UNIVERSITÄT DORTMUND

Konzeption und Realisierung eines Algorithmus für die *de novo*-Proteinidentifikation

Wolfgang Paul

Algorithm Engineering Report
TR06-2-004
Juli 2006
ISSN 1864-4503

Universität Dortmund
Fachbereich Informatik
Algorithm Engineering (LS 11)
44221 Dortmund / Germany
http://ls11-www.cs.uni-dortmund.de/

Diplomarbeit

Konzeption und Realisierung eines Algorithmus für die *de novo*-Proteinidentifikation

Universität Dortmund
Fachbereich Informatik

vorgelegt von
Wolfgang Paul

03. Mai 2006

Erstgutachter: Prof. Dr. Günter Rudolph
Zweitgutachterin: Prof. Dr. Petra Mutzel

Universität Dortmund
Fachbereich Informatik
Lehrstuhl für Algorithm Engineering (LS11)
Otto-Hahn-Str. 14
44227 Dortmund

Inhaltsverzeichnis

Vorwort

An dieser Stelle möchte ich mich bei meinen Betreuern vom Lehrstuhl 11, Frau Prof. Dr. Petra Mutzel, Herrn Prof. Dr. Günter Rudolph und Herrn Dr. Udo Feldkamp, bedanken. Ihre prompten Rückmeldungen auf meine Fragen und intensive Betreuung trugen maßgeblich zum Gelingen dieser Diplomarbeit bei.

Herrn Prof. Dr. Helmut E. Meyer vom Medizinischen Proteom-Center (MPC) an der Ruhr-Universität Bochum möchte ich für den Freiraum, der mir für die Bearbeitung gelassen wurde, danken.

Mein besonderer Dank gilt Kai A. Reidegeld vom MPC für die Überlassung des interessanten Themas und für seine Unterstützung während der gesamten Entstehungszeit dieser Diplomarbeit. Erst durch unsere vielen konstruktiven Diskussionen und die daraus entstandenen Ideen und Lösungsansätze gelang es mir die zu lösenden Problemstellungen erfolgreich bearbeiten zu können. Durch das von ihm gezeigte Interesse an der Diplomarbeit gelang es mir, die Motivation während der gesamten Bearbeitungszeit auf hohem Niveau zu halten.

Bei Cornelia Joppich vom MPC möchte ich mich herzlich für die enzymatische Aufbereitung und massenspektrometrische Analyse der Testdatensätze, die ich zur Evaluierung meiner Arbeit benötigte, bedanken.

Des Weiteren möchte ich auch Dr. Christian Stephan und den anderen Kollegen am MPC, insbesondere der Arbeitsgruppe Bioinformatik, für die produktive und angenehme Zusammenarbeit danken.

Außerdem danke ich meiner Frau, meinen Eltern und meiner Familie, die mich während meiner gesamten Ausbildung unterstützten und immer für mich da waren. Nicht zuletzt ihnen habe ich es zu verdanken, dass ich mein Studium erfolgreich abschließen konnte.

Kapitel 1

Einleitung

1.1 Motivation

Mitte der 80er Jahre gingen viele Biologen davon aus, dass sie durch die Bestimmung sämtlicher Erbinformationen eines Lebewesens dazu in die Lage versetzt würden, die in diesem Lebewesen ablaufenden biochemische Prozesse zu verstehen. Aus dieser Überzeugung heraus startete das US-amerikanische Energieministerium 1986 das Human Genom Project (HGP), ein drei Milliarden Dollar Projekt, welches es sich zur Aufgabe gemacht hatte das menschliche Genom zu sequenzieren. Das Projekt „beendete“ seine ursprüngliche Aufgabe im April 2003 [3, 4, 5, 6, 7, 8][1], nachdem im Februar 2001 bereits erste Zwischenergebnisse veröffentlicht worden waren [1]. Zum Zeitpunkt des Abschlusses des ursprünglichen Projektes hatte man 99% des menschlichen Genoms, welches aus mehreren Proben stammte, sequenziert und schickte sich an dies für die Erbinformationen weiterer Lebewesen zu tun. Neben den USA beteiligten sich noch Wissenschaftler aus vielen anderen Industrienationen, darunter China, Frankreich, Großbritannien, Japan und Deutschland daran.

