Clustermethoden für Massenspektren in proteomweiten statistischen Analysen

Dissertation

zur Erlangung des akademischen Grades Doktor der Naturwissenschaften
an der Fakultät Statistik der Technischen Universität Dortmund

vorgelegt von

Vera Rieder

Dortmund, 5. Februar 2018

Erstgutachter: Prof. Dr. Jörg Rahnenführer

Zweitgutachter: Prof. Dr. Claus Weihs

Inhaltsverzeichnis

1 Einleitung **5**

2 Biologischer Hintergrund und Datensätze **9**

2.1 Tandem-Massenspektrometrie in der Proteomik 9

2.2 Datensätze 14

2.2.1 MS/MS-Läufe 18

2.2.2 Datenbanksuche und De-Novo-Peptidsequenzierung 24

3 Methoden zur Gruppierung von Massenspektren **29**

3.1 Distanzmaße für Tandem-Massenspektren 30

3.2 Proteomweiter Abstand für LC-MS/MS-Läufe 33

3.2.1 DISMS2-Algorithmus 34

3.2.2 Parameteroptimierung von DISMS2 38

3.3 Distanzbasierte Clusterverfahren für Tandem-Massenspektren 40

3.3.1 Clusteralgorithmen 40

3.3.2 Bewertungsmaße der Qualität von Clusterlösungen 58

3.4 Visualisierung von Abständen 63

3.4.1 Dendrogramm 63

3.4.2 Multidimensionale Skalierung 63

3.4.3 Phylogenetischer Baum 66

4 Clusteranalyse von LC-MS/MS-Läufen **69**

4.1 Überblick der LC-MS/MS-Läufe 70

4.2 Optimierung und Visualisierung der Abstände von Läufen 76

4.3 Vergleich der Abstände mit Abständen basierend auf Annotationen . 83

4.4 Einfluss der Generierung und Vorverarbeitung von Spektren 91

4.5 Anmerkungen zu Laufzeiten und Speicherverbrauch 99

5 Clusteranalyse von Tandem-Massenspektren **103**
5.1 Vergleich von distanzbasierten Clusterverfahren 104
5.2 Interpretation einzelner Cluster 116
5.3 Clusterbildung von Massenspektren mehrerer Läufe 122
5.4 Verknüpfung der Spektrenclusterung mit DISMS2 126
5.5 Clusterung zur Peptidzuordnung fehlender Annotationen 129
5.6 Anmerkungen zu Laufzeiten und Speicherverbrauch 133

6 Zusammenfassung und Diskussion **137**

Literaturverzeichnis **145**

Eidesstattliche Erklärung **155**

A Ergänzende Tabellen **157**

B Ergänzende Abbildungen **173**

1 Einleitung

Das faszinierende Great Barrier Reef, das größte Korallenriff der Erde an der Küste vor Australien, bildet Lebensraum für unzählbare Arten. Im Jahr 2016 kam es dort zu einer verheerenden extremen Korallenbleiche. In 84 Riffen, die im nördlichen und zentralen Bereich untersucht wurden, waren im Mittel 35% der Korallen betroffen, die bereits tot oder am Sterben waren (BBC News, 2016). Nicht nur Korallen, sondern auch die Foraminiferen als Kalkbildner leiden unter erhöhten Wassertemperaturen, die eine Folge des Klimawandels sind, und bleichen (Stuhr et al., 2017b,c). Zur Begrenzung der globalen Erwärmung, die auch dem Artenschutz dient, wurde bei der UN-Klimakonferenz bereits im Jahr 2015 das Übereinkommen von Paris verabschiedet.

Der Großteil an Arten ist laut aktuellem Forschungsstand noch nicht einmal bekannt. Laut einer Schätzung von Mora et al. (2011) besiedeln ungefähr 8.7 Millionen (Standardfehler 1.3 Millionen) Eukaryonten die Erde. Davon sind 86% auf dem Land und 91% in den Meeren unbekannt. Zur Artenbestimmung hat sich ausgehend vom Genom das sogenannte DNA-Barcoding etabliert (Hebert et al., 2003). In den Jahren 2014-2017 förderte die Leibniz-Gemeinschaft das Projekt „(Reverse) Proteomics as novel tool for biodiversity research“. Es handelt sich um eine Kooperation zwischen dem Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V., dem Leibniz-Zentrum für Marine Tropenforschung, dem Senckenberg Biodiversität und Klima Forschungszentrum und der Fakultät Statistik der TU Dortmund. Es beschäftigt sich mit einer alternativen Herangehensweise zum DNA-Barcoding, der Analyse der Proteinzusammensetzung von Organismen zur Erforschung der biologischen Vielfalt auf der Erde, die durch den Eingriff des Menschen bedroht ist. Der Fokus liegt auf Arten mit unbekanntem Genom, den amöboiden Protisten, Foraminiferen, und Süßwasserschnecken, *Radix*. Im Rahmen des Projekts wurde ein erster Entwurf der Genomsequenz von *Radix auricularia* erstellt (Schell et al., 2017). Die Genomanalyse der Foraminiferen scheitert an der Gewinnung von Probenmaterial.

Im Jahr 1994 erwähnte Marc Wilkins erstmalig den Begriff Proteom, die Gesamtmenge aller Proteine in einem Organismus. In den letzten Jahrzehnten gewann die Proteomik immer mehr Bedeutung in der biochemischen Forschung. Sie profitiert dabei auch von neuartigen Methoden und technischen Gegebenheiten, die sich in einem fortwährenden Prozess wandeln (Fields, 2001, Method of the Year 2012, 2013). Die Grenzen der Genomik werden beim Vergleich des Phänotyps überschritten. Proben, die unter unterschiedlichen Bedingungen gewonnen wurden, oder Proben von unterschiedlichen Phänotypen können mithilfe von qualitativen und quantitativen Techniken analysiert werden.

Palmblad und Deelder (2012) und Yilmaz et al. (2016) zeigen, dass auch Biodiversitätsstudien von proteomweiten Analysen profitieren. Allein proteomweite Messungen dienten zur Rekonstruktion des eindeutig korrekten phylogenetischen Baumes der Menschenaffen und anderer Primaten (Palmblad und Deelder, 2012). Bisher nicht sequenzierte Spezies können bereits über eine proteombasierte Methode differentiell analysiert werden (Yilmaz et al., 2016) .

Heutzutage basieren die meisten Proteomarbeitsabläufe auf der Massenspektrometrie. Weit verbreitet ist die LC-MS/MS-Methode, also die Verwendung einer Flüssigchromatographie (LC) als Trennmethode in Kombination mit der Tandem-Massenspektrometrie (MS/MS). Daraus resultieren sogenannte MS/MS Spektren, die aus detektierten Intensitäten von vorkommenden Massen bestehen. Jene zu Peptidfragmenten korrespondierenden Massen dienen zur Identifikation von Peptiden und Proteinen, typischerweise durch datenbankabhängige Suchalgorithmen. Nichtsdestotrotz können neuartige Peptide, die in Datenbanken fehlen, mit diesem Ansatz nicht identifiziert werden. De-Novo-Peptidsequenzierungsalgorithmen sind unabhängig von Datenbanksuchen, aber sehr fehleranfällig.

Die Arbeit befasst sich mit Methoden für massenspektrometrische Analysen in der Biodiversitätsforschung. Zwei Themen, die Clusteranalyse von LC-MS/MS-Läufen, die jeweils aus tausenden Tandem-Massenspektren einer Probe bestehen, und die Clusteranalyse von einzelnen Tandem-Massenspektren, werden behandelt.

Das Hauptziel bei der Clusteranalyse von LC-MS/MS-Läufen ist es eine generelle Methode zur Verfügung zu stellen für den Vergleich von Proben, die sich im Phänotyp beeinflusst durch Faktoren wie beispielsweise Lebensraum, Zeit und Klima unterscheiden. DISMS2 (Rieder et al., 2017a) ist eine neu entwickelte Prozedur um Distanzen von mehreren Proteomen zu berechnen. Ohne die Kenntnis einer Peptididentifikation wird auf Grundlage geeigneter Distanzmaße von Massenspektren durch Aggregation der Information von tausenden Massenspektren eine

globale Distanz zwischen paarweisen LC-MS/MS-Läufen berechnet. Palmblad und Deelder (2012) haben zuvor einen einfachen Algorithmus vorgestellt, der zwischen Blutproben differenziert. DISMS2 stellt eine vielseitig einsetzbare, flexible Erweiterung dar. Die Auswahl der höchsten Peaks je Spektrum (`topn`), die Bingröße im Binning (`bin`), die Einschränkung bei dem Vergleich von Spektren auf zeitlich nahe Spektren (`ret`) mit ähnlicher Precursormasse (`prec`) und das Distanzmaß für Tandem-Massenspektren (`dist`) mit einem frei wählbaren Schwellenwert (`cdis`) können über mehrere Parameter variiert werden. Diese Parameter können durch Vorwissen, beispielsweise über Kenntnis der Auflösung des Massenspektrometers, angepasst werden. Alternativ wird ein Vorgehen zur Parameteroptimierung mithilfe eines permutationsbasierten nichtparametrischen Verfahrens zur Varianzanalyse vorgestellt. DISMS2 stellt eine Alternative zum Vergleich von Peptidlisten dar, die durch die Identifikation von Spektren in Datenbanksuchen erstellt werden. Der Identifikationsschritt wird nicht benötigt und es können auch neue Peptide, die in der Datenbank nicht vorhanden sind, mit der neuen Methode berücksichtigt werden. Ein großer Vorteil ist die Anwendung auf Proben von wenig erforschten Arten, beispielsweise Arten der Foraminiferen oder der *Radix*. Auch die Validierung der Ergebnisse über Datenbanksuchen bekannter oder nah verwandter Arten stellt eine Möglichkeit dar. Schließlich können phylogenetische Bäume, die eine Clusteranalyse von mehreren Läufen darstellen, auf Grundlage der DISMS2-Distanzen erstellt werden.

Das Hauptziel bei der Clusteranalyse einzelner Tandem-Massenspektren ist ein bisher in der Literatur fehlender umfassender Vergleich von Algorithmen, die für Tandem-Massenspektren etabliert (CAST, MS-Cluster, PRIDE Cluster), für große Datensätze bekannt (hierarchische Clusteranalyse, DBSCAN, Zusammenhangskomponenten eines Graphen) oder neu (Neighbor Clustering) sind (Rieder et al., 2017b). Der DISMS2-Datensatz (Rieder et al., 2017a) dient zur Evaluierung der sieben Verfahren, für die Parametereinstellungen variiert werden. Die Clusterlösungen werden untereinander und mit Peptidannotationen anhand mehrerer Bewertungsmaße der Qualität von Clusterlösungen verglichen. Unter anderem werden der adjustierte Rand-Index, die Reinheit der Cluster, der Anteil von Spektren in mehrelementigen Clustern, der Spektrenanteil ohne häufigste Annotation im jeweiligen Cluster, der verbleibende Anteil an Annotationen nach der Clusterbildung und der Anteil der Clusterrepräsentanten an allen Clustern verwendet. In einzelnen MS/MS-Läufen und Peptid-Spektralbibliotheken sind redundante Spektren zu finden. Es handelt sich um Spektren, die auf das gleiche Peptid zurückzuführen sind. Eine Clusterbildung der Spektren dient dem Entfernen duplizierter Spektren, sodass nur noch Re-

präsentanten eines Clusters, die sogenannten Konsensusspektren, vorliegen. Dieses Vorgehen ist implizit auch bei der häufig angewandten Peptidintentifikation enthalten. Massenspektren mit gleicher Annotation werden in Cluster gruppiert. Die Clusteranalyse von einzelnen Tandem-Massenspektren dient nicht nur als Vorschritt von DISMS2, sondern verfolgt weitere Ziele, die bei der Analyse unbekannter Proben in der Biodiversitätsforschung hilfreich sind. Als Vorschritt zur Datenbanksuche kann die Spektrenanzahl im Datensatz reduziert werden, sodass sich die Laufzeit der Peptidsuche eines Datensatzes verkürzt. Weitere Anwendungen sind die Entdeckung von Peptiden über Artengrenzen hinweg und eine Qualitätskontrolle. Ein Großteil der Spektren kann nicht annotiert werden, da beispielsweise die Ionensignale von Peptiden mit gleicher Gesamtmasse (Precursormasse) überlagert sind und technisch bedingt die Massengenauigkeit und Auflösung variiert. Der Abgleich unannotierter Spektren in Datenbanken verwandter Arten ist durch eine Clusteranalyse realisierbar. Zur Qualitätskontrolle können in einer Clusteranalyse die Annotationen falsch identifizierter Spektren korrigiert werden. Verwechslungen bei Datenbanksuchen sind häufig durch Hintergrundrauschen bedingt, sodass falsche Peaks gemessen werden.

Die Arbeit ist folgendermaßen gegliedert: In Kapitel 2 sind der biologische Hintergrund, der Einsatz von Tandem-Massenspektren in der Proteomik und massenspektrometrische Datensätze erläutert. In Kapitel 3 werden statistische Methoden zur Gruppierung von Tandem-Massenspektren vorgestellt. Distanzmaße für Tandem-Massenspektren bilden die Grundlage für alle weiteren Analysen. Die Vorgehensweise des neuen DISMS2-Algorithmus wird inklusive einer Parameteroptimierung erläutert. Für den Vergleich von Clusteralgorithmen für Tandem-Massenspektren werden die einzelnen Methoden, besonders das neue Neighbor Clustering, und deren Zusammenhänge sowie hilfreiche Bewertungsmaße erläutert. Zur Interpretation der Abstände von Läufen und Spektren werden Möglichkeiten der Visualisierung, Dendrogramm, multidimensionale Skalierung oder phylogenetischer Baum, vorgestellt. Die Daten werden in Kapitel 4 und 5 analysiert. In Kapitel 4 wird eine Clusteranalyse von Läufen durchgeführt. Kapitel 5 befasst sich anschließend mit der Clusteranalyse einzelner Spektren. Die Arbeit endet in Kapitel 6 mit einer umfangreichen Zusammenfassung und einem Ausblick für zukünftige Forschung.

2 Biologischer Hintergrund und Datensätze

Kapitel 2 beinhaltet einen Überblick über die Proteomforschung und gewährt einen Einblick in das dazu nützliche Tandem-Massenspektrometrie-Verfahren. Ausgehend von biochemischen Grundlagen wird der Bogen bis zur Generierung von massenspektrometrischen Daten gespannt. Abschnitt 2.1 umfasst Basiswissen zur Proteomik und das Prinzip der Tandem-Massenspektrometrie. Die Durchführung einer massenspektrometrischen Analyse wird häufig durch ein Identifikationsverfahren von Peptiden oder Proteinen ergänzt, das auf einer Datenbanksuche oder De-Novo-Sequenzierung basiert. In Abschnitt 2.2 werden die vorliegenden massenspektrometrischen Daten erläutert. Vier Datensätze (DISMS2, Foraminiferen, Biodiversität-Exactive, Biodiversität-Orbitrap) sind im Rahmen des Leibniz-Projekts generiert worden. Ergänzt werden sie durch den Palmblad-Datensatz aus der Literatur (Palmblad und Deelder, 2012). Zunächst werden die sogenannten MS/MS-Läufe, also massenspektrometrische Messungen einzelner Proben, beleuchtet (Abschnitt 2.2.1). Anschließend folgt ein kurzer Einblick in Möglichkeiten zur Peptidannotation von Tandem-Massenspektren (Abschnitt 2.2.2). Insbesondere wird eine Datenbanksuche im Datensatz DISMS2 und die De-Novo-Sequenzierung im Datensatz Foraminiferen beschrieben.

2.1 Tandem-Massenspektrometrie in der Proteomik

Es besteht großes Forschungsinteresse an Proteinen (Alberts et al., 2008; Lottspeich und Engels, 2006, S. 125ff, S. 995ff.), da sie als Enzyme, Transportmittel und zur Strukturbildung in Zellen fungieren. In Anlehnung an den zuvor etablierten Begriff Genom wird die Gesamtmenge aller Proteine in einer Zelle, einem Gewebe oder

einer Probe Proteom genannt. Marc Wilkins erwähnte den Begriff das erste Mal im Jahr 1994. Die Forschung im Bereich der Proteomik ist auf das sich dynamisch verändernde Proteom fokussiert.

Die Proteinsynthese (Alberts et al., 2008, S. 329ff) ist Teil der Genexpression, die den Transfer der genetischen Information von der DNA zur RNA (d. h. Transkription) und schließlich zum Protein (d. h. Translation) umfasst. Viele identische RNA-Kopien der DNA im Transkriptionsschritt ermöglichen die Synthese einer großen Menge an Proteinen im Translationsschritt am Ribosom. Nach der Translation können Veränderungen am Protein, sogenannte posttranslationale Modifikationen (PTMs) (Alberts et al., 2008, S. 189f) auftreten. Zum Beispiel die Phosphorylierung, das Anhängen einer Phosphatgruppe, ist wichtig für die Regulation in Zellen. Aufgrund dieser PTMs steigen die kombinatorischen Möglichkeiten und somit die Komplexität des Proteoms drastisch.

Die Primärstruktur von Proteinen ist eine Kette von Aminosäureresten, die die Aminosäuresequenz bilden. Stabilisiert von Wasserstoffbrückenbindungen zeigen sich lokale Strukturen, z. B. α-Helix, β-Faltblatt und Schleifen (Sekundärstruktur). Die räumliche Anordnung in drei Dimensionen (Tertiärstruktur) wird durch die Faltung bestimmt aufgrund eines hydrophoben Kerns, Salzbrücken, Disulfid- und Wasserstoffbrückenbindungen. Protein-Untereinheiten, bestehend aus vielen Proteinmolekülen, bilden ein Proteinkomplex (quaternäre Struktur).

Eine Aminosäure besteht aus einer Carboxylgruppe, einer Aminogruppe, einem Wasserstoffatom und einer variablen Seitenkette (funktionelle Gruppe R). Eine Peptidbindung verbindet zwei Aminosäuren durch Kondensation unter Abspaltung von Wasser (siehe Abbildung 2.1). Peptide werden aus einer Kette von Aminosäuren gebildet, die Polypeptidketten und Proteine hervorrufen. Zur Notation wird ein Ein-Buchstaben-Code von 20 häufigen Aminosäuren verwendet. Dieser startet gewöhnlich mit der Amino-terminalen (N-terminalen) und endet mit der Carboxy-terminalen (C-terminalen) Aminosäure (Tabelle A.1). Beispielsweise besteht das Peptid PEPTIDE aus sieben Aminosäuren mit einem N-terminalen Prolin (P) und der C-terminalen Glutaminsäure (E).

In den 1990ern musste eine erhebliche Probenmenge aufgereinigt werden zur Peptidsequenzierung via Edman-Abbau. Dieses Verfahren wurde abgelöst von einer Massenspektrometrie-basierten Peptidsequenzierung, die sensitiver und schneller ist. Der technologische Fortschritt ermöglicht Forschern heutzutage den Gebrauch von Massenspektrometrie zur Identifikation von Peptiden in einer Probe (Eidhammer et al., 2007, S. 119ff.). Da unterschiedliche Peptide, sogar von einem einzelnen Pro-

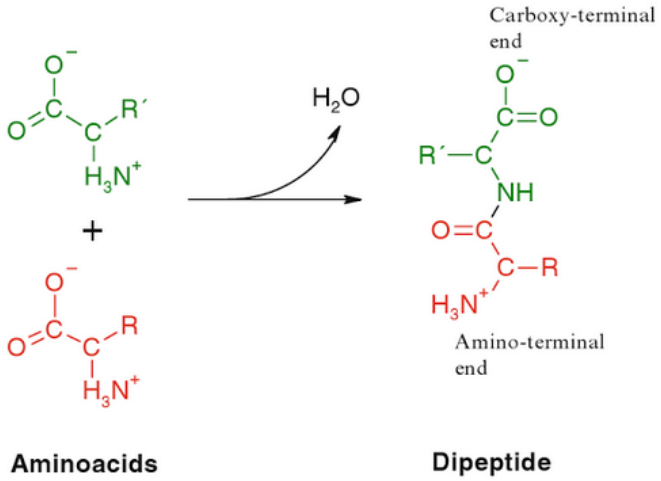

Abbildung 2.1: Zwei Aminosäuren mit Rest R und R' kondensieren unter Abspaltung von Wasser (H_2O) zu einem Dipeptid, einem Peptid mit zwei Aminosäuren (Buxbaum, 2015, Abbildung 2.1).

tein generierte, gleiche Peptidmassen haben können, ist die Information der Peptidmasse unzureichend zur Peptididentifikation. Der Gebrauch von der Peptidsequenz selbst ist ausreichend. Gleiche Peptide kommen selten in zwei Proteinen vor, sodass Proteinidentifikation via Peptidsequenzen möglich ist.

In Abbildung 2.2 ist ein Massenspektrometrie-Experiment in der Proteomik (oben) und das Prinzip ein MS/MS-Spektrum (unten) durch Massenspektrometrie zu erhalten näher dargestellt.

Bei der sogenannten Bottom-up-Analyse, einem Arbeitsablauf in der Proteomik, wird eine Proteinprobe nach mehreren Probenvorbereitungsschritten durch ein Enzym verdaut. Trypsin wird häufig verwendet, da es spezifisch nach Arginin (R) und Lysin (K) spaltet. Peptide mit C-terminalem Arginin oder Lysin werden generiert.

Die Flüssigchromatographie (liquid chromatography, LC) ist verbunden mit einem Massenspektrometer um die überwältigende Komplexität von biologischen Probenverdauen zu minimieren. Normalerweise werden Peptide bei der Umkehrphasen-Chromatographie auf Basis der Hydrophobizität getrennt und eluieren sukzessive von der Säule in einen flüssigen Zustand. ESI (Nobelpreis für Chemie für John B. Fenn im Jahr 2002) ist ein sanftes Ionisierungsverfahren, das geeignet ist für die Analyse von Biomolekülen und als Schnittstelle zwischen LC und Massenspektrometrie verwendet wird.

Ein Massenspektrometer wird in drei Hauptteile aufgeteilt, eine Ionenquelle (z. B. ESI), ein Massenanalysator und ein Detektor gefolgt von Datenverarbeitungselek-

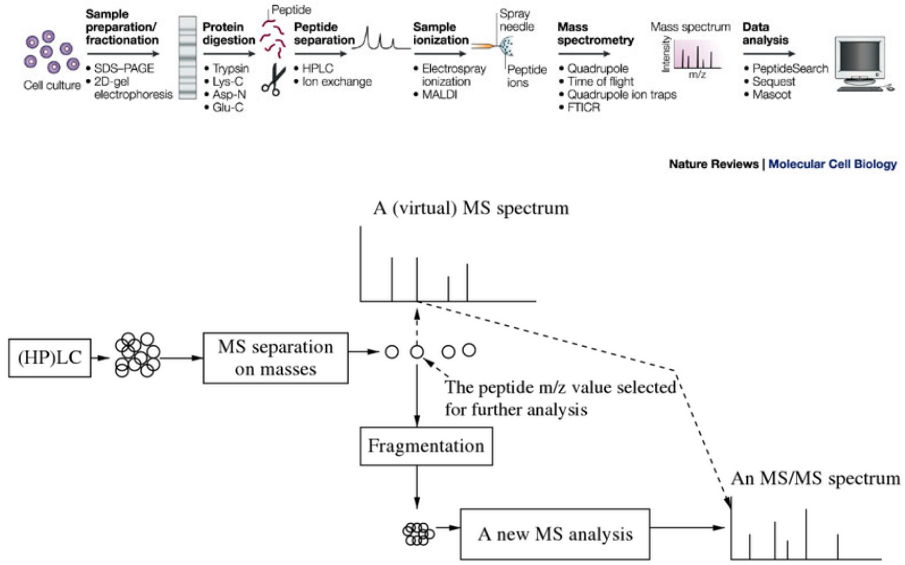

Abbildung 2.2: Massenspektrometrie-Experiment in der Proteomik (Steen und Mann, 2004, Abbildung 1) mit Probenvorbereitung, Proteinverdau, Probenionisierung, Massenspektrometrie und Datenanalyse (oben) und Details zum Prinzip ein MS/MS Spektrum (unten) durch Massenspektrometrie zu erhalten (Eidhammer et al., 2007, Abbildung 8.1).

tronik. Die Massenanalyse wird vom Massenanalysator (unter Vakuum) durchgeführt, sodass Peptidionen nach ihren Masse-zu-Ladung-Verhältnissen (m/z) getrennt werden. Um Peptidsequenzinformationen zu erhalten, wird Tandem-MS (MS/MS) durchgeführt. Dazu werden Peptid-/Precursorionen mit spezifischem m/z-Wert isoliert und anschließend getrennt, um Fragment-/Produktionen zu generieren.

Der Begriff Tandem-Massenspektrometrie stammt von zwei Massenanalysatoren, die hintereinander, in Tandem-Anordnung, geschaltet sind. Die ionisierten Peptide werden aufgrund ihres Masse-zu-Ladung-Verhältnisses (m/z) gefiltert. Unterschiedliche Typen von Ionenfallen werden als selektive Massenfilter verwendet. Tandem-Massenspektrometrie kann im Tandem im Raum (z. B. Triple-Quadrupol) oder im Tandem in der Zeit (z. B. Ionenfallen) ausgeführt werden (Johnson et al., 1990). Beispielsweise stellt Thermo Fisher Scientific Hybridgeräte her, unter anderem Q Exactive mit Quadrupol-Ionenfalle, Orbitrap Elite mit einer Orbitrap und Fusion mit einer Kombination aus Quadrupol-, linearer Ionenfalle und Orbitrap. Schließlich wird die Intensität von Ionen bei unterschiedlichen m/z-Werten detektiert. Ein MS-Spektrum besteht aus Intensitäten von Ionen in einem vorgegebenen m/z Bereich.

Zuerst wird ein bestimmter m/z Bereich durch Massenfilter ausgewählt. Normalerweise wird das MS1 Spektrum generiert, aber ein virtuelles MS1 Spektrum reicht aus, da die Intensitäten nicht benötigt werden. Die selektierten Peptidionen, die

Precursor- oder Parentionen genannt werden, werden fragmentiert und anschließend in einem zweiten Massenspektrometer analysiert. Ein MS/MS Spektrum bezieht sich auf Intensitäten der Masse-zu-Ladung von Fragmentionen.

In Abbildung 2.3 ist das Prinzip der Fragmentierung demonstriert. Kollisionsinduzierte Dissoziation (Collision induced dissociation, CID) wird häufig verwendet um Ionen zu fragmentieren. Die Ionen kollidieren mit neutralen Molekülen, z. B. Helium, mit hoher kinetischer Energie, die eine Dissoziation von Peptidbrücken in die Formation von speziellen Ionen vom b- und y-Typ bewirkt. Unterschiedliche Arten von Peptidfragmenten, z. B. interne Fragmente oder Immoniumfragmente, werden gebildet und Wasserverluste, Ammoniumverluste und Seitenkettenfragmente treten je nach Art der Fragmentierung auf.

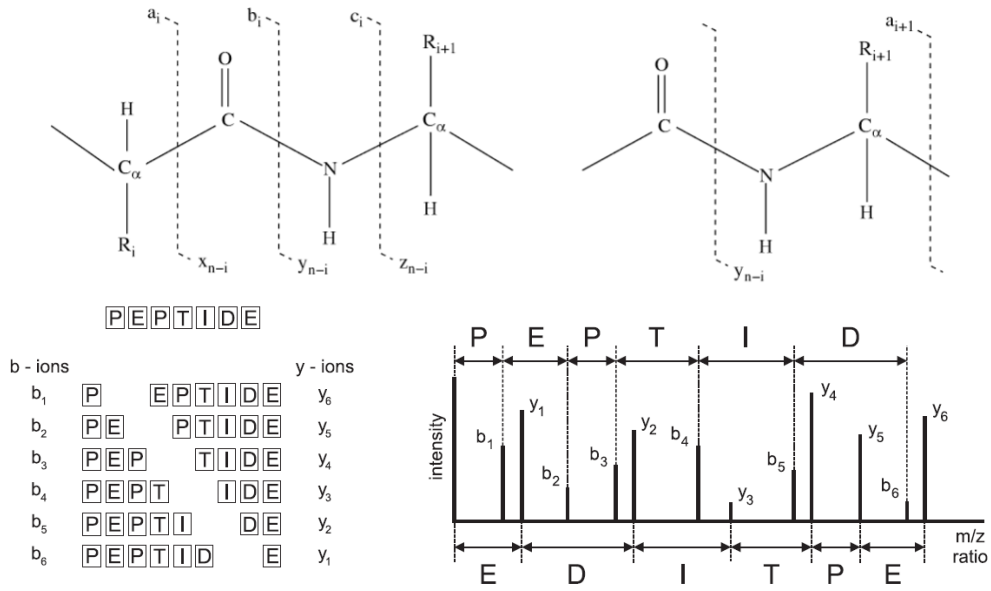

Abbildung 2.3: Roepstorff-Fohlmann-Biemann-Nomenklatur (Roepstorff und Fohlman, 1984; Biemann, 1992) für unterschiedliche Rückgratfragmente (oben links) und die Grundlage für ein Immoniumion (oben rechts) (Eidhammer et al., 2007, Abbildung 8.2). Das Prinzip der Fragmentierung (unten) von Peptid PEPTIDE mit $n = 7$ Resten in b- und y-Ionen ist dargestellt (Novak, 2013, Abbildung 2.6). Der Index i der Ionen bezieht sich auf den i-ten Rest des Peptids.

Aus einem Precursor, der mehr als eine Ladung trägt, werden Fragmentionen generiert. Falls der Precursor einfach geladen ist, werden ein einfach geladenes Schwesterfragment und ein neutrales Schwesterfragment gebildet. Letzteres wird neutraler Wasserverlust genannt, das im MS/MS Spektrum verborgen ist. Die Fragmentation entlang des Peptidrückgrats ergibt Rückgrat-Fragmente, die nach der Roepstorff-

Fohlmann-Biemann-Nomenklatur (Roepstorff und Fohlman, 1984; Biemann, 1992) mit Buchstaben a, b, und c für N-terminale Fragmente und mit x, y und z für C-terminale Fragmente benannt sind (Abbildung 2.3, oben links). Beispielsweise fragmentiert Peptid PEPTIDE in zwölf Ionen, jeweils sechs b- und y-Ionen. Zur Notation werden die Ionen nummeriert durch die Anzahl an Aminosäuren, z. B. b_3-Ion für das Fragment PEP. Die Interpretation von MS/MS Spektren durch Kenntnisse über den Zusammenhang zwischen der Masse von Fragmenttypen und der Summe der Rückstandsmassen (residue mass) der Aminosäuren von dem Fragmention kann dazu beitragen, spezifische Peptide in einer Probe zu identifizieren. Die monoisotopischen Massen von 20 häufigen Aminosäuren sind im Anhang aufgeführt (Tabelle A.1).

Ein Rückgratfragment mit Wasserverlust (Wassermolekül H_2O) oder Ammoniakverlust (Ammoniakmolekül NH_3) verliert eine Masse von 18.001 Da oder 17.027 Da. Kombinationen von b- und y-Typ Fragmenten, sogenannte interne Fragmente, entstehen durch doppelte Rückgratfragmentation. Immoniumfragmente sind spezielle interne Fragmente von a- und y-Ionen, die aus einer einzelnen Seitenkette bestehen. Daher ist die Masse dieser Ionen die Rückstandsmasse einer Aminosäure ohne Atome C und O (-28 Da). Eine zusätzliche Fragmentation der Seitenkette von einem Rückgrat wird Seitenkettenfragmentation genannt.

Die Datenanalyse von Massenspektren wird in den meisten Fällen durch Datenbanksuche, z. B. mithilfe von Mascot (Perkins et al., 1999) oder SEQUEST (Eng et al., 1994), durchgeführt, aber in den letzten Jahren wird die De-Novo-Peptidsequenzierung immer wichtiger. Zahlreiche Algorithmen, z. B. PEAKS (Ma et al., 2003), PepNovo (Frank und Pevzner, 2005) oder Novor (Ma, 2015) wurden entwickelt. Die Vorverarbeitung der Daten spielt eine wichtige Rolle für eine gute Identifizierung. Besonders Deisotoping, d. h. die Transformation der Hüllkurve eines Isotops in einen Peak, wird kombiniert mit der Vorhersage des Ladungszustands (Eidhammer et al., 2007, S. 135).

2.2 Datensätze

Die vorliegenden Datensätze beinhalten Ergebnisse von Tandem-Massenspektrometrie-Experimenten, sogenannten MS/MS-Läufen. Jeder MS/MS-Lauf ist eine Menge von tausenden MS/MS Spektren, die das Proteom einer spezifischen Probe, ein proteolytischer (tryptischer) Verdau, repräsentieren. Zusätzlich wurden teilweise eine

Datenanalyse, Datenbanksuche oder De-Novo-Peptidsequenzierung erstellt um Peptidannotationen von MS/MS Spektren zu erstellen.

Insgesamt fünf Datensätze, die sich in mehreren Punkten unterscheiden, werden untersucht (Tabelle 2.1). Das ISAS hat Massenspektrometrie-Instrumente von Thermo Fisher Scientific verwendet. Q Exactive, Orbitrap Elite und Fusion (Kapitel 2.1) sind aufeinanderfolgende Generationen von Geräten. Daher hängt die Qualität der Spektren von den verwendeten Geräten ab. Im Gegensatz zum ISAS haben Palmblad und Deelder (2012) (Palmblad-Datensatz) amazon speed ETD verwendet, ein Gerät von Bruker Corporation. Die Anzahl der MS/MS-Läufe N (27 oder 30) in jedem Datensatz ist begrenzt, da die Läufe nacheinander gemessen werden und jeder Lauf in der Regel zwischen 90 und 120 Minuten dauert. Tausende MS/MS Spektren werden je Lauf gemessen. Die Anzahl n der Spektren hängt ab von der Dauer, der Probenmenge, der Komplexität der Probe und der Präzision (engl. accuracy) der Messung. Die Datensätze DISMS2, Foraminiferen, Biodiversität-Exactive und Biodiversität-Orbitrap beinhalten alle gemessenen MS/MS-Spektren. Der Biodiversität-Exactive-Datensatz beinhaltet die meisten Spektren, gefolgt vom DISMS2- und Biodiversität-Orbitrap-Datensatz. Der Foraminiferen-Datensatz enthält nur wenige Spektren. Der Palmblad-Datensatz umfasst je Lauf nur 2000 Spektren, da Palmblad und Deelder (2012) eine Vorauswahl getroffen haben. Sie haben 2000 Spektren je Lauf mit der höchsten Gesamtionenintensität (total ion signal) ausgewählt.

Tabelle 2.1: Überblick über alle Datensätze. Tandem-Massenspektrometrie-Geräte, Anzahl N von MS/MS-Läufen, Anzahl n von Spektren und Datenformat (Thermo RAW oder mzXML).

Datensatz	MS/MS-Läufe			
	Gerät	N	n	Format
DISMS2	Q Exactive	27	30012 - 40236	RAW
Foraminiferen	Fusion	30	18054 - 31388	RAW
Biodiversität-Exactive	Q Exactive	30	37775 - 49600	RAW
Biodiversität-Orbitrap	Orbitrap Elite	30	31079 - 36418	RAW
Palmblad	amaZon speed ETD	27	2000 - 2000	mzXML

Deutsch (2012) hat eine Übersicht über die Vielzahl an Datenformaten erstellt, die bei einer massenspektrometrischen Analyse in der Proteomik verwendet werden. Ausgabedateien, die die Massenspektren von massenspektrometrischen Analysen beinhalten, werden in drei Kategorien eingeteilt. Herstellerspezifische Formate können in komplexe offene Formate und einfache Textformate konvertiert werden.

Die Human Proteome Organization Proteomics Standards Initiative (HUPO PSI) ist eine Gruppe, die Massenspektrometerhersteller, Softwareanbieter, Zeitschriftenredakteure, wissenschaftliche Softwareentwickler und Anwender vereint, um Standardformate zu etablieren. 2003 begann die Initiative offene Formate zu entwickeln. Zuerst wurde das mzXML-Format (Pedrioli et al., 2004) vom Institute for Systems Biology (ISB) vorgestellt, gefolgt vom mzData-Format von HUPO PSI. 2009 einigten sich die zwei Parteien auf das mzML-Format (Martens et al., 2011), das die besten Eigenschaften von mzXML und mzData vereint. Dennoch besteht kein Konsens, sodass auch weiterhin mzXML und mzData verwendet werden.

Die Massenspektren werden ursprünglich im Profilmodus (engl. profile mode, continuous) gespeichert, der die Peakform darstellt. Der Profilmodus wird dann häufig in Peaklisten (engl. peak lists, centroided, peak picked) überführt, die nur noch die ausgewählten detektierten Peaks enthalten. Im herstellerspezifischen Format von Thermo Fisher Scientific, dem RAW-Format, können auch beide Formen vorkommen.

Simple Textformate werden bei Datenbanksuchen zur einfachen und sicheren Übertragung von Massenspektren verwendet. Das gebräuchlichste Format ist das Mascot Generic Format (MGF). Es wurde von Matrix Science (London, Vereinigtes Königreich), den Herausgebern der Datenbanksuche Mascot (siehe Kapitel 2.2.2) entwickelt. Die Zusatzinformationen zu den Spektren gehen verloren, sodass lediglich die Massenspektren als Paar von m/z-Werten und Intensitäten gespeichert sind.

Mittlerweile wurde eine Reihe an Konvertierungstools entwickelt, sodass bei Bedarf die Datenformate ineinander überführt werden. Das frei verfügbare Tool MSConvertGUI (Chambers et al., 2012) von Proteowizard unterstützt unter anderem mzML, mzXML, MGF und das RAW-Format von Thermo Scientific Fisher. Im mzXML-Format können Daten mithilfe der statistischen Software R eingelesen (Gibb, 2015, Paket readmzXMLData) und analysiert werden.

Sequenzdateien werden bei der Datenbanksuche beispielsweise mit Mascot oder X!Tandem benötigt. Das einfache FASTA-Format beinhaltet die Sequenzen aus Proteindatenbanken, die zur Identifikation von Peptiden und Proteinen in massenspektrometrischen Daten verwendet werden.

Die Datensätze DISMS2, Foraminiferen, Biodiversität-Exactive und Biodiversität-Orbitrap sind Ergebnisse von „(Reverse) Proteomics as novel tool for biodiversity research“ (SAW-2014-ISAS-2-D), gefördert von dem Leibniz-Wettbewerb. Der DISMS2-Datensatz (Rieder et al., 2017a) ist mit dem Datensatzbezeichner PXD004824 im ProteomeXchange Consortium hinterlegt über das PRIDE Partner

Repository und verfügbar über ProteomeXchange. Die Proben des Foramiferen-Datensatzes wurden gesammelt und vorbereitet vom Leibniz-Zentrum für Marine Tropenforschung (ZMT). Vom Senckenberg Biodiversität und Klima Forschungszentrum (SBiK-F) wurden die Proben des Biodiversität-Exactive- und Biodiversität-Orbitrap-Datensatzes bereitgestellt. Anschließende Massenspektrometrie-Experimente wurden am Leibniz-Institut für Analytische Wissenschaften (ISAS) in Dortmund durchgeführt. Der Palmblad-Datensatz (Palmblad und Deelder, 2012) ist erhältlich über PRIDE (PRD000375). Tryptische Verdauungen von Blutseren von vier Menschenaffen und zwei anderen Primaten wurden analysiert, sodass sich ein Datensatz mit 27 MS/MS-Läufen ergibt.

Das PRIDE Consortium fördert den Austausch von Proteomikdatensätzen, um eine engere Zusammenarbeit von Wissenschaftlern zu ermöglichen. Die Sammlung der hier untersuchten Datensätze könnte erweitert werden. Zum Beispiel beinhalten PXD002193 (Dammeier et al., 2016) und PXD003625 (Yilmaz et al., 2016) zwei interessante Anwendungen in der Forensik und im Vergleich von zwei parasitären Bandwürmern.

Dammeier et al. (2016) präsentieren eine Methode, um verschiedene Rinderorgane (Niere, Lunge, Leber, Muskel und Herz) zu unterscheiden. In einem Experiment wurden die Gewebe mit einem Projektil durchquert und das biologische Material von der Projektiloberfläche wurde untersucht. Yilmaz et al. (2016) vergleichen zwei parasitäre Bandwürmer, *Taenia solium* und *Taenia hydatigena*. Eine Vorgehensweise zur differentiellen Analyse basierend auf Proteomik für bisher nicht sequenzierte Arten wird vorgestellt. Das vorgestellte Verfahren ermöglicht eine Unterscheidung zwischen einem für Menschen schädlichen Bandwurm, *Taenia solium*, und einem für Menschen unschädlichen Bandwurm, *Taenia hydatigena*.

Die Verwendung eines großen Benchmark-Datensatzes wäre optimal für die Evaluierung. Es bedarf jedoch einem erheblichen Zeit- und Ressourcenaufwand zur Analyse (siehe Kapitel 4.5 und 5.6). Zolg et al. (2017) haben bereits über 330 000 synthetische tryptische Peptide analysiert, die alle kanonischen humanen Genprodukte repräsentieren. Die Daten sind frei verfügbar und sollen in den kommenden Jahren auf über eine Millionen Peptide erweitert werden.

Wie bereits erörtert unterscheiden sich die Datensätze durch Probentypen, die untersuchten Arten und die Probenvorbereitung. In dem folgenden Abschnitt werden Charakteristika der MS/MS-Läufe je Datensatz im Detail erläutert. Anschließend folgt eine Beschreibung der Datenbanksuche der DISMS2-Daten und der De-Novo-Peptidsequenzierung der Foraminiferen-Daten.

2.2.1 MS/MS-Läufe

Ein MS/MS-Lauf umfasst die LC-MS/MS-Analyse einer Probe, deren Peptidzusammensetzung bestimmt werden soll. Die Daten bestehen aus einer Liste aller gemessenen Massenspektren, MS- und MS/MS-Spektren, die in zeitlicher Abfolge sortiert ist. Jedes Spektrum wird durch zwei Vektoren beschrieben, den gemessenen Signalstärken (Intensitäten) und der Masse ausgewählter Precursor- und Produktionen (siehe Kapitel 3.1). Zusätzlich enthält jedes Listenobjekt eine Vielzahl an Informationen, die bei der Datenanalyse hilfreich sind. Die Variablen Retentionszeit, Precursormasse und Precursorladung spielen im weiteren Verlauf eine tragende Rolle.

Peptide eluieren von der HPLC-Säule in Abhängigkeit bestimmter physikochemischer Eigenschaften. Der Zeitpunkt, zu dem ein Peptid sich von der Säule löst, wird Retentionszeit genannt. Sie ist in der Einheit Sekunden angegeben. Aus einem Precursorion können mehrere MS/MS-Spektren generiert werden. Die Information der Ladung und Masse des zugehörigen Precursors ist für jedes MS/MS-Spektrum verfügbar. Die Precursorladung ist eine natürliche Zahl und das Masse-zu-Ladung-Verhältnis (m/z) des Precursors eine positive reelle Zahl. Die Messung der MS/MS-Läufe, die als RAW- oder mzXML-Format vorliegen, wird im Folgenden näher erläutert.

Datensatz DISMS2

In Rieder et al. (2017a) wurde der DISMS2-Algorithmus eingeführt und anhand von 27 MS/MS-Läufen validiert. Ein Überblick der untersuchten Arten befindet sich in Tabelle 2.2 (siehe auch Tabelle A.2). Je drei technische Replikate von neun unterschiedlichen Arten wurden in permutierter Reihenfolge gemessen. Im SAW-Projekt der Leibniz-Institute ist seitens des ZMT Bremen ein Fokus auf Foraminiferen und seitens des SBiK Frankfurt ein Fokus auf *Radix* gelegt. Daher wurden zum einen *A. gibbosa* und *A. lessonii* der Foraminiferen-Gattung *Amphistegina* sowie MOTU (engl. Molecular taxonomic unit) 2 und MOTU 4 der Süßwasserschnecken *Radix* analysiert. Zusätzlich wurden stammesgeschichtlich sehr unterschiedliche Proben aus Laborpopulationen, Fadenwurm, Fruchtfliege, Mensch, Maus und Hefe hinzugefügt, für die es bewährte Proteindatenbanken gibt. Bei den Proben von Maus und Mensch handelt es sich um Zelllinien. HeLa-Zellen sind eine Zelllinie menschlicher Epithelzellen eines Zervixkarzinoms und C2C12 ist eine unsterbliche Maus-Myoblasten-Zelllinie (Muskelzellen). Detaillierte Informationen zu der Herkunft der

Proben sind in Rieder et al. (2017a, Zusatzdatei 1) enthalten. Die Datenbanksuche der Laborpopulationen (C, D, H, M, Y) ist in Kapitel 2.2.2 beschrieben.

Tabelle 2.2: Abkürzungen der untersuchten Arten und Nummerierung der Messreihenfolge der Proben im DISMS2-Datensatz.

Abkürzung	Art	Nummerierung
C1, C2, C3	Fadenwurm	7, 18, 23
D1, D2, D3	Fruchtfliege	5, 11, 27
H1, H2, H3	Mensch	1, 12, 25
M1, M2, M3	Maus	9, 16, 20
Y1, Y2, Y3	Hefe	3, 10, 26
Ag1, Ag2, Ag3	*A. gibbosa*	8, 14, 24
Al1, Al2, Al3	*A. lessonii*	4, 17, 21
R21, R22, R23	MOTU2	6, 13, 19
R41, R42, R43	MOTU4	2, 15, 22

In Abbildung 2.4 sind Fotografien von Vertretern der Süßwasserschnecken und Foraminiferen zu sehen, die eine Hauptrolle im Leibniz-Projekt „(Reverse) Proteomics as novel tool for biodiversity research“ spielen. Im Foraminiferen-Datensatz sind weitere Messungen von Foramiferen zu finden. Die Messungen in den Datensätzen Biodiversität-Exactive und Biodiversität-Orbitrap stammen größtenteils von *Radix*-Proben.

Abbildung 2.4: Fotografie einer Süßwasserschnecke (links, *Radix auricularia*, Schell et al. 2017, Abbildung 2, Bild von Markus Pfenninger) und einer Foraminifere (rechts, *Amphistegina gibbosa*, Stuhr et al. 2017c, Abbildung 2).

Mit einem Massenspektrometer von Thermo Fisher Scientific, der Q Exactive, wurden jeweils 2 μL je tryptischem Verdau analysiert. Für die Zelllinien (H1, H2, H3,

M1, M2, M3) und den Hefe-Stamm W303 wurde eine Lösung aus etwa 1 mg Zellen und für die Fadenwürmer 100 μL Lösung verwendet. Für die Fruchtfliegen wurden je 15 Tiere verwendet. Die *Radix*-Proben stammen von Inzuchtlinien aus Aquarien des SBiK Frankfurt. Nur aus einem Teil der Schnecken, dem Fuß, wurde Peptidmaterial gewonnen. Die *Amphistegina*-Proben sind Wildfänge aus Florida (Ag1, Ag2, Ag3) und Sansibar (Al1, Al2, Al3). Je 10 komplette Holobionten, also Foraminiferen-Wirt und endosymbiotische Mikroalgen, wurden vor der Messung noch einige Zeit am ZMT Bremen kultiviert.

Datensatz Foraminiferen

Der bisher unveröffentlichte Foraminiferen-Datensatz beinhaltet verschiedene Arten des Stamms der Foraminiferen der nicht näher verwandten Gattungen *Amphistegina* und *Marginopora*. Insgesamt werden vier verschiedene Arten untersucht, *Marginopora vertebralis* und die nah verwandten *A. gibbosa*, *A. lessonii* und *A. lobifera* der Gattung *Amphistegina*. Sie wurden zwischen 2014 und 2016 in Florida, Eilat und Sansibar gefangen. Die meisten Proben wurden einige Zeit vor der Messung am ZMT Bremen kultiviert. Art, Herkunft und Fangjahr der einzelnen untersuchten Proben sind zusammengefasst in Tabelle 2.3. Je drei biologische Replikate je Fang wurden direkt nacheinander in permutierter Reihenfolge gemessen. Kapitel 2.2.2 beinhaltet die Beschreibung der De-Novo-Peptidsequenzierung aller MS/MS-Läufe mit der in Blank-Landeshammer et al. (2017) beschriebenen Vorgehensweise.

Geologen nutzen fossile Foraminiferen aus einer antiken sauerstoffarmen Umwelt zum Auffinden von Erdölvorkommen (Gupta, 2002, S. 201). Gupta (2002) hat ein umfassendes Werk zu neuzeitlichen Foraminiferen herausgegeben. Neuzeitliche Foraminiferen spielen eine Rolle bei der globalen Erwärmung.

Weniger als ein Zehntel aller Foraminiferen leben in Endosymbiose mit Algen (Gupta, 2002, S. 123ff.). *Amphistegina* und *Marginopora* sind Teil dieser Foraminiferen. Vorteile der Symbiose mit den Algen ist die gewonnene Energie aus der Photosynthese, eine Erhöhung der Kalkbildung und die Aufnahme von Wirtsmetaboliten durch symbiotische Algen.

Foraminiferen könnten zur Prognose der Folgen des Klimawandels eingesetzt werden (Gupta, 2002, S. 137ff.). Die globale Erwärmung wird beeinflusst durch Treibhausgase. Auch die Meereschemie ist betroffen, da sie vom Kohlendioxid in der Atmosphäre beeinflusst wird. Für viele in Symbiose lebende Foraminiferen könnte die Erwärmung von Vorteil sein, da sich ihr Lebensraum erweitern würde. Schädliche UVB-Strahlen dringen zumindest einige Meter unter die Meeresoberfläche ein.

Tabelle 2.3: Abkürzungen, Herkunft und Fangjahr der untersuchten Arten der Gattungen *Amphistegina* und *Marginopora* und Nummerierung (Nr.) der Messreihenfolge der Proben im Foraminiferen-Datensatz.

Abkürzung	Art	Herkunft	Jahr	Nr.
AgiF141, AgiF143, AgiF144	*A. gibbosa*	Florida	2014	11, 24, 7
AgiF151, AgiF152, AgiF153	*A. gibbosa*	Florida	2015	19, 26, 29
AleE*161, AleE*162, AleE*163	*A. lessonii*	Eilat*	2016	3, 25, 8
AleE161, AleE162, AleE164	*A. lessonii*	Eilat	2016	1, 18, 9
AleZ141, AleZ142, AleZ143	*A. lessonii*	Sansibar	2014	2, 12, 16
AleZ151, AleZ152, AleZ154	*A. lessonii*	Sansibar	2015	14, 20, 4
AloE*161, AloE*162, AloE*163	*A. lobifera*	Eilat*	2016	13, 28, 21
AloE162, AloE163, AloE164	*A. lobifera*	Eilat	2016	6, 22, 15
AloZ152, AloZ153, AloZ154	*A. lobifera*	Sansibar	2015	27, 30, 10
MveZ141, MveZ142, MveZ143	*M. vertebralis*	Sansibar	2014	17, 23, 5

*direkt gemessen, nicht kultiviert.

Foraminiferen, die Gemeinschaften in subtropischem Flachwasser bilden, werden geschädigt. Insbesondere wird die Photosynthese und Proteinbiosynthese beeinflusst. Die Photosynthese spielt eine Rolle bei der Kalkbildung und beeinflusst infolgedessen den Rückgang der Korallenriffe und den globalen Kohlenstoffhaushalt.

Größere, neuzeitliche Foraminiferen sind Hauptproduzent des globalen Riffkarbonats (Gupta, 2002, S. 156ff.). Die Foraminiferen finden Schutz und Nährstoffe auf oder in der schwammigen Algendecke des Riffs. Foraminiferen, die in Symbiose mit Algen leben, sind angepasst an flaches, lichtdurchflutetes, nährstoffarmes Wasser von Korallenriffen und Riffhängen in tropischen Meeren. Die Gattung *Amphistegina* bildet die am weitesten verbreitete Gattung in Korallenriffen, in anderen flachen tropischen Gewässern oder auf hartem Untergrund. Sie zählen zu den größeren, am Meeresboden lebenden Foraminiferen (engl. larger benthic foraminifera). Sie finden auch Lebensraum im Seegras in der Nähe von Küsten und können sich an Brackwasser oder hoch salzhaltiges Wasser anpassen. Die zunehmende Wasserverschmutzung in Küstenregionen hat Auswirkungen auf das Überleben von Foraminiferen.

Mit einem Massenspektrometer vom Typ Fusion der Firma Thermo Fisher Scientific, wurden jeweils 330ng je tryptischem Verdau analysiert. Je Herkunftsort und Fangjahr wurden drei biologische Replikate aus je 8 kompletten Holobionten ausgewählt, die parallel zueinander vorverarbeitet und nacheinander massenspektrometrisch analysiert wurden. Die Vorverarbeitung ist in Stuhr et al. (2017a) beschrieben.

Nach dem Proteinverdau und einer Qualitätskontrolle wurde die sogenannte Aminosäureanalyse aller Proben durchgeführt. Infolgedessen wurde die Probenmenge korrigiert um gleiches Ausgangsmaterial für die LC-MS-Analyse zu gewährleisten.

Datensätze Biodiversität-Exactive und Biodiversität-Orbitrap

Die bisher unveröffentlichten Datensätze Biodiversität-Exactive und Biodiversität-Orbitrap beinhalten größtenteils Proben der Gattung *Radix* (Pfenninger et al., 2006). Diese wurden ergänzt durch eine andere Tellerschnecke (*Ancylus*), den gemeinen Regenwurm (*Lumbricus terrestris*) und Meeresfrüchte. Das SBiK Frankfurt hat die Auswahl an Proben mit großer Biodiversität getroffen und verblindet. In Tabelle 2.4 sind 30 Proben beschrieben. Sie wurden in permutierter Reihenfolge mit dem Massenspektrometer Q Exactive und anschließend als technische Replikate in gleicher Reihenfolge mit dem Massenspektrometer Orbitrap Elite analysiert.

Tabelle 2.4: Abkürzungen und teilweise Herkunft und Teile (Fuß, Geschlechtsorgan) der untersuchten Arten und Nummerierung (Nr.) der Messreihenfolge der Proben in den Datensätzen Biodiversität-Exactive und Biodiversität-Orbitrap.

Abkürzung	Beschreibung	Teil	Nr.
MOTU41, MOTU42, MOTU43	*R. auricularia*, ISAS		1, 2, 3
Wa1, Wb1, Wc1	*R. auricularia**, Taunus	G	25, 14, 27
Wa2, Wb2, Wc2	*R. auricularia**, Taunus	F	8, 7, 26
DGE15a1	*R. balthica**, Gelnhausen	G	13
DGE15a2	*R. balthica**, Gelnhausen	F	15
KAT1, KAT2, KAT3,	*Radix*		24, 9, 29
KAT4			11
KSHT2, KSHT3, KSHT6	*Radix*		21, 30, 5,
KSHT7, KSHT8, KSHT9,			4, 17, 10,
KSHT10, KSHT11, KSHT12,			16, 12, 6,
KSHT13			22
A1	*Ancylus**		28
LT1	*Lumbricus terrestris**		18
FDM1, FDM2, FDM3	Meeresfrüchte		23, 19, 20

*Kein DNA-Barcode, sondern Spezieseinteilung basiert auf dem Phänotyp.

Für einzelne Proben ist die Herkunft, eine Inzuchtlinie in einem Aquarium am ISAS, Wildfänge aus Gelnhausen und dem Taunus, bekannt. Teilweise wurden einzelne Teile, Fuß oder Geschlechtsorgan, extrahiert. Falls in der Tabelle keine Angabe

gemacht ist, wurden ganze Tiere (Schnecken, Würmer, Meeresfrüchte) verwendet. Um die Probenmenge konstant zu halten, wurden ganze Tiere der Schnecken gewählt und bei größeren Tieren, den Meeresfrüchten und Würmern, ein Teil der Probe ausgewählt, der der Größe einer Schnecke entspricht. Für viele Proben wurde für die Artenbestimmung kein DNA-Barcoding durchgeführt. Die Einteilung basiert in diesen Fällen auf dem Phänotyp.

Radix ist eine Gattung der Süßwasserschnecken (Pfenninger et al., 2006). Die Spezieseinteilung basiert historisch auf der Morphologie der Schalen und mittlerweile auf DNA-basierten Methoden. In Nordwesteuropa werden fünf *Radix* Spezies, *R. ampla*, *R. auricularia*, *R. balthica*, *R. labiata* und *R. lagotis* unterschieden. Eine genaue Einteilung der *Radix* ist wünschenswert, da sie als Indikator für Wasserqualität dienen und parasitäre Krankheiten über jene Gattung zum Menschen übertragen werden. DNA-Barcoding ermöglicht die Einteilung in fünf sogenannte MOTU (engl. molecular operational taxonomic unit). Die biologische Einheit MOTU2 ist mit dem taxonomischen Namen *R. balthica* assoziiert. MOTU4 wird in Verbindung mit *R. auricularia* gebracht.

Datensatz Palmblad

Der von Palmblad und Deelder (2012) beschriebene compareMS2-Algorithmus bildet die Grundlage für den neuen DISMS2-Algorithmus. Sie rekonstruierten den korrekten phylogenetischen Baum für Menschenaffen und andere Primaten anhand von massenspektrometrischen Analysen von Blutseren. Der Palmblad-Datensatz stammt aus Palmblad und Deelder (2012) und ist über das PRIDE-Repository (PRD000375, Accession-Nummer 16286-16312) im mzXML-Format frei verfügbar. Es wurden jeweils nur 2000 Tandem-Massenspektren veröffentlicht, die nach den höchsten Totalionensignalen ausgewählt wurden. Dies entspricht in etwa der Anzahl an identifizierten Spektren der humanen Proben in einer Sequenzdatenbank. Die Spektren wurden vorverarbeitet, sodass jedes Spektrum aus exakt 50 Peaks besteht. Die m/z-Werte des Precursors liegen zwischen 300 und 1300. Die Angabe der Precursorladung fehlt in den Zusatzinformationen zu den Spektren.

Es wurden je vier Blutseren von Westlicher Gorilla (*Gorilla gorilla*), Mensch (*Homo sapiens*), Javaneraffe (*Macaca fascicularis*), Rhesusaffe (*Macaca mulatta*) und Borneo-Orang-Utan (*Pongo pygmaeus*) sowie sechs Blutseren vom Gemeinen Schimpansen (*Pan troglodytes*) analysiert. Als Referenz wurde eine Probe des Bakteriums *Escherichia coli* hinzugefügt (Tabelle 2.5). Im weiteren Verlauf werden die Läufe anhand der Abkürzungen in Tabelle 2.5 (siehe auch Tabelle A.3) bezeich-

net. Jeweils 2 μL je tryptischem Verdau wurden mit einem Massenspektrometer der Bruker Corporation, amazon speed ETD, analysiert.

Tabelle 2.5: Abkürzungen der untersuchten Arten im Palmblad-Datensatz.

Abkürzung	Art
GG1, GG2, GG3, GG4	*Gorilla gorilla*
HS1, HS2, HS3, HS4	*Homo sapiens*
MF1, MF2, MF3, MF4	*Macaca fascicularis*
MM1, MM2, MM3, MM4	*Macaca mulatta*
PP1, PP2, PP3, PP4	*Pongo pygmaeus*
PT1, PT2, PT3, PT4, PT5, PT6	*Pan troglodytes*
EC1	*Escherichia coli*

2.2.2 Datenbanksuche und De-Novo-Peptidsequenzierung

Es werden zwei Vorgehensweisen bei der Datenbanksuche unterschieden. Zum einen erfolgt die Suche in einer Peptid-Spektralbibliothek (engl. peptide spectral library), einer Datenbank, die Massenspektren und deren Peptidannotationen enthält. Zum anderen werden Protein-Datenbanken genutzt. Die enzymatische Spaltung der Proteine in der Datenbank in theoretische Peptide erfolgt anhand fester Regeln, z. B. der spezifischen Spaltung von Trypsin nach Arginin (R) und Lysin (K) (siehe Abschnitt 2.1). Anschließend wird die Peptidfragmentierung simuliert, sodass theoretische Tandem-Massenspektren entstehen.

Die erste kommerziell erhältliche Datenbanksuche SEQUEST (Yates et al., 1996) verwendet zum Scoring die Kreuzkorrelation. Die beobachteten Spektren werden in mehreren Schritten vorverarbeitet. Die Intensitäten der Peaks werden normalisiert, Peaks mit niedrigen Intensitäten werden entfernt und die m/z-Werte werden auf ganze Zahlen abgerundet. Für die Erzeugung der theoretischen Spektren aus einer Protein-Datenbank werden vereinfachte Fragmentierungsregeln verwendet. Zwischen einem beobachteten Spektrum und allen theoretischen Spektren der Datenbank wird der Kreuzkorrelations-Score X_{corr} bestimmt. Für die Intensitätsvektoren $I_b = (I_{b,1}, \ldots, I_{b,n})'$ und $I_b = (I_{t,1}, \ldots, I_{t,n})'$ eines beobachteten und eines theoretischen Spektrums gilt in der aktuellen Implementierung (Eng et al., 2008):

$$X_{\text{corr}} = Corr(0) - \frac{1}{150} \sum_{x=-75, x\neq 0}^{75} Corr(x), \qquad Corr(x) = \sum_i I_{b,i}, I_{t,i+x}.$$

Ursprünglich ist im Korrekturfaktor, der durchschnittlichen Autokorrelation, $x = 0$ enthalten. Das Signal Corr(0) wurde aus dem Korrekturfaktor entfernt.

Die Grundlage für das Scoring der Datenbanksuche Mascot (Perkins et al., 1999) bildet wie bei SEQUEST die Anzahl gleicher Peaks. Dieser Score basiert jedoch auf Wahrscheinlichkeiten. Bei gegebener Anzahl der Peaks im Spektrum und der Verteilung der m/z-Werte der theoretischen Peaks wird die Wahrscheinlichkeit P für einen zufälligen Treffer berechnet. Um einen leicht zu interpretierbaren *Score* zu erhalten wird der P-Wert log-transformiert, sodass $Score = -10\log_{10}(P)$ gilt. Details des Scorings wurden nicht publiziert. Ein Beispiel dient zur Verdeutlichung der Werte des Scores. Es wird eine Toleranz für die Precursormasse vorgegeben, sodass nur 150 000 Peptide in der Datenbank in Frage kommen. Wird die Falsch-Positiv-Rate auf 0.05 festgelegt, so liegt die Chance für einen falsch positiven Treffer bei 1 zu 20. Aus der Wahrscheinlichkeit $P = \frac{1}{20 \cdot 150000} = \frac{1}{300000}$ ergibt sich dann der Score $65 (= -10\log_{10}(\frac{1}{300000}))$.

Die Open-Source-Datenbanksuche X!Tandem (Craig und Beavis, 2003) nutzt zur Bestimmung des Scores aus, dass die Treffer zufällig hypergeometrisch verteilt sind (Sadygov und Yates, 2003). Der Hyper-Score S_{hyper} beachtet die Anzahl n_b und n_y zugeordneter b- und y-Ionen:

$$S_{\text{hyper}} = (n_b! n_y!) \sum_i I_{b,i}, I_{t,i}$$

Das Ranking basiert auf Erwartungswerten (engl. expectation values). Bei diesem allgemeinen Konzept wird angenommen, dass ein gültiger Treffer erzielt wird, wenn der zugehörige Score maximal ist (Ghosh, 2015, S. 283ff.).

Im ersten Schritt wir die Verteilung $p(x) = f(x)/N$ der Scores x je Spektrum empirisch über die relativen Häufigkeiten $f(x)$ der N Scores aus den Daten bestimmt. Als Zweites wird die diskrete Survivalfunktion, $s(x) = P(X > x) = \sum_{i=x}^{\infty} p(i)$, berechnet. Anschließend wird an die Survivalfunktion im Bereich hoher Scores eine Modellverteilung angepasst. Mithilfe der angepassten Funktion wird die Survivalfunktion auf den gewünschten Score extrapoliert und dieser in den Erwartungswert umgewandelt. Der Erwartungswert

$$E(x) = n \cdot s(x)$$

gibt für n Peptidsequenzen die Anzahl der Treffer an, die mindestens den Score x haben.

Aus dem Vergleich der beobachteten und theoretischen Massenspektren ergibt sich je ein Peptid-Spektrum-Treffer PSM (engl. Peptide spectrum match). Eine da-

tenbankbasierte FDR-Korrektur der PSM-Liste hat sich etabliert. Dabei erfolgt die Suche zusätzlich zur Suche in der Target-Datenbank, die die gesuchten Proteine enthält, in einer künstlichen Decoy-Datenbank. Dazu wird die Reihenfolge der Aminosäuren verändert. Häufige Methoden sind die Umkehrung der Reihenfolge und die Zufallsreihenfolge der Aminosäuren.

Der DISMS2-Datensatz beinhaltet für Fadenwurm, Fruchtfliege, Mensch und Maus (`http://www.uniprot.org/`) sowie Hefe (`www.yeastgenome.org`, SGD) die Ergebnisse einer Datenbanksuche, die mit Proteome Discoverer 1.4 (Thermo Scientific) und Mascot 2.4 (Matrix Science) durchgeführt wurde. In einer Target-Decoy-Suche wurden die Proteinsequenzdatenbanken im FASTA-Format verwendet. Dabei wurden folgende Einstellungen verwendet: Enzym Trypsin, maximal zwei fehlende Spaltungen, Carbamidomethylierung von Cystein als feste Modifikationen, Oxidation von Methionin als dynamische Modifikationen, 10 ppm und 0.02 Dalton Toleranz der MS1- und MS2-Spektren, PSMs auf Platz 1 der Suche mit einer FDR $< 1\%$. Schell et al. (2017) haben einen ersten Entwurf der Sequenz des Genoms von *Radix auricularia* vorgestellt. Daher wurden nachträglich für R41, R42 und R43 Annotationen hinzugefügt.

Blank-Landeshammer et al. (2017) geben einen Überblick über Algorithmen zur De-Novo-Peptidsequenzierung und stellen eine Methode vor, die mehrere Algorithmen kombiniert, sodass bessere Ergebnisse in Form von einer kleineren empirischen False Discovery Rate (FDR) erzielt werden. Dazu wurden PEAKS, pNovo+ und NOVOR aus einer Vielzahl an Algorithmen ausgewählt (Bessant, 2016, S.30 ff.). PEAKS und NOVOR bestimmen für jede Aminosäure einen Score. Am häufigsten ist jedoch ein Peptidscore und eine Darstellung mit Massenlücken (engl. mass gap) für Bereiche, in denen es keine sicheres Ergebnis für die Sequenz gibt.

Bei PEAKS (Zhang et al., 2012) handelt es sich um eine Gesamtlösung, die die De-Novo-Peptidsequenzierung (Ma et al., 2003) in Pipelinelösungen mit Datenbanksuchalgorithmen integriert. Das Konzept der dynamischen Programmierung wird genutzt um effizient eine optimale Peptidsequenz zu berechnen. Es werden tausende Peptidkandidaten erzeugt, die mit stringenten Scoring-Funktionen neu bewertet werden. Schließlich werden die Konfidenzscores neu kalibriert. Für jede einzelne Aminosäure in der Sequenz wird schließlich ein lokaler Konfidenzscore angegeben.

Im Gegensatz zu der kommerziellen Software PEAKS sind pNovo+ und NOVOR frei verfügbar. Die Grundidee der De-Novo-Peptidsequenzierung kann durch einen gerichteten azyklischen Graph (engl. directed acyclic graph, DAG) dargestellt werden. Jeder gemessenen Masse eines Peptidfragments wird ein Knoten zugewiesen.

Eine gerichtete Kante zwischen zwei Massen wird hinzugefügt, falls die Massendifferenz der Knoten der Masse einer Aminosäure entspricht. Das Verfahren pNovo+ (Chi et al., 2013) zeichnet sich dadurch aus, dass der Algorithmus pDAG verwendet wird. Dieser findet in kurzer Zeit die k-längsten Pfade in einem DAG.

NOVOR (Ma, 2015) verwendet maschinelles Lernen zum Scoring, das auf einem Entscheidungsbaum basiert (Bessant, 2016, S.29 ff.). Es handelt sich um einen zweistufigen Algorithmus, einer dynamischen Programmierung und einem Verbesserungsschritt. Anhand einer Peptidspektralbibliothek mit über 300 000 Spektren hat der Algorithmus automatisch einen Entscheidungsbaum mit über 14 000 Knoten gelernt. Der Baum basiert auf 169 Features. Scoring-Features sind beispielsweise die Peakintensitäten, der Ladungszustand, die Entfernung der Aminosäure vom C- und N-Terminus. Für jede Aminosäure einer Kandidaten-Peptidsequenz wird ein Konfidenzscore berechnet. Daraus wird ein Peptidscore als ein gewichtetes Mittel der Aminosäurenscores erstellt. NOVOR ermöglicht eine Echtzeit-De-Novo-Peptidsequenzierung, da innerhalb einer Sekunde mehr als 300 Spektren auf einem Laptop sequenziert werden können.

