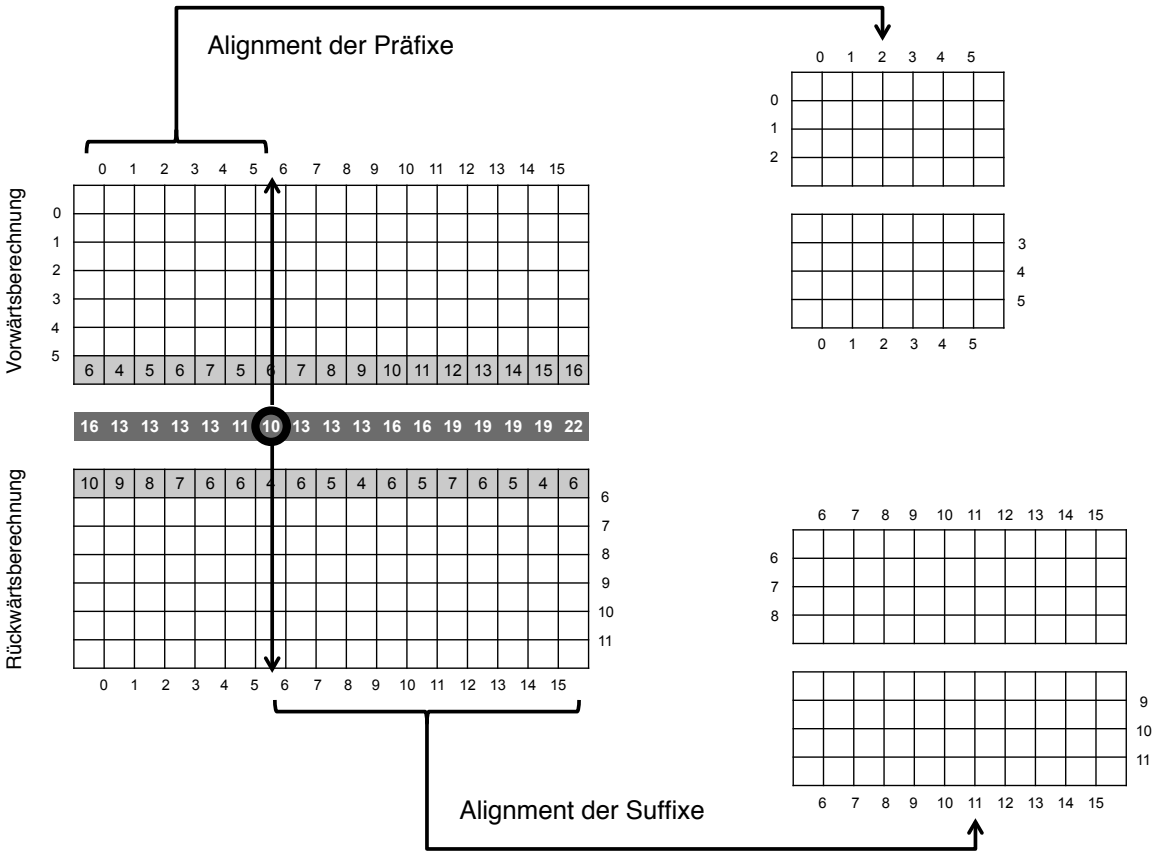

Abb. 12: LinDRTW: Teile und Herrsche.
Eine der beiden Eingabesequenzen wird etwa in der Mitte in ein Präfix und ein Suffix aufgeteilt. Für das Präfix und die gesamte andere Eingabesequenz wird ihre Unähnlichkeit mit Hilfe der dynamischen Programmierung ausgehend vom Anfang der Sequenzen (vorwärts) berechnet. Analog wird für das Suffix und die gesamte andere Eingabesequenz die Unähnlichkeit ausgehend von den Sequenzenden (rückwärts) berechnet. Die Ergebnisse der beiden Berechnungen sind Vektoren der jeweiligen Unähnlichkeitsdistanzen. Die beiden Vektoren werden summiert. Ein optimales Alignment der Eingabesequenzen an der Aufteilungsposition der ersten Eingabesequenz verläuft durch das Minimum des Summenvektors. Die zweite Eingabesequenz wird an dieser berechneten Position geteilt. Das Ausgangsproblem wird in zwei Unterprobleme aufgeteilt: das Alignment der Präfixe und das Alignment der Suffixe. Die beiden Unterprobleme werden nach dem gleichen Prinzip gelöst. Die rekursive Aufteilung in Unterprobleme erfolgt bis ihre Lösung trivial ist.

Durch den linearen Speicherbedarf $\mathcal{O}(n)$ ermöglicht der LinDRTW-Algorithmus die Berechnung des paarweisen Retentionszeitalignments von Sequenzen mit einigen Millionen Features ohne besondere Anforderungen an die Hardware. LinDRTW schränkt die Genauigkeit des berechneten Alignments gegenüber dem DRTW-Algorithmus nicht ein. Der Algorithmus weist formal eine quadratische Zeitkomplexität $\mathcal{O}(n^2)$ auf. Jedoch geht die Linearisierung des Speicherbedarfes mit einer moderaten Erhöhung der Laufzeit zur rekursiven Berechnung der Unterprobleme einher. Ein weiterer Vorteil des LinDRTW-

Algorithmus, der sich aus dem “teile und herrsche”-Prinzip ergibt, ist die Möglichkeit der einfachen Parallelisierung der Berechnung. Die Unterprobleme können unabhängig voneinander, simultan gelöst werden. Die Ausführung des Algorithmus lässt sich auf moderne Multiprozessor-Systeme (Mehrkern-Hauptprozessoren oder Vielkern-Graphikprozessoren) übertragen.

3.3.5 Fast Linear Dynamic Retention Time Warping

Die in den Kapiteln 3.3.4 und 3.3.5 beschriebenen Algorithmen FastDRTW und LinDRTW helfen Teilprobleme beim Retentionszeitalignment von Feature-Listen mit Hilfe des DRTW zu lösen. Durch eine Kombination ihrer Prinzipien zur Optimierung der Effizienz der dynamischen Programmierung lässt sich ein neuer Algorithmus für das paarweise Retentionszeitalignment von Feature-Sequenzen das Fast Linear Retention Time Warping (FastLinDRTW) definieren. FastLinDRTW schränkt analog zum FastDRTW-Algorithmus den Suchraum ein, indem zunächst das Alignment einer groben Repräsentation der Feature-Sequenzen berechnet wird. Das Ergebnis wird als Suchraumeinschränkung für das Alignment der weniger groben Datenrepräsentation projiziert und nach beiden Seiten um einen benutzerdefinierten Radius vergrössert. Die Projektion der Suchraumeinschränkung erfolgt auf eine gedachte Distanzmatrix, während der linearisierten dynamischen Programmierung wird tatsächlich jedoch der Bereich des jeweiligen Distanzvektors bearbeitet, der sich innerhalb des eingeschränkten Suchraumes befindet. Die Suchraumeinschränkung wird in mehreren Stufen mit der Steigerung der repräsentativen Datenmenge verfeinert. Schließlich wird das Alignment der Ausgangsdaten im iterativ eingeschränkten Suchraum berechnet. Das Voralignment anhand der gröbsten Datenabstraktion wird mit LinDRTW berechnet, alle folgenden Iterationen setzen den FastLinDRTW-Algorithmus ein.

Der FastLinDRTW-Algorithmus aligniert zwei Feature-Sequenzen auf linearem Speicher (Platzkomplexität $\mathcal{O}(n)$). Die schrittweise Verfeinerung der Suchraumeinschränkung reduziert die Laufzeit des Algorithmus und nähert diese theoretisch der linearen Zeitkomplexität $\mathcal{O}(n^2)$ an. Die Genauigkeit der Methode hängt von der Wahl eines nicht zu kleinen Fensterradius ab. Genau wie LinDRTW kann auch FastLinDRTW einfach parallelisiert werden. In der Praxis profitiert der FastLinDRTW-Algorithmus hinsichtlich der Gesamtlaufzeit von der iterativen Verfeinerung des Warping-Pfades mehr als FastDRTW, da bereits die Berechnung der Voralignments parallelisiert werden kann.

3.3.6 Effizienzvergleich der Retentionszeitalignmentalgorithmen

Implementierungen der vorgestellten Algorithmen DRTW, FastDRTW, LinDRTW und FastLinDRTW wurden auf ihre Effizienz hinsichtlich tatsächlicher Laufzeit und des Speicherverbrauchs untersucht. Der Effizienzvergleich basiert auf dem Alignment unterschiedlich großer Feature-Sequenzen unter Einhaltung gleicher Bedingungen. Abbildung

13 zeigt die gemittelten Messwerte der Laufzeit sowie des Speicherbedarfes der Algorithmen DRTW, FastDRTW, LinDRTW und FastLinDRTW in Abhängigkeit von der Größe der Eingabesequenz. Zur Speicherung der jeweils benötigten Datenstrukturen standen zur Laufzeit 12 GB Arbeitsspeicher zur Verfügung. Die Algorithmen wurden auf einem CPU-Kern (Xeon, 3.2GHz) ausgeführt. Zur besseren Vergleichbarkeit wurde auf die Parallelisierung der Algorithmen LinDRTW und FastLinDRTW verzichtet. Für die Warping-Pfad-Verfeinerung der Algorithmen FastDRTW und FastLinDRTW wurde der Radius-Parameter mit einem konstanten Wert von 500 verwendet. Die iterative Verfeinerung anhand von 1000, 2000, 5000, 10000, 20000, 50000 und 100000 repräsentativen Features durchgeführt, jeweils bis exklusive der Größe der Eingabesequenzen, z.B. 1000, 2000 und 5000 bei Eingabegrößen von 6000 bis 10000 Features je Sequenz. Jede Messung wurde dreimal wiederholt und die Laufzeit sowie der Speicherbedarf automatisiert aufgezeichnet.

Bei Eingabegrößen über 20000 Features überstieg der Speicherbedarf des DRTW-Algorithmus die verfügbaren Ressourcen. Durch die Einschränkungen der Implementierung einer dünnbesetzten Matrix für den FastDRTW-Algorithmus konnte dessen Effizienz bei Eingabegrößen über 30000 Features nicht erfasst werden. Sowohl die Laufzeit als auch der Speicherbedarf des nicht optimierten DRTW-Algorithmus steigen exponentiell mit der Größe der Eingabesequenzen an. Die Laufzeit des FastDRTW-Algorithmus steigt linear mit der Größe der Eingabesequenzen an. Der Speicherbedarf des FastDRTW steigt in weiten Bereichen linear bis maximal etwa 2 GB an, im Bereich der Eingabegrößen von 6000 bis 10000 Features kann jedoch ein stärkerer Anstieg des Speicherbedarfs beobachtet werden. Der Speicherbedarf der Algorithmen LinDRTW und FastLinDRTW ist linear und bleibt bei maximalen Eingabegrößen unter 100 MB. Die Laufzeit des LinDRTW Algorithmus bleibt in weiten Bereichen bis zur Eingabegröße von 30000 Features niedrig, steigt dann aber exponentiell an und erreicht 776 s bei der maximalen, getesteten Eingabegröße von 200000 Features. Die Laufzeit des FastLinDRTW-Algorithmus bleibt über den getesteten Bereich der Eingabegrößen linear. Für das Alignment von zwei Sequenzen von jeweils 200000 Features benötigt FastLinDRTW im Durchschnitt 30,7 s.

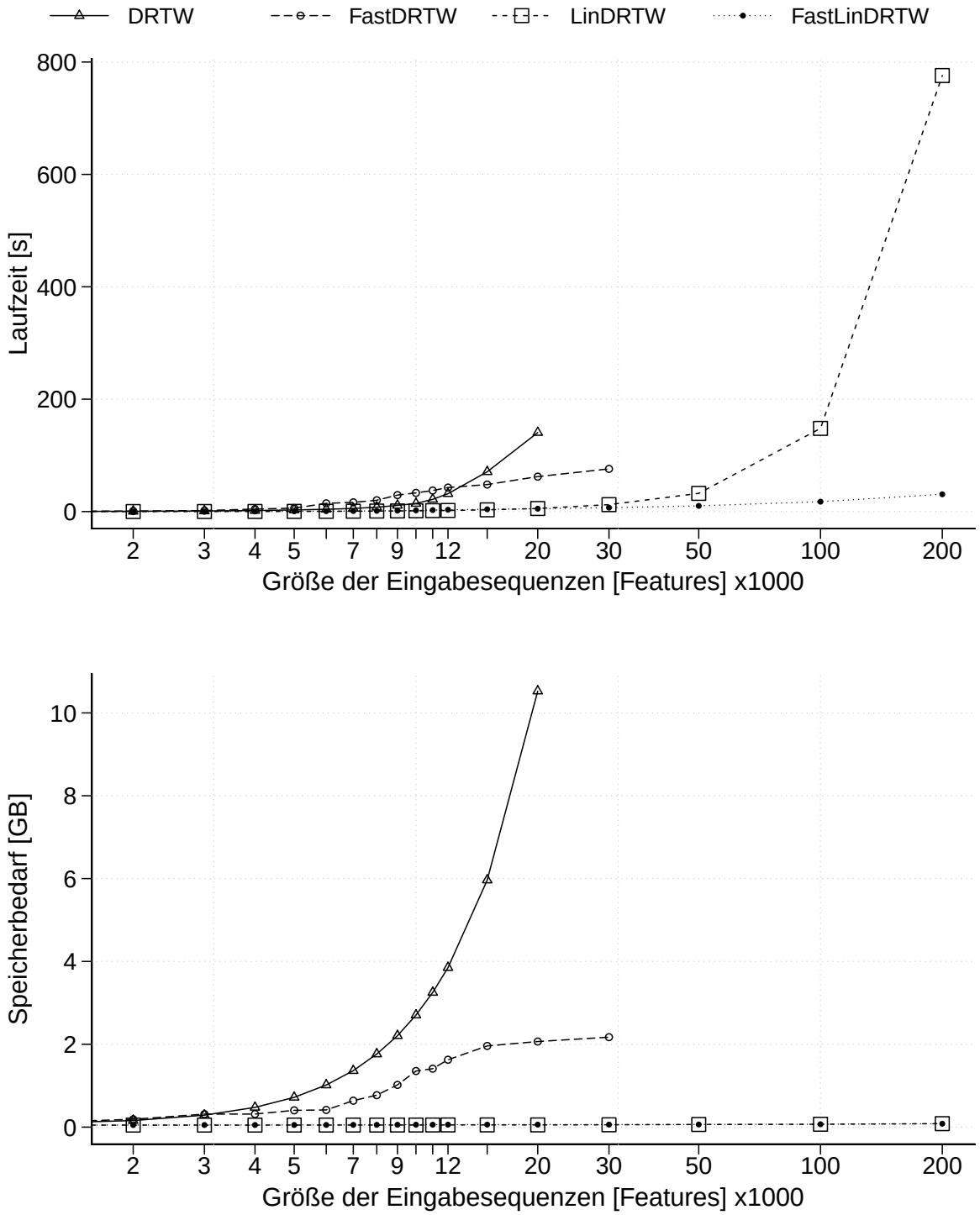

Abb. 13: Algorithmen für das Retentionszeitalignment im Effizienzvergleich.
Im oberen Teil der Abbildung werden die Laufzeiten und im unteren Teil der Speicherbedarf der Algorithmen DRTW, FastDRTW, LinDRTW und FastLinDRTW in Abhängigkeit von der Größe der Eingabesequenzen dargestellt. Da LinDRTW und FastLinDRTW den gleichen Speicherbedarf aufweisen, überlagern sich die entsprechenden Werte im unteren Teil der Abbildung.

3.3.7 Multiples Retentionszeitalignment

Das paarweise Retentionszeitalignment ermöglicht die Analyse von Zeitverschiebungen zwischen zwei beliebigen LC-MS-Messungen eines Experiments. Das Ergebnis ist die Angabe von Zeitverschiebungen einer LC-MS-Messung gegenüber einer anderen LC-MS-Messung zu den Zeitpunken, an denen übereinstimmende Features in beiden Messungen gefunden wurden. Um ein multiples Retentionszeitalignment zu erreichen, werden mit Hilfe des paarweisen Retentionszeitalignments Retentionszeitverschiebungen aller Messungen des Experiments gegenüber einer gemeinsamen Referenzmessung untersucht. Als Referenz wird die Messung mit den meisten detektierten Features ausgewählt. Für alle anderen Messungen des Experiments wird das entsprechende paarweise Retentionszeitalignment gegenüber dieser Referenz berechnet. Die Übereinstimmungen im paarweisen Alignment einer Messung ergeben ein Mapping der jeweiligen Retentionszeit auf die Referenz, für die Zeitpunkte zwischen den Übereinstimmungen wird die Zeitverzerrung durch lineare Interpolation geschätzt. Auf diese Weise kann zu jedem Zeitpunkt einer LC-MS-Messung des Experiments die entsprechende Referenz-Retentionszeit berechnet werden.

Die Referenz-Retentionszeit leitet sich bei dieser Vorgehensweise von der Auswahl der Referenzmessung ab und kann lokal starke Verschiebungen der Retentionszeit gegenüber anderen Messungen des Experiments aufweisen. Um den Einfluss des durch die Referenz eingebrachten Fehlers gering zu halten, können die Zeitverschiebungen über den gesamten Verlauf der Referenz-Retentionszeit um den Median der Verschiebung zu jedem Zeitpunkt normalisiert werden. Die berechneten Referenz-Retentionszeiten werden als Retentionszeiten der Features angenommen, so dass sich für die korrespondierenden Features in allen Messungen entsprechend ähnliche Retentionszeiten ergeben. Durch die lineare Interpolation können sich jedoch Zeitverschiebungen zwischen den Übereinstimmungen der Alignments ergeben. Abbildung 14 zeigt die Verschiebungen der Retentionszeit innerhalb eines Experimentes.

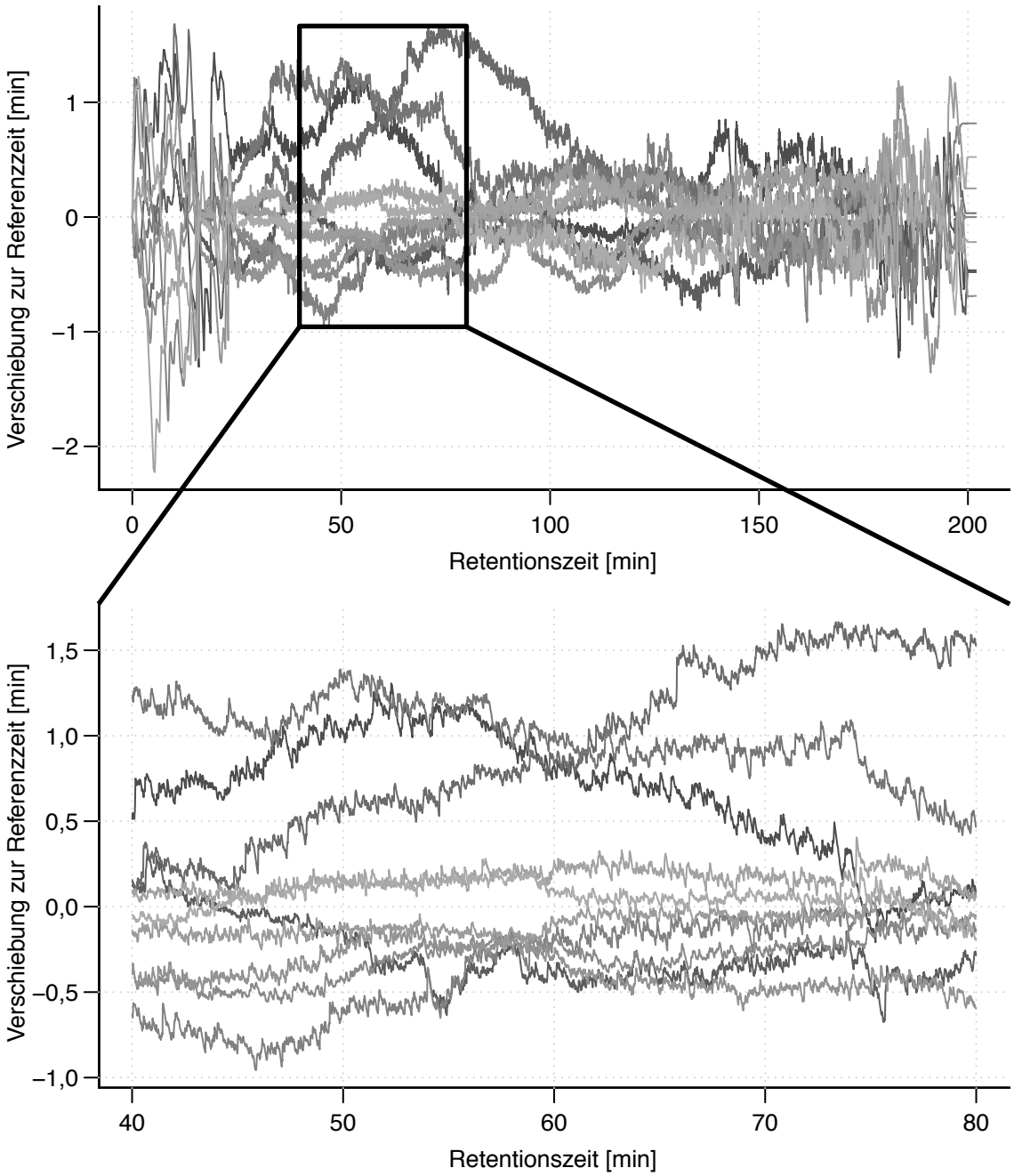

Abb. 14: Multiples Retentionszeitalignment.
Verzerrungen der Retentionszeiten innerhalb eines Experiments (HeLa-Hefe-*E.coli*-Metaproteom, 10 UDMS^{E} Messungen, zwei Proben mit je fünf technischen Replikaten, 200 min Laufzeit). Die paarweisen Retentionszeitalignments wurden mit dem FastLinDRTW-Algorithmus berechnet. Die Zeitverschiebungen zwischen den Übereinstimmungen wurden linear interpoliert. Der resultierende Zeitversatz zur Referenz wurde zu jedem Zeitpunkt zum Median aller Retentionszeitverschiebungen an diesem Zeitpunkt normalisiert. Im oberen Teil der Abbildung werden die Zeitverschiebungen aller 10 Messungen des Experiments über die gesamte Laufzeit dargestellt. Um die Sichtbarkeit der lokalen Verschiebungen zu verbessern, werden im unteren Teil der Abbildung die Zeitverschiebungen aller 10 Messungen des Experiments im eingeschränkten Retentionszeitbereich (Minuten 40 bis 80) dargestellt. Jede Linie repräsentiert die Funktion der Zeitverschiebung einer der 10 Messungen des Experiments.

3.4 Feature-Clustering

Mit Hilfe des Retentionszeitalignments wurden die Retentionszeiten von korrespondierenden Features in einen über alle Messungen gleichen Referenzzeitraum transformiert. Dies setzt sie jedoch nicht direkt zueinander in Beziehung. Mit dem sogenannten Feature-Clustering werden die Features aus unterschiedlichen Messungen auf ihre Übereinstimmung untersucht und in Clustern gruppiert.

Die korrespondierenden Features innerhalb eines Clusters weisen ähnliche Werte ihrer Massen, Retentionszeiten und Driftzeiten mit einer durch experimentelle Bedingungen vorgegebenen Varianz auf. Die Intensitätswerte sind proportional zu der Peptidmenge in der jeweiligen Probe und können in unterschiedlichen Messungen des Experiments stark voneinander abweichen. Die erwartete Abweichung der Retentionszeit hängt von der lokalen Qualität des Retentionszeitalignments ab und beträgt i.d.R. weniger als 0,15 bis 0,3 Minuten. Typischerweise sind die Abweichungen der Masse und der Driftzeit instrumentenabhängig. So kann die Massenabweichung bei Messungen mit dem Premier Q-TOF 10 bis 15 ppm und bei Messungen mit dem Synapt G2/G2S 4 bis 8 ppm betragen. Die Driftzeit hängt zudem von den Instrumenteneinstellungen ab, werden diese innerhalb eines Experimentes nicht verändert, so kann eine Abweichung der Driftzeit für das Synapt G2/G2S von höchstens zwei Einheiten beobachtet werden.

Die Abstraktion eines Features als EMRT erlaubt seine einfache Darstellung als Punkt in einem mehrdimensionalen geometrischen Raum. Die räumlichen Koordinaten werden von der Masse, der Retentionszeit sowie für IMS-MS-Daten der Driftzeit abgeleitet. Bei der Betrachtung aller Features eines Experiments in einem gemeinsamen geometrischen Raum zeichnen sich die zusammengehörigen Features durch räumliche Nähe zueinander aus. Diese Beobachtung erlaubt die Anwendung geometrischer Clustering-Verfahren für die Gruppierung übereinstimmender Features.

3.4.1 Preclustering

Auf Grund der großen Datenmengen in Proteomik-Experimenten wurde eine divisive Clustering-Methode Single-Link[162, 163] zur Unterteilung des Datenvolumens in Untergruppen entwickelt. Ordnet man eine Gruppe von Features nach einer ihrer Eigenschaften etwa der Masse, der Retentionszeit oder der Driftzeit, so kann eine Zusammengehörigkeit zweier benachbarten Features ausgeschlossen werden, wenn die Differenz der betrachteten

Eigenschaft einen benutzerdefinierten Schwellenwert übersteigt. An dieser Stelle können die Features der ursprünglichen Gruppe in zwei unabhängige Untergruppen aufgeteilt werden. Findet man in den vorliegenden Features alle solchen Lücken, können die Ausgangsdaten entsprechend in mehrere Untergruppen aufgeteilt werden. Diese zunächst eindimensionale divisive Clustering-Strategie wird sequentiell in der Dimension der Masse, der Retentionszeit und der Driftzeit angewandt, um die Feature-Gruppen zu verfeinern. Die sequenzielle Auftrennung in Untergruppen auf Basis unterschiedlicher Eigenschaften wird mehrfach wiederholt, bis keine neuen Gruppen abgespalten werden können.

Da bei jeder Iteration ein Feature lediglich mit seinen direkten Nachbarn verglichen wird, weist der Preclustering-Algorithmus eine lineare Platz- und Zeitkomplexität $\mathcal{O}(n)$ auf. Auf Grund von Verkettungseffekten, die als "single-link chaining phenomenon bekannt sind[164], kann diese Prozedur in komplexen Bereichen nicht die einzelnen Cluster korrespondierender Features isolieren, eignet sich jedoch sehr gut zur Partitionierung der Daten, um das Datenvolumen für die Anwendung komplexerer Clustering-Verfahren zu reduzieren.

3.4.2 Raumtransformation

Auf Grundlage experimentabhängiger Dispersion der Massen, Retentionszeiten und Driftzeiten werden alle Features in einen homogenen, mehrdimensionalen geometrischen Raum projiziert. Die absoluten, monoisotopischen Massen der Features werden in relative ppm Werte umgerechnet und anschließend mit einer benutzerdefinierten Auflösung in eine Massen-Koordinate überführt. Die absoluten Retentions- und Driftzeiten werden mit benutzerdefinierten Auflösungen in entsprechende Koordinaten umgewandelt. Dieser Transformationsalgorithmus berechnet für jedes Feature seine räumlichen Koordinaten (x, y, z).

$$x = \frac{log(mass)}{log(1 + R_{mass})}$$
$$y = \frac{rt}{R_{rt}}$$
$$z = \frac{dt}{R_{dt}}$$

mit $mass$ der Masse, rt der Retentionszeit und dt der Driftzeit eines Features sowie R_{mass}, R_{rt} und R_{dt} den entsprechenden benutzerdefinierten Massen, Retentions- und Driftzeitauflösungen. Für Daten ohne Ionenmobilität und für gemischte Daten mit und ohne

Ionenmobilität wird die Transformation in einen zweidimensionalen geometrischen Raum durchgeführt. Die Koordinaten (x, y) werden anhand der Masse und der Retentionszeit wie beschrieben berechnet.

3.4.3 Dichtebasiertes Clustering

Auf Grund der Annahme von Clustern in Regionen hoher Dichte und der Bereiche niedriger Dichte zwischen den Clustern wird dieser geometrische Raum mit dem dichtebasierten Verfahren Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[165] analysiert. Dabei wird die räumliche Entfernung zwischen zwei beliebigen Punkten A und B durch deren Euklidische Distanz als

$$d(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2}$$

und für Daten ohne Ionenmobilität als

$$d(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

berechnet. Das Verhalten von DBSCAN wird grundsätzlich durch zwei Parameter kontrolliert: eine Größe des Nachbarschaftsradius und eine benutzerdefinierte Mindestzahl der benachbarten Punkte für die Cluster-Expansion. Für das Feature-Clustering wird der Nachbarschaftsradius mit einem konstanten Wert von 1 angenommen, da seine Größe von der Beschaffenheit des geometrischen Raums abhängt und dieser bereits durch die benutzerdefinierten Auflösungen bei der Raumtransformation der Features beeinflusst wird. Bei korrekter Parametrisierung liefern die DBSCAN-Cluster akkurate Gruppierung von korrespondierenden Features über alle Messungen des Experiments. Die als Rauschen eingestuften Features werden als eigenständige Einpunkt-Cluster betrachtet. Die resultierenden Cluster beinhalten jeweils ein Feature aus jeder Messung des Experiments. Für einzelne Messungen können Features fehlen, wenn das entsprechende Peptid in der Messung nicht vorhanden ist, oder seiner niedrigen Abundanz die Signalintensität unterhalb des Detektionslimits liegt. Für manche Messungen können jedoch auch mehrere Features in einem Cluster auftreten, wenn Peptide mit gleicher Masse gleichzeitig eluieren oder wenn bei der Feature-Detektion ein Signal als mehrere Features fehlinterpretiert wird.