Aber bereits gegen Ende der 80er Jahre war immer deutlicher geworden, dass trotz enormer Fortschritte auf dem Gebiet der Molekularbiologie und trotz des Einsatz erprobter Methoden aus der Informatik, welche die Gewinnung, Verwaltung und Analyse der anfallenden großen Datenmengen überhaupt erst ermöglichten, eine Vielzahl alltäglicher biologischer Vorgänge auf Grund ihrer Komplexität noch immer nicht vollständig erklärt werden konnten. Die durch das Human Genome Project erzielten Fortschritte auf dem Gebiet der Genomforschung, führten zu der Erkenntnis, dass eines der berühmtesten Dogmen der Biologie, die Annahme, dass ein Eins-zu-eins-Verhältnis zwischen Genen, Proteinen und deren Funktion besteht, nicht länger haltbar war.

Die von vielen Biologen gehegte Hoffnung, durch die Sequenzierung der Erbinformationen ganzer Organismen umfassende Erkenntnisse über die in lebenden Zellen auf molekularer Ebene stattfindenden Prozess zu gewinnen, wurde enttäuscht. Es stellte sich vielmehr heraus, dass um diese Prozesse wirklich verstehen zu können, Wissen über Proteine, ihre Funktion und Lokalisation berücksichtigt werden musste. Richard Strohman formulierte diese Erkenntnis so:

> *Sequence information in DNA, by itself, contains insufficient information for determining how gene products (proteins) interact to produce a mechanism of any kind. The reason is that multicomponent complexes constructed from many proteins are themselves machines with rules of their own, rules not written in DNA.* [2]

Es sind also die aus der Erbinformation eines Lebewesens abgeleiteten Proteine und Proteinkomplexe, die für praktisch jeden der Prozesse, die in den Zellen eines Lebewesens stattfinden, verantwortlich sind. Da die Funktion einzelner Proteine und deren Rolle in der Interaktion mit anderen Proteinen aber nicht alleine aus der Kenntnis der Erbinformation eines Lebewesens abgeleitet werden kann, müssen diese Biomoleküle

[1] Die Frage, ob und wann das HGP seine eigentliche Arbeit wirklich beendete, ist schwierig und sehr kontrovers. Aufgrund ständiger technischer Weiterentwicklungen auf dem Gebiet der DNS-Analyse wurden die während des Projekts erzeugten Datenbestände mehrfach überarbeitet und korrigiert. Die letzte überarbeitete Version der Ergebnisse des HGP stammt aus dem Jahr 2005 [9].

folglich direkt untersucht werden. Es müssen Erkenntnisse über die verschiedenen Proteinenarten, deren Modifikationen und Konzentration gewonnen und so das bereits aus der Analyse der Gene erhaltene Wissen komplettiert werden. Diesen Bereich der Molekularbiologie, der sich mit der Erforschung der Proteine eines Lebewesens beschäftigt, nennt sich Proteomanalyse oder Proteomik.

Ebenso wie die Genomanalyse, die Erforschung und Sequenzierung des Erbguts eines Lebewesens, ist auch die Proteomanalyse ohne den Einsatz von Computern und geeigneter Software undenkbar. Die Katalogisierung und Zusammenfassung erzeugter Datensätze zu Gen- oder Proteindatenbanken, die Suche auf solchen Datenbänken oder die Identifikation einzelner molekularer funktionaler Einheiten wäre ohne die Unterstützung durch die Bioinformatik nicht zu leisten.

1.2 Zielsetzung

Diese Diplomarbeit entstand in Kooperation mit dem Medizinischen Proteom-Center (MPC) an der Ruhr-Universität Bochum, welches eines der in Deutschland führenden Forschungsinstitute im Bereich der Proteomforschung ist. Am MPC werden im Rahmen der Identifikation von Proteinen in biologischen Systemen verschiedene Formen der Massenspektrometrie in Kombination mit multidimensionalen Trennmethoden eingesetzt. Die eigentliche Proteinidentifikation geschieht über Algorithmen zur automatischen Suche auf Proteindatenbanken. Zu denen am MPC eingesetzten Algorithmen gehören Sequest [10, 11, 12], Mascot [13, 14], ProFound [15] und Phenyx [16, 17, 18].