Für alle MS/MS-Läufe des Foraminiferen-Datensatz wurde eine De-Novo-Peptidsequenzierung mit der von Blank-Landeshammer et al. (2017) vorgestellten Methode durchgeführt. Es wurden nur vollständige Sequenzen von übereinstimmenden Spektrenannoationen von PEAKS, pNovo+ und Novor verwendet, die eine erwartete False Discovery Rate (FDR) von 5% aufweisen. Die Distanz der Peptidlisten wird anhand des in Kapitel 3.2 beschriebenen Sørensen-Dice-Index bestimmt.

3 Methoden zur Gruppierung von Tandem-Massenspektren

Kapitel 3 gibt einen Überblick über statistische Methoden, die in Kapitel 4 und 5 in der Datenanalyse zur Gruppierung von Tandem-Massenspektren verwendet werden. Zunächst wird ein Massenspektrum mathematisch definiert und Distanzmaße zum Vergleich von Tandem-Massenspektren als Grundlage aller weiteren Verfahren vorgestellt (Abschnitt 3.1). Anschließend werden Möglichkeiten für einen proteomweiten Abstand für LC-MS/MS-Läufe, die aus tausenden Tandem-Massenspektren von der Messung einer Probe stammen, diskutiert (Abschnitt 3.2). Gebräuchlich sind Varianten des Abstands von Peptidlisten, für die Spektren annotiert werden müssen. Als neue Methode wird in Abschnitt 3.2.1 der DISMS2-Algorithmus vorgestellt, der ohne Peptidannotation auskommt. Da für diesen Algorithmus eine Vielzahl von Parametern gewählt werden müssen, wird ein Arbeitsablauf zur Parameteroptimierung empfohlen, der auf dem Bestimmtheitsmaß R^2 eines nichtparametrischen Verfahrens zur Varianzanalyse basiert (Abschnitt 3.2.2). Abschnitt 3.3 handelt von Clusterverfahren für Tandem-Massenspektren, die auf Distanzen basieren. Ausgehend von Grundlagen für geeignete Clusteralgorithmen folgt eine detaillierte Beschreibung unter Zuhilfenahme von Pseudocode von sechs etablierten Methoden (Hierarchische Clusteranalyse, CAST, DBSCAN, igraph, MS-Cluster und PRIDE Cluster) und der neuen Methode Neighbor Clustering (Abschnitt 3.3.1). Diese werden in der Datenanalyse in Kapitel 5 verglichen bezüglich mehrerer Bewertungsmaße zur Qualität von Clusterlösungen, die in Abschnitt 3.3.2 beschrieben sind. Das Kapitel endet mit Möglichkeiten zur Visualisierung von Abständen (Abschnitt 3.4). Statistische Verfahren, das Dendrogramm, die Visualisierung einer hierarchischen Clusterlösung, und die multidimensionale Skalierung, werden ergänzt durch Verfahren der Biodiversitätsforschung, den phylogenetischen Bäumen.

3.1 Distanzmaße für Tandem-Massenspektren

Im Folgenden werden Distanzmaße für den Vergleich von Tandem-Massenspektren diskutiert. Für einen MS/MS-Lauf i, der n_i MS/MS-Spektren enthält, wird das k-te Spektrum in Lauf i, S_{k_i}, als eine Menge zweier Vektoren $\boldsymbol{x}_{k_i}$ und $\boldsymbol{I}_{k_i}$ der Länge p_{k_i} definiert:

$$S_{k_i} = \{\boldsymbol{x}_{k_i}, \boldsymbol{I}_{k_i}\} = \left\{(x_{k_i,1}, \ldots, x_{k_i,p_{k_i}})', (I_{k_i,1}, \ldots, I_{k_i,p_{k_i}})'\right\}$$

Die Masse-zu-Ladung-Verhältnisse (m/z) $\boldsymbol{x}_{k_i}$ sind in aufsteigender Reihenfolge sortiert und die zugehörigen Peak-Intensitäten werden mit $\boldsymbol{I}_{k_i}$ bezeichnet.

Eine abgewandelte Version der Vektoren, die ein Spektrum definieren, ist hilfreich bei der Berechnung von Distanzen für Spektren. Abhängig von der experimentellen Auflösung wird dazu der Wertebereich der m/z-Verhältnisse in kleine Intervalle unterteilt, sodass jeder Peak genau einem Intervall zugeordnet werden kann. Der Eintrag an einer bestimmten Position des Vektors $\tilde{\boldsymbol{I}}_{k_i} = (\tilde{I}_{k_i,1}, \ldots, \tilde{I}_{k_i,\tilde{p}})'$ mit $\tilde{p}$ Einträgen ist die Peak-Intensität, falls ein Peak zugeordnet wurde, und ansonsten 0.

In Deza und Deza (2016, S. 2ff) sind die Begriffe Distanzfunktion, Semimetrik, Metrik und Ähnlichkeit voneinander abgegrenzt. Für eine Menge X, z. B. die Menge $X = \{S_{1,1}, \ldots S_{n_1,1}\}$ aller n_1 MS/MS-Spektren im ersten Lauf, ist $d : X \times X \to \mathbb{R}$ eine Distanzfunktion (Deza und Deza, 2016, S. 2), auch Distanzmaß genannt, falls für alle $x, y \in X$ folgende Axiome gelten:

$$d(x, y) \geq 0 \text{ (nichtnegativ)} \tag{3.1}$$

$$d(x, y) = d(y, x) \text{ (symmetrisch)} \tag{3.2}$$

$$d(x, x) = 0 \text{ (Reflexivität)} \tag{3.3}$$

Es handelt sich also um eine Funktion, die in den Bereich der positiven reellen Zahlen abbildet (3.1). Die Distanz von zwei identischen Argumenten ist 0 (3.3). Die Symmetrie der Funktion (3.2) gewährleistet, dass die Reihenfolge der Argumente (hier: Massenspektren) keine Rolle spielt.

Gilt zusätzlich zu den Axiomen der Distanzfunktion die Dreiecksungleichung, ist d eine Semimetrik (Deza und Deza, 2016, S. 2):

$$d(x, y) \leq d(x, z) + d(z, y) \text{ (Dreiecksungleichung)} \tag{3.4}$$

Es handelt sich bei d um eine Metrik, häufig ebenfalls als Distanz bezeichnet, falls zusätzlich zu den Axiomen der Distanzfunktion und der Dreiecksungleichung gilt, dass identische Argumente eine Distanz von 0 haben (Deza und Deza, 2016, S. 3):

$$d(x, y) = 0 \Leftrightarrow x = y \text{ (Identität der Ununterscheidbaren)} \tag{3.5}$$

Alternativ zu Distanzen von MS/MS-Spektren werden oft Ähnlichkeiten angegeben. Eine Funktion $s : X \times X \to \mathbb{R}$ wird Ähnlichkeit genannt, falls zusätzlich zur Nichtnegativität (vgl. 3.1) und Symmetrie (vgl. 3.2) folgendes gilt (Deza und Deza, 2016, S. 2):

$$s(x, y) \geq s(x, x) \quad \forall x, y \in X \tag{3.6}$$

$$s(x, y) = s(x, x) \Leftrightarrow x = y \tag{3.7}$$

Distanzen als Maße der Unähnlichkeit und Ähnlichkeiten stehen im Zusammenhang zueinander. Aus einer Ähnlichkeit lässt sich eine Distanz durch verschiedene Transformationen erzeugen, z. B. $d = 1 - s$ und $d = \arccos s$.

Die Kosinus-Distanz $d_{\cos}$ ist das meistverwendete Distanzmaß für den paarweisen Vergleich von Massenspektren (Kim und Zhang, 2013). Für die Kosinus-Ähnlichkeit von zwei Spektren $S_{k_i} = \{\boldsymbol{x}_{k_i}, \boldsymbol{I}_{k_i}\}$ und $S_{l_j} = \left\{\boldsymbol{x}_{l_j}, \boldsymbol{I}_{l_j}\right\}$ wird das Skalarprodukt der Intensitätsvektoren in der alternativen Definition durch das Produkt der euklidischen Norm jener Vektoren geteilt. Die Kosinus-Distanz ergibt sich dann, indem die Kosinus-Ähnlichkeit von 1 abgezogen wird:

$$d_{\cos}(S_{k_i}, S_{l_j}) = 1 - \frac{\left\langle \tilde{\boldsymbol{I}}_{k_i}, \tilde{\boldsymbol{I}}_{l_j} \right\rangle}{\left|\tilde{\boldsymbol{I}}_{k_i}\right| \left|\tilde{\boldsymbol{I}}_{l_j}\right|} = 1 - \frac{\sum_{q=1}^{\tilde{p}} \tilde{I}_{k_i,q} \cdot \tilde{I}_{l_j,q}}{\sqrt{\sum_{q=1}^{\tilde{p}} \tilde{I}_{k_i,q}^2} \cdot \sqrt{\sum_{q=1}^{\tilde{p}} \tilde{I}_{l_j,q}^2}}$$

Je nach Vorverarbeitung der Spektren, z. B. einer Auswahl der `topn` (`topn` $\in \mathrm{N}$) höchsten Peaks je Spektrum, ist eine Kosinus-Distanz, die den Wert der Intensitäten ignoriert, besser geeignet. Novak et al. (2013) definieren auf Grundlage der Original-Intensitätsvektoren die sogenannte Winkel-Distanz (engl. angle distance), eine alternativen Kosinus-Distanz:

$$d_{\text{angle}}(S_{k_i}, S_{l_j}, \epsilon) = \arccos\left(\frac{\sum_{q=1}^{p_{k_i}} \max_{q^*=1,\ldots,p_{l_j}} 1_{\left\{\left|x_{k_i,q} - x_{l_j,q^*}\right| \leq \epsilon\right\}}}{\sqrt{p_{k_i} \cdot p_{l_j}}}\right)$$

Für jeden Peak in Spektrum S_{k_i} wird aufsummiert, ob es mindestens einen Peak in Spektrum S_{l_j} mit einem maximalen Positionsabstand von ϵ gibt. Die Fehlertoleranz ϵ berücksichtigt so die Vernachlässigung von kleinen Messungenauigkeiten. Für $\epsilon = 0$ entspricht die Summe dem Zähler der Kosinus-Ähnlichkeit. Der Nenner ist allgemein identisch zum Nenner der Kosinus-Ähnlichkeit, da $p_{k_i} = \sum_{q=1}^{\tilde{p}} \tilde{I}_{k_i,q}^2$ gilt. Bei der Winkel-Distanz handelt es sich also um eine Transformation einer Kosinus-Ähnlichkeit durch die Arkuskosinus-Funktion, die im Wertebereich $[0, 1]$ streng monoton fallend ist. Der Zusammenhang des Wertebereichs der Kosinus-Distanz und der Winkel-Distanz für den Spezialfall $\epsilon = 0$, $d_{\text{angle}}(\epsilon = 0) = \arccos(1 - d_{\cos})$, ist in Abbildung 3.1 dargestellt. Es handelt sich um eine monoton wachsende Funktion oberhalb der Winkelhalbierenden, die Winkel-Distanz nimmt also höhere Werte an als die Kosinus-Distanz und ist nicht auf das Intervall $[0, 1]$ normiert, sondern auf $\left[0, \frac{\pi}{2}\right]$.

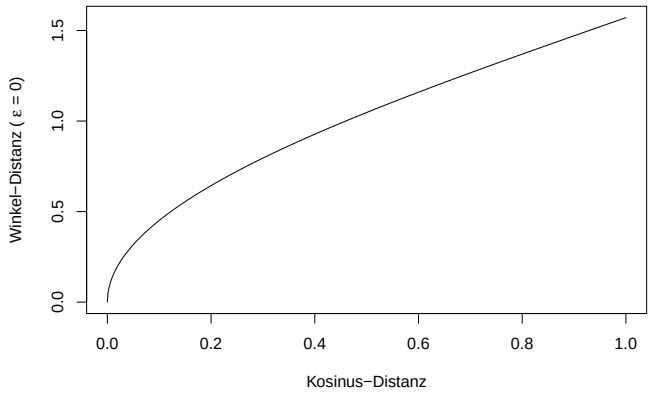

Abbildung 3.1: Zusammenhang der Wertebereiche zwischen Kosinus-Distanz und Winkel-Distanz mit Parameter $\epsilon = 0$.

Novak et al. (2013) haben die parametrisierte Hausdorff-Distanz vorgestellt:

$$d_{\text{PH}}(S_{k_i}, S_{l_j}, \delta, k) = \max(h(S_{k_i}, S_{l_j}, \delta, k), h(S_{l_j}, S_{k_i}, \delta, k)) \text{ mit}$$
$$h(S_{k_i}, S_{l_j}, \delta, k) = \frac{1}{p_{k_i}} \sum_{q=1}^{p_{k_i}} \left(\min_{q^*=1,\ldots,p_{l_j}} \left| x_{k_i,q} - x_{l_j,q^*} \right| 1_{\left\{ \left| x_{k_i,q} - x_{l_j,q^*} \right| > \delta \right\}} \right)^{1/k}.$$

Für eine gegebene Fehlertoleranz δ, mittelt h die k-te Wurzel der minimalen absoluten Distanz von der Position aller Peaks von Spektrum S_{k_i} im Vergleich zu allen Peaks von Spektrum S_{l_j}, falls diese größer als δ ist. Im Vergleich zur Winkel-Distanz wird also mit der Größe der Positionsdistanz gewichtet. Über den Parameter

k wird die Anzahl irrelevanter Peaks bestraft. Gibt es sehr viele Peaks in Spektrum S_{l_j}, die wenig Ähnlichkeit mit den Peaks in Spektrum S_{k_i}, so wird die Funktion h viel größer. Da es sich bei h aufgrund des Nenners p_{k_i} um eine gerichtete Distanz handelt, wird anschließend das Maximum der Funktion h mit je vertauschten Argumenten S_{k_i} und S_{l_j} gebildet. Es ergibt sich somit eine symmetrische Distanz d_{PH}.

Etliche andere Distanzmaße wurden diskutiert, u.a. auch der Korrelationskoeffizient (Pearson-Korrelation) und der Rangkorrelationskoeffizient nach Spearman (Kim und Zhang, 2013).

3.2 Proteomweiter Abstand für LC-MS/MS-Läufe

Eine übliche Vorgehensweise zum Vergleich des Proteoms mehrerer Proben ist die Auswertung von Peptid- oder Proteinlisten, die das Ergebnis einer Datenbanksuche oder einer De-Novo-Peptidsequenzierung sind (Kapitel 2.2.2). In jeder Liste kommen Peptide bzw. Proteine nur einfach und nicht doppelt vor, auch wenn mehrere Spektren das gleiche Suchergebnis liefern. Es gibt eine Vielzahl von Alternativen für einen Abstand von zwei Mengen A und B, die Peptide oder Proteine in der Trefferliste enthalten. Ludwig und Reynolds (1988) empfehlen drei Indizes, den Jaccard-Index, den Sørensen-Dice-Index und den Ochiai-Index. Beim Jaccard-Index s_{JI} wird die Schnittmenge in Relation zur Vereinigungsmenge gesetzt:

$$s_{JI}(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Wird die Größe der Schnittmenge in Relation zu den Größen der Mengen A und B gesetzt, so bildet der Sørensen-Dice-Index s_{DI} das harmonische Mittel und der Ochiai-Index s_{OI} das geometrische Mittel von jenen Quotienten, $|A \cap B|/A$ und $|A \cap B|/B$. Der Sørensen-Dice-Index s_{DI} lässt sich alternativ als Anteil der Schnittmenge an der mittleren Größe der Mengen A und B darstellen:

$$s_{DI}(A,B) = \frac{|A \cap B|}{\frac{1}{2}(|A| + |B|)} = \frac{2}{\frac{|A|}{|A \cap B|} + \frac{|B|}{|A \cap B|}} = \frac{2}{\frac{1}{\frac{|A \cap B|}{|A|}} + \frac{1}{\frac{|A \cap B|}{|B|}}}$$

Auch der Ochiai-Index s_{OI} lässt sich in einer ähnlichen Form darstellen, nämlich als Anteil der Schnittmenge an dem Produkt der Beträge der Größen von A und B:

$$s_{OI}(A,B) = \frac{|A \cap B|}{\sqrt{|A|}\sqrt{|B|}} = \sqrt{\frac{|A \cap B|}{|A|} \cdot \frac{|A \cap B|}{|B|}}$$

Werden die Mengen als binäre Vektoren dargestellt, so entspricht der Ochiai-Index der in Kapitel 3.1 beschriebenen Kosinus-Ähnlichkeit (Deza und Deza, 2016, S. 227).

Ergänzend zum harmonischen und geometrischen Mittel der Größe der Schnittmenge in Relation zu den Größen der Mengen A und B bildet der Kulczynski-Koeffizient s_K (Legendre und Legendre, 1998, S. 257) das arithmetische Mittel:

$$s_K(A,B) = \frac{|A \cap B| \cdot (|A| + |B|)}{2 \cdot |A| \cdot |B|} = \frac{1}{2}\left(\frac{|A \cap B|}{|A|} + \frac{|A \cap B|}{|B|}\right)$$

Die Kulczynski-Distanz d_K wird über eine Transformation, $d_K = 1 - s_K$ des Kulczynski-Koeffizienten bestimmt. Analog werden für den Jaccard-Index, den Sørensen-Dice-Index und Ochiai-Index Distanzen erzeugt.

Alternativ zur Auswertung von Peptid- oder Proteinlisten, die nicht immer erstellt werden können, wird in Rieder et al. (2017a) eine Distanz für LC-MS/MS-Läufe vorgestellt, die zum Vergleich des Proteoms mehrerer Proben verwendet werden kann. Er stellt eine Erweiterung des Algorithmus compareMS2 Palmblad und Deelder (2012) dar.

3.2.1 DISMS2-Algorithmus

Der Pseudocode des neuen Algorithmus zur Berechnung dieser Distanz ist in Algorithmus 1 auf Seite 37 dargestellt. Es handelt sich bei der Berechnung der paarweisen **DIS**tanzen von N **MS2**-Läufen (DISMS2) um ein Verfahren mit vier Schritten, einem Filtern von Spektren, der Prüfung von Bedingung für die Spektrenzuordnung, der Zuordnung von MS/MS-Spektren und der Berechnung einer Distanzmatrix. Eine Implementierung des Algorithmus ist frei verfügbar in der statistischen Programmiersprache R (R Core Team, 2016)(`https://www.statistik.tu-dortmund.de/genetics-publications-DISMS2.html`).

Im ersten Schritt werden die MS/MS-Spektren vorverarbeitet und gefiltert. Optional kann vorweg gefordert werden, dass nur Peaks der MS/MS-Spektren, die zu den `topn` höchsten Intensitäten zählen, verwendet werden. Anschließend wird für alle Spektren ein Binning mit einer frei wählbaren Bingröße `bin` durchgeführt. In compareMS2 (Palmblad und Deelder, 2012) ist ein Binning nur mit dem konstanten Wert `bin` $= 0.2$ möglich. Beim Binning handelt es sich um einen etablierten Vorschritt zur Distanzberechnung von Massenspektren (Keerthikumar und Mathivanan, 2016, S. 81). Es werden sogenannte Bins, d. h. kleine Intervalle $[n^* \cdot \texttt{bin}, (n^* + 1) \cdot \texttt{bin})$ ($n^* \in \mathrm{N}_0$), gebildet. Für die Wahl der Anzahl an Bins gibt es viele Varianten. Bei

DISMS2 wird die Anzahl der Bins fest gewählt, indem sie sich aus der Größe der Bins ergibt. In jedem der Intervalle werden nun die Intensitäten durch einen Repräsentanten ersetzt. Als Repräsentant wird die maximal gemessene Intensität gewählt und dem m/z-Wert $(n^* + 0.5) \cdot \texttt{bin}$, dem Intervallzentrum zugeordnet. Die Variante, die Summe der Intensitätswerte als Repräsentanten zu wählen (Frewen et al., 2006), wird nicht berücksichtigt, da die so konstruierten Intensitätswerte womöglich um ein vielfaches größer sind als reale Werte.

Die Idee des Verfahrens ist, alle MS/MS Spektren in Lauf i dem ähnlichsten Spektrum in Lauf j zuzuordnen und umgekehrt. Im zweiten Schritt wird die Anzahl möglicher Kandidaten für die Zuordnung durch die Prüfung von Bedingungen reduziert. Dies ist von Vorteil, da die Distanzberechnung aller Spektren viel Laufzeit und Speicher benötigen würde (siehe Kapitel 4.5). Zusätzlich gibt es auch inhaltliche Gründe für die Kandidatenauswahl, die im Folgenden erläutert werden.

Die drei Bedingungen werden nacheinander geprüft. Die Reihenfolge wurde so gewählt, dass bereits im ersten Schritt möglichst viele Spektren ausgeschlossen werden können. Die Bedingung (a) lautet, dass Spektren mit ähnlicher Retentionszeit ausgewählt werden. Inhaltlich ist dies durch die Chromatographie (HPLC) begründet, denn Peptide mit gleichen physikochemischen Eigenschaften eluieren zur gleichen Zeit von der Säule. Es gibt Verfahren, die sich speziell mit dem Retentionszeit-Alignment beschäftigen (Podwojski et al., 2009). Beim DISMS2-Algorithmus wurde jedoch eine Methode genutzt, die möglichst schnell ein grobes Filtern erreicht. Um die Laufzeit möglichst gering zu halten, werden der Einfachheit halber die Ränge der Scannummern verwendet. Eine Toleranz `ret` beschränkt die Anzahl der MS/MS-Spektren in Lauf j, die zeitlich vor und nach einem Spektrum in Lauf i liegen. Maximal $2 \cdot \texttt{ret} + 1$ MS/MS-Spektren bleiben übrig, wenn alle Kandidatenspektren mit einem Rang l, der im Intervall $[k - \texttt{ret}, k + \texttt{ret}]$ liegt, für eine Zuordnung zum Spektrum k in Lauf i ausgewählt werden. Die Wahl eines großen Werts für die Toleranz `ret` garantiert, dass die beste Zuordnung nicht verfehlt wird.

Da die Peaks der Massenspektren relativ zur Ladung der Ionen gemessen werden (m/z-Wert), ist der Vergleich von Spektren unterschiedlicher Precursorladungen möglich. Trotzdem wird in Bedingung (b) auf Gleichheit der Precursorladung geprüft, da Unterschiede aufgrund anderer Eigenschaften in Spektren verschiedener Ladungen auftreten.

Da die Masse von Peptiden mit gleicher Aminosäuresequenz und gleichen posttranslationalen Modifikationen nicht zu unterscheiden ist, wird in Bedingung (c) die Ähnlichkeit der Precursormassen geprüft.

Es werden für jedes Spektrum k in Lauf i mit Precursormasse m_{k_i} nur Spektren l in Lauf j ausgewählt, die im Intervall $[m_{k_i} \cdot (1 - 10^{-6}\texttt{prec}), m_{k_i} \cdot (1 + 10^{-6}\texttt{prec})]$ liegen. Somit wird eine maximale Genauigkeit der Precursormasse von `prec` ppm (parts per million) garantiert:

$$|(m_{l_j} - m_{k_i})/m_{k_i}| \cdot 10^6 \leq \texttt{prec}$$

Falls kein Spektrum in Lauf j alle Bedingungen für ein Spektrum k in Lauf i erfüllt, ist keine Zuordnung möglich. Dem Spektrum k wird dann in Schritt 4 eine Distanz, die größer als der Schwellenwert `cdis` ist, zugeordnet. Für alle anderen Spektren werden die Distanzen `dist` zu allen verbleibenden Kandidaten, die in der Menge M_i zusammengefasst sind, berechnet. In Kapitel 3.1 ist eine Auswahl an Distanzen für Massenspektren zu finden. Bisher wurden die Kosinus-Distanz $d_{\cos}$, die Winkel-Distanz d_{angle} und die parametrisierte Hausdorff-Distanz d_{PH} implementiert. Die Auswahl ist gegebenenfalls erweiterbar. Je nach Distanzmaß `dist` wird ein Schwellenwert `cdis` festgelegt, anhand dessen die Distanz in eine binäre Variable, gleiches oder unterschiedliches Spektrum, transformiert wird. Palmblad und Deelder (2012) konnten für die Kosinus-Distanz $d_{\cos}$ zeigen, dass $\texttt{cdis} = 0.2$ eine gute Wahl ist. Als Distanz zwischen Lauf i und j wird die relative Häufigkeit der Spektren in Lauf i bestimmt, für die es kein Trefferspektrum gibt:

$$d^*(i,j) = \frac{1}{n_i} \sum_{k=1}^{n_i} 1\left\{\left(\min_{l \in M_i} \texttt{dist}(S_{ki}, S_{lj})\right) > \texttt{cdis}\right\}$$

In Algorithmus 1 ist die Distanz über die Anzahl der Treffer (engl. Match) definiert. Ein Trefferspektrum liegt vor, wenn ein Spektrum aus der Menge M_i in Lauf j mit kleinster Distanz `dist` ausgewählt wird und diese kleiner als `cdis` ist. Die Distanz d^* ist nicht symmetrisch, da es sich um ein gerichtetes Verfahren handelt. Das Verfahren wird mit vertauschten Läufen i und j wiederholt und schließlich wird das arithmetische Mittel $d(i,j) = (d^*(i,j) + d^*(j,i))/2$ der gerichteten Distanzen gebildet. Für die Distanzmatrix D der gemittelten Distanzen werden $N \cdot (N-1)$ gerichtete Distanzen d^* erstellt.

Algorithmus 1 : DISMS2: **DIS**tanz von MS/MS (**MS2**) Läufen

Eingabe : N MS/MS Läufe; Parameter `topn`, `bin`, `ret`, `prec`, `dist` und `cdis`.

Ausgabe : Distanzmatrix mit Einträgen der paarweisen Distanzen der N MS/MS Läufe.

Zuerst wird Schritt 1 für jeden Lauf i ($i = 1, \ldots, N$) einzeln durchgeführt. Schritte 2 und 3 werden für jedes Paar (i, j) von MS/MS Läufen einzeln durchgeführt und wiederholt mit vertauschtem i und j. Zuletzt werden die Ergebnisse in Schritt 4 zusammengefasst.

1: Falls nur die Top `topn` Peaks der Spektren berücksichtigt werden sollen, filtere alle MS/MS-Spektren in Lauf i. Vollziehe ein Binning aller MS/MS-Spektren in Lauf i mit Bin-Größe `bin`.

2: Prüfe für jedes MS/MS-Spektrum in Lauf i ob die folgenden Bedingungen für jedes MS/MS-Spektrum in Lauf i erfüllt sind. Falls kein Spektrum alle Bedingungen erfüllt, zähle dies als einen Match mit einer Distanz, die größer als `cdis` ist. Sei k der Rang eines MS/MS Spektrums in Lauf i (zeitliche Ordnung) und l der Rang eines MS/MS-Spektrums in Lauf j. Bedingungen:

   (a) $k - \texttt{ret} \leq l \leq k + \texttt{ret}$.

   (b) Gleiche Precursorladung von Spektrum k und l.

   (c) Ähnliche Precursormasse:
   $m_{k_i} \cdot (1 - 10^{-6} \cdot \texttt{prec}) \leq m_{l_j} \leq m_{k_i} \cdot (1 + 10^{-6} \cdot \texttt{prec})$

3: Ordne MS/MS Spektrum mit Rang k in Lauf i dem MS/MS-Spektrum in Lauf j mit kleinster Distanz `dist` von allen MS/MS-Spektren, die Bedingung (a)–(c) in Schritt 2 erfüllen, zu (sogenannter Match). Als gerichtetes Distanzmaß $d^*(i, j)$ zwischen MS/MS-Lauf i und j berechne die Häufigkeit der Spektren in Lauf i, die keinen Match in Lauf j haben (alle Distanzen größer als `cdis`).

4: Erstelle eine Distanzmatrix d, deren Eintrag (i, j) die Distanz zwischen den MS/MS-Läufen i und j ist: $d(i, j) = (d^*(i, j) + d^*(j, i))/2$,
$d^*(i, j) = \#\{\text{Spektren in } i \text{ ohne Match in } j\} / \#\{\text{Spektren in } i\}$.

3.2.2 Parameteroptimierung von DISMS2

Ein Ziel bei der Evaluierung einer Distanzmatrix, die durch das DISMS2-Verfahren erstellt wurde, ist, dass Distanzen innerhalb von Gruppen kleiner als Distanzen zwischen Gruppen sind. Jene Gruppen werden bei den vorliegenden Daten durch technische oder biologische Replikate einer Spezies definiert, sie können allerdings in anderen Szenarien individuell angepasst werden. Es ist jedoch anzumerken, dass bei der massenspektrometrischen Analyse die datenabhängige Erfassung (engl. data dependent acquisition, DDA) aufgrund einer zufälligen Auswahl der Precursorionen die häufigsten Peptide einer komplexen Probe bevorzugt. Bei der DDA-Methode gibt es eine starke Streuung der Proteinidentifikation von technischen Replikaten einer Probe (Canterbury et al., 2014).

Eine etablierte Methode um Mittelwerte von Gruppen zu vergleichen ist eine Varianzanalyse (engl. analysis of variance, ANOVA, Fahrmeir et al., 1996a). Für eine ANOVA wird vorausgesetzt, dass $\{Y_{k1}, \ldots, Y_{kn_k}\}\,(k = 1, \ldots, p)$ p Mengen von Zufallsvariablen sind, die normalverteilt sind mit unbekanntem Erwartungswert μ_k und unbekannter Varianz $\sigma^2 > 0$. Um auf Unterschiede der Stichprobenmittelwerte zu testen, werden gleiche Erwartungswerte der Normalverteilungen in der Nullhypothese $H_0 : \mu_1 = \mu_2 = \ldots \mu_p$ angegeben. Die Alternative $H_1 : \exists k, l : \mu_k \neq \mu_l, k \neq l$ bedeutet, dass sich mindestens ein Stichprobenmittel μ_k von den anderen Stichprobenmitteln unterscheidet.

Eine ANOVA-Tabelle (Tabelle 3.1) fasst die Analyse zusammen. Es werden die Gruppenmittel von p Mengen von Zufallsvariablen $\overline{Y_{k.}} = \frac{1}{n_k} \sum_{m=1}^{n_k} Y_{km}$ mit dem Gesamtmittel $\overline{Y_{..}} = \frac{1}{N} \sum_{k=1}^{p} \sum_{m=1}^{n_i} Y_{km}$ verglichen. Die Gesamtvariation (engl. total sum of squares, SST) ist die Summe der Variation zwischen den Gruppen (SSB) und der Variation innerhalb der Gruppen (SSW), sodass SST = SSB + SSW gilt. Die Nullhypothese wird abgelehnt, d. h. es gibt einen signifikanten Unterschied zwischen Gruppenmitteln, falls das Verhältnis von MST und MSE, die F-Statistik

$$F = \frac{MSB}{MSW},$$

größer als das $1 - \alpha$-Quantil der F-Verteilung mit $p - 1$ und $N - p$ Freiheitsgraden ist.

Der Algorithmus DISMS2 berechnet Distanzen zwischen individuellen MS/MS-Läufen. Die Gruppenmittel $\overline{Y_{k.}}$ und das Gesamtmittel $\overline{Y_{..}}$, die für eine ANOVA benötigt werden, können nicht angegeben werden. Für Fälle, in denen für die Gruppen nur paarweise Distanzen zusammengefasst in einer Distanzmatrix vorliegen oder die Normalverteilungsannahme in den Gruppen nicht erfüllt ist, hat Anderson (2001)

Tabelle 3.1: Die ANOVA-Tabelle zeigt die Freiheitsgrade (engl. degrees of freedom, DF), Summe der Abweichungsquadrate (engl. sum of squares, SS) und mittlere Summe der Abweichungsquadrate (engl. mean square sum, MS) für Variationen zwischen den Gruppen (B), innerhalb der Gruppen (W) und gesamt (T).

Variation	DF	SS	MS
B	$p-1$	$SSB = \sum_{k=1}^{p} n_k(\overline{Y_{k.}} - \overline{Y_{..}})^2$	$MSB = SSB/(p-1)$
W	$N-p$	$SSW = \sum_{k=1}^{p}\sum_{m=1}^{n_k}(Y_{km} - \overline{Y_{k.}})^2$	$MSW = SSW/(N-p)$
T	$N-1$	$SST = \sum_{k=1}^{p}\sum_{m=1}^{n_k}(Y_{km} - \overline{Y_{..}})^2$	

ein permutationsbasiertes nichtparametrisches Verfahren vorgestellt. Den Lösungsansatz liefert die Aussage, dass in jeder Gruppe k die Summe der quadrierten Distanzen von individuellen Y_{km} zu deren Mittel $\overline{Y}_{k.}$ gleich der Summe der quadrierten Distanzen zwischen jedem Y_{km} in Gruppe k geteilt durch die Größe n_k ist:

$$\sum_{m=1}^{n_k}(Y_{km} - \overline{Y}_{k.})^2 = \frac{1}{n_k}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 1_{\{g_i=k, g_j=k\}}$$

Die Gruppenzugehörigkeiten g_i und g_j des (i,j)-ten Element d_{ij} $(i,j=1,\ldots,N)$ der Distanzmatrix D können dem Vektor $g = (g_1,\ldots,g_N)'$ entnommen werden. Es gilt $g_i = k$, falls das i-te Element der Gruppe $k(k=1,\ldots,p)$ zugeordnet ist. Somit können die Summen der Abweichungsquadrate alternativ über paarweise Distanzen dargestellt werden:

$$\widetilde{SST} = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2$$

$$\widetilde{SSW} = \frac{1}{\sum_{k=1}^{p} n_k}\sum_{i=1}^{N-1}\sum_{j=i-+1}^{N} d_{ij}^2 1_{\{g_i=k, g_j=k\}}$$

Für euklidische Distanzen entspricht F der Pseudo-F-Statistik $\widetilde{F}$:

$$\widetilde{F} = \frac{\widetilde{SST} - \widetilde{SSW}}{\widetilde{SSW}}\frac{N-p}{p-1}$$

Der Pseudo-F-Quotient $\widetilde{F}$ bildet sich aus dem Verhältnis der gemittelten Summe der quadrierten Distanzen zwischen und innerhalb der Gruppen. Ein Permutationstest wird verwendet, da die Verteilung der Teststatistik unbekannt ist. Die Verteilung der Teststatistik wird benötigt um den kritischen Wert oder p-Wert zu bestimmen. Die Verteilung wird geschätzt, indem die Gruppenzugehörigkeiten permutiert und alle möglichen Werte der Teststatistik berechnet werden. Mithilfe der

Funktion `adonis` im R-Paket `vegan` (Oksanen et al., 2016) können tausende Permutationen durchgeführt werden, um die Pseudo-F-Statistik zu berechnen.

Um schließlich verschiedene Parametereinstellungen von DISMS2 zu vergleichen, bietet sich als Gütemaß das Bestimmtheitsmaß R^2 an, also das Verhältnis der Summe der Abstandsquadrate zwischen den Gruppen und gesamt. Ein Wert nahe 1 wäre optimal, da der Anteil der gesamten Variabilität, der durch Haupteffekt zwischen Gruppen erklärt werden kann, bestimmt wird. Das Bestimmtheitsmaß $R^2 = SSB/SST$ (Fahrmeir et al., 1996c) ist äquivalent zum Effektmaß η^2 (Kennedy, 1970). Es macht eine aussagekräftige Angabe über die Effektstärke und eignet sich daher zum Vergleich bei unterschiedlichen Parametereinstellungen.

3.3 Distanzbasierte Clusterverfahren für Tandem-Massenspektren

Zunächst werden Grundlagen für Clusteralgorithmen, die auf Distanzen von Tandem-Massenspektren basieren, erläutert. Unter Einsatz von Pseudocode werden sechs bekannte Methoden (Hierarchische Clusteranalyse, CAST, DBSCAN, igraph, MS-Cluster und PRIDE Cluster) und das neue Neighbor Clustering beschrieben und Zusammenhänge hergestellt. Abschließend werden mehrere Bewertungsmaße zur Qualität von Clusterlösungen vorgestellt, die zum Vergleich der Clusteralgorithmen dienen.

3.3.1 Clusteralgorithmen

Im Allgemeinen ist die Clusteranalyse eine Methode des unüberwachten Lernens, wobei Beobachtungen in einem Datensatz in sogenannte Cluster gruppiert werden (Hastie et al., 2009, S. 507). Die Beschreibung der Beobachtungen durch Beziehungen zueinander, also Distanzen oder Ähnlichkeiten von Tandem-Massenspektren, reicht für die meisten Algorithmen aus, um eine Clusteranalyse durchzuführen. Im Folgenden werden im Detail Clusteralgorithmen beschrieben, die für Tandem-Massenspektren verwendet werden können. Die Herausforderung besteht darin, tausende Objekte mit begrenzten Mitteln bezüglich Laufzeit und Speicherverbrauch zu clustern. Methoden werden benötigt, die auf Distanzen basieren und für die die Anzahl der Cluster nicht im Vorfeld festgelegt werden muss.

Es werden partitionierende und hierarchische Verfahren unterschieden (Ester und Sander, 2000, S. 51ff.). In einem partitionierenden Verfahren erfolgt eine Zerlegung

in eine bestimmte Anzahl an Clustern, die je aus mindestens einem Objekt bestehen. Die Objekte werden dabei genau einem Cluster zugeordnet. Eine Zerlegung liegt bei hierarchischen Verfahren nicht vor. Sie bilden eine hierarchische Repräsentation, aus der die Clusterstruktur hergeleitet wird.

Neben der hierarchischen Clusteranalyse werden sechs partitionierende Verfahren, CAST, DBSCAN, igraph, Neighbor Clustering, MS-Cluster und PRIDE Cluster, untersucht, die auf Graphen basieren. Ein ungerichteter Graph $G = (V, E)$ wird erstellt, der aus Knoten V und Kanten E besteht. Die Objekte, also Tandem-Massenspektren, bilden die Knoten und ungerichtete Kanten zwischen Paaren von Knoten werden erstellt, falls die paarweise Distanz der Spektren einen Schwellenwert unterschreitet.

Das häufig verwendete R-Paket `igraph` (Csardi und Nepusz, 2006) dient der Analyse von einfachen Graphen und Netzwerken. Es enthält die Funktion `clusters`, die für einen ungerichteten Graphen die Zusammenhangskomponenten bestimmt. DBSCAN ist ein dichtebasiertes Verfahren, das sowohl mit hierarchischen Verfahren als auch mit `igraph` zusammen hängt. Es wird das neue Verfahren Neighbor Clustering vorgestellt, dessen Clusterbildung ebenso wie bei DBSCAN auf sogenannten ϵ-Nachbarschaften basiert. CAST, MS-Cluster und PRIDE Cluster sind beliebte Verfahren, die speziell zur Clusterbildung von Tandem-Massenspektren verwendet werden. Pep-Miner verwendet den CAST-Algorithmus, allerdings ist der Code nicht öffentlich zugänglich. PRIDE Cluster ist eine Erweiterung von MS-Cluster, das als approximatives hierarchisches Verfahren bezeichnet wird.

Die Auswahl an Algorithmen, die auf Graphen basieren, könnte um eine Vielzahl erweitert werden. Einige sind ebenfalls im R-Paket `igraph` enthalten. Außerdem sind Verfahren wie der viel zitierte HCS (Highly Connected Subgraphs) Clusteralgorithmus (Hartuv und Shamir, 2000) als auch spectral clustering (Hastie et al., 2009, S. 544ff.) relevant.

Zur Clusterbildung von Tandem-Massenspektren wurde auch die metrische Einbettung genutzt (Ramakrishnan et al., 2006; Dutta und Chen, 2007). MaRaCluster (The und Käll, 2016) zeichnet sich durch die Verwendung eines neuen Distanzmaßes aus. Es wird ebenfalls eine hierarchische Complete-Linkage Clusteranalyse verwendet. Diese Methode wird nicht weiter berücksichtigt, da der Einfluss des neuen Distanzmaßes nicht kontrolliert werden kann.

Eine Implementierung aller im Folgenden detailliert beschriebenen Clusteralgorithmen ist frei verfügbar in der statistischen Programmiersprache R (R Core Team, 2016)(siehe `https://www.statistik.tu-dortmund.de/genetics-publications-`

`clusspec.html`). Darin sind insbesondere Implementierungen von CAST (Kerschke, 2016) und DBSCAN (Schork, 2016) enthalten. Für MS-Cluster und PRIDE Cluster wurden R-Wrapper (R-Funktion `system3` aus dem Paket `bbmisc`, Bischl et al., 2016) erstellt. MS-Cluster v2 algorithm (Frank et al., 2011) ist ein von der Kommandozeile ausführbares Programm, das den Clusteralgorithmus MS-Cluster enthält. Für PRIDE Cluster liegt spectra-cluster-cli (Griss et al., 2016) vor, eine eigenständige Java-Applikation (spectra-cluster API Version 1.0 von Rui Wang und Johannes Griss). Der Ausgabewert der Funktionen ist ein Vektor, deren Länge der Anzahl an Spektren entspricht. Spektren, die dem gleichen Cluster i ($i = 1, \ldots, k$) zugeordnet sind, sind alle mit i kodiert. Die Ausgabewerte der Algorithmen in den folgenden Pseudocodes ist hingegen über eine Menge C definiert, die k Mengen $C_1, \ldots, C_k$ enthält. Ist ein Spektrum j ($j = 1, \ldots, n$) in Cluster i enthalten, so gilt $j \in C_i$. Der Ausgabewert eines Pseudocodes lässt sich über die Hilfsfunktion Helper (Abbildung B.1) in den Ausgabewert der zugehörigen R-Funktion transformieren.

Hierarchische Clusteranalyse

Eine hierarchische Darstellung, die Gruppen fusioniert oder aufteilt, ergibt sich aus einer hierarchischen Clusteranalyse (Hastie et al., 2009, Kapitel 14.2.12). Es werden zwei Ansätze verfolgt, von unten nach oben (engl. bottom-up, agglomerativ) und von oben nach unten (engl. top-down, divisiv). Ein agglomeratives Verfahren beginnt mit einzelnen Objekten als Cluster und fügt iterativ eine ausgewähltes Paar von Clustern zu einem Cluster zusammen. Ein divisives Verfahren hingegen teilt Cluster und beginnt mit einem großen Cluster, das alle Objekte beinhaltet. Da agglomeratives hierarchisches Clustern vorzugsweise erforscht ist, wird das bottom-up-Verfahren ausführlich erläutert.

Ein Distanzmaß für zwei Gruppen von Objekten, z. B. den Clustern, beeinflusst das Zusammenlegen von Gruppen. Für die Unähnlichkeit von Clustern werden im wesentlichen drei Varianten unterschieden, Single-Linkage (d_{SL}), Complete-Linkage (d_{CL}) und Average-Linkage (d_{AL}):

$$d_{SL}(C_l, C_m) = \min_{\substack{i \in C_l \\ j \in C_m}} d_{ij}$$

$$d_{CL}(C_l, C_m) = \max_{\substack{i \in C_l \\ j \in C_m}} d_{ij}$$

$$d_{AL}(C_l, C_m) = \frac{1}{|C_l| \cdot |C_m|} \sum_{i \in C_l} \sum_{j \in C_m} d_{ij}$$

Hier und im Folgenden bezeichnet $|C_l|$ die Mächtigkeit der Menge C_l, d. h. die Anzahl der Elemente von C_l. Das Single-Linkage fordert nur, dass eine einzelne Distanz von zwei Objekten aus den beiden Clustern klein genug ist. Dies führt zum sogenannten Verkettungseffekt (siehe auch Neighbor Clustering). Hingegen müssen beim Complete-Linkage alle paarweisen Distanzen klein genug sein. Das Average-Linkage, auch UPGMA (Unweighted Pair Group Method with Arithmetic Mean) genannt, bildet einen Kompromiss zwischen den Extremen, Single-Linkage und Complete-Linkage. Es wird die mittlere Distanz als Kriterium gewählt.

Der Pseudocode der agglomerativen hierarchischen Clusteranalyse ist in Algorithmus 2 dargestellt. Zunächst bilden die Objekte Einzelcluster. In maximal $n-1$ Schritten werden iterative Paare von Clustern zu einem neuen Cluster zusammengefügt, falls die Distanz zwischen den Clustern minimal ist. Der Algorithmus stoppt vorzeitig, falls die minimale Distanz eine feste Schranke h überschreitet. Für $h = 0$ werden $n-1$ Hierarchie-Ebenen erstellt. Zur Visualisierung eignet sich ein sogenanntes Dendrogramm (siehe Kapitel 3.4.1). Wird dieses Dendrogramm auf der Höhe h abgeschnitten, so ergibt sich die Clusterlösung, die in dem in Algorithmus 2 beschriebenen Verfahren beschrieben ist.

Algorithmus 2 : Hierarchische Clusteranalyse (Mardia et al., 1979, S. 369 ff.)

Eingabe : $n \times n$ Distanzmatrix D mit Einträgen $d_{ij} = d(S_i, S_j)$
Schwellenwert h

Ausgabe : Menge aller Cluster $C = \{C_1, \ldots, C_k\}$

1: $k = n$ Einzelcluster $C_1, \ldots C_k$
2: **while** $k > 1$ und $\min\limits_{i,j} d(C_i, C_j) \leq h$ **do**
3: Finde C_{i_1} und C_{i_2}: $d(C_{i_1}, C_{i_2}) = \min\limits_{i,j} d(C_i, C_j)$
4: Vereinige Cluster zu einem neuen Cluster: $C_{i_{\text{new}}} = C_{i_1} \cup C_{i_2}$
5: Entferne Zeile und Spalte von C_{i_1} und C_{i_2} in D
6: Für alle alten Cluster i_{old} berechne $d(C_{i_{\text{old}}}, C_{i_{\text{new}}})$ um eine Zeile und Spalte für $C_{i_{\text{new}}}$ in D einzufügen
7: $k \leftarrow k - 1$
8: **end while**
9: $C = \{C_1, \ldots, C_k\}$

Die R-Funktion `hclust` führt eine hierarchische Clusteranalyse durch. Einzelcluster im Vorhinein zu entfernen ermöglicht die Analyse auch für große Distanzmatrizen.

CAST

Clustering affinity search technique (CAST) ist ein Clusteralgorithmus, der ursprünglich zur Anwendung auf Genexpressionsdaten eingeführt wurde (Ben-Dor et al., 1999). Das von IBM entwickelte Tool Pep-Miner verwendet zur Clusterbildung von Massenspektren CAST (Beer et al., 2004). Zunächst wird die transitive Hülle gebildet. Zur Konstruktion der transitiven Hülle eines ungerichteten Graphen genügt es die Zusammenhangskomponenten eines Graphen (siehe igraph-Algorithmus) zu bestimmen (Leighton, 1992, S. 339). In einem zweiten Schritt erfolgt eine Clusterbildung für einzelne Cluster mithilfe von CAST. Letztendlich entstehen Cluster, die fast Cliquen bilden. Der Code von Pep-Miner ist leider nicht allgemein zugänglich (Frank et al., 2008).

CAST basiert auf einem graphentheoretischen stochastischen Modell, in dem wahre Clusterstrukturen durch zufällige Fehler verrauscht sind. Darin werden Cluster als Cliquen in einem ungerichteten Graphen repräsentiert. Ein heuristischer Ansatz wird vorgestellt, deren Pseudocode in Algorithmus 3 dargestellt ist. Dieser wechselt zwischen den Aktionen ADD (Algorithmus 4) und REMOVE (Algorithmus 5), also der Hinzufügung und Entfernung von Objekten über sogenannte Affinitäten.

Ein Schwellenwert t für die Distanz von Spektren wird in einen Schwellenwert $t^* = 1 - t$ für die Ähnlichkeit von Spektren transformiert. Die Affinität eines Spektrums i

$$a_i = \sum_{j \in C_{\text{open}}} (1 - d_{ij})$$

ist die Summe der Ähnlichkeit von i zu allen Objekten im aktuellen Cluster C_{open} und wird zusammengefasst im Vektor der Affinitäten $a = (a_1, \ldots, a_n)'$. Fällt die Wahl für das Distanzmaß auf die Kosinus-Distanz, so handelt es sich um die Kosinusähnlichkeit zur Berechnung der Affinitäten.

Zur Clusterbildung wird das erste Objekt zufällig gewählt, da alle Affinitäten gleich 0 sind und $|C_{\text{open}}| = 0$ gilt. Überschreitet bzw. unterschreitet die durchschnittliche Affinität eines Objekts i zu allen Objekten im aktuellen Cluster, $a_i/|C_{\text{open}}|$, einen Schwellenwert t*, so wird i hinzugefügt bzw. entfernt (Algorithmus 4 und 5). Abwechselnd werden Objekte mit hoher Affinität hinzugefügt und Objekte mit niedriger Affinität entfernt bis keine Änderungen mehr auftreten. Die Bildung des aktuellen Cluster ist abgeschlossen, da die Clusterzugehörigkeit nicht mehr wechselt. Anschließend startet der Algorithmus mit einem neuen zufälligen Objekt, das bis-

Algorithmus 3 : CAST (Ben-Dor et al., 1999)

Eingabe : $n \times n$ Distanzmatrix D mit Einträgen $d_{ij} = d(S_i, S_j)$
Schwellenwert t

Ausgabe : Menge aller Cluster $C = \{C_1, \ldots, C_k\}$

1: $t = 1 - t^*$
2: $C = \emptyset$
3: $U = \{1, \ldots, n\}$ {Unzugewiesene Punkte}
4: **while** $U \neq \emptyset$ **do**
5: $C_{\text{open}} = \emptyset$
6: $a = (0, \ldots, 0)'$ {Affinitäten}
7: **repeat**
8: ADD und REMOVE
9: **until** Es gibt keine Änderungen mehr.
10: $C = C \cup \{C_{\text{open}}\}$
11: **end while**

Algorithmus 4 : ADD

Eingabe : aktuelles Cluster C_{open}, Affinitäten a, Menge U
Distanzmatrix D
Schwellenwert t

Ausgabe : C_{open}, a, U

1: $t = 1 - t^*$
2: **while** $\max\{a_i | i \in U\} \geq t^* |C_{\text{open}}|$ **do**
3: wähle $j \in \{j \mid j \in U \wedge a_j = \max\{a_i | i \in U\}\}$
4: $C_{\text{open}} = C_{\text{open}} \cup \{j\}$
5: $U = U \setminus \{j\}$
6: **for all** $i \in U \cup C_{\text{open}}$ **do**
7: $a_i = a_i + 1 - d_{ij}$
8: **end for**
9: **end while**

Algorithmus 5 : REMOVE

Eingabe : aktuelles Cluster C_{open}, Affinitäten a, Menge U
Distanzmatrix D
Schwellenwert t

Ausgabe : C_{open}, a, U

1: $t = 1 - t^*$
2: **while** $\min\{a_i | i \in C_{\text{open}}\} < t^* |C_{\text{open}}|$ **do**
3: wähle $j \in \{j \mid j \in C_{\text{open}} \wedge a_j = \min\{a_i | i \in C_{\text{open}}\}\}$
4: $C_{\text{open}} = C_{\text{open}} \setminus \{j\}$
5: $U = U \cup \{j\}$
6: **for all** $i \in U \cup C_{\text{open}}$ **do**
7: $a_i = a_i - (1 - d_{ij})$
8: **end for**
9: **end while**

her noch nicht zugeordnet wurde. Schließlich endet die Clusterbildung, sobald alle Objekte zugeordnet wurden.

Letztendlich gilt für jedes Objekt, dass die durchschnittliche Affinität im zugehörigen Cluster mindestens t^* beträgt. Es ist jedoch möglich, dass die durchschnittliche Affinität zu einem anderen Cluster größer als zum zugehörigen Cluster ist.

DBSCAN

Die dichtebasierte räumliche Clusteranalyse mit Rauschen (engl. Density Based Spatial Clustering of Applications with Noise, DBSCAN) ist ein dichtebasierter Clusteralgorithmus (Ester et al., 1996). Der Algorithmus detektiert Cluster und zusätzlich sogenannte Rauschpunkte. Die Objekte werden so gruppiert, dass die Dichte der Objekte innerhalb eines Clusters größer ist als außerhalb des Clusters und die Dichte der Objekte innerhalb der Gruppe der sogenannten Rauschpunkte kleiner ist als zu jedem anderen Cluster. Drei Punkttypen, Kern-, Rand- und Rauschpunkte, werden unterschieden. Um einen Kernpunkt zu definieren müssen mindestens $MinPts$ Punkte in einem maximalen Radius von ϵ liegen. Die ϵ-Nachbarschaft eines Punktes p ist die Menge der Punkte mit maximaler Distanz ϵ:

$$N_\epsilon(p) = \{q | dist(p, q) \leq \epsilon\}\,.$$

Für die Clusterbildung von Tandem-Massenspektren sind als Distanz $dist(p, q)$ zweier Objekte p und q die in Kapitel 3.1 beschriebenen Abstände geeignet. Randpunkte liegen ebenfalls in der ϵ-Nachbarschaft, jedoch ist die Mindestanzahl an Objekten in ihrer ϵ-Nachbarschaft nicht erfüllt. Objekte, die weder Kern- noch Randpunkte sind, werden als Rauschpunkte bezeichnet und können als Einzelcluster interpretiert werden (Ester et al., 1996, Definition 6).

Für die Clusterbildung im DBSCAN-Algorithmus werden Bedingungen überprüft, die aus folgenden Beziehungen zwischen Objekten hergeleitet werden:

- *direkt Dichte-erreichbar* Ein Objekt p ist direkt Dichte-erreichbar von einem Objekt q, falls q ein Kernpunkt ist und p in der ϵ-Nachbarschaft von q enthalten ist.

- *Dichte-erreichbar* Ein Objekt p ist Dichte-erreichbar von q, falls es eine Kette von Objekten $p_1, \ldots, p_n$ von p $(p = p_1)$ nach q $(q = p_n)$ gibt und jedes Objekt p_{i+1} direkt Dichte-erreichbar vom vorigen Objekt p_i in der Kette ist.

- *Dichte-verbunden* Die Objekte p und q sind miteinander Dichte-verbunden, falls p und q direkt Dichte-erreichbar vom gleichen Objekt o sind.

Die Beziehung zwischen Kern- und Randpunkten ist in Abbildung 3.2 dargestellt. Der Kernpunkt q ist vom Randpunkt p nicht direkt Dichte-erreichbar oder Dichte-erreichbar, aber p und q sind über einen weiteren Kernpunkt o dichte-verbunden. Ein Cluster ist über zwei Bedingungen, Maximalität und Verbundenheit, in Abhängigkeit von ϵ und $MinPts$ definiert (Ester et al., 1996, Definition 5):

- *Maximalität* Falls q Dichte-erreichbar von einem im Cluster enthaltenen Objekt p ist, so ist q ebenfalls im Cluster enthalten.

- *Verbundenheit* Für alle Objekte in einem Cluster gilt, dass sie miteinander dichte-verbunden sind.

Eine alternative, äquivalente Definition eines Clusters für einen darin enthaltenen Kernpunkt p wird bei der Clusterbildung im DBSCAN-Algorithmus verwendet (Ester et al., 1996, Lemma 2). C_i ist die Menge aller Objekte, die von p Dichte-erreichbar sind:

$$C_i = \{o | o \text{ ist Dichte-erreichbar von } p\}$$

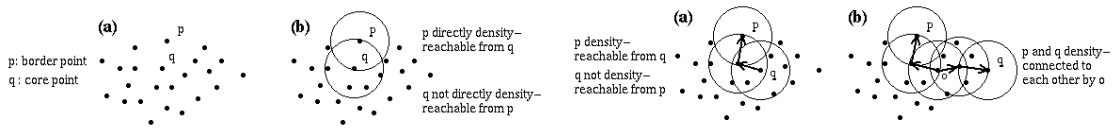

Abbildung 3.2: Grafische Darstellung (Abbildung 2 und 3 entnommen aus Ester et al., 1996) der Beziehung zwischen Kern- und Randpunkten q und p (engl. core point, border point). Trotz Dichte-Verbundenheit (engl. density-connected) von p und q über o gilt für q von p keine (direkte) Dichte-Erreichbarkeit (engl. (directly) density-reachable).

Ist der Radius ϵ einer Nachbarschaft und die minimale Anzahl an Objekten in einer Nachbarschaft eines Kernpunkt, $MinPts$, gegeben, so startet der DBSCAN-Algorithmus (Algorithmus 6) die Clusterbildung, indem ein zufälliges Objekt i ausgewählt wird. Zuerst werden die Nachbarn von i in der ϵ-Nachbarschaft herausgegriffen. Sie sind direkt Dichte-erreichbar von i. Falls weniger als $MinPts$ Nachbarn gefunden wurden, wird i als Rauschpunkt markiert. Ansonsten werden i und seine Nachbarn in einem neuen Cluster C_{open} verbunden. Weitere Objekte können in einer zweiten Suche in den Nachbarschaften der Nachbarn von i hinzugefügt werden. Falls diese Objekte die Bedingung für Kernpunkte erfüllen, d. h. mindestens $MinPts$ Punkte liegen in der ϵ-Nachbarschaft, werden ihre Nachbarn ebenfalls dem gleichen Cluster zugeordnet. Falls das aktuelle Cluster nicht mehr vergrößert werden kann, wird ein neues bisher unzugewiesenes Objekt zufällig ausgewählt. Das Verfahren startet erneut wie für das erste zufällige Objekt beschrieben. Sobald alle Objekte einem Cluster zugeordnet sind oder als Rauschpunkte markiert wurden, ist die Prozedur beendet. Da die Zuordnung der Randpunkte zu Clustern nicht eindeutig ist, werden doppelt zugeordnete Randpunkte entfernt. Die Rauschpunkte werden in dieser Variante als Einzelcluster der Clusterlösung hinzugefügt, sodass am Ende jedes Objekt genau einem Cluster zugeordnet ist.

DBSCAN mit $MinPts = 2$ oder $MinPts = 1$ entspricht dem Clusterergebnis, wenn bei hierarchischem Clustern mit Single-Link das Dendrogramm auf der Höhe von h abgeschnitten wird (Campello et al., 2015, Korollar 3.5). Bei $MinPts = 2$ werden Einzelcluster womöglich als Rauschen interpretiert und bei MinPts=1 gibt es einheitliche Cluster.

In Aggarwal (2015, Abbildung 6.15, S.182) ist eine Darstellung des Algorithmus zu finden, der DBSCAN graphenbasiert interpretiert. Zunächst werden Kernpunkte in einem Radius von ϵ in einem Graphen verbunden und anschließend die Zusammenhangskomponenten (s. u.) bestimmt. Anschließend werden Randpunkte den Zusammenhangskomponenten mit bester Verbindung zugeordnet. Für Kernpunkte ist die

Algorithmus 6 : DBSCAN (vgl. Viswanath und Babu, 2009)

Eingabe : $n \times n$ Distanzmatrix D mit Einträgen $d_{ij} = d(S_i, S_j)$ {für N_ϵ benötigt}
Schwellenwert ϵ
Mindestanzahl an Nachbarn $MinPts$

Ausgabe : Menge aller Cluster $C = \{C_1, \ldots, C_k\}$

$U = \{1, \ldots, n\}$ {unzugewiesene Punkte}
$Q = \emptyset$ {Warteschlange}
$C = \emptyset$ {Menge der Cluster}
while $U \neq \emptyset$ **do**
  wähle $i \in U$
  $U = U \setminus \{i\}$
  if $|N_\epsilon(i)| \geq MinPts$ **then**
    $C_{\text{open}} = \{j : j \in N_\epsilon(i)\}$
    $Q = Q \cup \{j : j \in N_\epsilon(i) \wedge j \in U\}$
    while $Q \neq \emptyset$ **do**
      for $j \in Q$ **do**
        $U = U \setminus \{j\}$
        if $|N_\epsilon(j)| \geq MinPts$ **then**
          $C_{\text{open}} = C_{\text{open}} \cup \{l : l \in N_\epsilon(j)\}$
          $Q = Q \cup \{l : l \in N_\epsilon(j) \wedge l \in U\}$
          $Q = Q \setminus \{j\}$
        end if
      end for
    end while
  end if
  $C = C \cup \{C_{\text{open}}\}$
end while
if $\exists j : j \neq l \wedge i \in C_j \wedge i \in C_l$ **then** {Entferne doppelt zugeordnete Randpunkte}
  for all $j : j \neq l \wedge i \in C_j \wedge i \in C_l$ **do**
    $C_j = C_j \setminus \{i\}$
  end for
end if
$N = \{1, \ldots, n\} \setminus \bigcup_{C_l \in C} C_l$ {Füge Rauschpunkte hinzu}
$C = C \cup \bigcup_{i \in N} \{i\}$

Bildung der Zusammenhangskomponenten identisch zum Single-Linkage-Verfahren, das im folgenden Abschnitt als Zusammenhangskomponente eines Graphen beschrieben ist. DBSCAN ist also eine Weiterentwicklung des Single-Linkage-Verfahrens mit einer Spezialbehandlung von Rand- und Rauschpunkten.

DBSCAN ist ein weit verbreiter Algorithmus zur Clusterung großer Datensätze. Der Algorithmus bildet die Grundlage für eine Vielzahl an Erweiterungen, z. B. OPTICS (Ankerst et al., 1999).

Zusammenhangskomponenten eines Graphen (igraph)

Eine naheliegende Idee der Clusterbildung basierend auf einem Graphen ist die Verwendung von Zusammenhangskomponenten. Im Folgenden wird die Kurzform igraph verwendet, die nach dem R-Paket `igraph` (Csardi und Nepusz, 2006) benannt ist, in dem die Funktion `clusters` die Zusammenhangskomponenten eines ungerichteten Graphen erstellt.

Zur Definition einer Zusammenhangskomponente werden Grundlagen der Graphentheorie benötigt (Kolaczyk und Csárdi, 2014, S.21 ff.). Ein Weg von einem Knoten v_0 nach einem Knoten v_l ist eine Folge von Knoten und Kanten $v_0, e_1, v_1, e_2, \ldots, v_{l-1}, e_l, v_l$. Dabei sind die Endpunkte der Kanten e_i die Knoten $\{v_{i-1}, v_i\}$. Ein Knoten ist von einem anderen Knoten erreichbar, falls es einen Weg vom einen zum anderen Knoten gibt. Ein Graph heißt verbunden, falls jeder Knoten voneinander erreichbar ist.

Die R-Funktion `clusters` bestimmt die Zusammenhangskomponenten eines Graphen, also einen maximal verbundenen Teilgraph. Das Hinzufügen eines beliebigen weiteren Knoten zu diesem Teilgraph führt dazu, dass der Teilgraph nicht mehr verbunden ist. Da es sich um einen ungerichteten Graphen handelt, wird nicht zwischen schwach und stark verbundenen Komponenten unterschieden. Für einen ungerichteten Graph liegt eine schwach verbundene Komponente vor, falls ein zugehöriger Graph, in dem die Richtung der Kanten keine Rolle spielt, verbunden ist. Die Komponente ist stark verbunden, falls jeder Knoten von einem anderen Knoten über einen gerichteten Weg erreichbar ist. Die Zusammenhangskomponente eines ungerichteten Graphen ist wie eine schwach verbundene Komponente eines gerichteten Graphen implementiert. Eine einfache Breitensuche (engl. breadth-first search) findet alle Knoten innerhalb einer Zusammenhangskomponente.

Der Pseudocode des igraph-Algorithmus (Algorithmus 7) ist eine vereinfachte Version von DBSCAN (Algorithmus 6) für $\epsilon =$ `cdis` und $MinPts = 1$. Die Bildung der Zusammenhangskomponenten ist außerdem identisch zum hierarchischen

Algorithmus 7 : igraph

Eingabe : $n \times n$ Distanzmatrix D mit Einträgen $d_{ij} = d(S_i, S_j)$ {für $N_{\texttt{cdis}}$ nötig}
Schwellenwert `cdis`

Ausgabe : Menge aller Cluster $C = \{C_1, \ldots, C_k\}$

1: $U = \{1, \ldots, n\}$ {unzugewiesene Punkte}
2: $Q = \emptyset$ {Warteschlange}
3: $C = \emptyset$ {Menge der Cluster}
4: **while** $U \neq \emptyset$ **do**
5: wähle $i \in U$
6: $U = U \setminus \{i\}$
7: $C_{\text{open}} = \{j : j \in N_{\texttt{cdis}}(i)\}$
8: $Q = Q \cup \{j : j \in N_{\texttt{cdis}}(i) \wedge j \in U\}$
9: **while** $Q \neq \emptyset$ **do**
10: **for** $j \in Q$ **do**
11: $U = U \setminus \{j\}$
12: $C_{\text{open}} = C_{\text{open}} \cup \{l : l \in N_{\texttt{cdis}}(j)\}$
13: $Q = Q \cup \{l : l \in N_{\texttt{cdis}}(j) \wedge l \in U\}$
14: $Q = Q \setminus \{j\}$
15: **end for**
16: **end while**
17: $C = C \cup \{C_{\text{open}}\}$
18: **end while**

Single-Linkage-Verfahren (s. o.) (Gross et al., 2013, S. 1332). Ausgehend von einem zufälligen Ausgangsknoten i werden Verbindungen zu Knoten in der `cdis`-Nachbarschaft gebildet. Ausgehend von den Knoten entstehen iterativ weitere Verzweigungen, wenn sich mindestens ein Knoten in deren `cdis`-Nachbarschaft befindet. Sobald keine neuen Wege mehr im Teilgraphen gefunden werden, ist das Cluster vollständig und das Prozedere beginnt mit einem neuen zufälligen Ausgangsknoten. Der igraph-Algorithmus ist beendet, sobald alle Objekte einem Cluster zugewiesen wurden. Im Vergleich zum DBSCAN-Algorithmus gibt es keine Rauschpunkte, da keine Mindestgröße für ein Cluster gefordert wird. Einzelcluster (engl. singletons) gibt es auf alle Fälle, da sich aus $d_{ii} = 0$ ergibt, dass eine beliebige Kante (i, i) immer

vorhanden ist. Außerdem gibt es auch keine Randpunkte, die mehreren Clustern zugeordnet werden.

Neighbor Clustering

Beim hierarchischen Clustern mit Single-Link, genauso beim identischen igraph-Algorithmus und DBSCAN im Spezialfall $MinPts = 1$ und $MinPts = 2$, tritt das Phänomen der Verkettung auf (Hastie et al., 2009, S. 524). Es werden Cluster gebildet, in denen sich eine Kette von Objekten mit paarweise kleinen Abständen bildet. Das Maximum der paarweisen Distanzen in jenen Clustern liegt jedoch sehr hoch, nahe 1. DBSCAN definiert Rauschpunkte, die in der implementierten Version Einzelcluster bilden, um diesen Effekt zu verhindern (Aggarwal, 2015, S. 171).

Das neue graphenbasierte Neighbor Clustering, das den Verkettungseffekt verhindert, wurde in Rieder et al. (2017b) vorgestellt. Die Grundidee lautet, dass ein Objekt im Zentrum eines Clusters viele Nachbarn in einer Umgebung mit Radius

Algorithmus 8 : Neighbor Clustering

Eingabe : $n \times n$ Distanzmatrix D mit Einträgen $d_{ij} = d(S_i, S_j)$ {für N_c nötig}
Schwellenwert c

Ausgabe : Menge der Cluster $C = \{C_1, \ldots, C_k\}$

1: $U = \{1, \ldots, n\}$ {Unzugewiesene Punkte}
2: $C = \emptyset$
3: **for all** $i = 1, \ldots, n$ **do**
4: **if** $|N_c(i)| = 1$ **then**
5: $U = U \setminus \{i\}$ {Entferne Einzelcluster}
6: **end if**
7: **end for**
8: **while** $U \neq \emptyset$ **do**
9: wähle $j \in M = \{j : |N_c(j)| = \max_{i \in U} |N_c(i)|\}$
10: $C_{\text{open}} = U \cap N_c(j)$
11: $U = U \setminus C_{\text{open}}$
12: $C = C \cup \{C_{\text{open}}\}$
13: **end while**
14: $N = \{1, \ldots, n\} \setminus \bigcup_{C_k \in C} C_k$
15: $C = C \cup \bigcup_{i \in N} \{i\}$ {Füge Einzelcluster hinzu}

c haben sollte. Der Pseudocode ist in Algorithmus 8 dargestellt. Zunächst wird für ein festes c für jedes Objekt die Anzahl der Nachbarn bestimmt. Objekte mit nur einem Nachbarn, sich selbst, werden vorweg als Einzelcluster markiert. Die größte Nachbarschaft, also das Objekt, das die meisten Nachbarn besitzt, inklusive seiner Nachbarn, bildet ein neues Cluster. Das Vorgehen wird immer wieder für alle Objekte wiederholt, die noch nicht zugeordnet wurden. Erneut wird ein Cluster basierend auf der größten Nachbarschaft gebildet. Sind alle Objekte einem Cluster zugeordnet, ist die Clusterbildung abgeschlossen. Die maximale Distanz von jedem Objekt zu seinem Clusterzentrum ist durch c beschränkt, sodass das Maximum der paarweisen Distanzen maximal doppelt so groß wie c ist. Wird c also nicht allzu groß gewählt, wird ein Verkettungseffekt nicht auftreten.