3.5 Normalisierung der Feature-Intensitäten

Während der einzelnen Messungen eines LFQ-LC-MS-Experimentes können die experimentellen Bedingungen Schwankungen unterliegen. Durch unterschiedliche Einflüsse können die Intensitäten der Features komplexe systematische Fehler enthalten. Unterschiedliche Abhängigkeiten dieser systematischen Fehler können beobachtet werden.

Die systematischen Fehler der Intensitätswerte werden anhand der sogenannten Logratios, der logarithmischen Verhältnisse der Feature-Intensitäten zu den Referenzintensitäten untersucht. Die Logratios werden auf Basis des Logarithmus dualis, wie folgt, berechnet.

$$Logratio = log_2(\frac{Intensität}{Referenzintensität})$$

Als Referenzintensität kann wahlweise entweder die durchschnittliche Intensität aller Features des jeweiligen gesamten Feature-Clusters, die durchschnittliche Intensität aller Features der technischen Replikate der jeweiligen Probe in einem Feature-Cluster, oder die Intensität einer bestimmten Messung verwendet werden.

Um Tendenzen der systematischen Fehler zu finden, werden Verteilungen der Logratios für jede einzelne Messung des Experiments in Abhängigkeit von den Intensitäten, den Retentionszeiten und den Massen untersucht. Unter der Annahme, dass die meisten Proteine (und deshalb auch ihre tryptischen Peptide) in den verglichenen Proben keine Unterschiede in ihrer Expression aufweisen, sollte der durchschnittliche Wert der Logratios für jeden Intensitätsbereich, zu jeder Retentionszeit und für jeden Massenbereich gleich null sein. Die Verschiebungen der jeweiligen Logratio-Verteilungen werden mit einer nichtlinearen Regression durch Locally Weighted Scatterplot Smoothing (LOWESS)[166] geschätzt. Die Ergebnisfunktion der LOWESS-Regression wird zur Normalisierung der Intensitätswerte eingesetzt. Die Intensität jedes Features einer Messung wird dann in einer Dimension, wie folgt, korrigiert:

$$\tilde{I}_p = \frac{I_p}{2^{R(x_p)}}$$

mit einem Feature p, seiner neuen (normalisierten) Intensität $\tilde{I}_p$, seiner alten Intensität vor Normalisierung I_p, dem Wert einer Eigenschaft x_p des Features p in der analysierten

Dimension und dem Wert der Logratio-Regression $R(x_p)$ am Punkt x_p. Abhängig von der analysierten Dimension kann der Parameter x_p den Wert der logarithmierten Intensität, der absoluten Masse oder der absoluten Retentionszeit eines Features einnehmen. Bei der multidimensionalen Normalisierung werden die Korrekturen der systematischen Fehler in unterschiedlichen Dimensionen sequentiell durchgeführt und so der Gesamtanteil der systematischen Fehler an der Intensität der Features reduziert. Der Algorithmus kann durch die Wahl der Referenz zur Logratio-Berechnung, die Abfolge der untersuchten Dimensionen und die jeweilige LOWESS-Bandbreite gesteuert werden. Der Einfluss der LOWESS-Bandbreite auf den Verlauf der Logratio-Fehlerfunktion wird in Abbildung 15 an einem Beispiel dargestellt. Durch LOWESS-Bandbreiten zwischen 0,0 und 1,0 tendiert der Algorithmus bei kleinen Bandbreiten entsprechend mehr zur Korrektur lokaler systematischer Fehler und bei großen Bandbreiten zur Korrektur globaler systematischer Fehler.

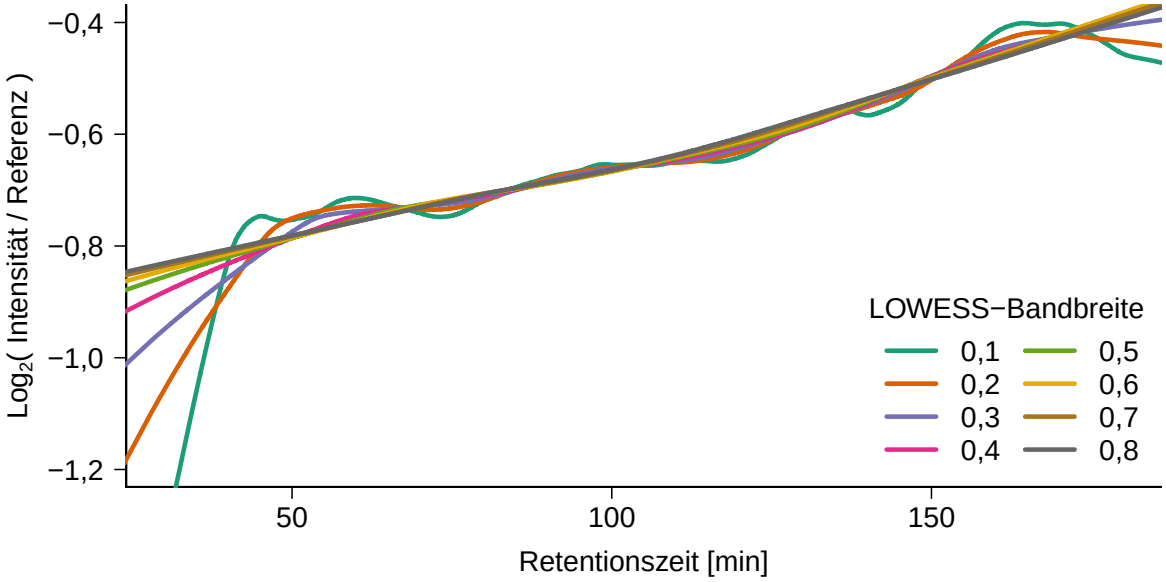

Abb. 15: Einfluss der LOWESS-Bandbreite auf die Logratio-Fehlerfunktion.
Die Abbildung zeigt den Einfluss der LOWESS-Bandbreite auf die Fehlerfunktion der Logratios vor der Normalisierung. Die Logratios wurden als $log_2(\frac{Intensität}{Referenz})$ mit der durchschnittlichen Intensität aller Features des entsprechenden Clusters als Referenz berechnet. Linien stellen den Verlauf der LOWESS-Regression mit unterschiedlichen Bandbreiten als Funktion der Retentionszeit dar. Für die Abbildung wurde ein Auszug aus einem im Kapitel 3.15.1 beschriebenen Testdatensatz (Metaproteom, Probe B, MS^E-Akquisition, 5. Replikatmessung) entsprechend während der Datenanalyse erfasst.

Abbildung 16 visualisiert am Beispiel einer einzelnen LC-MS-Messung die Logratios als Funktion der Intensität und der Retentionszeit vor und nach der multidimensionalen Normalisierung. Der Verlauf der Funktionen der entsprechenden systematischen Fehler werden mit LOWESS vor und nach der Normalisierung geschätzt.

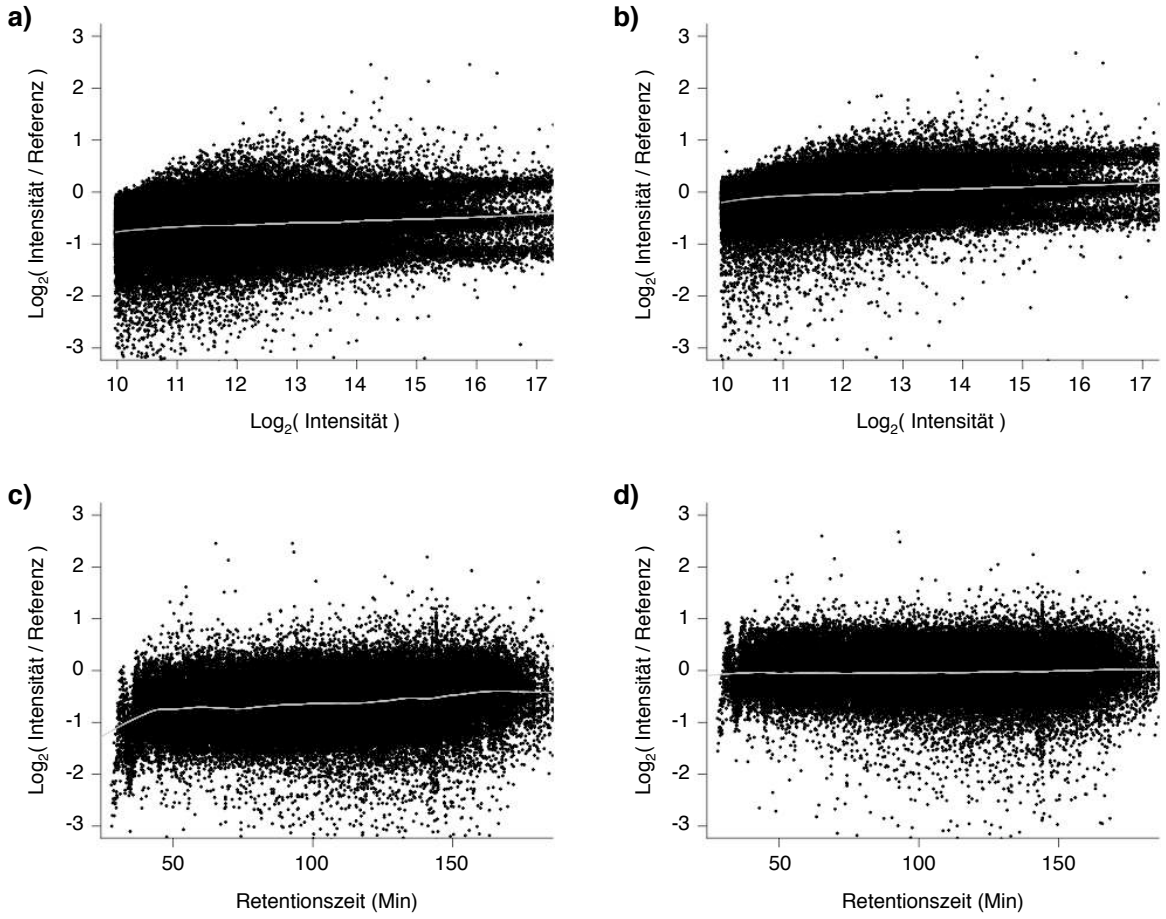

Abb. 16: Multidimensionale Normalisierung der Feature-Intensitäten.
Logratios der Signalintensitäten einer LC-MS-Messung werden als Funktion der logarithmierten Signalintensität (**a** und **b**) oder der Retentionszeit (**c** und **d**) vor der Normalisierung (**a** und **c**) sowie nach der multidimensionalen Normalisierung in der Intensitätsdomäne und der Retentionszeitdomäne (**b** und **d**) dargestellt. Die Logratios wurden als $log_2(\frac{Intensität}{Referenz})$ mit der durchschnittlichen Intensität aller Features des entsprechenden Clusters als Referenz berechnet. Linien stellen den Verlauf der LOWESS-Regression (Bandbreite=0,2) dar. Für die Abbildung wurde ein Auszug aus einem im Kapitel 3.15.1 beschriebenen Testdatensatz (Metaproteom, Probe B, MS^E-Akquisition, 5. Replikatmessung) entsprechend vor und nach der multidimensionalen Normalisierung erfasst.

3.6 Filterung der Peptididentifikationen

Bei der Peptididentifikation mit Identity^{E} in PLGS wird eine Teilmenge der Features einer Messung als Peptid erkannt. Die Peptididentifikation wird dabei für jede Messung unabhängig voneinander durchgeführt. Um gleiche Qualität der Peptididentifikation für alle Messungen des analysierten Experiments zu gewährleisten, werden für die Datenanalyse nur PLGS-Peptididentifikationen verwendet, die benutzerdefinierten Kriterien entsprechen. PLGS-Peptididentifikationen werden nach der Länge der Aminosäuresequenz, dem PLGS-Identifikationsscore, dem Typ der Peptididentifikation[1] und der Replikationsrate, d.h. der Häufigkeit seiner Identifikation, gefiltert. Ausschließlich PLGS-Identifikationen, die den Filterkriterien entsprechen, werden bei der weiteren Analyse berücksichtigt.

3.7 Annotation der Feature-Cluster

Die im Kapitel 3.4 beschriebene Feature-Clustering-Methode gruppiert übereinstimmende Features ohne Rücksicht auf ihre PLGS-Peptididentität. Die gemeinsame Auswertung der PLGS-Peptididentitäten der Features eines Clusters erlaubt die Ermittlung einer Konsensusidentität, die zur Annotation des gesamten Feature-Clusters genutzt wird. Während der Konsensusbildung können mehrere unterschiedliche Szenarien auftreten, die eine Annotation des gesamten Clusters erlauben oder verhindern können. Ein Cluster wird annotiert, wenn eine benutzerdefinierte Mindestzahl der darin enthaltenen Features durch PLGS identifiziert wurden und ihre Peptididentitäten übereinstimmen. Wird die Mindestzahl der Identifikationen nicht erreicht, gilt der Cluster und alle seine Features als nicht annotiert. Treten innerhalb eines Clusters mehrere unterschiedliche Peptididentitäten auf, wird bei der restriktiven Cluster-Annotation eine Fehlidentifikation seiner Features angenommen und die Annotation des Clusters verworfen. Die Regeln der restriktiven Cluster-Annotation werden in Abbildung 17 anhand mehrerer Beispielsituationen verdeutlicht.

[1] Nach der Art der Identifikation markiert PLGS die Peptide mit folgenden Identifikationstypen: `PEP_FRAG_1`, `PEP_FRAG_2`, `IN_SOURCE`, `MISSING_CLEAVAGE`, `VAR_MOD`, `NEUTRAL_LOSS-H20`, `NEUTRAL_LOSS-NH3`.

Cluster	PLGS-Peptididentifikation in LC-MS-Messungen 1	2	3	4	5	6	Cluster-Annotation
1	Peptid A	Peptid A	Peptid A	Peptid A	Peptid A	Peptid A	Peptid A
2	Peptid B	Peptid B	Peptid B	-	-	-	Peptid B
3	Peptid C	Peptid C	Peptid C	ni	ni	ni	Peptid C
4	Peptid D	Peptid D	-	-	ni	ni	Peptid D
5	Peptid E	-	-	ni	ni	ni	-
6	ni	ni	ni	ni	ni	ni	-
7	Peptid F	Peptid F	Peptid F	Peptid G	Peptid G	Peptid G	-
8	Peptid H	Peptid H	Peptid H	Peptid H	Peptid H	Peptid I	-
9	Peptid J	Peptid J	Peptid K	Peptid L	Peptid M	Peptid N	-

Abb. 17: Entscheidungsbeispiele der restriktiven Cluster-Annotation.
Anhand von fiktiven Beispielen werden die Entscheidungsmuster der restriktiven Cluster-Annotation mit mindestens zwei konfliktfrei identifizierten Features per Cluster visualisiert. Fehlende Features werden mit dem Zeichen "-", nicht identifizierte Features mit "ni" dargestellt. Cluster 1 bis 4 werden jeweils mit einer eindeutigen Peptididentität annotiert. In den Clustern 3 und 4 wird die Peptidannotation auf die nicht identifizierten Features übertragen. Cluster 5 bis 9 werden nicht annotiert. Für Cluster 5 reicht die schwache Evidenz durch die PLGS-Identifikation in lediglich einer Messung für die Cluster-Annotation nicht aus. Für Cluster 6 liegt keine Identifikation vor. Konflikte zwischen der Peptididentifikationen in Clustern 7 bis 9 verhindern eine eindeutige Annotation.

Durch eine benutzerdefinierte maximale Zahl unterschiedlicher Peptididentitäten pro Cluster kann alternativ zur restriktiven Konfliktauflösung eine kompetitive Cluster-Annotation durchgeführt werden. Übersteigt dabei die Zahl der unterschiedlichen Peptididentifikationen innerhalb eines Clusters den benutzerdefinierten Schwellenwert nicht, werden für jede Peptididentifikation die in unterschiedlichen Messungen erreichten PLGS-Identifikationsscores summiert und der Cluster durch das Peptid mit der höchsten Scoresumme annotiert.

In Abbildung 18 wird die Auswirkung der restriktiven Cluster-Annotation auf die Anzahl korrekter Peptid- und Proteinidentifikationen in Abhängigkeit von der FDR vor und nach der restriktiven Annotation der Feature-Cluster am Beispiel einzelner MS^E- sowie $HDMS^E$-Analysen des HeLa-Proteoms (180 min Gradientenzeit)[98] dargestellt.

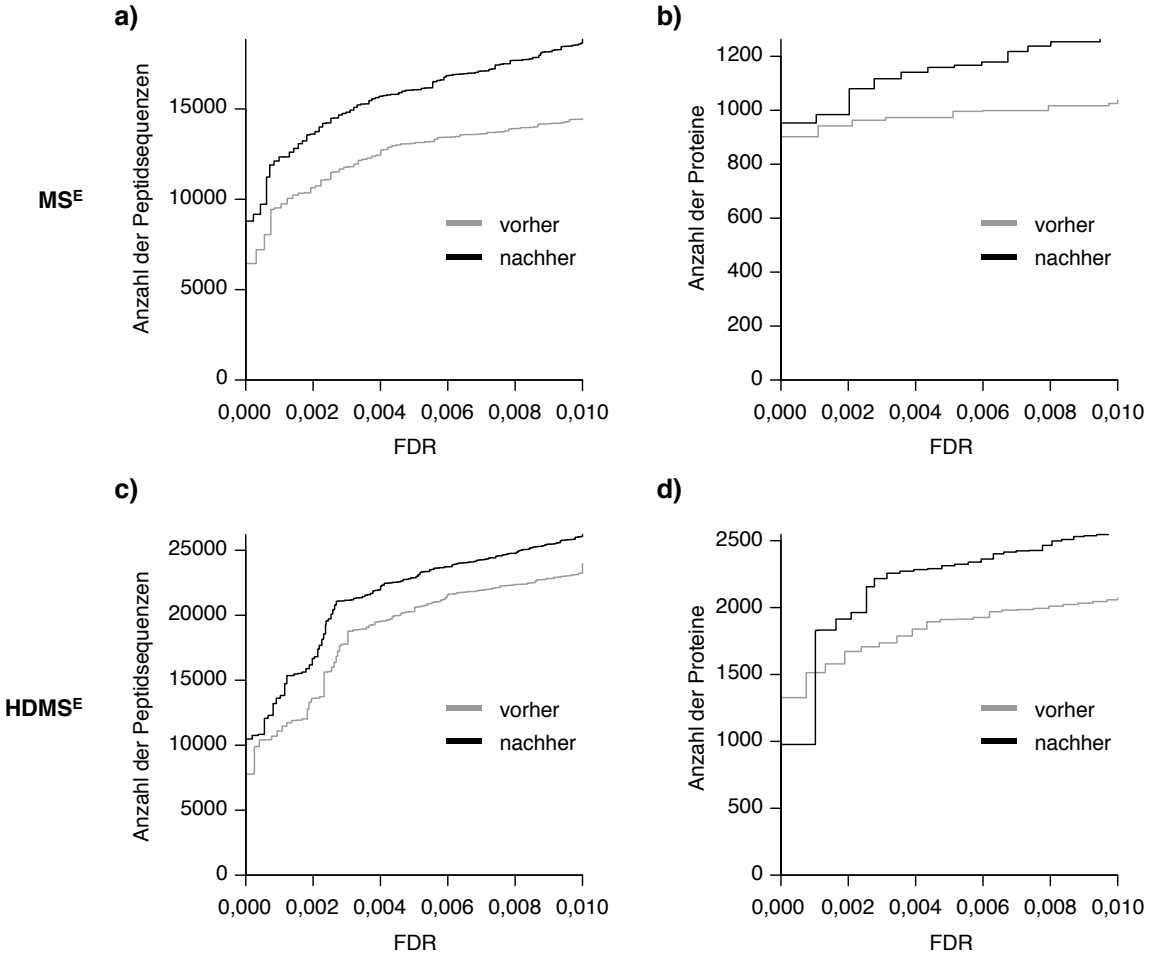

Abb. 18: Anzahl der korrekten Identifikationen als Funktion der FDR.
Die Anzahl der korrekt identifizierten, nicht-redundanten Peptidsequenzen (**a** und **c**) sowie die Anzahl der korrekt identifizierten Proteine (**b** und **d**) werden jeweils als Funktionen der geschätzten FDR (bis 1%) für das Proteom des HeLa-Lysats akquiriert mit MS^E (**a** und **b**) und $HDMS^E$ (**c** und **d**) mit 180 min Gradientenzeit vor (graue Linie) und nach der restriktiven Feature-Cluster-Annotation (schwarze Linie) dargestellt.

3.8 Filterung der Peptid-FDR

Die Peptid- und Proteinidentifikation in PLGS erfolgt für jede LC-MS-Messung eines Experiments unabhängig. Die Identifikationsergebnisse können dabei Anteile falscher Peptid- und Proteinidentifikationen enthalten, die aus dem Decoy-Teil der verwendeten Suchdatenbank stammen. Bei der Feature-Cluster-Annotation könnten falsche Identifikationen zwischen den Messungen übertragen werden. Um die Qualität der Peptididentifikation zu harmonisieren, wird für alle Feature-Cluster-Annotationen die FDR als Anteil der falsch-positiven Identifikationen an der Gesamtzahl der betrachteten Identifikationen auf Peptidebene berechnet und die Einhaltung eines benutzerdefinierten FDR-Schwellenwertes sichergestellt.

Für jedes Peptid wird zunächst sein über alle Messungen des Experiments maximal erreichter PLGS-Identifikationsscore ermittelt und seine Identifikation als wahr- oder falsch-positiv eingestuft. Wird ein Peptid in einer beliebigen Messung einem Decoy-Protein zugeordnet, wird es für die FDR-Berechnung als falsch-positive Identifikation markiert. Eine FDR-Verteilung wird als Funktion der Zuverlässigkeit der Identifikation einer Menge von Peptiden ermittelt, indem für jeden beobachteten Identifikationsscore die FDR ausschließlich anhand der Peptide berechnet wird, die mindestens mit diesem oder einem höheren Score identifiziert wurden. Der höchste Score, an dem der benutzerdefinierte FDR-Schwellenwert erreicht oder überschritten wird, dient als Filtergrenze. Ausschließlich Peptide, deren maximaler PLGS-Identifikationsscore über der FDR-Filtergrenze liegt, werden bei weiteren Schritten der Datenanalyse berücksichtigt.

3.9 Protein-Homologie-Filter

Das Proteininferenz-Problem (s. Kapitel 1.8.1) verhindert in vielen Fällen eindeutige Zuordnungen der Peptididentifikationen zu Proteinen. Mit einem neuen Algorithmus, dem Protein-Homologie-Filter wird die Komplexität der Peptid-Protein-Beziehungen untersucht und reduziert. Der Ablauf des Algorithmus wird in Abbildung 19 schematisch dargestellt.

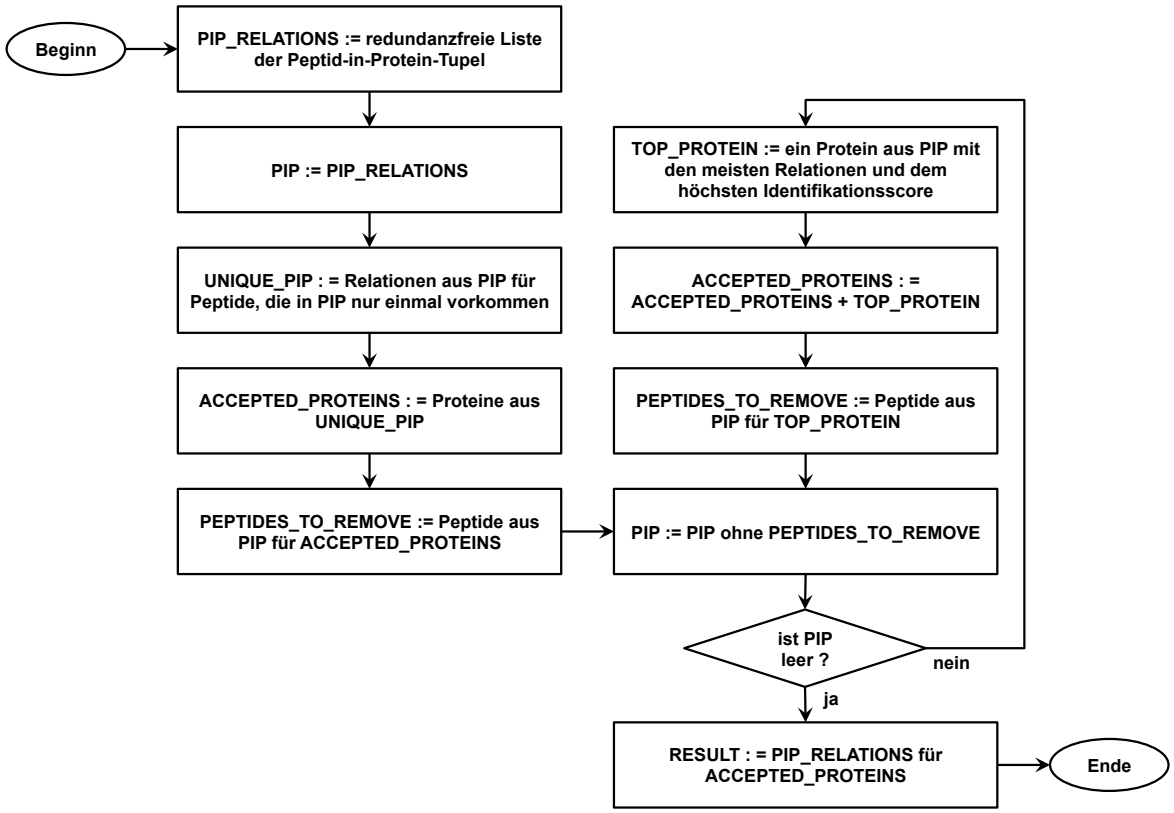

Abb. 19: Protein-Homologie-Filter.
Das Diagram beschreibt den Ablauf des Algorithmus Protein-Homologie-Filter. Aus der Liste aller Peptid-Protein-Beziehungen werden zunächst alle Proteine ausgewählt, die mindestens ein ausschließlich ihnen zugeordnetes Peptid haben. Diese eindeutig identifizierten Proteine werden der Ergebnisliste hinzugefügt. Alle direkten Beziehungen der Peptide, die eine Relation zu den eindeutig identifizierten Proteinen haben, werden aus der Liste der Peptid-Protein-Beziehungen entfernt. Im nächsten Schritt wird aus den verbleibenden Peptid-Protein-Beziehungen ein Protein ausgewählt, für das die höchste Wahrscheinlichkeit der richtigen Identifikation vorliegt. Dieses Protein wird der Ergebnisliste hinzugefügt. Alle direkten Relationen seiner Peptide werden aus der Liste der Peptid-Protein-Beziehungen entfernt. Die iterative Auswahl der am wahrscheinlichsten identifizierten Proteins und die Reduktion des Beziehungsnetzwerks werden solange wiederholt, bis alle Relationen aus der Liste Peptid-Protein-Beziehungen entfernt wurden.

Aus einer Liste aller vorhandenen Peptid-Protein-Beziehungen akzeptiert der Protein-Homologie-Filter zunächst nur die Proteine, für die es eindeutige Peptididentifikationen gibt. Entfernt man die Relationen ihrer Peptide aus der Liste der Peptid-Protein-Beziehungen, weisen die verbleibenden Proteine innerhalb einzelner Beziehungsnetzwerke Homologien auf der Ebene der Peptidsequenzen auf. Aus den verbleibenden Peptid-

Protein-Beziehungsnetzwerken werden dann iterativ einzelne Proteine ausgewählt, deren Identifikation innerhalb der jeweiligen Iteration am wahrscheinlichsten ist. Dies ist der Fall für das Protein mit der höchsten Zahl der verbleibenden Peptid-Relationen in der Liste der Peptid-Protein-Beziehungen. Liegen mehrere Proteine mit der gleichen Zahl der Relationen vor, wird aus ihnen das Protein mit dem höchsten PLGS-Identifikationsscore ausgewählt. Dieses Protein wird akzeptiert und die Beziehungen aller seiner Peptide werden aus der Liste der Peptid-Protein-Beziehungen entfernt. Die iterative Auswahl von Proteinen und die Reduktion des Peptid-Protein-Beziehungsnetzwerks werden solange fortgesetzt, bis alle Relationen aus der Liste Peptid-Protein-Beziehungen entfernt wurden. Als Konsequenz akzeptiert der Algorithmus neben den eindeutig identifizierten Proteinen jeweils nur ein Protein einer Familie homologer Proteine. Für die weitere Datenverarbeitung werden ausschließlich durch Protein-Homologie-Filter akzeptierte Proteine und ihre Peptidrelationen verwendet.

Der vorgestellte Algorithmus löst weitgehend das Proteininferenz-Problem und entfernt Peptid-Protein-Beziehungen, für die es keinen eindeutigen Beleg gibt. Netzwerke von homologen Proteinen mit unlösbaren Peptid-Protein-Beziehungen werden dabei auf jeweils nur ein repräsentatives Protein reduziert.

3.10 Verteilung der Peptidintensitäten

Ein identifiziertes Peptid kann in mehreren Proteinen vorkommen. Seine Signalintensität spiegelt seine Gesamtmenge wider, zu der jedes seiner Ursprungsproteine anteilig beiträgt. Zu welchen Anteilen die Signalintensität von den einzelnen Ursprungsproteinen herrührt, ist am betroffenen Peptid selbst nicht erkennbar. Für die Schätzung der Anteile der Ausgangsproteine an den Signalintensitäten geteilter Peptide werden für jede Messung die Mengenverhältnisse zwischen den zugehörigen Proteinen anhand der Signalintensitäten eindeutig zugeordneter Peptide geschätzt. Die Gesamtintensitäten geteilter Peptide werden zu den geschätzten Anteilen zwischen ihren Ursprungsproteinen verteilt. Nach der Verteilung der Gesamtintensität eines Peptides, das mehreren Proteinen zugeordnet ist, werden bei der Quantifizierung dieser Proteine nur ihre entsprechenden Anteile der Peptidintensität berücksichtigt.