Der datenbankbasierte Ansatz zur Proteinidentifikation unterliegt aber leider mehreren grundlegenden Problemen.

1. Nicht zu jedem Organismus gibt es Proteindatenbanken.
2. Die Größe der einzelnen Proteindatenbanken wächst seit Beginn der automatisierten Proteomanalyse zu Anfang der 90er Jahre exponentiell. Dies bedingt auch ein exponentielles Wachstum der Suchzeit auf diesen Datenbanken.
3. Proteindatenbanken enthalten zuweilen fehlerhafte Einträge wodurch es zu falsch positiven Proteinidentifikationen kommt.
4. Datenbanken decken im Allgemeinen nicht sämtliche zu einem Organismus gehörigen Proteine ab.
5. Mit zunehmender Größe der verwendeten Proteindatenbanken nimmt auch die Wahrscheinlichkeit einer falsch positiven Identifikation zu.

Daher soll ein Algorithmus für die so genannte *de novo*-Proteinidentifikation entwickelt werden, der die Limitationen der automatischen Proteinidentifikation via Datenbankabgleich überwindet. Der zu entwickelnde Proteinidentifikationsalgorithmus soll daher nicht auf bestehende Proteindatenbanken angewiesen sein. Vielmehr soll er dazu in der Lage sein, das zu identifizierende Protein auf Grund von experimentell ermittelten Daten aus der Massenspektrometrie zu bestimmen.

Die im Rahmen dieser Diplomarbeit erarbeiteten Ergebnisse und die daraus entstandene Software sollen in die Weiterentwicklung der am MPC entstehenden Software Peakardt [19, 20, 21] einfließen.

1.3 Gliederung

Nachdem im ersten Kapitel eine kurze Einleitung und Motivation der vorliegenden Aufgabenstellung erfolgte, widmet sich Kapitel Zwei der Einführung sämtlicher biologischer und molekularbiologischer Grundlagen, die für das Verständnis der vorliegenden Arbeit notwendig sind. Kapitel Drei stellt das grundlegende Vorgehen, wie es typischerweise bei der Analyse eines Proteins angewendet wird, exemplarisch vor. Da die Massenspektrometrie die wichtigste Technik der Datenakquisition in der Proteinanalytik darstellt, widmet sich ein großer Teil von Kapitel Drei ihren Grundlagen. Kapitel Vier gibt einen Überblick über die wichtigsten Aufgabengebiete der Bioinformatik innerhalb der Proteinanalytik und stellt das momentan wichtigste Anwendungsgebiet, die Interpretation von massenspektrometrischen Daten auf Basis von Sequenzdatenbanken inklusive der damit verbundenen Probleme, genauer dar. Durch Definition der

Anforderungen an einen *de novo*-Algorithmus für die Proteinidentifikation in Kapitel Fünf, richtet sich der Fokus dieser Arbeit dann wieder auf die eigentliche Aufgabenstellung. Kapitel Sechs beschreibt die zu Kapitel Fünf gehörige Implementierung des Algorithmus. Anschließend erfolgt in Kapitel Sieben die Evaluation des implementierten Algorithmus. Kapitel Acht fasst zum einen die Ergebnisse dieser Arbeit noch einmal kurz zusammen und gibt zum anderen einen Ausblick auf noch ausstehende Fragestellungen.

Kapitel 2

Biologische Grundlagen

Da die vorliegende Aufgabenstellung aus dem Bereich der Bioinformatik stammt, müssen zunächst einige Begrifflichkeiten aus der Biologie, insbesondere der Molekularbiologie, eingeführt werden.

2.1 Das Genom und die DNS

Unter dem Begriff des Genoms versteht man die Gesamtheit sämtlicher genetischer Informationen eines Organismus. Diese Erbinformationen sind in jeder Zelle eines Lebewesens gespeichert. Im übertragenen Sinne stellt das Genom den Bauplan eines Lebewesens dar.