MS-Cluster und PRIDE Cluster

Bei MS-Cluster (Frank, 2008; Frank et al., 2008) handelt es sich um einen approximativen hierarchischen Clusteralgorithmus. Wie die zuvor beschriebenen Algorithmen basiert die Clusterbildung auf der Distanz bzw. Ähnlichkeit von Massenspektren. Die Originalspektren müssen allerdings für die Clusterbildung vorliegen, da iterativ Ähnlichkeiten von Clustern über deren Repräsentanten, sogenannten künstlichen Konsensusspektren, gebildet werden.

Wichtige Bestandteile des Clusteralgorithmus MS-Cluster sind die Berechnung der Ähnlichkeit von Spektren und die Repräsentantenauswahl in den Clustern. Als Ähnlichkeitsmaß wird die in Kapitel 3.1 beschriebene Kosinus-Ähnlichkeit gewählt. Die Implementierung enthält jedoch eine Reihe an Vorverarbeitungsschritten der Spektren. Die ersten beiden Schritte entsprechen einer `topn`-Auswahl und einem Binning aus dem DISMS2-Algorithmus (Kapitel 3.2.1), sodass Peaks, die maximal 0.5 Dalton voneinander entfernt liegen, zusammengefasst werden. Es wird empfohlen den Parameter `topn` so zu wählen, dass in einem Massenfenster von 1000 Dalton 15 Peaks liegen. Um zu verhindern, dass eine geringe Anzahl an hohen Peaks die Ähnlichkeitsbestimmung zu stark beeinflusst, wird eine bestimmte Methode zur Skalierung der Peakintensitäten befürwortet. Die Peakintensitäten werden zunächst so normalisiert, dass die Gesamtintensität des Spektrums 1000 beträgt. Anschließend wird der natürliche Logarithmus dieser Intensitätswerte gebildet. Im Folgenden bezeichnet $s_{\text{MS-Cluster}}(S_i, S_j)$ die in MS-Cluster implementierte Ähnlichkeit zweier Spektren (oder Repräsentanten) S_i und S_j.

Eine Repräsentantenauswahl je Cluster spielt in dem MS-Cluster-Algorithmus eine große Rolle, da so vermieden wird vorab die Distanzmatrix aller Spektren zu

berechnen. Stattdessen wird je nach Bedarf nur die Ähnlichkeit von Paaren von Clustern über die Ähnlichkeit derer Repräsentanten bestimmt. Ein sogenanntes Konsensusspektrum dient als Repräsentant eines Clusters mit mehr als einem Clustermitglied. Einzelcluster werden selbstverständlich durch die Originalspektren repräsentiert. Zunächst werden die Peaks aller Spektren eines Clusters in einem neuen Spektrum zusammengefasst. Ist die Massendifferenz aufeinanderfolgender Peaks kleiner als eine vorgegebene Toleranz, fusionieren diese Peaks zu einem Peak. Die Masse des neuen Peaks entspricht dem gewichteten Mittel der einzelnen Peakmassen. Die Intensität des neuen Peaks wird über die Summe der Einzelintensitäten gebildet. In mehreren Iterationen fusionieren Peaks mit ansteigender Toleranz, die maximal 0.4 Dalton beträgt. Die Intensitäten der Peaks werden multipliziert mit einem sogenannten Skalierungsfaktor $\alpha = 0.95 + 0.05(1 + h)$, der von h, dem Quotienten der Anzahl der fusionierten Originalpeaks und der Gesamtanzahl an Spektren im Cluster abhängt. Der Zusammenhang von Werten zwischen 0 und 1 für h und α ist in Abbildung 3.3 dargestellt. Der Zusammenhang von h und α wird durch eine bei Null beginnende exponentiell steigende Linie beschrieben, die für $h = 1$ bei $\alpha = 2.55$ endet. Schließlich werden noch kleine Peaks ausgeschlossen, indem in einem gleitenden Fenster der Breite 100 Dalton nur die 5 höchsten Peaks beibehalten werden. Im Folgenden wird der Repräsentant des Clusters C_i Konsensusspektrum R_i genannt.

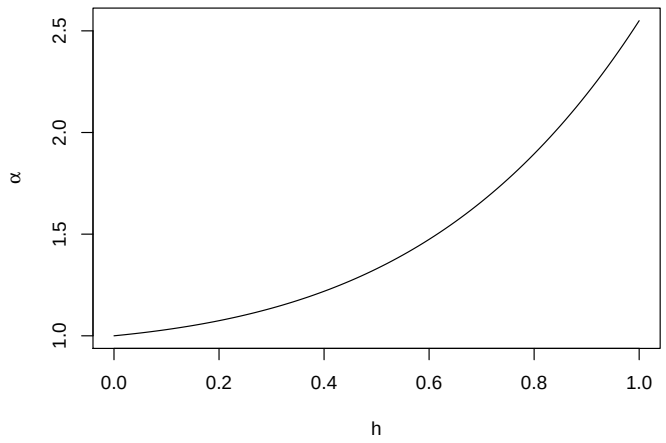

Abbildung 3.3: Zusammenhang der Wertebereiche zwischen dem Quotienten h und dem Skalierungsfaktor α bei der Konsensusspektrenberechnung im MS-Cluster-Algorithmus

Der Pseudocode von MS-Cluster ist in Algorithmus 9 dargestellt. Zwei Parameter, der Schwellenwert $\tau_{\min}$ für die Bewertung der Ähnlichkeit von zwei Konsensusspektren und die Anzahl der Runden r, können frei gewählt werden. Zunächst werden Einzelcluster $C_1, \ldots, C_n$ gebildet. Abhängig von den beiden Parametern $\tau_{\min}$

Algorithmus 9 : MS-Cluster (Frank et al., 2008, Abbildung 1)

Eingabe : Spektren $S_1, \ldots, S_n$
Schwellenwert $\tau_{\min}$
Anzahl an Runden r

Ausgabe : Menge der Cluster $C = \{C_1, \ldots, C_k\}$

1: $\delta = \frac{1-\tau_{\min}}{r}$
2: $C = \{\{1\}, \ldots, \{n\}\}$
3: $\tau = 1$
4: **do** r mal
5: $\tau = \tau - \delta$
6: **for all** $C_j \in C$ **do**
7: **for** $C_l \in C, l < j$ **do**
8: **if** $s_{\text{MS-Cluster}}(R_l, R_j) \geq \tau$ **then**
9: $C = C \setminus \{C_j, C_l\}$
10: $C_l = C_j \cup C_l$ {Füge Cluster zusammen}
11: $C = C \cup \{C_l\}$
12: **end if**
13: **end for**
14: **end for**

und r wird ein $\delta = (1 - \tau_{\min})/r$ bestimmt, um das der Schwellenwert τ in jeder Runde verringert wird. Zu Beginn nimmt τ den maximalen Wert 1 an. Der Algorithmus endet nach exakt r Iterationen. In jedem Schritt wird zunächst τ aktualisiert, indem die Größe δ abgezogen wird. Es werden nacheinander alle Cluster durchlaufen. Zu jedem Cluster C_j werden nur vorangehende Cluster $C_k (k < j)$ untersucht. Ist die oben beschriebene Ähnlichkeit $s_{\text{MS-Cluster}}$ von zwei Clusterrepräsentanten, den sogenannten Konsensusspektren R_k und R_j, mindestens so groß wie der Schwellenwert τ, fusioniert das vorangehende Cluster C_k mit dem Cluster C_j.

In dem Programm MS-Cluster v2 algorithm (Frank et al., 2011) werden der Schwellenwert für die Ähnlichkeit `similarity` und die Anzahl der Runden `rounds` festgelegt. Abweichend von Algorithmus 9 wird die erste Runde mit $\tau = 1$ durchgeführt und für $\delta = (1 - \texttt{similarity})/(\texttt{rounds} - 1)$ folgen weitere `rounds` - 1 Runden. Für `similarity`=0.8 und `rounds`=5 ergibt sich beispielsweise $\tau = 1.00, 0.95, 0.90, 0.85, 0.80$.

Die Implementierung des Algorithmus nutzt zwei Heuristiken aus, sodass die Anzahl der Ähnlichkeitsberechnungen und somit die Laufzeit verringert wird. Erstens wird die Ähnlichkeit nur dann berechnet, wenn es zwischen den jeweils fünf höchsten Peaks eine Übereinstimmung gibt. Frank (2008) beobachtete bei seiner Datenauswertung, dass Spektren, die vom gleichen Peptid stammen, in der Regel mindestens einen gemeinsamen Peak besitzen und es selten Fälle gab, in denen ein gemeinsamer Peak bei Spektren unterschiedlicher Peptide vorlag. Zweitens wurden Paare von Clustern mit extrem geringer Ähnlichkeit vermerkt, sodass sie in folgenden Runden des Algorithmus nicht erneut ausgewertet wurden.

PRIDE Cluster (Griss et al., 2013) basiert auf MS-Cluster. Die Ähnlichkeitsberechnung und die Bildung der Konsensusspektren wurden nicht verändert, aber es gibt einige Modifikationen im Algorithmus. Im Gegensatz zu MS-Cluster wird die Qualität der Massenspektren mit dem Signal-zu-Rauschen-Verhältnis von SpectraST (Lam et al., 2008) bewertet. Zusätzlich kann ein Cluster auch getrennt werden, wenn neue Spektren hinzugefügt werden. Das Zusammenfügen von Clustern erfolgt nur, wenn Cluster die größte Ähnlichkeit besitzen. Dadurch wird die Clusterbildung im Vergleich zu MS-Cluster verbessert, da im MS-Cluster-Algorithmus die iterativ zuerst identifizierten Cluster zusammengefügt werden, deren Ähnlichkeit einen festen Schwellenwert übersteigt.

Der Pseudocode von PRIDE Cluster ist in Algorithmus 10 dargestellt. Vor der Clusterbildung werden die Spektren bezüglich ihrer Qualität sortiert. Die Qualität Q eines Spektrums S_i wird über das verwendete Maß in SpectraST (Lam et al., 2008) definiert:

$$Q(S_i) = \frac{\frac{1}{5} \sum_{k=p-5}^{p-1} I_{(k)}}{I_{\text{med}}}$$

Die Qualität nimmt höhere Werte an, wenn der Mittelwert von fünf Intensitäten, die sechs größten Intensitäten abzüglich der größten Intensität, im Vergleich zum Median der p Intensitäten I_{med} groß ist.

Bei der Clusterbildung werden drei Schritte, die Clusterbildung der Spektren, das Zusammenfügen von Clustern und die Entfernung nicht passender Spektren, so lange wiederholt bis entweder die vorab festgelegte maximale Anzahl an Runden N erreicht ist oder je eine höhere Ähnlichkeit aller Spektren zu ihrem Clusterrepräsentanten vorliegt als den gewählten Schwellenwert t.

Am Anfang bildet jedes Spektrum ein Einzelcluster und der Zähler r wird auf Null gesetzt. Er erhöht sich zu Beginn jeder Runde um 1, sodass die Bedingung der

Algorithmus 10 : PRIDE Cluster

Eingabe : MS/MS Spektren $S_1, \ldots, S_n$ (sortiert nach Qualität)
Schwellenwert t
Maximale Anzahl an Runden N

Ausgabe : Menge der Cluster $C = \{C_1, \ldots, C_k\}$

1: r = 0
2: $C = \{\{1\}, \ldots, \{n\}\}$
3: **repeat**
4: r = r + 1
5: **for all** $i \in \{1, \ldots, n\}$ **do** {Clusterbildung der Spektren}
6: **if** $\exists j : s_{\text{MS-Cluster}}(S_i, R_j) > t$ **then**
7: $m : s_{\text{MS-Cluster}}(S_i, R_m) = \max_j s_{\text{MS-Cluster}}(S_i, R_j)$
8: $l(\neq m) : S_i \in C_l$
9: $C = C \setminus \{C_m, C_l\}$
10: $C_m = C_m \cup \{i\}$
11: $C_l = C_l \setminus \{i\}$
12: $C = C \cup \{C_m, C_l\}$
13: **else**
14: **if** $i \in C_m$ **then**
15: $C_m = C_m \setminus \{i\}$
16: $C = C \cup \{\{i\}\}$
17: **end if**
18: **end if**
19: **end for**
20: **for all** $C_j \in C$ **do** {Zusammenfügen von Clustern}
21: **if** $\exists l : s_{\text{MS-Cluster}}(R_j, R_l) > t$ **then**
22: $C = C \setminus \{C_j, C_l\}$
23: $C_j = C_j \cup C_l$
24: $C = C \cup \{C_j\}$
25: **end if**
26: **end for**
27: **for all** $i \in \{1, \ldots, n\}$ **do** {Entfernung nicht passender Spektren}
28: $j : i \in C_j$
29: **if** $s_{\text{MS-Cluster}}(S_i, R_j) \leq t$ **then**
30: $C = C \setminus \{C_j\}$
31: $C_j = C_j \setminus \{i\}$
32: $C = C \cup \{C_j, \{i\}\}$
33: **end if**
34: **end for**
35: **until** $r = N$ oder $\forall i \in \{1, \ldots, n\}, j(i \in C_j) : s_{\text{MS-Cluster}}(S_i, R_j) > t$

repeat-Schleife, dass maximal N Runden absolviert werden, geprüft werden kann. Jede Runde beginnt mit der Clusterbildung der Spektren. Dazu wird für jedes Spektrum geprüft, ob es Cluster gibt, dessen Konsensusspektrum eine Ähnlichkeit, die höher als t ist, zu jenem Spektrum aufweist. Unter allen Clustern, die diese Bedingung erfüllen, wird das Cluster mit der größten Ähnlichkeit gewählt und das Spektrum wechselt seine Zugehörigkeit zu diesem Cluster. Ansonsten bildet das Spektrum ein Einzelcluster.

Im zweiten Schritt jeder Runde werden Cluster zusammengefügt, falls die paarweise Ähnlichkeit der Repräsentanten größer als der Schwellenwert t ist. Im dritten Schritt werden Spektren entfernt, die nicht gut genug zu ihrem Cluster passen. Liegt die Ähnlichkeit eines Spektrums zu deren Konsensusspektrum maximal bei t, so wird das Spektrum entfernt und es bildet ein Einzelcluster.

Im Vergleich zu MS-Cluster werden Ähnlichkeitsberechnungen benötigt. Die Implementierung des PRIDE Clusteralgorithmus nutzt eine parallele Berechnung der Cluster. In der Java-Applikation spectra-cluster-cli (Griss et al., 2016) wird dazu zusätzlich der Parameter `precursor_tolerance` gewählt, der die Precursormassenfenstergröße beschränkt. Spektren werden in Gruppen je nach Precursormasse aufgeteilt. Um alle Spektren eines Peptids in einer Gruppe zu erfassen, werden um die Hälfte der Precursormasse überlappende Gruppen gebildet. Jedes Spektrum wird daher je zweimal geclustert. Anschließend werden die Cluster zusammengefügt, sodass jedes Spektrum genau einem Cluster angehört.

3.3.2 Bewertungsmaße der Qualität von Clusterlösungen

Zur Beurteilung der Validität von Clustern werden drei Arten von Kriterien unterschieden, nämlich interne, externe und relative Kriterien (Halkidi et al., 2001). Externe Kriterien ziehen bereits bekannte Clusterstruktren zurate, die möglichst nah an den wahren, jedoch unbekannten Gruppenaufteilungen sind. Interne Kriterien basieren allein auf Informationen aus den Clusterlösungen. Relative Kriterien beziehen sich auf den Vergleich verschiedener Clusterlösungen, die beispielsweise durch andere Parametereinstellungen in einem interessierenden Clusterverfahren erzeugt werden.

In Rieder et al. (2017b) werden Bewertungsmaße vorgestellt, die größtenteils bereits in anderen massenspektrometrischen Analysen zur Clusterbewertung verwendet wurden. Eine R-Implementierung dieser Maße ist frei verfügbar (`https://www.statistik.tu-dortmund.de/genetics-publications-clusspec.html`). Im Folgenden werden diese und andere interne, externe und relative Kriterien vorgestellt.

Interne Kriterien können speziell für hierarchische Clusterlösungen verwendet werden. Der kophenetische Korrelationskoeffizient wird verwendet, um die hierarchische Clusterlösung mit der grafischen Darstellung, dem sogenannten Dendrogramm (siehe Kapitel 3.4.1), zu vergleichen. Es gibt auch allgemeine Kriterien, wie beispielsweise den Dunn-Index oder Silhouetten-Index. In Rieder et al. (2017b) werden zwei Indizes, Proportion of clustered spectra und spectra remaining, benannt, die zur Kategorie der internen Kriterien gehören, jedoch je in Bezug auf ein externes Kriterium, interpretiert werden (siehe externe Kriterien).

Beim hierarchischen Clustern ist die kophenetische Distanz c_{ij} zwischen zwei Spektren i und j durch die Distanz zwischen den Gruppen definiert, bei welcher die Spektren i und j in einem Cluster verbunden werden. Der kophenetische Korrelationskoeffizient r_{koph} (Hastie et al., 2009, S.522f.) gibt die Korrelation zwischen den kophenetischen Distanzen c_{ij} und den Originaldistanzen d_{ij} von n Spektren mit mittlerer kophenetischer Distanz $\bar{c} = \frac{2}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} c_{ij}$ und mittlerer Originaldistanz $\bar{d} = \frac{2}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} d_{ij}$ an:

$$r_{\text{koph}} = \frac{\sum_{i=2}^{n} \sum_{j=1}^{i-1} (c_{ij} - \bar{c}) \left(d_{ij} - \bar{d}\right)}{\sqrt{\left(\sum_{i=2}^{n} \sum_{j=1}^{i-1} (c_{ij} - \bar{c})^2\right) \left(\sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(d_{ij} - \bar{d}\right)^2\right)}}$$

Die Begriffe Kompaktheit (engl. compactness) und Isolation (engl. separation) sind Ziele, die von Clustermethoden verfolgt werden (Celebi und Aydin, 2016). Ein Beispiel für Kompaktheit ist, dass die Distanzen der Objekte innerhalb der Cluster klein sein müssen. Dies kann durch die Summe der paarweisen Distanzen innerhalb jedes Clusters gemessen werden. Isolation hingegen besagt, dass zwischen Objekten, die aus unterschiedlichen Clustern stammen, größere Distanzen bestehen. Der Dunn-Index DI spiegelt Kompaktheit und Isolation von Clustern wider:

$$DI = \frac{\min\limits_{i,j=1,\ldots,k} \{ \min\limits_{p \in C_i, q \in C_j} d_{p,q} \}}{\max\limits_{l=1,\ldots,k} \{ \max\limits_{p,q \in C_k} d_{p,q} \}}$$

Im Zähler wird die minimale Distanz über alle Clusterpaare bestimmt, wobei die Distanz zwischen zwei Clustern die minimale Distanz zwischen einem Spektrum aus dem ersten und einem Spektrum aus dem zweiten Cluster ist. Im Nenner wird die maximale Breite aller Cluster berechnet. Innerhalb eines Clusters wird der maximale Abstand zwischen zwei Spektren ermittelt. Von all jenen Abständen wird dann das Maximum bestimmt.

Rousseeuw (1987) hat sogenannte Silhouetten vorgestellt, die beliebt sind, da sie auch zum Überblick der Clusterlösung in einer Grafik verwendet werden können. Sei $a_j = \frac{1}{n_i} \sum_{l,j \in C_i, l \neq j} d_{lj}$ die durchschnittliche Distanz zwischen Spektrum j und allen anderen Spektren in Cluster i. Falls j ein Einzelcluster bildet, so ist $a_j = 0$. Außerdem wird die Distanz $b_j = \min_{C_p, p \neq i} \frac{1}{n_p} \sum_{l \in C_p, j \in C_i} d_{lj}$ zwischen j und dem nächsten Cluster berechnet. Für die Silhouette sil von j gilt dann:

$$sil(j) = \frac{b_j - a_j}{\max\{a_j, b_j\}}$$

Ein Gütemaß für alle n Spektren ist der Silhouetten-Index I_{sil}, der Mittelwert aller Silhouetten (oder aller Silhouetten eines Clusters):

$$I_{sil} = \frac{1}{n} \sum_{j=1}^{n} sil(j)$$

Für externe Kriterien sind allgemein zusätzliche Informationen notwendig, die sich nicht direkt aus der Clusterlösung ergeben. Zur Validierung der Gruppen von Massenspektren ist eine Datenbankannotation oder alternativ auch eine De-Novo-Peptidsequenzierung hilfreich. Die Partition der Spektren, die sich aus deren Annotation ergibt, dient dabei als Leitbild. In Rieder et al. (2017b) werden der adjustierte Rand-Index, die Reinheit der Cluster, der mehrelementige Clusteranteil, der Spektrenanteil ohne häufigste Annotation, die Clusteranzahl in Relation zur Spektrenanzahl und der verbleibende Annotationsanteil näher erläutert. Spektren, bei denen die Annotation fehlt, werden bei diesen Maßzahlen nicht berücksichtigt. Die Kriterien können auch relativ interpretiert werden, um den Einfluss der Parametereinstellungen zu vergleichen.

Zum Vergleich von Partitionen, die sich aus geclusterten Objekten ergeben, eignet sich der adjustierte Rand-Index (Rand, 1971; Warrens, 2008). Die Klassifikation von Paaren von geclusterten Objekten wird für zwei gegebene Partitionen P_1 und P_2 bestimmt. Zum einen ist a (oder d) die Anzahl von Paaren die dem gleichen (oder unterschiedlichen) Clustern in beiden Partitionen, P_1 und P_2, zugeordnet werden. Zum anderen ist b (oder c) die Anzahl an Paaren, die dem gleichen Cluster in Partition P_1 (oder P_2) und verschiedenen Clustern in Partition P_2 (oder P_1) zugeordnet wird:

	gleiches Cluster (P_2)	verschiedene Cluster (P_2)
gleiches Cluster (P_1)	a	b
verschiedene Cluster (P_1)	c	d

Der Rand-Index RI (Rand, 1971; Warrens, 2008), der äquivalent zum Simple-Matching-Koeffizient ist, ist über den Anteil an konsistent klassifizierter Paare definiert:

$$RI = \frac{a+d}{a+b+c+d}$$

Da RI nicht jeden Wert im Intervall $[0,1]$ annimmt, gibt es eine Transformation, die den adjustierten Rand-Index (ARI) ergibt, der Cohen's Kappa (Hubert und Arabie, 1985; Warrens, 2008) entspricht:

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)} = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$$

Falls zwei Clusterlösungen identisch sind, ist der Wert von ARI gleich 1. Der Wert 0 ergibt sich, falls RI gleich dem Erwartungswert von unabhängigen Partitionen ist. Werte kleiner Null kommen zustande, wenn zwei Partitionen sich unähnlicher sind als unter Unabhängigkeit angenommen wird. Milligan und Cooper (1986) haben in einem Vergleich mit anderen externen Kriterien gezeigt, dass der weit verbreitete ARI am besten geeignet ist. ARI dient jedoch auch als relatives Kriterium. Eine Alternative zu ARI ist auch der Jaccard-Index (Kapitel 3.2).

Die Reinheit (engl. purity) eines Clusters oder einer gesamten Clusterlösung wird definiert über den größten Anteil an Spektren mit gleicher Peptidannotation (The und Käll, 2016; Griss et al., 2016). Angenommen, dass die häufigste Annotation $\tilde{A}_i$ in jedem Cluster korrekt ist, bezeichnet $\tilde{n}_i$ die Anzahl der richtig annotierten Spektren in Cluster i, $i = 1, \ldots, k$. Die Reinheit R ist definiert über den Quotienten der Summe der $\tilde{n}_i$ und n, der Gesamtanzahl an Spektren (Manning et al., 2008):

$$R = \frac{1}{n} \sum_{i=1}^{k} \tilde{n}_i$$

Die Reinheit kann für alle Spektren gemeinsam oder einzeln für Cluster (mit gleicher Clustergröße) bestimmt werden.

Es besteht ein Zielkonflikt zwischen einem großen Anteil an Spektren, die in mehrelementigen Clustern liegen, und einem geringen Anteil an Spektren, die bezüglich der Peptidannotation einem falschen Cluster zugeordnet werden. Eine gute Darstellung der Optimierung dieser zwei Kriterien ist ein Scatterplot (Griss et al., 2016). Auf der y-Achse wird der Anteil an Spektren in mehrelementigen Clustern (engl. proportion of spectra clustered (with at least one other spectrum)) abgetragen. Für den mehrelementigen Clusteranteil h_{clu} gilt:

$$h_{clu} = \frac{1}{n} \sum_{i=1}^{k} 1_{|C_i|>1}$$

Falls die Anzahl einelementiger Cluster (engl. singletons) klein ist, wird dieser Wert groß. Auf der x-Achse wird der Anteil an Spektren dargestellt, die im jeweiligen Cluster nicht zur Gruppe der häufigsten Annotation zählen (engl. proportion of spectra not identified as the most common annotation in the cluster). Für den Spektrenanteil ohne häufigste Annotation h_{inc} gilt:

$$h_{inc} = \frac{1}{n} \sum_{i=1}^{k} (|C_i| - \tilde{n}_i)$$

Die häufigste Annotation in einem Cluster verkörpert den Annotationsrepräsentanten des Clusters. Daher ist eine abweichende Annotation von geringerer Qualität. Bei der Vorstellung der Qualitätsmaße wurden die Begriffe *Rel. (incorrectly) clustered spectra* verwendet.

Zwei weitere konkurrierende Ziele enstehen durch die Auswahl der Clusterrepräsentanten und deren Annotation. Der Anteil verbliebener Annotationen im Vergleich zu Annotationen aller Spektren ($A_1, \ldots, A_n$) vor Durchführung der Clusterbildung (engl. retainment of identified spectra) liegt im Wettstreit mit dem Anteil der Clusterrepräsentanten an der Anzahl aller Spektren (engl. proportion of spectra remaining after clustering) (The und Käll, 2016). Auf der x-Achse wird in einem Scatterplot, $\frac{k}{n}$, die Clusteranzahl in Relation zu der Anzahl an Spektren dargestellt. Auf der zugehörigen y-Achse wird der Anteil verbliebener Annotationen abgetragen.

Die Relevanz der Kriterien wird beispielsweise bei einer Datenbanksuche deutlich. Ein kleiner Anteil an Clusterrepräsentanten ist hilfreich, um die Suche zu beschleunigen. Gleichzeitig sollte sich die Anzahl verbliebener Annotationen nicht verringern. Die Anzahl an Annotationen nach Durchführung der Clusterbildung wird durch die Anzahl unterschiedlicher Annotationen bezüglich der häufigsten Annotation je Cluster bestimmt. Daher wird der Anteil verbleibender Annotationen h_{rem} definiert als der Quotient jener Anzahl und der Gesamtanzahl an unterschiedlichen Annotationen vor der Clusterbildung:

$$h_{rem} = \frac{\left| \bigcup_{i=1}^{k} \tilde{A}_i \right|}{\left| \bigcup_{i=1}^{n} A_i \right|}$$

Falls zwei Peptide mit gleicher Häufigkeit auftreten, werden beide zur Menge der häufigsten Annotationen gezählt.

Wie bereits erwähnt wurde, können externe Kriterien wie ARI auch zum relativen Vergleich verwendet werden, sodass der Einfluss von Parametereinstellungen eines

Algorithmus analysiert werden kann. Zu den relativen Kriterien zählen auch interne Kriterien, wie der Dunn-Index.

3.4 Visualisierung von Abständen

Die zuvor beschriebenen Abstände von einzelnen Spektren und LC-MS/MS-Läufen können mit einer Reihe an Methoden dargestellt werden, die eine visuelle Interpretation von Clusterlösungen ermöglicht. Zunächst werden zwei bekannte statistische Verfahren, das Dendrogramm als Visualisierung einer hierarchischen Clusterlösung und die multidimensionale Skalierung vorgestellt. Anschließend werden phylognetische Bäume beschrieben, die in der Biodiversitätsforschung eine große Rolle spielen.

3.4.1 Dendrogramm

Eine hierarchische Clusterlösung lässt sich gut durch ein Dendrogramm visualisieren (Hastie et al., 2009, S. 521f.). Dazu wird ein verwurzelter binärer Baum erstellt. Alle Clusterobjekte bilden die Wurzel, von dem iterativ je zwei innere Knoten abgehen. Die Höhe eines inneren Knotens ist proportional zum Wert der Distanz zwischen den Gruppen, die an diesem Knoten vereinigt werden. Die Endknoten, auch Blätter genannt, repräsentieren die Clusterobjekte und werden auf der Höhe 0 dargestellt.

In Abbildung 3.4 ist ein Beispiel für ein Dendrogramm von acht Objekten dargestellt. Ein Abschneiden des Dendrogramms auf der Höhe $h = 0.6$ ergibt eine Clusterlösung mit vier Clustern $C_1 = \{4\}$, $C_2 = \{8\}$, $C_3 = \{2\}$ und $C_4 = \{1, 3, 5, 6, 7\}$.

Ein Dendrogramm wird häufig eigenständig als grafische Zusammenfassung der Daten verwendet. Abhängig von der Linkage-Methode können die Ergebnisse jedoch variieren. Der kophenetische Korrelationskoeffizient (Kapitel 3.3.2) dient bei dieser Problematik zur Bewertung, ob die Distanzen durch die hierarchische Abbildung gut repräsentiert werden.

3.4.2 Multidimensionale Skalierung

Verfahren wie selbstorganisierende Karten oder Flächen haben das Ziel Datenpunkte in eine Mannigfaltigkeit mit kleinerer Dimension abzubilden (Hastie et al., 2009, S. 570ff.). Die multidimensionale Skalierung (engl. Multidimensional Scaling, MDS) ist ein Verfahren mit ähnlichem Ziel und bietet den Vorteil, dass für die Berechnung die Objekte nicht selbst sondern nur Abstände von Objekten benötigt werden. Daher ist MDS geeignet für die Darstellung der Ähnlichkeit einer Menge von Massenspektren

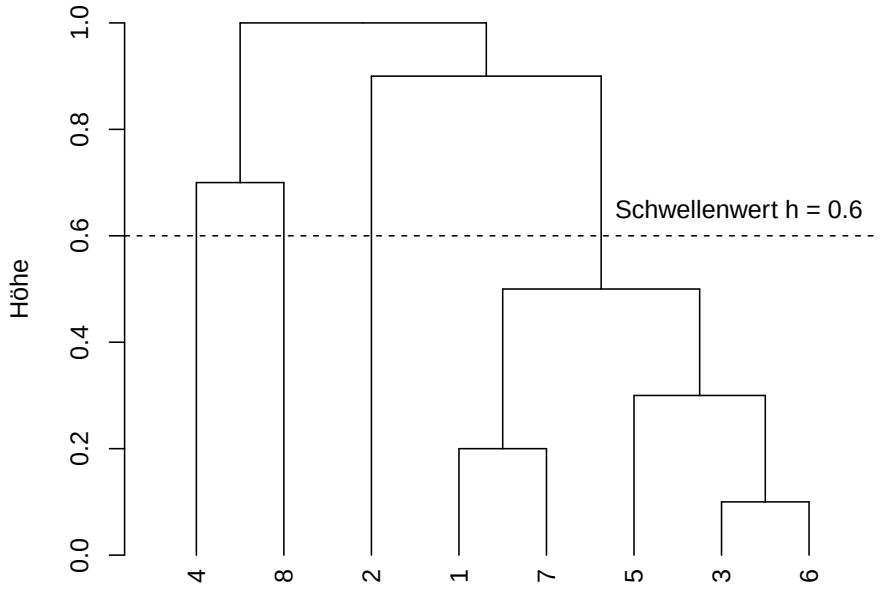

Abbildung 3.4: Beispiel für ein Dendrogramm für 8 geclusterte Objekte. Das Abschneiden des Dendrogramms auf der Höhe $h = 0.6$ führt zu vier Clustern $C_1 = \{4\}$, $C_2 = \{8\}$, $C_3 = \{2\}$ und $C_4 = \{1, 3, 5, 6, 7\}$.

in k Dimensionen. Für n Objekte werden n Werte $x_1, \ldots, x_n \in \mathbb{R}^k$ gesucht, die die Stressfunktion f_{stress} minimieren (Hastie et al., 2009, (14.98), S. 570):

$$f_{stress} = \sum_{i \neq j} \left(d_{ij} - ||x_i - x_j||_2\right)^2$$

Die Summe der Quadrate der Abweichungen zwischen Originaldistanzen und euklidischen Abständen der neuen Werte sollen minimal sein. Die Abstände in k Dimensionen sollen die ursprünglichen Abstände also so gut wie möglich widerspiegeln.

Synonyme für MDS sind Hauptkoordinatenanalyse, Torgerson-Scaling und Torgerson-Gower-Scaling (Gower, 1966; Torgerson, 1958). Ihre vorrangige Funktion ist die Visualisierung von Abstandsdaten. Ursprünglich wird sie als psychologisches Modell zur Analyse subjektiver Bewertungen der Ähnlichkeit von Objekten verwendet.

Die Bestimmung der Koordinaten der neuen Werte erfolgt über die quadrierte Distanzmatrix $D^{(2)}$ mit Einträgen d_{ij}^2. Die Matrix B wird erzeugt, die doppelt zentrierte quadrierte Distanzen enthält:

$$B = -\frac{1}{2} Z D^{(2)} Z \text{ mit } Z = I_n - \frac{1}{n} 1_n 1_n'.$$

I_n bezeichnet die n-dimensionale Einheitsmatrix und 1_n einen Vektor der Länge n mit Einträgen 1. Die doppelte Zentrierung bewirkt, dass der Mittelpunkt der MDS-

Anordnung der Ursprung wird. Für B wird anschließend die Eigenwertzerlegung in eine Diagonalmatrix Λ mit Eigenwerten von B und in die Matrix Q der zugehörigen Eigenvektoren bestimmt:

$$B = Q\Lambda Q'$$

Für die Koordinatenmatrix X der neuen Werte werden die k größten Eigenwerte, die größer als Null sind, in Λ_+ zusammengefasst und die zugehörigen k Spalten von Q in Q_+ kombiniert:

$$X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = Q_+\Lambda_+^{1/2}$$

Der beschriebene Algorithmus der klassischen MDS (Borg et al., 2012, S. 81ff.) ist in der R-Basisfunktion `cmdscale` implementiert.

Zur Visualisierung werden die $k = 2$ größten Eigenwerte verwendet. In Abbildung 3.5 ist ein Beispiel für eine MDS einer drei- bzw. vierdimensionalen Einheitsmatrix für $k = 2$ dargestellt. Für $n = 3$ bilden die Punkte die Ecken eines gleichseitigen Dreiecks. Bereits die Darstellung vier zueinander maximal entfernten Punkten im zweidimensionalen Raum führt zu Missverständnissen. Drei Punkte bilden erneut die Ecken eines Dreiecks, jedoch der vierte Punkt (Koordinaten -0.19, -0.2) liegt nur weit entfernt von zwei Punkten und sehr nah zu einem Punkt.

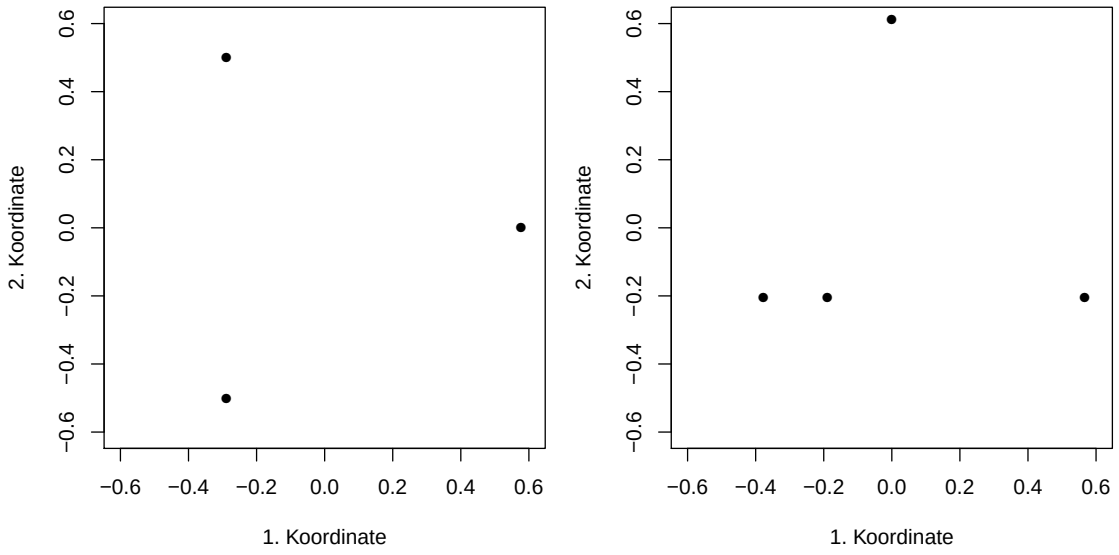

Abbildung 3.5: Beispiel für die Darstellung der Koordinaten einer MDS mit Parameter $k = 2$. Eine dreidimensionale (links) bzw. vierdimensionale (rechts) Einheitsmatrix wird in zwei Dimensionen abgebildet.

Zusätzlich ist bei der Interpretation einer MDS-Grafik das von Legendre und Legendre (1998, S. 465ff.) beschriebene Phänomen des Bogeneffekts zu beachten. Es wurde von ihnen beobachtet bei der Analyse von Umweltdaten, die beispielsweise von Umweltfaktoren kontrollierte Arten beinhalten. In der zweidimensionalen MDS-Grafik bilden die Punkte die Form von Bögen oder Hufeisen. Ein Bogeneffekt (engl. arch effect) liegt vor, wenn die Punkte einen Bogen bilden. Sind Punkte mit extremen Distanzen nach innen geklappt, so handelt es sich um den Hufeiseneffekt (engl. horseshoe effect).

3.4.3 Phylogenetischer Baum

Die Phylogenie befasst sich mit evolutionären Zusammenhängen. Phylogenetische Bäume dienen dabei zur Interpretation der evolutionären Differenzierung von Spezies. Es handelt sich um Baumgraphen, also zusammenhängende Graphen ohne Kreise. Die zu untersuchenden Arten bilden die terminalen Knoten und werden auch Blätter, d. h. ein Knoten ohne sogenannte Kindknoten, genannt. Es werden gewurzelte und ungewurzelte Bäume unterschieden. Als Synonym für einen gewurzelten Baum wird häufig auch der Begriff Dendrogramm (s. o.) verwendet. Bei einem gewurzelten Baum handelt es sich um einen gerichteten Baum. Dies bedeutet, dass nur ausgehend von der Wurzel ein gerichteter Weg zu allen anderen Knoten existiert. Es werden nur binäre Bäume betrachtet, d. h. jeder interne Knoten hat drei Kanten (Kolaczyk und Csárdi, 2014, S.24 ff., Hütt und Dehnert, 2016, S.218 ff.).

Zur Rekonstruktion phylogenetischer Bäume gibt es drei bekannte distanzbasierte Verfahren, UPGMA, den Neighbor-Joining-Algorithmus und die Minimum-Evolution-Methode. Unweighted Pair Group Method with Arithmetic mean (UPGMA) ist eine Variante der bereits erläuterten hierarchischen Clusteranalyse, nämlich dem Average-Linkage.

Der Neighbor-Joining-Algorithmus ist eine Split-Methode. Initial wird ein sternförmiger Baum gebildet. Innere Knote werden je nach Abständen zueinander hinzugefügt, sodass zwei Blätter mit dem Rest des Baumes verbunden sind. Beim Hinzufügen eines Knotens wird jenes Paar gewählt, für das die Summe der Kantenlängen minimal ist (Saitou und Nei, 1987; Studier und Keppler, 1988).

Die Minimum-Evolution-Methode dient ebenfalls zur Konstruktion eines ungewurzelten Baumes. Es werden zwei Schritte unterschieden, zuerst die Baumkonstruktion und anschließend die Modifikation des Baums zur Minimierung der Summe der Kantenlänge. Die Länge der Kanten zwischen zwei Blättern ist dabei proportional zu dem paarweisen Abstand. Gestartet wird mit drei Arten, zu denen sukzessive

Arten hinzugefügt werden. Dabei erfolgt die Anordnung so, dass die Summe der Kantenlänge minimal ist. Es werden Teilbäume vertauscht, zwischen denen genau drei Kanten liegen. Der Austausch weniger als drei Kanten führt zu äquivalenten Bäumen und der Austausch von mehr als drei Kanten wird aufgrund zu vieler Möglichkeiten nicht berücksichtigt (Desper und Gascuel, 2002).

Ein Beispiel für die Darstellung eines gewurzelten Baumes ist in der Diskussion des Dendrogramms (Abbildung 3.4) zu finden. Im Vergleich dazu ist in Abbildung 3.6 ein ungewurzelter Baum dargestellt. Beide Abbildungen beruhen auf der gleichen Distanzmatrix. Zur Erzeugung des ungewurzelten Baumes wurde eine Implementierung des Neighbor-Joining-Algorithmus aus dem R-Paket ape (Paradis et al., 2004) verwendet.

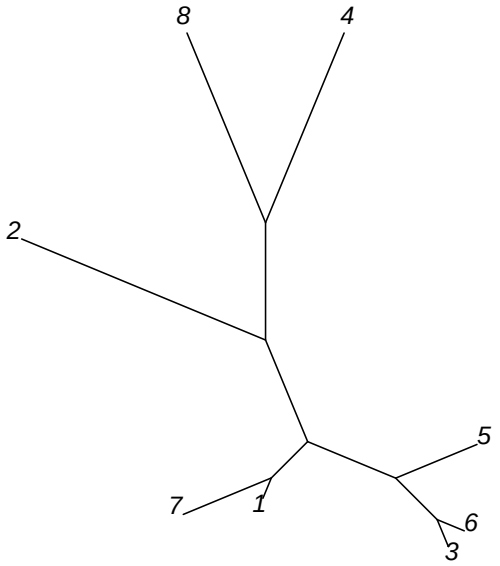

Abbildung 3.6: Neighbor-Joining-Algorithmus als Beispiel für einen ungewurzelten Baum (R-Paket ape).

Mithilfe der Funktion `dist.topo` aus dem R-Paket ape (Paradis et al., 2004) kann die topologische Distanz phylogenetischer Bäume berechnet werden. Es ist eine Modifikation der Methode von Robinson und Foulds (1981) implementiert (Rzhetsky und Nei, 1992).

4 Clusteranalyse von LC-MS/MS-Läufen

Dieses Kapitel handelt von einer neuen Methode DISMS2 zur Berechnung von Distanzen zwischen MS/MS-Läufen ohne Zuhilfenahme von Peptidannotationen. Zunächst werden zum Verständnis der folgenden Datenanalyse die in Kapitel 2.2 vorgestellten fünf Datensätze deskriptiv dargestellt (Abschnitt 4.1). Zusätzlich zu den Massenspektren werden auch Metadaten, wie beispielsweise Precursormasse, Precursorladung und Retentionszeit beschrieben. In Abschnitt 4.2 wird der Algorithmus DISMS2 angewendet auf die DISMS2- und Palmblad-Daten und die Parameter des Algorithmus werden optimiert. Die Distanzmatrizen werden mithilfe von Dendrogrammen, die eine hierarchische Clusterlösung visualisieren, und der multidimensionalen Skalierung grafisch dargestellt. Zur Bewertung der Abstände von Läufen erfolgt in Abschnitt 4.3 ein Vergleich zu Abständen basierend auf Peptidannotationen. Ergänzend zu dem DISMS2-Datensatz, für den bei einigen Arten eine Datenbanksuche vorliegt, wird der Foraminiferen-Datensatz analysiert, für den eine De-Novo-Sequenzierung nach dem Konzept von Blank-Landeshammer et al. (2017) durchgeführt wurde. Weitere Aspekte der Distanzberechnung werden in Abschnitt 4.4 untersucht. Die Messung und Vorverarbeitung der Massenspektren beeinflusst die Ergebnisse. Anhand der beiden Biodiversität-Datensätze erfolgt ein Vergleich von zwei Massenspektrometrie-Geräten. Bei der Vorverarbeitung wird die Wahl des Distanzmaßes und eine Selektion von Spektren und Peaks betrachtet. Insbesondere wird anhand des Foraminiferen-Datensatzes die bereits von Palmblad und Deelder angewandte Auswahl an Spektren mit intensivsten Signalen veranschaulicht. Abschließend werden beispielhaft Anmerkungen zu Laufzeiten und dem Speicherverbrauch gemacht (Abschnitt 4.5). Die in diesem Kapitel enthaltene Analyse des DISMS2-Datensatzes ist zum Teil bereits in Rieder et al. (2017a) veröffentlicht. Alle Berechnungen wurden mit der statistischen Software R (R Core Team, 2016, Version 3.4.2) erstellt.

4.1 Überblick der LC-MS/MS-Läufe

Zunächst werden die im RAW-Format vorliegenden Datensätze DISMS2, Foraminiferen, Biodiversität-Exactive und Biodiversität-Orbitrap mithilfe von MSConvert-GUI in das mzXML-Format konvertiert. Anschließend liegen alle Datensätze im mzXML-Format vor und können direkt mit dem R-Paket readMzXmlData eingelesen werden. Anhand der Retentionszeiten ist abzulesen, dass die Messung von einem Lauf des DISMS2-Datensatzes und der Biodiversität-Datensätze 125 Minuten dauerte. Im Foraminiferen-Datensatz war ursprünglich der sogenannte Waschschritt enthalten, in dem sich alle restlichen Peptide von der Chromatographiesäule lösen. Dieser wurde vor der Analyse entfernt, sodass nur noch Spektren berücksichtigt werden, die in den ersten 100 Minuten gemessen wurden.

In Tabelle 4.1 ist die Anzahl aller und im speziellen der annotierten MS/MS Spektren im DISMS2-Datensatz dargestellt. Zwischen 30 012 und 40 236 Spektren wurden je Lauf erzeugt. Für die beiden Foraminiferen-Proben (Ag, Al) und eine Radix-Probe (R2) ist keine Proteindatenbank verfügbar, sodass die Spektren der zugehörigen Läufe keine Annotation aufweisen. Bei den übrigen Läufen liegen zwischen 15 531 und 28 691 Peptidannotationen vor. Der Anteil variiert von 49.7% (R41) bis 71.3% (C2). Im Vergleich dazu sind im Foraminiferen-Datensatz unter Ausschluss des Waschschrittes insgesamt weniger Spektren, zwischen 17 176 und 30 238, generiert worden (Tabelle 4.2). Für alle Proben liegen je Lauf zwischen 2 257 und 5 473 De-Novo-Annotationen vor. Der Anteil annotierter Spektren ist also viel geringer als im DISMS2-Datensatz. Je Lauf weist nur ein Anteil von 13.1% (MaF142) bis 22.1% (AloZ152) Annotationen auf. Die Spektren der beiden Biodiversität-Datensätze wurden keiner Datenbanksuche unterzogen. Während mittels des Q-Exactive-Massenspektrometer zwischen 37 775 und 49 600 Spektren erzeugt werden, generiert das Orbitrap-Elite-Massenspektrometer nur 31 079 bis 36 418 Spektren (Tabelle A.4). Der Palmblad-Datensatz enthält nicht alle Spektren eines Laufs, sondern exakt jene 2 000 Spektren mit der höchsten Gesamtionenintensität.

Im DISMS2-Datensatz sind je Lauf durchschnittlich 49.6% (zwischen 43.6% und 55.0%) der Peptidannotationen auf ein Spektrum zurückzuführen. Die Anzahl der Spektren je Annotation ist rechtsschief verteilt. Gefolgt von einzelnen Annotationen sind doppelte Annotationen am häufigsten (im Mittel 11.4%). Maximal wurden in Lauf C2 101 Massenspektren mit dem gleichen Peptid (NMITGTSQADcAVLVVAcGTGEFEAGISK) annotiert. Die mediane Peptidlänge liegt bei 12. Insgesamt

Tabelle 4.1: Anzahl Spektren und Anzahl annotierter Spektren einzelner Läufe im DISMS2-Datensatz.

Lauf	Anzahl Spektren	Anzahl Annotationen
C1, C2, C3	39844, 40236, 39733	28256, 28691, 28085
D1, D2, D3	36410, 37282, 36238	21777, 22280, 21291
H1, H2, H3	36584, 37227, 38321	24348, 24392, 25129
M1, M2, M3	37470, 35647, 36770	24298, 22673, 23371
Y1, Y2, Y3	35372, 37000, 36557	16583, 17270, 16681
Ag1, Ag2, Ag3	31866, 32719, 34220	0, 0, 0
Al1, Al2, Al3	35802, 35350, 36149	0, 0, 0
R21, R22, R23	31495, 30012, 31758	0, 0, 0
R41, R42, R43	31227, 32387, 33662	15531, 16114, 16832

Tabelle 4.2: Anzahl Spektren und Anzahl de novo annotierter Spektren einzelner Läufe im Foraminiferen-Datensatz.

Lauf	Anzahl Spektren	Anzahl Annotationen
AgF141, AgF143, AgF144	27573, 19614, 22858	4934, 3719, 4055
AgF151, AgF152, AgF153	21905, 21248, 20112	4249, 4341, 4217
AlE161, AlE162, AlE163	22676, 20291, 21641	4538, 4301, 4583
AlEZ161, AlEZ162, AlEZ164	28928, 24448, 26061	5082, 5016, 4863
AlZ141, AlZ142, AlZ143	30238, 27396, 25831	5473, 5141, 4903
AlZ151, AlZ152, AlZ154	24869, 21266, 27713	4938, 3937, 5442
AloE161, AloE162, AloE163	25317, 19633, 20842	4728, 4005, 4206
AloEZ162, AloEZ163, AloEZ164	24018, 20443, 22601	4849, 4214, 4419
AloZ152, AloZ153, AloZ154	20052, 20106, 26079	4436, 4389, 5106
MaF141, MaF142, MaF143	19328, 17176, 26329	2566, 2257, 3651

gibt es Peptide der Länge 5 bis 49. Im Foraminiferen-Datensatz kommen je Lauf einzelne Annotationen häufiger vor (durchschnittlich 86.0%, zwischen 68.2% und 92.4%). Doppelte Annotationen bilden mit durchschnittlich 4.9% die zweitgrößte Gruppe. Die maximale Anzahl an Spektren, die mit dem gleichen Peptid je Lauf annotiert ist, beträgt hingegen nur 16. Ihre mediane Länge beträgt ebenfalls 12. Zwischen 4 und 30 Aminosäuren bilden ein Peptid.

Zu jeder Datenbankannotation im DISMS2-Datensatz liegt zusätzlich ein Score vor, der Auskunft über die Stärke der Übereinstimmung des Peptids mit dem Spektrum gibt. Eine bessere Übereinstimmung korrespondiert zu höheren Score-Werten

(siehe Abschnitt 2.2.2). In Abbildung 4.1 ist die rechtsschiefe Verteilung aller Scores dargestellt. Alle Scores liegen zwischen 10 und 250. Der Mittelwert beträgt 54.93.

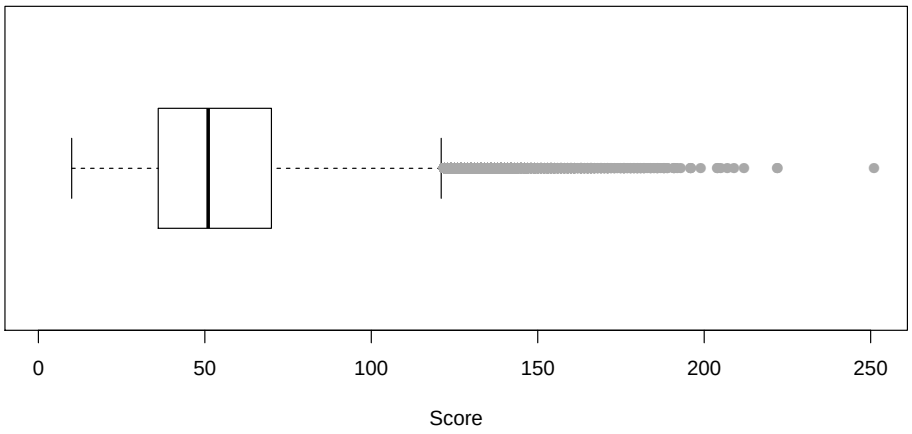

Abbildung 4.1: Boxplot des Scores (Mascot) der annotierten Spektren im DISMS2-Datensatz

Die einzelnen Tandem-Massenspektren werden durch detektierte Intensitäten zu bestimmten m/z-Werten, Masse-zu-Ladung-Verhältnissen, beschrieben. Die Anzahl der gemessenen Peaks gibt einen Hinweis auf die Länge der zugehörigen Peptide. Zum Vergleich der Verteilung der Anzahl an Peaks je Spektrum ist in Abbildung 4.2 je Datensatz des Leibniz-Projekts ein Histogramm dargestellt. Es fällt auf, dass nur bei Analysen mittels Q Exactive (Biodiversität-Exactive und DISMS2) mehrere Tausend Spektren mit weniger als 10 Peaks vorkommen. Im DISMS2-Datensatz liegt der Anteil bei 3.4%. Im Median besteht ein Spektrum aus 72 Peaks. Hingegen haben nur 0.3% der Spektren im Biodiversität-Exactive-Datensatz weniger als 10 Peaks. Im Vergleich der beiden Massenspektrometer zur Erstellung der Biodiversität-Datensätze sind deutliche Unterschiede in der Peakanzahl je Spektrum zu erkennen. Ein Spektrum im Datensatz Biodiversität-Exactive weist im Median 69 Peaks auf. Für den Biodiversität-Orbitrap-Datensatz beträgt die mediane Anzahl an Peaks hingegen 697. Auch die Verwendung der Fusion für zu einer höheren Anzahl an Peaks. Im Foraminiferen-Datensatz liegen im Median 213 Peaks je Spektrum vor.

Zur Bewertung der Qualität der Spektren wird außerdem das von Lam et al. (2008) vorgestellte Maß verwendet (siehe auch Abschnitt 3.3.1). Zum Vergleich der Verteilungen für einzelne Datensätze sind in Abbildung 4.3 Histogramme dargestellt. Die Verwendung des Q Exactive-Massenspektrometers führt zu ähnlichen Ergebnissen. Die Mittelwerte betragen 13.29 (Biodiversität-Exactive) und 15.53 (DISMS2). Die empirische Verteilung der Qualität ist im Foraminiferen-Datensatz

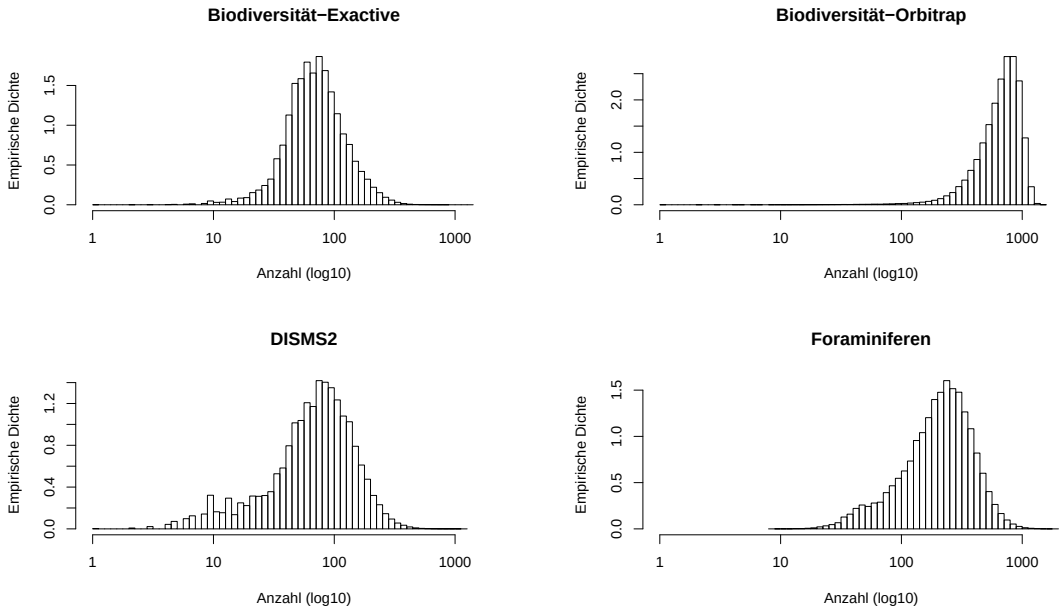

Abbildung 4.2: Histogramm der Anzahl an Peaks pro Spektrum auf einer log10-Skala in den vier Datensätzen des Leibniz-Projekts.

leicht und im Biodiversität-Orbitrap-Datensatz weiter nach rechts verschoben. Im Foraminiferen-Datensatz beträgt der Mittelwert 42.08. Die höchsten Werte wurden im Biodiversität-Orbitrap-Datensatz erzielt (Mittelwert 83.49). Zur Qualitätskontrolle eines Laufs wird die Intensität der gemessenen Peptidionen in Abhängigkeit der Retentionszeit dargestellt. In Abbildung 4.4 ist beispielhaft für Lauf H1 im DISMS2-Datensatz das sogenannte Gesamtionenchromatogramm dargestellt. Es ist zu erkennen, dass die detektierte Intensität der Ionen in den ersten Minuten stark ansteigt und ab ungefähr einer Stunde wieder rapide sinkt.

Die Fragmentionen, die in Tandem-Massenspektren detektiert werden, werden gebildet aus Precursorionen, also geladenen Peptiden. Masse und Ladung des Precursors sind hilfreiche Zusatzinformationen zu einem Tandem-Massenspektrum, die auch im DISMS2-Algorithmus verwendet werden. Bei der Erstellung der Datensätze im Leibniz-Projekt wurden die Massenspektrometer so eingestellt, dass die m/z-Werte des Precursors zwischen 300 und 1500 liegen (siehe auch Abbildung B.2). Precursorionen werden mehrfach geladen, sodass bei der Fragmentierung zwei Fragmentionen mit je mindestens einer Ladung entstehen. In Abbildung 4.5 sind die relativen Häufigkeiten der Precursorladungen je Lauf getrennt nach den vier Datensätzen mittels Boxplots dargestellt. Die Analyse mit dem Q Exactive-Massenspektrometer führt zu zwei-, drei- und vierfach geladenen Ionen. Im Biodiversität-Orbitrap-

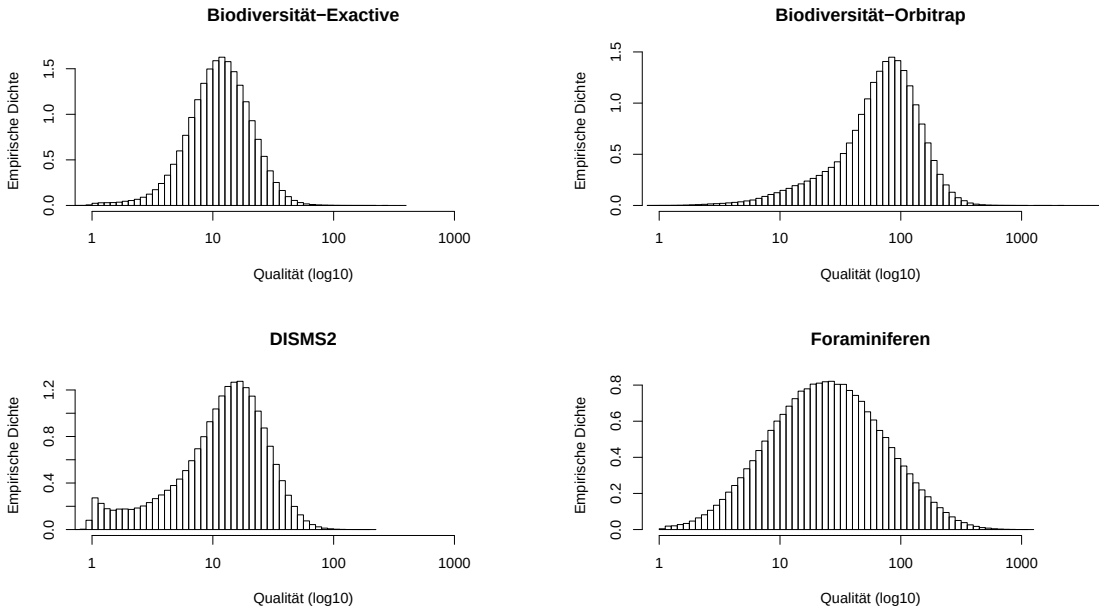

Abbildung 4.3: Histogramm der Qualität (SpectraST) auf einer log10-Skala in den vier Datensätzen des Leibniz-Projekts.

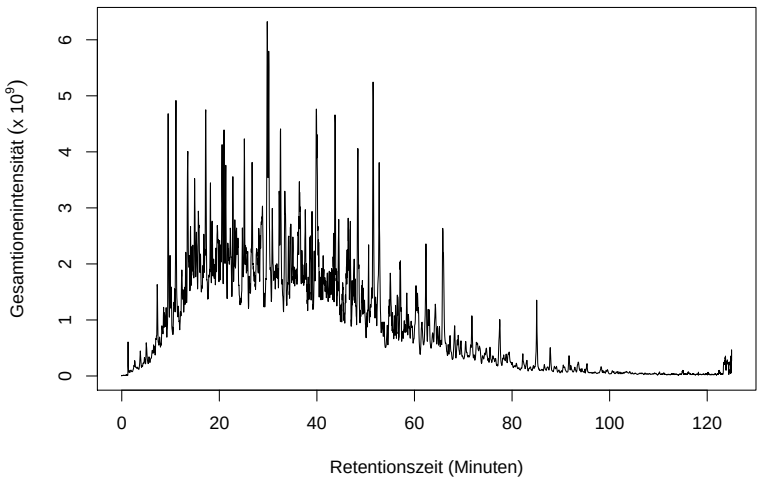

Abbildung 4.4: Gesamtionenchromatogramm des Laufs H1 im DISMS2-Datensatz.

und Foraminiferen-Datensatz kommen vereinzelt auch höhere Ladungen vor. Zweifach geladene Ionen bilden die häufigste Ausprägung. Im Median sind mehr als die Hälfte aller Ionen zweifach geladen. Die zweithäufigste Ausprägung bilden dreifach geladene Ionen. Etwa ein Drittel der Ionen ist im Median dreifach geladen.

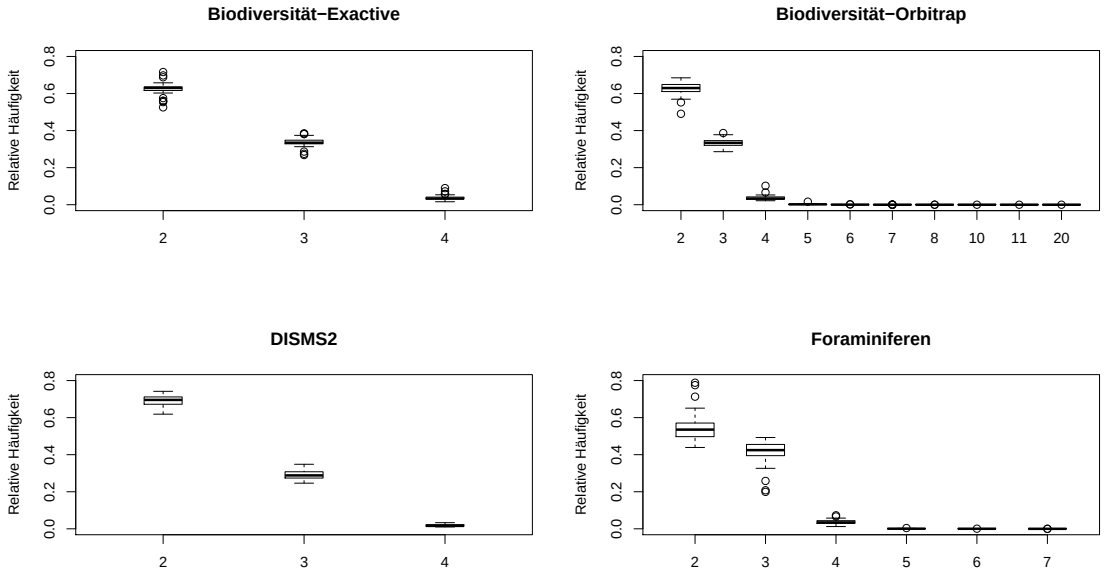

Abbildung 4.5: Boxplots der relativen Häufigkeiten der Precursorladung in den vier Datensätzen des Leibniz-Projekts.

Zusammenfassend ist festzustellen, dass die deskriptive Analyse der massenspektrometrischen Daten dem ersten Überblick dient. Der Anteil an Annotationen variiert stark je nach Art im DISMS2-Datensatz und ist allgemein höher als die De-Novo-Annotationen des Foraminiferen-Datensatzes. Selten sind mehrere Spektren mit gleicher Annotation. Die Peptidlänge beträgt im Median 12. Zugehörige Scores im DISMS2-Datensatz liegen im Mittel bei 54.93. Die Anzahl an Peaks und die Qualität der Spektren ist in den beiden Datensätzen DISMS2 und Biodiversität-Exactive kleiner als in den Datensätzen Biodiversität-Orbitrap und Foraminiferen. Die Precursorionen sind am häufigsten zweifach geladen und ihre m/z-Werte liegen zwischen 300 und 1500. Wie in Abschnitt 2.2.1 erwähnt, enthält der Palmblad-Datensatz je Lauf eine Auswahl an 2000 vorverarbeiteten Spektren mit je 50 Peaks. Die Ladung der Precursorionen ist nicht übermittelt und ihre m/z-Werte liegen zwischen 300 und 1300.

4.2 Optimierung und Visualisierung der Abstände von Läufen

Der in R implementierte DISMS2-Algorithmus wird zunächst angewandt auf alle 27 Läufe des DISMS2-Datensatzes. Ziel ist ein Vergleich der Proben von Mensch, Maus, Hefe, Fadenwurm, Fruchtfliege, zwei *Radix*- und zwei Foraminiferen-Arten.

Ein vollständiger Versuchsplan wurde verwendet um die Parametereinstellungen zu optimieren. Die Anzahl der Parameterkombinationen ist aufgrund der Laufzeit und dem Speicherverbrauch (siehe Abschnitt 4.5) begrenzt. Die Werte der Faktorstufen in Versuchsplan 1 wurden in Absprache mit Experten des ISAS gewählt (siehe Tabelle 4.3). Die maximale Precursormassenänderung `prec` wurde nicht variiert (konstant 10 ppm). Da die Winkel-Distanz auch Werte größer 1 annehmen kann, wurden in Versuchsplan 2 weitere Faktoren mit höheren Schwellenwerten `cdis` hinzugefügt. Der DISMS2-Algorithmus wurde angewandt auf insgesamt 81 Faktorkombinationen. Zur Optimierung wurden neun Gruppen mit technischen Replikaten einer Probe gebildet. Wie in Abschnitt 3.2.2 beschrieben, wurde in einem nichtparametrischen Verfahren zur Varianzanalyse ausgehend von den Distanzmatrizen jeweils das Bestimmtheitsmaß R^2 bestimmt. Für jede Parameterkombination wurde dazu der Permutationstest `adonis` aus dem R-Paket `vegan` (Oksanen et al., 2016) mit 10000 Permutationen durchgeführt, der auf Unterschiede zwischen den Gruppen prüft. Die Ergebnisse der Parameteroptimierung sind in Tabelle 4.4 zusammengefasst. Von besonderem Interesse ist die Wahl des Abstandsmaßes, daher wurden die besten Parameter je Abstandsmaß fett gedruckt.

Tabelle 4.3: Versuchsplan der Faktorstufen der Parameter im DISMS2-Algorithmus. Insgesamt gibt es 81 Kombinationen, davon 72 in Versuchsplan 1 und 9 in Plan 2. `topn` $= \infty$ bedeutet, dass alle Peaks eines Spektrums berücksichtigt werden (in Anlehnung an Rieder et al., 2017a, Tabelle 1).