3.11 Absolute Proteinquantifizierung

Die absoluten Proteinmengen werden anhand einer abgewandelten Top3-Methode[136] ermittelt. Bei der TopX genannten Methode wird für die Quantifizierung eines Proteins statt der durchschnittlichen Intensität der drei Peptide mit der höchsten Abundanz die durchschnittliche Intensität einer benutzerdefinierten Anzahl von Peptiden mit der höchsten Abundanz verwendet. Die berechneten TopX-Intensitäten korrespondieren mit den absoluten molaren Mengen der Proteine in der Probe. Durch Multiplikation der TopX-Werte mit den jeweiligen molekularen Massen der Proteine werden entsprechend relative Gewichtsmengen der Proteine in der Probe berechnet. Die Summe dieser relativen Gewichtsmengen entspricht der Gesamtmenge der Proteine in der massenspektrometrischen Analyse. Durch eine Relation der relativen Gewichtsmenge einzelner Proteine zur Gesamtmenge können ihre relativen Gewichtsanteile ander Gesamtmenge z.B. in ppm oder fmol/µg berechnet werden. Durch die Berücksichtigung der TopX-Intensität eines Standardproteins zu seiner bekannten Menge in der Probe wird die Umrechnung der TopX-Intensitäten der erfassten Proteine in absolute Mengen wie fmol oder ng ermöglicht.

3.12 Filterung der Protein-FDR

Um die Qualität der Proteinidentifikation zu harmonisieren, wird die FDR als Anteil der falsch-positiven Identifikationen an der Gesamtzahl der betrachteten Identifikationen auf Proteinebene berechnet und die Einhaltung eines benutzerdefinierten FDR-Schwellenwertes sichergestellt. Die FDR-Verteilung wird als eine Funktion der Zuverlässigkeit der Proteinidentifikation ermittelt, indem für jeden beobachteten PLGS-Proteinidentifikationsscore die FDR ausschließlich anhand der Proteine berechnet wird, die mindestens mit diesem oder einem höheren Score identifiziert wurden. Der höchste Score, an dem der benutzerdefinierte FDR-Schwellenwert erreicht oder überschritten wird, dient als Filtergrenze. Das Ergebnis der Datenanalyse umfasst so ausschließlich Proteine, deren maximaler PLGS-Identifikationsscore an und oberhalb des Scores an der FDR-Filtergrenze liegt.

3.13 Implementierung des Analyseworkflows - ISOQuant

Der beschriebene Workflow (s. Kapitel 3.1) samt der vorgestellten Methoden (s. Kapitel 3.3 bis 3.12) für die quantitative Analyse labelfreier MS^E/$HDMS^E$/$UDMS^E$-Daten wurden als Teile der Software ISOQuant implementiert. Für die Implementierung der Softwarekomponenten wurde die plattformunabhängige Programmiersprache Java verwendet. ISOQuant kann unter allen modernen Betriebssystemen mit grafischer Benutzungsoberfläche wie Windows, Mac OS X oder Linux und einer installierten Java Virtual Machine (ab Version 1.6) genutzt werden. Die Software wird über eine grafische Benutzungsoberfläche bedient. Die Parametrisierung der implementierten Methoden erfolgt global über eine Konfigurationsdatei. Für die Speicherung der Daten wird für jedes Experiment ein neues Abbild der im Kapitel 3.2 beschriebenen relationalen Datenbank in einem Datenbankmanagementsystem erstellt. Für diesen Zweck wird das Datenbankmanagementsystem MySQL (ab Version 5.1, oder kompatibel) verwendet, das entweder lokal installiert oder im lokalen Netzwerk zur Laufzeit verfügbar sein muss. Zur Implementierung einzelner Funktionen in ISOQuant wurden frei-verfügbare externe Java-Bibliotheken verwendet, die in der Tabelle 3 zusammengefasst werden.

Tab. 3: Die in ISOQuant genutzten externen Java-Bibliotheken.

Bibliothek	Version	Typ der Lizenz	Verwendungszweck
JDOM	1.1.3	BSD	XML-Zugriff
Tagsoup	1.2.1	Apache v2.0	XML-Zugriff
MySQL Connector/J	5.1.13	GPLv2 mit FOSS Ausnahme	Datenbankverbindung
JSiX	1.0	BSD	Java-Erweiterungen
Apache POI	3.8	Apache v2.0	Excel-Dateien
DOM4J	1.6.1	BSD	POI-Abhängigkeit
StAX	1.0.1	Apache v2.0	POI-Abhängigkeit

ISOQuant wurde nicht-kommerziell unter einer 4-Klausel-BSD-Lizenz als quelloffene Software über die Internetpräsenz www.isoquant.net kostenfrei veröffentlicht. Quellcode sowie Installationspakete für Windows, Mac OS X und eine portable Version für andere Betriebssysteme sind verfügbar.

3.14 Vergleich mit PLGS

Um Analyseergebnisse des entwickelten Analyseworkflows auf die Reproduzierbarkeit der Proteinidentifikation und -quantifizierung gegenüber den Analyseergebnissen von PLGS zu untersuchen, wurde ein tryptischer Verdau von einem HeLa-Zelllysat unter unterschiedlichen Bedingungen mit MS^E-, $HDMS^E$- und $UDMS^E$-Methoden erfasst[98]. Jede Messung wurde dreimal wiederholt. Die Rohdaten wurden zunächst mit PLGS prozessiert und die PLGS-Ergebnisse anschließend mit ISOQuant analysiert. Alle Replikatmessungen einer Akquisitionsmethode wurden jeweils einem separaten LC-MS-Experiment zugeordnet. Sowohl die Ergebnisse der Peptid- und Proteinidentifikation als auch die Ergebnisse der Quantifizierung wurden vor und nach der ISOQuant-Analyse erfasst. Abbildung 20 und Tabelle 4 vergleichen die Anzahl der korrekt identifizierten Peptide und Proteine der HeLa-Zellen bei FDR $\leq 1\%$ in den MS^E-, $HDMS^E$- und $UDMS^E$-Datensätzen vor und nach der ISOQuant-Analyse.

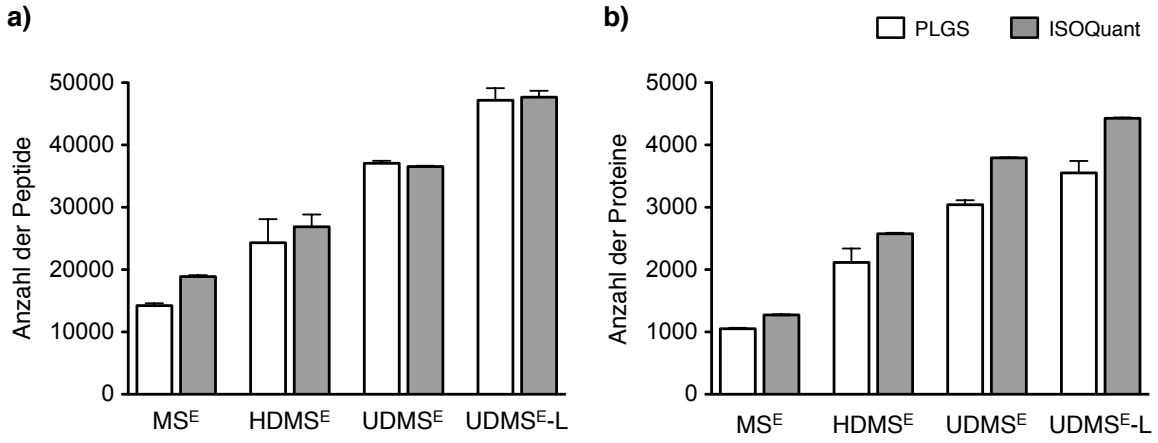

Abb. 20: Peptid- und Proteinidentifikationen nach PLGS- und ISOQuant-Analyse.
Anzahl der bei FDR $\leq 1\%$ in einer LC-MS-Messung identifizierten Peptide (**a**) und Proteine (**b**) vom tryptischen Verdau eines HeLa-Zelllysats, das mit MS^E, $HDMS^E$, $UDMS^E$ (200 ng, 90 min) und $UDMS^E$ (300 ng, 180 min) ($UDMS^E$-L) akquiriert wurde, vor der ISOQuant-Analyse (PLGS, weisse Balken) und nach der ISOQuant-Analyse (ISOQuant, graue Balken). Die Balken repräsentieren die Durchschnittswerte von jeweils drei technischen Replikaten. Die Whiskerenden repräsentieren die entsprechenden Maximalwerte.

Tab. 4: Peptid- und Proteinidentifikationen nach PLGS- und ISOQuant-Analyse.
Die durchschnittliche Anzahl der bei FDR $\leq$ 1% in einer LC-MS-Messung identifizierten Peptide und Proteine eines tryptischen Verdaus von HeLa-Zellen (s. Abbildung 20) werden für die Akquisition mit MS^E, $HDMS^E$, $UDMS^E$ (200 ng, 90 min) und $UDMS^E$ (300 ng, 180 min) ($UDMS^E$-L) vor der ISOQuant-Analyse (PLGS) und nach ISOQuant-Analyse (ISOQuant) in der Tabelle aufgeführt.

Datensatz	Peptide PLGS	Peptide ISOQuant	Proteine PLGS	Proteine ISOQuant
MS^E	14235	18893	1052	1274
$HDMS^E$	24309	26889	2114	2575
$UDMS^E$	37055	36551	3042	3795
$UDMS^E$-L	47173	47665	3552	4428

Gegenüber PLGS stieg die durchschnittliche Anzahl der Peptididentifikationen nach der ISOQuant-Analyse der MS^E-Daten um 32,7%, der $HDMS^E$-Daten um 10,6% und der $UDMS^E$-Daten mit 180 min Gradientenzeit um 1% an, fiel jedoch bei den $UDMS^E$-Daten mit 90 min Gradientenzeit um 1,4% ab. Die durchschnittliche Anzahl der Proteinidentifikationen stieg bei allen Datensätzen nach der ISOQuant-Analyse um 21,1% bis 24,8% im Vergleich zum PLGS-Ergebnis an.

Abbildung 21 zeigt Übereinstimmungen der Peptid- und Proteinidentifikationen zwischen den drei technischen Replikaten des mit $UDMS^E$ (300 ng, 180 min) akquirierten HeLa-Proteoms vor und nach der Datenanalyse mit ISOQuant. Durch die Datenanalyse mit ISOQuant fiel die Anzahl der mit PLGS im LC-MS-Experiment identifizierten Peptide von insgesamt 68663 auf 43463 um 36,7%. Die Anzahl der identifizierten Proteine wurde von 4270 auf 3833 um 10,2% reduziert. Dabei konnte eine signifikante Verbesserung der Reproduzierbarkeit der Peptid- und Proteinidentifikation nach der ISOQuant-Analyse beobachtet werden. So stieg der Anteil der Peptide, die in allen drei technischen Replikaten identifiziert wurden, von 23,9% auf 69,3%. Der Anteil der in allen drei technischen Replikaten übereinstimmenden Proteinidentifikationen stieg von 48,5% auf 97,8%. Die gegenüber PLGS geringeren Gesamtzahlen an Peptid- und Proteinidentifikationen werden bei der ISOQuant-Analyse hauptsächlich durch die benutzerdefinierte Filterung der PLGS-Identifikationen verursacht. Die Reproduzierbarkeit der Peptid- und Proteinidentifikation wird mit der Übertragung von Peptididentifikationen zwischen den Messungen durch die Cluster-Annotation gesteigert.

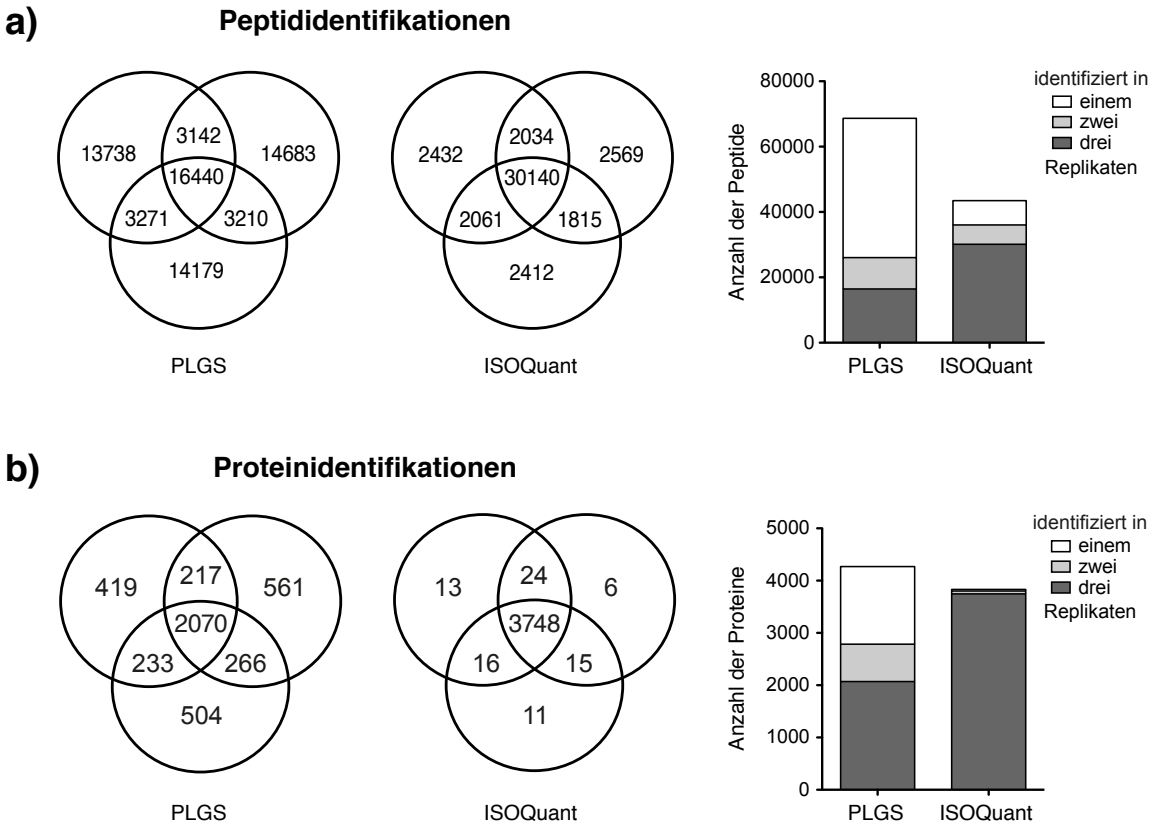

Abb. 21: Reproduzierbarkeit der Identifikation nach PLGS- und ISOQuant-Analyse.
Das HeLa-Proteom wurde in drei technischen Replikaten mit UDMS^{E} (300 ng, 180 min) massenspektrometrisch erfasst und mit PLGS und ISOQuant analysiert[98]. In der Abbildung wird die Übereinstimmung der Peptid- (**a**) und der Proteinidentifikationen (**b**) zwischen den drei technischen Replikaten dargestellt. Die Balkendiagramme zeigen jeweils die Anzahl der Peptide bzw. Proteine, die in nur einem, in zwei oder in allen drei technischen Replikaten identifiziert wurden.

Um die Reproduzierbarkeit der Proteinquantifizierung eines LC-MS-Experimentes zu bewerten, wird für jedes quantifizierte Protein die Varianz der berechneten Proteinmengen zwischen den technischen Replikaten betrachtet. Abbildung 22 und Tabelle 5 zeigen die Verteilungen von Variationskoeffizienten (CV)[1] der Proteinmengen für das mit MS^E, $HDMS^E$ und $UDMS^E$-Methoden akquirierte HeLa-Proteom vor und nach der ISOQuant-Analyse.

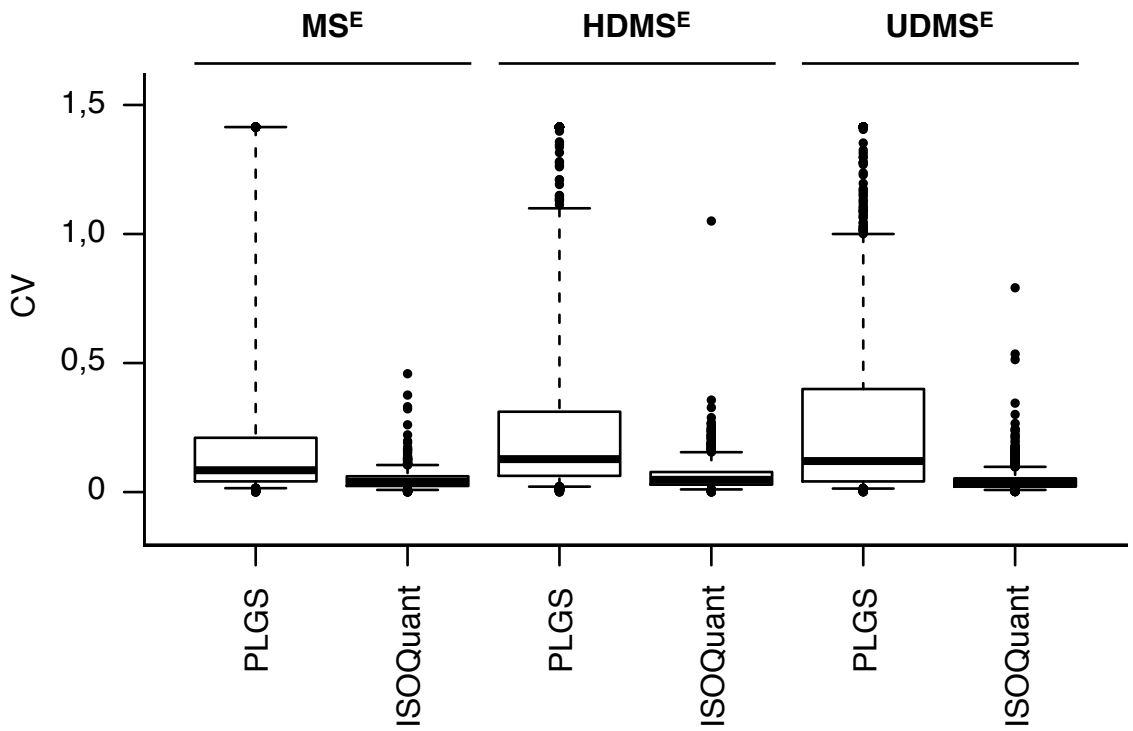

Abb. 22: Varianz der Proteinquantifizierung nach PLGS- und ISOQuant-Analyse.
Das HeLa-Proteom wurde mit MS^E, $HDMS^E$ und $UDMS^E$ in jeweils drei technischen Replikaten (200 ng, 90 min) analysiert. Der Boxplot visualisiert die Varianz der Proteinquantifizierung anhand der Verteilungen der Variationskoeffizienten von ermittelten Proteinmengen zwischen den technischen Replikaten der jeweiligen Akquisitionsmethode nach der Datenanalyse mit PLGS und ISOQuant. Die Boxen repräsentieren das 25. und das 75. Percentil, die Linien in den Boxen den Median und die Whiskerenden das 5. und 95. Percentil der jeweiligen Verteilungen.

[1] Der Variationskoeffizient wird oft auch als die relative Standardabweichung (RSD, engl. relative standard deviation) bezeichnet.

Tab. 5: Varianz der Proteinquantifizierung nach PLGS- und ISOQuant-Analyse.
Die Tabelle zeigt die Varianz der Proteinquantifizierung anhand der CV-Verteilungen von ermittelten Proteinmengen zwischen den technischen Replikaten der jeweiligen Akquisitionsmethode nach der Datenanalyse mit PLGS und ISOQuant. Die CV-Verteilungen werden hier anhand der statistischen Lagemaße, der 0,05-, 0,25-, 0,5- (Median), 0,75- und 0,95-Quantile beschrieben.

Datensatz	Software	$Q_{0,05}$	$Q_{0,25}$	Median	$Q_{0,75}$	$Q_{0,95}$
MS^E	PLGS	0,0152	0,0415	0,0847	0,2105	1,4142
MS^E	ISOQuant	0,0085	0,0237	0,0400	0,0612	0,1049
$HDMS^E$	PLGS	0,0211	0,0630	0,1278	0,3115	1,0996
$HDMS^E$	ISOQuant	0,0103	0,0288	0,0475	0,0778	0,1546
$UDMS^E$	PLGS	0,0137	0,0413	0,1203	0,3993	1,0000
$UDMS^E$	ISOQuant	0,0084	0,0210	0,0350	0,0532	0,0979

Für alle Akquisitionsmethoden reduzierte die ISOQuant-Analyse die Varianz von Proteinmengen zwischen den technischen Replikaten in einem LC-MS-Experiment gegenüber der Datenanalyse mit PLGS deutlich. Nach der PLGS-Analyse verteilen sich 95% der Variationskoeffizienten unter 1,41 in MS^E-, unter 1,1 in $HDMS^E$- und unter 1,0 in $UDMS^E$-Daten. Nach der ISOQuant-Analyse verteilen sich die 95% der Variationskoeffizienten unter 0,1 in MS^E-, 0,15 in $HDMS^E$- und 0,1 in $UDMS^E$-Daten. Im Vergleich zu den entsprechenden Variationskoeffizienten nach der PLGS-Analyse wurden die mittleren Variationskoeffizienten nach der Datenanalyse mit ISOQuant um 52,8% in MS^E-, um 62,8% in $HDMS^E$- und um 70,9% in $UDMS^E$-Daten reduziert. Eine Reduktion der Varianz von Proteinmengen zwischen den technischen Replikaten wird bei der Datenanalyse mit ISOQuant hauptsächlich durch die Analyseschritte der restriktiven Annotation der Feature-Cluster und die multidimensionale Normalisierung der Feature-Intensitäten erreicht.

3.15 Vergleich mit Progenesis QI for Proteomics und synapter

Um den Einfluss der Datenanalyse auf das Ergebnis der labelfreien Quantifizierung erfassen zu können, wurde ein Satz aus zwei Evaluierungsproben entwickelt sowie mit MS^E- und $UDMS^E$-Methoden massenspektrometrisch analysiert. Anhand der resultierenden Evaluierungsdatensätze wurde die analytische Performance des implementierten Analyseworkflows (ISOQuant) bewertet sowie mit der analytischen Performance der Softwarepakete synapter und Progenesis QIP verglichen.

3.15.1 HeLa-Hefe-*E.coli*-Metaproteom

Bei einem typischen, komparativen Experiment zur labelfreien Proteinquantifizierung werden Proteinmengen in unterschiedlichen Proben ermittelt und verglichen. Dabei stützen sich viele Analysemethoden auf die Annahme, dass der Großteil der Proteine in einem Experiment keiner Regulation unterliegt. So werden gleiche Mengen dieser Hintergrundproteine in allen Proben des Experiments vorausgesetzt. Lediglich für eine Untermenge der Proteine wird ein Unterschied in der Expression und damit unterschiedliche Konzentrationen in den einzelnen Proben angenommen. Werden mehrere Proben bekannter Komposition analysiert, so kann der Einfluss der Datenanalyse auf das Ergebnis der labelfreien Quantifizierung nachvollzogen werden. Solche Evaluierungsdaten können z.B. durch die Zugabe einzelner Proteine in bekannten, in einzelnen Proben unterschiedlichen Mengen zu einem bekannten Hintergrund-Proteom erstellt werden[167]. Die statistische Signifikanz einer solchen Evaluierung wird jedoch durch die geringe Zahl der verwendeten Standardproteine reduziert. Die Gesamtmenge der Proteine in den Evaluierungsproben sollte konstant gehalten werden, um zusätzlichen systematischen Fehlern bei der Datenanalyse vorzubeugen. Unter Berücksichtigung dieser Aspekte wurden zwei Metaproteomproben bekannter Komposition generiert, die die tryptisch verdauten Proteome von drei unterschiedlichen Organismen (Mensch, Weinhefe und *Escherichia coli*) in exakt bekannten Anteilen enthalten. Um eine große Zahl von unregulierten Hintergrundproteinen zu simulieren, bestehen die beiden Proben (Probe A und Probe B) jeweils zu 65% der Gesamtproteinmenge aus gleichen Mengen tryptisch verdauter Proteine menschlicher Epithelzellen eines Zervixkarzinoms (HeLa-Zellen). Für die Simulation der in der Probe A gegenüber der Probe B vierfach herunterregulierten Proteine, wurden tryptisch verdaute Proteine des Bakteriums *E.coli* entsprechend 5% der Gesamtproteinmenge zur Probe A und entsprechend 20% zur Probe B hinzugefügt. Die

Gesamtproteinmengen der beiden Proben wurden durch die Zugabe der tryptisch verdauten Hefeproteine ausgeglichen. Neben dem Ausgleich der Gesamtproteinmengen simuliert das Hefeproteom durch seine Mengenanteile von 30% in der Probe A und 15% in der Probe B zweifach hochregulierte Proteine in der Probe A gegenüber der Probe B. Beide Proben wurden jeweils in den Akquisitionsmodi MS^E und $UDMS^E$ massenspektrometrisch analysiert. Jede Probe wurde dabei in fünf technischen Replikaten gemessen. Somit liegen innerhalb des LC-MS-Experiments insgesamt fünf Rohdatensätze jeder Probe zu jeder Akquisitionsart vor. In Abbildung 23 werden die Zusammensetzung der Metaproteomproben und der Ablauf der LC-MS-Analyse schematisch dargestellt.

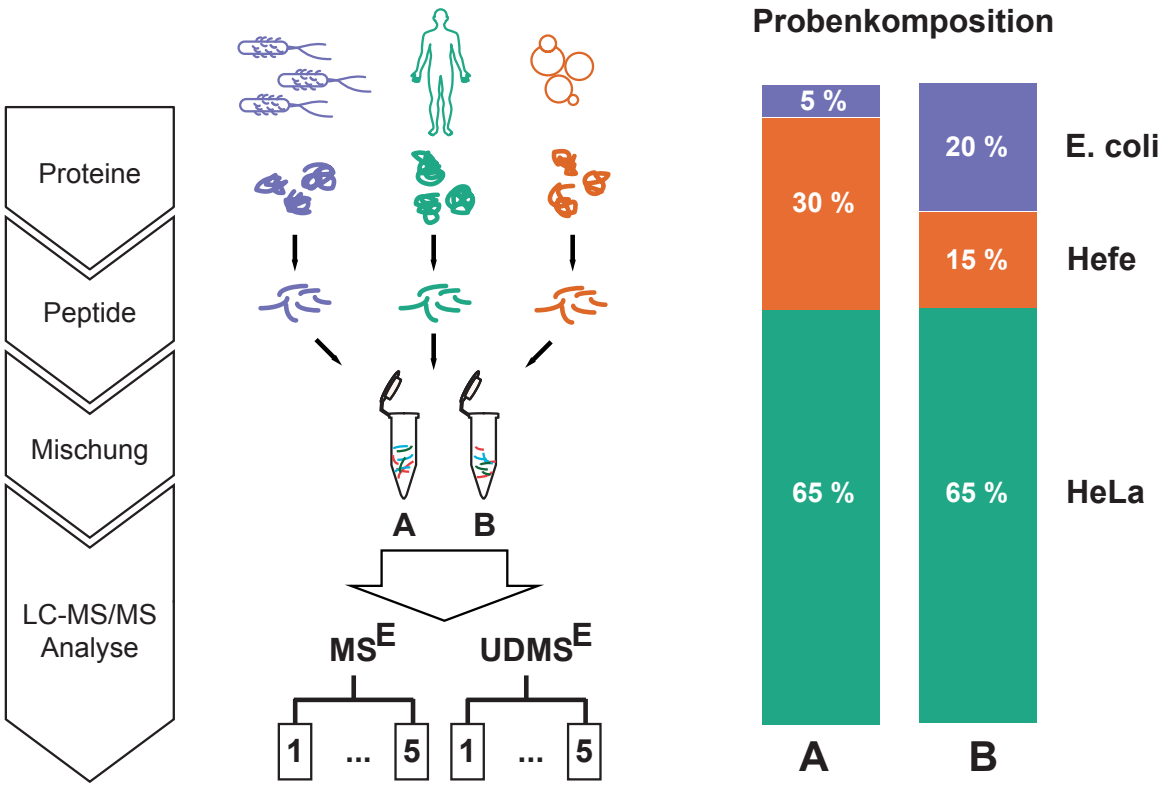

Abb. 23: Komposition und LC-MS-Akquisition der Metaproteomproben.
Die Metaproteomproben A und B bestehen zu 65% aus humanem (HeLa) Proteom. Die Anteile des bakteriellen (*E.coli*) Proteomes betragen 5% in der Probe A und 20% in der Probe B. Entsprechend ist der Anteil des Hefeproteoms 30% in der Probe A und 15% in der Probe B. Beide Metaproteome wurden mit MS^E- und $UDMS^E$-Akquisition in fünf technischen Replikaten massenspektrometrisch analysiert.

3.15.2 Analyseworkflows

Aus den generierten Rohdaten der Metaproteomproben wurden drei Testdatensätze zusammengestellt: MS^E, $UDMS^E$ sowie ein kombinierter Datensatz aus MS^E und $UDMS^E$-Daten. Die MS^E- und $UDMS^E$-Datensätze enthalten die Daten aus den jeweiligen 10 Messungen. Entsprechend umfasst der kombinierte Datensatz die Daten der insgesamt 20 Messungen. Die drei Datensätze wurden der Analyse mit Progenesis, synapter und ISOQuant unterzogen. In Abbildung 24 werden die Analyseworkflows von Progenesis QIP, synapter und ISOQuant aus Sicht des Anwenders dargestellt.