Dieser Bauplan wird durch DNS-Moleküle kodiert. Aus Sicht der Chemie stellt sich ein solches Desoxyribonukleinsäure-Molekül als eine Doppelhelix (siehe Abbildung 2.1) zweier einzelner Stränge dar. Die beiden Einzelstränge bestehen aus Ketten von so genannten Nukleotiden. Nukleotide sind Untereinheiten der DNS und bestehen aus je einem Zuckermolekül, einer so genannten Phosphatgruppe und einer der vier Basen Adenin, Cytosin, Guanin und Thymin (Abbildung 2.2).

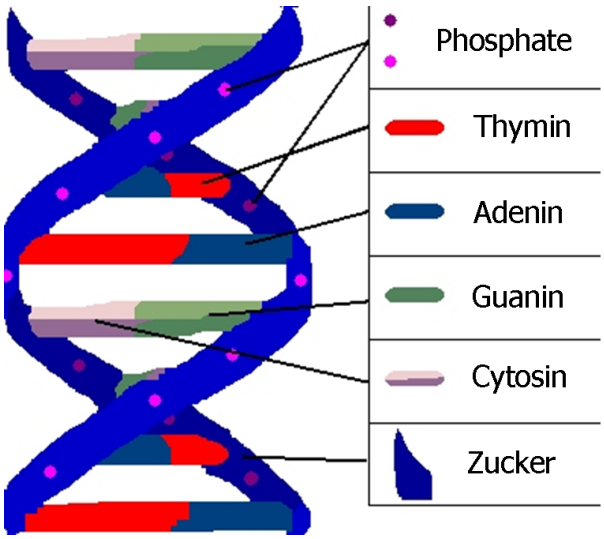

Abbildung 2.1: Graphisches Darstellung der Doppelhelixstruktur eines DNS-Moleküls. Quelle: [22]

Im Kontext der Genomforschung und der Bioinformatik, wird die Struktur solcher Nukleotidketten aber vereinfacht als Zeichenketten über dem Alphabet $\Sigma = \{A, C, G, T\}$ dargestellt. Die Zeichen des Alphabets Σ entsprechen dabei den Basen Adenin, Cytosin, Guanin und Thymin.

Die Nukleotide zweier solcher Stränge stehen sich paarweise gegenüber und sind über ihre Basen miteinander verbunden. Bei der Bindung der Basen sind nur Paarungen zwischen Adenin und Thymin bzw. Guanin und Cytosin möglich. Dies bedingt, dass die beiden Stränge bezüglich ihres Informationsgehalts

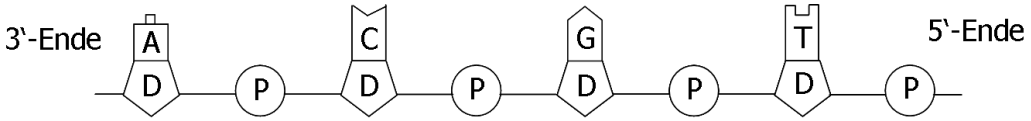

Abbildung 2.2: Beispiel für einen Nukleotidstrang. (P: Phosphatgruppen, D: Zuckermoleküle, A, C, G, T: Basen)

komplementär zu einander sind.

Ziel der Genomik ist es mittels der Analyse genetischer Informationen das Genom einzelner Lebewesen zu identifizieren, sowie die Funktionen der einzelnen zu diesem Genom gehörigen Gene zu bestimmen. Die Sequenzierung des Genoms eines Lebewesens geschieht über die Identifikation codierender Abschnitte auf den Nukleotidsträngen einzelner DNS-Moleküle. Genau diese Abschnitte sind es, die die Gene eines Lebewesens beschreiben. Die Funktion eines genkodierenden Abschnitts auf der DNS lässt sich aber nicht direkt aus der Kenntnis der zugehörigen Basensequenz ableiten. Um diese zu bestimmen, muss man sich die aus diesem Gen abgeleiteten Produkte und ihre Aufgabe im Organismus anschauen.