Parameter	Versuchsplan 1	Versuchsplan 2
`topn`	$20, 50, \infty$	$20, 50, \infty$
`bin`	0.01, 0.2	0.2
`ret`	1000, 3000	3000
`prec`	10	10
`dist`	$d_{\text{angle}}(\epsilon = 0.05), d_{\cos}, d_{\text{PH}}(\delta = 0.05, k = 50)$	$d_{\text{angle}}(\epsilon = 0.05)$
`cdis`	0.1, 0.3	0.4, 0.5, 0.6

Tabelle 4.4: Parameteroptimierung im DISMS2-Datensatz mithilfe des Bestimmtheitsmaßes R^2. In allen 81 Kombinationen gilt `prec = 10`. Optimale Werte bezüglich verschiedener Parameter sind durch fett gedruckte Zeilen gekennzeichnet (Rieder et al., 2017a, Tabelle 2).

Rang	`topn`	`bin`	`ret`	`dist`	`cdis`	R^2
1	∞	**0.20**	**3000**	$d_{\mathbf{cos}}$	**0.3**	**0.923**
2	50	0.20	3000	$d_{\cos}$	0.3	0.923
3	20	0.20	3000	$d_{\cos}$	0.3	0.923
4	∞	0.20	3000	$d_{\cos}$	0.1	0.892
5	50	0.20	3000	$d_{\cos}$	0.1	0.892
6	20	0.01	3000	$d_{\cos}$	0.3	0.890
7	20	0.20	3000	$d_{\cos}$	0.1	0.890
8	50	0.01	3000	$d_{\cos}$	0.3	0.890
9	∞	0.01	3000	$d_{\cos}$	0.3	0.890
10	**20**	**0.01**	**3000**	$d_{\mathbf{PH}}$	**0.3**	**0.879**
11	∞	0.20	1000	$d_{\cos}$	0.3	0.878
⋮	⋮	⋮	⋮	⋮	⋮	⋮
23	**20**	**0.20**	**3000**	$d_{\mathbf{angle}}$	**0.6**	**0.808**
⋮	⋮	⋮	⋮	⋮	⋮	⋮
81	∞	0.20	3000	d_{angle}	0.3	0.308

Um den Einfluss der Parameter auf R^2 zu quantifizieren, wurde ein Regressionsbaum mithilfe des R-Pakets `rpart` (Therneau et al., 2015) erstellt (Abbildung 4.6). Bei einem Regressionsbaum handelt es sich um einen Entscheidungsbaum für eine stetige Zielvariable (Hastie et al., 2009, S. 305ff.). Mithilfe des CART (engl. classification and regression tree)-Algorithmus stratifiziert der Baum anhand einer Vielzahl an binären Aufteilungen der Einflussgrößen die Beobachtungen in Untergruppen mit hohen und niedrigen Werten der Zielgröße. Ausgehend von einem Knoten, der alle Beobachtungen enthält, werden zwei Gruppen gebildet. Die binäre Aufteilung erfolgt so, dass die Klassifikation der gegebenen Daten optimal ist. In beiden Gruppen wird die Zielvariable über den Mittelwert modelliert. Die binäre Aufteilung wird in beiden Gruppen und daraus resultierenden Gruppen immer wieder wiederholt, bis ein Abbruchkriterium erfüllt ist. Beispielsweise wird eine minimale Anzahl an Knoten festgelegt. Der Regressionsbaum, der die Beobachtungen bezüglich R^2 in Abhängigkeit der Parameter des DISMS2-Algorithmus klassifiziert, beinhaltet Aufteilungen der Distanzmaße und des Schwellenwerts. Eine Änderung des Distanzmaßes hat den

größten Einfluss auf das Ergebnis. Die Kosinus-Distanz schneidet besser ab als die Winkel- und Hausdorff-Distanz. Die Konkurrenzfähigkeit der Hausdorff-Distanz zur Kosinus-Distanz wird verbessert, wenn zusätzlich ein großer Schwellenwert `cdis` gewählt wird.

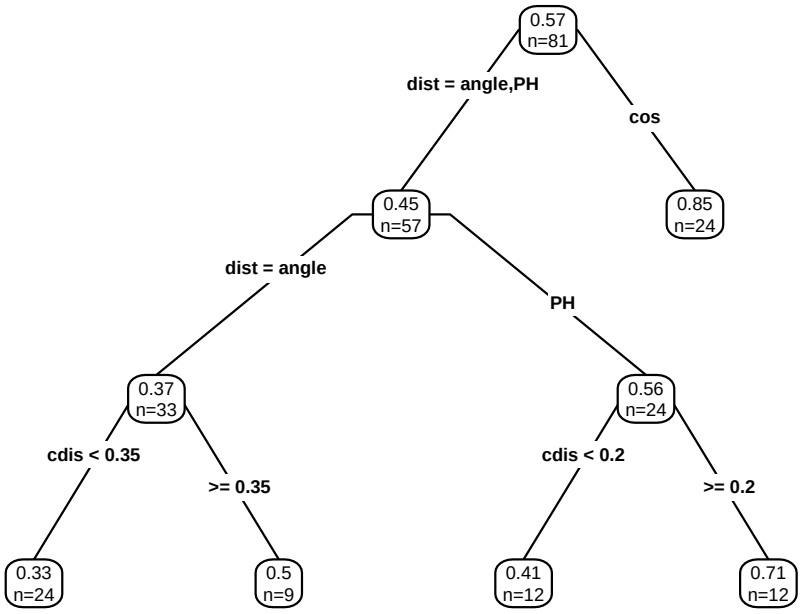

Abbildung 4.6: Angepasster Regressionsbaum zur Erklärung des Bestimmtheitsmaß R^2 (adonis) basierend auf Kombinationen von Parametereinstellungen der DISMS2-Parameter topn, bin, ret, dist, und cdis. Jeder Knoten zeigt das mittlere R^2 im Knoten (oben) und die Anzahl der Beobachtungen, die im Knoten liegen (unten). Die Klassifikation wurde mithilfe des R-Pakets rpart (Therneau et al., 2015) durchgeführt (Rieder et al., 2017a, Abbildung 1).

Zur Visualisierung der aus dem DISMS2-Algorithmus resultierenden Distanzmatrix aller 27 Läufe wird als phylogenetischer Baum ein Dendrogramm erstellt, das auf einer hierarchischen Clusteranalyse mit Average-Link basiert (Abbildung 4.7). Der kophenetische Korrelationskoeffizient liegt bei 0.999, d. h. die Distanzen in der Originalmatrix werden sehr gut widergespiegelt. Der durchschnittliche Abstand technischer Replikate liegt in etwa bei 0.30. Bei drei Knoten, die unterschiedliche Gruppen derselben Art verbinden, wird eine höhere Ähnlichkeit beobachtet. Der Knoten zwischen Mensch und Maus liegt auf der Höhe von 0.62. Auch der Knoten zwischen *Radix* MOTU2 und MOTU4 (Abstand 0.67) und der Knoten zwischen *A. gibbosa* und *A. lessonii* (Abstand 0.76) fällt auf.

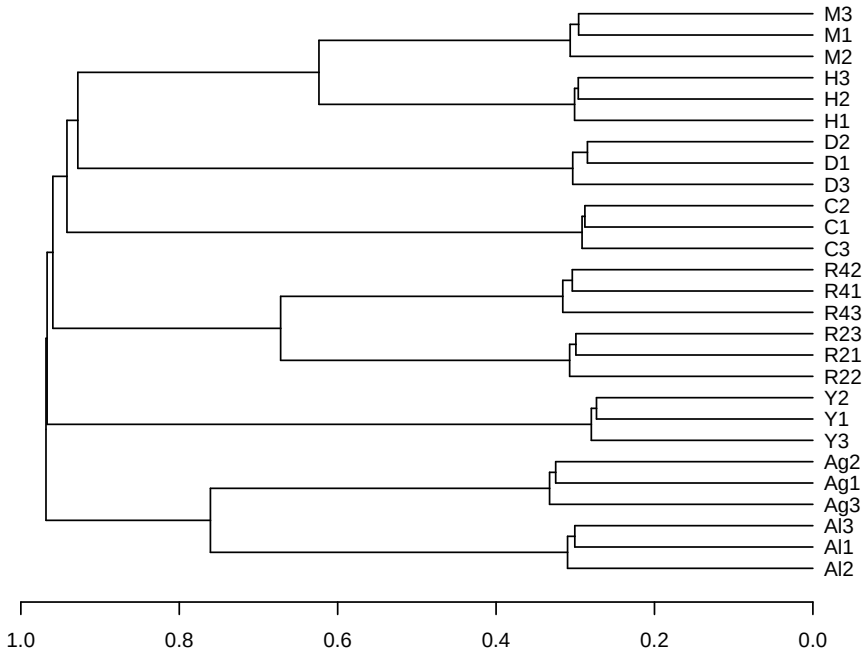

Abbildung 4.7: Dendrogramm basierend auf einer hierarchischen Clusteranalyse mit Average-Linkage für die DISMS2-Distanzen aller 27 Läufe des DISMS2-Datensatzes (DISMS2.f) mit optimierten Parametereinstellungen (Rieder et al., 2017a, Abbildung 5).

Auf die anderen Datensätze des Leibniz-Projekts (Foraminiferen, Biodiversität-Exactive, Biodiversität-Orbitrap) wurde ebenfalls der DISMS2-Algorithmus angewendet (siehe Abschnitt 4.3 und 4.4). Der Foraminiferen-Datensatz wurde mithilfe der gleichen Versuchspläne wie für den DISMS2-Datensatz (Tabelle 4.3) optimiert. In den Biodiversität-Datensätzen wurde keine weitere Optimierung der Parameter durchgeführt, sondern die im DISMS2-Datensatz optimierten Parameter wurden verwendet. Die Umsetzbarkeit einer Gruppeneinteilung, die zur Optimierung nötig ist, ist nicht möglich. Denn es ist fraglich um welche Spezies es sich bei vielen Proben handelt und mehrere Arten sind nur durch einen Lauf repräsentiert.

Eine Analyse des Palmblad-Datensatzes dient zusätzlich zur Evaluierung des DISMS2-Algorithmus. Palmblad und Deelder (2012) konnten mithilfe des compareMS2-Algorithmus, der als Vorlage für den DISMS2-Algorithmus dient, den phylogenetischen Baum der Menschenaffen und anderer Primaten erstellen (siehe links in Abbildung 4.8).

Um die Parametereinstellungen zu optimieren, wurde ein neuer Versuchsplan erstellt. Der bei der Optimierung des DISMS2-Datensatzes angepasste Regressionsbaum (siehe Abbildung 4.6) enthält Aufteilungen der Distanzmaße (`dist`) und

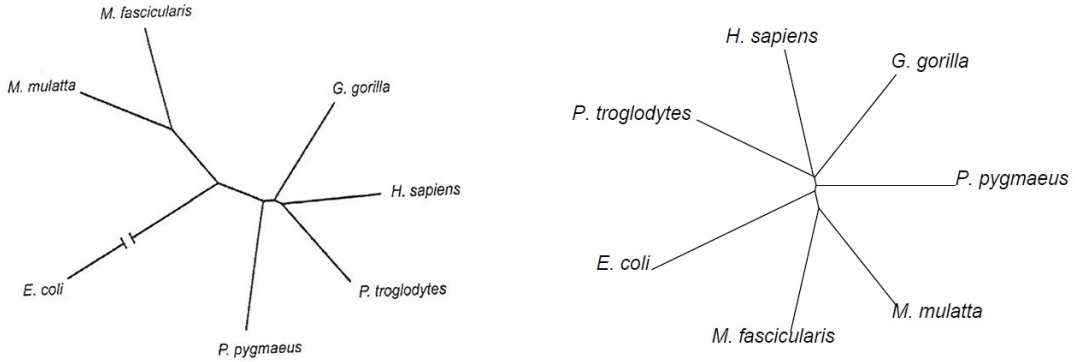

Abbildung 4.8: Phylogenetischer Baum von sechs Primaten und *E. coli* in Datensatz Palmblad. Links: Abbildung 4 in Palmblad und Deelder (2012). Rechts: DISMS2.f mit optimierten Parametern. Die Durchschnittliche Spezies-Distanzen basieren auf allen paarweisen Probenvergleichen und sind als ungerichteter phylogenetischer Baum dargestellt.

des Schwellenwerts (`cdis`). Daher enthält der Versuchsplan (siehe Tabelle 4.5) verschiedene Faktorstufen für `dist` und `cdis`. Die maximale Precursormassenänderung wurde sehr hoch gewählt (`prec`= 10000), da die Auflösung des Massenspektrometers nicht so gut ist. Für die anderen Parameter wurden die optimalen Werte (`topn`= ∞, `bin`= 0.2, `ret`= 3000) des DISMS2-Datensatzes gewählt. Die Berücksichtigung aller Peaks (`topn`= ∞) entspricht einer Auswahl der 50 höchsten Peaks (`topn`= 50), da je Spektrum exakt 50 Peaks vorliegen.

Tabelle 4.5: Versuchsplan der Faktorstufen der Parameter im DISMS2-Algorithmus für den Palmblad-Datensatz. Insgesamt gibt es 12 Kombinationen, davon 9 in Versuchsplan 1 und 3 in Plan 2.

Parameter	Versuchsplan 1	Versuchsplan 2
`topn`	50	50
`bin`	0.2	0.2
`ret`	3000	3000
`prec`	10000	10000
`dist`	$d_{\text{angle}}(\epsilon = 0.05), d_{\cos}, d_{\text{PH}}(\delta = 0.05, k = 50)$	$d_{\text{angle}}(\epsilon = 0.05)$
`cdis`	0.1, 0.2, 0.3	0.4, 0.5, 0.6

Auf die 12 resultierenden Faktorstufen wurde der DISMS2-Algorithmus angewendet. Die Gruppen wurden bezüglich der Individuen einer Art gewählt. Es handelt sich also nicht um technische Replikate einer Probe, sondern Blutseren verschiedener Individuen wurden analysiert. Die Gruppengrößen variieren. Bei fünf Arten (GG, HS, MF, MM, PP) liegen je vier Läufe vor und für die Schimpansen (PT)

gibt es sogar sechs Läufe. Nur ein Lauf wurde für die Referenz (EC1) erstellt. Das Bestimmtheitsmaß R^2 wurde in dem Permutationstest `adonis` mit 10000 Permutationen bestimmt (siehe Tabelle A.5). Die optimalen Werte der drei unterschiedlichen Distanzmaße sind deutlich niedriger als im DISMS2-Datensatz. Die Wahl des Schwellenwerts (`cdis`) hat keinen Einfluss auf das R^2. Das R^2 der optimalen Parametereinstellung (`dist`$= d_{\cos}$) liegt nur bei 0.385.

Ebenfalls wird ein phylogenetischer Baum aller 27 Läufe erstellt (siehe Abbildung 4.9). Das Dendrogramm, das das Ergebnis einer hierarchischen Clusteranalyse mit Average-Linkage ist, spiegelt die Distanzen in der Originalmatrix sehr gut wider. Der kophenetische Korrelationskoeffizient nimmt nämlich den Wert 0.975 an. Die Distanz innerhalb der Gruppen liegt bei über 0.67. Wie zu erwarten ist, liegt die Referenz EC1 am weitesten entfernt von allen anderen. Eine klare Trennung der übrigen Gruppen ist nicht zu erkennen. Für einen Schwellenwert von 0.85 ergeben sich drei Cluster. Die Referenz EC1 bildet ein Einzelcluster. In einem weiteren Cluster befinden sich die Arten (MF, MM) der Gattung der Makaken (*Macaca*). Eine Unterteilung der Makaken in Javaneraffe (MF) und Rhesusaffe (MM) ist anhand der Clusterlösung nicht möglich. Die mittlere Distanz von GG3 zu allen anderen Läufen im dritten Cluster ist am größten. Bis auf diese Ausnahme sind in dem Cluster die Abstände innerhalb der Arten minimal kleiner als zwischen den Arten. Diese Aufteilung der Läufe in Gruppen wird auch mithilfe einer multidimensionalen Skalierung ($k = 2$) deutlich (siehe Abbildung B.3). Zusätzlich zu der im Dendrogramm aufgezeigten Aufteilung ist eine Abgrenzung der Orang-Utans (PP) zum dritten Cluster zu erkennen.

Zum direkten Vergleich der vorliegenden Distanzen mit den Ergebnissen von Palmblad und Deelder (2012) werden die Distanzen innerhalb der Gruppen ebenfalls gemittelt. Mithilfe des Neighbor-Joining-Algorithmus wird ebenfalls ein ungewurzelter Baum erstellt und dem anderen Baum gegenübergestellt (siehe Abbildung 4.8). Die Verzweigungen der beiden Bäume sind identisch. Allerdings unterscheiden sich die Astlängen. Die Entfernung zum nächsten Verwandten ist bei Verwendung des DISMS2-Algorithmus größer als beim compareMS2-Algorithmus.

Zusammenfassend ist festzustellen, dass bei der Anwendung des DISMS2-Algorithmus auf die Daten DISMS2 und Palmblad die Parameter mithilfe geeigneter Versuchspläne bezüglich des R^2 in einem nichtparametrischen Verfahren zur Varianzanalyse (Permutationstest `adonis`) optimiert wurden. Für die optimalen Parametereinstellungen der Daten DISMS2 ($R^2 = 0.923$) und Palmblad ($R^2 = 0.385$) wurden die Abstände mittels Dendrogrammen dargestellt. Eine deutliche Abgren-

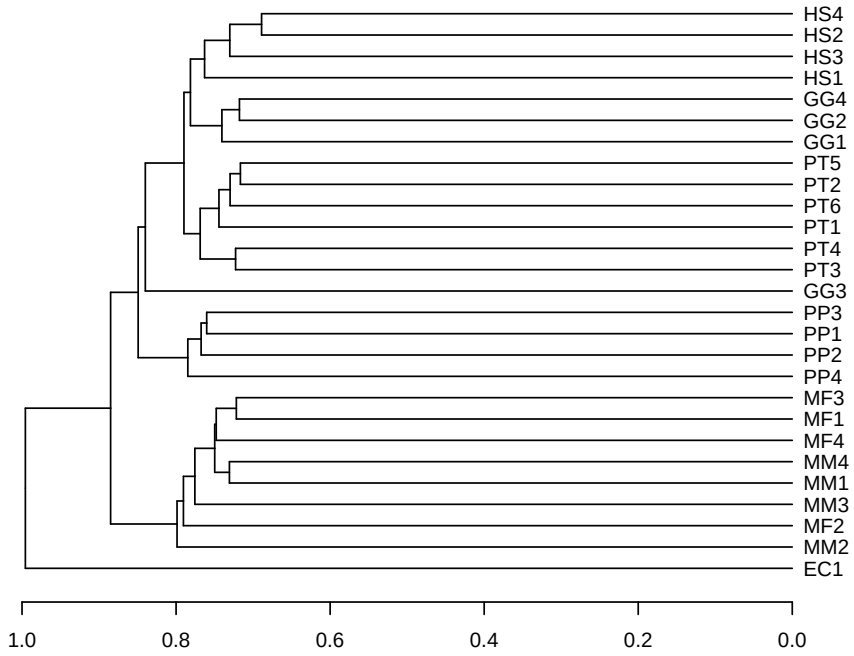

Abbildung 4.9: Dendrogramm basierend auf einer hierarchischen Clusteranalyse mit Average-Linkage für die DISMS2-Distanzen aller 27 Läufe des Palmblad-Datensatzes (DISMS2.f) mit optimierten Parametereinstellungen.

zung der Gruppen wird im DISMS2-Datensatz deutlich. Hingegen sind die Abstände im Palmblad-Datensatz viel größer, sodass beispielsweise keine Unterscheidung der beiden Makakenarten, Javaneraffe und Rhesusaffe, möglich ist.

Mehrere Gründe können für die schlechtere Anpassung angeführt werden. Es wurde ein technisch anderes Massenspektrometer zur Generierung der Spektren verwendet. Anstatt der Originaldaten liegt eine Auswahl an 2000 Spektren je Lauf vor und die Spektren wurden zudem vorverarbeitet, sodass exakt 50 Peaks je Spektrum vorliegen. Die Aufgabe verschiedene Primaten zu unterscheiden ist schwieriger als zwischen den nicht näher verwandten Arten im DISMS2-Datensatz zu differenzieren. Hinzu kommt, dass zur Gruppenbildung bei der Optimierung anstatt technischer Replikate mehrere Individuen einer Art gewählt wurden. Da bei einer massenspektrometrischen Analyse in einem Lauf immer nur ein Teil des kompletten Proteoms gemessen werden kann, sollten Replikate gemessen werden. Die Kombination mehrerer Replikate respräsentiert dann das wahre Proteom besser. Die Datenqualität des Palmblad-Datensatzes ist daher für diese phylogenetische Anwendung nicht gut geeignet. Zusätzlich hat die Darstellungsart des phylogenetischen Baumes einen Einfluss auf die Interpretation. Palmblad und Deelder (2012) bilden durchschnittliche

Distanzen in den einzelnen Gruppen für die Darstellung als ungerichteter Baum. Mit einer analogen Darstellung der Ergebnisse des DISMS2-Algorithmus ergeben sich identische Verzweigungen (siehe Abbildung 4.8).

4.3 Vergleich der Abstände mit Abständen basierend auf Annotationen

Zunächst erfolgt für den DISMS2-Datensatz ein ausführlicher Vergleich mit Abständen, die auf Peptidannotationen basieren. Ein direkter Vergleich des vorgestellten Algorithmus mit einem klassischen Annotationsabstand ist unfair, da es Unterschiede in mehreren Schritten gibt. Der DISMS2-Algorithmus zeichnet sich insbesondere durch das Herausfiltern potentieller Kandidaten (Filterkontrolle) aus und bei der Datenbanksuche spielt die Annotation eines Spektrums eine besondere Rolle (Annotationskontrolle). Um den Einfluss einzelner Schritte besser zu verstehen, werden verschiedene Algorithmen verglichen, die algorithmische Schritte unterschiedlich kombinieren. Tabelle 4.6 beinhaltet eine Liste der verglichenen Algorithmen.

Tabelle 4.6: Überblick über Algorithmen zum Vergleich der Abstände des DISMS2-Algorithmus mit Abständen basierend auf Peptidannotationen (Rieder et al., 2017a, Tabelle 3).

Name	Such-methode	Universum der Spektren	Annotations-kontrolle	Filter-kontrolle	Duplikate entfernt
DB.ra	*Datenbank*	reduziert	ja	nein	nein
DB.ra.nodup	*Datenbank*	reduziert	ja	nein	ja
DISMS2.f	*Distanz*	gesamt	nein	ja	nein
DB.a	*Datenbank*	gesamt	ja	nein	nein
DISMS2.af	*Distanz*	gesamt	ja	ja	nein
DB.af	*Datenbank*	gesamt	ja	ja	nein

Wie im DISMS2-Algorithmus wird für alle Algorithmen der Durchschnitt von zwei gerichteten Suchen gebildet, in denen die relative Häufigkeit der Spektren in einem Lauf ohne Trefferspektrum im anderen Lauf bestimmt wird. Für den Vergleich werden zwei Listen von Spektren, die jeweils aus einem Lauf stammen, verwendet. Sukzessive wird jedes Spektrum der einen Liste mit der anderen Liste abgeglichen. Die Algorithmen unterscheiden sich in der Definition eines Treffers, dem Universum

der Spektren, mögliche Annotations- und Filterkontrollen sowie einer eventuellen Entfernung von Duplikaten.

Bei der Suchmethode *Datenbank* wird ein Treffer erzielt, wenn in der Kandidatenspektrenliste das gleiche Peptid annotiert ist. Die andere Suchmethode, *Distanz*, zählt einen Treffer, falls die Kosinus-Distanz zwischen einem Spektrum und einem zugehörigen Kandidatenspektrum kleiner als `cdis` = 0.3 ist. Das Universum der Spektren ist die Bezeichnung für die Menge aller Spektren in den jeweiligen Läufen. Das gesamte Universum umfasst alle Spektren kompletter Läufe. In dem reduzierten Universum sind nur Spektren enthalten, für die eine Peptidannotation vorliegt. Bei einer Annotationskontrolle werden Treffer nur berücksichtigt, wenn beide Spektren annotiert sind. Unter Umständen werden also übereinstimmende Kandidaten nicht als Treffer gewertet. Wie bereits im ersten Abschnitt des Kapitels erwähnt, ist ein großer Anteil an Spektren nicht annotiert. Die Filterkontrolle umfasst eine Prüfung bezüglich Retentionszeit, Precursormasse und Ladung, wie im zweiten Schritt des DISMS2-Algorithmus beschrieben. Um einen Treffen zu erhalten, müssen alle Bedingungen erfüllt sein. Normalerweise wird das gesamte Spektrenuniversum berücksichtigt, also auch Duplikate von Peptidannotationen werden beibehalten. Bei Entfernung dieser doppelten Annotationen (nodup) wird eine Trefferliste von allen Peptiden verwendet, deren Annotation mindestens einmal vorkommt.

In Tabelle 4.6 sind insgesamt sechs Algorithmen vermerkt. DISMS2.f und DB.ra bezeichnen zwei Standardmethoden. DISMS2.f analysiert das gesamte Universum an Spektren und beinhaltet die Filterkontrolle. Hingegen fehlt bei DB.ra die Filterkontrolle und es wird nur das reduzierte Universum untersucht. Zusätzlich werden Spektren mit gleicher Annotation bei der Methode DB.ra.nodup durch einen Repräsentanten ersetzt, d. h. es werden doppelte Annotationen weggelassen. Die Berechnung der Distanz bei der Methode DB.ra.nodup entspricht der Kulczynski-Distanz. Es wird das arithmetische Mittel des Anteils der Peptide, die in beiden Listen vorkommen, an allen Peptiden in der ersten oder zweiten Liste bestimmt. Es gibt Verknüpfungen zwischen der Filterkontrolle und der *Distanz*-Suchmethode sowie zwischen der Annotationskontrolle und der *Datenbank*-Suchmethode. DISMS2.f und DB.a unterscheiden sich nur durch die Annotations- und Filterkontrolle. DB.a berücksichtigt im Gegensatz zur vergleichbaren Methode DB.ra das gesamte Universum der Spektren und nicht nur das reduzierte Universum. Für einen fairen Vergleich wurden auch zwei Varianten erstellt, die die Annotations- und gleichzeitig die Filterkontrolle umfassen. Unter Einbezug aller Spektren im Universum unterscheiden sich DISMS2.af und DB.af nur durch die Suchmethode.

Zum Vergleich der Algorithmen wird die durchschnittliche Distanz der Läufe innerhalb und zwischen fünf annotierten Gruppen, die sich bezüglich der betrachteten Arten ergeben, untersucht (Abbildung 4.10 und Tabelle 4.7). Die zugehörigen Standardfehler, die die Streuung zwischen technischen Replikaten beschreibt, sind zu vernachlässigen (siehe Tabelle A.6). Sie liegen innerhalb der Arten unter 0.008 und sind maximal 0.002 in Vergleichen zwischen Arten. Die durchschnittliche Distanz ist innerhalb der Arten klein und zwischen den Arten groß. Die meisten Werte in Vergleichen zwischen den Arten liegen über 0.9. Eine Ausnahme stellt der Vergleich von Mensch und Maus dar. Die Methoden mit einem reduzierten Universum an Spektren, DB.ra (0.551) und DB.ra.nodup (0.623), und die neue Methode DISMS2.f (0.624) führen zu den kleinsten Werten. Innerhalb der Arten lässt sich eine Rangfolge der Algorithmen erstellen. Für DB.ra und DB.ra.nodup ergeben sich die kleinsten Werte, gefolgt von DISMS2.f. Die Werte von DB.a liegen viel höher. Das Schlusslicht bilden die beiden Varianten mit Annotations- und Filterkontrolle, DB.af und DISMS2.af.

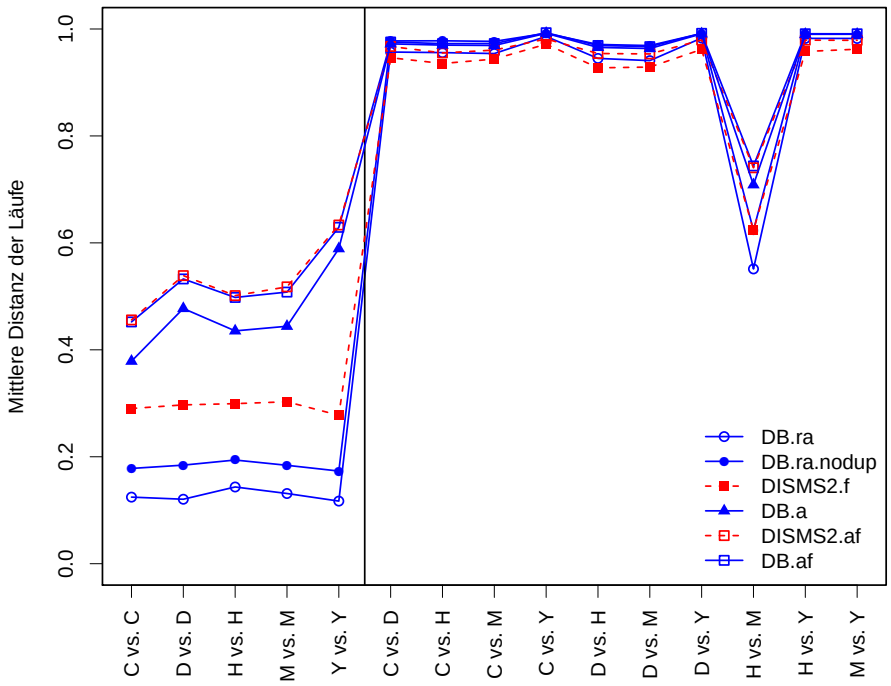

Abbildung 4.10: Mittlerer relativer Anteil an Partnern für unterschiedliche Methoden des Proteomvergleichs. Es werden Kombinationen von Suchmethoden, Spektrumuniversen, Annotationskontrolle, Filterkontrolle und ggf. die Entfernung duplizierter Spektren verglichen zwischen (rechts) und innerhalb (links) der Arten Fadenwurm (C), Fruchtfliege (D), Mensch (H), Maus (M) und Hefe (Y) (in Anlehnung an Rieder et al., 2017a, Abbildung 3).

Tabelle 4.7: Durchschnittliche Distanz der Läufe für unterschiedliche Proteomvergleichmethoden zwischen und innerhalb von Spezies. Die Methoden unterscheiden sich in Algorithmenschritten (Suchmethode, Universum der Spektren, Annotations- und Filterkontrolle, Entfernung von Duplikaten). Die durchschnittliche Distanz von Läufen zwischen (unten) und innerhalb (oben) der Spezies Fadenwurm (C), Fruchtfliege (D), Mensch (H), Maus (M) und Hefe (Y) (Rieder et al., 2017a, Tabelle 4).

	DB.ra	DB.ra.nodup	DISMS2.f	DB.a	DISMS2.af	DB.af
C vs. C	0.125	0.178	0.290	0.379	0.456	0.452
D vs. D	0.121	0.184	0.297	0.477	0.539	0.533
H vs. H	0.143	0.194	0.299	0.435	0.501	0.498
M vs. M	0.131	0.184	0.303	0.444	0.518	0.508
Y vs. Y	0.117	0.173	0.278	0.589	0.633	0.629
C vs. D	0.957	0.978	0.946	0.972	0.967	0.976
C vs. H	0.956	0.978	0.935	0.970	0.955	0.973
C vs. M	0.954	0.977	0.944	0.969	0.961	0.973
C vs. Y	0.986	0.992	0.972	0.992	0.981	0.993
D vs. H	0.945	0.971	0.927	0.966	0.954	0.969
D vs. M	0.941	0.969	0.929	0.963	0.953	0.967
D vs. Y	0.983	0.992	0.962	0.991	0.979	0.992
H vs. M	0.551	0.623	0.624	0.709	0.740	0.744
H vs. Y	0.982	0.991	0.958	0.990	0.979	0.991
M vs. Y	0.982	0.991	0.962	0.990	0.979	0.991

Vor einer Überinterpretation ist zu warnen, da das Proteom der Proben unbekannt ist. Datenbankannotationen als Mittel zur Beschreibung der Zusammensetzung von Peptiden können unvollständig und fehlerhaft sein. Dies führt zu einer Ungenauigkeit bei der Interpretation der Größe der mittleren Distanz. Zwei der Algorithmen, DB.ra und DB.ra.nodup, basieren auf einem reduzierten Universum der Spektren. Eine Auswahl der annotierten Spektren schließt Spektren mit geringer Qualität aus, da diese in der Regel nicht annotiert werden können. Infolgedessen werden viele große Distanzen entfernt, die auf Vergleiche mit Spektren geringer Qualität zurückzuführen sind. Fakt ist, dass die mittleren Distanzen innerhalb der Arten bei DB.ra und DB.ra.nodup unterhalb 0.2 liegen und somit am besten abschneiden.

Die Ergebnisse der zwei Methoden mit Annotations- und Filterkontrolle, DB.af und DISMS2.af, lassen sich direkt miteinander vergleichen. Die Anzahl möglicher Treffer ist viel kleiner, da die Spektren in zwei Kontrollen nicht aussortiert wer-

den dürfen. Allein der Anteil fehlender Annotationen liegt zwischen 30% und 60%, sodass für viele Spektren der ersten Liste gar keine Suche nach einem Treffer durchgeführt wird. In der zweiten Liste gibt es zusätzlich einen beachtlichen Anteil an Spektren, die keinen Treffer erzielen, da sie entweder die Annotations- oder die Filterkontrolle nicht überstehen (siehe Tabelle A.7). Dieser Anteil liegt zwischen 8% und 13% (innerhalb von Arten) und bei 30% bis 50% (zwischen Arten). Zum Vergleich der Suchmethoden *Distanz* und *Datenbank* wird beispielhaft für den Vergleich von zwei humanen Proben, H1 und H2, der Algorithmus DB.af betrachtet (siehe Abbildung 4.11). Haben Paare von Spektren die gleiche Annotation, so wird eine kleine Kosinus-Distanz erwartet. In den Daten ist eine rechtsschiefe Verteilung (dunkelgrau markiert) mit Modus nahe 0 der Kosinus-Distanz dieser Paare an Spektren zu beobachten. Ist die Annotation verschieden, so ergibt sich eine linksschiefe Verteilung und die meisten Distanzen liegen oberhalb des Schwellenwerts 0.3. Es gibt also wenige Spektren, die trotz großer Kosinus-Ähnlichkeit nicht gleich annotiert sind. Ein Grund dafür können falsche oder fehlende Annotationen sein.

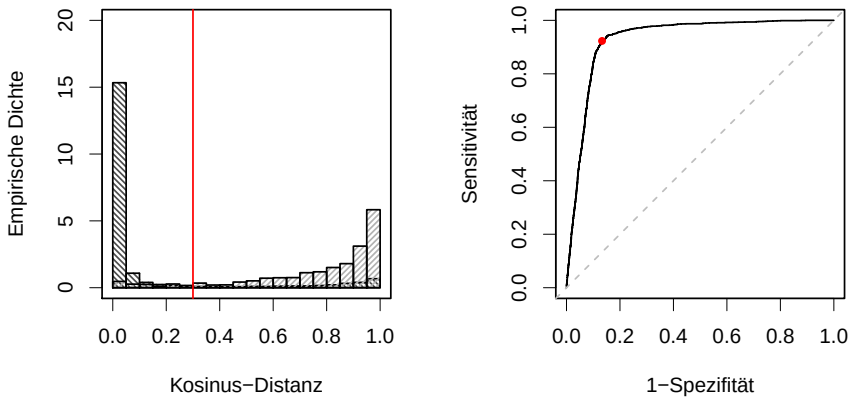

Abbildung 4.11: Histogramm von Tandem-Massenspektren und korrespondierende ROC-Kurve. Das Histogramm (links) zeigt Distanzen von Tandem-Massenspektren von Paaren mit gleicher Peptidannotation in DB.af (dunkelgrau) und von den verbleibenden Spektrenpaaren mit unterschiedlicher Peptidannotation in DB.af (hellgrau) beim Vergleich der Proben H1 und H2. Die ROC-Kurve (rechts) zeigt die Performance der Kosinus-Distanz als binärer Klassifikator gleicher oder unterschiedlicher Peptide für unterschiedliche Schwellenwerte `cdis`. Der Schwellenwert `cdis`= 0.3 ist rot markiert (in Anlehnung an Rieder et al., 2017a, Abbildung 4).

Die Kosinus-Distanz ist also eine gute Wahl für einen binären Klassifikator, der zwischen Gruppen gleicher und unterschiedlicher Peptide klassifiziert. Die Performanz der Klassifikation ist an einer ROC-Kurve (engl. receiver operating characteristic) abzulesen (Hastie et al., 2009, S. 313 ff.). Zur Bestimmung einer ROC-Kurve

werden für jeden Schwellenwert der Kosinus-Distanz Sensitivität und Spezifität berechnet. Die Sensitivität beschreibt den Anteil richtig positiv klassifizierter an allen positiven Beobachtungen. Hingegen wird der Anteil richtig negativ klassifizierter an allen negativen Beobachtungen Spezifität genannt. Für den optimierten Schwellenwert `cdis` = 0.3 ergibt sich eine Sensitivität von 0.923 und eine Spezifität von 0.867. In einer ROC-Kurve, die die Sensitivität (y-Achse) in Abhängigkeit von 1-Spezifität (x-Achse) darstellt, kann zusätzlich das AUC (engl. area under the curve), die Fläche unter der Kurve, bestimmt werden. Das AUC kann Werte zwischen 0 und 1 annehmen. Für eine zufällige Klassifikation ergibt sich der Wert 0.5 und für eine perfekte Klassifikation der Wert 1. Für die gegebenen Daten liegt der AUC-Wert hoch bei 0.93.

Die allgemein verwendete Standardmethode DB.ra.nodup unterscheidet sich von allen anderen betrachteten Algorithmen. Spektren, deren Annotation Duplikate sind, gehen mit einem anderen Gewicht in die Distanzberechnung der Läufe ein. Die mittlere Distanz innerhalb von Arten von DB.ra.nodup liegt ungefähr bei 18% und somit niedriger als ungefähr 30%, dem Wert der neuen Methode DISMS2.f.

Das Dendrogramm der Methode DISMS2.f (Abbildung 4.7) wurde bereits in Abschnitt 4.2 erläutert. Weitere Dendrogramme (Abbildung B.4, B.5, B.6, B.7, B.8) wurden für 15 annotierte Läufe mithilfe der übrigen Methoden erstellt. Die resultierenden Distanzmatrizen von DB.af und DISMS2.af sind annähernd identisch. Im Vergleich zu den anderen Methoden ist die Trennung zwischen den Arten nicht so deutlich.

Für den direkten Vergleich von zwei Methoden wurde zunächst die absolute Differenz der Distanzen von zwei Methoden berechnet und anschließend der Variationskoeffizient, d. h. der Quotient aus Standardabweichung und Mittelwert (Tabellen A.8 und A.9). Ein Wert kleiner 0.5 weist auf einen relevanten Unterschied der jeweiligen beiden Methoden hin. In den meisten Fällen gibt es relevante Unterschiede. DISMS1.af und DB.af hingegen zeigen große Ähnlichkeit, da die Variationskoeffizienten relativ groß sind und einige sogar deutlich größer als 0.5.

Ein Vergleich mit Abständen, die auf Peptidannotationen basieren, ist auch im Foraminiferen-Datensatz möglich. Die De-Novo-Peptidsequenzierung liefert jedoch nur wenige Annotationen der Spektren (siehe Abschnitt 4.1). Im Zusammenhang mit der Vorverarbeitung von Spektren ist im folgenden Abschnitt 4.4 die Parameteroptimierung im DISMS2-Algorithmus und die Auswahl einer Teilmenge der Spektren enthalten. Es wird auch ein Vergleich zu Distanzen erstellt, die auf der Grundlage von De-Novo-Annotationen erstellt wurden. Zur Bestimmung der Ähnlichkeit von Pep-

tidlisten wurde der Sørensen-Dice-Index anstatt des Kulczynski-Koeffizienten verwendet. In Abbildung 4.12 ist ein Vergleich des mittleren relativen Anteils an Partnern der beiden Methoden DISMS2.f und DB.ra.nodup im Foraminiferen-Datensatz dargestellt. Beide Methoden wurden leicht verändert. DISMS2.f enthält zusätzlich eine Vorauswahl an Spektren. Je Lauf werden nur die 2000 Spektren mit der höchsten Gesamtionen-Intensität berücksichtigt. DB.ra.nodup beinhaltet nicht die Berechnung des Kulczynski-Koeffizienten, sondern die Berechnung des populäreren Sørensen-Dice-Index. Links sind die Distanzen innerhalb der Arten (Agi, Ale, Alo und Mve) und rechts sechs Vergleiche zwischen Paaren von unterschiedlichen Arten abgebildet.

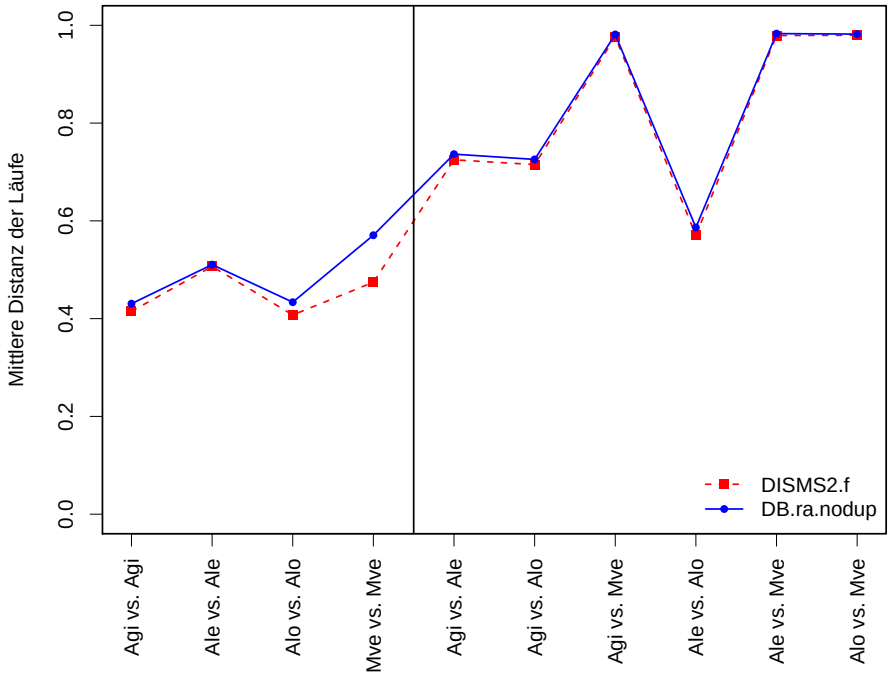

Abbildung 4.12: Mittlerer relativer Anteil an Partnern für zwei Methoden des Proteomvergleichs der Arten *Amphistegina gibbosa* (Agi), *Amphistegina lessonii* (Ale), *Amphistegina lobifera* (Alo) und *Marginopora vertebralis* (Mve).

Die Resultate der beiden Methoden sind sehr ähnlich (Abbildung 4.12). Die mittlere Distanz der Läufe schwankt innerhalb der Arten zwischen 0.41 und 0.57. Die mittlere Distanz (0.57) der Art *M. vertebralis* (Mve) liegt bei der *Datenbank*-Methode DB.ra.nodup in etwa auf der Höhe der mittleren Distanz (0.59) zwischen zwei *Amphistegina* Arten (Ale und Alo). Das Proteom der Arten *A. lessonii* und *A. lobifera* hat am meisten Ähnlichkeit im Vergleich unterschiedlicher Arten. Die Distanz dieser beiden *Amphistegina* Arten zu *A. gibbosa* liegt im Durchschnitt un-

gefähr bei 0.72. Es ist plausibel, dass die paarweisen Distanzen zwischen Arten innerhalb derselben Gattung *Amphistegina* geringer sind als die Distanzen zwischen *M. vertebralis* und einer der anderen Arten, die immer oberhalb 0.97 liegen.

Die Standardfehler sind im Vergleich zum DISMS2-Datensatz viel höher (siehe Tabelle A.10). Sie liegen maximal bei 0.040 (DISMS2.f) bzw. 0.056 (DB.ra.nodup) innerhalb der Art *Marginopora vertebralis* (Mve). Bei der Analyse von biologischen statt technischen Replikaten ist eine größere Streuung zu erwarten. Bei der Probenvorbereitung musste ein neues Verfahren entwickelt werden um genug Material aus den winzigen Kalkschalen zu extrahieren. Daher wurden je Lauf acht Organismen gepoolt. Es sind in jedem Lauf auch die Symbionten enthalten, da die Trennung der Algen von den Foraminiferen bisher nicht möglich ist. Eine Variation der Proteinzusammensetzung innerhalb der Arten ist auch durch die zum Teil unterschiedliche Herkunft zu erklären. Außerdem wurde im Foraminiferen-Datensatz im Vergleich zum DISMS2-Datensatz das Massenspektrometer Fusion statt Q Exactive verwendet.

Zusammenfassend ist in diesem Abschnitt festzustellen, dass zum Vergleich der Abstände des Algorithmus DISMS2 mit Abständen mittels Peptidannotationen der DISMS2-Datensatz analysiert wurde. Im direkten Vergleich des neu vorgestellten Algorithmus DISMS2 (DISMS2.f), der allein auf einem Massenspektrenvergleich basiert, mit der Standardmethode mittels Annotationen (DB.ra.nodup) schneidet DISMS2.f etwas schlechter ab als DB.ra.nodup. Die Algorithmen unterscheiden sich in mehreren Schritten. Durch die Betrachtung weiterer Algorithmen, die mehrere gleiche Schritte enthalten, wird ein fairer Vergleich ermöglicht. Besonders im Vergleich von DISMS2.af und DB.af sind keine deutlichen Unterschiede zwischen der Suchmethode *Datenbank* und *Distanz* aufzudecken. Eine gute Wahl für einen binären Klassifikator, der zwischen gleichen und verschiedenen Peptiden in der Datenbank unterscheidet, ist also die Kosinus-Distanz der Tandem-Massenspektren. Auch anhand des Foraminiferen-Datensatzes wird die Konkurrenzfähigkeit von DISMS2 mit einer Distanzberechnung, die auf Peptidannotationen aus Datenbanken basiert, deutlich. Die Datenqualität ist nicht so gut wie im DISMS2-Datensatz, sodass die Trennung der Arten nicht so deutlich wird. Im folgenden Abschnitt 4.4 werden die Daten der Foraminiferen näher analysiert und auch in Bezug auf das Fangjahr und die Herkunft untersucht.

4.4 Einfluss der Generierung und Vorverarbeitung von Tandem-Massenspektren

Weitere Aspekte, die Einfluss auf die Bestimmung der Distanz von LC-MS/MS-Läufen haben, werden in diesem Abschnitt anhand der Foraminiferen- und Biodiversität-Datzensätze erläutert. Bei der Optimierung der Parameter des DISMS2-Algorithmus (siehe Abschnitt 4.2) wird deutlich, dass die Wahl eines geeigneten Distanzmaßes für den Vergleich von Tandem-Massenspektren, aber auch die Vorverarbeitung der Spektren, die Auswahl an Peaks mit höchsten Intensitäten und das Binning, Bedeutung haben. Im Palmblad-Datensatz fällt auf, dass je Lauf 2000 vorverarbeitete Spektren mit je 50 Peaks ausgewählt wurden. Es wurden jeweils nur die Spektren mit dem höchsten Totalionensignal veröffentlicht. Diese Möglichkeit der Vorverarbeitung der Tandem-Massenspektren beeinflusst das Spektren-Universum und schließlich auch die resultierende Distanzmatrix des DISMS2-Algorithmus. Wie in Abschnitt 4.3 erläutert ist die Unterscheidung der Arten im Foraminiferen-Datensatz erschwert. Zur Verbesserung der Resultate wird im Folgenden die Möglichkeit der Auswahl einer Teilmenge an Spektren mit hohem Gesamtionensignal analysiert. Die Datensätze dieser Arbeit enthalten Tandem-Massenspektren, die mithilfe unterschiedlicher Massenspektrometer generiert wurden. Die beiden Biodiversität-Datensätze eignen sich sehr gut zum Vergleich der Generierung von Massenspektren, da jede Probe im Q Exactive und Orbitrap Elite Massenspektrometer analysiert wurde. Am Ende dieses Abschnitts werden die Ergebnisse der Biodiversität-Datensätze diskutiert.

Im Foraminiferen-Datensatz wurden zunächst Distanzen für alle Spektren berechnet. Anschließend wurden in jedem Durchlauf 2000 Spektren mit dem höchsten Gesamtionensignal extrahiert. Eine Vorauswahl von 2000 besonderen Spektren wurde bereits vorher von Palmblad und Deelder (2012) verwendet. Um optimale Parametereinstellungen für alle Spektren im Foraminiferen-Datensatz zu finden, wurde das gleiche vollfaktorielle Design wie in Rieder et al. (2017a) verwendet. Von insgesamt 81 Faktorkombinationen wurde die Kombination mit dem höchsten R^2 (0.6183) gewählt (siehe auch Tabelle A.11). Die optimierten Parameter entsprechen den Werten im DISMS2-Datensatz (siehe Abschnitt 4.2).

Für die optimierten Parameter wird ein Dendrogramm erstellt, das auf einer hierarchischen Clusteranalyse mit Average-Link basiert (Abbildung 4.13, oben). Wie bereits in Abschnitt 4.3 erwähnt, sind die Distanzen im Foraminiferen-Datensatz höher als im DISMS2-Datensatz. Zur Verbesserung wird die von Palmblad und Deelder

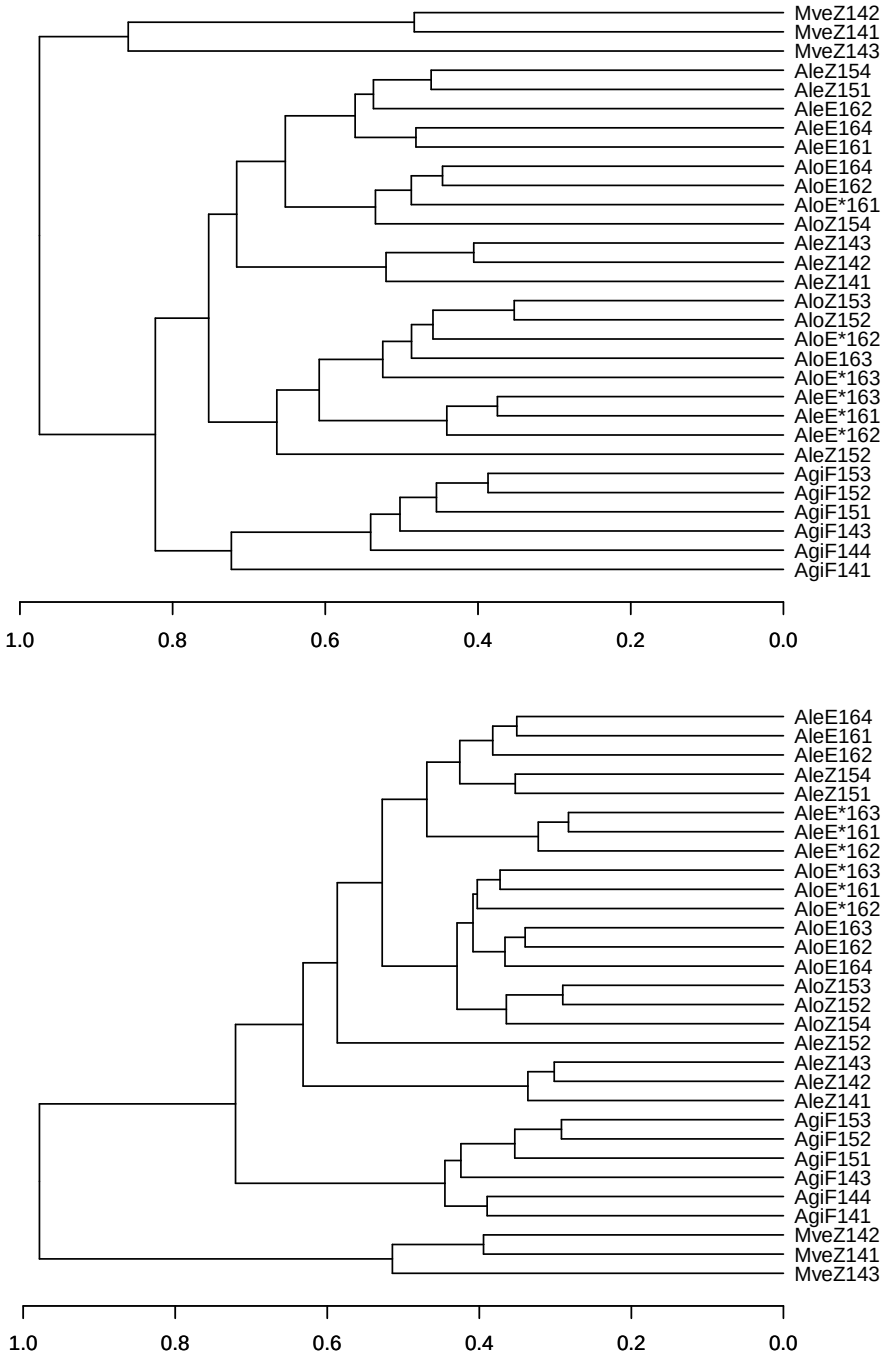

Abbildung 4.13: Gegenüberstellung von zwei Dendrogrammen des Foraminiferen-Datensatz generiert mittels dem DISMS2-Algorithmus. Die Parameter im DISMS2-Algorithmus wurden optimiert und es wurden je Lauf alle (oben) oder nur die 2000 (unten) Spektren mit dem höchsten Gesamtionensignal verwendet.

(2012) verwendete Auswahl einer Teilmenge an Spektren mit hohem Gesamtionensignal vorgenommen. Werden je Lauf nur Tandem-Massenspektren mit den 2000 höchsten Gesamtionensignalen berücksichtigt, so werden die paarweisen Distanzen zwischen Läufen kleiner und das Ergebnis der hierarchischen Clusteranalyse mit Average-Link ändert sich (Abbildung 4.13, unten). Die Distanzen beider Originalmatrizen werden gut durch die Dendrogramme dargestellt, denn der kophenetische Korrelationskoeffizient liegt bei 0.901 (alle Spektren) und 0.989 (2000 Spektren). Wird das untere Dendrogramm in Abbildung 4.13 auf der Höhe von 0.7 abgeschnitten, so ergeben sich drei Cluster. Die Läufe der Art *Marginopora vertebralis* und *Amphistegina gibbosa* bilden je ein Cluster. Die Läufe von *Amphistegina lessonii* und *Amphistegina lobifera* bilden das dritte Cluster. Wird ein niedrigerer Schwellenwert auf der Höhe von 0.52 gewählt, so liegen sechs Cluster vor. Das dritte Cluster wird in vier Cluster aufgeteilt. Die Läufe der Art *Amphistegina lobifera* bilden eines dieser Cluster. Die Läufe der Art *Amphistegina lessonii* sind auf drei Cluster aufgeteilt. Alle Proben aus Sansibar aus dem Jahr 2014 (AleZ14) bilden ein Cluster. Eine Probe aus Sansibar aus dem Jahr 2015, AleZ152, bildet ein Einzelcluster. Alle weiteren Läufe der Art *Amphistegina lessonii* bilden ein Cluster. Zusätzlich zu den direkt gemessenen und kultivierten Proben aus Eilat liegen zwei Proben aus Sansibar aus dem Jahr 2015 (AleZ151 und AleZ154) in dem Cluster.

Proben der Art *A. gibbosa* wurden nur in Florida entnommen. Ein Unterschied zwischen den Fangjahren 2014 und 2015 ist in der Hierarchie nicht eindeutig zu bestimmen. Für die Art *A. lobifera* ist in der Hierarchie eine Trennung nach der Herkunft, Eilat oder Sansibar zu erkennen. Bei der Art *A. lessonii* hingegen liegen zwei von drei biologischen Replikaten aus Sansibar (2015) im Cluster mit den Proben aus Eilat (2016). Die direkt gemessenen Proben aus Eilat (E*) bilden ein Cluster und weisen weniger Ähnlichkeit mit den kultivierten Proben aus Eilat (E) auf. Dieser Unterschied zwischen direkt gemessenen und kultivierten Proben fällt bei der Art *A. lobifera* nicht auf. Ein großer Unterschied bei der Art *A. lessonii* ist bei den Proben aus Sansibar zwischen den Fangjahren 2014 und 2015 zu beobachten. Besonders die ein Einzelcluster bildende Probe AleZ152 zeigt exemplarisch, dass die Ähnlichkeit von biologischen Replikaten zum Teil kleiner ist als zwischen verschiedenen Arten. Die Distanz 0.529 zwischen den biologischen Replikaten AleZ151 und AleZ152 der Art *A. lessonii* aus Sansibar (2015) ist größer als die Distanz 0.518 zwischen zwei Läufen AleZ151 und AloZ152 unterschiedlicher Art, *A. lessonii* und *A. lobifera*, aus Sansibar (2015). Das Genom der symbiontischen Algen wurde bisher noch nicht sequenziert. Ähnlichkeiten von Läufen gleicher Herkunft und glei-

chen Fangjahres können gegebenenfalls in Zukunft auch durch Artenzugehörigkeit der Algen erklärt werden.

Für alle MS/MS-Läufe des Foraminiferen-Datensatz liegt eine De-Novo-Peptidsequenzierung mit der von Blank-Landeshammer et al. (2017) vorgestellten Methode vor. Die Distanz der Peptidlisten wurde mit der Sørensen-Dice-Distanz bestimmt. Die Auswahl der intensivsten Signale und die Anwendung des DISMS2-Algorithmus im Foraminiferen-Datensatz führt zu ähnlichen Ergebissen wie die auf De-Novo-Annotationen basierte Distanzbestimmung. In der direkten Gegenüberstellung der Dendrogramme der resultierenden Distanzmatrizen ist zu erkennen, dass die hierarchische Clusterung bis auf zwei Vertauschungen zu gewurzelten Bäumen mit identischen Verzweigungen führt (Abbildung 4.14). Je zwei biologische Replikate der Art *A. lobifera*, gefangen in Eilat im Jahr 2016, AloE*161 und AloE*162 sowie AloE162 und AloE164, sind vertauscht. Außerdem ist die Korrelation der beiden Originaldistanzmatrizen sehr hoch (0.995).

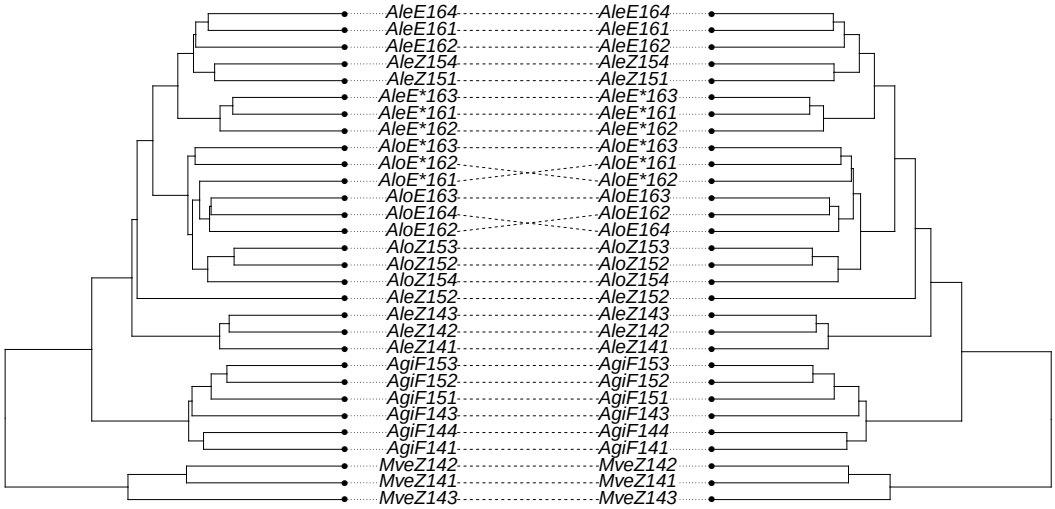

Abbildung 4.14: Gegenüberstellung von zwei Dendrogrammen generiert mittels einer De-Novo-Annotationsabstandsmethode (links) und dem DISMS2-Algorithmus (rechts). Die Parameter im DISMS2-Algorithmus wurden optimiert und es wurden je Lauf nur die 2000 Spektren mit dem höchsten Gesamtionensignal verwendet.

Zwei Verfahren für die Bestimmung des Proteomabstands, die Anwendung des DISMS2-Algorithmus auf eine Vorauswahl von je 2000 Spektren je Lauf mit höchster Gesamtionenintensität und die Sørensen-Dice-Distanz von De-Novo-Peptidlisten, führen zu sehr ähnlichen Dendrogrammen. Es stellt sich daher die Frage, ob ein Zusammenhang zwischen der Höhe der Gesamtionenintensität eines Tandem-Massenspektrums und einer De-Novo-Annotation besteht. In Abbildung 4.15 sind für alle

annotierten und alle nicht annotierten Tandem-Massenspektren im Foraminiferen-Datensatz Boxplots der Gesamtionenintensität dargestellt. Im Median liegt die Intensität der annotierten Spektren bei 3 885 249 und somit höher als 1 290 306, dem Median der Gruppe der nicht annotierten Spektren. Die Auswahl der annotierten Spektren führt insgesamt zu einer Verschiebung der Intensitätsverteilung nach rechts. Median, unteres und oberes Quartil und unterer und oberer Whisker der Intensitäten der nicht annotierten Spektren liegen jeweils unterhalb der annotierten Spektren. Die beiden Verteilungen überschneiden sich jedoch. Es kann also kein Schwellenwert für die Gesamtionenintensität bestimmt werden, sodass die annotierten von den nicht annotierten Spektren klar getrennt werden.

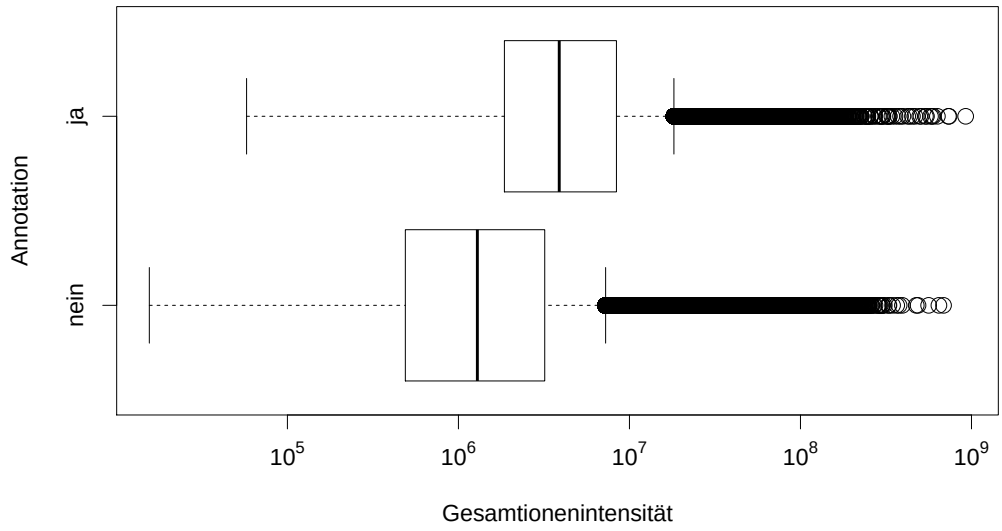

Abbildung 4.15: Boxplots der Gesamtionenintensität ($\log_2$-Skala) annotierter und nicht annotierter Tandem-Massenspektren im Foraminiferen-Datensatz.

Eine Gegenüberstellung der Biodiversität-Datensätze dient zum Vergleich gleicher Proben, die mit unterschiedlichen Massenspektrometern gemessen wurden. Q Exactive und Orbitrap Elite sind zwei aufeinander folgende Generationen von Massenspektrometern der Firma Thermo Scientific. Je Lauf generierte die Orbitrap Elite weniger Spektren als die Q Exactive. Die Auflösung der Orbitrap Elite ist jedoch höher als die der Q Exactive. Die Parameter des DISMS2-Algorithmus wurden nicht optimiert, sondern gemäß der Optimierung des DISMS2-Datensatzes gewählt. Die Wahl der Gruppen für eine Optimierung der Parametereinstellungen ist nämlich unklar, da die Probenauswahl sehr heterogen ist. Für mehrere Arten, beispielsweise Regenwurm und Tellerschnecke, wurde nur ein Repräsentant gewählt. Einzelne Tei-

le, Fuß und Geschlechtsorgan, oder komplette Individuen unterschiedlicher Herkunft der Gattung *Radix* wurden untersucht. Teilweise basiert die Artenbestimmung nicht auf einem DNA-Barcoding, sondern nur auf dem Phänotyp. Die Datenqualität weist insgesamt also eine Reihe an Schwächen auf.

Zum Vergleich von Proben, für die mittels zweier Massenspektrometer in Läufen Massenspektren generiert wurden, wurde der Proteomabstand mithilfe des DISMS2-Algorithmus erstellt. Die Dendrogramme, die aus einer hierarchischen Clusteranalyse mit Average-Link der beiden Distanzmatrizen resultieren, sind in Abbildung 4.16 gegenübergestellt. Der kophenetische Korrelationskoeffizient liegt bei 0.991 (Exactive) und 0.968 (Orbitrap), d. h. die Distanzen beider Originalmatrizen werden sehr gut durch die Dendrogramme dargestellt. Es fällt auf, dass die hierarchische Aufteilung nur bei 16 von 30 Läufen identisch ist. Es gibt zahlreiche Vertauschungen und andere Verzweigungen. Die Abstände des Orbitrap-Datensatzes sind allgemein kleiner als die Abstände des Exactive-Datensatzes. Die Korrelation der beiden Originaldistanzmatrizen liegt bei 0.963. Insgesamt sind die Unterschiede also nicht gravierend.

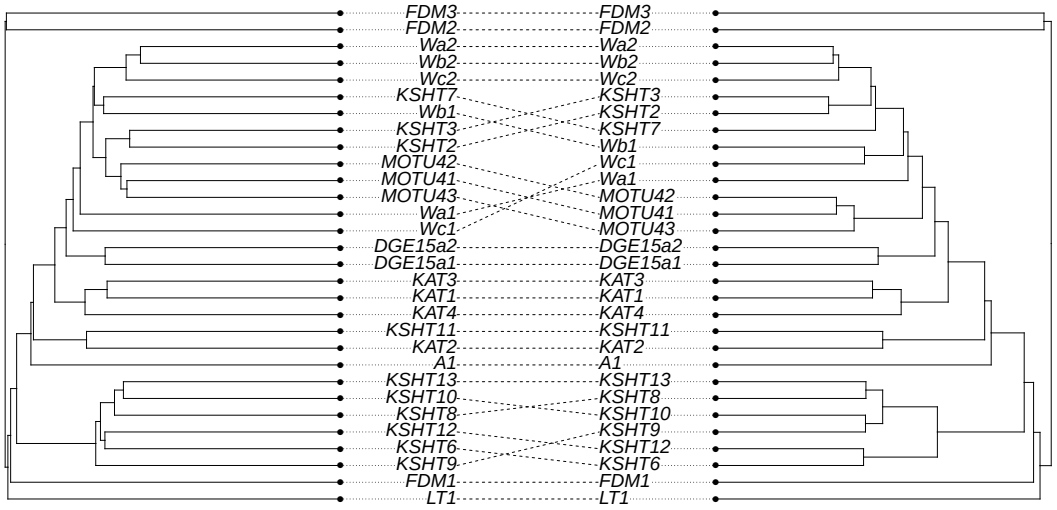

Abbildung 4.16: Gegenüberstellung von zwei Dendrogrammen, die aus einer hierachischen Clusteranalyse mit Average-Link stammen, der Biodiversität-Datensätze (links Exactive, rechts Orbitrap) generiert mittels dem DISMS2-Algorithmus. Die Parameter wurden im DISMS2-Datensatz optimiert.