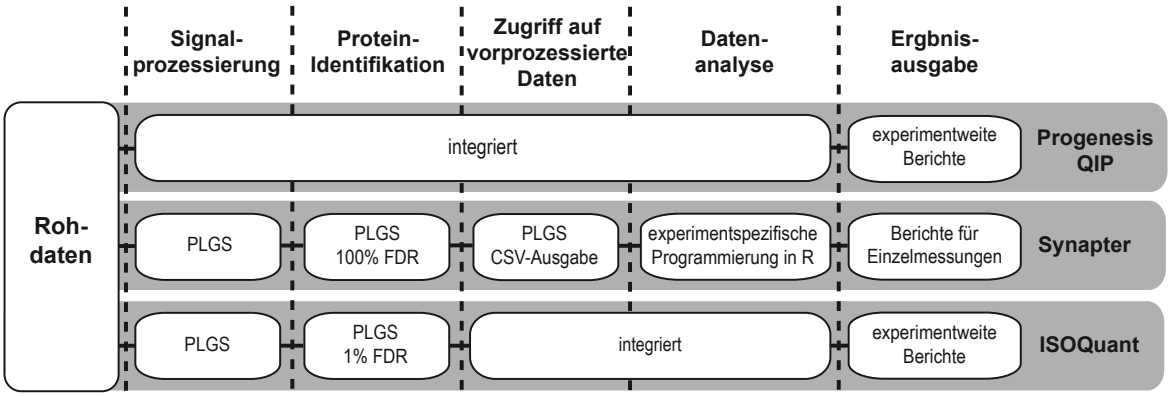

Abb. 24: Analyseworkflows der getesteten Softwarelösungen.
Die Metaproteom-Rohdaten wurden mit drei Analyseworkflows zur labelfreien Proteinquantifizierung analysiert. Während Progenesis QIP alle Analyseschritte beinhaltet, müssen die Rohdaten für synapter und ISOQuant mit PLGS vorprozessiert werden. Die Peptid- und Proteinidentifikation in PLGS erfolgte für synapter mit 100% FDR und für ISOQuant mit 1% FDR. Der direkte Zugriff auf die PLGS-Datenstrukturen ist in ISOQuant integriert. Für synapter wurden die PLGS-Ergebnisse im CSV-Format exportiert und die Analyse experimentspezifisch in R programmiert. Die Ergebnisse der labelfreien Quantifizierung in synapter werden für jede Messung separat berichtet. Progenesis QIP und ISOQuant exportieren die Analyseergebnisse jeweils für ein gesamtes Experiment.

Das kommerzielle Progenesis QIP integriert alle nötigen Analyseschritte in einer Software. Für die freien Softwarepakete synapter und ISOQuant wird eine Vorprozessierung der Rohdaten mit PLGS vorausgesetzt, die sowohl die Signalprozessierung als auch die Peptid- und Proteinidentifikation beinhaltet. Die Analyseworkflows von Progenesis QIP, synapter und ISOQuant verwenden gleiche Algorithmen für die Signalprozessierung (Apex3D, Waters), weitere Schritte der Datenanalyse unterscheiden sich jedoch. Während ISOQuant keine besonderen Anforderung an die PLGS-Analyse stellt, benötigt synapter eine Peptid- und Proteinidentifikation mit 100% FDR sowie aktivierter Zusatzausgabe der PLGS-Ergebnisse in CSV-Dateien. Nach der PLGS-Analyse werden die PLGS-Datenstrukturen von ISOQuant automatisch durchsucht. Das Design des jeweiligen Experimentes wird durch Benutzerinteraktion festgelegt. Anschließend werden die Experimentdaten in die relationale

Datenbank importiert und der Analyseworkflow automatisiert abgearbeitet. Die synapter-Analyse wird experimentspezifisch mit einem R-Programm definiert und parametrisiert. Dabei werden die Dateisystempfade zu den aus PLGS erzeugten CSV-Dateien angegeben. Analyseergebnisse der labelfreien Quantifizierung mit synapter werden für jede Messung separat generiert. Progenesis QIP und ISOQuant berichten ihre Analyseergebnisse auf Basis des gesamten Experiments.

Zur besseren Vergleichbarkeit wurden die Analyseergebnisse der Softwarepakete in ein einheitliches Format überführt. Für den kombinierten Datensatz werden dabei nur die Quantifizierungswerte der MS^E-Daten bei der Evaluierung berücksichtigt.

3.15.3 Filterung der Analyseergebnisse

Die verwendeten Analyseworkflows entscheiden nach unterschiedlichen, internen Kriterien, welche Proteine quantifiziert werden. Um eine einheitliche, vergleichbare Qualität der Analyseergebnisse durch Progenesis QIP, synapter und ISOQuant sicherzustellen, wurden die Ergebnisse der labelfreien Quantifizierung zusätzlich gefiltert. Dabei wurde die Quantifizierung eines Proteins nur dann akzeptiert, wenn dieses durch mindestens zwei Peptide identifiziert wurde (Peptidfilter) und wenn das Protein mindestens in zwei technischen Replikaten in einer der beiden Proben quantifiziert wurde (Replikationsfilter).

3.15.4 Proteinidentifikation

Um den Einfluss des Peptid- und Replikationsfilters auf die Proteinidentifikation zu untersuchen, wurden diese Filter sequentiell angewandt. Die Anzahl der identifizierten Proteine wurde jeweils für alle Datensätze und Softwarepakete vor und nach Anwendung des Peptidfilters sowie nach der anschließenden Anwendung des Replikationsfilters erfasst. Dabei konnten auf Basis gleicher Eingabedaten deutliche Unterschiede in der Anzahl der identifizierten Proteine zwischen Progenesis QIP, synapter und ISOQuant bei MS^E, $UDMS^E$ und kombinierten Datensätzen beobachtet werden. Wie in Abbildung 25 gezeigt, hat die Datenfilterung mittels Peptid- und Replikationsfilter einen erheblichen Einfluss auf die Anzahl der Proteinidentifikationen.

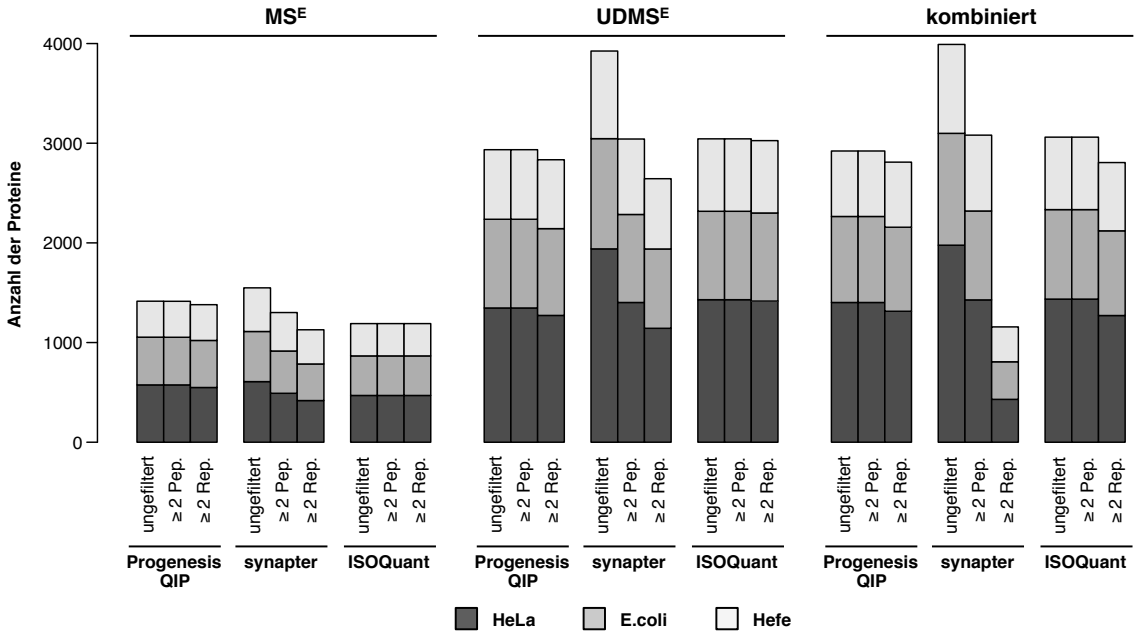

Abb. 25: Einfluss des Peptid- und Replikationsfilters auf die Proteinidentifikation.
Vor Filterung identifiziert synapter die meisten Proteine in jedem Testdatensatz. Progenesis QIP und ISOQuant identifizieren eine vergleichbare Anzahl der Proteine in allen Testdatensätzen, dabei findet Progenesis QIP in MS^E-Daten 17% Proteine mehr als ISOQuant. Die zusätzliche Datenfilterung reduziert die Anzahl identifizierter Proteine in allen Ergebnissen.

Die synapter-Strategie zur Maximierung der Proteinidentifikationen liefert die höchste Zahl an identifizierten Proteinen für alle drei Testdatensätze vor der zusätzlichen Filterung der Daten. Progenesis QIP und ISOQuant berichten eine vergleichbare Anzahl der identifizierten Proteine für alle Testdatensätze, dabei findet Progenesis QIP in MS^E-Daten 17% mehr Proteine als ISOQuant. Die Peptid- und Replikationsfilter reduzieren die Zahlen identifizierter Proteine für jede Software und jeden Testdatensatz. Durch die Filter wurde die Anzahl der Proteine in Ergebnissen von Progenesis QIP um bis zu 3,8% und in Ergebnissen von ISOQuant um bis zu

8,8% reduziert. Im Fall von synapter konnte eine Reduktion der Proteinidentifikationen um 27% in MS^E- und um 33% in $UDMS^E$-Daten beobachtet werden. Im kombinierten Datensatz identifizierte synapter 3994 Proteine, diese Zahl sank nach Anwendung des Peptidfilters auf 3084 Proteine und nach Anwendung des Replikationsfilters auf 1159 Proteine. Dies bedeutet einen Gesamtverlust von 71% der Identifikation durch die zusätzliche Filterung der Daten in synapter-Ergebnissen für den kombinierten Datensatz.

Wie in Abbildung 26 dargestellt, wurde eine hohe Übereinstimmung der identifizierten Proteine in allen Datensätzen zwischen den Analyseergebnissen der drei Softwarepakete nach der Anwendung der Peptid-und Replikationsfilter beobachtet.

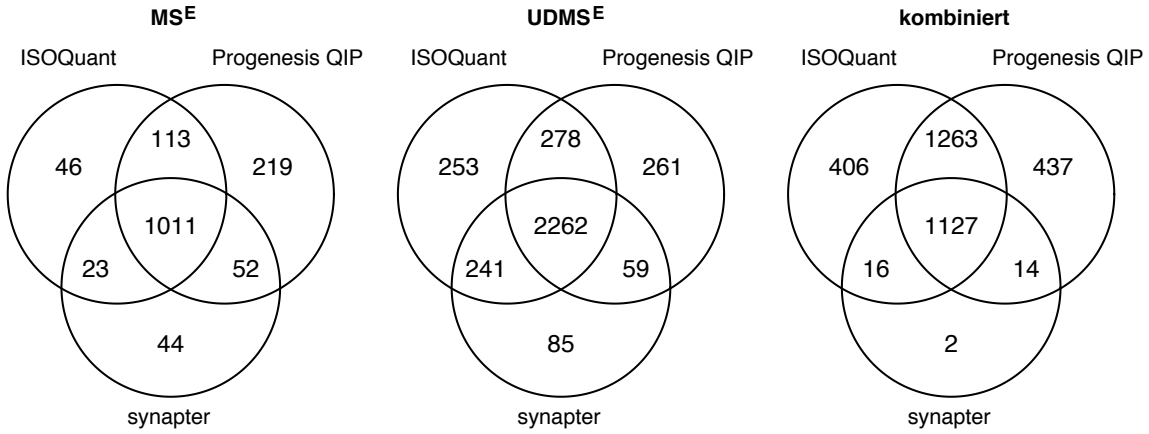

Abb. 26: Übereinstimmung der Proteinidentifikation zwischen den Softwarelösungen.
Nach der Peptid-und Replikationsfilterung konnten 1011 Proteinidentifikationen im MS^E- und 2262 Proteinidentifikationen im $UDMS^E$-Datensatz übereinstimmend in allen drei Softwarepaketen identifiziert werden. Für den kombinierten Datensatz betrug die Zahl der übereinstimmenden Proteinidentifikationen 1127.

Im MS^E-Datensatz wurden 1011 (67%) Proteine von Progenesis QIP, synapter und ISOQuant übereinstimmend identifiziert. Im $UDMS^E$-Datensatz betrug die Übereinstimmung 2262 (66%) Proteine. Limitiert durch die synapter-Identifikationen stimmten im kombinierten Datensatz 1127 (35%) Proteine in den Ergebnissen der drei Softwarepakete überein. 2390 (73%) Proteine wurden im kombinierten Datensatz durch Progenesis QIP und ISOQuant übereinstimmend identifiziert.

Die Übereinstimmungen zwischen den Analyseergebnissen der einzelnen Testdatensätze werden in Abbildung 27 als Venn-Diagramme für die einzelnen Softwarepakete dargestellt. Die Analyseergebnisse innerhalb eines Softwarepaketes weisen jeweils hohe Übereinstimmungen der Proteinidentifikationen über alle drei Testdatensätze auf: 1311 Proteine in den Ergebnissen von Progenesis QIP, 1108 in den Ergebnissen von synapter und 1157 in den Ergebnissen von ISOQuant. $UDMS^E$ liefert eine höhere Proteomabdeckung

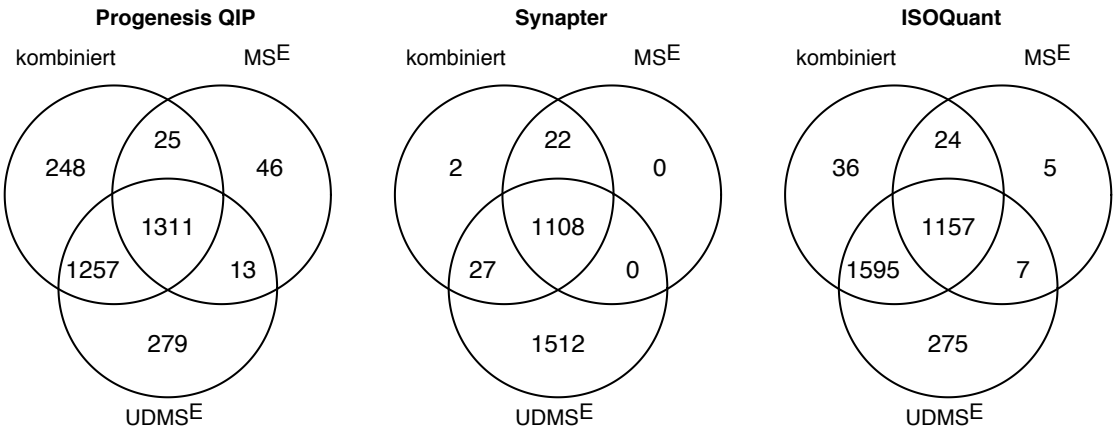

Abb. 27: Übereinstimmung der Proteinidentifikation zwischen den Testdatensätzen.
1311 Identifikationen stimmten in Progenesis, 1108 in synapter und 1157 in ISOQuant-Analyseergebnissen zwischen MS^E, $UDMS^E$ und den kombinierten Daten überein. Zwischen dem $UDMS^E$- und dem kombinierten Datensatz beträgt die Übereinstimmung der Identifikationen jeweils 2568 Proteine für Progenesis, 1135 Proteine für synapter und 2752 Proteine für ISOQuant.

als MS^E. Lediglich bis zu 1,5% der MS^E-Proteine wurden nicht in $UDMS^E$- oder den kombinierten Datensätzen erkannt. Die Zahl liegt im Bereich der bei der Peptid- und Proteinidentifikation verwendeten FDR-Schranke (1%), so dass es sich mit hoher Wahrscheinlichkeit um falsch positive Proteinidentifikationen handelt. Zwischen den $UDMS^E$ und den kombinierten Datensätzen stimmen die Identifikationen weitgehend überein: 2568 Proteine in den Ergebnissen von Progenesis QIP, 1135 Proteine in den Ergebnissen von synapter und 2752 Proteine in den Ergebnissen von ISOQuant.

3.15.5 Präzision der relativen labelfreien Quantifizierung

Hohe technische Reproduzierbarkeit ist eine Voraussetzung für die Präzision der labelfreien Proteinquantifizierung. In einem typischen Experiment übt die Reproduzierbarkeit der Probenaufbereitung einen signifikanten Einfluss auf die Präzision der labelfreien Proteinquantifizierung aus. Die Analysewokflows der drei Softwarepakete gleichen sich zwar in der initialen Signaldetektion, unterscheiden sich jedoch in den darauffolgenden Schritten, die sich auf die Präzision der labelfreien Proteinquantifizierung auswirken. Da die Metaproteomproben sich aus definierten Anteilen vorverdauter Proteome zusammensetzen, erlauben die Testdatensätze den Einfluss der Datenanalyse auf die Präzision der labelfreien Quantifizierung durch Erfassung der Reproduzierbarkeit der quantitativen Ergebnisse zu analysieren. Hierfür wird für jedes Protein die Varianz der berechneten Proteinmengen zwischen den technischen Replikaten innerhalb einer Probe als CV geschätzt. In Abbildung 28 und Tabelle 6 wird anhand der CV-Verteilungen aller quantifizierten Proteine innerhalb der einzelnen Proben der Testdatensätze der Einfluss der getesteten Analyseworkflows auf die Präzision der labelfreien Quantifizierung visualisiert.

Sowohl die Analysesoftware als auch die Art der Datenakquisition beeinflussen die resultierende Verteilung der Variationskoeffizienten. Für alle Datensätze konnte eine niedrige, vergleichbare Varianz der Proteinquantifizierung in den Ergebnissen durch die Analyse mit Progenesis QIP und ISOQuant erzielt werden. Die niedrigsten Mediane der CV-Verteilungen resultieren aus den Analysen der MS^E-Daten, gefolgt von den Medianen der CV-Verteilungen des $UDMS^E$-Datensatzes. Für den kombinierten Datensatz wurden für jede Software jeweils die höchste Varianz der Proteinquantifizierung beobachtet. Insgesamt weisen die Ergebnisse von synapter in allen Testdatensätzen die jeweils höchste Varianz der Proteinquantifizierung auf.

Tab. 6: Varianz der Proteinquantifizierung in verschiedenen Softwarelösungen.
Die Tabelle listet die Mediane der CV-Verteilungen der durch verschiedene Softwarelösungen berechneten Proteinmengen innerhalb der technischen Replikate einer Probe in den Ergebnissen der drei Testdatensätze.

Datensatz	Progenesis QIP	synapter	ISOQuant
MS^E	0,0489	0,2628	0,0451
$UDMS^E$	0,0610	0,1983	0,0975
kombiniert	0,0702	0,2628	0,1204

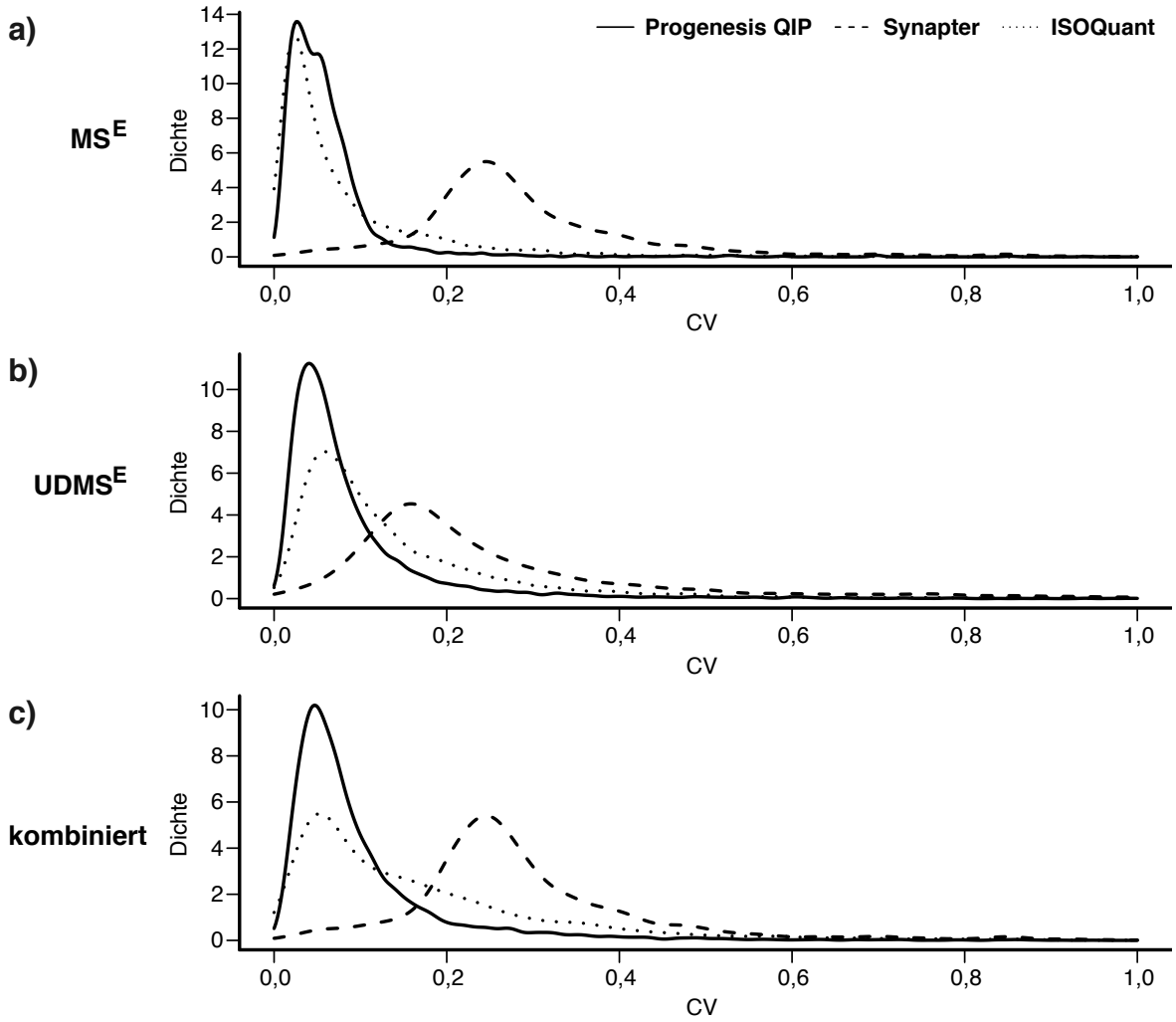

Abb. 28: Präzision der labelfreien Proteinquantifizierung.
Die Abbildung stellt die technische Reproduzierbarkeit der labelfreien Proteinquantifizierung durch Progenesis QIP (durchgezogene Linien), synapter (gestrichelte Linien) und ISOQuant (gepunktete Linien) jeweils für den MS^E (**a**), den $UDMS^E$ (**b**) und den kombinierten Datensatz (**c**) dar. Die Varianz der Proteinquantifizierung wird hier mit einer Kerndichteschätzung (engl. kernel density estimation) mit Gaußkern der CV-Verteilungen der berechneten Proteinmengen innerhalb der technischen Replikate einer Probe abgebildet. Die mittleren Werte der CV-Verteilungen werden in der Tabelle 6 aufgeführt.

3.15.6 Richtigkeit der relativen labelfreien Quantifizierung

Die Kombination der tryptischen Proteine dreier unterschiedlicher Organismen in exakt definierten Mengenverhältnissen ermöglicht eine Simulation von identischen, relativen Unterschieden der Expression für Hunderte von Proteinen gleichzeitig. Obwohl diese Parallelität der Proteinexpression nicht natürlich erscheint, dient die hohe Komplexität der Testdatensätze als eine Art Stresstest für die Fähigkeiten der getesteten Softwarepakete zur akkuraten Quantifizierung ganzer Proteome. Durch die synthetische Natur der Testproben werden für die Proteine einzelner Spezies definierte, relative Mengenverhältnisse zwischen den beiden Proben erwartet. Durch die Abweichung der nach Datenanalyse berechneten Mengenverhältnisse zu den erwarteten Werten, kann die Richtigkeit der labelfreien Quantifizierung beurteilt werden. Für jedes Protein wird sein relativer Unterschied als Logratio geschätzt, welches als das binär-logarithmierte Verhältnis der durchschnittlichen Menge in der Probe A zur durchschnittlichen Menge in der Probe B (Log_2(A:B)) berechnet wird. Da HeLa-Proteine zu gleichen Anteilen in beiden Proben vertreten sind, sollten sich ihre Logratios um den Wert null verteilen. Der Erwartungswert für die Logratios der Hefeproteine zwischen den Proben A und B beträgt 1,0 und für die Logratios der *E.coli*-Proteine -2,0. Abbildung 29 zeigt die berechneten Logratios aller quantifizierten Proteine entlang der logarithmierten Intensität für die Ergebnisse der labelfreien Quantifizierung der MS^E-, $UDMS^E$- und der kombinierten Daten durch Progenesis QIP, synapter und ISOQuant. Dabei werden nur Logratios der Proteine dargestellt, die sowohl in der Probe A als auch in der Probe B quantifiziert werden konnten. Die Logratios wurden um den Median der Logratios von HeLa-Proteinen zentriert.

Während sich die Logratios in den MS^E-Ergebnissen von synapter und ISOQuant akkurat um die entsprechenden Erwartungswerte verteilen, unterschätzte Progenesis QIP die relativen Mengenverhältnisse für Hefe- und *E.coli*-Proteine systematisch. Auch im $UDMS^E$- und dem kombinierten Datensatz wurden von Progenesis QIP die relativen Mengenverhältnisse für Hefe- und *E.coli*-Proteine systematisch unterschätzt.

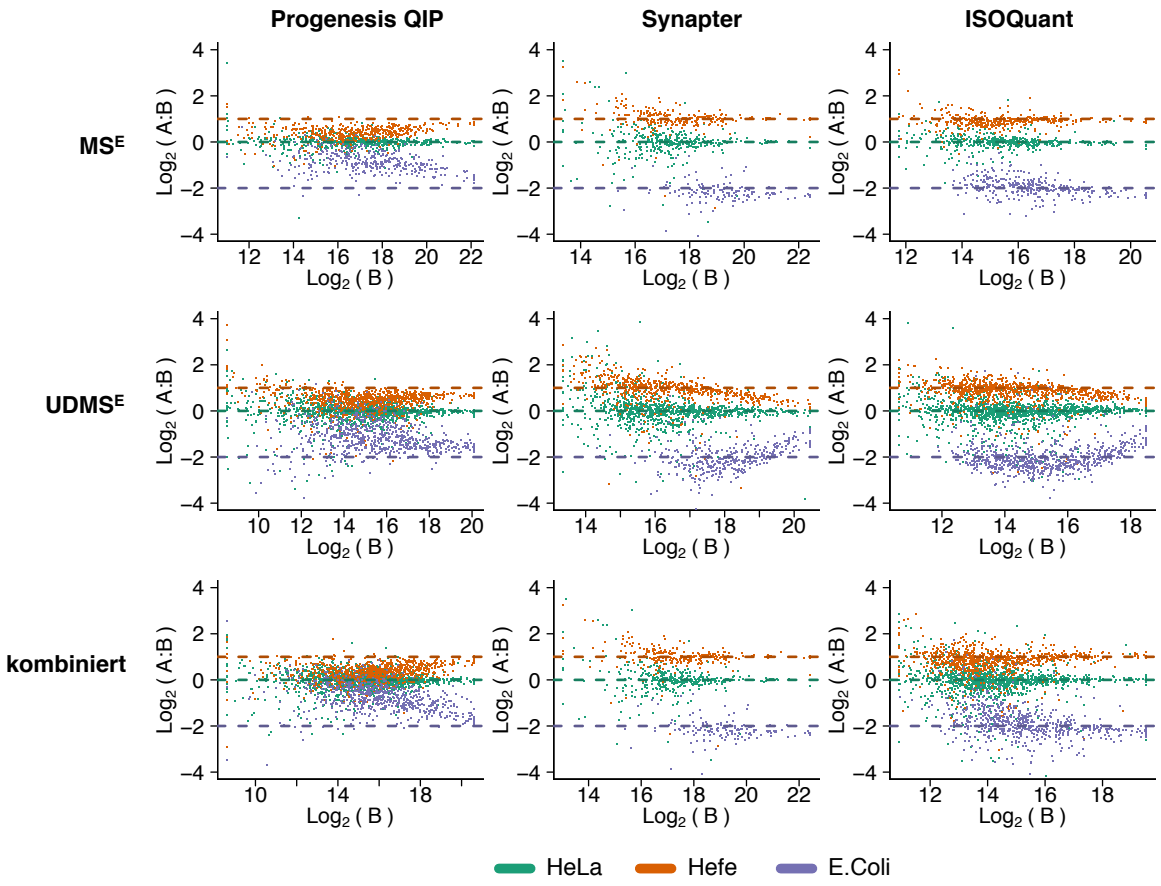

Abb. 29: Richtigkeit der labelfreien Proteinquantifizierung.
Die Ergebnisse der labelfreien Quantifizierung der drei Testdatensätze durch Progenesis, synapter und ISOQuant werden als Punktdiagramme der Logratios dargestellt. Die Logratios wurden als logarithmierte Verhältnisse der Probe A zu Probe B (Log_2(A:B)) für jedes Protein berechnet, das in beiden Proben quantifiziert wurde. Die Menge eines Proteins in einer Probe wurde anhand der durchschnittlichen Mengen innerhalb der technischen Replikate dieser Probe geschätzt. Die Strichlinien stellen die erwarteten Logratios dar: 0 für HeLa, +1 für Hefe und -2 für *E.coli*-Proteine.