2.2 Von der DNS zum Protein: Die Proteinsynthese

Als Proteine bezeichnet man lange Ketten von Aminosäuren, die über so genannte Peptidbindungen miteinander verbunden sind. Die Information zu ihrem Zusammenbau ist in der Abfolge der DNS-Basen der Gene gespeichert. Die beiden Enden der Aminosäurekette bezeichnet man als N- bzw. C-Terminus des Proteins. Die Leserichtung der zugehörigen Aminosäuresequenz entspricht der Abfolge der Aminosäuren vom N- zum C-Terminus. Wie bereits erwähnt gibt es vier verschiedene DNS-Basen. Dabei codieren jeweils drei zusammenhängende Basen, ein so genanntes Codon, die Information für eine Aminosäure. Da es insgesamt nur 20 verschiedene Aminosäuren[1] in der Natur gibt (siehe Tabelle 2.1), ist der Genetische Code zur Codierung der Aminosäuren redundant (siehe Tabelle 2.2). Kurze Aminosäureketten aus zwei bis neun Aminosäuren werden als Oligopeptide bezeichnet, längere Ketten von Aminosäuren mit zehn bis etwa 100 Aminosäuren als Polypeptide[2]. Aminosäureketten, die noch länger sind, nennt man Proteine [22].

Das zentrale Dogma der molekularen Biologie (siehe Abbildung 2.3) besagt, dass die Merkmale eines Organismus im Wesentlichen durch seine Proteine festgelegt werden. Diese bestimmen direkt oder indirekt seine Eigenschaften. Praktisch alle in den Zellen eines Lebewesens ablaufenden Prozesse werden direkt oder indirekt von Proteinen ausgeführt und gesteuert (siehe Tabelle 2.3).

Die Anweisungen zur Herstellung dieser Proteine sind auf der DNS in verschlüsselter Form gespeichert. Das Ablesen dieser Information und die anschließende Herstellung von Proteinen, nennt man Proteinsynthese. Aufgabe der Proteinsynthese ist es, die auf der DNS, in der Form von Genen, gespeicherten genetischen Informationen zu exprimieren [23].

In Abbildung 2.3 ist das zentrale Dogma zusammenfassend dargestellt. Die DNS besitzt die Fähigkeit sich mit Hilfe einer Vielzahl unterschiedlicher Enzyme selbst replizieren zu können, dieses ist notwendig um sicherzustellen, dass Zellen sich erfolgreich teilen und tote und zerstörte Zellen ersetzen können. Des Weiteren besitzen lebende Zellen die Möglichkeit Proteine zu exprimieren, dazu dient der Mechanismus der Proteinsynthese, welcher aus zwei Phasen, der Transkription und der Translation, besteht.

In der ersten Phase, der Phase der Transkription (siehe Abbildung 2.4), wird der so genannte codogene Strang eines DNS-Moleküls abgelesen und als mRNS-Molekül (Messenger-Ribonukleinsäure) nachgebildet. Dies bedeutet, dass ein spezifischer Gen-Abschnitt eines DNS-Stranges gelesen wird und die gele-

[1] Der Begriff Aminosäure wird meistens als Synonym für die proteinogenen Aminosäuren verwendet, die für die meisten bekannten Organismen als grundlegende Bausteine ihrer Proteine dienen. Insgesamt sind bisher 23 proteinogene Aminosäuren bekannt. Das Spektrum der Klasse der Aminosäuren geht aber weit über diese hinaus. So sind bisher 250 nicht-proteinogene Aminosäuren bekannt. Neben den hier aufgezählten 20 proteinogenen Aminosäuren, die im menschlichen sowie im Organismus vieler anderer Lebewesen für die Erzeugung essentieller Proteine verantwortlich sind, gibt es noch drei weitere, für den Menschen nicht-proteinogene Aminosäuren, die für den Stoffwechsel von einigen Bakterien essentiell sind. Die 21. proteinogene Aminosäure, heißt Selenocystein und wurde 1986 entdeckt, die 22. wurde 2002 in dem Archaebakterium Methanosarcina barkeri entdeckt und trägt den Namen Pyrrolysin. Die 23. proteinogene Aminosäure heißt Selenomethionin.