In Abbildung 4.17 ist das Dendrogramm des Orbitrap-Datensatzes vergrößert dargestellt. Die paarweisen Abstände liegen zwischen 0.323 und 0.978. Wird das Dendrogramm auf der Höhe von 0.7 abgeschnitten, so ergeben sich acht Cluster. Regenwurm, Tellerschnecke und die drei Meeresfrüchte bilden jeweils Einzelcluster und werden, wie zu erwarten ist, nicht zu den Radixproben gruppiert. Diese teilen

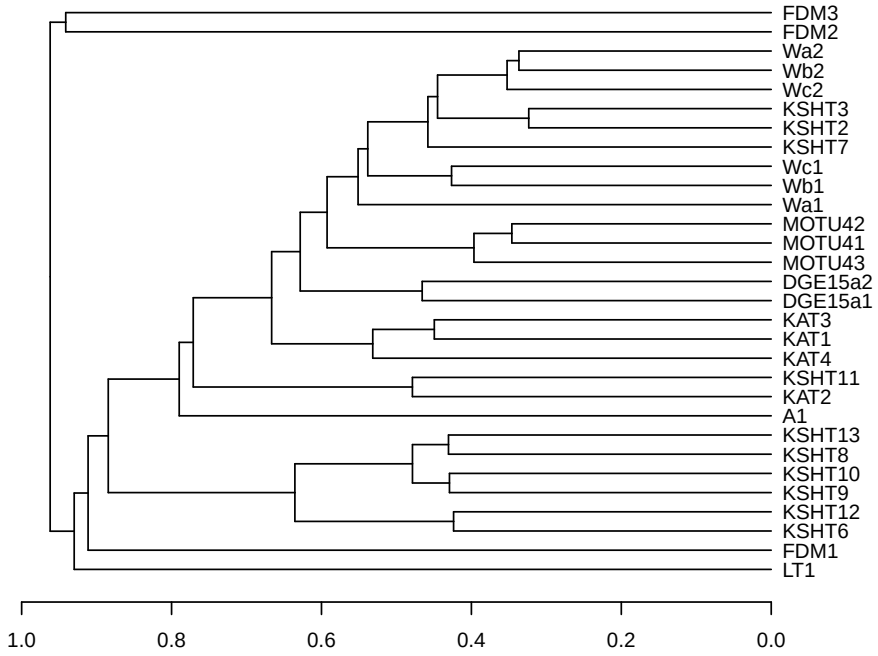

Abbildung 4.17: Dendrogramm des Biodiversität-Orbitrap-Datensatz generiert mittels einer hierarchischen Clusteranalyse mit Average-Link der Distanzmatrix aus dem DISMS2-Algorithmus. Die Parameter wurden im DISMS2-Datensatz optimiert.

sich in drei Cluster der Größe 2, 6 und 17 auf. Zwei Radixproben, KAT2 und KSHT11, bilden ein kleines Cluster. Sechs weitere Radixproben (KSHT6, -8, -9, -10, -12 und -13) bilden ein Cluster. Die übrigen 17 Radixproben bilden ein großes Cluster. Es sind Proben aus dem Taunus, Gelnhausen und vom ISAS enthalten sowie einzelne Teile der Schnecken, Fuß und Geschlechtsorgan.

Wird eine niedrigerer Schwellenwert 0.5 zum Abschneiden des Dendrogramms gewählt, so ergeben sich 15 statt 8 Cluster. Die beiden größten Cluster mit 6 und 17 Radixproben werden feiner gruppiert. Zum einen ergeben sich zwei Cluster mit zwei (KSHT6 und -12) und vier (KSHT8, -9, -10, -13) Radixproben. Das größte Cluster wird in sieben Cluster aufgeteilt. Drei Proben, die vom ISAS stammen, bilden ein Cluster und es gibt zwei Einzelcluster, Wa1 und KAT4. Drei Cluster der Größe 2 bilden zwei Geschlechtsorganproben aus dem Taunus (Wb1 und Wb2), zwei Teilproben, Fuß und Geschlechtsorgan, aus Gelnhausen (DGE15a1 und -2) und zwei weitere Radixproben (KAT1 und -3). Das größte Cluster bilden sechs Proben, drei Fußproben aus dem Taunus (Wa1, Wb1 und Wc2) und drei weitere Radixproben (KSHT2, -3 und -7).

Zur Validierung der Clusterlösung dient die Artenzugehörigkeit, die Herkunft und die Angabe, ob eine komplette Schnecke oder ein Teil der Zwitter, Geschlechtsorgan oder Fuß, untersucht wurden. Die biologischen Replikate von *R. auricularia*, die von einer Inzuchtlinie aus einem Aquarium am ISAS stammen, haben einen maximalen Abstand von 0.397. Die Artenbestimmung der übrigen Proben ist unklar und basiert nur auf dem Phänotyp. Der Wildfang aus Gelnhausen ist von den anderen Proben unterscheidbar. Geschlechtsorgan und Fuß dieser Schnecke, die nach dem Phänotyp zur Art *R. balthica* gehört, liegen in einem Cluster der Größe 2. Fuß und Geschlechtsorgan der drei Wildfänge aus dem Taunus, die nach dem Phänotyp zur Art *R. auricularia* angehören, sind hingegen zu unterscheiden. Die Fußproben liegen mit drei weiteren Radixproben ganzer Schnecken in einem Cluster. Die Geschlechtsorganproben der zwittrigen Schnecken teilen sich in zwei Cluster, einem Einzelcluster und einem Cluster der Größe 2. Die Läufe der Wildfänge aus dem Taunus werden auf der Höhe von 0.593 mit den Läufen der Inzuchtlinie vom ISAS zu einem Cluster verbunden. In diesem Cluster ist nach DNA-Barcoding oder Phänotyp die *R. auricularia* enthalten. Drei nicht näher bestimmte Radixproben unbekannter Herkunft (KSHT2,-3-7) gehören ebenfalls diesem Cluster an. Die mangelnde Probenbezeichnung erschwert also die Interpretation der Ergebnisse.

Zusammenfassend ist in diesem Abschnitt festzustellen, dass der Einfluss der Generierung und Vorverarbeitung von Tandem-Massenspektren auf die Bestimmung des Proteomabstands mittels des DISMS2-Algorithmus untersucht wurde. Im Foraminiferen-Datensatz wurde eine Vorauswahl von 2000 Spektren je Lauf durchgeführt. Die bereits von Palmblad und Deelder (2012) durchgeführte Auswahl an Spektren mit hoher Gesamtionenintensität wurde vor Anwendung des DISMS2-Algorithmus vorgenommen. Das resultierende Dendrogramm konnte durch eine De-Novo-Methode bestätigt werden. Für Peptidlisten, die anhand des von Blank-Landeshammer et al. (2017) vorgestellten Verfahrens zur De-Novo-Annotation erstellt wurden, wurde die populäre Sørensen-Dice-Distanz berechnet. Die Arten *A. gibbosa* und *M. vertebralis* können deutlich voneinander getrennt werden. Zwischen *A. lessonii* und *A. lobifera* ist eine höhere Ähnlichkeit zu erkennen. Die Bestimmung der Symbiontenarten könnte dazu beitragen die Unterscheidung zwischen Herkunft und Fangjahr zu verbessern. Im Median ist die Gesamtionenintensität annotierter Spektren (3 885 249) höher als die nicht annotierter Spektren (1 290 306). Eine klare Trennung der Verteilungen liegt im Foraminiferen-Datensatz jedoch nicht vor.

Die Generierung der Massenspektren durch unterschiedliche Massenspektrometer hat einen Einfluss auf die Bestimmung des Proteomabstands. Im Vergleich der

Biodiversität-Datensätze konnten Unterschiede der Höhe der Distanzen und bei den Verzweigungen im Dendrogramm festgestellt werden. Wie zu erwarten ist, können die Radixproben von anderen Arten unterschieden werden. Allgemein ist die Interpretation der resultierenden hierarchischen Clusterlösung der Radixproben aufgrund der mangelnden Datenqualität unklar. Eine Unterscheidung der Herkunft oder der Teile der Schnecken ist nicht eindeutig. Die Generierung stabiler Ergebnisse kann bei verschiedenen Massenspektrometern unter schwierigen Versuchsbedingungen, beispielsweise mangelnder Datenqualität, nicht gewährleistet werden. Zur Verbesserung der Distanzbestimmung sind weitere Vorverarbeitungsschritte denkbar. Das in Kapitel 4.1 bestimmte Qualitätsmaß könnte alternativ zur Gesamtionenintensität für die Vorauswahl von Spektren verwendet werden. Auch eine Erweiterung der grundlegenden Distanzberechnung einzelner Spektren ist erfolgversprechend. Im flexiblen Algorithmus können Maße ergänzt werden. Die von Novak et al. (2013) vorgestellte parametrisierte Hausdorff-Distanz weist in einer früheren Publikation (Novak und Hoksza, 2010) einen zusätzlichen Parameter, den power modifier m, auf. Es wird schließlich die Potenz m von d_{PH} gebildet.

4.5 Anmerkungen zu Laufzeiten und Speicherverbrauch

Die Implementierung von DISMS2 ist konzipiert für die Verwendung mit wenig Speicherverbrauch. Anstatt Berechnungen aller Distanzen von zwei Läufen vorab zu erstellen, werden Distanzen nur nach Bedarf kalkuliert. Eine Reduktion der Anzahl an Berechnungen wird durch die Überprüfung von Bedingungen erreicht. Es wird nur gespeichert, ob letztendlich ein Trefferspektrum gefunden wurde. Eventuell müssen also Distanzen einzelner Spektren mehrfach berechnet werden. In der Regel führt die Verwendung der Überprüfung von Bedingungen jedoch zu einer drastischen Reduktion an Distanzberechnungen. Bei der Parameteroptimierung des DISMS2-Algorithmus der DISMS2-Daten mit 27 Läufen wurde die Laufzeit unter Verwendung des R-Pakets `BatchJobs` (Bischl et al., 2015) gemessen. Es wurden 3 GB RAM auf einem Kern eines Intel Xeon E5-2630 (8 Kerne, 2.4 GHz, 128 GB RAM, Debian Linux 8.3.0 Betriebssystem) angefordert. Die mediane Laufzeit für eine von insgesamt 81 Faktorkombinationen beträgt 13.41 Stunden (Spannweite 4.49 - 20.01 Stunden). Für die beste Parametereinstellung (siehe Tabelle 4.4) sind 15.73 Stunden notwendig.

Die Überprüfung der Bedingungen (a) - (c) in Schritt 2 des DISMS2-Algorithmus, besonders die Wahl von `ret`, reduziert die Laufzeit drastisch. In Abbildung 4.18 ist die Laufzeit des DISMS2-Algorithmus für die Berechnung der Proteomabstandsmatrix im Datensatz DISMS2 in Abhängigkeit von verschiedenen Parametereinstellungen für `ret` dargestellt. Es wurden die Werte 1 000, 3 000, 5 000, 10 000, 20 000, 30 000 und 40 000 für `ret` gewählt. Die übrigen Parametereinstellungen entsprechen den optimierten Werten. Die Laufzeit steigt zunächst stark an, um dann in etwa konstant bei über 26 Stunden zu bleiben. Wird `ret` in der besten Parametereinstellung von 3 000 auf 40 000 erhöht, erhöht sich die Laufzeit von fast 16 Stunden auf über 26 Stunden. Dies entspricht einer Steigerung um den Faktor 1.65.

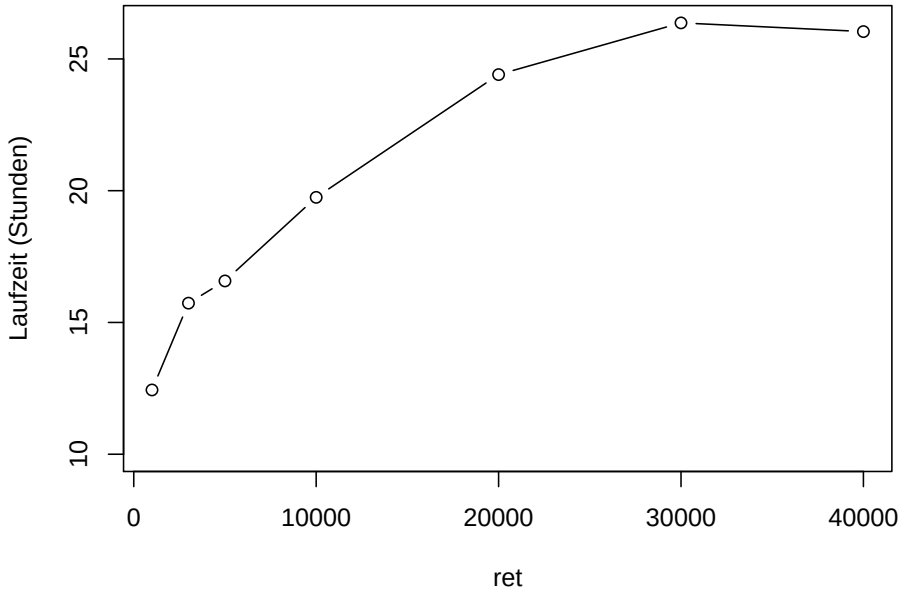

Abbildung 4.18: Laufzeit für die Berechnung der Proteomabstandsmatrix im Datensatz DISMS2 mithilfe des DISMS2-Algorithmus für verschiedene Werte von `ret` (1 000, 3 000, 5 000, 10 000, 20 000, 30 000, 40 000) und ansonsten optimalen Parametereinstellungen.

Die Filterkontrolle bei der Suchmethode Distanz, wie im DISMS2-Algorithmus verwendet, verringert die Laufzeit erheblich (Methode DISMS.f). Zur Veranschaulichung wurde für zwei humane Proben, H1 und H2, DISMS2 ohne Filterkontrolle berechnet. Im Vergleich zu den anderen Methoden (siehe Tabelle 4.8) ist DISMS2 konkurrenzfähig, sogar zu den Methoden mit einem reduzierten Universum an Spektren.

Das Weglassen aller Bedingungen führt dazu, dass ohne Parallelisierung für die Berechnung einer einzelnen Distanz $d(\mathrm{H1}, \mathrm{H2})$ die Laufzeit insgesamt auf etwa zwei Tage steigt. Angenommen, dass die Laufzeit für den Vergleich unterschied-

Tabelle 4.8: Distanzen von zwei humanen Läufen H1 und H2: Gerichteter Abstand d^* und Durchschnitt d der gerichteten Distanzen. DISMS2 bedeutet DISMS2.f ohne Filterkontrolle (Rieder et al., 2017a, Tabelle 5).

Name	$d^*(\text{H1}, \text{H2})$	$d^*(\text{H2}, \text{H1})$	$d(\text{H1}, \text{H2})$
DB.ra	0.1467	0.1487	0.1477
DB.ra.nodup	0.1971	0.2010	0.1990
DISMS2.f	0.2969	0.3050	0.3009
DB.a	0.4321	0.4422	0.4371
DISMS2.af	0.4968	0.5087	0.5027
DB.af	0.4935	0.5068	0.5002
DISMS2	0.1616	0.1645	0.1631

licher Läufe nicht schwankt, würde sich die Laufzeit bei $N = 27$ Läufen für die $N \cdot (N-1)/2 = 351$ Distanzen auf etwa zwei Jahre erhöhen. Nur für die Anzahl an Kandidaten, die nach Prüfung der Bedingungen übrig bleiben, müssen Distanzen berechnet und geprüft werden. Beispielsweise verringert sich beim Vergleich von H1 und H2 die Kandidatenanzahl von mehr als 35 000 im Durchschnitt auf 2.5.

Zusammenfassend ist in diesem Abschnitt festzustellen, dass die Laufzeit und der Speicherverbrauch der Implementierung des DISMS2-Algorithmus im DISMS2-Datensatz untersucht wurde. Der Speicherverbrauch ist gering (unter 3 GB), da Distanzen nur bei Bedarf nach Prüfung mehrerer Bedingungen berechnet werden. Die Laufzeit beträgt für die optimalen Einstellungen aus Abschnitt 4.2 fast 16 Stunden. Besonders eine Lockerung der Einschränkung auf zeitlich nahe Spektren (`ret` = 40 000 anstelle von `ret` = 3 000) führt zu einer Steigerung um den Faktor 1.65. Beispielhaft wurde für die zwei humanen Läufe, H1 und H2, die Filterkontrolle weggelassen. Die resultierende Distanz ist ähnlich zu den Distanzen, die auf Annotationen basieren. Die Kandidatenanzahl erhöht sich von im Mittel 2.5 auf über 35 000. Für den Vergleich aller 27 Läufe würde sich somit die Laufzeit auf über zwei Jahre erhöhen. Die Implementierung des DISMS2-Algorithmus eignet sich hervorragend zur Parallelisierung. Einerseits können einzelne Einträge der Distanzmatrix unabhängig voneinander berechnet werden. Andererseits können die Treffer für Einträge der ersten Liste parallel zueinander bestimmt werden. Insgesamt könnte also die Laufzeit noch erheblich reduziert werden.

5 Clusteranalyse von Tandem-Massenspektren

Im letzten Kapitel wurden tausende Massenspektren mehrerer Läufe zur Berechnung von Distanzen aggregiert. Dabei wurde noch nicht berücksichtigt, dass redundante Spektren, die auf das gleiche Peptid zurückzuführen sind, in einem oder mehreren Läufen technisch bedingt auftreten. Clusterverfahren ermöglichen das Entfernen duplizierter Spektren, sodass nur noch Konsensusspektren, also Repräsentanten eines Clusters, vorliegen. Zur Wahl eines geeigneten Clusterverfahrens wurde in Rieder et al. (2017b) ein bisher in der Literatur fehlender umfassender Vergleich von insgesamt sieben Algorithmen (CAST, MS-Cluster, PRIDE Cluster, hierarchische Clusteranalyse, DBSCAN, igraph und Neighbor Clustering) auf realen Daten (DISMS2-Datensatz) mithilfe mehrerer Gütemaße durchgeführt. Die in diesem Kapitel enthaltene Analyse besteht aus Ergebnissen, die in Rieder et al. (2017b) veröffentlicht sind, sowie darauf aufbauenden Analysen und Ergänzungen. Der einzige wesentliche algorithmische Unterschied der hier präsentierten Analysen und der Analysen in Rieder et al. (2017b) liegt beim PRIDE Clusterverfahren. Es wurde die neu erschienene Version 1.0.3 anstatt Version 1.0.1 der Java-Applikation spectra-cluster-cli verwendet. Die Änderung der Default-Einstellung des Parameters `x_min_comparisons` von 0 (Version 1.0.1) auf 10000 (Version 1.0.3) führt zu einer deutlichen Verbesserung der Clusterergebnisse.

Dieses Kapitel beginnt in Abschnitt 5.1 mit dem Vergleich der sieben Algorithmen für die Daten DISMS2. Zur Evaluierung mithilfe von Gütemaßen wird die Information der Annotationen verwendet. In Abschnitt 5.2 wird eine Auswahl an interessanten Clustern näher beschrieben. Die zunächst nur auf einzelne Läufe beschränkte Clusterbildung von Spektren wird in Abschnitt 5.3 auf mehrere Läufe erweitert. Zu beachten sind dabei die Anmerkungen zu Laufzeit und Speicherverbrauch am Ende des Kapitels (Abschnitt 5.6). Die Clusterbildung von Massenspektren wird in Abschnitt 5.4 als Vorschritt des DISMS2-Algorithmus verwendet. Der Einfluss der

Repräsentantenauswahl wird also untersucht. Abschnitt 5.5 enthält eine Anwendung der Clusterlösungen zur Peptidzuordnung von Spektren mit fehlenden Annotationen im DISMS2-Datensatz. Alle Berechnungen wurden mit der statistischen Software R (R Core Team, 2016, Version 3.4.2) erstellt.

5.1 Vergleich von distanzbasierten Clusterverfahren

Dieser Abschnitt handelt von der Clusterbildung von Tandem-Massenspektren einzelner Läufe des DISMS2-Datensatzes. Jeder der insgesamt 27 Läufe beinhaltet zwischen 30 012 und 40 236 Tandem-Massenspektren (Tabelle 4.1). Die Clusteralgorithmen sind in der statistischen Programmiersprache R implementiert (siehe `https://www.statistik.tu-dortmund.de/genetics-publications-clusspec.html`). Es werden sieben verschiedene Clusteralgorithmen verglichen, die in Abschnitt 3.3.1 näher beschrieben sind. Für CAST, DBSCAN, eine hierarchische Clusteranalyse (Complete-Linkage), igraph, Neighbor Clustering, MS-Cluster und PRIDE Cluster werden die Parametereinstellungen variiert (siehe Tabelle 5.1). Zusätzlich wurde im MS-Cluster-Algorithmus eine feste Anzahl an Runden (Parametereinstellung `rounds`=5) gewählt. Im Gegensatz zu der Darstellung in Rieder et al. (2017b) wurde für das PRIDE Clusterverfahren Version 1.0.3 statt Version 1.0.1 der Java-Applikation spectra-cluster-cli verwendet. Eine deutliche Verbesserung der Clusterlösungen wird durch die Änderung der Default-Einstellung des Parameters `x_min_comparisons` von 0 (Version 1.0.1) auf 10000 (Version 1.0.3) erzielt. Bei einer festen Anzahl an Runden `rounds` = 5, dem Parameter `precursor_tolerance` = 5 und der Default-Einstellung `fragment_tolerance`= 0.5 wird absteigend von `threshold_start` = 1 bis *threshold_end* der Schwellenwert für die Ähnlichkeit von Spektren variiert.

Es wurden gleiche Werte für die Parameter c, $cdis$, ϵ, h und t gewählt, da sie sich alle auf den Schwellenwert für die Kosinus-Distanz zwischen Massenspektren beziehen. Die Werte werden ausgehend von 0.2 iterativ halbiert (0.2, 0.1, 0.05, 0.025). Die Parameter *similarity* und *threshold_end* (für die Algorithmen MS-Cluster und PRIDE Cluster) sind hingegen Schwellenwerte für die Ähnlichkeit von Massenspektren. Daher werden die Werte bezüglich der Transformation $d = 1 - s$ gewählt (0.8, 0.9, 0.95, 0.975).

Um einen fairen Vergleich der Algorithmen zu gewährleisten, wird eine einheitliche Berechnung der Distanz von Tandem-Massenspektren benötigt. Zur Vorver-

Tabelle 5.1: Parametereinstellungen der verwendeten Clusteralgorithmen (Rieder et al., 2017b, Tabelle 3).

Algorithmus	Parameter	Werte
CAST	t	0.025, 0.05, 0.1, 0.2
DBSCAN	ϵ	0.025, 0.05, 0.1, 0.2
	$minPts$	2, 3, 5, 10
hclust	h	0.025, 0.05, 0.1, 0.2
igraph	$cdis$	0.025, 0.05, 0.1, 0.2
N-Cluster	c	0.025, 0.05, 0.1, 0.2
MS-Cluster	$similarity$	0.975, 0.95, 0.9, 0.8
PRIDE Cluster	$threshold\_end$	0.975, 0.95, 0.9, 0.8

arbeitung der Originalspektren wurde ein Binning mit fester Bingröße `bin`= 0.2 angewandt. Die Kosinus-Distanz aller Paare von Tandem-Massenspektren wurde berechnet. Diese Parametereinstellungen (`bin`=0.2, `dist`=$d_{\cos}$ und `topn` = ∞) haben sich bereits bei der Clusterung von Läufen in Abschnitt 4.2 bewährt. Die abgespeicherten resultierenden Distanzmatrizen sind Eingaben aller Clusteralgorithmen bis auf zwei Ausnahmen. Vorverarbeitungsschritte und spezielle Distanzberechnungen sind Teil von MS-Cluster und PRIDE Cluster, sodass im direkten Vergleich mit den anderen Algorithmen die Clusterergebnisse nicht auf den gleichen Distanzen von Massenspektren beruhen.

Zur Verbesserung der Clusterlösungen wird zusätzlich eine zweite Variante der Distanzmatrixkonstruktion berücksichtigt, die den DISMS2-Filter verwendet. Novak et al. (2013) haben bei der Anwendung von SimTandem, einer Methode, die mit den modernen state-of-the-art Tools zur Peptididentifizierung OMSSA und X!Tandem verglichen wird, einen Precursor-Massenfilter verwendet. Diese Idee taucht auch im DISMS2-Filter wieder auf. Außerdem kam eine heuristische Auswahl von Peaks zum Einsatz. Um über den gesamten m/z-Bereich die intensivsten Signale je Spektrum zu filtern, wurden die m/z-Werte in Bereiche von je 50 Dalton aufgeteilt. In jedem der so gewählten Fenster wurden die fünf größten Peaks herausgefiltert. Anschließend wurden von allen herausgefilterten Peaks die insgesamt 50 größten Peaks je Spektrum gewählt. Wie im zweiten Schritt des DISMS2-Algorithmus werden beim DISMS2-Filter Bedingungen geprüft. Falls im paarweisen Vergleich von Spektren beim Abgleich von Zusatzinformationen bezüglich Retentionszeit (maximale Abweichung der geordneten Scannummern um `ret` = 3000), Precursorladung, Precursormasse (maximale Verschiebung der Precursormasse um `prec`=10 ppm) keine Übereinstimmung

vorliegt, wird die zugehörige Distanz durch den Maximalwert 1 ersetzt. Dies hat zur Folge, dass jene Tandem-Massenspektren nicht in einem gemeinsamen Cluster liegen können. Die Wahl der Retentionszeitfenstergröße `ret` hängt von der Reproduzierbarkeit der HPLC ab und ist nur anwendbar für gleiche Versuchsbedingungen bei der Chromatographie. Eine notwendige Voraussetzung für Peptide mit den gleichen Aminosäuresequenzen und den gleichen posttranslationalen Modifikationen sind dieselben Precursormassen. Die Clusterbildung von Tandem-Massenspektren wird durch das Herausfiltern ähnlicher Precursormassen vereinfacht.

Cluster von Spektren können auch gebildet werden, indem Spektren mit gleicher Peptidannotation zusammengefasst werden. Die vorliegenden Annotationen von 18 Läufen im DISMS2-Datensatz der sechs Arten (C, D, H, M, Y, R4) mit je drei technischen Replikaten werden hier verwendet. Die zusätzliche Annotationsclusterlösung gibt einen Hinweis auf die wahre Clusterlösung der Spektren. Unannotierte Spektren werden beim Annotationsclustering nicht berücksichtigt. Werden im Folgenden andere Clusterlösungen mit annotationsbasierten Clusterlösungen verglichen, so werden die nicht annotierten Spektren ignoriert.

Insgesamt 1962 Clusterlösungen von einzelnen Läufen wurden gebildet. Je Lauf wurden 32 Clusterlösungen auf Basis der Kosinus-Distanz und 32 Lösungen mit zusätzlichem DISMS2-Filter erstellt. Zusammen mit acht Clusterlösungen von MS-Cluster und PRIDE mit anderer Distanzberechnung und gegebenenfalls einem Annotationsclustering ergeben sich 72 bis 73 Clusterlösungen je Lauf. Für die 18 annotierten und 9 unannotierten Läufe des DISMS2-Datensatzes sind es insgesamt also $73 \cdot 18 + 72 \cdot 9 = 1962$ Clusterlösungen. Die Lösungen von igraph und DBSCAN bei der Wahl von $minPts = 2$ sind identisch. Denn es handelt sich in beiden Fällen um die Lösung, wenn bei einem hierarchischen Clusterverfahren mit Single-Link das Dendrogramm auf der Höhe von c oder ϵ abgeschnitten wird (siehe Abschnitt 3.3.1).

Zur Bewertung der Clusterlösungen werden die in Abschnitt 3.3.2 vorgestellten Bewertungsmaße der Qualität von Clusterlösungen verwendet. Zunächst wird die Ähnlichkeit verschiedener Lösungen anhand des ARI beurteilt. Im Vergleich einer Clusterlösung mit der Annotationsclusterlösung bedeuten größere Werte (Maximalwert 1) eine bessere Qualität. In Abbildung 5.1 ist der über 18 Läufe gemittelte ARI jeder Clusterlösung im Vergleich zur Annotation dargestellt. Die Werte schwanken zwischen 0.00 und 0.65. In Tabelle 5.2 sind die höchsten mittleren Werte des ARI einzelner Läufe für verschiedene Clustermethoden und verschiedene Distanzberechnungen dargestellt. Zusätzlich sind in der letzten Spalte die jeweiligen Werte für Clusterlösungen mehrerer Läufe (siehe Abschnitt 5.3) aufgeführt.

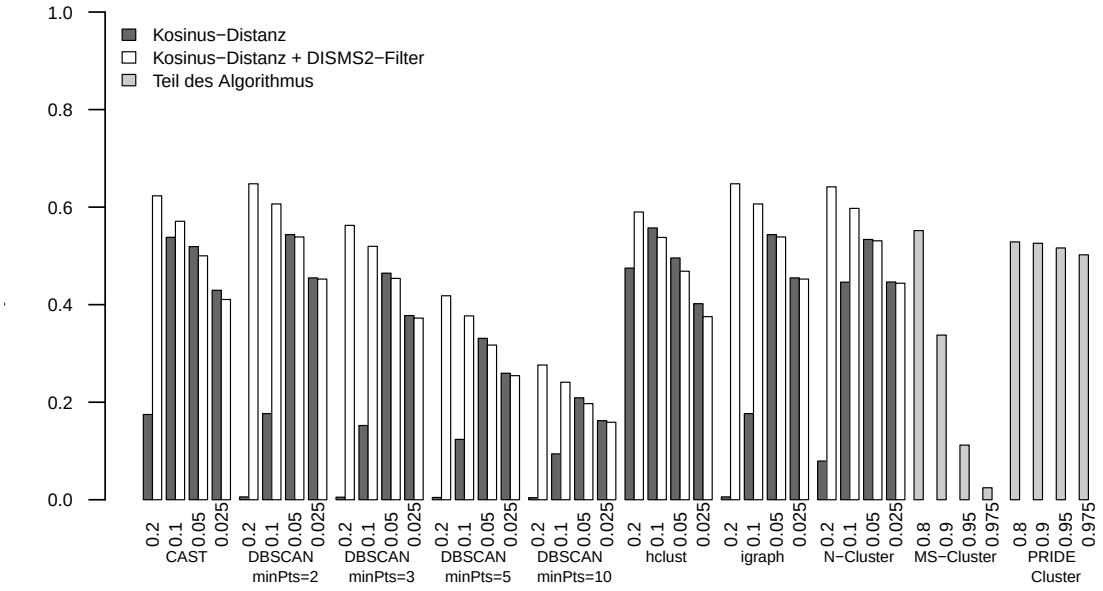

Abbildung 5.1: Durchschnittlicher adjustierter Rand-Index zwischen der Peptidannotation und Clusterlösungen von 18 annotierten Läufen. Die Kosinus-Distanz wurde ohne und mit DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster. Für PRIDE Cluster wurde die neue Version 1.0.3 verwendet (in Anlehnung an Rieder et al., 2017b, Abbildung 1).

Tabelle 5.2: Durchschnittlicher adjustierter Rand-Index von einzelnen und mehreren Läufen (in Anlehnung an Rieder et al., 2017b, Tabelle 4).

Distanz-berechnung	Clusterverfahren	einzelne Läufe	mehrere Läufe
Kosinus-Distanz	CAST ($t = 0.1$)	0.538	0.642
	DBSCAN ($\epsilon = 0.05, minPts = 2$)	0.544	0.667
	hclust ($h = 0.1$)	0.557	0.608
	igraph ($cdis = 0.05$)	0.544	0.667
	N-Cluster ($c = 0.05$)	0.534	0.649
Kosinus-Distanz mit DISMS2-Filter	CAST ($t = 0.2$)	0.623	0.721
	DBSCAN ($\epsilon = 0.2, minPts = 2$)	0.648	0.743
	hclust ($h = 0.2$)	0.590	0.648
	igraph ($cdis = 0.2$)	0.648	0.743
	N-Cluster ($c = 0.2$)	0.642	0.730
Teil des Algorithmus	MS-Cluster ($similarity = 0.8$)	0.552	0.593
	PRIDE Cluster ($threshold\_end = 0.8$)	0.529	0.678

Im Vergleich der beiden auf Tandem-Massenspektren spezialisierten Algorithmen, MS-Cluster und PRIDE Cluster, schneidet MS-Cluster für Einzelläufe im Mittel etwas besser ab. Sind Kosinus-Distanzen die Eingabe der Algorithmen, so liegen bei adäquater Parameterwahl die gemittelten ARI-Werte der fünf Algorithmen auf gleicher Höhe zwischen 0.53 und 0.56. Das beste Resultat liefert das hierarchische Verfahren ($h = 0.1$). Die neue Methode N-Cluster ($c = 0.05$) kann mit den etablierten Verfahren mithalten. Der DISMS2-Filter führt allgemein zu höheren mittleren ARI-Werten. Die beste Wahl ist DBSCAN ($\epsilon = 0.2, minPts = 2$) bzw. igraph ($cdis = 0.2$) mit einem mittleren Wert von 0.65, direkt gefolgt von 0.64 der neuen Methode N-Cluster ($c = 0.2$).

Je nach Datensatz variieren die Ergebnisse deutlich. In Abbildung 5.2 sind daher Boxplots des ARI für die besten Einstellungen pro Algorithmus aus Tabelle 5.2 dargestellt. Beispielsweise ist die Streuung von N-Cluster in der Variante ohne DISMS2-Filter hoch. Die mittleren 50% der ARI-Werte liegen zwischen 0.49 und 0.57. Im Vergleich dazu liegen für N-Cluster mit DISMS2-Filter unteres Quartil (0.64) und oberes Quartil (0.67) näher beieinander, sodass der Interquartilsabstand nur 0.03 beträgt. Zum Vergleich der Algorithmen ist ein Vergleich der Anzahl der Cluster von Bedeutung, da es für Clusterverfahren üblich ist, die Clusterlösung für eine feste Clusteranzahl anzugeben. Daher sind in Abbildung 5.3 Boxplots der Clusteranzahl für die Algorithmen mit Parametereinstellungen dargestellt, die bezüglich des mittleren ARI optimiert wurden. Die Anzahl der Cluster schwankt für die Clusterlösungen zwischen 24 252 und 34 122. Die Mediane liegen in etwa bei 30 000. Im PRIDE Algorithmus werden im Median 28 783.5 Cluster gebildet. Aus der Annotation der Spektren ergeben sich deutlich weniger Cluster, zwischen 10 129 und 19 572. Exemplarisch ist für den Lauf C2 in Abbildung 5.4 die Höhe des ARI in Abhängigkeit der Clusteranzahl dargestellt. Es ist kein linearer Trend zu erkennen. Der ARI liegt bei Anwendung des Filters höher und die Clusteranzahl streut weniger.

Der adjustierte Rand-Index ist auch nützlich zum direkten Vergleich der unterschiedlichen Clusterlösungen unabhängig von der Annotation. Im Anhang sind Heatmaps dargestellt, die auf mittleren Werten aller 27 Läufe oder nur der 18 annotierten Läufe für Distanzberechnungen mit und ohne DISMS2-Filter basieren (Abbildungen B.9, B.10, B.11 und B.12). Gruppen mit ähnlichen Clusterlösungen sind in einer Matrix hervorgehoben. Die Höhe des mittleren ARI ist mittels einer blauen Farbskala visualisiert. Je dunkler die Blautöne sind, desto höher liegen die Werte. Die einzelnen Zeilen und Spalten repräsentieren unterschiedliche Clusterverfahren. Eine Sortierung der Zeilen erfolgt anhand des links von der Matrix dargestellten

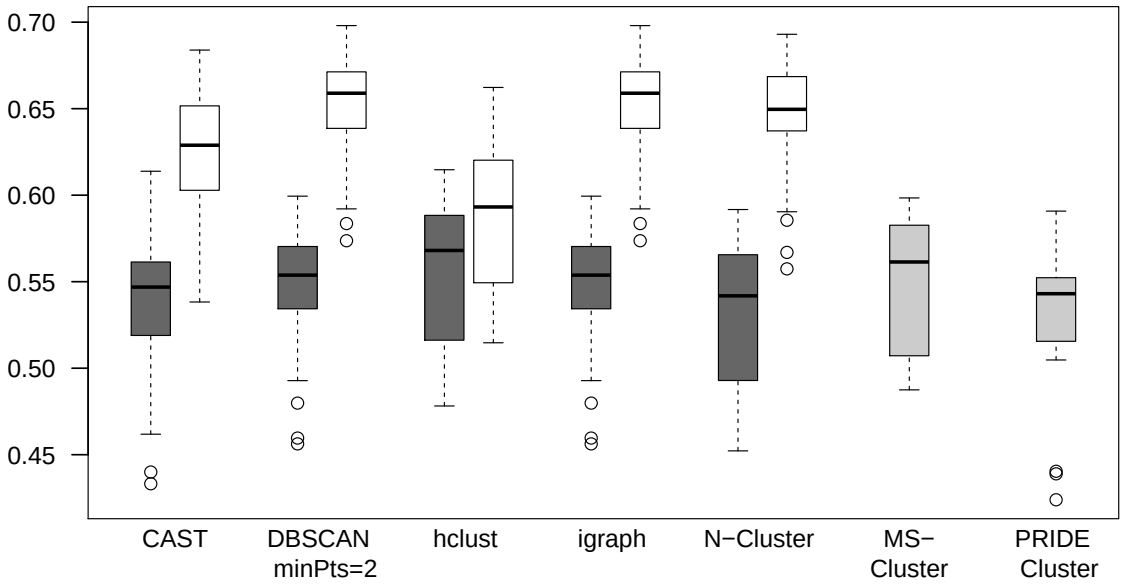

Abbildung 5.2: Boxplots des adjustierten Rand-Indexes zwischen der Peptidannotation und ausgewählten Clusterlösungen von 18 annotierten Läufen. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

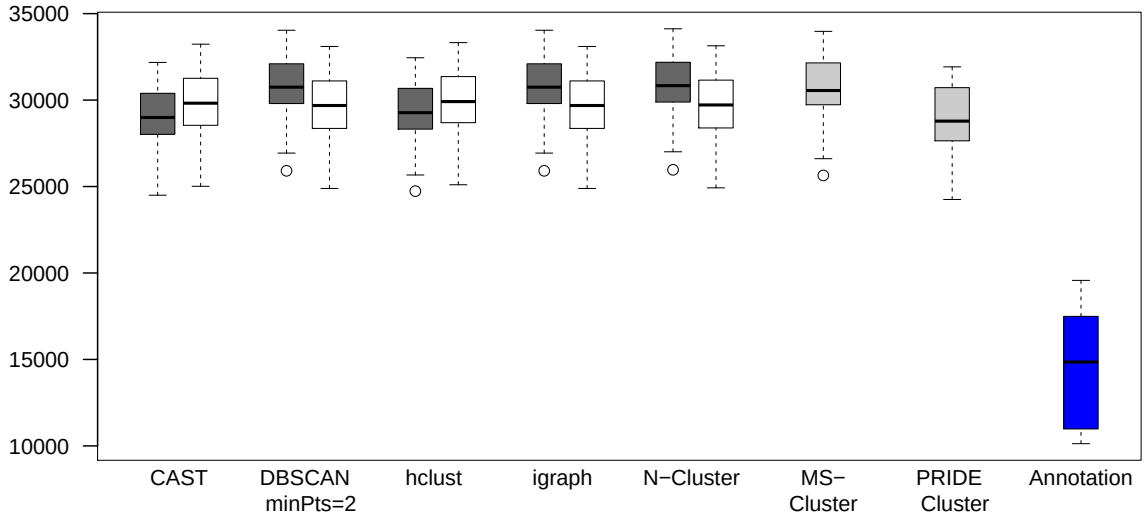

Abbildung 5.3: Boxplots der Clusteranzahl von ausgewählten Clusterlösungen von 18 annotierten Läufen und von Annotationsclusterlösungen (blau). Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

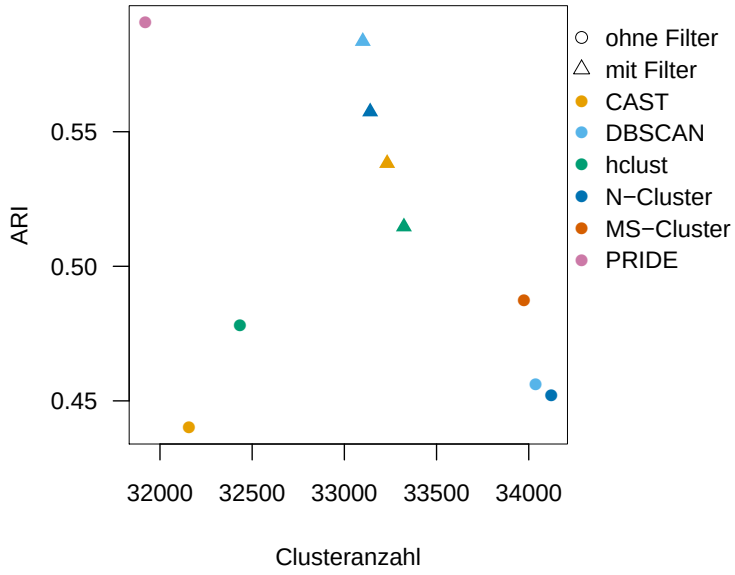

Abbildung 5.4: Adjustierter Rand-Index zwischen der Peptidannotation und den Clusterlösungen von Lauf C2 in Abhängigkeit der Anzahl an Clustern. Die Kosinus-Distanz wurde ohne und mit DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster.

Dendrogramms, das mittels eines hierarchischen Clusterverfahrens mit Complete-Linkage erstellt wurde. Es sind keine großen Unterschiede zwischen den Heatmaps zu erkennen, die auf gemittelten ARI-Werten aller 27 Läufe oder nur der 18 annotierten Läufe basieren. Auffällig ist der Einfluss der Distanzberechnung auf die Clusterlösung, denn die Anwendung des DISMS2-Filters führt zu höheren Ähnlichkeiten zwischen den Algorithmen. Abbildung 5.5 zeigt eine Heatmap für die sieben Algorithmen mit je optimierten Parametereinstellungen. Bei der Aufteilung in drei Cluster werden die unterschiedlichen Distanzberechnungen voneinander getrennt. Das auf Annotationen basierende Clusterverfahren bildet ein Einzelcluster. Die beiden Clusterverfahren MS-Cluster und PRIDE, die eine ähnliche Distanzberechnung im Algorithmus beinhalten, bilden das zweite nicht so homogene Cluster. Im dritten Cluster sind die Algorithmen enthalten, für die Kosinus-Distanzen mit DISMS2-Filter berechnet wurden. Innerhalb des dritten Clusters ergeben sich in der Hierarchie zwei Untergruppen. Das hierarchische Clusterverfahren weist eine hohe Ähnlichkeit zu CAST auf. Das neue einfache Verfahren N-Cluster ist sehr ähnlich zu einer hierarchischen Clusterlösung mit Single-Linkage (DBSCAN mit $\epsilon = 0.2$ und

$minPts = 2$ oder igraph mit $cdis = 0.2$). Zu letzteren drei Verfahren ist CAST allerdings auch ähnlich.

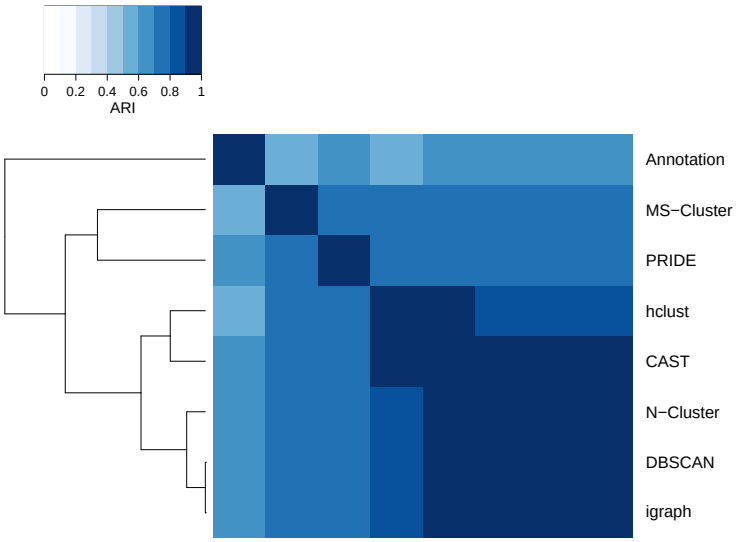

Abbildung 5.5: Heatmap des mittleren adjustierten Rand-Index von 18 Läufen im Datensatz DISMS2 für ausgewählte Clusterlösungen. Alle Lösungen (ausgenommen MS-Cluster und PRIDE Cluster) wurden auf Basis des DISMS2-Filters und der Kosinus-Distanz von Spektren generiert. Es sind nur die sieben Algorithmen mit je optimierten Parametereinstellungen dargestellt.

Im Folgenden wird nur noch ein Teil der Clusterlösungen analysiert. Je Clusteralgorithmus werden nur noch die Parameter mit dem höchsten mittleren ARI aus Tabelle 5.2 berücksichtigt. Die Reinheit der Cluster ist der Anteil der Spektren je Cluster mit der häufigsten Annotation. Die durchschnittliche Reinheit ist in Abbildung 5.6 in Abhängigkeit der Clustergröße dargestellt. Bei fast allen Algorithmen ist die durchschnittliche Reinheit hoch, selbst bei großen Clustergrößen. Die neue Methode N-Cluster ($c = 0.05$), die einen Verkettungseffekt von Spektren vermeidet, eignet sich gut für größere Cluster mit Werten über 0.9. Die Werte von Clustern des CAST-Algorithmus ($t = 0.1$) hingegen nehmen mit zunehmender Clustergröße ab. Die Anwendung des DISMS2-Filter führt meist zu besseren Ergebnissen, insbesondere Cluster von CAST ($t = 0.2$) profitieren davon. Nur bei sehr großen hierarchischen Clustern ($h = 0.2$) mit einer Größe zwischen 32 und 63 liegt die durchschnittliche Reinheit unter 0.7.

Ein Scatterplot des Anteils verbleibender Annotationen nach der Clusterbildung in Abhängigkeit der Clusteranzahl in Relation zur Anzahl an Spektren ist in Abbildung 5.7 dargestellt. Einerseits sollen keine Annotationen verloren gehen, anderer-

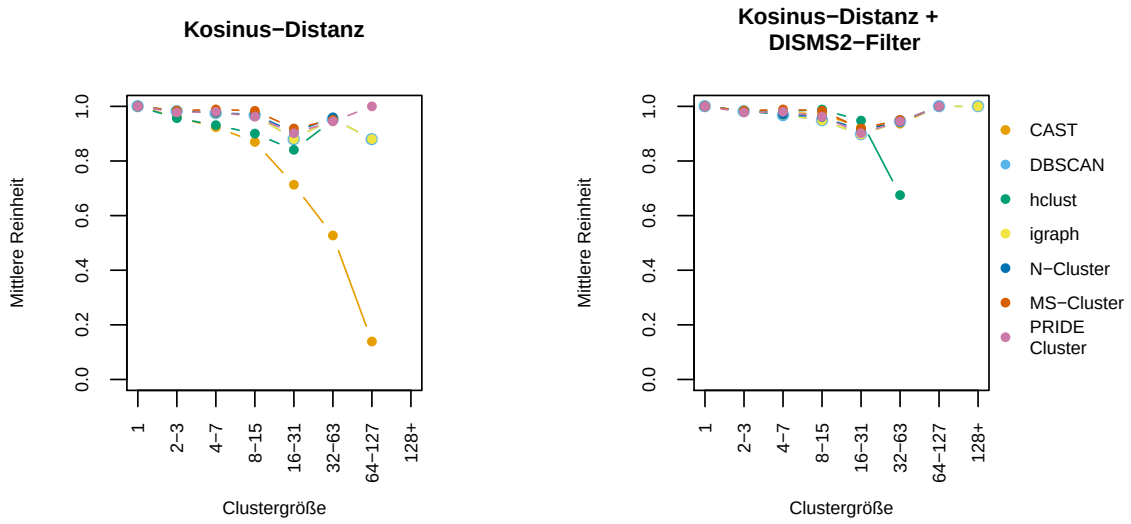

Abbildung 5.6: Reinheit gruppiert nach Clustergröße der Clusterlösungen mit bestem durchschnittlichen ARI je Clusteralgorithmus basierend auf Kosinus-Distanzen der Spektren (links) und zusätzlich mit dem DISMS2-Filter (rechts) (in Anlehnung an Rieder et al., 2017b, Abbildung 2).

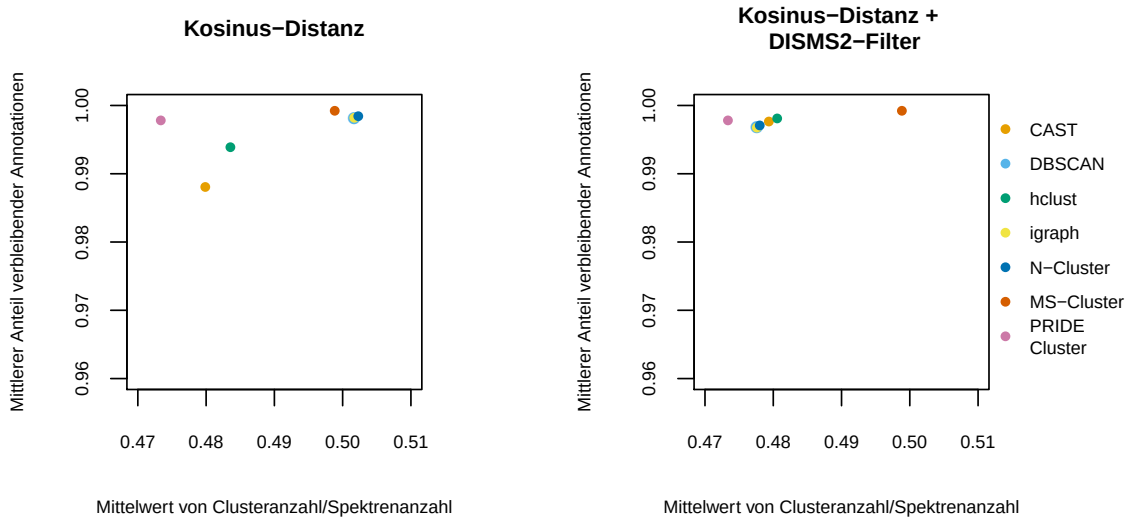

Abbildung 5.7: Mittlerer Anteil verbleibender Annotationen in Abhängigkeit des Mittelwerts der Clusteranzahl in Relation zur Spektrenanzahl (in Anlehnung an Rieder et al., 2017b, Abbildung 3).

seits ist die Bildung möglichst großer Cluster wünschenswert. Liegen die zugehörigen Punkte möglichst weit links oben im Scatterplot, so werden beide Maßzahlen optimiert. Die Werte wurden jeweils über 18 annotierte Läufe gemittelt. Bei allen Algorithmen wird im Durchschnitt die Anzahl der Spektren etwa halbiert. Dabei geht nur ein sehr kleiner Anteil an Annotationen verloren. Über 98% der Peptide bleiben erhalten. PRIDE (*threshold_end*) hat einen Anteil von 99.8% verbleibender Annotationen bei einem Anteil von 47.3% verbleibender Spektren. Wird zusätzlich der DISMS2-Filter angewendet (Abbildung 5.7, rechts), sind die Ergebnisse sehr ähnlich. Die Werte auf der y-Achse liegen nun sogar über 0.99. Zum Beispiel werden für N-Cluster ($c = 0.2$) 99.7% der Annotationen beibehalten, bei einem Anteil von 47.8% der Spektren. Der DISMS2-Filter führt dazu, dass mehr Peptidannotationen erhalten und weniger Spektren übrig bleiben. Da sich die Werte je Lauf unterscheiden, sind Boxplots der Clusteranzahl in Relation zur Spektrenanzahl (Abbildung B.13) und des Anteils verbleibender Annotationen (Abbildung B.14) dargestellt. Die Streuung der Clusteranzahl in Relation zur Spektrenanzahl ist in jedem Verfahren hoch. Je nach Lauf liegt der Quotient zwischen 0.357 und 0.609. Beim Anteil verbleibender Annotationen hingegen gibt es fast keine Streuung. Beispielsweise liegt der Interquartilsabstand für MS-Cluster unter 0.0003.

Ein Scatterplot von Durchschnittswerten des mehrelementigen Clusteranteils in Abhängigkeit des Spektrenanteils ohne häufigste Annotation ist in Abbildung 5.8 abgebildet. Analog zu Abbildung 5.7 sind Punkte, die oben links liegen, wünschenswert. Da die meisten Spektren Einzelcluster bilden (siehe Abbildung B.15), ist der mehrelementige Clusteranteil niedrig (25%-32%). PRIDE ($threshold\_end = 0.8$) erzeugt hohe Werte (32.4%) und der Spektrenanteil ohne häufigste Annotation ist niedrig (2.5%). Der DISMS2-Filter hat einen positiven Einfluss auf den Spektrenanteil ohne häufigste Annotation. Beispielsweise sinkt bei CAST der Anteil von 7% auf 2%. Da die Werte je nach Lauf variieren, sind im Anhang Boxplots des mehrelementigen Clusteranteils (Abbildung B.16) und des Spektrenanteils ohne häufigste Annotation (Abbildung B.17) dargestellt. Die Streuung des mehrelementigen Clusteranteils ist für alle Verfahren in etwa konstant. Der Anteil liegt zwischen 0.210 und 0.383. Die Streuung des Spektrenanteils ohne häufigste Annotation ist viel kleiner. Mit der Ausnahme von zwei Verfahren (CAST und hclust ohne DISMS2-Filter) liegen die Werte unterhalb von 0.03. Durch die Anwendung des DISMS2-Filters sinkt der mediane Anteil beim hierarchischen Clusterverfahren von 0.057 auf 0.018 und der Interquartilsabstand von 0.018 auf 0.004.

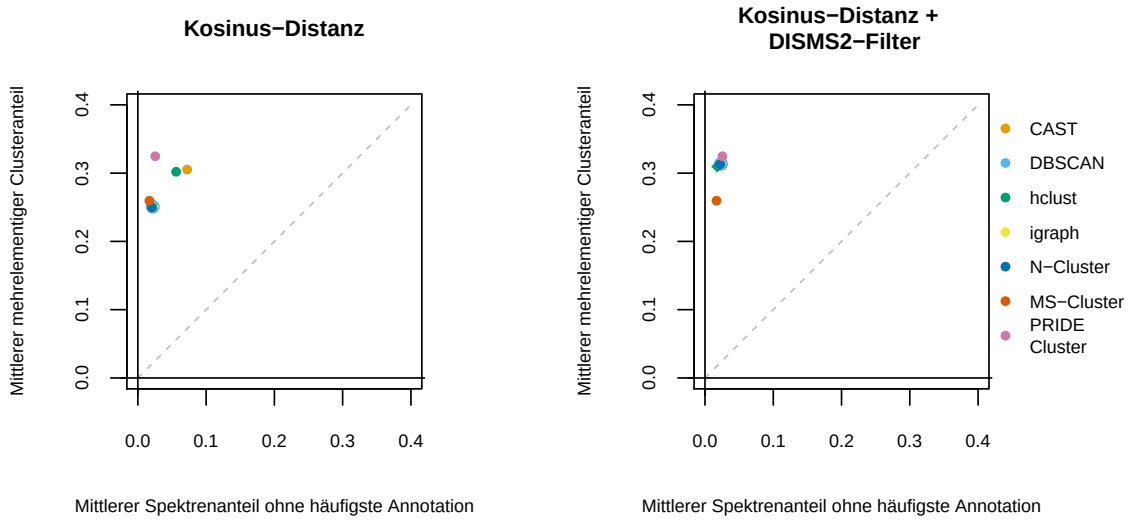

Abbildung 5.8: Mittlerer mehrelementiger Clusteranteil in Abhängigkeit des mittleren Spektrenanteils ohne häufigste Annotation von Clusterlösungen basierend auf Kosinus-Distanzen der Spektren (links) und zusätzlich mit dem DISMS2-Filter (rechts) (in Anlehnung an Rieder et al., 2017b, Abbildung 4).

In Tabelle 5.3 ist ein Teil der Ergebnisse zusammengefasst. Für jeden Clusteralgorithmus wurden die Parametereinstellungen so gewählt, dass der mittlere adjustierte Rand-Index maximal ist. Bei fünf Verfahren, deren Eingabe eine Distanzmatrix ist, führen Kosinus-Distanzen mit DISMS2-Filter im Mittel zu höheren ARI-Werten. Zusätzlich sind die mittleren Werte von h_{clu}, h_{inc}, h_{rem} und k/n aufgeführt. Im Vergleich aller Maße gibt es keinen eindeutigen Sieger. Während mittels DBSCAN der höchste mittlere ARI-Wert generiert wird, ist bei PRIDE der mehrelementige Clusteranteil am höchsten und die Clusteranzahl in Relation zur Spektrenanzahl am kleinsten. Mittels MS-Cluster ist der Anteil verbleibender Annotationen am größten. Der Spektrenanteil ohne häufigste Annotation liegt bei MS-Cluster (0.017) deutlich unter den Anteilen für DBSCAN (0.023) und PRIDE (0.025). Die zugehörigen Standardfehler zu Tabelle 5.3 sind in Tabelle 5.4 enthalten. Die Unterschiede zwischen den Algorithmen sind sehr klein. Bei k/n liegt der Standardfehler bei 0.020. Für ARI (0.010) und h_{clu} (0.008) sind die Werte geringer. Am kleinsten ist der Standardfehler bei h_{inc} (0.001) und h_{rem} (< 0.001).

Zusammenfassend ist in diesem Abschnitt festzustellen, dass Peptidannotationen mit Clusterlösungen von Tandem-Massenspektren des DISMS2-Datensatzes ver-

Tabelle 5.3: Zusammenfassung der gemittelten Werte von Gütemaßen für Clusterverfahren mit optimalen Parametereinstellungen und DISMS2-Filter einzelner Läufe. Die Qualität der Clusterlösungen wird anhand des adjustierten Rand-Index (ARI), dem mehrelementigen Clusteranteil (h_{clu}), dem Spektrenanteil ohne häufigste Annotation (h_{inc}), dem Anteil verbleibender Annotationen (h_{rem}) und der Clusteranzahl in Relation zur Spektrenanzahl (k/n) bewertet. * DBSCAN (0.2, 2) entspricht igraph (0.2).

	ARI	h_{clu}	h_{inc}	h_{rem}	k/n
CAST (0.2)	0.623	0.311	0.020	0.998	0.479
DBSCAN (0.2, 2)*	0.648	0.314	0.023	0.997	0.478
hclust (0.2)	0.590	0.310	0.018	0.998	0.481
N-Cluster (0.2)	0.642	0.313	0.022	0.997	0.478
MS-Cluster (0.8)	0.552	0.260	0.017	0.999	0.499
PRIDE (0.8)	0.529	0.324	0.025	0.998	0.473

Tabelle 5.4: Standardfehler bezüglich der Werte in Tabelle 5.3 für Clusterverfahren mit optimalen Parametereinstellungen und DISMS2-Filter einzelner Läufe. * DBSCAN (0.2, 2) entspricht igraph (0.2).

	ARI	h_{clu}	h_{inc}	h_{rem}	k/n
CAST (0.2)	0.010	0.008	0.001	< 0.001	0.020
DBSCAN (0.2, 2)*	0.009	0.008	0.001	< 0.001	0.020
hclust (0.2)	0.010	0.008	0.001	< 0.001	0.020
N-Cluster (0.2)	0.010	0.008	0.001	< 0.001	0.020
MS-Cluster (0.8)	0.009	0.007	0.001	< 0.001	0.020
PRIDE (0.8)	0.011	0.009	0.001	< 0.001	0.020

schiedener Algorithmen und unterschiedlicher Parametereinstellungen verglichen wurden. Besonders die Ähnlichkeit von Clusterlösungen und Datenbank-Annotationen steht im Fokus der Analyse, da die Peptidannotationen einen Hinweis auf die wahre Clusterlösung geben. Es konnte gezeigt werden, dass etablierte Methoden und der neu vorgestellte Neighbor-Clustering-Algorithmus (N-Cluster) mindestens genauso gut sind wie die aus der Proteomik stammenden Verfahren MS-Cluster und PRIDE Cluster. Eine deutliche Verbesserung der Clusterlösungen wird durch den DISMS2-Filter erreicht. Die Berücksichtigung von Precursorladung, Precursormasse und Retentionszeit führen zu ähnlichen Ergebnissen wie die Clusterlösung der Datenbankannotationen. Zunächst wurden die Parametereinstellungen bezüglich des höchsten mittleren adjustierten Rand-Indexes im Vergleich zur Peptidannotation

optimiert. Keines der Verfahren ist im Mittel optimal bezüglich aller betrachteten Gütemaße. Der mehrelementige Clusteranteil ist bei PRIDE am größten und auch die Clusteranzahl in Relation zur Spektrenanzahl am kleinsten. DBSCAN ist maximal bezüglich der ARI-Werte. Den geringsten Spektrenanteil ohne häufigste Annotation und den größten Anteil verbleibender Annotationen liefert MS-Cluster. Auch die Streuung der Werte wurde berücksichtigt. Zwischen den Algorithmen gibt es nur geringe Unterschiede und insgesamt liegen die Standardfehler unter 0.02. Die Reproduzierbarkeit der Analyse ist gewährleistet, da die Implementierung der Clusterverfahren frei verfügbar ist. Eine Erweiterung auf andere Clusteralgorithmen, andere Distanzberechnungen von Tandem-Massenspektren und die Erweiterung der Evaluierungsmethoden ist außerdem möglich. Aufgrund der Limitierung der Laufzeit und des maximalen Speicherverbrauchs wurden einzelne MS/MS-Läufe analysiert und eine Vorauswahl für die Parametereinstellungen der Algorithmen getroffen. Eine Erweiterung auf mehrere Läufe ist in Abschnitt 5.3 zu finden und eine Analyse einzelner Cluster ist im folgenden Abschnitt 5.2 enthalten. Analog könnte der Vergleich auf andere Datensätze erweitert werden. Von den vorliegenden Daten kommt der Foraminiferen-Datensatz in Frage, der für einige Spektren De-Novo-Annotationen beinhaltet. Da je Lauf jedoch nur 13.1% bis 22.1% der Spektren annotiert sind, ist die Datenqualität für eine Evaluation der Algorithmen zu gering.

5.2 Interpretation einzelner Cluster

Die zuvor generierten Cluster werden in diesem Abschnitt näher beschrieben. Die Clusterlösungen, die auf Annotationen basieren, wurden implizit bereits im Überblick über den DISMS2-Datensatz in Abschnitt 4.1 erläutert. Im letzten Abschnitt sind die Clusteranzahl (siehe Abbildung 5.3) und die Clustergrößen (siehe Abbildung B.15) beschrieben. In Abbildung 5.9 sind Boxplots der maximalen Clustergröße dargestellt. Cluster, die auf Annotionen basieren, beinhalten im Median 50.5 Spektren und maximal 101. Für Verfahren ohne DISMS2-Filter werden meist deutlich größere Cluster gebildet. Bei DBSCAN ohne DISMS2-Filter liegt der Median der Clustergröße bei 113. Die Anwendung des Filters führt zu deutlich kleineren Clustern, die im Median mit 49.5 in etwa so viele Spektren wie die Annotationscluster beinhalten. Der Zusammenhang zwischen den gebildeten Clustern und den vorliegenden Annotationen wird mittels des ARI, der Reinheit, des Anteils verbleibender Annotationen und des Spektrenanteils ohne häufigste Annotation beschrieben. Es stellt sich die Frage, in wie vielen Clustern eine Aussage über die Annotation gemacht werden

kann. In Abbildung 5.10 ist ein Überblick über die Anzahl der mehrelementigen Cluster aufgeführt, die mindestens eine Annotation enthalten. Die Anzahl liegt für PRIDE Cluster im Median bei 3 175.5 und ist somit im Vergleich am größten. Die Streuung ist jedoch wie auch bei den anderen Verfahren groß, denn die Anzahl liegt insgesamt zwischen 2 249 und 4 116. Die Verwendung des Filters führt bei CAST und hclust im Median dazu, dass die Anzahl sinkt. Bei N-Cluster liegt die Anzahl im Median bei 2 356.5 ohne DISMS2-Filter und bei 2 785.5 mit DISMS2-Filter. Eine Erhöhung der Anzahl ist auch bei DBSCAN (igraph) zu beobachten. Im Folgenden werden ausgewählte einzelne Cluster interpretiert.

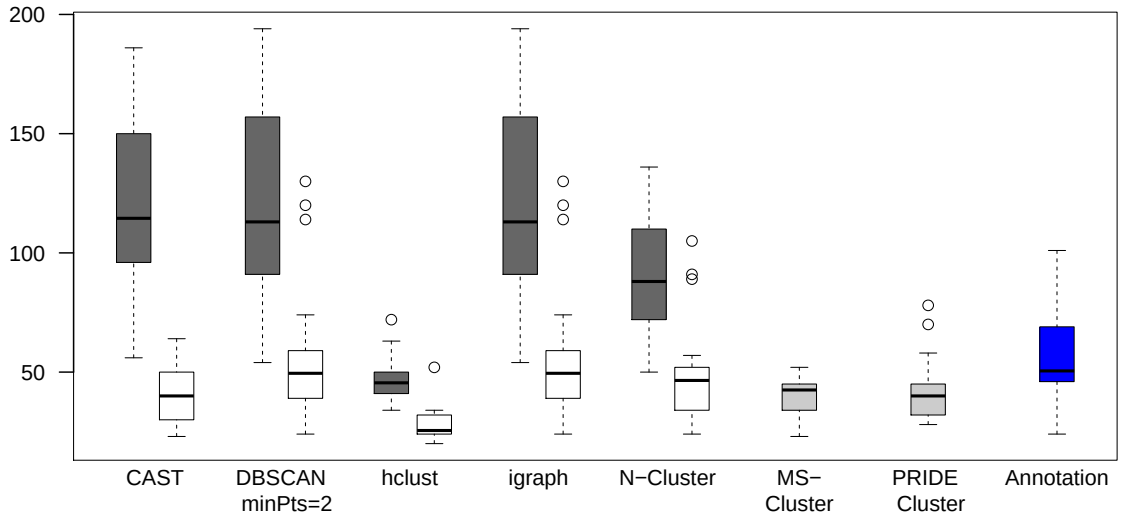

Abbildung 5.9: Boxplots der maximalen Clustergröße von ausgewählten Clusterlösungen von 18 annotierten Läufen und von Annotationsclusterlösungen (blau). Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

Ein wichtiges Ziel bei der Clusteranalyse von Massenspektren ist die Qualitätskontrolle. Im Vergleich von Clusterlösungen, die auf Basis der Kosinus-Distanz von Tandem-Massenspektren erstellt wurden, liegt der mittlere ARI für eine hierarchische Clusteranalyse mit Parameter $h = 0.1$ am höchsten (siehe Abschnitt 5.1). Zunächst werden zwei Cluster Clus_1 und Clus_2 des Laufs H1 analysiert, die mithilfe des Algorithmus hclust ($h = 0.1$) erstellt wurden. Das viertgrößte Cluster Clus_1 enthält ein fälschlicherweise nicht annotiertes Spektrum und die Spektren im mittelgroßen Cluster Clus_2 sind entweder nicht annotiert oder mit unterschiedlichen Peptiden annotiert. Clus_1 beinhaltet 24 Spektren. Davon ist ein Spektrum

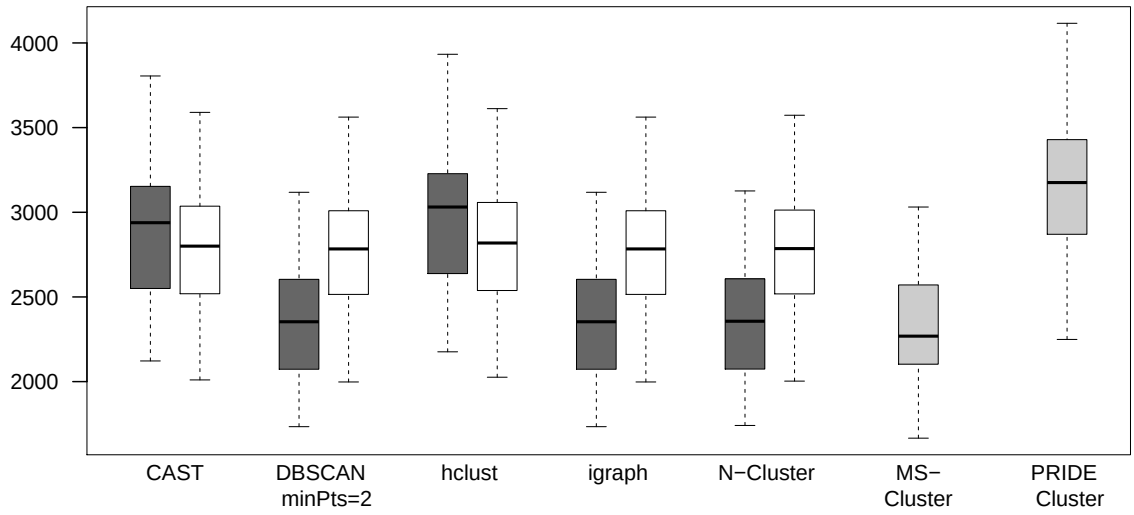

Abbildung 5.10: Boxplots der Anzahl mehrelementiger Cluster mit mindestens einer Annotation von ausgewählten Clusterlösungen von 18 annotierten Läufen. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

nicht annotiert. Alle anderen Spektren sind mit einer von zwei Peptiden, entweder HQGVmVGMGQK oder HQGVMVGmGQK, annotiert. Diese unterscheiden sich nur bezüglich der Oxidation von Methionin an der fünften oder achten Position des Peptids. Auffällig ist, dass das unannotierte Spektrum große Ähnlichkeit zu den anderen annotierten Spektren aufweist (Abbildung 5.11, links). Die maximale Distanz aller Spektrenpaare liegt unter 0.1. Ein Annotationsscore liegt bei 43. Das zugehörige Spektrum liegt etwas abseits der anderen Spektren. Die Scores der Annotationen der übrigen Spektren liegen zwischen 51 und 76. Die nächsten Nachbarn des unannotierten Spektrums sind mit HQGVmVGMGQK annotiert und die Distanz des unannotierten Spektrums zum nächsten Nachbarn liegt bei 0.013 (Abbildung B.18). Eine Zuordnung des Peptids HQGVmVGMGQK zu dem unannotierten Spektrum liegt daher nahe. Die fehlende Annotation ist das Resultat einer falschen Precursormassenkorrektur bei der Datenbankannotation. In der Clusterlösung von H1 mittels hclust($h = 0.1$) gibt es viele ähnlich geartete Cluster wie Clus_1. Insgesamt gibt es 83 Cluster mit mindestens einem unannotierten Spektrum, Mindestclustergröße 5 und maximal drei unterschiedlichen Annotationen.