Die Ergebnisse aller drei Softwarepakete visualisieren das in anderen Studien berichtete Sättigungsproblem in den IMS-Daten[144, 168]. Dies kann vor Allem in den synapter-Ergebnissen für den UDMS^E-Datensatz beobachtet werden, während in den Ergebnissen von ISOQuant die intensitätsabhängigen Tendenzen durch die multidimensionale Normalisierung weitgehend unterdrückt wurden. Für den kombinierten Datensatz verteilen sich in den synapter- und ISOQuant-Ergebnissen die Logratios für *E.coli*- und Hefeproteine erneut akkurat um die Erwartungswerte. Hier werden jedoch die Effekte des Replikationsfilters auf die synapter-Ergebnisse sichtbar, so dass für synapter deutlich weniger Logratios berechnet und abgebildet werden konnten. Alle drei Softwarelösungen waren in der Lage die Proteine einzelner Spezies in MS^E- und UDMS^E-Daten weitgehend voneinander zu trennen.

Obwohl Progenesis QIP in allen drei Datensätzen die Regulationsverhältnisse der Proteine stark unterschätzte, wiesen die Ergebnisse der labelfreien Proteinquantifizierung in Progenesis QIP die wenigsten Lücken auf, d.h. die Menge der meisten Proteine konnte für beide Proben berechnen werden. In Abbildung 30 werden für die drei Softwarepakete und die drei Testdatensätze jeweils die Anzahlen der Proteine dargestellt, für die es nicht möglich war ein Logratio zu berechnen, da sie lediglich in einer der beiden Proben quantifiziert wurden. Weniger als 0,1% der Proteine wurden durch Progenesis QIP in nur einer der beiden Proben quantifiziert. ISOQuant quantifizierte 1,5% im MS^E-, 2% im $UDMS^E$- sowie 5,4% Proteine im kombinierten Datensatz entweder nur in der Probe A oder nur in der Probe B. Synapter quantifizierte 31,6% solcher Proteine im MS^E-, 25,8% im $UDMS^E$- und 32,3% im kombinierten Datensatz.

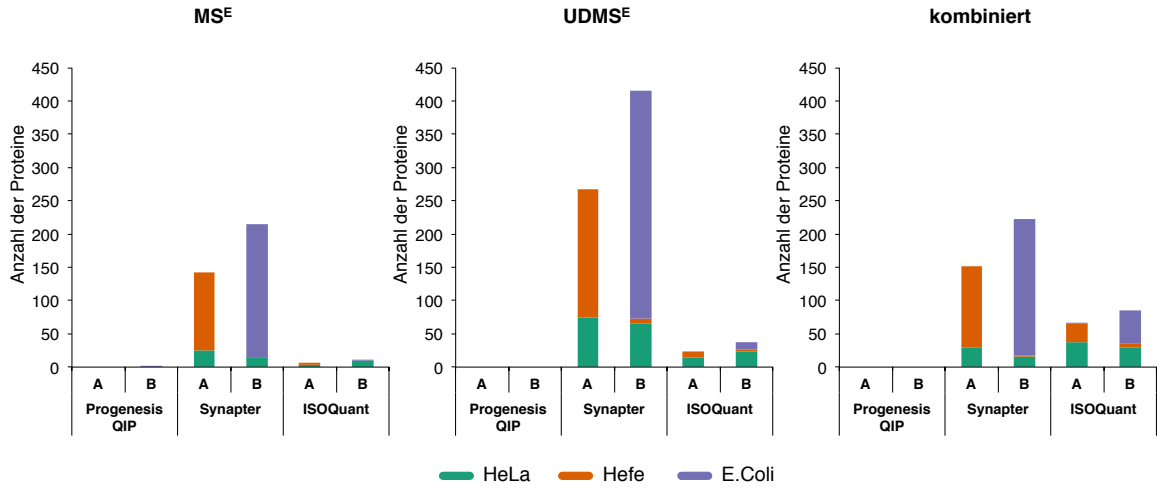

Abb. 30: Exklusive Proteinquantifizierung in verschiedenen Softwarelösungen.
Für Progenesis QIP, synapter und ISOQuant werden für MS^E-, $UDMS^E$- und den kombinierten Datensatz die Anzahlen der Proteine dargestellt, die exklusiv in nur einer der beiden Proben quantifiziert wurden. Für diese Proteine kann kein relatives Mengenverhältnis zwischen den Proben berechnet werden.

4 Diskussion

Neben den Markierungstechniken zur Proteinquantifizierung hat sich die labelfreie Proteinquantifizierung für viele Forscher zur Methode der Wahl entwickelt. Insbesondere die labelfreie Proteinquantifizierung auf Basis datenunabhängiger Akquisitionsmethoden erfreut sich zunehmender Beliebtheit. Die datenunabhängige Akquisitionsmethode MS^E sowie ihre IMS-Varianten $HDMS^E$ und $UDMS^E$ heben sich durch eine hohe Samplingrate gegenüber anderen Akquisitionsmethoden hervor, indem sie mit sehr kurzen Zykluszeiten nach jedem Vorläuferionenmassenspektrum lediglich ein Fragmentionenmassenspektrum erfassen. Auf diese Weise werden alle Peptidionen und ihre Fragmente ohne Verluste quantitativer Informationen erfasst. Dadurch eignen sich diese Akquisitionsmethoden besonders für eine Kombination mit hochauflösenden Trenntechniken wie der UPLC. Die resultierenden LC-MS-Daten weisen eine hohe Komplexität auf, so dass für ihre Verarbeitung spezielle Lösungsansätze benötigt werden, die den besonderen Herausforderungen im Zusammenhang mit der MS^E/$HDMS^E$/$UDMS^E$-Datenanalyse gewachsen sind.

In dieser Arbeit wurde eine Strategie und eine Reihe von dedizierten Algorithmen zur messungsübergreifenden, weiterführenden Analyse labelfreier MS^E/$HDMS^E$/$UDMS^E$-Daten entwickelt, die mit der Software PLGS vorprozessiert wurden. Die entwickelte Analysestrategie und die benötigten Algorithmen wurden als Bestandteile der Software ISOQuant implementiert. Nach Beginn dieser Arbeit wurden andere, alternative Lösungsansätze in Form der kommerziellen Software Progenesis QIP(http://www.nonlinear.com/) und eines frei verfügbaren R-Paketes synapter[144] vorgestellt. In einem Vergleich der quantitativen Analyseergebnisse von ISOQuant mit den Ergebnissen von Progenesis QIP und synapter konnte die Effektivität des entwickelten Analyseworkflows gezeigt werden.

4.1 Datenzugriff und Zusammenführung der PLGS-Ergebnisse

Der entwickelte Analyseworkflow setzt zur Speicherung der Daten für die Post-PLGS-Datenanalyse die im Kapitel 3.2 vorgestellte relationale Datenbank ein. Gegenüber der Datenspeicherung in Klartext-, CSV- oder XML-Dateien, wie dies bei vielen Analysepipelines wie SuperHirn[145], OpenMS/TOPP[146–148], etc. üblich ist, bietet eine relationale Datenbank Vorteile bei der Zugriffsgeschwindigkeit und einfache Möglichkeiten des wahlfreien, parallelen Zugriffs auf mehrere Datensätze zur gleichen Zeit. Auf diese Weise wird eine messungsübergreifende Zusammenführung der bisher unabhängig voneinander prozessierten Daten zu einem zusammenhängenden Experiment vereinfacht. Durch die Datenabfragesprache SQL können dabei komplizierte, logische Zusammenhänge der gespeicherten Daten direkt abgefragt werden. So müssen keine externen Logiken entwickelt werden um Daten nach bestimmten Kriterien zu filtern. Das Schema der relationalen Datenbank ist angelehnt an PLGS-Datenstrukturen und eignet sich deshalb speziell zur Speicherung der datenunabhängigen MS^E/$HDMS^E$/$UDMS^E$-Daten eines typischen labelfreien Schrotschuss-Proteomik-LC-MS-Experimentes, bei dem mehrere Proben in mehreren Replikaten erfasst und vergleichend untersucht werden. Durch den hohen Grad der Spezialisierung auf PLGS-vorprozessierte, labelfreie MS^E/$HDMS^E$/$UDMS^E$-Daten kann das entwickelte Datenbankschema ohne Adaptation keine weiteren Strategien der MS-basierten Proteomik-Experimente abbilden. Bei der künftigen Entwicklung werden unterschiedliche datenunabhängige, datenabhängige und gezielte Akquisitionsmethoden sowie ihre Kombinationen, aber auch Markierungsmethoden wie SILAC[125] berücksichtigt.

4.2 Retentionszeitalignment

Das Retentionszeitalignment spielt bei einer messungsübergreifenden Analyse massenspektrometrischer Daten eine wichtige Rolle[140]. Viele verschiedene Methoden zur Ermittlung und Korrektur von linearen und nichtlinearen Fluktuationen der Retentionszeit in LC-MS-Messungen wurden in der Literatur beschrieben und evaluiert[169–172]. Die Retentionszeitalignmentalgorithmen können in mehrere Klassen abhängig von der verwendeten Eingabeinformation eingeteilt werden. Eine Gruppe dieser Algorithmen schließt von bereits vorhandenen Peptididentifikationen auf die Retentionszeitverschiebungen. Jedoch können die Peptididentifikationen in einzelnen LC-MS-Messungen eines Experiments lückenhaft sein oder gänzlich fehlen (s. Kapitel 1.8.1). Eine andere Gruppe der Algorithmen

berechnet das Retentionszeitalignment anhand der Korrelationen chromatographischer Signale. Dabei werden rechenintensiv Korrelationen von XIC, TIC (engl. total ion current) oder Signalprofilen ausgewertet. Die Reduktion von detektierten LC-MS-Signalen auf die abstrakte Repräsentation als EMRT hat einen teilweisen Verlust chromatographischer Informationen zur Folge. So gehören der vorgestellte Algorithmus DRTW und seine effizienteren Ableger FastDRTW, LinDRTW sowie FastLinDRTW in eine weitere Klasse von Retentionszeitalignmentalgorithmen, die das Alignment anhand der extrahierten Features berechnen und sich damit auf die Qualität der vorangegangenen Feature-Detektion verlassen. Die in dieser Arbeit entwickelten Algorithmen werten extrahierte Features in Form von EMRTs aus und benötigen keine detaillierten chromatographischen Informationen. Sie setzen keine Peptididentifikation voraus und eignen sich prinzipiell gleichermaßen für das Retentionszeitalignment unterschiedlicher Arten zeitaufgelöster Daten speziell LC-MS-Daten in verschiedenen Bereichen der Proteomik, Lipidomik oder Metabolomik.

Zur Findung eines optimalen Alignments wird bei den entwickelten Algorithmen das Prinzip der dynamischen Programmierung verwendet, deren Eignung zu einer optimalen Lösung des Problems des Retentionszeitalignments mehrfach untersucht und bestätigt wurde[158, 170, 173–177].

Bei einem MS^E/$HDMS^E$/$UDMS^E$-Experiment können mehrere hunderttausend Features in einer einziger Messung vorliegen. Die Alignment-Algorithmen auf Basis der dynamischen Programmierung haben ursprünglich eine quadratische Zeit- und Speicherkomplexität und stellen dadurch in Verbindung mit großen Datenmengen hohe Anforderungen an die Ausführungsumgebung. Sie eignen sich unter Umständen nicht für das Alignment von solchen Datenmengen auf handelsüblicher Hardware. Neben der Berechnung eines optimalen Retentionszeitalignments stand deshalb bei der Entwicklung der Algorithmen deren Effizienz im Vordergrund. Die Algorithmen FastDRTW, LinDRTW bedienen sich unterschiedlicher Strategien zur Optimierung der dynamischen Programmierung. FastLinDRTW kombiniert auf eine neuartige Weise die Vorteile der beiden Optimierungstechniken. Dadurch erreicht FastLinDRTW bei gleicher Qualität des Retentionszeitalignments eine überlegene Effizienz gegenüber DRTW, FastDRTW und LinDRTW auf (s. Kapitel 3.3.6). Der FastLinDRTW-Algorithmus ermöglicht eine schnelle Alignmentberechnung für sehr große Eingabesequenzen auf handelsüblicher Hardware in linearer Zeit und auf linearem Speicher. Durch eine problemspezifische Adaption der Distanzfunktion wäre eine Anwendung des FastLinDRTW-Algorithmus auf andere wissenschaftliche Fragestellungen denkbar, die eine

ressourcenschonende Berechnung paarweiser Alignments von sehr großen Eingabesequenzen voraussetzen.

Durch das paarweise Retentionszeitalignment wird eine Untermenge der möglichen Übereinstimmungen zwischen zwei LC-MS-Messungen gefunden. Die Zeitverschiebungen für alle Features werden durch lineare Interpolation zwischen den Zeitverschiebungen an korrespondierenden Feature-Paaren interpretiert. Zur Steigerung der Genauigkeit böte sich die Anwendung nichtlinearer Interpolationsmethoden an, z.B. der Bezierkurven oder der Spline-Interpolation[178, 179]. Zugunsten der Effizienz wird an der linearen Interpolation als ausreichend genauen Lösung festgehalten, zumal die potentiellen lokalen Interpolationsfehler keinen erkennbaren Einfluss auf das Ergebnis des Feature-Clustering haben.

Das paarweise Retentionszeitalignment aller Messungen eines Experiments gegen eine experimentweite Referenz ermöglicht ein multiples Retentionszeitalignment, bei dem messungsspezifische nichtlineare Zeitverschiebungen gegenüber einem gemeinsamen Referenzzeitraum analysiert werden. Analog zu Lösungen in anderen Studien dient dabei eine der Messungen des Experiments als Referenz[169]. Der resultierende Referenzzeitraum ergibt sich somit unmittelbar aus der Wahl der Referenzmessung. Die Retentionszeiten alignierter Messungen verteilen sich folglich über die Dauer der Referenz. Ein Alleinstellungsmerkmal der in dieser Arbeit entwickelten Algorithmen für das paarweise Retentionszeitalignment ist die Fähigkeit, das Retentionszeitalignment für MS-Messungen mit LC-Gradientzeiten zu berechnen, die stark voneinander abweichen, z.B. Alignment von Messungen mit Gradientzeiten von 90 min und 180 min. Durch die Verwendung einer einheitlichen Referenzmessung ergibt sich die Möglichkeit innerhalb eines Experiments Messungen mit unterschiedlichen Gradientenzeiten zu kombinieren. Bei einem solchen multiplen Retentionszeitalignment werden die Gradientenzeiten aller Messungen des Experiments auf die Gradientenzeit der Referenzmessung skaliert. Eine weitere Verbesserung dieser Vorgehensweise könnte durch die Abbildung der skalierten Gradientzeiten auf eine benutzerdefinierte Standard-Gradientenzeit erreicht werden. Dies würde die Abhängigkeit der resultierenden Gradientzeiten von der Auswahl der Referenzmessung entkoppeln und damit die Parametrisierung weiterer Analyseschritte vereinfachen.

4.3 Feature-Clustering

Die labelfreie quantitative Schrotschuss-Proteomik stützt sich auf den Vergleich der in verschiedenen LC-MS-Messungen erfassten Intensitäten gleicher Peptide, die vor der

Identifikation als Features vorliegen. Um Features messungsübergreifend zu vergleichen, wird ihre Zusammengehörigkeit mit der im Kapitel 3.4 vorgestellten Methode analysiert. Nach dem multiplen Retentionszeitalignment sind die Elutionszeiten korrespondierender Features jeweils über alle Messungen des Experiments angenähert, ihre weiteren Eigenschaften wie das *m/z* oder die Driftzeit können systemabhängig von Messung zu Messung unterschiedlich stark variieren. Die Unschärfe der Übereinstimmung von Eigenschaften korrespondierender Features kann mit Hilfe der Clusteranalyse zur Gruppierung der entsprechenden Features überwunden werden. Für diesen Zweck wurden bereits Algorithmen wie das Hierarchische Complete-Linkage- oder das k-Means-Clustering eingesetzt[141, 180–182]. Eine Anwendung dieser Algorithmen bringt neue Herausforderungen mit sich. Die Hierarchische Clusteranalyse ordnet alle untersuchten Elemente systematisch zu einer baumartigen Struktur. Um Features mit Hilfe der Hierarchischen Clusteranalyse in unabhängige Gruppen nach ihrer Übereinstimmung aufzuteilen, muss der Ablauf der Analyse beim Erreichen einer bestimmten, zunächst unbekannten Clustering-Tiefe unterbrochen werden. Der Mehraufwand zur Schätzung der korrekten Tiefe sowie die quadratische Komplexität des Algorithmus schränken den Einsatz der Hierarchischen Clusteranalyse bei komplexen Daten ein. Die k-Means-Methode kann zwar eine bessere Zeit- und Speichereffizienz als die Hierarchische Clusteranalyse vorweisen, setzt jedoch die Kenntnis der Anzahl von Zielclustern voraus. Die Berechnung der optimalen Anzahl der Zielcluster für k-Means stellt ein eigenständiges, informatisches Problem dar und erhöht zumindest die praktische Zeitkomplexität. Die in dieser Arbeit eingesetzte, dichtebasierte Clustering-Methode - DBSCAN[165] benötigt kein Abbruchkriterium und keine Heuristik über die Anzahl der Cluster. DBSCAN nimmt an, dass Objekte, die zusammengehören, mit einer gewissen Dichte nah beieinander im untersuchten geometrischen Raum zu finden sind. Neben der eigentlichen Aufgabe zur Gruppierung der korrespondierenden Features ist DBSCAN zusätzlich in der Lage alleinstehende Features, für die es keine entsprechenden Features in anderen Messungen gibt, als Rauschen zu klassifizieren. Durch die Definition einer Mindestzahl der erwarteten Features in einem Cluster kann der Benutzer auf den Ausgang der DBSCAN-Clusteranalyse Einfluss nehmen.

Die Überführung der Features in einen mehrdimensionalen, homogenen, geometrischen Raum dient der Optimierung und Vereinfachung der Evaluierung von Distanzen zwischen den Features während der Clusteranalyse. Die instrument- und experimentspezifischen Auflösungsparameter für die Masse, Retentionszeit und die Driftzeit beeinflussen zusätzlich das Ergebnis des Feature-Clusterings. Hieraus resultiert ein für alle Clustering-Methoden

gemeinsames Problem. Bei einer zu hoch gewählten Experiment-spezifischen Auflösung kann die Größe der einzelnen Cluster durch die Clustering-Methode unterschätzt und bei einer zu niedrigen Auflösung entsprechend überschätzt werden. Im ersten Fall werden zusammengehörige Features in mehrere Cluster aufgeteilt, so dass im Ergebnis viele kleine Cluster auftreten. Im zweiten Fall werden unterschiedliche Features in einen Cluster eingruppiert, so dass im Ergebnis wenige jedoch große Cluster auftreten. Das Problem könnte durch eine algorithmische Schätzung der Auflösungsparameter und eine nachträgliche Analyse der resultierenden Cluster auf ihre Plausibilität umgangen werden. Möglicherweise könnte die Parameterabhängigkeit des Clustering-Ergebnisses durch Einsatz anderer Clustering-Algorithmen reduziert werden, die in der Lage sind Cluster unterschiedlicher Dichte zu erkennen, wie z.B. OPTICS[183] oder DeLi-Clu[184].

Der DBSCAN-Algorithmus besitzt theoretisch eine lineare Zeit- und Speicherkomplexität, jedes Element im untersuchten Raum wird vom Algorithmus nur einmal analysiert. Die Ermittlung der Nachbarschaft für jedes betrachtete Element kann die Gesamtkomplexität des Algorithmus bis hin zur quadratischen Komplexität steigern. Bei der beschriebenen Clustering-Prozedur werden Features über alle Messungen des Experiments gleichzeitig in einem gemeinsamen geometrischen Raum analysiert. Bei der Analyse komplexer MS^E/$HDMS^E$/$UDMS^E$-Daten könnte der Speicherbedarf die zur Verfügung stehenden Ressourcen somit übersteigen. Um den tatsächlichen Speicherbedarf für das Feature-Clustering zu reduzieren, wurde eine effiziente Preclustering-Methode entwickelt, die alle Features des Experiments grob in Subgruppen unterteilt. Hierzu wird eine simple Single-Linkage[162, 163] Clustering-Strategie iterativ auf einzelne Eigenschaften der Features angewandt. Die für Single-Linkage typischen Verkettungseffekte[185] führen zunächst zur Unterteilung des gesamten Datensatzes in große Cluster. Precluster werden dann unabhängig voneinander dem finalen Clustering mit DBSCAN zugeführt. Durch die Partitionierung des Gesamtdatensatzes ergibt sich neben der Speicherbedarfoptimierung eine Möglichkeit der Parallelisierung. So können mehrere Precluster simultan in voneinander unabhängigen Clustering-Prozessen bearbeitet werden.

4.4 Normalisierung der Feature-Intensitäten

Die Signalintensitäten massenspektrometrisch erfasster Peptide können systematische Fehler aufweisen, deren Ursprung auf die Varianz der Bedingungen während der Durchführung eines Experiments sowie auf Instabilitäten des Elektrosprays oder der

chromatographischen Methode zurückgeführt werden kann. Unterschiedliche Techniken zur Normalisierung der als Signalintensitäten erfassten Peptidmengen wurden in der Literatur beschrieben[186, 187]. Einige dieser Methoden korrigieren globale Fehler der Gesamtintensitäten aller Signale einzelner LC-MS-Messungen oder verwenden interne Standards[136, 188–190], andere untersuchen lokale Abhängigkeiten der Signalintensitäten, z.B. von der Retentionszeit[145].

Im Kapitel 3.5 wird eine neuartige multidimensionale Normalisierungsmethode zur Korrektur systematischer Fehler der Feature-Intensitäten vorgestellt, die ohne Einsatz interner Standards, datenabhängig lokale und globale systematische Fehler der Signalintensitäten korrigiert. Anhand der LOWESS-Regression untersucht die Methode nacheinander lokale Tendenzen der systematischen Fehler in Abhängigkeit von mehreren Parametern (Signalintensität, Retentionszeit und Masse) und korrigiert dann entsprechend die Intensitäten betroffener Features. So kann durch eine Änderung der LOWESS-Bandbreite eine Balance zur Korrektur der lokalen oder globalen Fehlertrends eingestellt werden.

Die Funktionsweise der entwickelten Normalisierungsmethode stützt sich auf die Grundannahme, dass beim Vergleich mehrerer Proteome die Proteine und damit ihre enzymatischen Peptide mehrheitlich keiner Regulation unterliegen. Somit kann davon ausgegangen werden, dass die massenspektrometrisch erfassten Peptide zu jedem Zeitpunkt der chromatographischen Methode, in jedem Intensitätsbereich der detektierten Signale oder in jedem Massenbereich im Durchschnitt mehrheitlich keine Unterschiede in der Expression aufweisen. Diese Annahme ermöglicht eine Analyse der systematischen Fehler der Signalintensität in Abhängigkeit von der Höhe der Signalintensität, der Retentionszeit des Peptides oder der Masse bzw. des Masse-zu-Ladung-Verhältnisses des erfassten Peptidions. Vor Allem scheinen die Höhe der Signalintensität sowie die Retentionszeit des entsprechenden Peptides den stärksten Einfluss auf den systematischen Fehler der Signalintensität auszuüben. Die Methode der multidimensionalen Normalisierung der Feature-Intensitäten setzt durch die oben beschriebene Grundannahme eine gewisse Mindestkomplexität der zu normalisierenden Daten voraus und eignet sich deshalb besonders für die Normalisierung labelfrei akquirierter Proteomikdaten. Die Methode lässt sich ohne weitere Adaptationen auf Metabolomik- oder Lipidomik-Daten ausreichender Komplexität direkt anwenden.

4.5 Filterung der Peptididentifikationen und Annotation der Feature-Cluster

Die messungsübergreifende Gruppierung korrespondierender Features durch das Feature-Clustering berücksichtigt nicht die Peptididentität der Features. Im Kapitel 3.7 wurde eine Vorgehensweise zur Annotation der Feature-Cluster vorgestellt. In durch PLGS unabhängig voneinander prozessierten Daten fehlen oft in einzelnen Messungen Identifikationen einzelner Features, deren korrespondierende Features in anderen Messungen erfolgreich identifiziert werden konnten. Unterschiedliche Möglichkeiten zur Übertragung von Peptididentifikation auf Features anderer Messungen wurden bereits untersucht[144, 148, 191, 192]. So werden bei der Annotation der Feature-Cluster in eindeutigen Fällen Identifikationen der mit PLGS identifizierten Features auf alle korrespondierenden Features ausgeweitet, indem das entsprechende gesamte Feature-Cluster durch ein Peptid annotiert wird.

Eine Definition von Mindestkriterien für die Akzeptanz von Peptididentifikationen gilt in der Proteomik als unumgänglich. Üblicherweise werden nur Peptide zur weiteren Datenanalyse zugelassen, die eine Mindestlänge der Aminosäuresequenz, einen Mindestidentifikationsscore, etc. aufweisen. Filterung der Peptididentifikationen nach benutzerdefinierten Kriterien vor der Annotation der Feature-Cluster sorgt für eine einheitliche Qualität der Peptididentifikationen im gesamten Experiment. Besonders die Einhaltung einer minimalen Replikationsrate einer Peptididentifikation ermöglicht dabei eine Steigerung der Gesamtqualität der Peptididentifikationen[190].

Neben der Übertragung von Peptididentifikationen werden während der Annotation der Feature-Cluster potentielle Identifikationskonflikte aufgelöst. Die restriktive Cluster-Annotation steigert zwar die Qualität der Peptididentifikationen durch Vermeidung von Identifikationskonfliken, reduziert jedoch gleichzeitig die Gesamtanzahl der resultierenden Peptididentifikationen. Durch Ausweichen auf die kompetitive Annotation der Feature-Cluster kann eine Balance zwischen dem Verlust der Peptididentifikationen und der Qualitätssteigerung erreicht werden. Die kompetitive Annotation wertet zur Lösung der Konflikte PLGS-Scores der Peptididentifikationen aus. Durch Einbeziehung von Eigenschaften der Features wie der Fragmentierungsinformation zur Plausibilitätsprüfung ihrer Identität könnte die Konfliktlösung weiter verbessert werden.

4.6 Filterung der False Discovery Rate

Bei der entwickelten Datenanalysestrategie werden Ergebnisse mehrerer, unabhängig voneinander vorprozessierter LC-MS-Messungen zusammengeführt und die Anzahl der identifizierten Peptide und Proteine gegenüber einzelnen Messungen verändert. Zusätzlich ändert sich die Anzahl der Identifikationen durch Datenfilter und Auflösung der Konflikte bei der Annotation der Feature-Cluster. Eine Änderung der Komposition identifizierter Peptide oder Proteine in einem Experiment zieht potentiell eine Änderung der FDR nach sich. Mit den in Kapiteln 3.8 und 3.12 vorgestellten Analyseschritten werden jeweils simple Vorgehensweisen zur separaten FDR-Beschränkung auf Ebene der Peptide und der Proteine beschrieben. Die Filter der Peptid-FDR und der Protein-FDR werden an unterschiedlichen Stellen des Analyseworkflows angewandt, sobald der vorangegangene Analyseschritt die Zahl der entsprechenden Identifikationen potentiell verändert.

Bei der Filterung der FDR erfolgt eine Klassifizierung der Identifikationen danach, ob sie aus dem Decoy-Teil der Proteindatenbank stammen. Bei Peptiden, die mehreren Proteinen zugeordnet wurden, wird von einer Decoy- oder falsch positiven Identifikation ausgegangen, sobald eines der Ursprungsproteine aus dem Decoy-Teil der Suchdatenbank stammt. Diese konservative Klassifizierung birgt zwar das Risiko der Fehleinstufung einer wahren Peptididentifikation, steigert jedoch die Qualität der resultierenden Peptididentifikationen und das Vertrauen des Anwenders in die Qualität der Analyseergebnisse.

4.7 Protein-Homologie-Filter

In Experimenten der Schrotschuss-Proteomik erschwert das Problem der Proteininferenz Rückschlüsse von identifizierten Peptiden auf die Ausgangsproteine[142]. Viele Datenbanksuchalgorithmen bewerten zwar einzelne Proteinidentifikationen mit einem Zuverlässigkeitsindex oder -score, überlassen jedoch die Entscheidung in uneindeutigen Fällen der Interpretation des Anwenders. Mehrere Ansätze zur Auflösung nichteindeutiger Peptid-Protein-Beziehungen und zur Lösung des Proteininferenz-Problems werden in der Literatur beschrieben und diskutiert[193]. Einige Algorithmen setzen spezifische Lösungen für bestimmte Datenbanksuchalgorithmen um oder benötigen herstellerspezifische Datenformate. Dadurch lassens sie sich nicht direkt auf MS^E/$HDMS^E$/$UDMS^E$-Daten übertragen. Andere Methoden berechnen rechenintensiv über komplizierte Modelle die Wahrscheinlichkeit der einzelnen Peptid- und Proteinidentifikationen.