[2] oder abgekürzt als Peptide

Aminosäure	**Dreibuchstaben-Code**	**Einbuchstaben-Code**
Alanin	Ala	A
Asparaginsäure	Asp	D
Histidin	His	H
Methionin	Met	M
Serin	Ser	S
Arginin	Arg	R
Glutamin	Gln	Q
Isoleucin	Ile	I
Phenylalanin	Phe	F
Tryptophan	Trp	W
Cystein	Cys	C
Glutaminsäure	Glu	E
Leucin	Leu	L
Prolin	Pro	P
Tyrosin	Tyr	Y
Asparagin	Asn	N
Glycin	Gly	G
Lysin	Lys	K
Threonin	Thr	T
Valin	Val	V

Tabelle 2.1: Zusammenstellung sämtlicher proteinogener Aminosäuren und ihrer Drei- und Einbuchstaben-Codes

Ala	GCU, GCC, GCA, GCG	**Leu**	UUA, UUG, CUU, CUC, CUA, CUG
Arg	CGU, CGC, CGA, CGG, AGA, AGG	**Lys**	AAA, AAG
Asn	AAU, AAC	**Met**	AUG
Asp	GAU, GAC	**Phe**	UUU, UUC
Cys	UGU, UGC	**Pro**	CCU, CCC, CCA, CCG
Gln	CAA, CAG	**Ser**	UCU, UCC, UCA, UCG, AGU, AGC
Glu	GAA, GAG	**Thr**	ACU, ACC, ACA, ACG
Gly	GGU, GGC, GGA, GGG	**Trp**	UGG
His	CAU, CAC	**Tyr**	UAU, UAC
Ile	AUU, AUC, AUA	**Val**	GUU, GUC, GUA, GUG
Start	AUG, GUG	**Stopp**	UAG, UGA, UAA

Tabelle 2.2: Codon-Tabelle des genetischen Codes. Diese Tabelle zeigt die 20 proteinogenen Aminosäuren, die zur Ableitung von Proteinen verwendet werden, und die zugehörigen Codons, die diese Aminosäuren codieren. **Start** und **Stopp** dienen als Abkürzung für die Codierungen der Stopp- und Start-Codons einzelner Gene (siehe unten).

senen Basen als Vorlage zur Synthese eines neuen RNS-Stranges dienen. Ribonukleinsäure oder RNS , ist wie DNS ebenfalls eine Nukleinsäure, allerdings enthalten ihre Moleküle im Unterschied zur DNS einen anderen Typ Zuckermolekül (RNS enthält Ribose, während DNS den so genannten Zweifachzucker Desoxyribose enthält) und die Basen Adenin, Cytosin, Guanin und Uracil (abgekürzt U). Wird daher während der Transkription eine Adenin-Base ausgelesen, so wird diese durch eine Uracil-Base in der mRNS-Repräsentation des abzulesenden Gens substituiert. Ist die Transkription abgeschlossen, so wird das Transkript zu den Ribosomen der Zellen transportiert. Dies ist eine spezielle Zellorganelle, die zur Herstellung von Proteinen dient. In den Zellen höherer Lebewesen findet an dieser Stelle noch ein Zwischenschritt statt, der Spleißen genannt wird. Dabei werden Teile der abgeschriebenen Informationen aus der mRNS entfernt und die übrigen Teile zu einem neuen mRNS-Molekül zusammengefügt. Genbestandteile, deren mRNS-Entsprechungen nach der Transkription entfernt werden, nennt man Introns, die anderen Exons [24]. Für die so gewonnenen Exons gibt es verschiedene Kombinationsmöglichkeiten: So können Exons vorne oder hinten an ein mRNS-Molekül angehängt oder aber auch aus der Mitte einer Gensequenz enfernt werden. Dies wird als alternatives Spleißen bezeichnet.