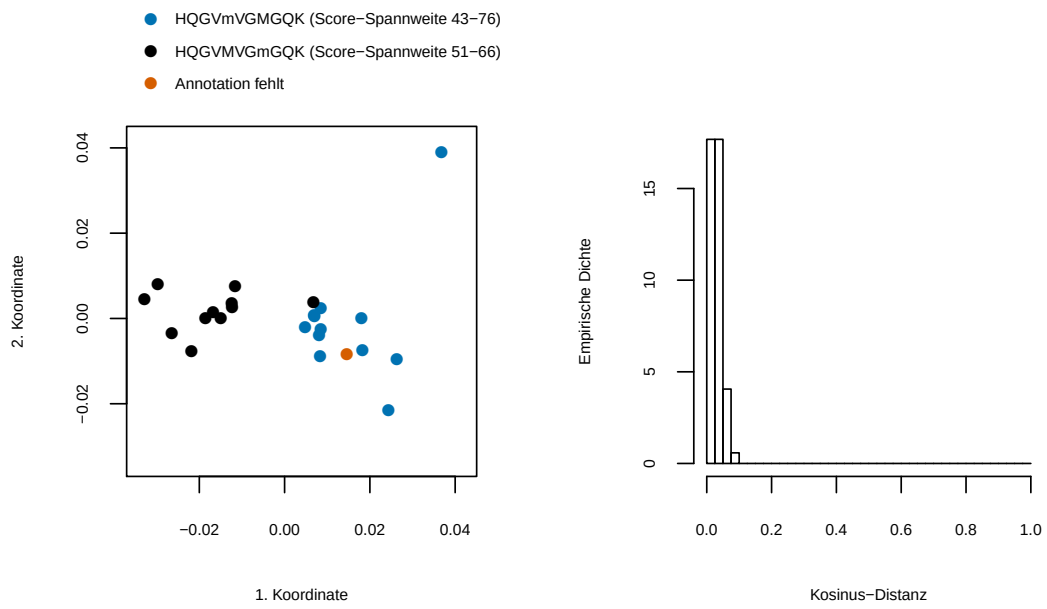

Abbildung 5.11: Grafische Darstellung des viertgrößten Clusters Clus_1 (24 Spektren) der Clusterbildung von H1 mit hierarchischem Clustern ($h = 0.1$), visualisiert durch MDS (links) inklusive Scores (Mascot) und durch ein Histogramm (rechts) aller paarweisen Kosinus-Distanzen der 24 Spektren (in Anlehnung an Rieder et al., 2017b, Abbildung 5).

13 Spektren liegen in Clus_2. Davon sind sieben nicht annotiert und die anderen Spektren weisen unterschiedliche Annotationen auf (Abbildung 5.12). Die Scores der Annotationen sind niedrig. Die Werte liegen zwischen 28 und 48. Jede Sequenz umfasst maximal 8 Aminosäuren und beginnt mit Phenylalanin (F). Zwar sind die Distanzen klein, aber die Precursormassen zwischen 619 und 1393 Dalton variieren. Es fällt auf, dass alle Spektren einen sehr großen Peak bei 120 m/z, der Masse des Immoniumions Phenylalanin, aufweisen (Abbildung B.19). Der dominante Peak beeinflusst sehr stark die Kosinus-Distanzberechnung der Spektren, sodass die Clusterbildung falsch ist. Der DISMS2-Filter ist ein Ausweg um diese Art an falschen Clustern zu vermeiden, da die Precursormasse bei der Distanzberechnung berücksichtigt wird. Es kommen weitere Cluster wie Clus_2 in der Clusterlösung von H1 mittels hclust($h = 0.1$). Es gibt 169 Cluster mit mindestens einer Annotation und einer Mindestclustergröße 5. In 38 Clustern (22%) liegt die relative Häufigkeit der häufigsten Annotation im Verhältnis zur Spektrenanzahl bei unter einem Drittel. Die relative Häufigkeit von Clus_2 liegt sogar nur bei 7%.

In Abschnitt 5.5 werden Cluster untersucht, in denen ein Teil der Spektren annotiert ist und ein anderer Teil der Spektren nicht annotiert ist. Die Spektren ohne

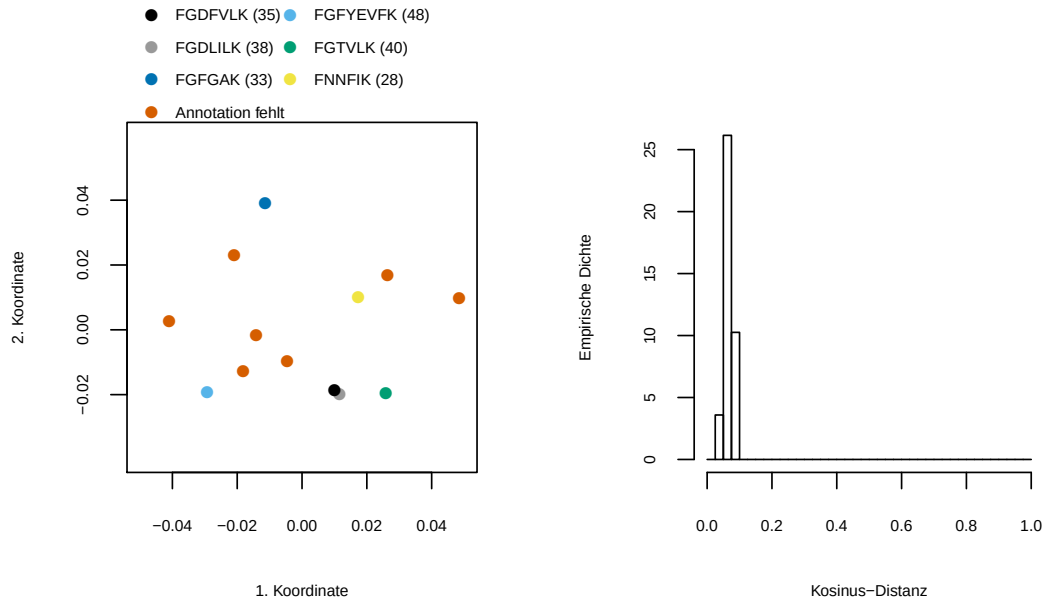

Abbildung 5.12: Grafische Darstellung des Cluster Clus_2 (13 Spektren) der Clusterbildung von H1 mit hierarchischem Clustern ($h = 0.1$), visualisiert durch MDS (links) inklusive Scores (Mascot) und durch ein Histogramm (rechts) aller paarweisen Kosinus-Distanzen der 13 Spektren (in Anlehnung an Rieder et al., 2017b, Abbildung 6).

Annotation sollen mittels Clusterbildung nachträglich einem Peptid des Clusters zugeordnet werden. Dabei werden Cluster mit einer und mehreren Peptidannotationen unterschieden. Beispiele für Cluster mit nur einer Annotation sind in den Abbildungen 5.13 und B.31 dargestellt. Das in Abbildung 5.13 abgebildete Cluster umfasst 17 Spektren. Sechs der Spektren sind dem Peptid VIAHTQmK zugewiesen, die übrigen neun Spektren sind nicht annotiert. Es fällt auf, dass die paarweisen Distanzen klein sind und in der MDS-Grafik keine Trennung der Spektren mit und ohne Annotation vorliegt. Hingegen liegt in dem in Abbildung B.31 dargestellten Cluster ein nicht annotiertes Spektrum getrennt von 14 weiteren Spektren mit der Annotation HQGVmVGmGQK. Der Abstand ist allerdings nicht sehr groß, denn alle paarweisen Distanzen liegen unter 0.2.

In den Abbildungen 5.14, B.32 und B.33 sind Cluster mit mehreren Annotationen abgebildet. Das in Abbildung 5.14 dargestellte Cluster umfasst fünf Spektren, wovon drei mit langen Peptiden und zwei nicht annotiert sind. Zwei der Spektren sind mit dem Peptid NSQEDDDLTIGASPDAGLAFHFVQPSDANVVR annotiert und ein weiteres mit dem Peptid NHGEDEEVTEQVELAAMETEASDSIVDNVPK. Die paarweisen Distanzen sind zum Teil größer, denn es werden auch Werte

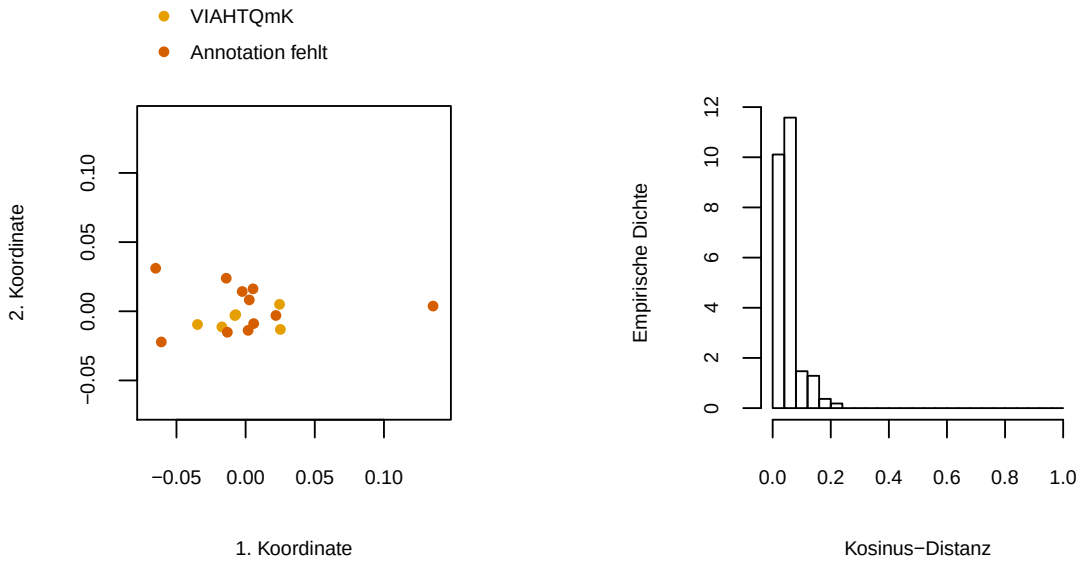

Abbildung 5.13: Grafische Darstellung eines Clusters mit 17 Spektren der Clusterbildung von C2 mit DBSCAN ($\epsilon = 0.2, minPts = 2$), visualisiert durch MDS (links) und durch ein Histogramm (rechts) aller paarweisen Kosinus-Distanzen der 17 Spektren.

um 0.4 angenommen. Die Annotationen des Clusters aus Abbildung B.32 bestehen aus wenigen Aminosäuren. Je einem Spektrum ist das Peptid IIQLK oder IQIIK zugewiesen. Die restlichen vier der insgesamt sechs Spektren im Cluster sind nicht annotiert. Da Leucin und Isoleucin im Massenspektrometer nicht unterschieden werden können, unterscheiden sich die Aminosäuresequenzen nur an der zweiten und dritten Position durch die Vertauschung der Aminosäuren I und Q. Es fällt auf, dass in der MDS-Grafik die annotierten Spektren von den anderen umschlossen werden. Insgesamt liegen die Spektren nah beieinander, denn die paarweisen Distanzen liegen unter 0.2. In Abbildung B.33 ist ein Cluster mit ähnlichen Eigenschaften abgebildet. Zwei Spektren sind mit IEIIK annotiert und ein weiteres Spektrum mit der fast identischen Aminosäuresequenz IIEIK. Im Vergleich ist ebenfalls die zweite und dritte Position vertauscht. Ebenfalls sind die annotierten Spektren von den anderen Spektren in der MDS-Grafik umgeben und die paarweisen Distanzen sind klein.

Zusammenfassend ist festzustellen, dass einzelne Cluster interpretiert wurden. Das Cluster Clus_1 der Größe 24 beinhaltet ein Spektrum, das wegen eines Fehlers der Software zur Datenbankannotation nicht annotiert wurde. In dem Cluster Clus_2 wurden Spektren mit sehr unterschiedlichen Annotationen zusammengefasst. Ein

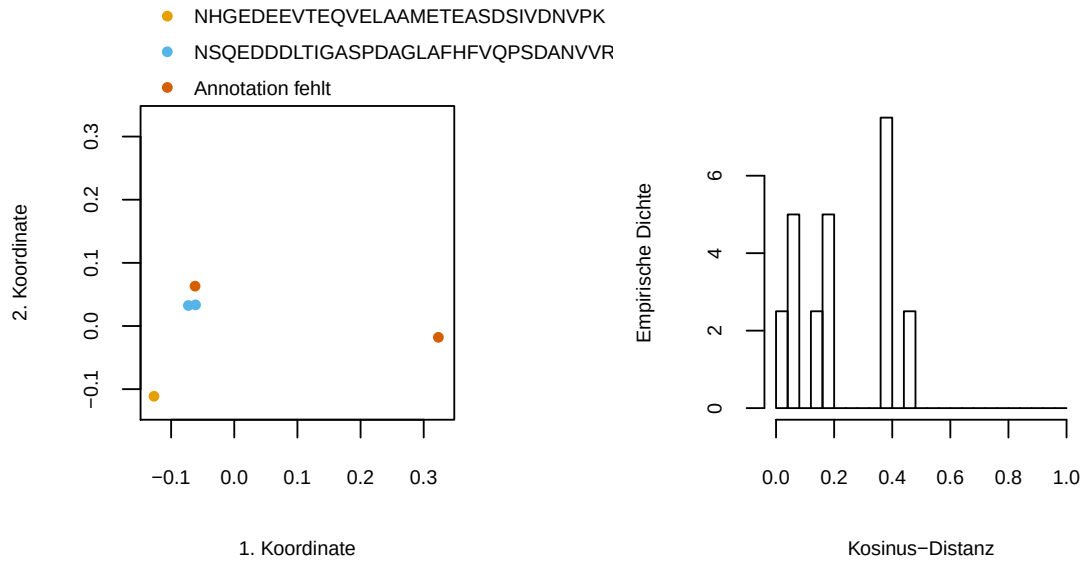

Abbildung 5.14: Grafische Darstellung eines Clusters mit 5 Spektren der Clusterbildung von C2 mit PRIDE Cluster ($threshold\_end = 0.8$), visualisiert durch MDS (links) und durch ein Histogramm (rechts) aller paarweisen Kosinus-Distanzen der 5 Spektren.

hoher Peak eines Immoniumions in den Spektren hat fälschlicherweise zu kleinen Kosinusdistanzen geführt. Spektren, die bis auf einen Peak wenig Ähnlichkeit aufweisen, bilden daher fälschlicherweise ein Cluster. Weitere Cluster mit einer Annotation oder mehreren Annotationen wurden beschrieben, die in Abschnitt 5.5 näher analysiert werden. In einem Verfahren werden nicht annotierte Spektren aufgrund ihrer Lage in einem Cluster nachträglich mit einem Peptid annotiert.

5.3 Clusterbildung von Massenspektren mehrerer Läufe

In den vorherigen Abschnitten des Kapitels 5 wurden einzelne MS/MS-Läufe beurteilt. Allerdings erfordern Anwendungen der Clusterbildung von Massenspektren, beispielsweise die Identifizierung neuartiger Peptide über Artengrenzen hinweg, die gleichzeitige Analyse mehrerer Läufe. Die Analyse großer Datenmengen ist limitiert, da eine hohe Rechengeschwindigkeit und viel Speicherplatz notwendig sind. Anmerkungen zu Laufzeit und Speicherverbrauch sind später in Abschnitt 5.6 zu finden.

Drei technische Replikate je Art (C, D, H, M, Y, und R4) wurden jeweils gleichzeitig analysiert. Bis zu 85 032 annotierte Spektren wurden berücksichtigt (Rieder et al., 2017b, Tabelle S-2). Aufgrund beschränkter Ressourcen wurden die optimierten Parametereinstellungen der Einzelläufe aus Abschnitt 5.1 verwendet. Die Stichprobe der gemessenen Peptide ist also größer, da je Lauf immer nur ein Teil des Proteoms gemessen wird. Die Grundgesamtheit der Peptide einer Probe wird also besser repräsentiert durch die Mehrfachmessungen.

Insgesamt liegen $72 (= 6 \cdot 13 - 6)$ Clusterlösungen vor. Für jedes Triplikat von Läufen einer Art gibt es fünf Lösungen basierend auf Kosinus-Distanzen und fünf Lösungen mit zusätzlichem DISMS2-Filter. Zusammen mit MS-Cluster, PRIDE Cluster und der Annotationsclusterung ergeben sich 13 Lösungen. Bei der hierarchischen Clusterbildung wurde das Speicherlimit für Triplikate von Maus, Mensch und Fadenwurm überschritten, sodass in diesen Fällen keine Lösung vorliegt (siehe Abschnitt 5.6).

Die Größe der Cluster ist erhöht bei mehreren Läufen. In Abbildung B.20 sind Boxplots der maximalen Clustergröße dargestellt. In Clustern, die auf Annotationen basieren, sind bis zu 250 Spektren enthalten. Bei den auf Spektren basierenden Clustern ist die Clustergröße deutlich kleiner. Im Median ist die maximale Größe bei PRIDE mit 79 am geringsten. Am häufigsten werden Cluster mit zwei bis drei Spektren gebildet (siehe Abbildung 5.15). Einzelcluster sind also im Vergleich zu Einzelläufen nicht mehr in der Mehrheit. Dies ist bei technischen Replikaten zu erwarten, da ein Großteil der Peptide in jedem Lauf erfasst wird. Im Idealfall wären die technischen Replikate identisch, sodass jedes Spektrum mindestens dreimal vorkäme und daher jedes Cluster mindestens drei Spektren enthielte.

Die Anzahl mehrelementiger Cluster liegt mit einer Ausnahme, der hierarchischen Clusteranalyse, im Median bei allen Verfahren um 16 000 (Abbildung B.21). Boxplots der Clusteranzahl sind in Abbildung B.22 dargestellt. Es fällt auf, dass die Anzahl der Annotationscluster im Median geringer ist. Es werden zwischen 13 692 und 24 572 Cluster gebildet. Der Medianwert 19 285 liegt beispielsweise deutlich unter dem Medianwert 23 832 von PRIDE Cluster.

Im direkten Vergleich der Einzelläufe mit Triplikaten von Läufen fällt auf, dass die durchschnittlichen ARI-Werte mehrerer Läufe bis zu 14.9% höher liegen (Tabelle A.12). In Abbildung 5.16 sind die durchschnittlichen ARI-Werte aus Tabelle 5.2 dargestellt. Am besten schneidet wieder DBSCAN ($\epsilon = 0.2, minPts = 2$) bzw. igraph ($cdis = 0.2$) ab mit einem Mittelwert von 0.743. PRIDE Cluster, das besonders für große Datensätze konzipiert ist, verbessert sich deutlich auf 0.678. Anhand der

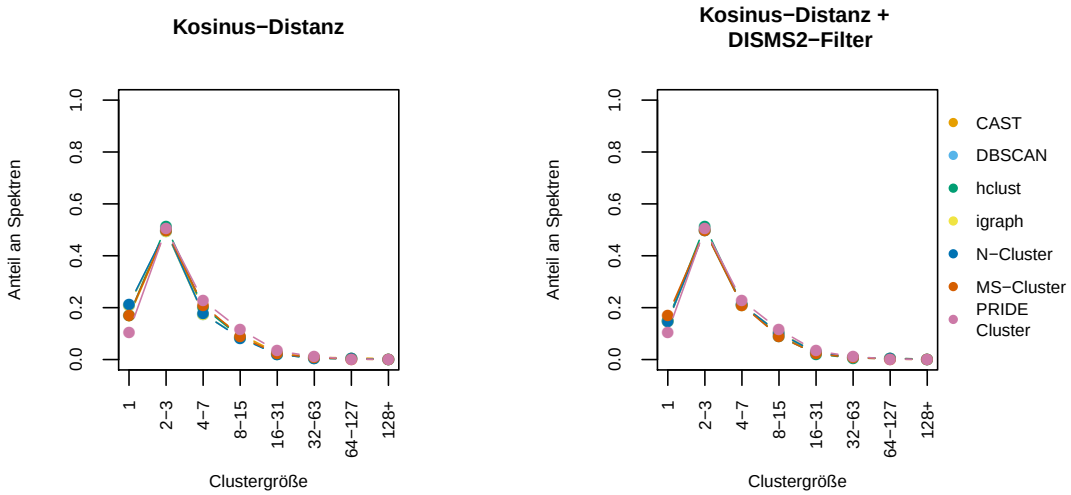

Abbildung 5.15: Anteil Spektren gruppiert nach Clustergröße der Clusterlösungen mehrerer Läufe.

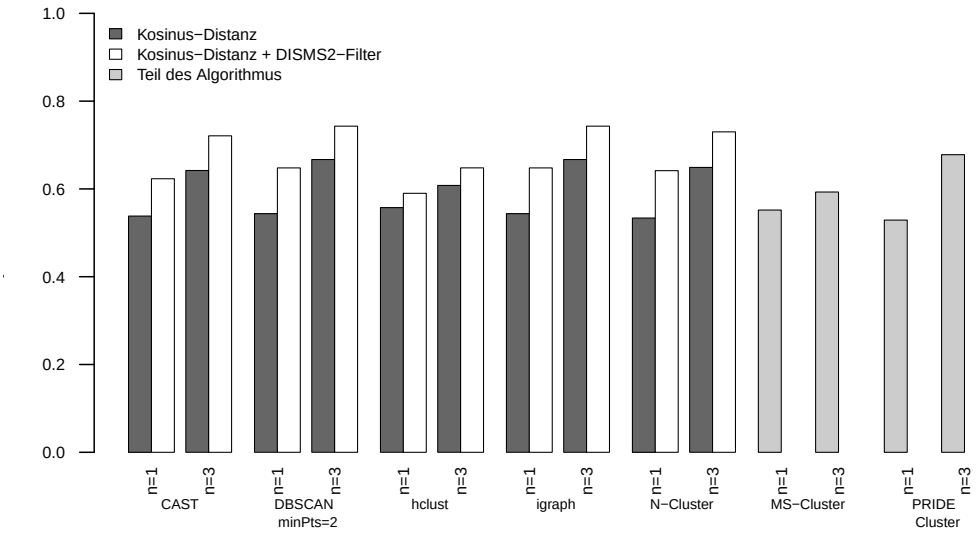

Abbildung 5.16: Durchschnittlicher adjustierter Rand-Index zwischen der Peptidannotation und Clusterlösungen von einzelnen annotierten (n=1) und mehreren (n=3) Läufen. Die Kosinus-Distanz wurde ohne und mit DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (in Anlehnung an Rieder et al., 2017b, Abbildung S-8).

Boxplots in Abbildung B.23 ist die Verteilung der Werte visualisiert. Eine höhere Streuung der Werte liegt nur bei der hierarchischen Clusterbildung und MS-Cluster vor. Bei der Weiterentwicklung von MS-Cluster, PRIDE Cluster, ist die Varianz der Werte geringer.

Die Ergebnisse der durchschnittlichen Reinheit, gruppiert nach Clustergröße, sind ähnlich zu denen für Einzelläufe (Abbildung B.24). Die durchschnittliche Clusteranzahl in Relation zur Anzahl an Spektren sinkt und der durchschnittliche Anteil verbleibender Annotationen ist nur etwas geringer (Tabellen A.13 und A.14, Abbildungen B.25, B.26, B.27). Die technischen Replikate sind im besten Fall identisch. Dies bedeutet, dass die Clustergröße bei mindestens drei liegt. Daraus folgt, dass die Clusteranzahl in Relation zur Spektrenanzahl im Idealfall maximal ein Drittel beträgt. Bei den vorliegenden Clusterlösungen liegt das Verhältnis zwischen 0.338 und 0.459. Der mehrelementige Clusteranteil liegt im Mittel bei mindestens 78% (Abbildung B.28). Trotz dieser deutlichen Steigerung ist der Spektrenanteil ohne häufigste Annotation gering (Tabellen A.15 und A.16, Abbildungen B.29 und B.30).

Zusammenfassend ist festzustellen, dass in diesem Abschnitt die Clusteranalyse auf mehrere Läufe von drei technischen Replikaten des DISMS2-Datensatzes erweitert wurde. Das Proteom einer Art ist also besser repräsentiert durch die Auswertung von Mehrfachmessungen. Das hierarchische Clusterverfahren lieferte für Fadenwurm, Maus und Mensch keine Lösung, da das Speicherlimit von 64 GB überschritten war. In folgenden Analysen ist die hierarchische Clusterung in der aktuellen Implementierung also nicht zu empfehlen. Im Vergleich zur Evaluierung einzelner Läufe in Abschnitt 5.1 ist eine drastische Verbesserung der Werte der Bewertungsmaße zu beobachten. Eine weitere Erweiterung der Analyse ist wünschenswert, um mehrere Millionen Spektren unterschiedlicher Arten gemeinsam zu clustern. Dadurch können Peptide über Artengrenzen hinweg identifiziert werden. Die Umsetzbarkeit scheitert zurzeit bis auf zwei Ausnahmen (MS-Cluster und PRIDE Cluster) an der Begrenzung des Speicherplatzes und der Rechenzeit. Nähere Details dazu sind in Abschnitt 5.6 zu finden. Die Auswertung eines großen Benchmark-Datensatzes sollte durchgeführt werden. Beispielsweise gibt es eine aktuelle Studie von Zolg et al. (2017), die sich mit der Analyse von bereits über 330 000 synthetischen tryptischen Peptiden befasst, die alle kanonischen humanen Genprodukte repräsentieren. In den kommenden Jahren sollen die frei verfügbaren Daten auf über eine Millionen Peptide erweitert werden.

5.4 Verknüpfung der Spektrenclusterung mit DISMS2

Im Ausblick der Veröffentlichung von Rieder et al. (2017b), die den DISMS2-Algorithmus zur Erstellung phylogenetischer Bäume mittels LC-MS/MS-Läufen vorstellt, wurde angemerkt, dass eine vorangestellte Clusterbildung der Spektren zu einer Verbesserung der Ergebnisse beitragen könnte. In der Masterarbeit von Ottenheym (2017) wurde die Rekonstruktion von jenen phylogenetischen Bäumen und dabei unter anderem der Einfluss der in Rieder et al. (2017b) erwähnten Clusterverfahren untersucht. Zur Repräsentantenauswahl, d. h. der Bestimmung eines Repräsentantenspektrums von Clustern, wurden vier Varianten, u. A. Medoide, berücksichtigt. Außerdem wurden drei bekannte distanzbasierte Verfahren zur Rekonstruktion phylogenetischer Bäume, UPGMA, der Neighbor-Joining-Algorithmus und die Minimum-Evolution-Methode analysiert. In einer Datenanalyse der Datensätze DISMS2 und Foraminiferen wurde festgestellt, dass die Repräsentantenauswahl die geringste Auswirkung hat. Ein Einfluss der Clusterverfahren wurde beobachtet, jedoch gibt es keinen großen Unterschied zwischen phylogenetischen Bäumen, die auf dem DISMS2-Algorithmus mit und ohne vorangestellter Clusterbildung basieren.

Im Folgenden wird der Einfluss einer vorangestellten Clusterbildung auf die Distanzen von Läufen, die auf dem DISMS2-Algorithmus beruhen, untersucht. Es wird erneut der DISMS2-Datensatz untersucht. Die Repräsentantenauswahl wird nicht variiert, da sie laut Ottenheym (2017) eine zu vernachlässigende Auswirkung hat. Ein Cluster wird durch den Medoiden repräsentiert. Der Medoid eines Clusters ist jenes Spektrum im Cluster, das die Summe der Distanzen zu allen anderen Spektren im Cluster minimiert (Hastie et al., 2009, S. 155ff.). Es handelt sich also um ein gemessenes Spektrum, für das auch Zusatzinformationen, wie beispielsweise Precursormasse und -ladung, vorliegen. Stellvertretend werden nur zwei Clusterverfahren, PRIDE Cluster und hclust, mit optimierten Parametereinstellungen ($threshold\_end = 0.8$ und $h = 0.1$) betrachtet. Das PRIDE-Clusterverfahren wird von Anwendern aus dem Proteomikbereich häufig verwendet. Es ist im Vergleich der Laufzeit und des Speicherverbrauchs herausragend (siehe Abschnitt 5.6) und bezüglich des adjustierten Randindex konkurrenzfähig. Im Vergleich aller Clusterverfahren schneidet das hierarchische Clustern mit Complete-Linkage sehr gut ab (Rieder et al., 2017b). Unter allen Clusterlösungen, die auf Kosinus-Distanzen ohne zusätzlichen DISMS2-Filter beruhen, ist der adjustierte Rand-Index für die Parametereinstellung $h = 0.1$ für einzelne Läufe im Mittel am größten (siehe Abbildung 5.1). Distanzen mit DISMS2-

Filter sind in diesem Vergleich zu vernachlässigen, da die anschließende Anwendung des DISMS2-Algorithmus diesen Filter impliziert.

Mittels der Clusterlösungen der Algorithmen PRIDE Cluster ($threshold\_end = 0.8$) und hclust ($h = 0.1$) wurden jeweils die Medoide in allen 27 Läufen des DISMS2-Datensatzes bestimmt. Auf die Repräsentantenspektren erfolgt die Anwendung des DISMS2-Algorithmus. Die in Abschnitt 4.2 optimierten Parameter des DISMS2-Algorithmus wurden hierbei verwendet. Zur Beantwortung der Frage, ob eine Repräsentantenauswahl hilfreich ist, wird wie in Abschnitt 4.3 der mittlere Anteil an Partnern innerhalb und zwischen den Gruppen betrachtet (siehe Abbildung 5.17). Zwischen den beiden verwendeten Clusteralgorithmen sind keine relevanten Unterschiede zu erkennen. Ob eine Repräsentantenauswahl durchgeführt wurde, ist zwischen Gruppen in der mittleren Distanz der Läufe nicht zu erkennen. Jedoch unterscheiden sich innerhalb der Gruppen der Algorithmus DISMS2.f und die beiden Varianten DISMS2.f.hclust und DISMS2.f.PRIDE, die zusätzlich eine vorangestellte Clusterbildung beinhalten. Es fällt auf, dass die Distanzen zwischen technischen Replikaten der einzelnen Arten größer sind. Die zugehörigen Standardfehler liegen innerhalb der Arten maximal bei 0.005 und im Vergleich zwischen den Arten unter 0.002. Die leichte Erhöhung der Distanzen innerhalb der Gruppen bei der vorangestellten Clusterbildung wird beispielsweise auch in der Darstellung des Dendrogramms für die DISMS2-Distanzen mit vorangestellter Repräsentantenauswahl mittels des PRIDE-Clusterverfahren in den einzelnen Läufen deutlich (siehe Abbildung 5.18). Die Distanz innerhalb der Gruppen liegt im Mittel bei 0.36. Ansonsten hat die Gestalt des Dendrogramms große Ähnlichkeit mit der Variante ohne Repräsentantenauswahl.

In dieser beispielhaften Datenanalyse konnte DISMS2 durch die Clusterbildung von Massenspektren nicht verbessert werden. Die Repräsentantenauswahl mittels PRIDE Cluster und hclust hat größere Distanzen von Läufen innerhalb der Gruppen zur Folge. Dies ist durch den großen Anteil an Einzelclustern zu erklären. Die Clusterbildung in einzelnen Läufen ist limitiert. Einen Großteil der Cluster bilden einzelne Spektren. Dies ist zu erwarten, da im DISMS2-Datensatz je Lauf durchschnittlich die Hälfte der Peptidannotationen auf ein Spektrum zurückzuführen sind (siehe Abschnitt 4.1). Ein großer Anteil an Einzelclustern ist jedoch auch ein Hinweis darauf, dass viele Spektren von schlechter Qualität sind. Trotz guter Datenbanken und einer fortlaufenden technischen Verbesserung der Generierung von Massenspektren, können viele Spektren nicht annotiert werden. Auch fehlerhafte und verrauschte Spektren sind in jedem Lauf enthalten. Diese Spektren von schlechter Qualität bilden

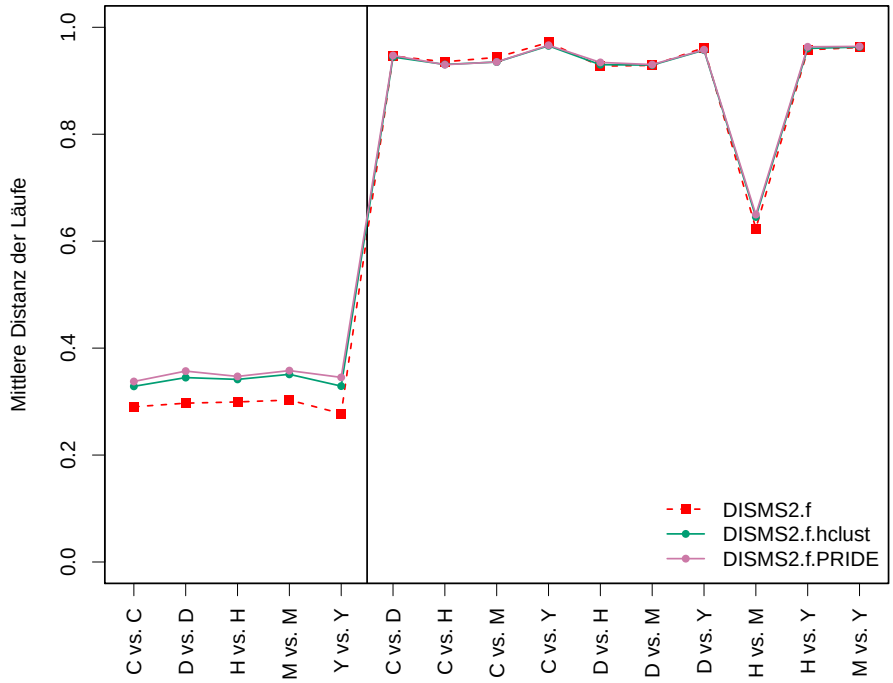

Abbildung 5.17: Mittlerer relativer Anteil an Partnern für unterschiedliche Methoden des Proteomvergleichs. Im Vergleich zu DISMS2.f wird vorweg eine Repräsentantenauswahl mittels hierarchischer (hclust) oder PRIDE-Clusteranalyse durchgeführt.

Einzelcluster, da sie wenig Ähnlichkeit zu anderen Spektren haben. Angenommen, dass qualitativ gute Spektren in einem Cluster zusammengefasst werden, werden gleiche Spektren entfernt. Dadurch wird seltener ein guter Partner zu diesen Spektren gefunden. Folglich ist das Gewicht an der gemittelten Distanz von zwei Läufen verringert. Umgekehrt ist das Gewicht von verrauschten Spektren größer. Letztendlich kann die veränderte Gewichtung dazu führen, dass die Distanzen innerhalb der Gruppen steigen.

Ein möglicher Ausweg wäre die Entfernung von Einzelclustern. Dies würde jedoch einen großen Informationsverlust bedeuten, da auch gute Spektren entfernt würden. Die Hälfte der Peptidannotationen basiert in den vorliegenden Daten auf einem einzelnen Spektrum. Eine weitere Möglichkeit wäre ausschließlich die Analyse nur der annotierten Spektren. Dieses Vorgehen ist allerdings nicht praxisrelevant, da der Algorithmus ohne die Zuhilfenahme einer Datenbank Anwendung findet. Eine gute Idee ist die Evaluation auf größeren Datensätzen. Bei der Clusterbildung mehrerer Läufe sinkt die Anzahl an Einzelclustern (siehe Abschnitt 5.3). Eine umgekehrte Änderung der Distanzen ist daher denkbar. Bei der Clusterung sehr vieler

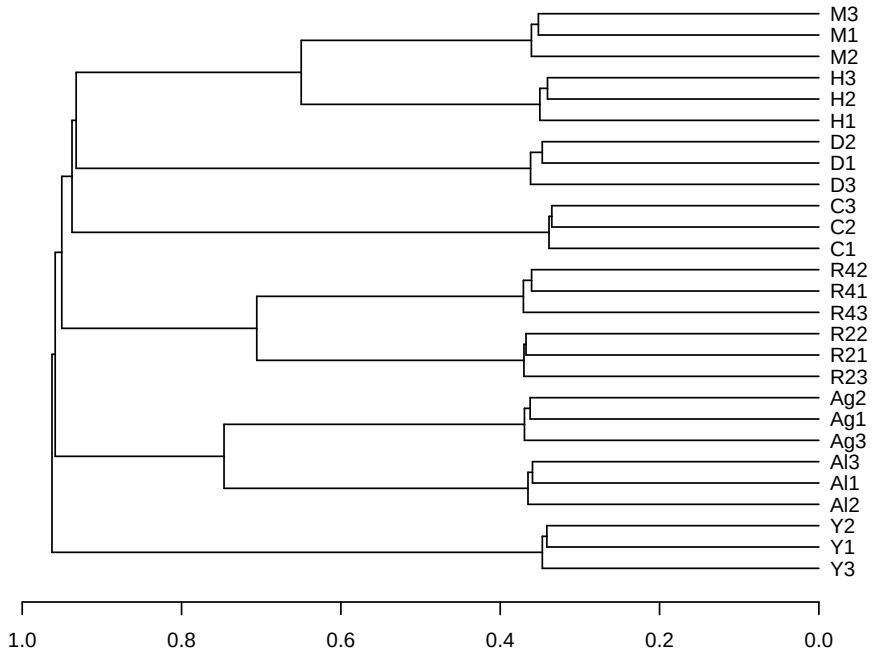

Abbildung 5.18: Dendrogramm basierend auf einer hierarchischen Clusteranalyse mit Average-Linkage für die DISMS2-Distanzen aller 27 Läufe des DISMS2-Datensatzes (DISMS2.f.PRIDE) mit optimierten Parametereinstellungen und einer vorangestellten Repräsentantenauswahl in den einzelnen Läufen mittels des PRIDE-Clusterverfahren.

Läufe, wobei auch Replikate enthalten sind, wäre die Entfernung von Einzelclustern hilfreich, da qualitativ gute Spektren in mehrelementigen Clustern lägen.

Zusammenfassend ist festzustellen, dass die in diesem Abschnitt untersuchte Verknüpfung der Clusterbildung von Massenspektren mit dem DISMS2-Algorithmus keine Verbesserung des ursprünglichen Algorithmus darstellt. Die Wahl des Repräsentanten eines Clusters hat einen geringen Einfluss. Die Verwendung einer Clusteranalyse (hclust oder PRIDE) im Voraus zur Anwendung des DISMS2-Algorithmus hat keinen Vorteil. Die Distanz innerhalb der Gruppen vergrößert sich sogar.

5.5 Clusterung zur Peptidzuordnung fehlender Annotationen

In Abschnitt 5.2 wurden beispielhaft einzelne Cluster betrachtet. Ein Ziel der Clusteranalyse von Tandem-Massenspektren ist es, Informationen bezüglich unannotierter Spektren zu gewinnen. Liegen diese in einem Cluster mit annotierten Spektren, so liegt die Idee nahe, eine Peptidzuordnung fehlender Annotationen vorzunehmen.

Berendes (2017) beschäftigt sich in ihrer Bachelorarbeit mit der Fragestellung, inwiefern die Anzahl an Annotationen erhöht werden kann, wenn Clusterlösungen mit Informationen einer Datenbanksuche verbunden werden. Diese Analyse basiert auf Clusterlösungen von H1 für die in Rieder et al. (2017b) optimierten Parametereinstellungen bezüglich Kosinus-Distanzen ohne DISMS2-Filter und den vorliegenden Datenbankannotationen, die teilweise durch mögliche Peptidannotationen mit niedrigem Score der unannotierten Spektren ergänzt wurden. Bei der Analyse von Clustern, die mindestens fünf Spektren und mindestens ein unannotiertes Spektrum enthalten, konnten zusätzlich nur wenige Spektren annotiert werden. Die Auswertung ist in Cluster mit nur einer Annotation und in Cluster mit mindestens zwei Annotationen unterteilt. Für Cluster ohne Annotation kann keine Peptidzuordnung erfolgen.

Unannotierte Spektren in Clustern mit nur einer Peptidannotation werden jenem Peptid zugeordnet, falls es Bestätigung durch einen Treffer in der Datenbank mit einem eventuell niedrigen Score gibt und die Zugehörigkeit zum Cluster gut ist. Die Lage im Cluster wird dabei über einen Vergleich der mittleren Distanz eines unannotierten zu allen annotierten Spektren bestimmt. In Clustern mit mehreren Annotationen ist die Zuordnung schwieriger, da es mehr Möglichkeiten gibt. Es handelt sich um Gruppen von Spektren, deren Annotation variiert. Die Clusterlösung spiegelt in diesen Fällen also nicht die Gruppierung wider, die sich aus der Annotation ergibt. Es wurden womöglich zu große Cluster gebildet, in denen eine feinere Struktur mittels der Annotation zu erkennen ist.

Im Folgenden wird eine Mindestanzahl von zwei annotierten Spektren bei der Auswahl an Clustern gefordert. Zusätzlich sollen die Cluster mindestens ein Spektrum mit fehlender Annotation beinhalten. Jenen unannotierten Spektren können über ihre Clusterzugehörigkeit Peptide zugeordnet werden. Außerdem werden die gefilterten Distanzen verwendet, da die daraus resultierenden Clusterlösungen im Vergleich zur Peptidannotation zu einem höheren adjustierten Rand-Index führen (siehe Abschnitt 5.1). Exemplarisch wird Lauf C2 ausgewählt, da der Anteil annotierter Spektren am höchsten ist. Es werden stellvertretend nur zwei Clusteralgorithmen mit optimierten Parametereinstellungen betrachtet. Einerseits wird PRIDE Cluster ($threshold\_end = 0.8$) als eine für den Anwender relevante Methode ausgewählt, die bei einer konkurrenzfähigen Qualität der Clusterlösungen (ARI = 0.529) durch eine kurze Rechenzeit und wenig Speicherverbrauch überzeugt. Andererseits wird der Algorithmus mit dem im Vergleich zur Peptidannotation im Mittel höchs-

ten ARI ausgewählt. Dies ist der Algorithmus DBSCAN ($\epsilon = 0.2, minPts = 2$), der igraph ($cdis = 0.2$) entspricht, mit einem Wert von 0.648 (siehe Tabelle 5.2).

Für DBSCAN kommen 103 Cluster infrage, in denen insgesamt 571 Spektren liegen. Davon sind 167 Spektren nicht annotiert. 254 Cluster mit insgesamt 1389 Spektren und 254 nicht annotierten Spektren sind für PRIDE relevant. Die Clustergröße beträgt maximal 46 (DBSCAN) und 78 (PRIDE).

In Abbildung 5.19 ist die Anzahl richtiger Annotationen gegen die Anzahl falscher Annotationen in einem Scatterplot mit transparenten Punkten dargestellt. Eine richtige Annotation bezeichnet die häufigste Annotation in dem jeweiligen Cluster. Die falschen Annotationen beinhalten auch fehlende Annotationen. Da in jedem der ausgewählten Cluster mindestens ein Spektrum mit fehlender Annotation enthalten ist, kommt in jedem Cluster mindestens eine falsche Annotation vor. Bei DBSCAN gibt es viele solcher Cluster mit maximal zehn richtigen Annotationen. Heterogene Cluster, die also mehrere falsche Annotationen enthalten, sind seltener. Bei PRIDE gibt es insgesamt mehr falsche Annotationen. In diesen Clustern bildet eine Annotation also keine deutliche Mehrheit. In dem Fall, dass mehr falsche als richtige Annotationen in einem Cluster liegen, liegen die Punkte unterhalb der Winkelhalbierenden. In nur 19 (DBSCAN) oder 35 (PRIDE) aller Cluster liegt dieses Verhältnis vor.

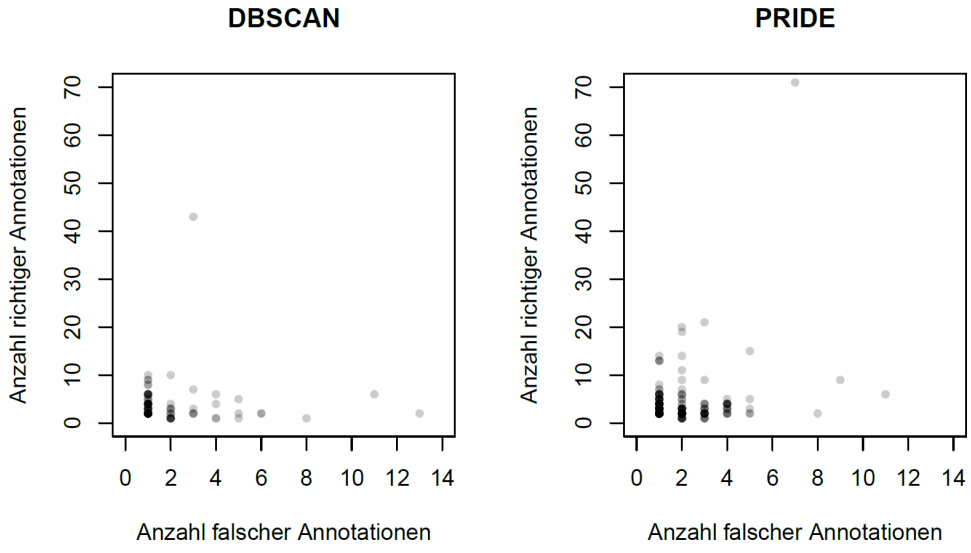

Abbildung 5.19: Anzahl richtiger Annotationen in Abhängigkeit der Anzahl falscher Annotationen in Clustern von C2, die mittels DBSCAN ($\epsilon = 0.2, minPts = 2$) (links) oder PRIDE Cluster ($threshold\_end = 0.8$) (rechts) generiert wurden.

Zunächst werden Cluster mit nur einer Annotation untersucht. Bei DBSCAN kommt in 83 von 103 Clustern nur eine Annotation vor. Bei PRIDE sind es sogar 225 von 254 der Cluster. Die Lage eines nicht annotierten Spektrums wird über die durchschnittliche Distanz (a_j) zwischen allen annotierten Spektren zu einem nicht annotierten Spektrum bestimmt, die für die Berechnung von Silhouetten (siehe Abschnitt 3.3.2) verwendet wird. Die Lage der annotierten Spektren wird analog bestimmt. Falls es mindestens ein annotiertes Spektrum gibt, das im Vergleich zum nicht annotierten Spektrum weiter am Rand des Clusters liegt, wird dem nicht annotierten Spektrum nachträglich die Peptidannotation, die in dem Cluster vorkommt, zugeordnet. Insgesamt können bei DBSCAN 51 von 120 Spektren und bei PRIDE 149 von 377 Spektren nachträglich einem Peptid zugeordnet werden.

In Abschnitt 5.2 wurden zwei Cluster beschrieben, die nur eine Annotation enthalten. In dem in Abbildung 5.13 dargestellten Cluster werden neun der elf nicht annotierten Spektren dem Peptid VIAHTQmK zugeordnet. Nur die in der MDS-Grafik oben links und weit rechts liegenden Spektren bleiben unannotiert, da die durchschnittlichen Distanzen im Vergleich zu groß sind. Auch das unannotierte Spektrum, das in dem in Abbildung B.31 dargestellten Cluster liegt, wird nachträglich annotiert. Die durchschnittliche Distanz des nicht annotierten Spektrums zu allen anderen Spektren beträgt 0.05. Da es mindestens ein Spektrum gibt, dessen durchschnittliche Distanz zu allen anderen Spektren größer als 0.05 ist, wird die Aminosäuresequenz HQGVmVGmGQK empfohlen.

In einem zweiten Schritt werden Cluster mit mehr als einer Annotation analysiert. Jedes nicht annotierte Spektrum wird einmal einer der Annotationsgruppen zugeordnet und die Silhouette wird berechnet. Liegt die größte Silhouette über einem Schwellenwert von 0.5, so wird das nicht annotierte Spektrum nachträglich dem Peptid mit dem höchsten Silhouettenwert zugeordnet. Es kommen nur 29 Cluster (PRIDE) und 20 Cluster (DBSCAN) mit mehr als einer Annotation infrage. Nachträglich können bei DBSCAN insgesamt 25 von 47 Spektren und bei PRIDE 20 von 44 Spektren annotiert werden.

Vier Cluster sind in Abschnitt 5.2 dargestellt, die zwei Annotationen enthalten. In dem in Abbildung 5.14 dargestellten Cluster wird ein Spektrum ohne Annotation, das in der MDS-Grafik weit rechts liegt, nicht nachträglich annotiert. Das oben links liegende unannotierte Spektrum wird mit einem Silhouettenwert von 0.745 dem Peptid NSQEDDDLTIGASPDAGLAFHFVQPSDANVVR zugeordnet. Auch in dem in Abbildung B.32 dargestellten Cluster wird nur ein Spektrum nachträglich annotiert. Das in der MDS-Grafik oben links liegende Spektrum wird mit einem Silhouetten-

wert von 0.781 der Annotation IIQLK zugeordnet. Anhand der in der MDS-Grafik dargestellten Distanzen ist die Zuordnung nicht ersichtlich, da der dargestellte Abstand zu dem Spektrum mit der Annotation IQIIK kleiner ist. Auch die nachträgliche Annotation eines Spektrums in dem in Abbildung B.33 dargestellten Cluster ist anhand der in der MDS-Grafik dargestellten Abstände nicht direkt ersichtlich. Das mittig liegende unannotierte Spektrum (1. Koordinate 0.02, 2. Koordinate 0.00) wird mit einem knapp über dem Schwellenwert von 0.5 liegenden Silhouettenwert 0.549 dem Peptid IEIIK zugeordnet. Die anderen Spektren, die weiter außerhalb liegen, können keinem der beiden Peptide IEIIK und IIEIK zugeordnet werden. Das in Clus_1 (siehe Abbildung 5.11) nicht annotierte Spektrum kann der Annotation HQGVmVGMGQK mit einem Silhouettenwert von 0.486, der knapp unter dem Schwellenwert 0.5 liegt, nicht zugeordnet werden.

Zusammenfassend ist in diesem Abschnitt festzustellen, dass Cluster untersucht wurden, in denen mindestens zwei Spektren annotiert sind und ein Spektrum nicht annotiert ist. Die unannotierten Spektren wurden mittels zwei Clusterlösungen des Laufs C2, DBSCAN ($\epsilon = 0.2, minPts = 2$) mit DISMS2-Filter und PRIDE Cluster ($threshold\_end = 0.8$), nachträglich einem Peptid des Clusters zugeordnet. In Clustern mit nur einer Annotation wurden 51 von 120 Spektren (DBSCAN) und 149 von 377 Spektren (PRIDE) nachträglich annotiert. Falls das Spektrum nicht am Rand des Clusters lag, wurde es annotiert. Dabei wurde die Lage mittels der durchschnittlichen Distanz bestimmt. In Clustern mit mehreren Annotationen wurden 25 von 47 Spektren (DBSCAN) und 20 von 44 Spektren (PRIDE) nachträglich annotiert. Zur Zuordnung zu einer Annotationsgruppe wurde der Maximalwert der Silhouetten bestimmt und zusätzlich das Überschreiten des Schwellenwerts 0.5 gefordert. In Anbetracht der hohen Anzahl an nicht annotierten Spektren, konnte die nachträgliche Zuordnung nur in wenigen Fällen angewendet werden. Anhand der angeführten Beispielcluster ist jedoch ersichtlich, dass das Vorgehen in Einzelfällen nützlich ist.

5.6 Anmerkungen zu Laufzeiten und Speicherverbrauch

Zur Berechnung der Clusterlösungen wurde das LiDOng Cluster der TU Dortmund verwendet. Je nach Datengröße wurden bis zu 64 GB RAM auf Knoten mit Intel Xeon E7340 (2.4 GHz) CPUs angefordert. Zur Clusterung der Tandem-Massenspektren wurden zunächst Vektoren der Distanzmatrizen, die als Eingabe der Algorithmen

dienen, erzeugt und in einem R-Datenformat (.rds) abgespeichert. Für n Spektren werden $n \cdot (n-1)/2$ Distanzen berechnet. Bei den Einzelläufen liegt die Größe der Dateien zwischen 2.89 GB (R21) und 5.40 GB (C2). Für die annotierten Triplikate steigt die Größe auf 8.38 GB (R41, R42, R43) bis 26.33 GB (C1, C2, C3).

Die Berechnung der Clusterlösung von Einzelläufen der humanen Proben benötigt bis zu 13.5 GB RAM und etwa 13 Minuten (Tabelle A.17). Die Laufzeiten und der Speicherverbrauch variieren je nach Clusteralgorithmus. Am besten schneiden MS-Cluster und PRIDE Cluster ab. Sie unterbieten die anderen Algorithmen mit Abstand. Es müssen keine Distanzen berechnet werden, sondern die beiden Algorithmen greifen direkt auf die Originaldaten im Format MGF zu. In wenigen Minuten werden Clusterlösungen des PRIDE-Algorithmus bei einem maximalen Speicherverbrauch von 0.14 GB generiert.

Zur Clusterung mehrerer Läufe steigen Speicherverbrauch und Laufzeit an. Maximal wurden 60 GB RAM bei einer Laufzeit von ungefähr 67 Minuten verbraucht (Tabellen A.18 und A.19). Bei der hierarchischen Clusteranalyse wurde das Speicherlimit von 64 GB in einigen Fällen überschritten. Für die Läufe C1, C2, C3, H1, H2, H3, M1, M2 und M3 gibt es also keine Clusterlösungen. Die Implementierung der hierarchischen Clusteranalyse über die R-Funktion `hclust()` benötigt als Eingabe ein sogenanntes `dist`-Objekt. Dieses beinhaltet neben einem Vektor, der die Einträge der Distanzmatrix beinhaltet, weitere Attribute. Mithilfe der Funktion `structure()` kann ein `dist`-Objekt erstellt werden, indem den vorliegenden Kosinusdistanzen Attribute hinzugefügt werden. Jedoch reicht der angeforderte Speicherverbrauch für die großen Distanzvektoren von Fadenwurm (26.33 GB), Mensch (19.83 GB) und Maus (17.95 GB) nicht aus.

In Abbildung 5.20 sind Scatterplots dargestellt, die die Laufzeiten in Abhängigkeit des Speicherverbrauchs darstellen. Je Clusteralgorithmus ist ein linearer Zusammenhang zwischen Speicherverbrauch und Laufzeit zu erkennen. Bei CAST ist die Steigerung der Laufzeit in Abhängigkeit des Speicherverbrauchs am größten. Bei igraph und DBSCAN mit dem Parameter $minPts = 2$ handelt es sich um zwei Implementierungen unterschiedlicher Algorithmen, die zur gleichen Clusterlösung führen. Es wird eine hierarchische Clusterlösung mit Single-Link berechnet, wenn das Dendrogramm auf der Höhe von $cdis$ oder ϵ abgeschnitten wird. Im Vergleich der Laufzeiten fällt auf, dass die Clusterlösung mittels igraph schneller generiert wird. Allerdings benötigt DBSCAN ($minPts = 2$) weniger Speicher. Daher wird die Implementierung von DBSCAN ($minPts = 2$) bevorzugt. Besonders bei den gefiltertern Distanzen der annotierten Replikate des Fadenwurms (C1, C2, C3) ist

dies von Bedeutung. Es werden 58.66 GB RAM von igraph und 38.13 GB RAM von DBSCAN ($minPts = 2$) benötigt. Der Speicherbedarf ist also ungefähr um die Hälfte erhöht. Die Laufzeit sinkt jedoch von etwa 37 auf ungefähr 21 Minuten.

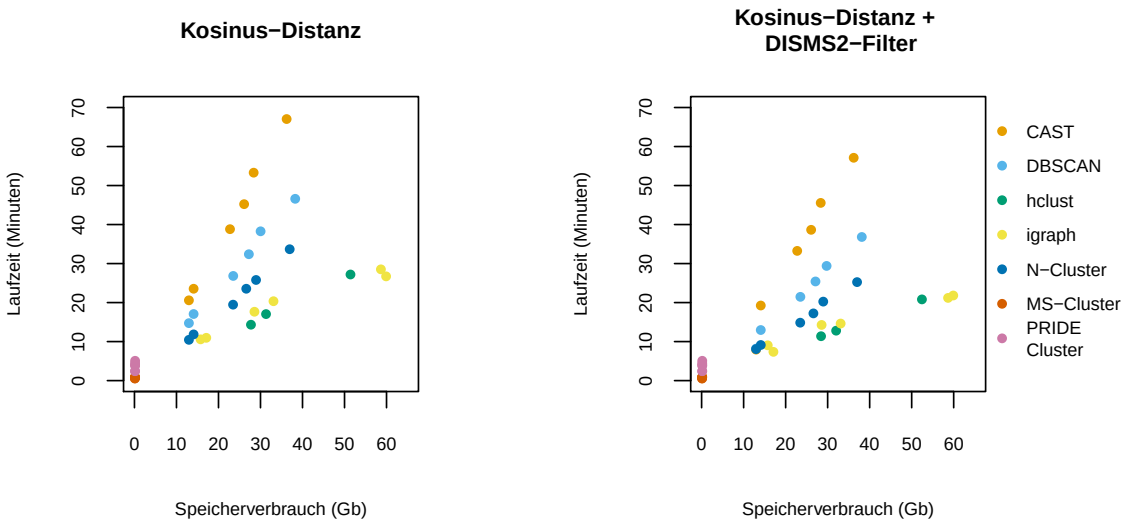

Abbildung 5.20: Laufzeit (Minuten) und Speicherverbrauch (Gb) für die Clusterung von je drei Läufen. Pro Algorithmus sind sechs Punkte für die Arten C, D, H, M, Y und R4 dargestellt.

Zusammenfassend ist in diesem Abschnitt festzustellen, dass die Laufzeit und der Speicherverbrauch von Lösungen unterschiedlicher Implementierungen von Clusterverfahren aus Abschnitt 5.1 und 5.3 untersucht wurden. Für die Algorithmen MS-Cluster und PRIDE Cluster liegen von der Kommandozeile aus ausführbare Programme vor. Die Laufzeiten des MS-Cluster v2 algorithm und der eigenständigen Java-Applikation spectra-cluster-cli liegen bei unter 10 Minuten bei einem Speicherbedarf von maximal 0.15 GB. Für die übrigen Algorithmen wurden Implementierungen in R verwendet, deren Eingabe ein Distanzmatrixvektor ist. Für die Clusterung mehrerer Läufe wurden maximal 60 GB RAM und etwa 67 Minuten benötigt. Da die in R implementierten Algorithmen zum Teil bessere Ergebnisse liefern als MS-Cluster und PRIDE Cluster, sollten effizientere Implementierungen gewählt werden, um bessere Alternativen zu MS-Cluster und PRIDE Cluster anzubieten. Außerdem sind Implementierungen der Algorithmen nötig, die weniger Speicherverbrauch und eine kürzere Laufzeit bieten, um in Zukunft auch größere Datensätze analysieren zu können.

6 Zusammenfassung und Diskussion

Die Arbeit handelt von Clustermethoden für massenspektrometrische Analysen in der Biodiversitätsforschung. In den Jahren 2014-2017 förderte die Leibniz-Gemeinschaft diese Dissertationsarbeit im Rahmen des Projekts „(Reverse) Proteomics as novel tool for biodiversity research“ (SAW-2014-ISAS-2-D). Alternativ zur Artenbestimmung mittels DNA-Barcoding, wird die Analyse der Proteinzusammensetzung von Organismen verwendet. Der Schwerpunkt liegt auf Arten unbekannten Genoms, sogenannten Foraminiferen und Süßwasserschnecken der Gattung *Radix*. Die Mehrheit der Proteinanalytik basiert mittlerweile auf der sogenannten LC-MS/MS-Methode. Dabei wird eine Flüssigchromatographie (LC) als Trennmethode mit der Tandem-Massenspektrometrie (MS/MS) kombiniert. Tandem-Massenspektren, die aus detektierten Intensitäten von vorkommenden Massen bestehen, dienen zur Identifikation von Peptiden und Proteinen mittels Datenbanksuchalgorithmen. Neuartige unbekannte Peptide werden mittlerweile über De-Novo-Peptidsequenzierungsalgorithmen detektiert, doch diese Verfahren sind sehr fehleranfällig. Alternativ zu Annotationsverfahren wird die direkte Clusteranalyse der Tandem-Massenspektren behandelt. Zwei Aspekte, die Clusteranalyse sogenannter Läufe als Objekte, die tausende Spektren einer Proteinprobe umfasst, und die Clusteranalyse von Tandem-Massenspektren als Objekte wurden untersucht.

Zunächst werden die im Projekt erstellten Datensätze (DISMS2, Foraminiferen, Biodiversität-Exactive, Biodiversität-Orbitrap) und ein öffentlich zugänglicher Datensatz (Palmblad) vorgestellt. Die Verfahren basieren auf vorgestellten Distanzmaßen für Tandem-Massenspektren. Ein neuer flexibler Algorithmus zur Clusteranalyse von Läufen, DISMS2, wird eingeführt. Parameter können frei gewählt werden, sodass die Auswahl der höchsten Peaks je Spektrum (`topn`), die Bingröße im Binning (`bin`), die Einschränkung bei dem Vergleich von Spektren auf zeitlich nahe Spektren (`ret`) mit ähnlicher Precursormasse (`prec`) und das Distanzmaß für

Tandem-Massenspektren (`dist`) mit einem frei wählbaren Schwellenwert (`cdis`) variieren. Zur Parameterwahl wird ein Vorgehen zur Optimierung vorgestellt, das das Bestimmtheitsmaß R^2 eines nichtparametrischen Verfahrens zur Varianzanalyse verwendet. Replikate werden in Gruppen eingeteilt, innerhalb derer kleinere Distanzen zu erwarten sind als zwischen Gruppen. Für den Vergleich von Algorithmen zur Clusteranalyse von einzelnen Massenspektren wird eine Auswahl an Methoden aus der Literatur (Hierarchische Clusteranalyse, CAST, DBSCAN, igraph, MS-Cluster und PRIDE Cluster) ergänzt durch das neue Neighbor Clustering vorgestellt. Vergleichende Bewertungsmaße werden diskutiert und Visualisierungstechniken werden erläutert.

Eine Clusteranalyse sogenannter Läufe wird für alle Datensätze mithilfe der neuen Methode DISMS2 (Rieder et al., 2017a), die ohne Zuhilfenahme von Peptidannotationen Distanzen zwischen MS/MS-Läufen bestimmt, durchgeführt. Jene Distanzen werden mittels einer hierarchischen Clusteranalyse als Dendrogramme dargestellt, sodass phylogenetische Bäume entstehen. Als Erweiterung des compareMS2-Algorithmus (Palmblad und Deelder, 2012) wird eine Alternative zum Vergleich von Peptidlisten, die auf der Identifikation von Spektren in Datenbanksuchen basieren, ermöglicht. Ein erster Überblick über die massenspektrometrischen Daten erfolgt in einer deskriptiven Analyse. Zusätzlich zu den Massenspektren werden auch Metadaten, wie beispielsweise Precursormasse, Precursorladung und Retentionszeit beschrieben. Für den DISMS2-Datensatz liegen für einige Arten Peptidannotationen vor, die aus einer Datenbanksuche stammen. Der Foraminiferen-Datensatz enthält De-Novo-Annotationen, die mithilfe der Methode von Blank-Landeshammer et al. (2017) erstellt wurden. Die Optimierung der Parametereinstellungen im DISMS2-Algorithmus wurde exemplarisch für die Datensätze DISMS2 ($R^2 = 0.923$) und Palmblad ($R^2 = 0.385$) beschrieben. Im Dendrogramm des DISMS2-Datensatz, der je drei technische Replikate der Proben von Mensch, Maus, Hefe, Fadenwurm, Fruchtfliege, Foraminiferen und Süßwasserschnecken beinhaltet, ist eine deutliche Abgrenzung der Gruppen zu erkennen. Im Palmblad-Datensatz, der Blutseren von Menschenaffen und andere Primaten umfasst, sind die Abstände viel größer. Eine Unterscheidung der beiden Makakenarten, Javaneraffe und Rhesusaffe, ist nicht möglich.

Die vorliegenden Peptidannotationen im DISMS2-Datensatz wurden zur Validierung der DISMS2-Distanzen verwendet. In einem direkten Vergleich ergeben sich Unterschiede zwischen Annotationsabständen und DISMS2-Abständen, jedoch sind die Algorithmenschritte nicht vergleichbar. In einem fairen Vergleich, in dem die Schrit-

te bis auf die Suchmethode *Datenbank* (Peptidannotation) oder *Distanz* (DISMS2) gleich sind, ergaben sich sehr ähnliche Abstände zwischen Läufen. Aus der Kosinus-Distanz der Tandem-Massenspektren lässt sich ein binärer Klassifikator erstellen, der zwischen gleichen und verschiedenen Peptiden aus der Datenbank unterscheidet. Für den Schwellenwert `cdis`= 0.3 liegen beim Vergleich von zwei humanen Läufen die Sensitivität bei 0.923 und die Spezifität bei 0.867. Die Konkurrenzfähigkeit von DISMS2- mit Annotationsabständen wurde in der Analyse des Foraminiferen-Datensatzes bestätigt. Die Unterscheidung nah verwandter Foraminiferenarten, für die je drei biologische Replikate vorliegen, ist jedoch schwieriger und nicht immer eindeutig.

Die Generierung und Vorverarbeitung von Tandem-Massenspektren beeinflusst die DISMS2-Abstände. Vor Anwendung des Algorithmus wurde im Foraminiferen-Datensatz die von Palmblad und Deelder (2012) erwähnte Auswahl an 2000 Spektren mit hoher Gesamtionenintensität vorgenommen. Das resultierende Dendrogramm konnte durch eine De-Novo-Methode bestätigt werden. Für Peptidlisten, die mittels des Verfahrens von Blank-Landeshammer et al. (2017) erstellt wurden, wurde der Sørensen-Dice-Index berechnet. Eine deutliche Trennung der Foraminiferenarten *A. gibbosa* und *M. vertebralis* wurde beobachtet. Die Arten *A. lessonii* und *A. lobifera* zeigten mehr Ähnlichkeit. Bei den gemessenen Proben handelt es sich um Holobionten, also Foraminiferen-Wirte und endosymbiotische Mikroalgen. Eine Bestimmung der Symbiontenarten sollte vorgenommen werden, um eine Unterscheidung zwischen Herkunft und Fangjahr der Proben zu verbessern. Eine Sequenzierung der Algen ist vom Leibniz-Zentrum für Marine Tropenforschung bereits geplant.

Die Verwendung verschiedener Massenspektrometer hat einen erheblichen Einfluss auf die DISMS2-Abstände. Die Biodiversität-Datensätze enthalten Proben der Gattung *Radix* und zur Referenz Proben von Meeresfrüchten, einer Tellerschnecke und einem Regenwurm. Technische Replikate wurden mit Massenspektrometern vom Typ Q Exactive und vom Typ Orbitrap Elite analysiert. In den Dendrogrammen wurden Unterschiede bei der Höhe der Distanzen und bei den Verzweigungen festgestellt. Eine deutliche Trennung wurde zwischen den *Radix* und anderen Proben beobachtet, jedoch nicht zwischen einzelnen Arten der Gattung *Radix*, der Herkunft oder verschiedener Teile der Schnecken.