Mit dem im Kapitel 3.9 vorgestellten Protein-Homologie-Filter wurde eine einfache Lösung des Problems der Proteininferenz entwickelt. Der Protein-Homologie-Filter analysiert in mehreren Iterationen alle Peptid-Protein-Beziehungen innerhalb eines LC-MS-Experiments. Der Algorithmus reduziert Proteinidentifikationen auf Proteine, deren Identifikation durch einzelne Peptide eindeutig belegt wird. Zusätzlich werden aus Netzwerken von nicht eindeutig identifizierten Proteinen einzelne Proteine in das Ergebnis übernommen, deren Identifikation im entsprechenden Netzwerk am wahrscheinlichsten ist. Auf diese Weise berücksichtigt der Protein-Homologie-Filter die tatsächlich beobachteten, partiellen Homologien der Proteine basierend auf den identifizierten Peptidsequenzen. Für die Berechnung der Wahrscheinlichkeit von Proteinidentifikationen innerhalb eines Beziehungsnetzwerkes werden lediglich die Anzahl der dem Protein zugeordneten Peptididentifikationen und der jeweilige PLGS-Identifikationsscore des Proteins herangezogen. Gruppen sequenzhomologer Proteine werden nach dem Protein-Homologie-Filter durch ein Protein repräsentiert. Evaluierung weiterer Eigenschaften wie der erreichten Sequenzabdeckung oder der eventuell vorhandenen Heuristiken bzw. Nachweise bestimmter Proteine aus vorangegangenen Experimenten würde die Zuverlässigkeit der Bestimmung des jeweils am wahrscheinlichsten identifizierten Proteins potentiell steigern. Die Evaluierung des jeweils am wahrscheinlichsten identifizierten Proteins kann durch Anwendung eines anwendungsspezifischen Rankings anstelle des PLGS-Identifikationsscores generalisiert werden. Der generalisierte Algorithmus kann zur Analyse des Proteininferenz-Problems in beliebigen weiteren Schrotschuss-Proteomik-Daten eingesetzt werden.

4.8 Verteilung der Peptidintensitäten

Stammt ein Peptid von mehreren Proteinen, trägt jedes seiner Ursprungsproteine zur Gesamtmenge dieses Peptides bei. Die Signalintensität spiegelt die Gesamtmenge dieses Peptides und wider und lässt zunächst keine Schlüsse über die Anteile der Ursprungsproteine an der Gesamtmenge zu. Üblicherweise werden solche geteilten Peptide von der Proteinquantifizierung ausgenommen. Ihr Ausschluss kann jedoch die Quantifizierung einzelner Proteine nachteilig beeinflussen[194]. Im Kapitel 3.10 wird eine Vorgehensweise zur anteiligen Aufteilung der Gesamtintensitäten solcher geteilter Peptide auf ihre Ursprungsproteine beschrieben. Die Methode berechnet für jede Messung die Anteile eines geteilten Peptides an entsprechenden Proteinmengen anhand der relativen Verhältnissen eindeutig zugeordneter Peptide. Werden die aufgeteilten

Peptidintensitäten zur Proteinquantifizierung herangezogen, ändern sie das Mengenverhältnis ihrer Ausgangsproteine innerhalb einer Messung nicht. Die Methode bietet eine einfache und intuitive Möglichkeit, geteilte Peptide in die Quantifizierung der Proteine einzubeziehen.

4.9 Absolute Proteinquantifizierung

Im Kapitel 3.11 wird mit TopX eine Erweiterung der verbreiteten Top3-Methode[136] zur absoluten Proteinquantifizierung auf Basis von Signalintensitäten vorgestellt. Die Top3-Methode zeigt eine ähnliche Performance der absoluten Proteinquantifizierung, wie sie mit moderneren Ansätzen erreicht wird[139], wie z.B. der iBAQ-Methode, die neben den Signalintensitäten auch die theoretische Anzahl der massenspektrometrisch erfassbaren, tryptischen Peptide eines Proteins berücksichtigt[138]. Die Berechnung der Top3-Werte gestaltet sich jedoch wesentlich einfacher als die Berechnung der iBAQ-Werte, da keine Annahmen über die theoretische Detektierbarkeit der tryptischen Peptide getroffen werden müssen. Für die Berechnung eines Top3-Wertes für ein Protein werden lediglich die drei Peptide des betrachteten Proteins mit den höchsten Signalintensitäten herangezogen. Die ursprüngliche Beschränkung der Top3-Methode auf drei Peptide hat vermutlich praktische Gründe. Im Gegensatz dazu erlaubt die TopX-Variante die Verwendung einer benutzerdefinierten Anzahl von Peptiden zur Quantifizierung eines Proteins. Wie bei der Ausgangsmethode werden mit TopX zunächst relative Proteinmengen berechnet, erst durch die Relation mit den TopX-Werten und der vordefinierten Menge eines Standardproteins werden absolute Proteinmengen für jedes identifizierte Protein ermittelt. Eine konkrete Auswirkung der Anzahl von Peptiden, die bei der TopX-Methode zur Quantifizierung eines Proteins herangezogen werden, wurde bislang nicht untersucht. Ein direkter Vergleich der unterschiedlich parametrisierten TopX-Methode mit anderen Methoden zur absoluten Proteinquantifizierung ist Gegenstand künftiger Untersuchungen. Neben den bekannten Methoden auf Basis der MS1-Signalintensität könnte eine Eignung der vielversprechenden, robusten Quantifizierungsmethoden auf Basis von MS2-Signalintensitäten wie RIBAR und xRIBAR[195] im Kontext des entwickelten Analyseworkflows evaluiert werden.

4.10 Analyseworkflow und Implementierung

Mit dem im Kapitel 3.1 vorgestellten Analyseworkflow wird eine schrittweise Strategie zur Lösung der bekannten Herausforderungen bei der Analyse von datenunabhängigen, labelfreien

MS^E/$HDMS^E$/$UDMS^E$-Daten (s. Kapitel 1.8.1) umgesetzt. Mangels frei verfügbarer Alternativen zu Waters-Algorithmen zur Vorprozessierung von MS^E/$HDMS^E$/$UDMS^E$-Rohdaten und zur proteinzentrischen Peptid- und Proteinidentifikation setzt der entwickelte Analyseworkflow nach PLGS ein und führt eine messungsübergreifende Analyse der PLGS-Ergebnisse durch. Die Implementierung des Workflows samt der beschriebenen Algorithmen wurde als quelloffenes Softwarepaket ISOQuant über eine Webpräsenz (www.isoquant.de) anderen Wissenschaftlern zur Verfügung gestellt. ISOQuant wird bereits vielerorts routinemässig eingesetzt. Eine Liste der Publikationen, die sich auf Datenanalysen mit ISOQuant stützen, wird in Tabelle 7 aufgeführt.

Im Gegensatz zu den alternativen Analyseworkflows wie Progenesis QIP oder synapter reduziert ISOQuant die Benutzerinteraktionen auf ein Minimum, indem der Zugriff auf PLGS-Daten und alle weiteren Analyseschritte automatisiert ausgeführt werden. Damit eignet sich ISOQuant besonders für routinemäßige, ressourcenschonende Hochdurchsatzanalysen mit einem minimalen Personalaufwand. Als eine eigenständige Applikation mit einer graphischen Benutzungsschnittstelle orientiert sich ISOQuant am Endanwender und setzt für die Konfiguration lediglich Grundkenntnisse über die zu analysierenden Daten voraus.

Die Spezialisierung des Analyseworkflows ermöglicht momentan ausschließlich Analysen datenunabhängiger MS^E/$HDMS^E$/$UDMS^E$-Daten, die aus einem typischen labelfreien Schrotschuss-Proteomik-Experiment stammen. Eine Generalisierung der verwendeten Datenstrukturen und Algorithmen würde den Einsatz des Workflows für die Analyse beliebiger DIA, DDA oder gezielt akquirierter, labelfreier LC-MS-Daten ermöglichen. Teile des Workflows insbesondere das Retentionszeitalignment, Feature-Clustering und die multidimensionale Normalisierung der Signalintensitäten können disziplinübergreifend für die Signalanalyse von Metabolomik- oder Lipidomik-LC-MS-Daten verwendet werden. Eine Übertragung des Workflows auf die quantitative Analyse von LC-MS-Daten, die mit Markierungstechniken wie SILAC[125], ICAT[124] oder iTRAQ[127] generiert werden, bedürfte hauptsächlich spezifischer Änderungen im Bereich der Signalanalyse. Um dem Analyseworkflow mehr Flexibilität zu verleihen und künftige Weiterentwicklungen der vorgestellten Algorithmen sowie ihren Einsatz im Kontext anderer Experimenttypen anzuregen, könnten einerseits die einzelnen Analyseschritte in Form eigenständiger Module auf etablierte Analyseplattformen wie etwa OpenMS/TOPP[146–148] übertragen werden, andererseits kann ISOQuant in Richtung einer dynamischer Analysepipeline für mehrere, unterschiedliche Typen von LC-MS-Daten weiterentwickelt werden.

Tab. 7: Einsatz von ISOQuant in verschiedenen Studien.

Publikation	Art der Analyse	Referenz
Patzig et al., 2011	Qualitative und quantitative Analyse des murinen Myelin-Proteoms	[196]
Tenzer et al., 2011	Qualitative und quantitative Analyse der Proteinkorona von anorganischen Nanopartikeln unterschiedlicher Größe im humanen Blutplasma	[197]
Michel et al., 2013	Qualitative und quantitative Analyse der Proteome myeloider Suppressorzellen in unterschiedlichen Mausstämmen	[198]
Tenzer et al., 2013	Qualitative und quantitative Analyse zeitlicher Änderung der Proteinkorona von anorganischen Nanopartikeln im humanen Blutplasma	[199]
Tenzer et al., 2013	Qualitative und quantitative Analyse des Interaktoms des RNA-bindenden Proteins RALY	[200]
Distler et al., 2014	Qualitative und quantitative Analyse des HeLa-Proteoms	[98]
Distler et al., 2014	Qualitative und quantitative Analyse der Proteine der postsynaptischen Dichte im murinen Hippocampus	[201]
Docter et al., 2014	Qualitative und quantitative Analyse zeitlicher Änderung der Proteinkorona von anorganischen Nanopartikeln im humanen Blutplasma	[202]
Schick et al., 2014	Qualitative Analyse der Proteinkorona von anorganischen Nanopartikeln im humanen Blutplasma	[203]
Kuharev et al., 2015	Vergleich von ISOQuant mit anderen Analyseworkflows anhand der qualitativen und quantitativen Analyse von HeLa-Hefe-*E.coli*-Metaproteomdatensätzen	[153]
Ritz et al., 2015	Qualitative und quantitative Analyse der Proteinkorona von anorganischen Nanopartikeln im humanen Blutplasma unter verschiedenen Bedingungen	[204]

4.11 Vergleich mit PLGS

Im Kapitel 3.14 werden qualitative und quantitative Ergebnisse der Datenanalyse mit PLGS und der anschließenden Datenanalyse mit ISOQuant gegenübergestellt. Da die Datenanalyse mit ISOQuant auf die Datenanalyse mit PLGS aufbaut, können die in ISOQuant verwendeten Algorithmen die Gesamtanzahl der Peptid- und Proteinidentifikationen prinzipiell nur reduzieren. So reduzierte ISOQuant bei der Analyse des mit UDMS^{E} und 180 min Gradientenzeit akquirierten HeLa-Proteoms die Gesamtanzahl der Peptididentifikationen über alle technischen Replikate um 36,7% und der Proteinidentifikationen um 10,2% gegenüber PLGS. Der Einbruch der Gesamtanzahlen von Identifikationen lässt sich vor Allem auf Verlust von schwach belegten Identifikationen durch Datenfilter und Konfliktauflösungen bei der restriktiven Annotation der Feature-Cluster zurückführen. Während die Gesamtanzahl der Peptid- und Proteinidentifikationen über alle technischen Replikate durch die Analyse mit ISOQuant sinkt, wird in einer einzelnen LC-MS-Messung im Durchschnitt meist eine höhere Anzahl an Peptid und Proteinidentifikationen als nach der PLGS-Analyse erreicht. Die Abweichung der Anzahl von Peptididentifikationen pro LC-MS-Messung betrug nach ISOQuant-Analyse des HeLa-Proteoms abhängig von der Akquisitionsmethode zwischen -1% und 33%. Die durchschnittliche Anzahl der Proteinidentifikationen in einer Messung stieg jedoch unabhängig der verwendeten Akquisitionsmethode um über 20% an. Die merkliche Steigerung der durchschnittlichen Anzahlen von Peptid- und Proteinidentifikationen pro LC-MS-Messung wird hauptsächlich durch Schließen der Identifikationslücken mit der Übertragung der Peptididentifikationen zwischen den Messungen des Experiments während der Annotation der Feature-Cluster erreicht. Die Steigerung der durchschnittlichen Anzahlen von Identifikationen geht mit einer signifikanten Verbesserung der Reproduzierbarkeit der Peptid- und Proteinidentifikation einher, so dass der Anteil der in allen Messungen des Experiments übereinstimmend identifizierten Peptide und Proteine durch die ISOQuant-Analyse gegenüber PLGS jeweils signifikant erhöht wird (s. Abbildung 21 auf Seite 81). Neben der signifikant gesteigerten Reproduzierbarkeit der Identifikation reduziert ISOQuant die Varianz der berechneten Proteinquantitäten zwischen den technischen Replikaten. Dies wird hauptsächlich durch die restriktive Cluster-Annotation und die Korrektur systematischer Fehler durch die multidimensionale Normalisierung der Feature-Intensitäten erreicht. Da jedoch die Qualität der Normalisierung eng mit dem Erfolg des Feature-Clusterings und damit auch des Retentionszeitalignments zusammenhängt, leisten alle Schritte der Signalanalyse jeweils einen wichtigen Beitrag zur Reduktion der Varianz von Proteinmengen zwischen

den technischen Replikaten. Die Datenanalyse mit ISOQuant steigert somit signifikant die Reproduzierbarkeit der Proteinquantifizierung.

4.12 Vergleich mit synapter und Progenesis

Neben der massenspektrometrischen Methode und der praktischen Aufbereitung biologischer Proben haben die Schritte der Datenanalyse bzw. die für diesen Zweck verwendete Software einen kritischen Einfluss auf die Ergebnisse eines LC-MS-Experiments[205]. Im Kapitel 3.15 wird der Einfluss des Analyseworkflows systematisch untersucht und mit Ergebnissen von Progenesis QIP und synapter verglichen. Dabei wird die Gesamtperformance der Proteinidentifikation und -quantifizierung der drei Analyseworkflows anhand der erzielten Analyseergebnisse auf Basis identischer LC-MS-Daten erfasst.

Typischerweise werden zur Visualisierung der quantitativen Performance von Analysetools einzelne Proteine in bekannten Mengen zu einem Hintergrundproteom hinzugefügt[167]. Eine Betrachtung nur weniger Proteine reduziert jedoch die statistische Aussagekraft der Untersuchungen stark. Um diesem Problem vorzubeugen und eine statistisch gesicherte Vergleichbarkeit der quantitativen Analyseergebnisse zu gewährleisten, wurde ein Satz aus zwei Metaproteomen entworfen, welche Proteome dreier Spezies in bekannten Mengen enthalten. Durch gleiche Mengen an humanen Proteinen in beiden Proben wird ein komplexer Hintergrund unregulierter Proteine simuliert, während durch unterschiedliche Mengen bakterieller und mykotischer Proteine jeweils vordefinierte Regulationsverhältnisse zwischen den beiden Proben simuliert werden. Die beiden Metaproteome wurden mit MS^E und $UDMS^E$ in mehreren Replikatmessungen erfasst und die resultierenden Rohdaten so kombiniert, dass jeweils drei Evaluierungsdatensätze mit jeder der drei Softwarelösungen analysiert werden konnten.

Die drei verglichenen Analyseworkflows unterscheiden sich in ihren Analyseschritten und dem Informationsgehalt der ausgegebenen Analyseergebnisse, wodurch ein direkter Vergleich erschwert wird. Auf Grund unterschiedlicher Scoring-Algorithmen unterscheiden sich die Identifikationsscores. Die Sequenzabdeckung der Proteine wird durch synapter und ISOQuant ausgegeben, in Progenesis QIP jedoch nicht. Die Ergebnisse von synapter werden für jede Messung einzeln in separate Dateien ausgegeben und mussten manuell zu messungsübergreifenden Berichten zusammengefasst werden. Während ISOQuant für jedes Protein die Anzahlen der ihm zugeordneten Peptide separat nach ihrer Klassifikation ausgibt, werden von Progenesis QIP für jedes Protein jeweils die Gesamtanzahl der Peptide

für seine Identifikation und für seine Quantifizierung angegeben. In synapter-Ergebnissen mussten Peptididentifikationen nachträglich für jedes Protein gezählt werden. Informationen über Proteingruppen werden nur von Progenesis QIP und ISOQuant berichtet, da synapter das Problem der Proteininferenz nicht behandelt. Auf Grund der vielen Unterschiede der Softwarepakete in ihrer Datenhandhabung wurden alle Analyseergebnisse nachträglich in ein einheitliches Format überführt und wurden zwecks Vergleichbarkeit zusätzlich nach gleichen Qualitätskriterien gefiltert.

In den Ergebnissen der drei Softwarepakete konnte für alle drei Datensätze eine Reduktion der Proteinidentifikationen durch die Applikation der Peptid- und Replikationsfilter beobachtet werden. Vor der zusätzlichen Filterung berichtete synapter in allen drei Datensätzen jeweils die meisten Proteinidentifikationen. Interessanterweise brachen jedoch die Anzahlen der Identifikationen nach der Filterung stark ein, so dass in synapter-Ergebnissen jeweils die wenigsten Proteinidentifikationen erhalten blieben. Eine hohe Übereinstimmung der Proteinidentifikationen konnte zwischen den Ergebnissen der drei Softwarepakete für alle Datensätze aber auch zwischen den Datensätzen für jede Software separat beobachtet werden. Die Höhe der Übereinstimmung der Identifikationen zwischen UDMS^{E} und den kombinierten Daten belegt eine erfolgreiche Übertragung der Proteinidentifikationen von UDMS^{E}- auf MS^{E}-Daten durch alle Softwarepakete.

Bei einer guten technischen Reproduzierbarkeit bzw. der Präzision der Proteinquantifizierung erwartet man eine niedrige Varianz der berechneten Mengen eines Proteins innerhalb der technischen Replikate einer Probe. Auch wenn in Ergebnissen der verglichenen Softwarelösungen für alle drei Datensätze jeweils eine mit anderen Studien vergleichbare Varianz der Proteinquantifizierung beobachtet wurde[167, 206], konnten zum Teil große Unterschiede der Reproduzierbarkeit der Proteinquantifizierung zwischen den Analyseworkflows und den einzelnen Datensätzen beobachtet werden. Angesichts der hohen Komplexität der Metaproteomproben erreichten Progenesis QIP und ISOQuant eine außergewöhnliche Präzision der Quantifizierung bei der Analyse des MS^{E}-Datensatzes mit jeweils mittleren Variationskoeffizienten der berichteten Proteinmengen von unter 5%. Progenesis QIP quantifizierte die UDMS^{E}- und die kombinierten Daten präziser als ISOQuant oder synapter. Die synapter-Ergebnisse wiesen für jeden Datensatz jeweils die höchste Varianz der Proteinquantifizierung auf, vermutlich, weil synapter keine Normalisierungsroutine beinhaltet.

Durch einen Vergleich der bei der Datenanalyse erzielten, relativen Verhältnisse der Proteinmengen zwischen den beiden Proben in Form von Logratios mit den entsprechenden Erwartungswerten konnte die Richtigkeit der Proteinquantifizierung untersucht werden. Zwischen den beiden Metaproteomen wurden für *E.coli*-Proteine ein Logratio von -2,0 und für Hefeproteine ein Logratio von 1,0 erwartet. Die quantitativen Analyseergebnisse von synapter und ISOQuant entsprachen weitgehend den Erwartungswerten für die relativen Mengenverhältnissen zwischen den Metaproteomproben. Im Gegensatz dazu unterschätzte Progenesis QIP in allen Datensätzen die relativen Mengenverhältnisse zwischen den Proben deutlich. Beispielweise betrugen nach der Progenesis-Analyse der MS^E-Daten die Logratios der *E.coli*-Proteine im Mittel -0,84 (Erwartungswert -2,0) und die Logratios der Hefeproteine im Mittel 0,46 (Erwartungswert 1,0). Ähnliche Werte wurden auch bei der Analyse der beiden anderen Datensätze durch Progenesis QIP erreicht. Bei der Suche nach dem Grund für dieses Verhalten, konnte beobachtet werden, dass Progenesis QIP die Ausdehnung der Features in der Retentionszeit überschätzt und so gelegentlich ein Feature durch die Integration der Intensitäten mehrerer, unabhängiger Signale quantifiziert. Dies trägt potentiell zur Unterschätzung der Regulationsverhältnisse bei, ähnlich den Beobachtungen in iTRAQ-Experimenten, wenn mehrere koeluierende Peptide zusammen fragmentiert werden[207]. Die Tendenz zur Unterschätzung der Regulationsverhältnisse erklärt zum Teil auch die sehr hohe Präzision der Proteinquantifizierung, die bei der Datenanalyse mit Progenesis QIP erreicht wurde (s Kapitel 3.15.5, Abbildung 28) – ein gutes Beispiel dafür, dass die Präzision und die Richtigkeit der Proteinquantifizierung nicht unabhängig voneinander evaluiert werden sollten.

Bei der Betrachtung der Logratios im $UDMS^E$-Datensatz konnte in Ergebnissen von synapter und ISOQuant eine systematische Verfälschung der Regulationsverhältnisse hochabundanter *E.coli*- und Hefeproteine beobachtet werden, die vermutlich durch das bereits bekannte Problem der Kompression des dynamischen Bereichs in IMS-MS-Daten verursacht wird[168]. Dieser Effekt könnte durch Entwicklung spezieller auf diesen Datentyp abgestimmter Quantifizierungsmethoden oder eventuell durch Anwendung anderer Quantifizierungsmethoden als TopX beseitigt werden, z.B. durch Proteinquantifizierung auf Basis der Fragmentionen[208]. Die Eignung dieser Quantifizierungsmethoden zur Lösung des Problems müsste jedoch in weiteren Studien geprüft werden. Bei der Analyse des kombinierten Datensatzes konnte das Problem der systematischen Verfälschung der Regulationsverhältnisse hochabundanter Proteine durch Workflows von synapter und ISOQuant weitestgehend korrigiert bzw. umgangen werden, da für den Vergleich

ausschließlich die quantitativen Ergebnisse aus dem MS^E-Anteil des kombinierten Datensatzes verwendet wurden.

Manche Proteine konnten bei der Datenanalyse in nur einer der beiden Proben quantifiziert werden. Solche exklusive Identifikation oder Quantifizierung eines Proteins in einer Probe, welches aber eigentlich in allen Proben des Experiments vorhanden ist, kann in einer Studie zur Ermittlung von Interaktionspartnern oder bei der Suche nach Biomarkern zu Fehlinterpretationen der Analyseergebnisse führen, so dass bei einem solchen Protein fälschlicherweise von einem relevanten Kandidaten ausgegangen wird. Während die Richtigkeit der Proteinquantifizierung durch Progenesis QIP nicht das Niveau von synapter oder ISOQuant erreichte, überzeugte Progenesis QIP durch auffällig wenige Lücken in der Proteinquantifizierung, so dass weniger als 0,1% der Proteine in nur einer der beiden Metaproteomproben quantifiziert wurden. Im Gegensatz zu Progenesis QIP weisen die Analyseergebnisse von synapter recht viele Lücken auf, zwischen 26% und 32% wurden in nur einer der beiden Proben quantifiziert. Vermutlich entsteht dieses Problem bei synapter durch Mängel beim Retentionszeitalignment oder bei der Übertragung der Peptididentifikationen zwischen den Messungen. In den Ergebnissen von ISOQuant wurden 1,5% bis 5,4% aller Proteine in nur einer der beiden Proben quantifiziert. Die Lücken wurden dabei hauptsächlich bei *E.coli*- und Hefeproteinen beobachtet. Erwartungsgemäß sollten die wenigsten Lücken bei der Identifikation und der Quantifizierung der humanen Proteine berichtet werden, da sie als Hintergrundproteine in beiden Proben mit gleichen Anteilen enthalten sind. Die Detektion der Hintergrundproteine in nur einer Probe kann potentiell als eine falsch positive Identifikation gedeutet werden. Die Anzahl solcher Proteine sollte in einem Experiment deshalb im Bereich der festgelegten FDR-Grenze oder darunter liegen. Diese Bedingung wird in den Analyseergebnissen von Progenesis QIP und ISOQuant für MS^E und $UDMS^E$-Datensätze erfüllt.

In der Gesamtbetrachtung konnte eine akkurate Quantifizierung selbst niedrig-abundanter Proteine durch synapter und ISOQuant erreicht werden, wodurch die Validität der labelfreien Analysestrategien bestätigt wird, bei denen Identifikationen von Features aus einzelnen Messungen auf korrespondierende Features in anderen Messungen des Experiments übertragen werden können. Künftigen Versionen der Software Progenesis QIP könnten durch Verbesserungen bei der Feature-Detektion eine Steigerung der Richtigkeit der Proteinquantifizierung erzielen.

Mit dem Vergleich der quantitativen Analyseperformance konnte unter Anderem gezeigt werden, dass mit labelfreien datenunabhängigen Proteomuntersuchungen trotz hoher Komplexität der biologischen Proben hohe Präzision und Richtigkeit der Proteinquantifizierung erreicht werden können. Die erreichte Qualität der Proteinquantifizierung erfüllt bereits die Erwartungen systembiologischer Untersuchungen auf Basis von Abundanzen der Proteinklassen unterschiedlicher Funktionen oder Zellkompartimente[209]. Es ist dennoch anzumerken, dass die Präzision der labelfreien Quantifizierung in einem direkten Verhältnis zur Reproduzierbarkeit der Probenvorbereitung steht. Die akkurate relative Proteinquantifizierung bleibt nach wie vor ein nicht triviales, vielfältiges Problem, das nicht nur durch die verwendeten Instrumente sondern auch durch die verwendete Analysesoftware stark beeinflusst wird[210, 211].

Workflows zur Analyse labelfreier, massenspektrometrischer Proteomikdaten könnten künftig von der Modellierung fehlender Werte durch geeignete Modelle profitieren. Die entwickelten Metaproteomproben können als eine Art hochqualitative Ressource für die Evaluation der quantitativen Analyse mit künftigen Softwareentwicklungen im Bereich der MS-gestützten Proteomik eingesetzt werden. Sie eignen sich außerdem zur Ermittlung der Performance und zum Vergleich unterschiedlicher MS-Plattformen und Methoden für die labelfreie Proteinquantifizierung.

Zusammenfassung

Moderne ESI-LC-MS/MS-Techniken erlauben in Verbindung mit Bottom-up-Ansätzen eine qualitative und quantitative Charakterisierung mehrerer tausend Proteine in einem einzigen Experiment. Für die labelfreie Proteinquantifizierung eignen sich besonders datenunabhängige LC-MS-Akquisitionsmethoden wie MS^E und die entsprechenden IMS-Varianten $HDMS^E$ und $UDMS^E$. Durch ihre hohe Komplexität stellen die so erfassten Daten besondere Anforderungen an die Analysesoftware. Eine quantitative Analyse der MS^E/$HDMS^E$/$UDMS^E$-Daten blieb bislang wenigen kommerziellen Lösungen vorbehalten.

In der vorliegenden Arbeit wurden eine Strategie und eine Reihe neuartiger Methoden zur messungsübergreifenden, quantitativen Analyse von MS^E/$HDMS^E$/$UDMS^E$-Daten entwickelt und als Bestandteile der quelloffenen Software ISOQuant in der Programmiersprache Java implementiert. Für die ersten Schritte der Datenanalyse (Signaldetektion, Peptid- und Proteinidentifikation) wird die kommerzielle Software PLGS verwendet. Der entwickelte Analyseworkflow überträgt automatisiert die PLGS-Ergebnisse in eine relationale Datenbank. Zur schrittweisen Lösung datenspezifischer Probleme beinhaltet der Analyseworkflow Algorithmen für paarweises und multiples Retentionszeitalignment, Gruppierung korrespondierender Features, mehrstufige Datenfilterung, Annotation der Feature-Cluster, Normalisierung der Feature-Intensitäten, Analyse des Proteininferenz-Problems, Umverteilung der Peptidintensitäten und absolute Quantifizierung der Proteine. Mit einer ISOQuant-Analyse von HeLa-Proteomdaten konnte eine signifikante Steigerung der Reproduzierbarkeit der Proteinidentifikation und -quantifizierung gegenüber der Datenanalyse mit PLGS gezeigt werden. Um die Performance der quantitativen Datenanalyse mit anderen Lösungen zu vergleichen, wurde ein Satz aus zwei exakt definierten Metaproteomproben entworfen, massenspektrometrisch erfasst und mit Progenesis QIP, synapter und ISOQuant analysiert. Der entwickelte Analyseworkflow zeigte dabei eine hohe Reproduzierbarkeit der Analyseergebnisse und erreichte neben Progenesis QIP eine hohe Performance bei der Proteinidentifikation. Bei der Qualität der Proteinquantifizierung übertrafen die Ergebnisse der ISOQuant-Analyse die Analyseergebnisse von Progenesis QIP und synapter.

Als Bestandteile der Software ISOQuant ermöglichen die entwickelten Algorithmen und der Analyseworkflow akkurate, reproduzierbare qualitative und quantitative Proteomanalysen. Sie stellen somit ein effizientes Werkzeug für routinemäßige Hochdurchsatzanalysen labelfreier MS^E/$HDMS^E$/$UDMS^E$-Daten dar.