Laufzeit und Speicherverbrauch der Implementierung des Algorithmus in R wurden bei der Optimierung im DISMS2-Datensatz untersucht. Der Speicherverbrauch war gering (unter 3 GB), da Distanzen nur bei Bedarf nach Prüfung mehrerer Bedingungen berechnet werden. Die Laufzeit betrug für die optimalen Einstellungen fast

16 Stunden. Durch Parallelisierung sollte sie reduziert werden, da einzelne Einträge der Distanzmatrix unabhängig voneinander und die Treffer für Einträge der ersten Liste parallel zueinander bestimmt werden.

Besonders eine Analyse von wenig erforschten Arten profitiert von der neuen Methode DISMS2. Ein großes Problem ist die Ungewissheit darüber, welche Peptide tatsächlich gemessen wurden. Eine Übereinstimmung zwischen De-Novo- oder Datenbankannotationen und dem DISMS2-Algorithmus wurde beobachtet. Die Kombination der Suchmethoden *Datenbank* und *Distanz* ist also vielversprechend für zukünftige Analysen. Denn eine De-Novo-Annotation liefert einen Hinweis auf das wahre Peptid, der mittels der Auswertung der Spektrendistanzen bestätigt werden kann. Bei der Auswertung der unterschiedlichen Datensätze wurde deutlich, dass die Qualität der durchgeführten Experimente inklusive Probenvorbereitung und technischer Messung im Massenspektrometer schwankt. Es wird empfohlen technische und biologische Replikate zu berücksichtigen, um die Varianz innerhalb von Arten abschätzen zu können. Leider war in einigen Fällen die Varianz der Replikate so hoch, dass die Unterscheidung verschiedener Arten nicht möglich war. Der DISMS2-Algorithmus bietet viele Möglichkeiten zur Verbesserung. Die Vorauswahl an Spektren und die zugrundeliegende Distanzberechnung von Spektrenpaaren können beliebig erweitert werden.

Zur Clusteranalyse von einzelnen Tandem-Massenspektren wurde ein bisher in der Literatur fehlender umfassender Vergleich von sieben Algorithmen erstellt, die für Tandem-Massenspektren etabliert (CAST, MS-Cluster, PRIDE Cluster), für große Datensätze bekannt (hierarchische Clusteranalyse, DBSCAN, Zusammenhangskomponenten eines Graphen) oder neu (Neighbor Clustering) sind. Die Evaluierung basiert auf dem DISMS2-Datensatz und mehreren Gütemaßen. Die Qualität der Cluster wird bewertet anhand des adjustierten Rand-Indexes (ARI), der Reinheit der Cluster, des Anteils von Spektren in mehrelementigen Clustern, des Spektrenanteils ohne häufigste Annotation im jeweiligen Cluster, des verbleibenden Anteils an Annotationen nach der Clusterbildung und der Clusteranzahl in Relation zur Spektrenanzahl. In Erweiterung zu dem in Rieder et al. (2017b) publizierten Vergleich wurde die Clusterbildung von Massenspektren mit dem DISMS2-Algorithmus verknüpft und die Clusterbildung von Spektren zur Peptidzuordnung von Spektren mit fehlender Annotation verwendet. Ein wesentlicher algorithmischer Unterschied ist zudem, dass für das PRIDE Clusterverfahren die neu erschienene Version 1.0.3 statt Version 1.0.1 der Java-Applikation spectra-cluster-cli verwendet wurde. Die Änderung der Default-Einstellung des Parameters `x_min_comparisons` von 0 (Ver-

sion 1.0.1) auf 10 000 (Version 1.0.3) führte zu einer deutlichen Verbesserung der Clusterergebnisse.

Zuerst wurden Peptidannotationen mit Clusterlösungen von Tandem-Massenspektren des DISMS2-Datensatzes der sieben Algorithmen und unterschiedlicher Parametereinstellungen verglichen. Da Peptidannotationen ein Indikator für die wahre Clusterlösung sind, wurden die Clusterlösungen in Bezug zu den Datenbankannotationen gesetzt. Es konnte gezeigt werden, dass etablierte Methoden und der neu vorgestellte Neighbor-Clustering-Algorithmus (N-Cluster) mindestens genauso gut sind wie die aus der Proteomik stammenden Verfahren MS-Cluster und PRIDE Cluster. Die Distanzberechnung ist Teil des Algorithmus von MS-Cluster und PRIDE Cluster. Für die anderen Verfahren wurden die Distanzen auf gleiche Weise vorab berechnet und es wird eine deutliche Verbesserung der Clusterlösungen durch den DISMS2-Filter erreicht. Die Berücksichtigung von Precursorladung, Precursormasse und Retentionszeit führt zu ähnlichen Ergebnissen wie die Clusterlösung der Datenbankannotationen. Zunächst wurden die Parametereinstellungen bezüglich des höchsten mittleren ARI im Vergleich zur Peptidannotation optimiert. Keines der Verfahren ist im Mittel optimal bezüglich aller betrachteten Gütemaße. Der mehrelementige Clusteranteil ist bei PRIDE am größten und auch die Clusteranzahl in Relation zur Spektrenanzahl am kleinsten. DBSCAN ist maximal bezüglich der ARI-Werte. Den geringsten Spektrenanteil ohne häufigste Annotation und den größten Anteil verbleibender Annotationen liefert MS-Cluster. Insgesamt gibt es nur geringe Unterschiede in der Streuung der Werte (Standardfehler unter 0.02). Die Reproduzierbarkeit der Analyse ist gewährleistet, da die Implementierung der Clusterverfahren frei verfügbar ist. Eine Erweiterung auf andere Clusteralgorithmen, andere Distanzberechnungen von Tandem-Massenspektren und die Erweiterung der Evaluierungsmethoden ist außerdem möglich. Aufgrund der Limitierung der Laufzeit und des maximalen Speicherverbrauchs wurden einzelne MS/MS-Läufe analysiert und eine Vorauswahl für die Parametereinstellungen der Algorithmen getroffen.

Bei der Interpretation einzelner Cluster wurde festgestellt, dass in einem Cluster ein Spektrum wegen eines Fehlers der Software zur Datenbankannotation nicht annotiert wurde. In einem anderen Cluster wurden Spektren mit sehr unterschiedlichen Annotationen und Precursormassen zusammengefasst. Ursache war ein hoher Peak eines Immoniumions in den Spektren, der fälschlicherweise zu kleinen Kosinusdistanzen geführt hat. Der DISMS2-Filter verhindert diesen Fehler, denn Spektren mit unterschiedlichen Precursormassen werden nicht in einem Cluster zusammengefasst.

Die Clusteranalyse wurde anschließend auf mehrere Läufe von drei technischen Replikaten des DISMS2-Datensatzes erweitert. Das Proteom einer Art ist also besser repräsentiert durch die Auswertung von Mehrfachmessungen. Das hierarchische Clusterverfahren lieferte für Fadenwurm, Maus und Mensch keine Lösung, da das Speicherlimit überschritten wurde. In folgenden Analysen ist die hierarchische Clusterung in der aktuellen Implementierung also nicht zu empfehlen. Im Vergleich zur Evaluierung einzelner Läufe ist eine drastische Verbesserung der Werte der Bewertungsmaße zu beobachten. Eine weitere Erweiterung der Analyse ist wünschenswert, um mehrere Millionen Spektren unterschiedlicher Arten gemeinsam zu clustern. Dadurch können Peptide über Artengrenzen hinweg identifiziert werden. Die Umsetzbarkeit scheitert zurzeit bis auf zwei Ausnahmen (MS-Cluster und PRIDE Cluster) an der Begrenzung des Speicherplatzes und der Rechenzeit.

Schließlich wurde die Clusterbildung von Massenspektren mit dem DISMS2-Algorithmus verknüpft. Die Verwendung einer Clusteranalyse (hierarchische Clusteranalyse oder PRIDE) der Spektren vor der Anwendung des DISMS2-Algorithmus führte dazu, dass sich die Distanz innerhalb der Gruppen vergrößert hat. Dies ist durch den großen Anteil an Einzelclustern zu erklären. Ein Ausweg ist die Evaluation auf größeren Datensätzen mit Replikaten, bei deren Clusterbildung die Anzahl an Einzelclustern sinkt. Eine Clusterbildung in Kombination mit einer anschließenden Entfernung der Einzelcluster ist erfolgversprechend. Die in mehreren Proben vorkommenden Peptide würden somit in Cluster zusammengefasst und Spektren von schlechter Qualität entfernt. Letztendlich würden qualitativ gute Spektren mit gleichem Gewicht in die Berechnung der DISMS2-Distanz eingehen.

Eine Anwendung der Clusterbildung von Spektren ist die Peptidzuordnung von Spektren mit fehlender Annotation. Es wurden Cluster untersucht, in denen mindestens zwei Spektren annotiert sind und ein Spektrum nicht annotiert ist. Die unannotierten Spektren wurden mittels zweier Clusterlösungen eines Laufs des Fadenwurms, DBSCAN ($\epsilon = 0.2$, $minPts = 2$) mit DISMS2-Filter und PRIDE Cluster ($threshold\_end = 0.8$), nachträglich einem Peptid des Clusters zugeordnet. In Clustern mit nur einer Annotation wurden 51 von 120 Spektren (DBSCAN) und 149 von 377 Spektren (PRIDE) nachträglich annotiert. Falls das Spektrum nicht am Rand des Clusters lag, wurde es annotiert. Dabei wurde die Lage mittels der durchschnittlichen Distanz bestimmt. In Clustern mit mehreren Annotationen wurden 25 von 47 Spektren (DBSCAN) und 20 von 44 Spektren (PRIDE) nachträglich annotiert. Zur Zuordnung zu einer Annotationsgruppe wurde der Maximalwert der Silhouetten bestimmt und zusätzlich das Überschreiten des Schwellenwerts 0.5 gefordert. In

Anbetracht der hohen Anzahl an nicht annotierten Spektren, konnte die nachträgliche Zuordnung nur in wenigen Fällen angewendet werden. Anhand der angeführten Beispielcluster ist jedoch ersichtlich, dass das Vorgehen in Einzelfällen nützlich ist. In Clusteranalysen mehrerer Läufe sinkt die Anzahl von Einzelclustern, sodass die Anzahl der in Frage kommenden Cluster für eine nachträgliche Peptidannotation womöglich steigt.

Im Vergleich der Laufzeit und des Speicherverbrauchs von Lösungen unterschiedlicher Implementierungen von Clusterverfahren sind MS-Cluster und PRIDE Cluster am besten. Für sie gibt es von der Kommandozeile aus ausführbare Programme, MS-Cluster v2 algorithm und die eigenständige Java-Applikation spectra-cluster-cli. Die Clusterlösungen dieser Arbeit wurden in unter 10 Minuten bei einem Speicherbedarf von maximal 0.15 GB generiert. Für die übrigen Algorithmen wurden Implementierungen in R verwendet, deren Eingabe ein Distanzmatrixvektor ist. Für die Clusterung mehrerer Läufe wurden maximal 60 GB RAM und etwa 67 Minuten benötigt. Da die in R implementierten Algorithmen zum Teil bessere Ergebnisse liefern als MS-Cluster und PRIDE Cluster, sollten effizientere Implementierungen gewählt werden, um bessere Alternativen zu MS-Cluster und PRIDE Cluster anzubieten. Außerdem sind Implementierungen der Algorithmen nötig, die weniger Speicherverbrauch und eine kürzere Laufzeit bieten, um in Zukunft auch größere Datensätze analysieren zu können.

Entgegen der Schlussfolgerung in Rieder et al. (2017b) ist die Anwendung des Algorithmus PRIDE Cluster mit den Default-Einstellungen der neuen Version 1.0.3 zu empfehlen. Die Qualität der Clusterlösungen des DISMS2-Datensatzes war bezüglich der betrachteten Gütemaße sehr gut oder zumindest konkurrenzfähig mit anderen Algorithmen. Besonders bei der gleichzeitigen Analyse mehrerer Läufe ist PRIDE Cluster besser. Vorteile sind auch die kurze Laufzeit und der geringe Speicherbedarf der Implementierung. Die Verwendung der anderen bekannten Algorithmen ist bei Verwendung des DISMS2-Filters besonders gut. Daher sollte untersucht werden, ob PRIDE zusätzlich von der Verwendung der Zusatzinformationen, die für die Spektren vorliegen, profitiert.

Ein Ansatz zur Verbesserung der Clusterlösung ist das sogenannte Consensus Clustering (Abu-Jamous et al., 2015, S. 295f.), das im biometrischen Bereich bei der Genexpressionsanalyse zur Clusterung von Microarray-Daten Anwendung findet. Wird wie in dem durchgeführten Vergleich eine Vielzahl an unterschiedlichen Clusterlösungen generiert, ist ein Ziel die Bildung einer einzigen Lösung, die einen Konsens darstellt. Neben dem Einsatz von Resampling-Verfahren ist der von Swift

et al. (2004) vorgestellte Consensus Clustering Algorithmus erfolgversprechend. Basierend auf dem ARI werden zwei Algorithmen vorgestellt zur Bildung von robusten Clustern. Nur Objekte, die bei allen oder vielen der ausgewählten Algorithmen in dem gleichen Cluster liegen, werden dem gleichen Cluster hinzugefügt. Ein Abgleich der betrachteten Algorithmen könnte daher zur Generierung homogener Cluster beitragen.

In dem durchgeführten Vergleich der Clusteralgorithmen wurde die Qualität der Clusterlösungen einzeln anhand verschiedener Gütemaße bewertet. Die Parameter in den Algorithmen, die Schwellenwerte für Distanzen oder Ähnlichkeiten von Spektrenpaaren sind, wurden so gewählt, dass der ARI im Mittel über alle betrachteten Läufe am höchsten ist. Es handelt sich also um eine Optimierung bezüglich eines Kriteriums. Eine multikriterielle Optimierung, auch Pareto-Optimierung genannt, findet eine Clusterlösung, die optimal für mehrere Zielkriterien ist. Die Kriterien, die sich aus den vorgestellten Gütemaßen ergeben, sollten in Zukunft erweitert werden und dann gemeinsam optimiert werden.

In dieser Arbeit wurden relativ kleine Datensätze verwendet. Die Auswertung eines großen Benchmark-Datensatzes sollte folgen. Geeignet ist eine aktuelle Studie von Zolg et al. (2017), die sich mit der Analyse von bereits über 330 000 synthetischen tryptischen Peptiden befasst, die alle kanonischen humanen Genprodukte repräsentieren. In den kommenden Jahren sollen die frei verfügbaren Daten auf über eine Millionen Peptide erweitert werden. Zur Simulation von Tandem-Massenspektren müssen zahlreiche Eigenschaften von Proteinen und deren Verhalten bei der massenspektrometrischen Messung beachtet werden. Vaughan et al. (2009) stellen sogenannte Plasmode-Datensätze als Alternative zu Simulationen vor. Sie werden in natürlichen biologischen Prozessen, jedoch unter kontrollierbaren Laborbedingungen erzeugt.

Literaturverzeichnis

Abu-Jamous, B., Fa, R., und Nandi, A. K. (2015): *Integrative cluster analysis in bioinformatics*. John Wiley & Sons, Chichester, 1. Auflage.

Aggarwal, C. C. (2015): *Data mining: the textbook*. Springer, 1. Auflage.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., und Walter, P. (2008): *Molecular biology of the cell*. Garland Science, New York, 5. Auflage.

Anderson, M. (2001): A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26:32–46.

Ankerst, M., Breunig, M. M., peter Kriegel, H., und Sander, J. (1999): Optics: Ordering points to identify the clustering structure. In *ACM Sigmod record* 49–60. ACM.

BBC News (2016): Great barrier reef: Bleaching 'kills 35% of area's coral'. `http://www.bbc.com/news/world-australia-36410767` (besucht am 28.11.2017).

Beer, I., Barnea, E., Ziv, T., und Admon, A. (2004): Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics*, 4(4):950–960.

Ben-Dor, A., Shamir, R., und Yakhini, Z. (1999): Clustering gene expression patterns. *Journal of computational biology*, 6(3-4):281–297.

Berendes, I.-M. (2017): Statistische Analyse von Clusterergebnissen von Massenspektren. Bachelorarbeit, TU Dortmund.

Bessant, C. (2016): *Proteome Informatics*. New Developments in Mass Spectrometry. Royal Society of Chemistry, 1. Auflage.

Biemann, K. (1992): Mass spectrometry of peptides and proteins. *Annu. Rev. Biochem.*, 61:977–1010.

Bischl, B., Lang, M., Bossek, J., Horn, D., Richter, J., und Surmann, D. (2016): *BBmisc: Miscellaneous Helper Functions for B. Bischl.* R package version 1.10.

Bischl, B., Lang, M., Mersmann, O., Rahnenführer, J., und Weihs, C. (2015): BatchJobs and BatchExperiments: Abstraction mechanisms for using R in batch environments. *Journal of Statistical Software*, 64(11):1–25.

Blank-Landeshammer, B., Kollipara, L., Biß, K., Pfenninger, M., Malchow, S., Shuvaev, K., Zahedi, R. P., und Sickmann, A. (2017): Combining de novo peptide sequencing algorithms, a synergistic approach to boost both identifications and confidence in bottom-up proteomics. *Journal of Proteome Research*, 16(9):3209–3218. PMID: 28741358.

Borg, I., Groenen, P., und Mair, P. (2012): *Applied Multidimensional Scaling.* Springer Briefs in Statistics. Springer, Berlin.

Buxbaum, E. (2015): *Fundamentals of Protein Structure and Function.* Springer, 2. Auflage.

Campello, R. J., Moulavi, D., Zimek, A., und Sander, J. (2015): Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1):5.

Canterbury, J. D., Merrihew, G. E., MacCoss, M. J., Goodlett, D. R., und Shaffer, S. A. (2014): Comparison of data acquisition strategies on quadrupole ion trap instrumentation for shotgun proteomics. *J. Am. Soc. Mass Spectrom.*, 25(12):2048–2059.

Celebi, M. E. und Aydin, K. (2016): *Unsupervised Learning Algorithms.* Springer, Cham.

Chambers, M. C., Maclean, B., Burke, R., und others, . (2012): A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*, 30(10):918–920.

Chi, H., Chen, H., He, K., Wu, L., Yang, B., Sun, R. X., Liu, J., Zeng, W. F., Song, C. Q., He, S. M., und Dong, M. Q. (2013): pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra. *J. Proteome Res.*, 12(2):615–625.

Craig, R. und Beavis, R. C. (2003): A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.*, 17(20):2310–2316.

Csardi, G. und Nepusz, T. (2006): The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Dammeier, S., Nahnsen, S., Veit, J., Wehner, F., Ueffing, M., und Kohlbacher, O. (2016): Mass-Spectrometry-Based Proteomics Reveals Organ-Specific Expression Patterns To Be Used as Forensic Evidence. *J. Proteome Res.*, 15(1):182–192.

Desper, R. und Gascuel, O. (2002): Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.*, 9(5):687–705.

Deutsch, E. W. (2012): File formats commonly used in mass spectrometry proteomics. *Mol. Cell Proteomics*, 11(12):1612–1621.

Deza, M. M. und Deza, E. (2016): *Encyclopedia of Distances.* Springer, Berlin, 4. Auflage.

Dutta, D. und Chen, T. (2007): Speeding up tandem mass spectrometry database search: metric embeddings and fast near neighbor search. *Bioinformatics*, 23(5):612–618.

Eidhammer, I., Flikka, K., Martens, L., und Mikalsen, S. (2007): *Computational Methods for Mass Spectrometry Proteomics.* John Wiley & Sons.

Eng, J. K., Fischer, B., Grossmann, J., und Maccoss, M. J. (2008): A fast SEQUEST cross correlation algorithm. *J. Proteome Res.*, 7(10):4598–4602.

Eng, J. K., McCormack, A. L., und Yates, J. R. (1994): An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, 5(11):976–989.

Ester, M., Kriegel, H.-P., Sander, J., und Xu, X. (1996): A density-based algorithm for discovering clusters in large spatial databases with noise. In Simoudis, E., Han, J., und Fayyad, U. (Hrsg.), *Proceedings of the Second Knowledge Discovery and Data Mining Conference,* 226–231. The AAAI Press.

Ester, M. und Sander, J. (2000): *Knowledge Discovery in Databases: Techniken und Anwendungen.* Springer, Berlin, Heidelberg, 1. Auflage.

Fahrmeir, L., Hamerle, A., und Nagl, W. (1996a): *Varianz- und Kovarianzanalyse* 169–238. In Fahrmeir et al. (1996b), 2. Auflage.

Fahrmeir, L., Hamerle, A., und Tutz, G. (Hrsg.) (1996b): *Multivariate statistische Verfahren*. de Gruyter, Berlin, 2. Auflage.

Fahrmeir, L., Kaufmann, H., und Kredler, C. (1996c): *Regressionsanalyse* 93–168. In Fahrmeir et al. (1996b), 2. Auflage.

Fields, S. (2001): Proteomics. Proteomics in genomeland. *Science*, 291(5507):1221–1224.

Frank, A. und Pevzner, P. (2005): PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.*, 77(4):964–973.

Frank, A. M. (2008): *Algorithms for Tandem Mass Spectrometry-based Proteomics*. Dissertation, University of California at San Diego, La Jolla, CA, USA. AAI3307704.

Frank, A. M., Bandeira, N., Shen, Z., Tanner, S., Briggs, S. P., Smith, R. D., und Pevzner, P. A. (2008): Clustering millions of tandem mass spectra. *J. Proteome Res.*, 7(1):113–122.

Frank, A. M., Monroe, M. E., Shah, A. R., Carver, J. J., Bandeira, N., Moore, R. J., Anderson, G. A., Smith, R. D., und Pevzner, P. A. (2011): Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat. Methods*, 8(7):587–591.

Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S., und MacCoss, M. J. (2006): Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.*, 78(16):5678–5684.

Ghosh, P. (2015): *Introduction to Protein Mass Spectrometry*. Elsevier Science, 1. Auflage.

Gibb, S. (2015): *readMzXmlData: Reads Mass Spectrometry Data in mzXML Format*. R package version 2.8.1.

Gower, J. C. (1966): Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Griss, J., Foster, J. M., Hermjakob, H., und Vizcaino, J. A. (2013): PRIDE Cluster: building a consensus of proteomics data. *Nat. Methods*, 10(2):95–96.

Griss, J., Perez-Riverol, Y., Lewis, S., Tabb, D. L., Dianes, J. A., Del-Toro, N., Rurik, M., Walzer, M. W., Kohlbacher, O., Hermjakob, H., Wang, R., und Vizcaino, J. A. (2016): Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat. Methods*, 13(8):651–656.

Gross, J., Yellen, J., und Zhang, P. (2013): *Handbook of Graph Theory*. CRC Press, 2. Auflage.

Gupta, B. (2002): *Modern Foraminifera*. Springer.

Halkidi, M., Batistakis, Y., und Vazirgiannis, M. (2001): On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145.

Hartuv, E. und Shamir, R. (2000): A clustering algorithm based on graph connectivity. *Information processing letters*, 76(4-6):175–181.

Hastie, T. J., Tibshirani, R. J., und Friedman, J. H. (2009): *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York, 2. Auflage.

Hebert, P. D., Cywinska, A., Ball, S. L., und deWaard, J. R. (2003): Biological identifications through DNA barcodes. *Proc. Biol. Sci.*, 270(1512):313–321.

Hütt, M. T. und Dehnert, M. (2016): *Methoden der Bioinformatik*. Springer Spektrum, Berlin, Heidelberg, 2. Auflage.

Hubert, L. und Arabie, P. (1985): Comparing partitions. *Journal of classification*, 2(1):193–218.

Johnson, J. V., Yost, R. A., Kelley, P. E., und Bradford, D. C. (1990): Tandem-in-space and tandem-in-time mass spectrometry: triple quadrupoles and quadrupole ion traps. *Analytical Chemistry*, 62(20):2162–2172.

Keerthikumar, S. und Mathivanan, S. (2016): *Proteome Bioinformatics*. Methods in Molecular Biology. Springer, New York.

Kennedy, J. J. (1970): The eta coefficient in complex anova designs. *Educational and Psychological Measurement*, 30(4):885–889.

Kerschke, L. (2016): Clustern von massenspektrometrischen Daten. Masterarbeit, TU Dortmund.

Kim, S. und Zhang, X. (2013): Comparative analysis of mass spectral similarity measures on peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry. *Comput Math Methods Med*, 2013:509761.

Kolaczyk, E. D. und Csárdi, G. (2014): *Statistical analysis of network data with R.* Springer, 1. Auflage.

Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., Stein, S. E., und Aebersold, R. (2008): Building consensus spectral libraries for peptide identification in proteomics. *Nature Methods*, 5(10):873–875.

Legendre, P. und Legendre, L. (1998): *Numerical ecology.* Elsevier, Amsterdam, 2. Auflage. Developments in Environmental Modelling 20.

Leighton, F. (1992): *Introduction to Parallel Algorithms and Architectures: Arrays · Trees · Hypercubes.* Morgan Kaufmann Publishers, San Mateo, California.

Lottspeich, F. und Engels, J. (2006): *Bioanalytik.* Springer Spektrum, 2. Auflage.

Ludwig, J. A. und Reynolds, J. F. (1988): *Statistical ecology: a primer in methods and computing*, volume 1. John Wiley & Sons.

Ma, B. (2015): Novor: real-time peptide de novo sequencing software. *J. Am. Soc. Mass Spectrom.*, 26(11):1885–1894.

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., und Lajoie, G. (2003): PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.*, 17(20):2337–2342.

Manning, C. D., Raghavan, P., und Schütze, H. (2008): *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

Mardia, K., Kent, J., und Bibby, J. (1979): *Multivariate Analysis.* Probability and mathematical statistics. Academic Press, London.

Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Rompp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P. A., und Deutsch, E. W. (2011): mzML–a community standard for mass spectrometry data. *Mol. Cell Proteomics*, 10(1):R110.000133.

Method of the Year 2012 (2013): Method of the year 2012. *Nature Methods*, 10(1):1.

Milligan, G. W. und Cooper, M. C. (1986): A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. *Multivariate Behav Res*, 21(4):441–458.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G., und Worm, B. (2011): How many species are there on Earth and in the ocean? *PLoS Biol.*, 9(8):e1001127.

Novak, J. (2013): *Similarity Search in Mass Spectra Databases*. Dissertation, Charles University Praque, Czech Republic.

Novak, J. und Hoksza, D. (2010): Parametrised hausdorff distance as a non-metric similarity model for tandem mass spectrometry. In *DATESO* 1–12.

Novak, J., Sachsenberg, T., Hoksza, D., Skopal, T., und Kohlbacher, O. (2013): On comparison of SimTandem with state-of-the-art peptide identification tools, efficiency of precursor mass filter and dealing with variable modifications. *J Integr Bioinform*, 10(3):228.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., und Wagner, H. (2016): *vegan: Community Ecology Package*. R package version 2.3-4.

Ottenheym, T. (2017): Rekonstruktion phylogenetischer Bäume aus Proteomdaten. Masterarbeit, TU Dortmund.

Palmblad, M. und Deelder, A. M. (2012): Molecular phylogenetics by direct comparison of tandem mass spectra. *Rapid Commun. Mass Spectrom.*, 26(7):728–732.

Paradis, E., Claude, J., und Strimmer, K. (2004): APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290.

Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W., und Aebersold, R. (2004): A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.*, 22(11):1459–1466.

Perkins, D. N., Pappin, D. J., Creasy, D. M., und Cottrell, J. S. (1999): Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567.

Pfenninger, M., Cordellier, M., und Streit, B. (2006): Comparing the efficacy of morphologic and dna-based taxonomy in the freshwater gastropod genus radix (basommatophora, pulmonata). *BMC Evolutionary Biology*, 6(1):100.

Podwojski, K., Fritsch, A., Chamrad, D. C., Paul, W., Sitek, B., Stühler, K., Mutzel, P., Stephan, C., Meyer, H. E., Urfer, W., Ickstadt, K., und Rahnenführer, J. (2009): Retention time alignment algorithms for LC/MS data must consider non-linear shifts. *Bioinformatics*, 25(6):758–764.

R Core Team (2016): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramakrishnan, S. R., Mao, R., Nakorchevskiy, A. A., Prince, J. T., Willard, W. S., Xu, W., Marcotte, E. M., und Miranker, D. P. (2006): A fast coarse filtering method for peptide identification by mass spectrometry. *Bioinformatics*, 22(12):1524–1531.

Rand, W. M. (1971): Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Rieder, V., Blank-Landeshammer, B., Stuhr, M., Schell, T., Biß, K., Kollipara, L., Meyer, A., Pfenninger, M., Westphal, H., Sickmann, A., und Rahnenführer, J. (2017a): DISMS2: A flexible algorithm for direct proteome-wide distance calculation of lc-ms/ms runs. *BMC Bioinformatics*, 18(1):148.

Rieder, V., Schork, K. U., Kerschke, L., Blank-Landeshammer, B., Sickmann, A., und Rahnenführer, J. (2017b): Comparison and evaluation of clustering algorithms for tandem mass spectra. *Journal of Proteome Research*, 16(11):4035–4044.

Robinson, D. F. und Foulds, L. R. (1981): Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2):131–147.

Roepstorff, P. und Fohlman, J. (1984): Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.*, 11(11):601.

Rousseeuw, P. J. (1987): Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Rzhetsky, A. und Nei, M. (1992): A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol.*, 9(5):945–967.

Sadygov, R. G. und Yates, J. R. (2003): A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.*, 75(15):3792–3798.

Saitou, N. und Nei, M. (1987): The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425.

Schell, T., Feldmeyer, B., Schmidt, H., Greshake, B., Tills, O., Truebano, M., Rundle, S. D., Paule, J., Ebersberger, I., und Pfenninger, M. (2017): An annotated draft genome for Radix auricularia (Gastropoda, Mollusca). *Genome Biol Evol.*

Schork, K. U. (2016): Verbesserte Annotation von Massenspektren mit Algorithmen der Clusteranalyse. Masterarbeit, TU Dortmund.

Steen, H. und Mann, M. (2004): The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.*, 5(9):699–711.

Studier, J. A. und Keppler, K. J. (1988): A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.*, 5(6):729–731.

Stuhr, M., Blank-Landeshammer, B., Reymond, C. E., Kollipara, L., Sickmann, A., Kucera, M., und Westphal, H. (2017a): Disentangling thermal stress responses in a reef-calcifier and its photosymbionts by shotgun proteomics. Manuskript zur Veröffentlichung eingereicht bei Scientific Reports.

Stuhr, M., Meyer, A., Reymond, C. E., Narayan, G. R., Rieder, V., Rahnenführer, J., Kucera, M., Westphal, H., und Hallock, P. (2017b): Variable thermal stress tolerance of the reef-associated symbiont-bearing foraminifera *Amphistegina* linked to differences in symbiont type. Manuskript zur Veröffentlichung eingereicht bei Coral Reefs.

Stuhr, M., Reymond, C. E., Rieder, V., Hallock, P., Rahnenführer, J., Westphal, H., und Kucera, M. (2017c): Reef calcifiers are adapted to episodic heat stress but vulnerable to sustained warming. *PLoS ONE*, 12(7):e0179753.

Swift, S., Tucker, A., Vinciotti, V., Martin, N., Orengo, C., Liu, X., und Kellam, P. (2004): Consensus clustering and functional interpretation of gene-expression data. *Genome Biol.*, 5(11):R94.

The, M. und Käll, L. (2016): MaRaCluster: A Fragment Rarity Metric for Clustering Fragment Spectra in Shotgun Proteomics. *J. Proteome Res.*, 15(3):713–720.

Therneau, T., Atkinson, B., und Ripley, B. (2015): *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-10.

Torgerson, W. (1958): *Theory and methods of scaling*. Wiley, New York.

Vaughan, L. K., Divers, J., Padilla, M., Redden, D. T., Tiwari, H. K., Pomp, D., und Allison, D. B. (2009): The use of plasmodes as a supplement to simulations: A simple example evaluating individual admixture estimation methodologies. *Comput Stat Data Anal*, 53(5):1755–1766.

Viswanath, P. und Babu, V. S. (2009): Rough-dbscan: A fast hybrid density based clustering method for large data sets. *Pattern Recognition Letters*, 30(16):1477–1488.

Warrens, M. J. (2008): On the equivalence of cohen's kappa and the hubert-arabie adjusted rand index. *Journal of Classification*, 25(2):177–183.

Yates, J. R., Eng, J. K., Clauser, K. R., und Burlingame, A. L. (1996): Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides. *J. Am. Soc. Mass Spectrom.*, 7(11):1089–1098.

Yilmaz, S., Victor, B., Hulstaert, N., Vandermarliere, E., Barsnes, H., Degroeve, S., Gupta, S., Sticker, A., Gabriel, S., Dorny, P., Palmblad, M., und Martens, L. (2016): A Pipeline for Differential Proteomics in Unsequenced Species. *J. Proteome Res.*, 15(6):1963–1970.

Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A., und Ma, B. (2012): PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell Proteomics*, 11(4):M111.010587.

Zolg, D. P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Bailey, D. J., Gessulat, S., Ehrlich, H. C., Weininger, M., Yu, P., Schlegl, J., Kramer, K., Schmidt, T., Kusebauch, U., Deutsch, E. W., Aebersold, R., Moritz, R. L., Wenschuh, H., Moehring, T., Aiche, S., Huhmer, A., Reimer, U., und Kuster, B. (2017): Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods*, 14(3):259–262.

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Dissertation ist bisher keiner anderen Fakultät vorgelegt worden. Ich erkläre, dass ich bisher kein Promotionsverfahren erfolglos beendet habe und dass keine Aberkennung eines bereits erworbenen Doktorgrades vorliegt.

Dortmund, 5. Februar 2018 Vera Rieder

A Ergänzende Tabellen

Tabelle A.1: Monoisotopische Masse und Einbuchstabencode von Aminosäuren (Lottspeich und Engels, 2006, Tabelle 1 in Anhang 2, S. 1093)

Aminosäure	Einbuchstabencode	Monoisotopische Masse
Alanin	A	71.03711
Arginin	R	156.10111
Asparagin	N	114.04293
Asparaginsäure	D	115.02694
Cystein	C	103.00919
Glutaminsäure	E	129.04259
Glutamin	Q	128.05858
Glycin	G	57.02146
Histidin	H	137.05891
Isoleucin	I	113.08406
Leucin	L	113.08406
Lysin	K	128.09496
Methionin	M	131.04049
Phenylalanin	F	147.06841
Prolin	P	97.05276
Serin	S	87.03203
Threonin	T	101.04768
Tryptophan	W	186.07931
Tyrosin	Y	163.06333
Valin	V	99.06841

Tabelle A.2: Zuordnung der verwendeten Abkürzung und der Bezeichnung der Datensätze im PRIDE-Repository vom Datensatz DISMS2.

Bezeichnung des Datensatzes	Abkürzung
qExHF01_02580	H1
qExHF01_02581	R41
qExHF01_02582	Y1
qExHF01_02583	Al1
qExHF01_02584	D1
qExHF01_02585	R21
qExHF01_02586	C1
qExHF01_02587	Ag1
qExHF01_02588	M1
qExHF01_02590	Y2
qExHF01_02591	D2
qExHF01_02592	H2
qExHF01_02593	R22
qExHF01_02594	Ag2
qExHF01_02595	R42
qExHF01_02596	M2
qExHF01_02597	Al2
qExHF01_02598	C2
qExHF01_02600	R23
qExHF01_02601	M3
qExHF01_02602	Al3
qExHF01_02603	R43
qExHF01_02604	C3
qExHF01_02605	Ag3
qExHF01_02606	H3
qExHF01_02607	Y3
qExHF01_02608	D3

Tabelle A.3: Zuordnung der verwendeten Abkürzung und der Bezeichnung der Datensätze im PRIDE-Repository vom Datensatz Palmblad.

Bezeichnung des Datensatzes	Abkürzung
PRIDE_Exp_Complete_Ac_16286	EC1
PRIDE_Exp_Complete_Ac_16287	GG1
PRIDE_Exp_Complete_Ac_16288	GG2
PRIDE_Exp_Complete_Ac_16289	GG3
PRIDE_Exp_Complete_Ac_16290	GG4
PRIDE_Exp_Complete_Ac_16291	HS1
PRIDE_Exp_Complete_Ac_16292	HS2
PRIDE_Exp_Complete_Ac_16293	HS3
PRIDE_Exp_Complete_Ac_16294	HS4
PRIDE_Exp_Complete_Ac_16295	MF1
PRIDE_Exp_Complete_Ac_16296	MF2
PRIDE_Exp_Complete_Ac_16297	MF3
PRIDE_Exp_Complete_Ac_16298	MF4
PRIDE_Exp_Complete_Ac_16299	MM1
PRIDE_Exp_Complete_Ac_16300	MM2
PRIDE_Exp_Complete_Ac_16301	MM3
PRIDE_Exp_Complete_Ac_16302	MM4
PRIDE_Exp_Complete_Ac_16303	PP1
PRIDE_Exp_Complete_Ac_16304	PP2
PRIDE_Exp_Complete_Ac_16305	PP3
PRIDE_Exp_Complete_Ac_16306	PP4
PRIDE_Exp_Complete_Ac_16307	PT1
PRIDE_Exp_Complete_Ac_16308	PT2
PRIDE_Exp_Complete_Ac_16309	PT3
PRIDE_Exp_Complete_Ac_16310	PT4
PRIDE_Exp_Complete_Ac_16311	PT5
PRIDE_Exp_Complete_Ac_16312	PT6

Tabelle A.4: Anzahl Spektren einzelner Läufe in den Datensätzen Biodiversität-Exactive (E) und Biodiversität-Orbitrap (O).

Lauf	Anzahl Spektren (E)	Anzahl Spektren (O)
MOTU41	47319	33349
MOTU42	49600	33275
MOTU43	49044	35382
Wa1	42428	31930
Wb1	45325	33336
Wc1	45554	33424
Wa2	44414	33332
Wb2	45830	33116
Wc2	43623	33674
DGE15a1	47068	33486
DGE15a2	45395	32631
KAT1	47968	33478
KAT2	45029	32661
KAT3	45318	33023
KAT4	44400	32698
KSHT2	46394	33585
KSHT3	45936	33752
KSHT6	42103	31917
KSHT7	44338	33162
KSHT8	43576	34277
KSHT9	46304	32772
KSHT10	48929	33988
KSHT11	42756	31677
KSHT12	40554	31609
KSHT13	45425	32923
A1	45744	33833
LT1	45818	33206
FDM1	43054	36418
FDM2	37775	31079
FDM3	39916	31845

Tabelle A.5: Parameteroptimierung im Palmblad-Datensatz mithilfe des Bestimmtheitsmaßes R^2. In allen 12 Kombinationen gilt `prec = 10000`, `topn` $= 50$, `bin` $= 0.2$ und `rtol` $=$ 3000. Optimale Werte bezüglich verschiedener Parameter sind durch fett gedruckte Zeilen gekennzeichnet.

Rang	`dist`	`cdis`	R^2
1	$\mathbf{d_{cos}}$	**0.1**	**0.3853**
2	$d_{\cos}$	0.2	0.3853
3	$d_{\cos}$	0.3	0.3853
4	$\mathbf{d_{PH}}$	**0.1**	**0.2319**
5	d_{PH}	0.2	0.2319
6	d_{PH}	0.3	0.2319
7	$\mathbf{d_{angle}}$	**0.1**	**0.2308**
8	d_{angle}	0.2	0.2308
9	d_{angle}	0.3	0.2308
10	d_{angle}	0.4	0.2308
11	d_{angle}	0.5	0.2308
12	d_{angle}	0.6	0.2308

Tabelle A.6: Standardfehler der Distanz der Läufe für verschiedene Methoden im DISMS2-Datensatz (Rieder et al., 2017a, Tabelle S2)

	DB.ra	DB.ra.nodup	DISMS2.f	DB.a	DISMS2.af	DB.af
C vs. C	0.00114	0.00176	0.00164	0.00157	0.00133	0.00164
D vs. D	0.00537	0.00725	0.00663	0.00473	0.00448	0.00663
H vs. H	0.00220	0.00267	0.00155	0.00121	0.00109	0.00155
M vs. M	0.00241	0.00319	0.00363	0.00245	0.00338	0.00363
Y vs. Y	0.00188	0.00297	0.00254	0.00196	0.00253	0.00254
C vs. D	0.00038	0.00010	0.00167	0.00019	0.00070	0.00167
C vs. H	0.00026	0.00006	0.00065	0.00020	0.00031	0.00065
C vs. M	0.00016	0.00019	0.00178	0.00006	0.00077	0.00178
C vs. Y	0.00005	0.00007	0.00121	0.00004	0.00023	0.00121
D vs. H	0.00033	0.00023	0.00085	0.00017	0.00078	0.00085
D vs. M	0.00024	0.00016	0.00076	0.00009	0.00013	0.00076
D vs. Y	0.00017	0.00006	0.00036	0.00008	0.00022	0.00036
H vs. M	0.00079	0.00087	0.00138	0.00082	0.00161	0.00138
H vs. Y	0.00010	0.00009	0.00125	0.00007	0.00019	0.00125
M vs. Y	0.00012	0.00009	0.00050	0.00008	0.00022	0.00050

Tabelle A.7: Zusatzinformationen zum Algorithmus DB.af. Mittlerer relativer Anteil an Partnern (gleiches Peptid), anderer Peptide, fehlender Annotationen in Liste 1 und kein Spektrum, das die Kriterien der Filterkontrolle in Liste 2 erfüllt (Rieder et al., 2017a, Tabelle S3).

	Gleiches Peptid	Anderes Peptid	Liste 1: fehlende Annotation	Liste 2: Kein Spektrum erfüllt Kriterien
C vs. C	0.548	0.034	0.290	0.128
D vs. D	0.467	0.019	0.406	0.108
H vs. H	0.502	0.031	0.341	0.126
M vs. M	0.492	0.026	0.360	0.122
Y vs. Y	0.371	0.013	0.535	0.082
C vs. D	0.024	0.135	0.348	0.493
C vs. H	0.027	0.160	0.316	0.497
C vs. M	0.027	0.146	0.325	0.501
C vs. Y	0.007	0.114	0.412	0.466
D vs. H	0.031	0.132	0.373	0.464
D vs. M	0.033	0.122	0.383	0.462
D vs. Y	0.008	0.097	0.470	0.424
H vs. M	0.256	0.087	0.350	0.307
H vs. Y	0.009	0.112	0.438	0.441
M vs. Y	0.009	0.105	0.447	0.438

Tabelle A.8: Variationskoeffizient des Betrags der Differenz zwischen zwei Proteomvergleichmethoden (*Datenbank*) innerhalb (oben) und zwischen (unten) Arten von Fadenwurm (C), Fruchtfliege (D), Mensch (H), Maus (M) und Hefe (Y). Werte kleiner 0.5 sind markiert (*) und kennzeichnen relevante Unterschiede zwischen den entsprechenden Paaren (Rieder et al., 2017a, Tabelle S4)

Method A vs. Method B	DB.ra.nodup DB.ra	DB.a DB.ra	DB.af DB.a
C vs. C	0.0201	0.0034	0.0103
D vs. D	0.0511	0.0031	0.0439
H vs. H	0.0167	0.0116	0.0112
M vs. M	0.0260	0.0107	0.0127
Y vs. Y	0.0376	0.0014	0.0128
C vs. D	0.0441	0.0373	0.3762
C vs. H	0.0358	0.0154	0.0619
C vs. M	0.0231	0.0216	0.5253*
C vs. Y	0.0384	0.0110	0.4577
D vs. H	0.0236	0.0267	0.1336
D vs. M	0.0190	0.0221	0.0675
D vs. Y	0.0426	0.0339	0.1248
H vs. M	0.0124	0.0125	0.0163
H vs. Y	0.0138	0.0156	0.2553
M vs. Y	0.0185	0.0198	0.2273

Tabelle A.9: Variationskoeffizient des Betrags der Differenz zwischen zwei Proteomvergleichmethoden innerhalb (oben) und zwischen (unten) Arten von Fadenwurm (C), Fruchtfliege (D), Mensch (H), Maus (M) und Hefe (Y). Werte kleiner 0.5 sind markiert (*) und kennzeichnen relevante Unterschiede zwischen den entsprechenden Paaren (Rieder et al., 2017a, Tabelle S5)

Method A vs. Method B	DISMS2.af DB.af	DISMS2.af DISMS2.f	DISMS2.f DB.ra	DISMS2.f DB.ra.nodup	DB.a DISMS2.f
C vs. C	0.4448	0.0047	0.0065	0.0062	0.0031
D vs. D	0.2715	0.0156	0.0193	0.0270	0.0235
H vs. H	0.9901*	0.0224	0.0161	0.0309	0.0277
M vs. M	0.3774	0.0056	0.0140	0.0111	0.0261
Y vs. Y	0.6739*	0.0080	0.0072	0.0143	0.0037
C vs. D	0.1631	0.1949	0.4332	0.1568	0.1841
C vs. H	0.0404	0.0794	0.0884	0.0491	0.0555
C vs. M	0.1614	0.2436	0.5458*	0.1774	0.2142
C vs. Y	0.0952	0.4221	0.2509	0.1832	0.1779
D vs. H	0.1246	0.0630	0.1267	0.0512	0.0610
D vs. M	0.0212	0.0753	0.1360	0.0499	0.0595
D vs. Y	0.0597	0.0657	0.0337	0.0330	0.0303
H vs. M	0.8931*	0.0303	0.0389	16.6491*	0.0242
H vs. Y	0.0519	0.1780	0.1468	0.1111	0.1134
M vs. Y	0.0557	0.0918	0.0781	0.0560	0.0552

Tabelle A.10: Standardfehler der Distanz der Läufe für zwei Methoden im Foraminiferen-Datensatz

	DISMS2.f	DB.ra.nodup
Agi vs. Agi	0.0133	0.0095
Ale vs. Ale	0.0121	0.0099
Alo vs. Alo	0.0068	0.0059
Mve vs. Mve	0.0402	0.0565
Agi vs. Ale	0.0034	0.0022
Agi vs. Alo	0.0024	0.0020
Agi vs. Mve	0.0010	0.0011
Ale vs. Alo	0.0066	0.0051
Ale vs. Mve	0.0008	0.0007
Alo vs. Mve	0.0007	0.0008

Tabelle A.11: Parameteroptimierung im Foraminiferen-Datensatz mithilfe des Bestimmtheitsmaßes R^2. In allen 81 Kombinationen gilt `prec = 10`. Optimale Werte bezüglich verschiedener Parameter sind durch fett gedruckte Zeilen gekennzeichnet.

Rang	`topn`	`bin`	`ret`	`dist`	`cdis`	R^2
1	∞	**0.20**	**3000**	$d_{\mathbf{cos}}$	**0.3**	**0.618**
2	∞	0.20	3000	$d_{\cos}$	0.1	0.618
3	50	0.20	3000	$d_{\cos}$	0.1	0.617
4	50	0.20	3000	$d_{\cos}$	0.3	0.617
5	20	0.20	3000	$d_{\cos}$	0.1	0.614
6	20	0.20	3000	$d_{\cos}$	0.3	0.614
7	20	0.01	3000	$d_{\cos}$	0.1	0.579
8	20	0.01	3000	$d_{\cos}$	0.3	0.579
9	50	0.01	3000	$d_{\cos}$	0.1	0.578
10	50	0.01	3000	$d_{\cos}$	0.3	0.578
11	∞	0.01	3000	$d_{\cos}$	0.1	0.576
12	∞	0.01	3000	$d_{\cos}$	0.3	0.576
13	**20**	**0.01**	**3000**	$d_{\mathbf{PH}}$	**0.1**	**0.572**
⋮	⋮	⋮	⋮	⋮	⋮	⋮
14	20	0.01	3000	d_{PH}	0.3	0.572
⋮	⋮	⋮	⋮	⋮	⋮	⋮
43	**20**	**0.01**	**3000**	$d_{\mathbf{angle}}$	**0.1**	**0.354**
⋮	⋮	⋮	⋮	⋮	⋮	⋮
81	∞	0.20	3000	d_{angle}	0.6	0.310

Tabelle A.12: Adjustierter Rand-Index zwischen Annotation und Clusterlösungen von drei technischen Replikaten der Spezies C, D, H, M, Y und R4 (in Anlehnung an Rieder et al., 2017b, Tabelle S-6). Zum Teil konnte keine Clusterlösung berechnet werden (* Speicherlimit überschritten).

Distanz	Clusterverfahren	C	D	H	M	Y	R4
Kosinus-	CAST (0.1)	0.578	0.653	0.662	0.659	0.653	0.650
Distanz	DBSCAN (0.05, 2)	0.609	0.664	0.691	0.683	0.707	0.646
	hclust (0.1)	*	0.496	*	*	0.698	0.631
	igraph (0.05)	0.609	0.664	0.691	0.683	0.707	0.646
	N-Cluster (0.05)	0.601	0.645	0.660	0.665	0.694	0.627
Kosinus-	CAST (0.2)	0.672	0.713	0.740	0.733	0.760	0.708
Distanz mit	DBSCAN (0.2, 2)	0.703	0.730	0.780	0.747	0.770	0.728
DISMS2-	hclust (0.2)	*	0.520	*	*	0.738	0.684
Filter	igraph (0.2)	0.703	0.730	0.780	0.747	0.770	0.728
	N-Cluster (0.2)	0.683	0.725	0.749	0.741	0.766	0.715
Teil des	MS-Cluster (0.8)	0.614	0.483	0.652	0.519	0.692	0.598
Algorithmus	PRIDE Cluster (0.8)	0.647	0.677	0.675	0.700	0.714	0.656

Tabelle A.13: Anteil verbleibender Annotationen von drei technischen Replikaten der Spezies C, D, H, M, Y und R4 (in Anlehnung an Rieder et al., 2017b, Tabelle S-9). Zum Teil konnte keine Clusterlösung berechnet werden (* Speicherlimit überschritten).

Distanz	Clusterverfahren	C	D	H	M	Y	R4
Kosinus-	CAST (0.1)	0.968	0.973	0.969	0.970	0.970	0.976
Distanz	DBSCAN (0.05, 2)	0.987	0.990	0.989	0.989	0.992	0.994
	hclust (0.1)	*	0.981	*	*	0.978	0.982
	igraph (0.05)	0.987	0.990	0.989	0.989	0.992	0.994
	N-Cluster (0.05)	0.990	0.992	0.990	0.990	0.993	0.995
Kosinus-	CAST (0.2)	0.987	0.989	0.987	0.989	0.988	0.991
Distanz mit	DBSCAN (0.2, 2)	0.983	0.986	0.983	0.985	0.984	0.988
DISMS2-	hclust (0.2)	*	0.991	*	*	0.990	0.992
Filter	igraph (0.2)	0.983	0.986	0.983	0.985	0.984	0.988
	N-Cluster (0.2)	0.985	0.987	0.985	0.987	0.986	0.990
Teil des	MS-Cluster (0.8)	0.987	0.938	0.989	0.952	0.995	0.991
Algorithmus	PRIDE Cluster (0.8)	0.984	0.986	0.985	0.988	0.989	0.990

Tabelle A.14: Clusteranzahl in Relation zur Anzahl an Spektren von drei technischen Replikaten der Spezies C, D, H, M, Y und R4 (in Anlehnung an Rieder et al., 2017b, Tabelle S-10). Zum Teil konnte keine Clusterlösung berechnet werden (* Speicherlimit überschritten).

Distanz	Clusterverfahren	C	D	H	M	Y	R4
Kosinus-Distanz	CAST (0.1)	0.411	0.395	0.422	0.416	0.389	0.404
	DBSCAN (0.05, 2)	0.450	0.432	0.459	0.455	0.426	0.447
	hclust (0.1)	*	0.440	*	*	0.400	0.414
	igraph (0.05)	0.450	0.432	0.459	0.455	0.426	0.447
	N-Cluster (0.05)	0.454	0.437	0.463	0.459	0.430	0.452
Kosinus-Distanz mit DISMS2-Filter	CAST (0.2)	0.404	0.382	0.414	0.407	0.378	0.389
	DBSCAN (0.2, 2)	0.398	0.375	0.409	0.401	0.372	0.383
	hclust (0.2)	*	0.428	*	*	0.383	0.394
	igraph (0.2)	0.398	0.375	0.409	0.401	0.372	0.383
	N-Cluster (0.2)	0.400	0.377	0.411	0.403	0.374	0.385
Teil des Algorithmus	MS-Cluster (0.8)	0.426	0.387	0.441	0.415	0.413	0.426
	PRIDE Cluster (0.8)	0.355	0.338	0.373	0.364	0.342	0.349

Tabelle A.15: Anteil an Spektren in mehrelementigen Clustern von drei technischen Replikaten der Spezies C, D, H, M, Y und R4 (in Anlehnung an Rieder et al., 2017b, Tabelle S-7). Zum Teil konnte keine Clusterlösung berechnet werden (* Speicherlimit überschritten).

Distanz	Clusterverfahren	C	D	H	M	Y	R4
Kosinus-Distanz	CAST (0.1)	0.835	0.839	0.831	0.832	0.851	0.825
	DBSCAN (0.05, 2)	0.791	0.796	0.786	0.785	0.812	0.773
	hclust (0.1)	*	0.824	*	*	0.849	0.823
	igraph (0.05)	0.791	0.796	0.786	0.785	0.812	0.773
	N-Cluster (0.05)	0.787	0.792	0.784	0.782	0.809	0.770
Kosinus-Distanz mit DISMS2-Filter	CAST (0.2)	0.848	0.855	0.843	0.847	0.868	0.846
	DBSCAN (0.2, 2)	0.851	0.859	0.847	0.850	0.872	0.850
	hclust (0.2)	*	0.839	*	*	0.867	0.844
	igraph (0.2)	0.851	0.859	0.847	0.850	0.872	0.850
	N-Cluster (0.2)	0.850	0.858	0.846	0.849	0.871	0.848
Teil des Algorithmus	MS-Cluster (0.8)	0.825	0.853	0.815	0.838	0.833	0.805
	PRIDE Cluster (0.8)	0.897	0.898	0.886	0.892	0.905	0.885

Tabelle A.16: Spektrenanteil ohne häufigste Annotation von drei technischen Replikaten der Spezies C, D, H, M, Y und R4 (in Anlehnung an Rieder et al., 2017b, Tabelle S-8). Zum Teil konnte keine Clusterlösung berechnet werden (* Speicherlimit überschritten).

Distanz	Clusterverfahren	C	D	H	M	Y	R4
Kosinus-Distanz	CAST (0.1)	0.029	0.024	0.026	0.026	0.026	0.020
	DBSCAN (0.05, 2)	0.011	0.009	0.009	0.010	0.006	0.007
	hclust (0.1)	*	0.018	*	*	0.015	0.013
	igraph (0.05)	0.011	0.009	0.009	0.010	0.006	0.007
	N-Cluster (0.05)	0.009	0.008	0.008	0.008	0.006	0.006
Kosinus-Distanz mit DISMS2-Filter	CAST (0.2)	0.009	0.009	0.009	0.008	0.008	0.007
	DBSCAN (0.2, 2)	0.012	0.012	0.012	0.011	0.011	0.009
	hclust (0.2)	*	0.013	*	*	0.006	0.006
	igraph (0.2)	0.012	0.012	0.012	0.011	0.011	0.009
	N-Cluster (0.2)	0.011	0.011	0.011	0.010	0.009	0.009
Teil des Algorithmus	MS-Cluster (0.8)	0.011	0.133	0.008	0.115	0.005	0.006
	PRIDE Cluster (0.8)	0.018	0.016	0.018	0.015	0.013	0.014

Tabelle A.17: Laufzeit in Minuten (und Speicherverbrauch in Gb) von Clusterlösungen einzelner Läufe H1, H2, H3 (in Anlehnung an Rieder et al., 2017b, Tabelle S-3).

Distanz	Clusterverfahren	H1	H2	H3
Kosinus-Distanz	CAST (0.1)	12.38 (7.40)	12.23 (7.66)	13.02 (8.12)
	DBSCAN (0.05, 2)	7.62 (7.40)	8.00 (7.66)	9.50 (8.11)
	hclust (0.1)	4.75 (7.40)	4.87 (7.66)	5.13 (8.12)
	igraph (0.05)	3.13 (12.30)	3.25 (12.73)	3.48 (13.49)
	N-Cluster (0.05)	4.18 (7.40)	4.33 (7.66)	4.60 (8.12)
Kosinus-Distanz mit DISMS2-Filter	CAST (0.2)	10.87 (7.40)	10.68 (7.66)	11.52 (8.12)
	DBSCAN (0.2, 2)	7.13 (7.40)	7.25 (7.66)	7.63 (8.12)
	hclust (0.2)	2.90 (7.40)	2.98 (7.66)	3.17 (8.12)
	igraph (0.2)	1.77 (12.30)	1.78 (12.73)	1.98 (13.49)
	N-Cluster (0.2)	2.85 (7.40)	2.88 (7.66)	3.10 (8.12)
Teil des Algorithmus	MS-Cluster (0.8)	0.75 (0.14)	0.78 (0.15)	0.82 (0.15)
	PRIDE Cluster (0.8)	2.97 (0.14)	2.80 (0.14)	6.15 (0.14)

Tabelle A.18: Laufzeit in Minuten (und Speicherverbrauch in Gb) von Clusterlösungen von drei Replikaten der Arten C, D und H (in Anlehnung an Rieder et al., 2017b, Tabelle S-4). Zum Teil konnte keine Clusterlösung berechnet werden (* Speicherlimit überschritten).

Distanz	Clusterverfahren	C1, C2, C3	D1, D2, D3	H1, H2, H3
Kosinus-Distanz	CAST (0.1)	67.05 (36.22)	38.82 (22.74)	53.30 (28.37)
	DBSCAN (0.05, 2)	46.60 (38.26)	26.83 (23.51)	38.27 (30.02)
	hclust (0.1)	*	27.20 (51.42)	*
	igraph (0.05)	28.52 (58.66)	17.65 (28.56)	26.72 (59.90)
	N-Cluster (0.05)	33.68 (36.95)	19.47 (23.47)	25.80 (28.94)
Kosinus-Distanz mit DISMS2-Filter	CAST (0.2)	57.12 (36.20)	33.25 (22.75)	45.55 (28.35)
	DBSCAN (0.2, 2)	36.82 (38.13)	21.47 (23.51)	29.40 (29.75)
	hclust (0.2)	*	20.82 (52.44)	*
	igraph (0.2)	21.20 (58.66)	14.27 (28.56)	21.82 (59.90)
	N-Cluster (0.2)	25.25 (37.00)	14.85 (23.47)	20.25 (28.95)
Teil des Algorithmus	MS-Cluster (0.8)	0.92 (0.15)	0.80 (0.12)	0.88 (0.12)
	PRIDE Cluster (0.8)	3.80 (0.15)	4.00 (0.14)	5.08 (0.15)

Tabelle A.19: Laufzeit in Minuten (und Speicherverbrauch in Gb) von Clusterlösungen von drei Replikaten der Arten M, Y und R4 (in Anlehnung an Rieder et al., 2017b, Tabelle S-5). Zum Teil konnte keine Clusterlösung berechnet werden (* Speicherlimit überschritten).

Distanz	Clusterverfahren	M1, M2, M3	Y1, Y2, Y3	R41, R42, R43
Kosinus-Distanz	CAST (0.1)	45.23 (26.10)	23.55 (14.08)	20.58 (12.96)
	DBSCAN (0.05, 2)	32.40 (27.23)	17.07 (14.08)	14.73 (12.96)
	hclust (0.1)	*	17.05 (31.32)	14.33 (27.73)
	igraph (0.05)	20.35 (33.09)	10.97 (17.09)	10.58 (15.74)
	N-Cluster (0.05)	23.55 (26.61)	11.85 (14.08)	10.45 (12.96)
Kosinus-Distanz mit DISMS2-Filter	CAST (0.2)	38.65 (26.07)	19.25 (14.08)	7.94 (12.96)
	DBSCAN (0.2, 2)	25.42 (27.10)	12.97 (14.08)	8.21 (12.96)
	hclust (0.2)	*	12.80 (32.02)	11.38 (28.43)
	igraph (0.2)	14.63 (33.09)	7.37 (17.09)	9.05 (15.74)
	N-Cluster (0.2)	17.23 (26.61)	9.13 (14.08)	8.08 (12.96)
Teil des Algorithmus	MS-Cluster (0.8)	0.82 (0.12)	0.53 (0.10)	0.60 (0.10)
	PRIDE Cluster (0.8)	4.42 (0.15)	2.42 (0.13)	2.43 (0.15)

B Ergänzende Abbildungen

Eingabe : Clusterung $C = \{C_1, \ldots, C_k\}$

Ausgabe : Vektor der Cluster-IDs $cid \in \mathbb{N}_+^n$

1: $cid = (0, \ldots, 0)$' {Erstelle n-dimensionalen Nullvektor cid}
2: **for** $i = 1, \ldots, n$ **do**
3: **if** $i \in C_j$ **then**
4: $cid_{[i]} = j$
5: **end if**
6: **end for**

Abbildung B.1: Hilfsfunktion Helper zur Transformation von Ausgabewerten der Clusteralgorithmen im Pseudocode in den Ausgabewert der zugehörigen R-Funktion.

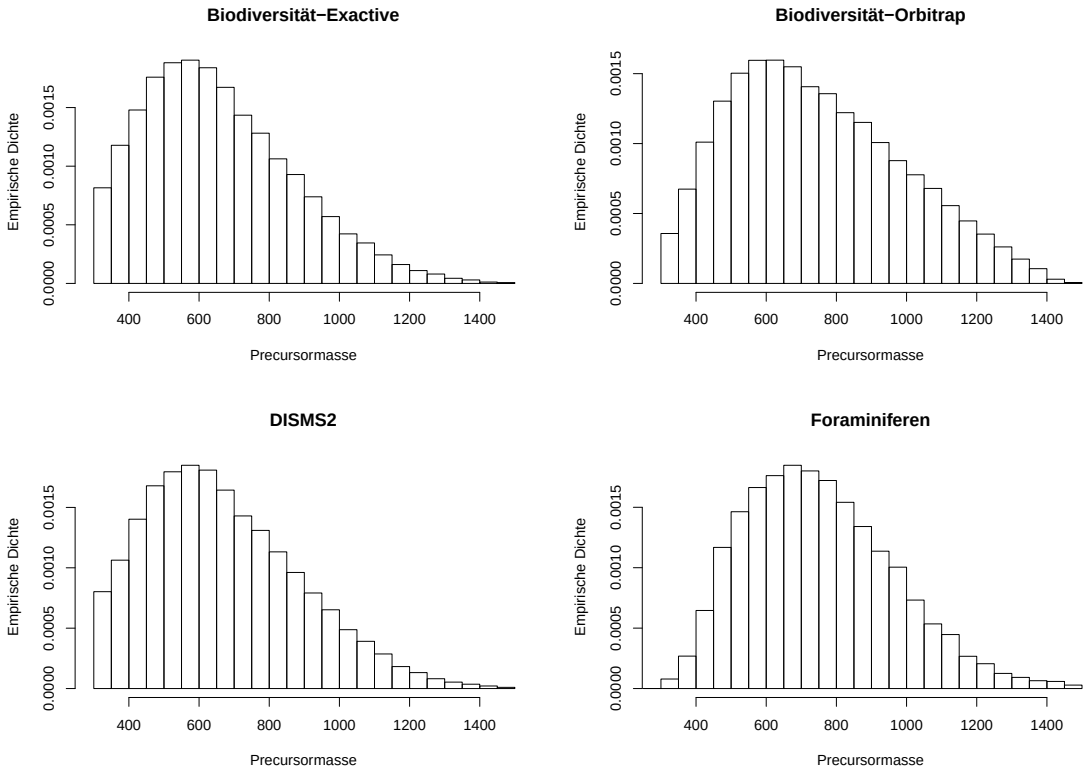

Abbildung B.2: Histogramm des Precursormasse-zu-Ladung-Verhältnis (m/z) in den vier Datensätzen des Leibniz-Projekts.

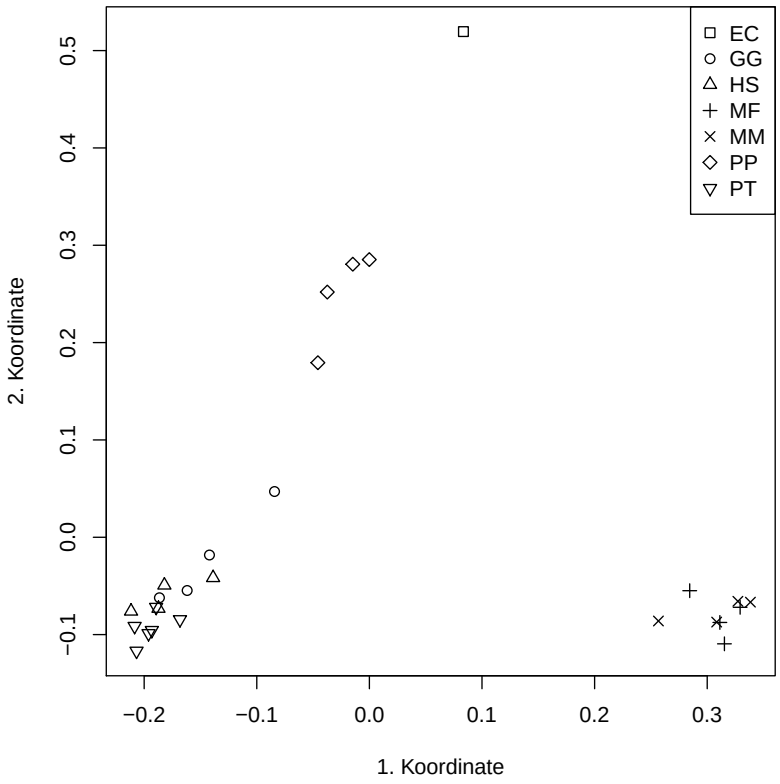

Abbildung B.3: Multidimensionale Skalierung der Distanzen von LC-MS/MS-Läufen im Datensatz Palmblad.

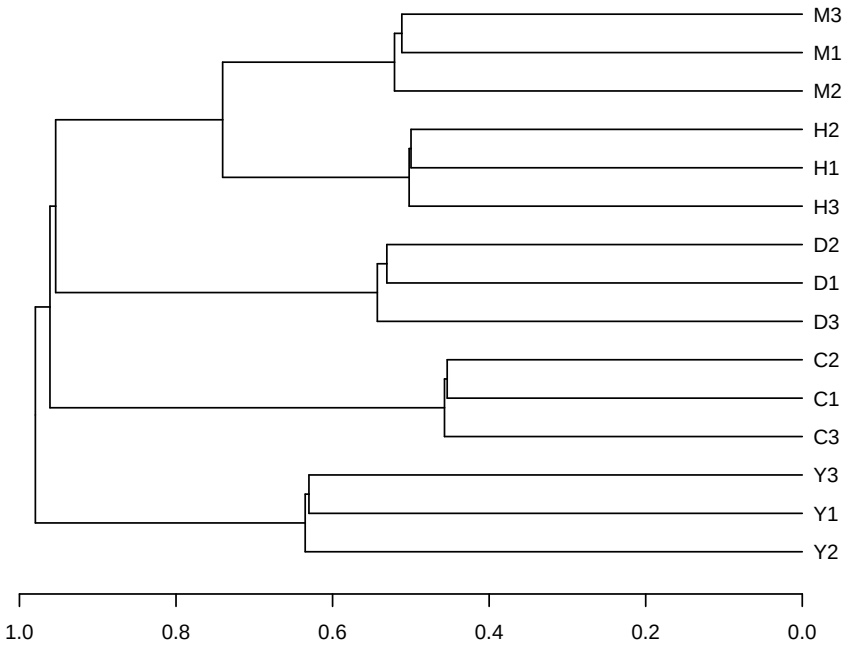

Abbildung B.4: Dendrogramm basierend auf einer hierarchischen Clusteranalyse mit Average-Linkage für die DISMS2-Distanzen von 15 annotierten Läufen des DISMS2-Datensatzes (DISMS2.af) mit optimierten Parametereinstellungen (Rieder et al., 2017a, Abbildung 6).

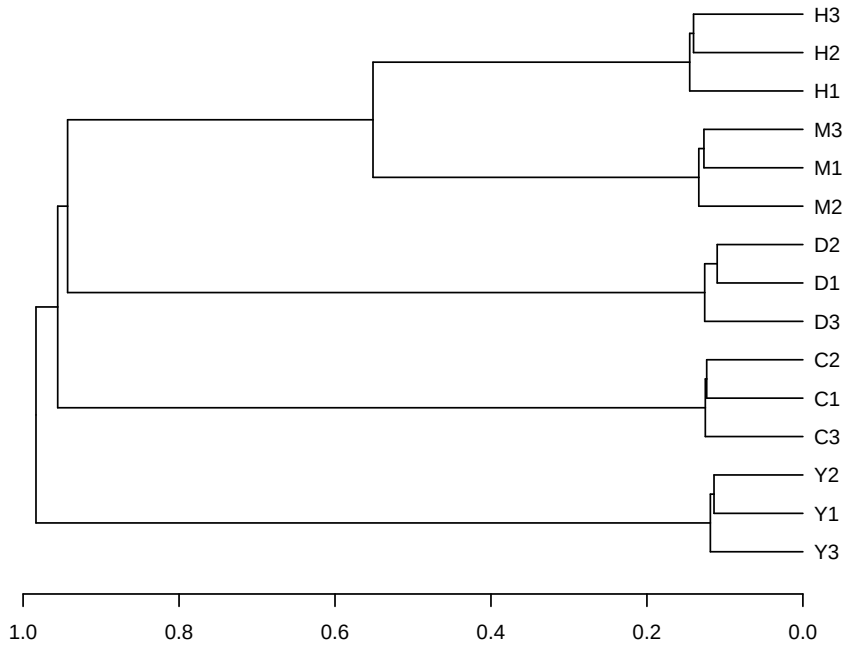

Abbildung B.5: Dendrogramm basierend auf einer hierarchischen Clusteranalyse mit Average-Linkage für die DISMS2-Distanzen von 15 annotierten Läufen des DISMS2-Datensatzes (DB.ra) mit optimierten Parametereinstellungen (Rieder et al., 2017a, Abbildung S1).

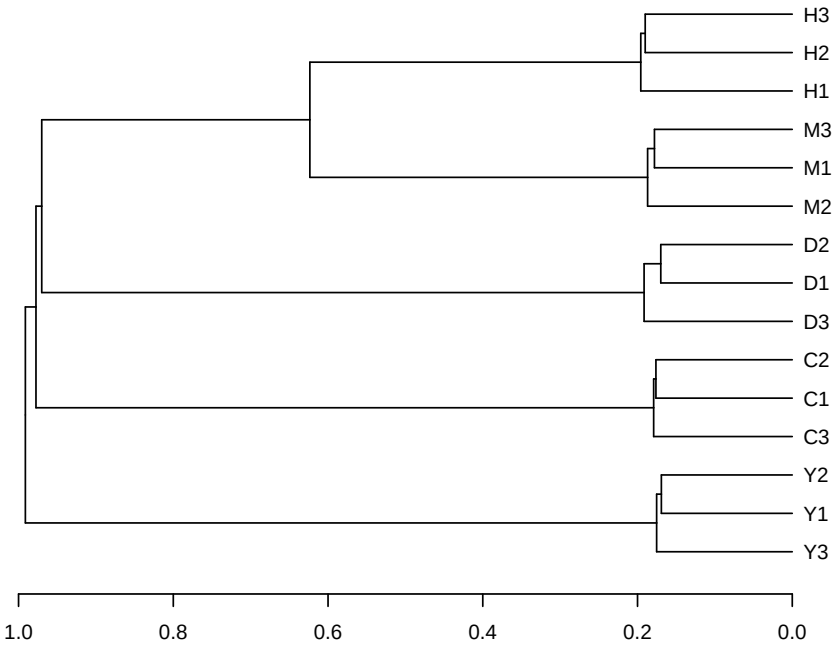

Abbildung B.6: Dendrogramm basierend auf einer hierarchischen Clusteranalyse mit Average-Linkage für die DISMS2-Distanzen von 15 annotierten Läufen des DISMS2-Datensatzes (DB.ra.nodup) mit optimierten Parametereinstellungen (Rieder et al., 2017a, Abbildung S2).

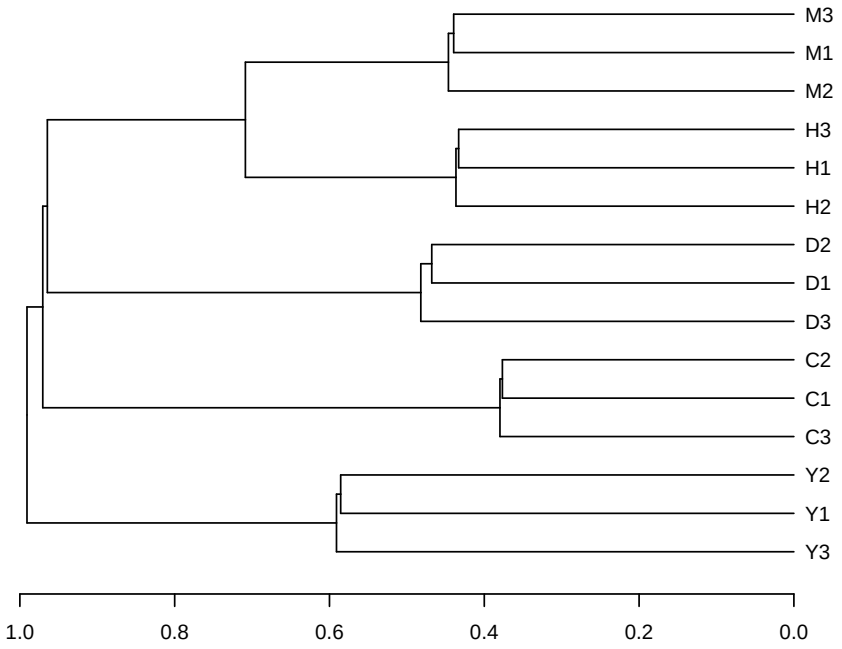

Abbildung B.7: Dendrogramm basierend auf einer hierarchischen Clusteranalyse mit Average-Linkage für die DISMS2-Distanzen von 15 annotierten Läufen des DISMS2-Datensatzes (DB.a) mit optimierten Parametereinstellungen (Rieder et al., 2017a, Abbildung S3).

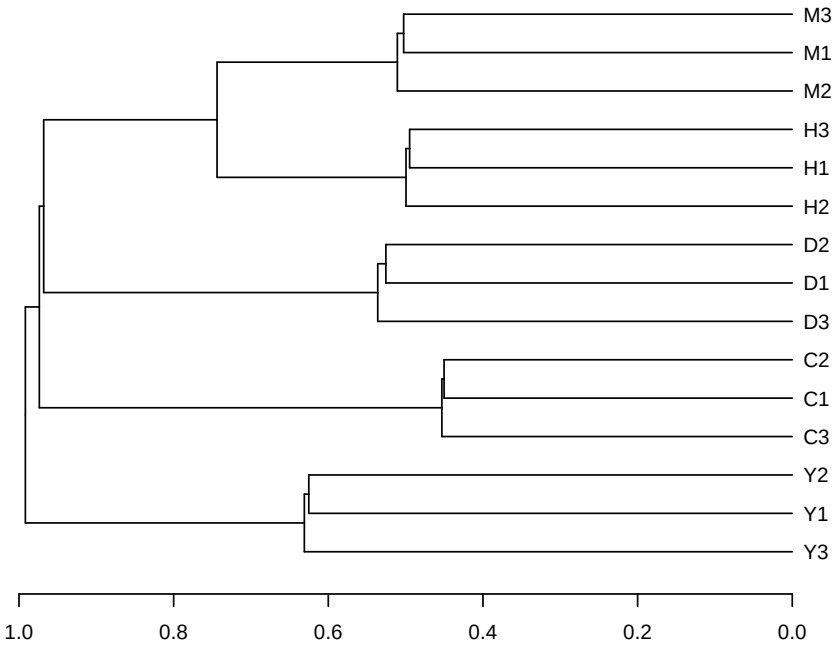

Abbildung B.8: Dendrogramm basierend auf einer hierarchischen Clusteranalyse mit Average-Linkage für die DISMS2-Distanzen von 15 annotierten Läufen des DISMS2-Datensatzes (DB.af) mit optimierten Parametereinstellungen (Rieder et al., 2017a, Abbildung S4).

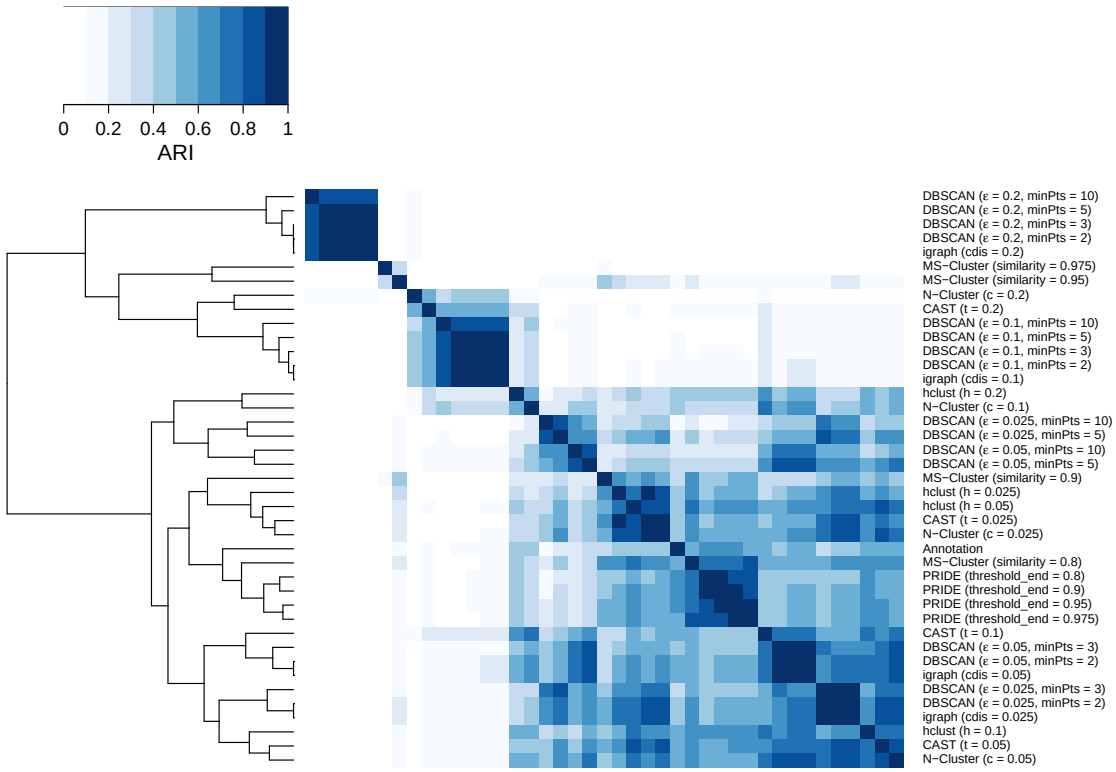

Abbildung B.9: Heatmap des mittleren adjustierten Rand-Index von 18 Läufen im Datensatz DISMS2. Alle Clusterlösungen (ausgenommen MS-Cluster und PRIDE Cluster) wurden auf Basis der Kosinus-Distanz von Spektren generiert (in Anlehnung an Rieder et al., 2017b, Abbildung S-1).

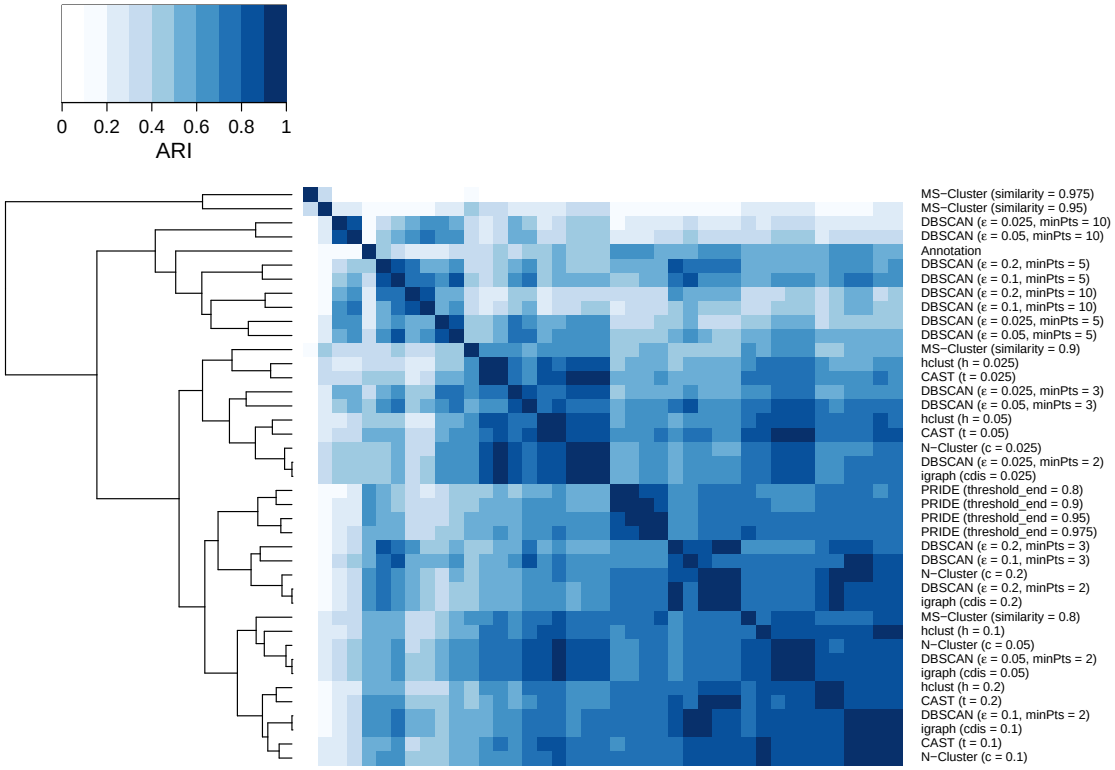

Abbildung B.10: Heatmap des mittleren adjustierten Rand-Index von 18 Läufen im Datensatz DISMS2. Alle Clusterlösungen (ausgenommen MS-Cluster und PRIDE Cluster) wurden auf Basis des DISMS2-Filters und der Kosinus-Distanz von Spektren generiert (in Anlehnung an Rieder et al., 2017b, Abbildung S-2).

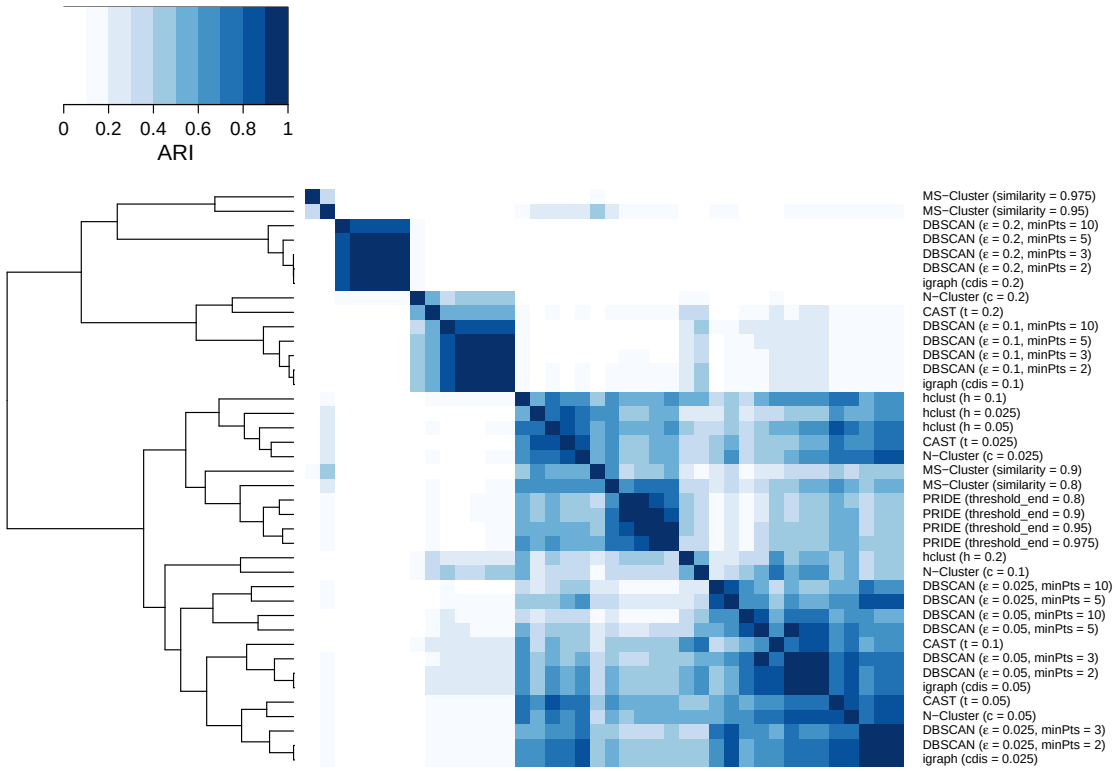

Abbildung B.11: Heatmap des mittleren adjustierten Rand-Index von 27 Läufen im Datensatz DISMS2. Alle Clusterlösungen (ausgenommen MS-Cluster und PRIDE Cluster) wurden auf Basis der Kosinus-Distanz von Spektren generiert (in Anlehnung an Rieder et al., 2017b, Abbildung S-3).

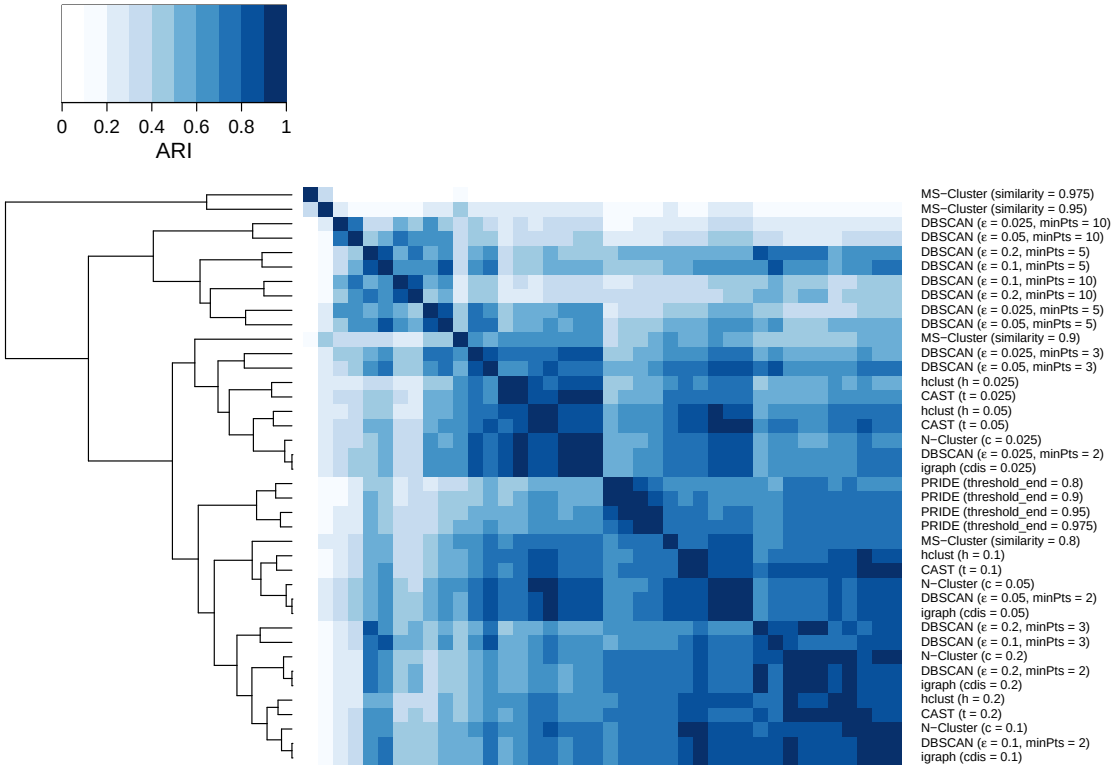

Abbildung B.12: Heatmap des mittleren adjustierten Rand-Index von 27 Läufen im Datensatz DISMS2. Alle Clusterlösungen (ausgenommen MS-Cluster und PRIDE Cluster) wurden auf Basis des DISMS2-Filters und der Kosinus-Distanz von Spektren generiert (in Anlehnung an Rieder et al., 2017b, Abbildung S-4).

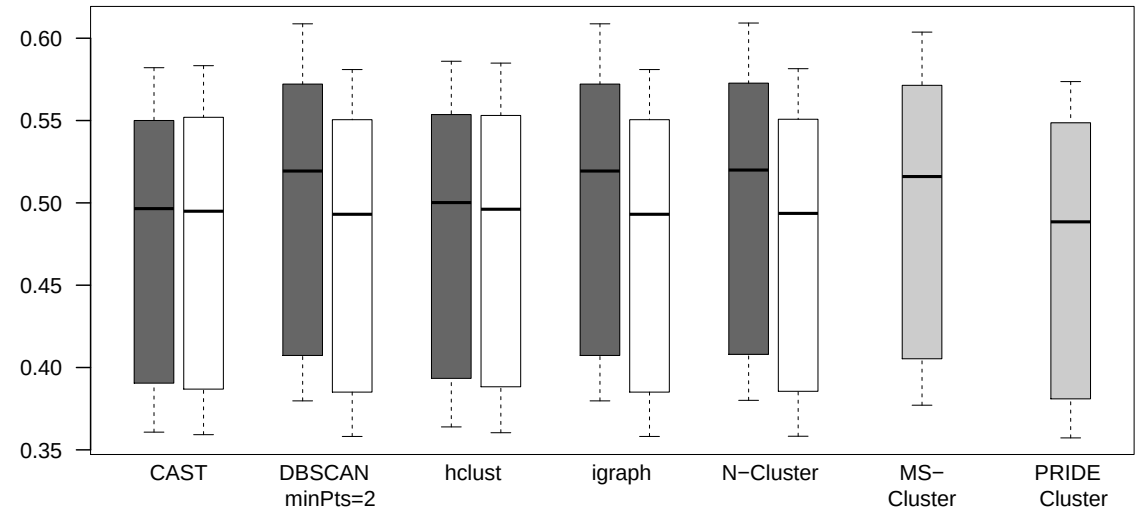

Abbildung B.13: Boxplots der Custeranzahl in Relation zur Spektrenanzahl für einzelne Läufe. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

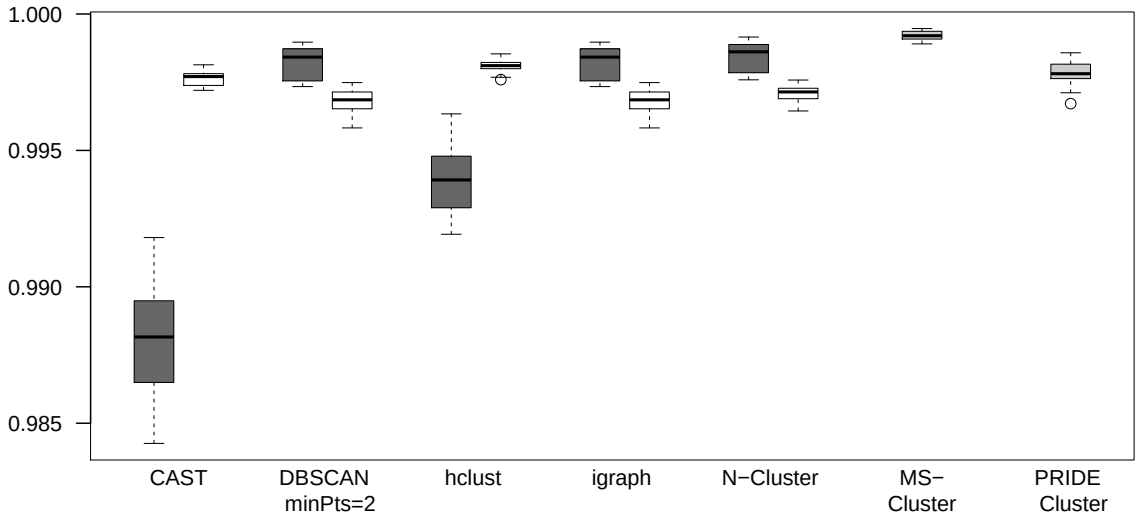

Abbildung B.14: Boxplots des Anteils verbleibender Annotationen für einzelne Läufe. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

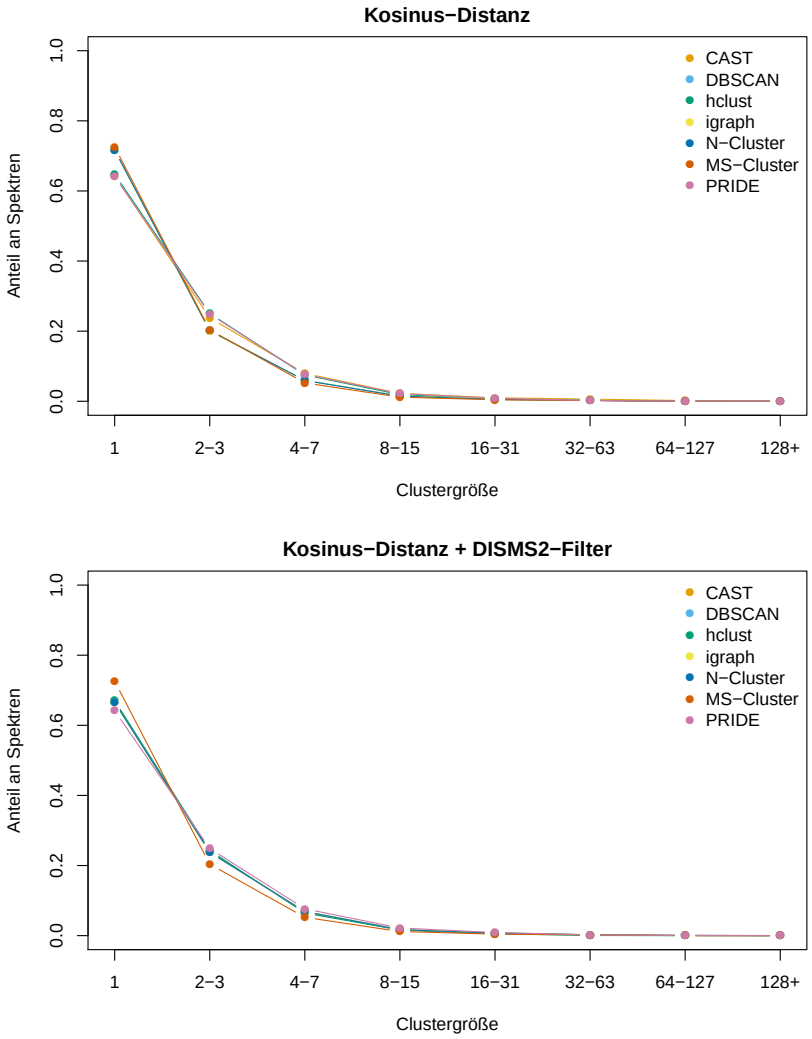

Abbildung B.15: Anteil Spektren gruppiert nach Clustergröße der Clusterlösungen einzelner Läufe mit bestem durchschnittlichen ARI je Clusteralgorithmus auf Basis der Kosinus-Distanz von Spektren (oben) oder zusätzlichem DISMS2-Filter (unten) (in Anlehnung an Rieder et al., 2017b, Abbildung S-5).

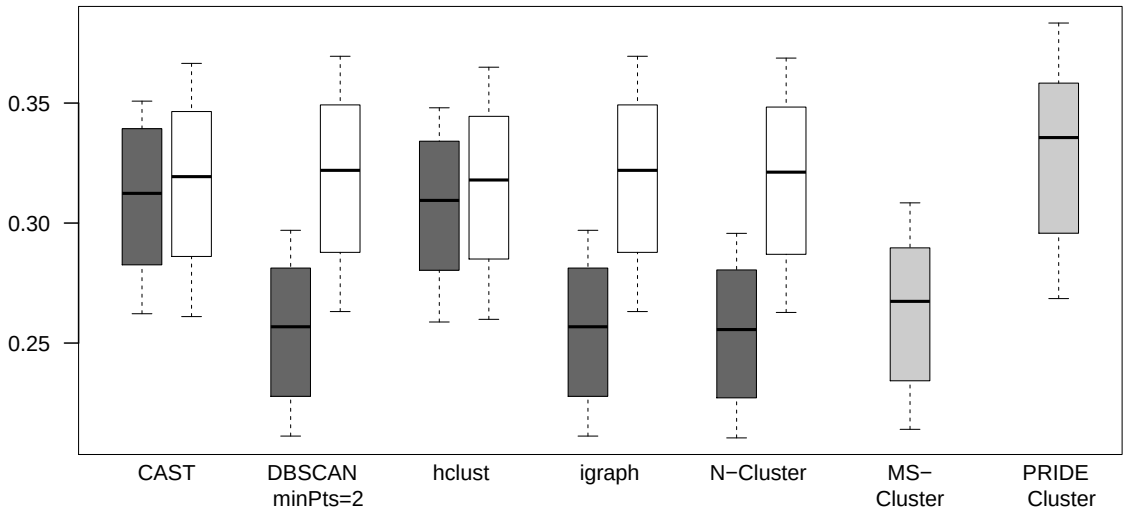

Abbildung B.16: Boxplots des mehrelementigen Clusteranteils für einzelne Läufe. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

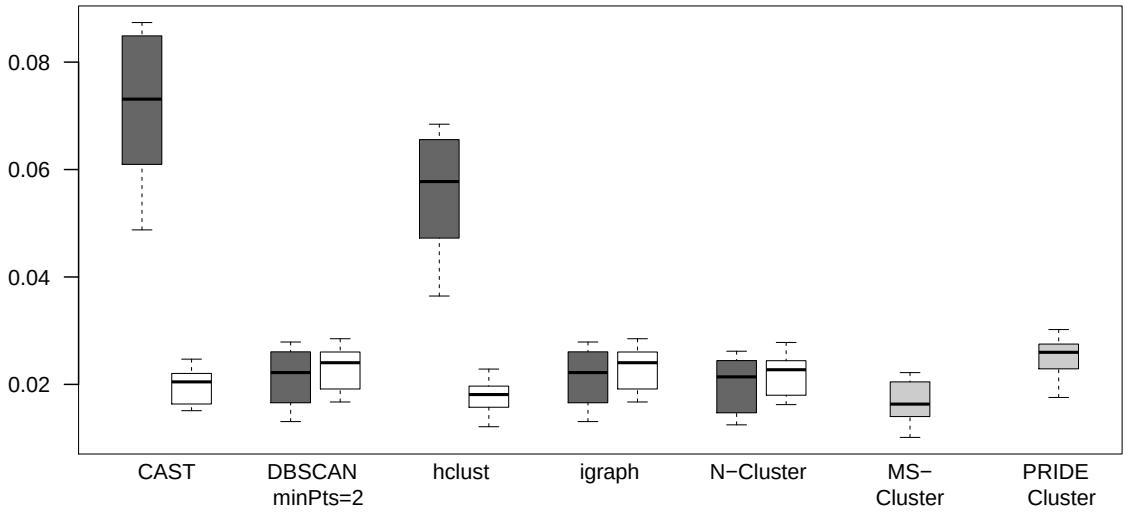

Abbildung B.17: Boxplots des Spektrenanteils ohne häufigste Annotation für einzelne Läufe. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

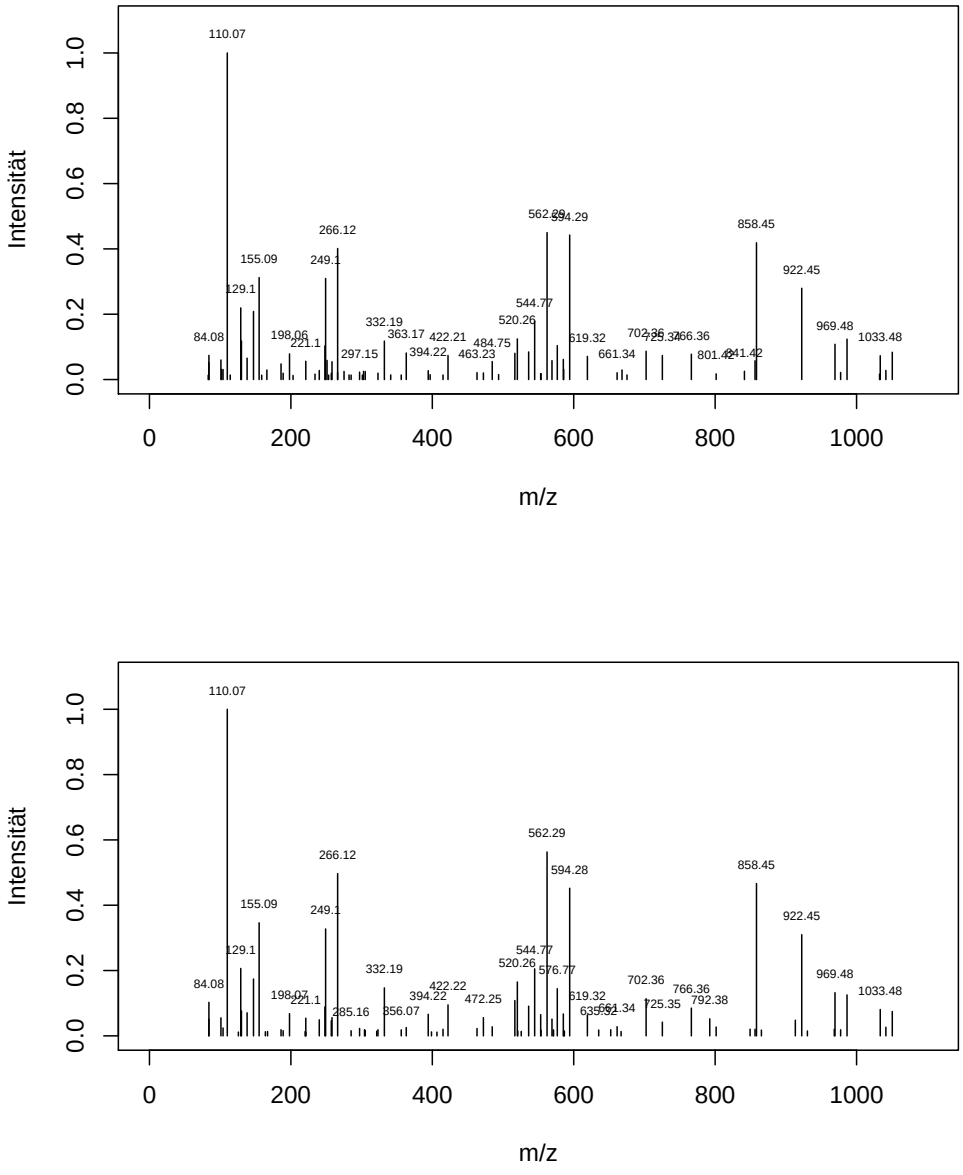

Abbildung B.18: Beispiel für Massenspektren des Clusters in Abbildung 5.11. Die Distanz des Spektrum mit fehlender Annotation (oben) und dem nächsten Spektrum mit Annotation HQGVmVGMGQK (unten) ist 0.012. (Rieder et al., 2017b, Abbildung S-6).

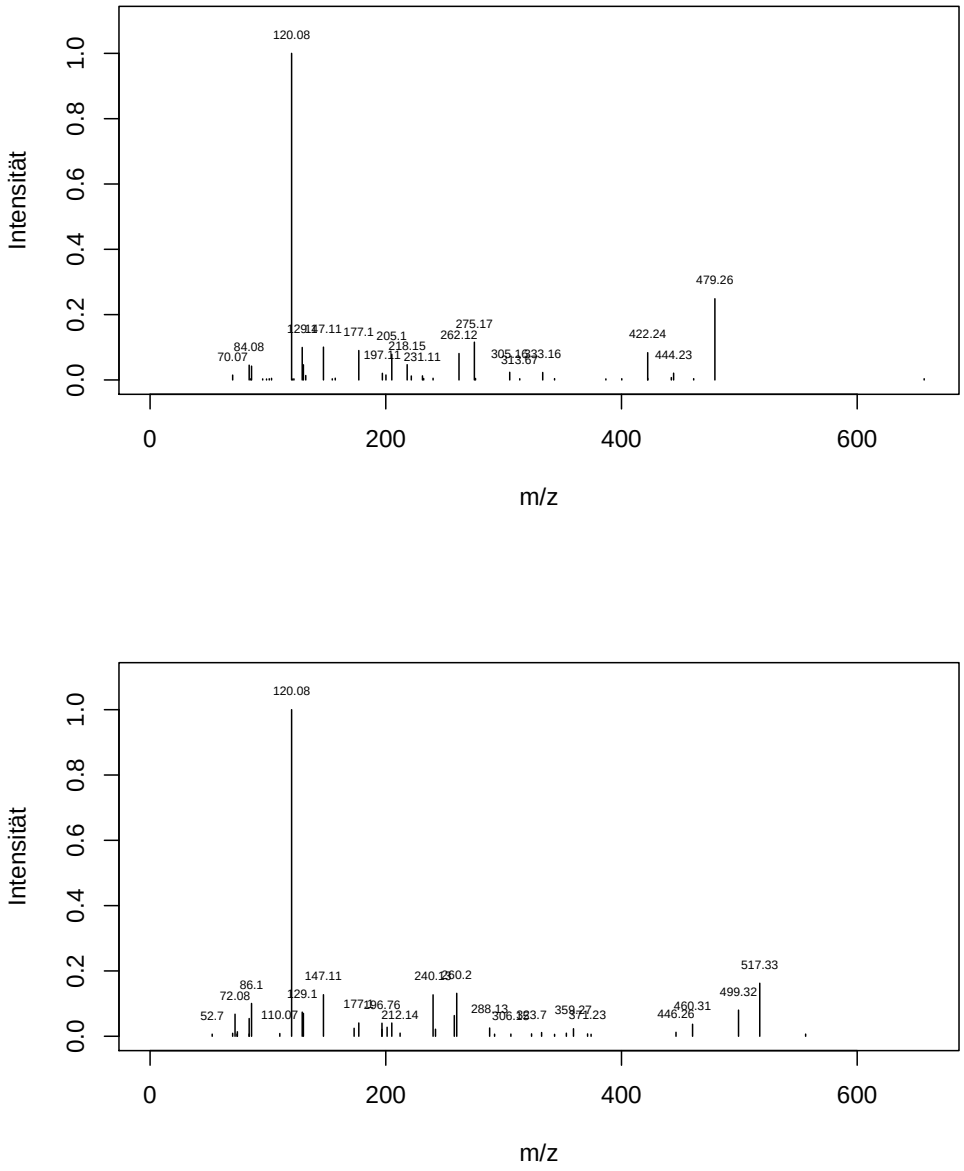

Abbildung B.19: Beispiel für Massenspektren des Clusters in Abbildung 5.12. Bei beiden Spektren, dem Spektrum mit Annotation FGFGAK (oben) und dem Spektrum mit Annotation FGTVLK (unten), liegt der höchste Peak bei 120 m/z. (Rieder et al., 2017b, Abbildung S-7).

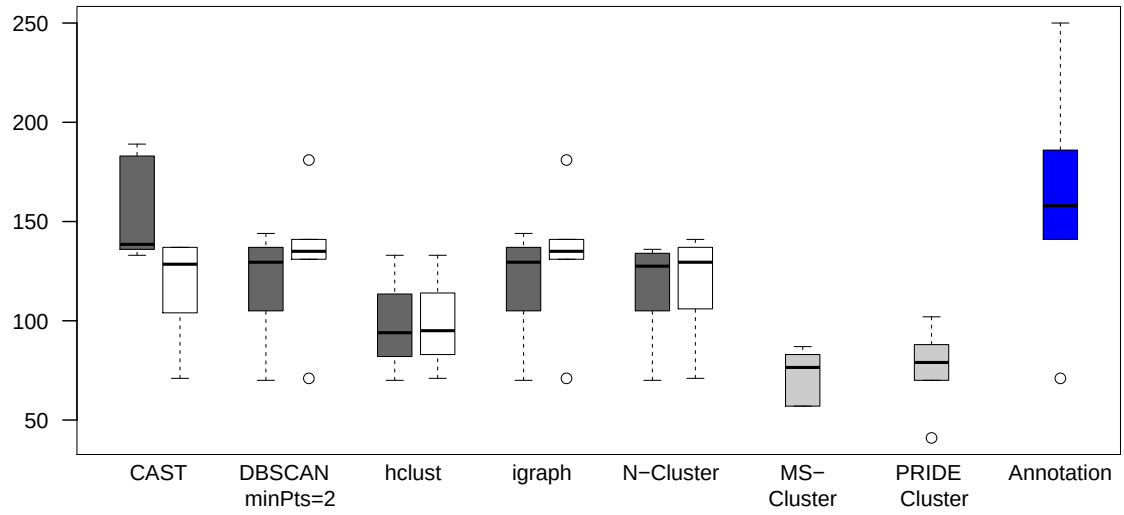

Abbildung B.20: Boxplots der maximalen Clustergröße von ausgewählten Clusterlösungen und Annotationsclusterlösungen (blau) mehrerer Läufe. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

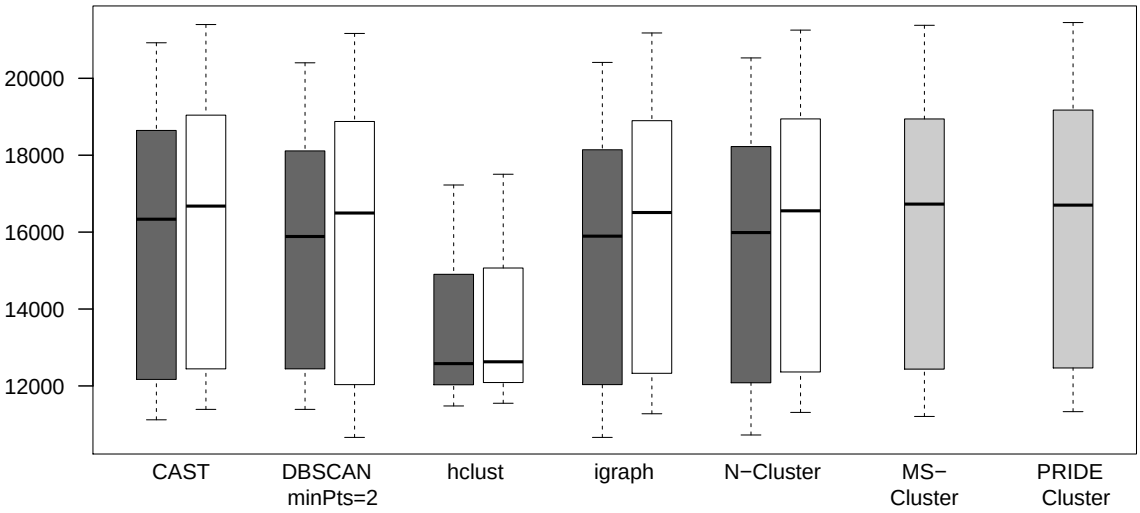

Abbildung B.21: Boxplots der Anzahl mehrelementiger Cluster mit mindestens einer Annotation von ausgewählten Clusterlösungen mehrerer Läufe. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

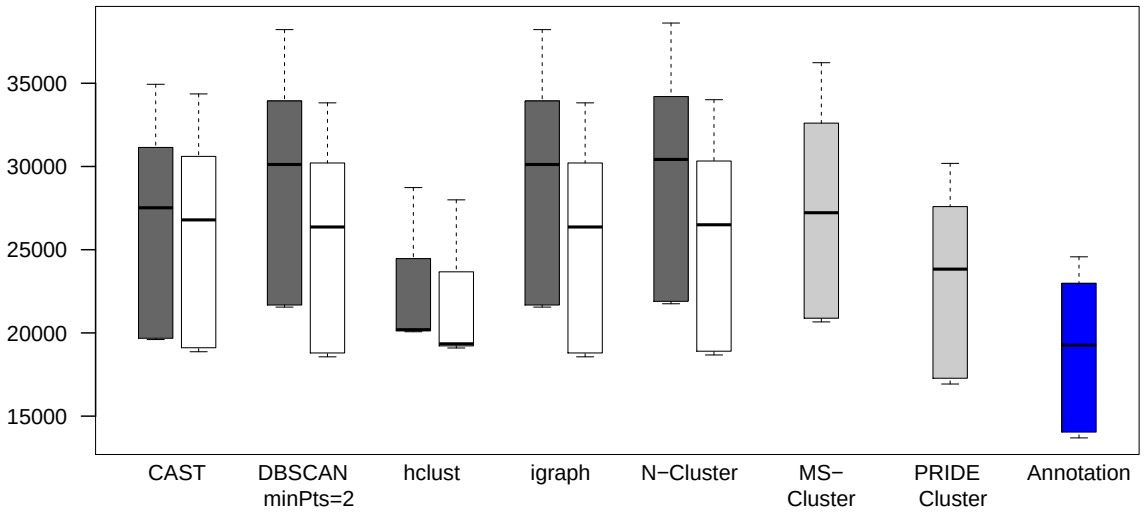

Abbildung B.22: Boxplots der Clusteranzahl von ausgewählten Clusterlösungen und Annotationsclusterlösungen (blau) mehrerer Läufe. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

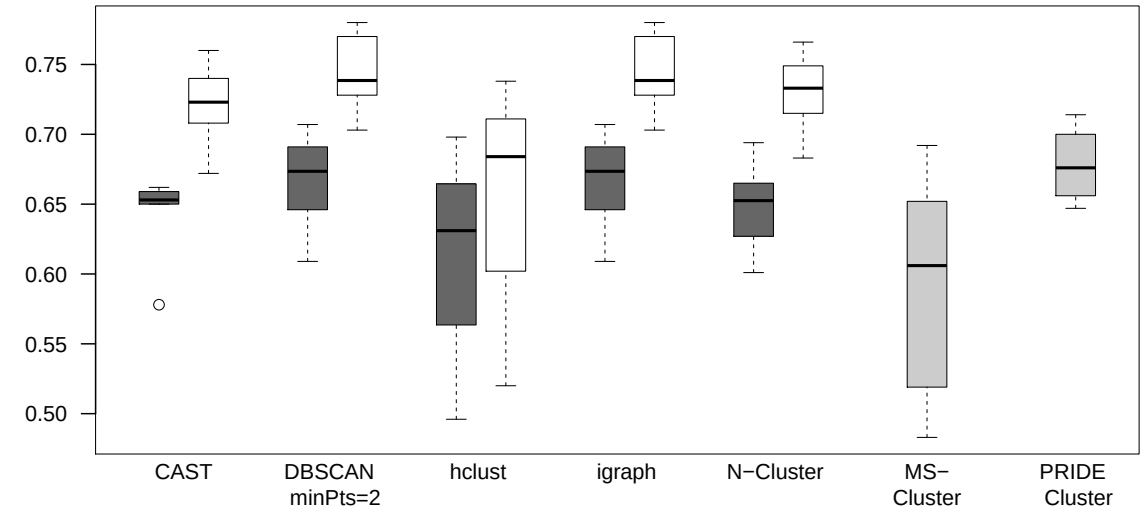

Abbildung B.23: Boxplots des adjustierten Rand-Indexes zwischen der Peptidannotation und ausgewählten Clusterlösungen für mehrere Läufe. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

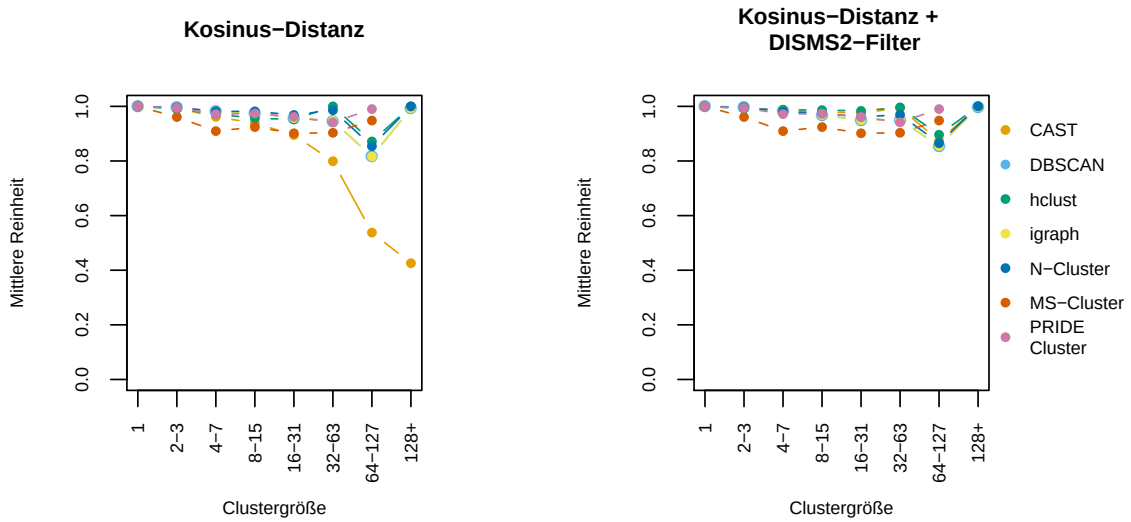

Abbildung B.24: Reinheit gruppiert nach Clustergröße der Clusterlösungen mehrerer Läufe mit bestem durchschnittlichen ARI je Clusteralgorithmus basierend auf Kosinus-Distanzen der Spektren (links) und zusätzlich mit dem DISMS2-Filter (rechts) (in Anlehnung an Rieder et al., 2017b, Abbildung S-9).

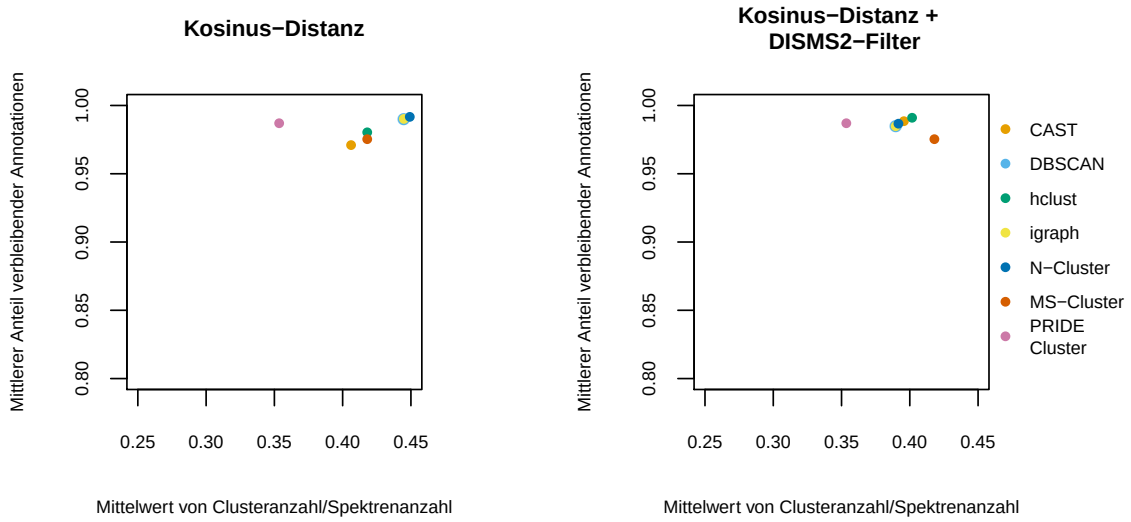

Abbildung B.25: Mittlerer Anteil verbleibender Annotationen in Abhängigkeit des Mittelwerts der Clusteranzahl in Relation zur Spektrenanzahl für mehrere Läufe (in Anlehnung an Rieder et al., 2017b, Abbildung S-10).

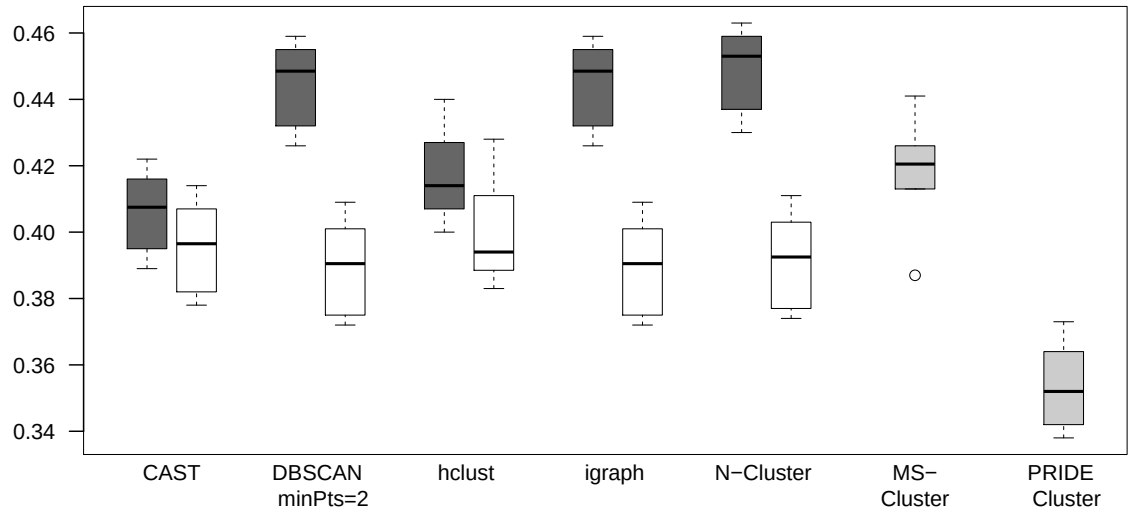

Abbildung B.26: Boxplots der Clusteranzahl in Relation zur Spektrenanzahl von ausgewählten Clusterlösungen für mehrere Läufe. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

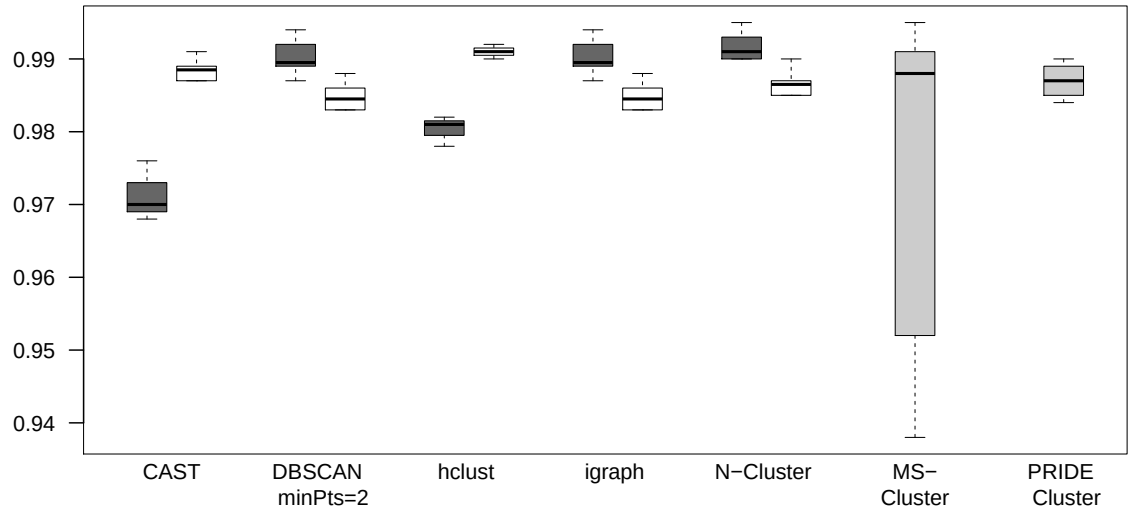

Abbildung B.27: Boxplots des Anteils verbleibender Annotationen von ausgewählten Clusterlösungen für mehrere Läufe. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

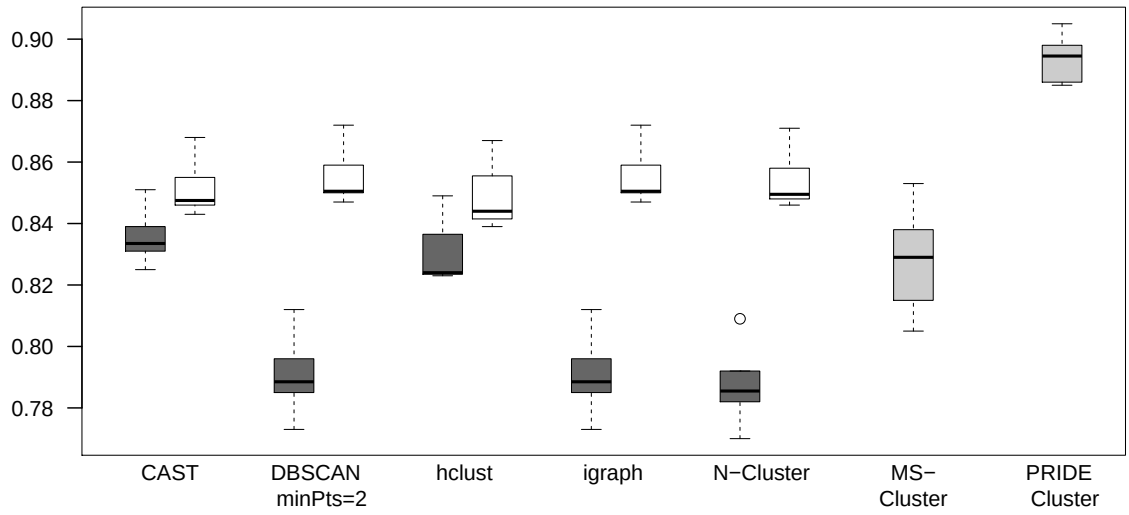

Abbildung B.28: Boxplots des mehrelementigen Clusteranteils von ausgewählten Clusterlösungen für mehrere Läufe. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

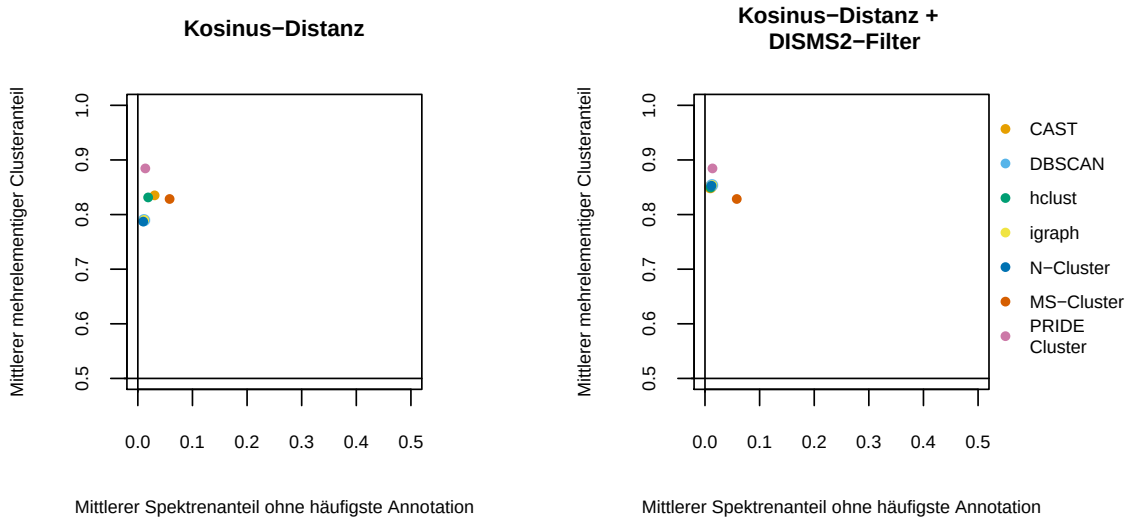

Abbildung B.29: Mittlerer mehrelementiger Clusteranteil in Abhängigkeit des mittleren Spektrenanteils ohne häufigste Annotation von Clusterlösungen mehrerer Läufe basierend auf Kosinus-Distanzen der Spektren (links) und zusätzlich mit dem DISMS2-Filter (rechts) (in Anlehnung an Rieder et al., 2017b, Abbildung S-11).

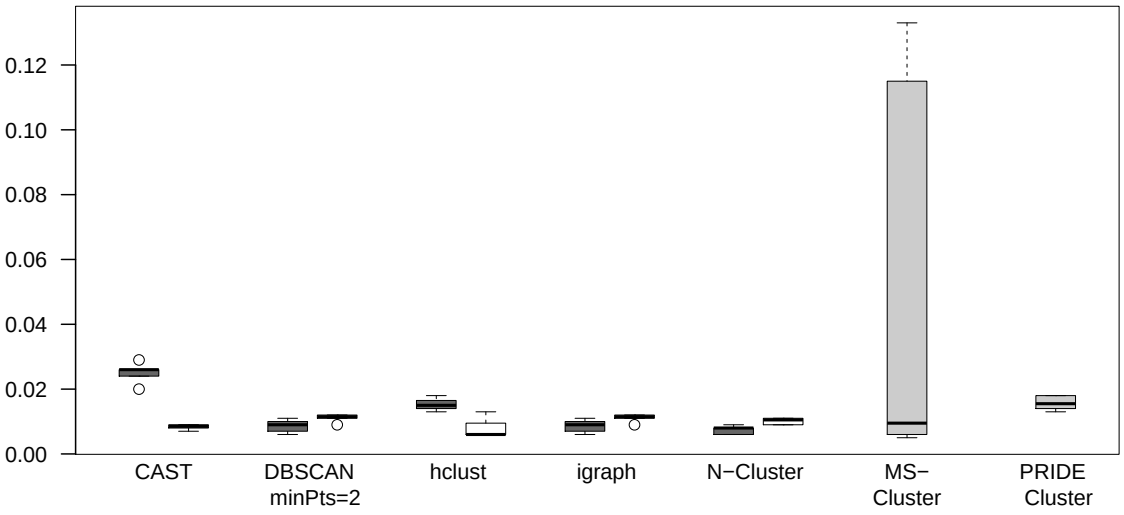

Abbildung B.30: Boxplots des Spektrenanteils ohne häufigste Annotation von ausgewählten Clusterlösungen für mehrere Läufe. Die Kosinus-Distanz wurde ohne (dunkelgrau) und mit (weiß) DISMS2-Filter verwendet. Die Vorverarbeitung und Distanzberechnung ist Teil des Algorithmus bei MS-Cluster und PRIDE Cluster (hellgrau).

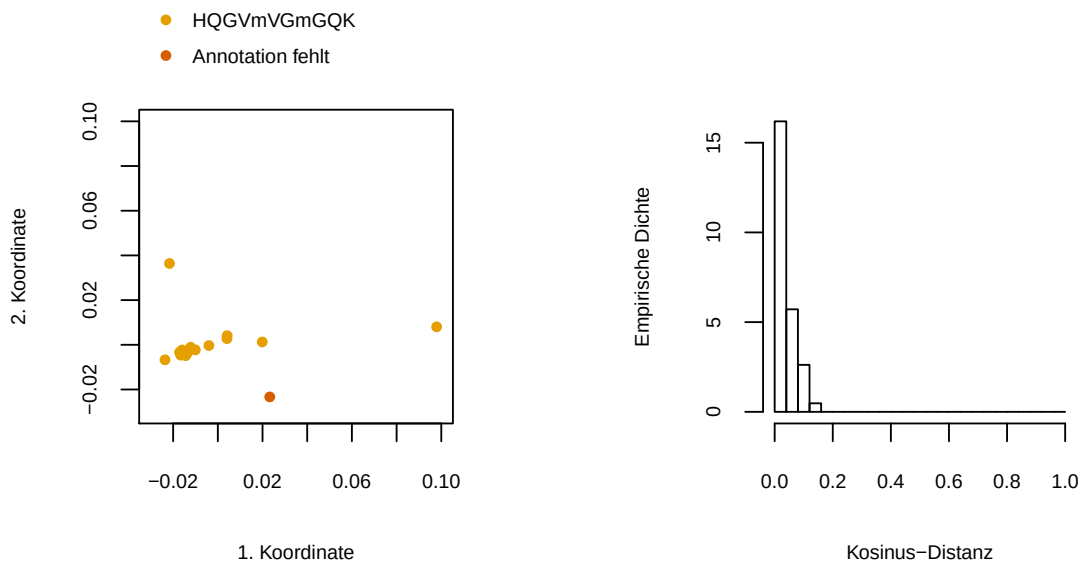

Abbildung B.31: Grafische Darstellung eines Clusters mit 15 Spektren der Clusterbildung von C2 mit PRIDE Cluster ($threshold\_end = 0.8$), visualisiert durch MDS (links) und durch ein Histogramm (rechts) aller paarweisen Kosinus-Distanzen der 15 Spektren.

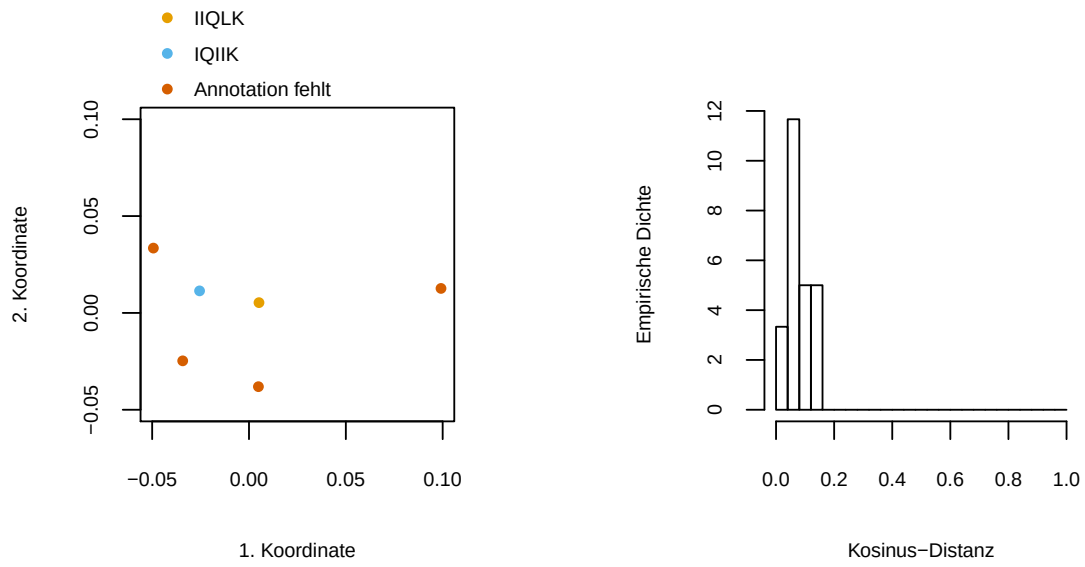

Abbildung B.32: Grafische Darstellung eines Clusters mit 6 Spektren der Clusterbildung von C2 mit DBSCAN ($\epsilon = 0.2, minPts = 2$), visualisiert durch MDS (links) und durch ein Histogramm (rechts) aller paarweisen Kosinus-Distanzen der 6 Spektren.

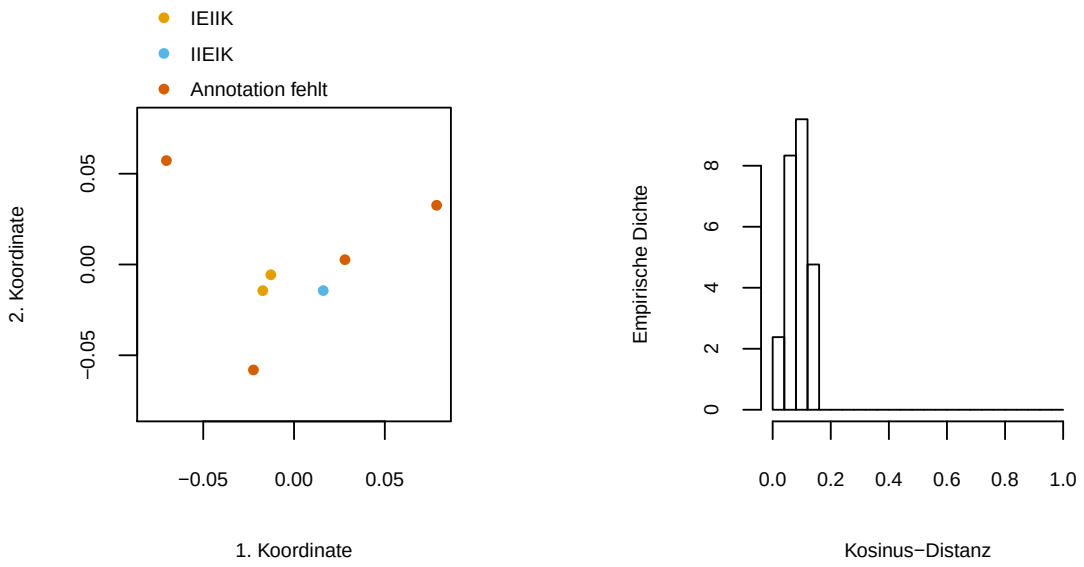

Abbildung B.33: Grafische Darstellung eines Clusters mit 7 Spektren der Clusterbildung von C2 mit DBSCAN ($\epsilon = 0.2, minPts = 2$), visualisiert durch MDS (links) und durch ein Histogramm (rechts) aller paarweisen Kosinus-Distanzen der 7 Spektren.