Eigene Publikationen

Im Rahmen dieser Arbeit sind folgende Publikationen entstanden:

Patzig, J., Jahn, O., Tenzer, S., Wichert, S. P., Monasterio-Schrader, P. de, Rosfa, S., **Kuharev, J.**, Yan, K., Bormuth, I., Bremer, J., Aguzzi, A., Orfaniotou, F., Hesse, D., Schwab, M. H., Möbius, W., Nave, K.-A., & Werner, H. B. (2011). Quantitative and Integrative Proteome Analysis of Peripheral Nerve Myelin Identifies Novel Myelin Proteins and Candidate Neuropathy Loci. *The Journal of Neuroscience*, *31*(45), 16369–16386. http://doi.org/10.1523/JNEUROSCI.4016-11.2011

Tenzer, S., Docter, D., Rosfa, S., Wlodarski, A., **Kuharev, J.**, Rekik, A., Knauer, S. K., Bantz, C., Nawroth, T., Bier, C., Sirirattanapan, J., Mann, W., Treuel, L., Zellner, R., Maskos, M., Schild, H., & Stauber, R. H. (2011). Nanoparticle Size Is a Critical Physicochemical Determinant of the Human Blood Plasma Corona: A Comprehensive Quantitative Proteomic Analysis. *ACS Nano*, *5*(9), 7155–7167. http://doi.org/10.1021/nn201950e

Michel, A., Schüler, A., Friedrich, P., Döner, F., Bopp, T., Radsak, M., Hoffmann, M., Relle, M., Distler, U., **Kuharev, J.**, Tenzer, S., Feyerabend, T. B., Rodewald, H.-R., Schild, H., Schmitt, E., Becker, M., & Stassen, M. (2013). Mast Cell–deficient KitW-sh "Sash" Mutant Mice Display Aberrant Myelopoiesis Leading to the Accumulation of Splenocytes That Act as Myeloid-Derived Suppressor Cells. *The Journal of Immunology*, *190*(11), 5534–5544. http://doi.org/10.4049/jimmunol.1203355

Tenzer, S., Docter, D., **Kuharev, J.**, Musyanovych, A., Fetz, V., Hecht, R., Schlenk, F., Fischer, D., Kiouptsi, K., Reinhardt, C., Landfester, K., Schild, H., Maskos, M., Knauer, S. K., & Stauber, R. H. (2013). Rapid formation of plasma protein corona critically affects nanoparticle pathophysiology. *Nature Nanotechnology*, *8*(10), 772–781. http://doi.org/10.1038/nnano.2013.181

Tenzer, S., Moro, A., **Kuharev, J.**, Francis, A. C., Vidalino, L., Provenzani, A., & Macchi, P. (2013). Proteome-Wide Characterization of the RNA-Binding Protein RALY-Interactome Using the in Vivo-Biotinylation-Pulldown-Quant (iBioPQ) Approach. *Journal of Proteome Research*, *12*(6), 2869–2884. http://doi.org/10.1021/pr400193j

Distler*, U., **Kuharev*, J.**, Navarro, P., Levin, Y., Schild, H., & Tenzer, S. (2014). Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nature Methods*, *11*(2), 167–170. http://doi.org/10.1038/nmeth.2767, **geteilte Erstautorenschaft*

Distler, U., **Kuharev, J.**, & Tenzer, S. (2014). Biomedical applications of ion mobility-enhanced data-independent acquisition-based label-free quantitative proteomics. *Expert Review of Proteomics*, *11*(6), 675–684. http://doi.org/10.1586/14789450.2014.971114

Distler, U., Schmeisser, M. J., Pelosi, A., Reim, D., **Kuharev, J.**, Weiczner, R., Baumgart, J., Boeckers, T. M., Nitsch, R., Vogt, J., & Tenzer, S. (2014). In-depth protein profiling of the postsynaptic density from mouse hippocampus using data-independent acquisition proteomics. *PROTEOMICS*, *14*(21-22), 2607–2613. http://doi.org/10.1002/pmic.201300520

Docter, D., Distler, U., Storck, W., **Kuharev, J.**, Wünsch, D., Hahlbrock, A., Knauer, S. K., Tenzer, S., & Stauber, R. H. (2014). Quantitative profiling of the protein coronas that form around nanoparticles. *Nature Protocols*, *9*(9), 2030–2044. http://doi.org/10.1038/nprot.2014.139

Kuharev, J., Navarro, P., Distler, U., Jahn, O., & Tenzer, S. (2015). In-depth evaluation of software tools for data-independent acquisition based label-free quantification. *PROTEOMICS*. http://doi.org/10.1002/pmic.201400396

Ritz, S., Schöttler, S., Kotman, N., Baier, G., Musyanovych, A., **Kuharev, J.**, Landfester, K., Schild, H., Jahn, O., Tenzer, S., & Mailänder, V. (2015). The Protein Corona of Nanoparticles: Distinct Proteins Regulate the Cellular Uptake. *Biomacromolecules*. http://doi.org/10.1021/acs.biomac.5b00108

Literaturverzeichnis

[1] Anderson, N. G., & Anderson, N. L., Twenty years of two-dimensional electrophoresis: Past, present and future. *ELECTROPHORESIS*, *17*(3), 443–453. doi:10.1002/elps.1150170303

[2] Gygi, S. P., Rochon, Y., Franza, B. R., & Aebersold, R., Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, *19*(3), 1720–1730. Retrieved from http://mcb.asm.org/content/19/3/1720

[3] Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J.-C., Yan, J. X., Gooley, A. A., Hughes, G., Humphery-Smith, I., Williams, K. L., & Hochstrasser, D. F., From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Nature Biotechnology*, *14*(1), 61–65. doi:10.1038/nbt0196-61

[4] Apweiler, R. et al., Approaching clinical proteomics: Current state and future fields of application in cellular proteomics. *Cytometry. Part A: the journal of the International Society for Analytical Cytology*, *75*(10), 816–832. doi:10.1002/cyto.a.20779

[5] Campbell, N. A., Reece, J. B., & Markl, J., *Biologie* (6th ed.). Heidelberg u.a.: Spektrum Akademischer Verlag.

[6] Cravatt, B. F., Simon, G. M., & Yates, J. R., The biological impact of mass-spectrometry-based proteomics. *Nature*, *450*(7172), 991–1000. doi:10.1038/nature06525

[7] DonaldáSedgwick, R., & others, Fast atom bombardment of solids (FAB): A new ion source for mass spectrometry. *Journal of the Chemical Society, Chemical Communications*, (7), 325–327.

[8] Munson, M. S. B., & Field, F. H., Chemical ionization mass spectrometry. i. general introduction. *Journal of the American Chemical Society*, *88*(12), 2621–2630. doi:10.1021/ja00964a001

[9] Carroll, D. I., Dzidic, I., Stillwell, R. N., Haegele, K. D., & Horning, E. C., Atmospheric pressure ionization mass spectrometry. corona discharge ion source for use in a liquid chromatograph-mass spectrometer-computer analytical system. *Analytical Chemistry*, *47*(14), 2369–2373. doi:10.1021/ac60364a031

[10] Karas, M., Bachmann, D., Bahr, U., & Hillenkamp, F., Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International Journal of Mass Spectrometry and Ion Processes*, *78*, 53–68. doi:10.1016/0168-1176(87)87041-6

[11] Yamashita, M., & Fenn, J. B., Electrospray ion source. another variation on the free-jet theme. *The Journal of Physical Chemistry*, *88*(20), 4451–4459. Retrieved from http://pubs.acs.org/doi/abs/10.1021/j150664a002

[12] Paul, W., & Steinwedel, H., Ein neues massenspektrometer ohne magnetfeld. *Zeitschrift Naturforschung Teil A*, *8*, 448.

[13] Church, D. A., Storage-ring ion trap derived from the linear quadrupole radio-frequency mass filter. *Journal of Applied Physics*, *40*(8), 3127–3134. doi:10.1063/1.1658153

[14] Comisarow, M. B., & Marshall, A. G., Fourier transform ion cyclotron resonance spectroscopy. *Chemical Physics Letters*, *25*(2), 282–283. doi:10.1016/0009-2614(74)89137-2

[15] Makarov, A., Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*, *72*(6), 1156–1162. doi:10.1021/ac991131p

[16] Stephens, W., A pulsed mass spectrometer with time dispersion. In *Physical review* (Vol. 69, pp. 691–691). AMERICAN PHYSICAL SOC ONE PHYSICS ELLIPSE, COLLEGE PK, MD 20740-3844 USA.

[17] McLafferty, F. W., & Bockhoff, F. M., Separation/identification system for complex mixtures using mass separation and mass spectral characterization. *Analytical Chemistry*, *50*(1), 69–76. doi:10.1021/ac50023a021

[18] Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., & Watanabe, C., Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings of the National Academy of Sciences*, *90*(11), 5011–5015. Retrieved from http://www.pnas.org/content/90/11/5011

[19] James, P., Quadroni, M., Carafoli, E., & Gonnet, G., Protein identification by mass profile fingerprinting. *Biochemical and Biophysical Research Communications*, *195*(1), 58–64. doi:10.1006/bbrc.1993.2009

[20] Mann, M., Højrup, P., & Roepstorff, P., Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biological Mass Spectrometry*, *22*(6), 338–345. doi:10.1002/bms.1200220605

[21] Pappin, D. J. C., Hojrup, P., & Bleasby, A. J., Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology*, *3*(6), 327–332. doi:10.1016/0960-9822(93)90195-T

[22] Yates, J. R., Speicher, S., Griffin, P. R., & Hunkapiller, T., Peptide mass maps: A highly informative approach to protein identification. *Analytical Biochemistry*, *214*(2), 397–408. doi:10.1006/abio.1993.1514

[23] Yates III, J. R., Mass spectrometry: From genomics to proteomics. *Trends in Genetics*, *16*(1), 5–8. doi:10.1016/S0168-9525(99)01879-X

[24] Taylor, G., Disintegration of water drops in an electric field. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *280*(1382), 383–397. doi:10.1098/rspa.1964.0151

[25] Rayleigh, L., XX. on the equilibrium of liquid conducting masses charged with electricity. *Philosophical Magazine Series 5*, *14*(87), 184–186. doi:10.1080/14786448208628425

[26] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science*, *246*(4926), 64–71. doi:10.1126/science.2675315

[27] Ho, C., Lam, C., Chan, M., Cheung, R., Law, L., Lit, L., Ng, K., Suen, M., & Tai, H., Electrospray ionisation mass spectrometry: Principles and clinical applications. *The Clinical Biochemist Reviews*, *24*(1), 3–12. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1853331/

[28] Wilm, M. S., & Mann, M., Electrospray and taylor-cone theory, dole's beam of macromolecules at last? *International Journal of Mass Spectrometry and Ion Processes*, *136*(23), 167–180. doi:10.1016/0168-1176(94)04024-9

[29] Wilm, M., & Mann, M., Analytical properties of the nanoelectrospray ion source. *Analytical Chemistry*, *68*(1), 1–8. doi:10.1021/ac9509519

[30] Juraschek, R., Dülcks, T., & Karas, M., NanoelectrosprayMore than just a minimized-flow electrospray ionization source. *Journal of the American Society for Mass Spectrometry*, *10*(4), 300–308. doi:10.1016/S1044-0305(98)00157-3

[31] Tang, L., & Kebarle, P., Effect of the conductivity of the electrosprayed solution on the electrospray current. factors determining analyte sensitivity in electrospray mass spectrometry. *Analytical Chemistry*, *63*(23), 2709–2715. doi:10.1021/ac00023a009

[32] Tang, L., & Kebarle, P., Dependence of ion intensity in electrospray mass spectrometry on the concentration of the analytes in the electrosprayed solution. *Analytical Chemistry*, *65*(24), 3654–3668. doi:10.1021/ac00072a020

[33] Grace, J. M., & Marijnissen, J. C. M., A review of liquid atomization by electrical means. *Journal of Aerosol Science*, *25*(6), 1005–1019. doi:10.1016/0021-8502(94)90198-8

[34] Cloupeau, M., & Prunet-Foch, B., Electrohydrodynamic spraying functioning modes: A critical review. *Journal of Aerosol Science*, *25*(6), 1021–1036. doi:10.1016/0021-8502(94)90199-6

[35] De La Mora, J. F., & Loscertales, I. G., The current emitted by highly conducting taylor cones. *Journal of Fluid Mechanics*, *260*, 155–184. doi:10.1017/S0022112094003472

[36] Noymer, P. D., & Garel, M., STABILITY AND ATOMIZATION CHARACTERISTICS OF ELECTROHYDRODYNAMIC JETS IN THE CONE-JET AND MULTI-JET MODES. *Journal of Aerosol Science*, *31*(10), 1165–1172. doi:10.1016/S0021-8502(00)00019-7

[37] Zeleny, J., Instability of electrified liquid surfaces. *Physical Review*, *10*(1), 1.

[38] Juraschek, R., & Röllgen, F. W., Pulsation phenomena during electrospray ionization. *International Journal of Mass Spectrometry*, *177*(1), 1–15. doi:10.1016/S1387-3806(98)14025-3

[39] Valaskovic, G. A., Murphy III, J. P., & Lee, M. S., Automated orthogonal control system for electrospray ionization. *Journal of the American Society for Mass Spectrometry*, *15*(8), 1201–1215. doi:10.1016/j.jasms.2004.04.033

[40] Gapeev, A., Berton, A., & Fabris, D., Current-controlled nanospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry*, *20*(7), 1334–1341. doi:10.1016/j.jasms.2009.03.007

[41] Straub, R. F., & Voyksner, R. D., Negative ion formation in electrospray mass spectrometry. *Journal of the American Society for Mass Spectrometry*, *4*(7), 578–587. doi:10.1016/1044-0305(93)85019-T

[42] Mamyrin, B., Karataev, V., Shmikk, D., & Zagulin, V., The massreflect ron, a new non-magnetic time-of-flight mass spectrometer with high resolution. *Zh. Eksp. Teor. Fiz*, *64*, 82–89.

[43] Olsen, J. V., Godoy, L. M. F. de, Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., & Mann, M., Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a c-trap. *Molecular & Cellular Proteomics*, *4*(12), 2010–2021. doi:10.1074/mcp.T500030-MCP200

[44] Mabud, M. A., Dekrey, M. J., & Graham Cooks, R., Surface-induced dissociation of molecular ions. *International Journal of Mass Spectrometry and Ion Processes*, *67*(3), 285–294. doi:10.1016/0168-1176(85)83024-X

[45] Martin, S. A., Hill, J. A., Kittrell, C., & Biemann, K., Photon-induced dissociation with a four-sector tandem mass spectrometer. *Journal of the American Society for Mass Spectrometry*, *1*(1), 107–109. doi:10.1016/1044-0305(90)80013-D

[46] Little, D. P., Speir, J. P., Senko, M. W., O'Connor, P. B., & McLafferty, F. W., Infrared multiphoton dissociation of large multiply charged ions for biomolecule sequencing. *Analytical Chemistry*, *66*(18), 2809–2815. doi:10.1021/ac00090a004

[47] Zubarev, R. A., Kelleher, N. L., & McLafferty, F. W., Electron capture dissociation of multiply charged protein cations. a nonergodic process. *Journal of the American Chemical Society*, *120*(13), 3265–3266.

[48] Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., & Hunt, D. F., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(26), 9528–9533. doi:10.1073/pnas.0402700101

[49] Biemann, K., Mass spectrometry of peptides and proteins. *Annual Review of Biochemistry*, *61*(1), 977–1010. doi:10.1146/annurev.bi.61.070192.004553

[50] Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., & Mann, M., Higher-energy c-trap dissociation for peptide modification analysis. *Nature Methods*, *4*(9), 709–712. doi:10.1038/nmeth1060

[51] Mann, M., & Wilm, M., Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical chemistry*, *66*(24), 4390–4399.

[52] Morris, H. R., Panico, M., Barber, M., Bordoli, R. S., Sedgwick, R. D., & Tyler, A., Fast atom bombardment: A new mass spectrometric method for peptide sequence analysis. *Biochemical and Biophysical Research Communications*, *101*(2), 623–631. doi:10.1016/0006-291X(81)91304-8

[53] Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S., & Hauer, C. R., Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences*, *83*(17), 6233–6237. Retrieved from http://www.pnas.org/content/83/17/6233

[54] Biemann, K., Contributions of mass spectrometry to peptide and protein structure. *Biomedical & environmental mass spectrometry*, *16*(1-12), 99–111.

[55] Biemann, K., & Papayannopoulos, I. A., Amino acid sequencing of proteins. *Accounts of Chemical Research*, *27*(11), 370–378. doi:10.1021/ar00047a008

[56] Horn, D. M., Zubarev, R. A., & McLafferty, F. W., Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proceedings of the National Academy of Sciences*, *97*(19), 10313–10317. doi:10.1073/pnas.97.19.10313

[57] Hill, H. H., Siems, W. F., & St. Louis, R. H., Ion mobility spectrometry. *Analytical Chemistry*, *62*(23), 1201A–1209A. doi:10.1021/ac00222a001

[58] McDaniel, E. W., Martin, D. W., & Barnes, W. S., Drift tube-mass spectrometer for studies of low-energy ion-molecule reactions. *Review of Scientific Instruments*, *33*(1), 2–7. doi:10.1063/1.1717656

[59] McAfee Jr, K., & Edelson, D., Identification and mobility of ions in a townsend discharge by time-resolved mass spectrometry. *Proceedings of the Physical Society*, *81*(2), 382.

[60] Kanu, A. B., Dwivedi, P., Tam, M., Matz, L., & Hill, H. H., Ion mobilitymass spectrometry. *Journal of Mass Spectrometry*, *43*(1), 1–22. doi:10.1002/jms.1383

[61] Pringle, S. D., Giles, K., Wildgoose, J. L., Williams, J. P., Slade, S. E., Thalassinos, K., Bateman, R. H., Bowers, M. T., & Scrivens, J. H., An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. *International Journal of Mass Spectrometry*, *261*(1), 1–12. doi:10.1016/j.ijms.2006.07.021

[62] Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., & Aebersold, R., Evaluation of two-dimensional gel electrophoresis based proteome analysis technology. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(17), 9390–9395. Retrieved from http://www.jstor.org/stable/123473

[63] Hamdan, M., & Righetti, P. G., Modern strategies for protein quantification in proteome analysis: Advantages and limitations. *Mass Spectrometry Reviews*, *21*(4), 287–302. doi:10.1002/mas.10032

[64] Yamashita, M., & Fenn, J. B., Negative ion production with the electrospray ion source. *The Journal of Physical Chemistry*, *88*(20), 4671–4675. doi:10.1021/j150664a046

[65] Whitehouse, C. M., Dreyer, R. N., Yamashita, M., & Fenn, J. B., Electrospray interface for liquid chromatographs and mass spectrometers. *Analytical Chemistry*, *57*(3), 675–679. doi:10.1021/ac00280a023

[66] Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., & Yates, J. R., Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, *17*(7), 676–682. doi:10.1038/10890

[67] Swartz, M. E., & Murphy, B. J., Ultra performance liquid chromatography: Tomorrows HPLC technology today. *LabPlus Int*, *18*(3), 6–9.

[68] Schey, K. L., Anderson, D. M., & Rose, K. L., Spatially-directed protein identification from tissue sections by top-down LC-MS/MS with electron transfer dissociation. *Analytical Chemistry*, *85*(14), 6767–6774. doi:10.1021/ac400832w

[69] Whitelegge, J., Halgand, F., Souda, P., & Zabrouskov, V., Top-down mass spectrometry of integral membrane proteins. *Expert Review of Proteomics*, *3*(6), 585–596. doi:10.1586/14789450.3.6.585

[70] Taverna, S. D., Ueberheide, B. M., Liu, Y., Tackett, A. J., Diaz, R. L., Shabanowitz, J., Chait, B. T., Hunt, D. F., & Allis, C. D., Long-distance combinatorial linkage between methylation and acetylation on histone h3 n termini. *Proceedings of the National Academy of Sciences*, *104*(7), 2086–2091. doi:10.1073/pnas.0610993104

[71] Siuti, N., & Kelleher, N. L., Decoding protein modifications using top-down mass spectrometry. *Nature Methods*, *4*(10), 817–821. doi:10.1038/nmeth1097

[72] Anderson, S., Shotgun DNA sequencing using cloned DNase i-generated fragments. *Nucleic Acids Research*, *9*(13), 3015–3027. doi:10.1093/nar/9.13.3015

[73] Wu, C. C., & MacCoss, M. J., Shotgun proteomics: Tools for the analysis of complex biological systems. *Current opinion in molecular therapeutics*, *4*(3), 242–250.

[74] MacCoss, M. J., McDonald, W. H., Saraf, A., Sadygov, R., Clark, J. M., Tasto, J. J., Gould, K. L., Wolters, D., Washburn, M., Weiss, A., Clark, J. I., & Yates, J. R., 3rd, Shotgun identification of protein modifications from protein complexes and lens tissue. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(12), 7900–7905. doi:10.1073/pnas.122231399

[75] Nagaraj, N., Kulak, N. A., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., & Mann, M., System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top orbitrap. *Molecular & Cellular Proteomics*, *11*(3), M111.013722. doi:10.1074/mcp.M111.013722

[76] McLafferty, F. W., Tandem mass spectrometry. *Science*, *214*(4518), 280–287. doi:10.1126/science.7280693

[77] Washburn, M. P., Wolters, D., & Yates, J. R., Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, *19*(3), 242–247. doi:10.1038/85686

[78] Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., & Gygi, S. P., Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *Journal of Proteome Research*, *2*(1), 43–50. doi:10.1021/pr025556v

[79] Liu, H., Sadygov, R. G., & Yates, J. R., A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry*, *76*(14), 4193–4201. doi:10.1021/ac0498563

[80] Geromanos, S. J., Vissers, J. P. C., Silva, J. C., Dorschel, C. A., Li, G.-Z., Gorenstein, M. V., Bateman, R. H., & Langridge, J. I., The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *PROTEOMICS*, *9*(6), 1683–1695. doi:10.1002/pmic.200800562

[81] Michalski, A., Cox, J., & Mann, M., More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *Journal of Proteome Research*, *10*(4), 1785–1793. doi:10.1021/pr101060v

[82] Panchaud, A., Jung, S., Shaffer, S. A., Aitchison, J. D., & Goodlett, D. R., Faster, quantitative, and accurate precursor acquisition independent from ion count. *Analytical Chemistry*, *83*(6), 2250–2257. doi:10.1021/ac103079q

[83] Kondrat, R. W., McClusky, G. A., & Cooks, R. G., Multiple reaction monitoring in mass spectrometry/mass spectrometry for direct analysis of complex mixtures. *Analytical Chemistry*, *50*(14), 2017–2021. doi:10.1021/ac50036a020

[84] Picotti, P., & Aebersold, R., Selected reaction monitoring-based proteomics: Workflows, potential, pitfalls and future directions. *Nature methods*, *9*(6), 555–566. doi:10.1038/nmeth.2015

[85] Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S., & Coon, J. J., Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Molecular & cellular proteomics: MCP*, *11*(11), 1475–1488. doi:10.1074/mcp.O112.020131

[86] Holman, S. W., Sims, P. F. G., & Eyers, C. E., The use of selected reaction monitoring in quantitative proteomics. *Bioanalysis*, *4*(14), 1763–1786. doi:10.4155/bio.12.126

[87] Kiyonami, R., Schoen, A., Prakash, A., Peterman, S., Zabrouskov, V., Picotti, P., Aebersold, R., Huhmer, A., & Domon, B., Increased selectivity, analytical precision, and throughput in targeted proteomics. *Molecular & Cellular Proteomics*, *10*(2), M110.002931. doi:10.1074/mcp.M110.002931

[88] Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., & Aebersold, R., Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Molecular & cellular proteomics: MCP*, *11*(6), O111.016717. doi:10.1074/mcp.O111.016717

[89] Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A., & Yates, J. R., Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods*, *1*(1), 39–45. doi:10.1038/nmeth705

[90] Panchaud, A., Scherl, A., Shaffer, S. A., Haller, P. D. von, Kulasekara, H. D., Miller, S. I., & Goodlett, D. R., Precursor acquisition independent from ion count: How to dive deeper into the proteomics ocean. *Analytical Chemistry*, *81*(15), 6481–6488. doi:10.1021/ac900888s

[91] Silva, J. C., Denny, R., Dorschel, C. A., Gorenstein, M., Kass, I. J., Li, G.-Z., McKenna, T., Nold, M. J., Richardson, K., Young, P., & Geromanos, S., Quantitative proteomic analysis by accurate mass retention time pairs. *Analytical chemistry*, *77*(7), 2187–2200. doi:10.1021/ac048455k

[92] Geiger, T., Cox, J., & Mann, M., Proteomics on an orbitrap benchtop mass spectrometer using all-ion fragmentation. *Molecular & Cellular Proteomics*, *9*(10), 2252–2261. doi:10.1074/mcp.M110.001537

[93] Plumb, R. S., Johnson, K. A., Rainville, P., Smith, B. W., Wilson, I. D., Castro-Perez, J. M., & Nicholson, J. K., UPLC/MSE; a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Communications in Mass Spectrometry*, *20*(13), 1989–1994. doi:10.1002/rcm.2550

[94] Bern, M., Finney, G., Hoopmann, M. R., Merrihew, G., Toth, M. J., & MacCoss, M. J., Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry. *Analytical Chemistry*, *82*(3), 833–841. doi:10.1021/ac901801b

[95] Wong, J. W., Schwahn, A. B., & Downard, K. M., ETISEQ an algorithm for automated elution time ion sequencing of concurrently fragmented peptides for mass spectrometry-based proteomics. *BMC Bioinformatics*, *10*(1), 244. doi:10.1186/1471-2105-10-244

[96] Li, G.-Z., Vissers, J. P. C., Silva, J. C., Golick, D., Gorenstein, M. V., & Geromanos, S. J., Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. *Proteomics*, *9*(6), 1696–1719. doi:10.1002/pmic.200800564

[97] Blackburn, K., Mbeunkui, F., Mitra, S. K., Mentzel, T., & Goshe, M. B., Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation. *Journal of Proteome Research*, *9*(7), 3621–3637. doi:10.1021/pr100144z

[98] Distler, U., Kuharev, J., Navarro, P., Levin, Y., Schild, H., & Tenzer, S., Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nature Methods*, *11*(2), 167–170. doi:10.1038/nmeth.2767

[99] Orchard, S., Montechi-Palazzi, L., Deutsch, E. W., Binz, P.-A., Jones, A. R., Paton, N., Pizarro, A., Creasy, D. M., Wojcik, J., & Hermjakob, H., Five years of progress in the standardization of proteomics data 4th annual spring workshop of the HUPO-proteomics standards initiative april 2325, 2007 ecole nationale supérieure (ENS), lyon, france. *PROTEOMICS*, *7*(19), 3436–3440. doi:10.1002/pmic.200700658

[100] Pedrioli, P. G. A. et al., A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology*, *22*(11), 1459–1466. doi:10.1038/nbt1031

[101] Deutsch, E., mzML: A single, unifying data format for mass spectrometer output. *PROTEOMICS*, *8*(14), 2776–2777. doi:10.1002/pmic.200890049

[102] Windig, W., Phalp, J. M., & Payne, A. W., A noise and background reduction method for component detection in liquid chromatography/Mass spectrometry. *Analytical Chemistry*, *68*(20), 3602–3606. doi:10.1021/ac960435y

[103] Biller, J. E., & Biemann, K., Reconstructed mass spectra, a novel approach for the utilization of gas chromatographMass spectrometer data. *Analytical Letters*, *7*(7), 515–528. doi:10.1080/00032717408058783

[104] Dromey, R. G., Stefik, M. J., Rindfleisch, T. C., & Duffield, A. M., Extraction of mass spectra free of background and neighboring component contributions from gas chromatography/mass spectrometry data. *Analytical Chemistry*, *48*(9), 1368–1375. doi:10.1021/ac50003a027

[105] Cappadona, S., Nanni, P., Benevento, M., Levander, F., Versura, P., Roda, A., Cerutti, S., & Pattini, L., Improved label-free LC-MS analysis by wavelet-based noise rejection. *BioMed Research International*, *2010*, e131505. doi:10.1155/2010/131505

[106] Ueno, T., Sueyoshi, T., Tanaka, E., Jinkawa, R., Hamada, A., & Takegami, Y., Computer-aided deduction of mass spectra detected on a photographic plate. III. background and sample background (undesired components in a mixture) subtraction. *Shitsuryo Bunseki*, *22*, 109.

[107] Savitzky, A., & Golay, M. J. E., Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627–1639. doi:10.1021/ac60214a047

[108] Zhang, Z., & Marshall, A. G., A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *Journal of the American Society for Mass Spectrometry*, *9*(3), 225–233. doi:10.1016/S1044-0305(97)00284-5

[109] Andreev, V. P., Rejtar, T., Chen, H.-S., Moskovets, E. V., Ivanov, A. R., & Karger, B. L., A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain. *Analytical Chemistry*, *75*(22), 6314–6326. doi:10.1021/ac0301806

[110] Sykes, M. T., & Williamson, J. R., Envelope: Interactive software for modeling and fitting complex isotope distributions. *BMC Bioinformatics*, *9*(1), 446. doi:10.1186/1471-2105-9-446

[111] Leinonen, R., Diez, F. G., Binns, D., Fleischmann, W., Lopez, R., & Apweiler, R., UniProt archive. *Bioinformatics*, *20*(17), 3236–3237. doi:10.1093/bioinformatics/bth191

[112] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., & Apweiler, R., The international protein index: An integrated database for proteomics experiments. *Proteomics*, *4*(7), 1985–1988.

[113] Eng, J. K., McCormack, A. L., & Yates, J. R., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, *5*(11), 976–989. doi:10.1016/1044-0305(94)80016-2

[114] Perkins, D. N., Pappin, D. J. C., Creasy, D. M., & Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS*, *20*(18), 3551–3567. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

[115] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., & Bryant, S. H., Open mass spectrometry search algorithm. *Journal of Proteome Research*, *3*(5), 958–964. doi:10.1021/pr0499491

[116] Craig, R., & Beavis, R. C., TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics*, *20*(9), 1466–1467. doi:10.1093/bioinformatics/bth092

[117] Kim, S., Gupta, N., & Pevzner, P. A., Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. *Journal of Proteome Research*, *7*(8), 3354–3363. doi:10.1021/pr8001244

[118] Corporation, W., Waters: ProteinLynx global SERVER (PLGS). Retrieved from http://www.waters.com/waters/nav.htm?cid=513821

[119] Benjamini, Y., & Hochberg, Y., Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

[120] Weatherly, D. B., Atwood, J. A., Minning, T. A., Cavola, C., Tarleton, R. L., & Orlando, R., A heuristic method for assigning a false-discovery rate for protein identifications from mascot database search results. *Molecular & Cellular Proteomics*, *4*(6), 762–772. doi:10.1074/mcp.M400215-MCP200

[121] Asara, J. M., Christofk, H. R., Freimark, L. M., & Cantley, L. C., A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen. *PROTEOMICS*, *8*(5), 994–999. doi:10.1002/pmic.200700426

[122] Ong, S.-E., & Mann, M., Mass spectrometry-based proteomics turns quantitative. *Nature chemical biology*, *1*(5), 252–262. doi:10.1038/nchembio736

[123] Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., & Kuster, B., Quantitative mass spectrometry in proteomics: A critical review. *Analytical and Bioanalytical Chemistry*, *389*(4), 1017–1031. doi:10.1007/s00216-007-1486-6

[124] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., & Aebersold, R., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, *17*(10), 994–999. doi:10.1038/13690

[125] Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., & Mann, M., Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics: MCP*, *1*(5), 376–386.

[126] Kuhn, E., Wu, J., Karl, J., Liao, H., Zolg, W., & Guild, B., Quantification of c-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards. *Proteomics*, *4*(4), 1175–1186. doi:10.1002/pmic.200300670

[127] Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., & Pappin, D. J., Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics: MCP*, *3*(12), 1154–1169. doi:10.1074/mcp.M400129-MCP200

[128] Allet, N. et al., In vitro and in silico processes to identify differentially expressed proteins. *Proteomics*, *4*(8), 2333–2351. doi:10.1002/pmic.200300840

[129] Zhang, B., VerBerkmoes, N. C., Langston, M. A., Uberbacher, E., Hettich, R. L., & Samatova, N. F., Detecting differential and correlated protein expression in label-free shotgun proteomics. *Journal of Proteome Research*, *5*(11), 2909–2918. doi:10.1021/pr0600273

[130] Rappsilber, J., Ryder, U., Lamond, A. I., & Mann, M., Large-scale proteomic analysis of the human spliceosome. *Genome Research*, *12*(8), 1231–1245. doi:10.1101/gr.473902

[131] Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., & Mann, M., Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & cellular proteomics: MCP*, *4*(9), 1265–1272. doi:10.1074/mcp.M500061-MCP200

[132] Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B., & Aebersold, R., Full dynamic range proteome analysis of s. cerevisiae by targeted proteomics. *Cell*, *138*(4), 795–806. doi:10.1016/j.cell.2009.05.051

[133] Beynon, R. J., Doherty, M. K., Pratt, J. M., & Gaskell, S. J., Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nature Methods*, *2*(8), 587–589. doi:10.1038/nmeth774

[134] Pratt, J. M., Simpson, D. M., Doherty, M. K., Rivers, J., Gaskell, S. J., & Beynon, R. J., Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nature Protocols*, *1*(2), 1029–1043. doi:10.1038/nprot.2006.129

[135] Rappsilber, J., Ishihama, Y., Mittler, G., Mortensen, P., Foster, L., & Mann, M., Approximate relative abundance of proteins within a mixture determined from LC-MS data. *Proceedings of the 51st American Society for Mass Spectrometry Conference on Mass Spectrometry*.

[136] Silva, J. C., Gorenstein, M. V., Li, G.-Z., Vissers, J. P. C., & Geromanos, S. J., Absolute quantification of proteins by LCMSE: A virtue of parallel MS acquisition. *Molecular & Cellular Proteomics: MCP*, *5*(1), 144–156. doi:10.1074/mcp.M500230-MCP200

[137] Forner, F., Foster, L. J., Campanaro, S., Valle, G., & Mann, M., Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Molecular & Cellular Proteomics*, *5*(4), 608–619. doi:10.1074/mcp.M500298-MCP200

[138] Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., & Selbach, M., Global quantification of mammalian gene expression control. *Nature*, *473*(7347), 337–342. doi:10.1038/nature10098

[139] Wilhelm, M. et al., Mass-spectrometry-based draft of the human proteome. *Nature*, *509*(7502), 582–587. doi:10.1038/nature13319

[140] Podwojski, K., Fritsch, A., Chamrad, D. C., Paul, W., Sitek, B., Stühler, K., Mutzel, P., Stephan, C., Meyer, H. E., Urfer, W., Ickstadt, K., & Rahnenführer, J., Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics*, *25*(6), 758–764. doi:10.1093/bioinformatics/btp052

[141] Groot, J. C. W. de, Fiers, M. W. E. J., Ham, R. C. H. J. van, & America, A. H. P., Post alignment clustering procedure for comparative quantitative proteomics LC-MS data. *Proteomics*, *8*(1), 32–36. doi:10.1002/pmic.200700707

[142] Nesvizhskii, A. I., & Aebersold, R., Interpretation of shotgun proteomic data: The protein inference problem. *Molecular & cellular proteomics: MCP*, *4*(10), 1419–1440. doi:10.1074/mcp.R500012-MCP200

[143] Geromanos, S. J., Hughes, C., Ciavarini, S., Vissers, J. P. C., & Langridge, J. I., Using ion purity scores for enhancing quantitative accuracy and precision in complex proteomics samples. *Analytical and Bioanalytical Chemistry*, *404*(4), 1127–1139. doi:10.1007/s00216-012-6197-y

[144] Bond, N. J., Shliaha, P. V., Lilley, K. S., & Gatto, L., Improving qualitative and quantitative performance for MS e-based label-free proteomics. *Journal of proteome research*.

[145] Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M.-Y., Vitek, O., Aebersold, R., & Müller, M., SuperHirn a novel tool for high resolution LC-MS-based peptide/protein profiling. *PROTEOMICS*, *7*(19), 3470–3480. doi:10.1002/pmic.200700057

[146] Sturm, M., Bertsch, A., Gröpl, C., Hildebrandt, A., Hussong, R., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., Zerck, A., Reinert, K., & Kohlbacher, O., OpenMS an open-source software framework for mass spectrometry. *BMC Bioinformatics*, *9*(1), 163. doi:10.1186/1471-2105-9-163

[147] Kohlbacher, O., Reinert, K., Gröpl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O., & Sturm, M., TOPPthe OpenMS proteomics pipeline. *Bioinformatics*, *23*(2), e191–e197.

[148] Weisser, H., Nahnsen, S., Grossmann, J., Nilse, L., Quandt, A., Brauer, H., Sturm, M., Kenar, E., Kohlbacher, O., Aebersold, R., & Malmström, L., An automated pipeline for high-throughput label-free quantitative proteomics. *Journal of Proteome Research*, *12*(4), 1628–1644. doi:10.1021/pr300992u

[149] Navarro, P., Hahlbrock, J., Kuharev, J., Distler, U., & Tenzer, S., Peptide-centric database search engines applied to data independent acquisition UDMSE data. In. Madrid.

[150] Wiśniewski, J. R., Zougman, A., Nagaraj, N., & Mann, M., Universal sample preparation method for proteome analysis. *Nature Methods*, *6*(5), 359–362. doi:10.1038/nmeth.1322

[151] Olsen, J. V., Ong, S.-E., & Mann, M., Trypsin cleaves exclusively c-terminal to arginine and lysine residues. *Molecular & Cellular Proteomics*, *3*(6), 608–614. doi:10.1074/mcp.T400003-MCP200

[152] R Core Team, *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

[153] Kuharev, J., Navarro, P., Distler, U., Jahn, O., & Tenzer, S., In-depth evaluation of software tools for data-independent acquisition based label-free quantification. *PROTEOMICS*. doi:10.1002/pmic.201400396

[154] Bellman, R., On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, *38*(8), 716.

[155] Bellman, R., & Kalaba, R., On adaptive control processes. *Automatic Control, IRE Transactions on*, *4*(2), 1–9.

[156] Itakura, F., Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *23*(1), 67–72. doi:10.1109/TASSP.1975.1162641

[157] Sakoe, H., & Chiba, S., Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *26*(1), 43–49.

[158] Ahmad, I., Suits, F., Hoekman, B., Swertz, M. A., Byelas, H., Dijkstra, M., Hooft, R., Katsubo, D., Breukelen, B. van, Bischoff, R., & Horvatovich, P., A high-throughput processing service for retention time alignment of complex proteomics and metabolomics LC-MS data. *Bioinformatics (Oxford, England)*, *27*(8), 1176–1178. doi:10.1093/bioinformatics/btr094

[159] Chu, S., Keogh, E., Hart, D., & Pazzani, M., Iterative deepening dynamic time warping for time series. In *Proceedings of the 2002 SIAM international conference on data mining*, Proceedings (pp. 195–212). Society for Industrial; Applied Mathematics. Retrieved from http://epubs.siam.org/doi/abs/10.1137/1.9781611972726.12

[160] Salvador, S., & Chan, P., FastDTW: Toward accurate dynamic time. *Warping in Linear Time and Space*.

[161] Hirschberg, D. S., A linear space algorithm for computing maximal common subsequences. *Commun. ACM*, *18*(6), 341–343. doi:10.1145/360825.360861

[162] Florek, K., \Lukaszewicz, J., Perkal, J., Steinhaus, H., & Zubrzycki, S., Sur la liaison et la division des points d'un ensemble fini. In *Colloquium mathematicae* (Vol. 2, pp. 282–285). Institute of Mathematics Polish Academy of Sciences.

[163] Florek, K., Lukaszewicz, J., Perkal, J., Steinhaus, H., & Zubrzycki, S., On wroclaw taxonomy. *Przeglad Antropologiczny*, *17*, 193–211.

[164] Sibson, R., SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, *16*(1), 30–34. doi:10.1093/comjnl/16.1.30

[165] Ester, M., Kriegel, H. P., Sander, J., & Xu, X., A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, & U. Fayyad (Eds.), *Second international conference on knowledge discovery and data mining* (pp. 226–231). Portland, Oregon: AAAI Press.

[166] Cleveland, W. S., Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*(368), pp. 829–836. Retrieved from http://www.jstor.org/stable/2286407

[167] Zhang, R., Barton, A., Brittenden, J., Huang, J. T.-J., & Crowther, D., Evaluation for computational platforms of LC-MS based label-free quantitative proteomics: A global view. *Journal of Proteomics & Bioinformatics*, *03*(09), 260–265. doi:10.4172/jpb.1000149

[168] Shliaha, P. V., Bond, N. J., Gatto, L., & Lilley, K. S., Effects of traveling wave ion mobility separation on data independent acquisition in proteomics studies. *Journal of proteome research*.

[169] Lange, E., Tautenhahn, R., Neumann, S., & Gröpl, C., Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, *9*(1), 375. doi:10.1186/1471-2105-9-375

[170] Tomasi, G., Berg, F. van den, & Andersson, C., Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, *18*(5), 231–241. doi:10.1002/cem.859

[171] Vandenbogaert, M., Li-Thiao-Té, S., Kaltenbach, H.-M., Zhang, R., Aittokallio, T., & Schwikowski, B., Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *PROTEOMICS*, *8*(4), 650–672. doi:10.1002/pmic.200700791

[172] Nederkassel, A. M. van, Xu, C. J., Lancelin, P., Sarraf, M., MacKenzie, D. A., Walton, N. J., Bensaid, F., Lees, M., Martin, G. J., Desmurs, J. R., Massart, D. L., Smeyers-Verbeke, J., & Vander Heyden, Y., Chemometric treatment of vanillin fingerprint chromatograms: Effect of different signal alignments on principal component analysis plots. *Journal of Chromatography A*, 29th international symposium on high performance liquid phase separations and related techniques part II, *1120*(12), 291–298. doi:10.1016/j.chroma.2005.11.134

[173] Nielsen, N.-P. V., Carstensen, J. M., & Smedsgaard, J., Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A*, *805*(12), 17–35. doi:10.1016/S0021-9673(98)00021-1

[174] Christin, C., Smilde, A. K., Hoefsloot, H. C. J., Suits, F., Bischoff, R., & Horvatovich, P. L., Optimized time alignment algorithm for LC-MS data: Correlation optimized warping using component detection algorithm-selected mass chromatograms. *Analytical Chemistry*, *80*(18), 7012–7021. doi:10.1021/ac800920h

[175] Bylund, D., Danielsson, R., Malmquist, G., & Markides, K. E., Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatographymass spectrometry data. *Journal of Chromatography A*, *961*(2), 237–244. doi:10.1016/S0021-9673(02)00588-5

[176] Prince, J. T., & Marcotte, E. M., Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Analytical Chemistry*, *78*(17), 6140–6152. doi:10.1021/ac0605344

[177] Sadygov, R. G., Martin Maroto, F., & Hühmer, A. F. R., ChromAlign: A two-step algorithmic procedure for time alignment of three-dimensional LC-MS chromatographic surfaces. *Analytical Chemistry*, *78*(24), 8207–8217. doi:10.1021/ac060923y

[178] Forrest, A. R., Interactive interpolation and approximation by bézier polynomials. *The Computer Journal*, *15*(1), 71–79.

[179] Schumaker, L., On shape preserving quadratic spline interpolation. *SIAM Journal on Numerical Analysis*, *20*(4), 854–864. doi:10.1137/0720057

[180] MacQueen, J., Some methods for classification and analysis of multivariate observations. In. The Regents of the University of California. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200512992

[181] Lance, G. N., & Williams, W. T., A general theory of classificatory sorting strategies 1. hierarchical systems. *The Computer Journal*, *9*(4), 373–380. doi:10.1093/comjnl/9.4.373

[182] Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., & Le, Q.-T., Sample classification from protein mass spectrometry, by peak probability contrasts. *Bioinformatics*, *20*(17), 3034–3044. doi:10.1093/bioinformatics/bth357

[183] Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J., Optics: Ordering points to identify the clustering structure. In *ACM sigmod record* (Vol. 28, pp. 49–60). ACM.

[184] Achtert, E., Böhm, C., & Kröger, P., DeLi-clu: Boosting robustness, completeness, usability, and efficiency of hierarchical clustering by a closest pair ranking. In *Advances in knowledge discovery and data mining* (pp. 119–128). Springer.

[185] Kuiper, F. K., & Fisher, L., 391: A monte carlo comparison of six clustering procedures. *Biometrics*, 777–783.

[186] Callister, S. J., Barry, R. C., Adkins, J. N., Johnson, E. T., Qian, W.-j., Webb-Robertson, B.-J. M., Smith, R. D., & Lipton, M. S., Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *Journal of Proteome Research*, *5*(2), 277–286. doi:10.1021/pr050300l

[187] America, A. H. P., & Cordewener, J. H. G., Comparative LC-MS: A landscape of peaks and valleys. *PROTEOMICS*, *8*(4), 731–749. doi:10.1002/pmic.200700694

[188] Katajamaa, M., & Orešič, M., Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics*, *6*(1), 179. doi:10.1186/1471-2105-6-179

[189] Silva, J. C., Denny, R., Dorschel, C., Gorenstein, M. V., Li, G.-Z., Richardson, K., Wall, D., & Geromanos, S. J., Simultaneous qualitative and quantitative analysis of the escherichia coli proteome: A sweet tale. *Molecular & cellular proteomics: MCP*, *5*(4), 589–607. doi:10.1074/mcp.M500321-MCP200

[190] Vissers, J. P. C., Langridge, J. I., & Aerts, J. M. F. G., Analysis and quantification of diagnostic serum markers and protein signatures for gaucher disease. *Molecular & Cellular Proteomics*, *6*(5), 755–766. doi:10.1074/mcp.M600303-MCP200

[191] Cox, J., & Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, *26*(12), 1367–1372. doi:10.1038/nbt.1511

[192] Thakur, S. S., Geiger, T., Chatterjee, B., Bandilla, P., Fröhlich, F., Cox, J., & Mann, M., Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Molecular & Cellular Proteomics*, *10*(8), M110.003699. doi:10.1074/mcp.M110.003699

[193] Huang, T., Wang, J., Yu, W., & He, Z., Protein inference: A review. *Briefings in bioinformatics*, bbs004.

[194] Jin, S., Daly, D. S., Springer, D. L., & Miller, J. H., The effects of shared peptides on protein quantitation in label-free proteomics by LC/MS/MS. *Journal of Proteome Research*, *7*(1), 164–169. doi:10.1021/pr0704175

[195] Colaert, N., Gevaert, K., & Martens, L., RIBAR and xRIBAR: Methods for reproducible relative MS/MS-based label-free protein quantification. *Journal of Proteome Research*, *10*(7), 3183–3189. doi:10.1021/pr200219x

[196] Patzig, J., Jahn, O., Tenzer, S., Wichert, S. P., Monasterio-Schrader, P. de, Rosfa, S., Kuharev, J., Yan, K., Bormuth, I., Bremer, J., Aguzzi, A., Orfaniotou, F., Hesse, D., Schwab, M. H., Möbius, W., Nave, K.-A., & Werner, H. B., Quantitative and integrative proteome analysis of peripheral nerve myelin identifies novel myelin proteins and candidate neuropathy loci. *The Journal of Neuroscience*, *31*(45), 16369–16386. doi:10.1523/JNEUROSCI.4016-11.2011

[197] Tenzer, S., Docter, D., Rosfa, S., Wlodarski, A., Kuharev, J., Rekik, A., Knauer, S. K., Bantz, C., Nawroth, T., Bier, C., Sirirattanapan, J., Mann, W., Treuel, L., Zellner, R., Maskos, M., Schild, H., & Stauber, R. H., Nanoparticle size is a critical physicochemical determinant of the human blood plasma corona: A comprehensive quantitative proteomic analysis. *ACS Nano*, *5*(9), 7155–7167. doi:10.1021/nn201950e

[198] Michel, A., Schüler, A., Friedrich, P., Döner, F., Bopp, T., Radsak, M., Hoffmann, M., Relle, M., Distler, U., Kuharev, J., Tenzer, S., Feyerabend, T. B., Rodewald, H.-R., Schild, H., Schmitt, E., Becker, M., & Stassen, M., Mast celldeficient KitW-sh Sash mutant mice display aberrant myelopoiesis leading to the accumulation of splenocytes that act as myeloid-derived suppressor cells. *The Journal of Immunology*, *190*(11), 5534–5544. doi:10.4049/jimmunol.1203355

[199] Tenzer, S., Docter, D., Kuharev, J., Musyanovych, A., Fetz, V., Hecht, R., Schlenk, F., Fischer, D., Kiouptsi, K., Reinhardt, C., Landfester, K., Schild, H., Maskos, M., Knauer, S. K., & Stauber, R. H., Rapid formation of plasma protein corona critically affects nanoparticle pathophysiology. *Nature Nanotechnology*, *8*(10), 772–781. doi:10.1038/nnano.2013.181

[200] Tenzer, S., Moro, A., Kuharev, J., Francis, A. C., Vidalino, L., Provenzani, A., & Macchi, P., Proteome-wide characterization of the RNA-binding protein RALY-interactome using the in vivo-biotinylation-pulldown-quant (iBioPQ) approach. *Journal of Proteome Research*, *12*(6), 2869–2884. doi:10.1021/pr400193j

[201] Distler, U., Schmeisser, M. J., Pelosi, A., Reim, D., Kuharev, J., Weiczner, R., Baumgart, J., Boeckers, T. M., Nitsch, R., Vogt, J., & Tenzer, S., In-depth protein profiling of the postsynaptic density from mouse hippocampus using data-independent acquisition proteomics. *PROTEOMICS*, *14*(21-22), 2607–2613. doi:10.1002/pmic.201300520

[202] Docter, D., Distler, U., Storck, W., Kuharev, J., Wünsch, D., Hahlbrock, A., Knauer, S. K., Tenzer, S., & Stauber, R. H., Quantitative profiling of the protein coronas that form around nanoparticles. *Nature Protocols*, *9*(9), 2030–2044. doi:10.1038/nprot.2014.139

[203] Schick, I., Lorenz, S., Gehrig, D., Tenzer, S., Storck, W., Fischer, K., Strand, D., Laquai, F., & Tremel, W., Inorganic janus particles for biomedical applications. *Beilstein journal of nanotechnology*, *5*(1), 2346–2362.

[204] Ritz, S., Schöttler, S., Kotman, N., Baier, G., Musyanovych, A., Kuharev, J., Landfester, K., Schild, H., Jahn, O., Tenzer, S., & Mailänder, V., The protein corona of nanoparticles: Distinct proteins regulate the cellular uptake. *Biomacromolecules*. doi:10.1021/acs.biomac.5b00108

[205] Nahnsen, S., Bielow, C., Reinert, K., & Kohlbacher, O., Tools for label-free peptide quantification. *Molecular & Cellular Proteomics*, *12*(3), 549–556. doi:10.1074/mcp.R112.025163

[206] Röst, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S. M., Schubert, O. T., Wolski, W., Collins, B. C., Malmström, J., Malmström, L., & Aebersold, R., OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology*, *32*(3), 219–223. doi:10.1038/nbt.2841

[207] Altelaar, A. F. M., Frese, C. K., Preisinger, C., Hennrich, M. L., Schram, A. W., Timmers, H. T. M., Heck, A. J. R., & Mohammed, S., Benchmarking stable isotope labeling based quantitative proteomics. *Journal of Proteomics*, Special issue: New horizons and applications for proteomics [EuPA 2012], *88*, 14–26. doi:10.1016/j.jprot.2012.10.009

[208] Daly, C. E., Ng, L. L., Hakimi, A., Willingale, R., & Jones, D. J. L., Qualitative and quantitative characterization of plasma proteins when incorporating traveling wave ion mobility into a liquid chromatographyMass spectrometry workflow for biomarker discovery: Use of product ion quantitation as an alternative data analysis tool for label free quantitation. *Analytical Chemistry*, *86*(4), 1972–1979. doi:10.1021/ac403901t

[209] Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., & Bähler, J., Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell*, *151*(3), 671–683. doi:10.1016/j.cell.2012.09.019

[210] Noble, W. S., & MacCoss, M. J., Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput Biol*, *8*(1), e1002296. doi:10.1371/journal.pcbi.1002296

[211] Matzke, M. M., Brown, J. N., Gritsenko, M. A., Metz, T. O., Pounds, J. G., Rodland, K. D., Shukla, A. K., Smith, R. D., Waters, K. M., McDermott, J. E., & Webb-Robertson, B.-J., A comparative analysis of computational approaches to relative protein quantification using peptide peak intensities in label-free LC-MS proteomics experiments. *PROTEOMICS*, *13*(3-4), 493–503. doi:10.1002/pmic.201200269

Abkürzungsverzeichnis

AIF	all-ion fragmentation
AMRT	accurate mass retention time pairs
APCI	atmospheric pressure chemical ionization
BSD	Berkeley Software Distribution
CAD	collisionally activated dissociation
CHAPS	3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate
CI	chemical ionization
CID	collision-induced dissociation
CPU	central processing unit
CSV	comma separated value
CURE	clustering using representatives
CV	coefficient of variation
DBSCAN	density-based spatial clustering of applications with noise
DDA	data dependant acquisition
DIA	data-independent acquisition
DIGE	difference gel electrophoresis
DNA	deoxyribonucleic acid
DRTW	dynamic retention time warping
DTT	Dithiothreitol
DTW	dynamic time warping
ECD	electron capture dissociation
EDV	elektronische Datenverarbeitung
emPAI	exponentially modified protein abundance index
EMRT	exact mass retention time pairs
ESI	Elektrospray-Ionisation
ETD	electron-transfer dissociation
FAB	fast atom bombardment
FASP	filter-aided sample preparation
FastDRTW	fast dynamic retention time warping
FastDTW	fast dynamic time warping
FastLinDRTW	fast linear dynamic retention time warping

FDR	false discovery rate
FOSS	free and open source software
FPR	Falsch-Positiv-Rate
FT-ICR	Fourier transform ion cyclotron resonance
FWHM	full width at half maximum
GB	Gigabyte
GPL	general public license
HCD	higher-energy collisional dissociation oder higher-energy C-trap dissociation
$HDMS^E$	high definition MS^E
HPLC	high performance liquid chromatography
HUPO	Human Proteome Organization
IAA	iodoacetic acid
iBAQ	intensity based absolute quantification
ICAT	Isotope-coded affinity tag
IDDTW	iterative deepening dynamic time warping
IDE	integrated development environment
IMS	Ionenmobilitätsspektrometrie
IMS-MS	Ionenmobilitätsspektrometrie-Massenspektrometrie
IPI	international protein index
IRPMD	infrared multiphoton dissociation
iTRAQ	isobaric tags for relative and absolute quantitation
JDK	Java SE Development Kit
JRE	Java SE Runtime Environment
LC	liquid chromatography
LC-MS	liquid chromatography – mass spectrometry
LFQ	label free quantification
LinDRTW	linear dynamic retention time warping
LOWESS	locally weighted scatterplot smoothing
m/z	Masse-zu-Ladung-Verhältnis
MALDI	matrix-assisted laser desorption/ionization
MB	Megabyte
MRM	multiple reaction monitoring

mRNA	messenger RNA
MS	Massenspektrometrie
MS/MS	Tandem-Massenspektrometrie
MSE	mass spectrometry by elevated energy
OPTICS	ordering points to identify the clustering structure
PAcIFIC	precursor acquisition independent from ion count
PAI	protein abundance index
PID	photon-induced dissociation
PLGS	ProteinLynx Global Server
PMF	Peptidmassenfingerprint-Methode
ppm	parts per million
PRM	parallel reaction monitoring
Progenesis QIP	Progenesis QI for Proteomics software
PSI	Proteomics Standards Initiative
PTM	posttranslationale Modifikation
QconCAT	quantification concatamer
QMS	Quadrupol-Massenspektrometer
RNA	Ribonukleinsäure, engl. ribonucleic acid
RSD	relative standard deviation
SID	surface-induced dissociation
SILAC	stable isotope labeling by/with amino acids in cell culture
SPC/ISB	Seattle Proteome Center/Institute for Systems Biology
SQL	structured query language
SRM	selected reaction monitoring / single reaction monitoring
SWATH-MS	sequential window acquisition of all theoretical fragment-ion spectra mass spectrometry
TFA	trifluoroacetic acid
TIC	total ion current
TM	trade mark
TOF	time of flight
TOPP	The OpenMS Proteomics Pipeline
TRIS	Tris(hydroxymethyl)-aminomethan
TWIG	Travelling-Wave-Ion-Guide

TWIMS	travelling wave ion mobility mass spectrometry
UDMS^{E}	ultra definition MS^{E}
UniProt	universal protein database
UPLC	ultra performance liquid chromatography
XIC	extracted ion chromatogram, extracted ion current
XML	extensible markup language
xPAI	extracted ion intensity-based protein abundance

Abbildungsverzeichnis

Abb. 1 Aufbau eines einstufigen und eines hybriden Massenspektrometers 4
Abb. 2 Funktionsweise der Elektrospray-Ionisation 5
Abb. 3 Schematischer Aufbau des Waters Synapt G2-S Massenspektrometers 10
Abb. 4 Peptid-Isotopenmuster 19
Abb. 5 Identifizierbare und Quantifizierbare Anteile eines Proteoms 22
Abb. 6 Ablauf der Post-PLGS-Analyse von MS^E/$HDMS^E$/$UDMS^E$-Daten 46
Abb. 7 Relationale Datenbank zur Speicherung von PLGS-Ergebnissen 48
Abb. 8 Ablauf des Algorithmus Dynamic Retention Time Warping (DRTW) 52
Abb. 9 DRTW-Distanzmatrix und rekurrente Distanzminimierung 53
Abb. 10 Sakoe-Chiba-Bande und Itakura-Parallelogramm 54
Abb. 11 FastDRTW: Iterative Verfeinerung des Warping-Pfads 56
Abb. 12 LinDRTW: Teile und Herrsche 58
Abb. 13 Algorithmen für das Retentionszeitalignment im Effizienzvergleich 62
Abb. 14 Multiples Retentionszeitalignment 64
Abb. 15 Einfluss der LOWESS-Bandbreite auf die Logratio-Fehlerfunktion 69
Abb. 16 Multidimensionale Normalisierung der Feature-Intensitäten 70
Abb. 17 Entscheidungsbeispiele der restriktiven Cluster-Annotation 72
Abb. 18 Anzahl der korrekten Identifikationen als Funktion der FDR 73
Abb. 19 Protein-Homologie-Filter 75
Abb. 20 Peptid- und Proteinidentifikationen nach PLGS- und ISOQuant-Analyse 79
Abb. 21 Reproduzierbarkeit der Identifikation nach PLGS- und ISOQuant-Analyse .. 81
Abb. 22 Varianz der Proteinquantifizierung nach PLGS- und ISOQuant-Analyse 82
Abb. 23 Komposition und LC-MS-Akquisition der Metaproteomproben 85
Abb. 24 Analyseworkflows der getesteten Softwarelösungen 86
Abb. 25 Einfluss des Peptid- und Replikationsfilters auf die Proteinidentifikation 88
Abb. 26 Übereinstimmung der Proteinidentifikation zwischen den Softwarelösungen . 89
Abb. 27 Übereinstimmung der Proteinidentifikation zwischen den Testdatensätzen ... 90
Abb. 28 Präzision der labelfreien Proteinquantifizierung 92
Abb. 29 Richtigkeit der labelfreien Proteinquantifizierung 94
Abb. 30 Exklusive Proteinquantifizierung in verschiedenen Softwarelösungen 95

Tabellenverzeichnis

Tab. 1 Die im Zusammenhang mit dieser Arbeit verwendeten Geräte 42

Tab. 2 Die im Zusammenhang mit dieser Arbeit verwendeten Softwarepakete 43

Tab. 3 Die in ISOQuant genutzten externen Java-Bibliotheken 78

Tab. 4 Peptid- und Proteinidentifikationen nach PLGS- und ISOQuant-Analyse 80

Tab. 5 Varianz der Proteinquantifizierung nach PLGS- und ISOQuant-Analyse 83

Tab. 6 Varianz der Proteinquantifizierung in verschiedenen Softwarelösungen 91

Tab. 7 Einsatz von ISOQuant in verschiedenen Studien 108