Nach der Transkription erfolgt in der nächsten Phase die Translation der mRNS in ein Protein (siehe Abbildung 2.5). Dabei hilft eine weitere Form der RNS, die tRNS (Transfer-Ribonukleinsäure), welche die

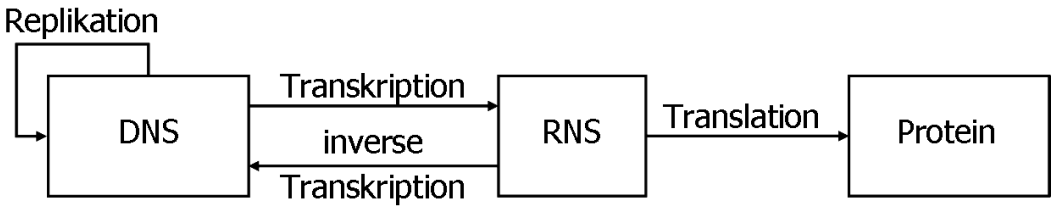

Abbildung 2.3: Zentrales Dogma der Molekularbiologie

Aminosäuren aus denen das neue Protein besteht zu den Ribosomen transportieren. Bei der Herstellung von Proteinen werden die in mRNS übersetzten Gen-Informationen von den Ribosomen ausgelesen. Da aus den abgelesenen Geninformationen nicht notwendiger Weise nur ein Protein abgeleitet werden kann, gibt es spezielle Start- und Stopp-Codons, die mit ausgelesen werden und die den Proteinherstellungsprozess steuern. Bei der eigentlichen Proteinsynthese gleiten die Ribosomen an der transkribierten mRNS entlang und lesen immer jeweils ein Codon aus. Dieses Codon benennt die nächste, an das bisher erzeugt Protein anzuhängende, Aminosäure. Damit dies gelingt, besitzen die tRNS-Moleküle spezielle Anti-Codons, die zu einer kleineren Anzahl von Codons, die alle die gleiche Aminosäure codieren, passen. Wurde ein Codon auf der mRNS ausgelesen, so fügt eines der an den Ribosomen vorhandenen tRNS-Moleküle eine passende Aminosäure an die letzte Stelle an. Die Übersetzung eines Proteins ist beendet, sobald ein Stopp-Codon gelesen wird. Das fertige Protein löst sich von der mRNS ab. Zu diesem Zeitpunkt, wie zu jedem anderen beliebigen Zeitpunkt in der Existenz eines Proteins, können so genannte post-translationale Modifikationen (oft als PTMs abgekürzt) an dem fertigen Protein vorgenommen werden, welche nicht in der DNS des ursprünglichen Gens kodiert waren. Ist der Gesamtprozess abgeschlossen, so nimmt das fertige Protein eine dreidimensionale Struktur an und begibt sich an seinen Einsatzort [24].

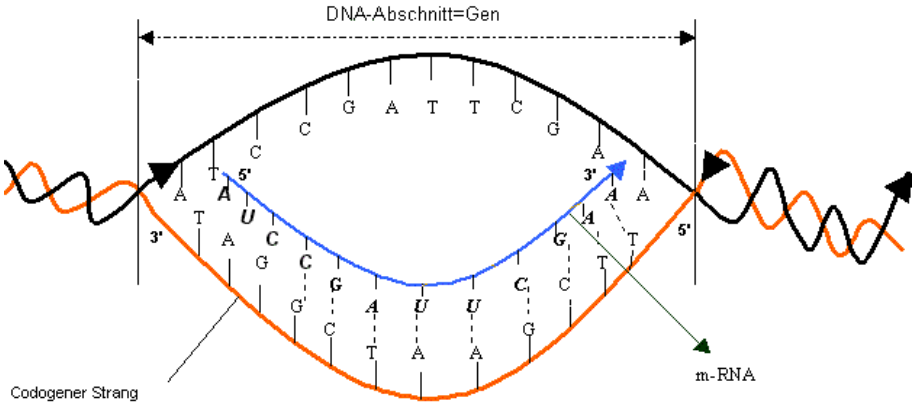

Abbildung 2.4: Schematische Darstellung der ersten Phase der Proteinsynthese. Quelle: http://www.scheffel.og.bw.schule.de (Stand vom 21.09.04)

Proteine erfüllen eine Vielzahl von Aufgaben (siehe Tabelle 2.3). Aus ihnen bestehen wichtige Gewebetypen, wie Sehnen, Fingernägel oder Muskeln oder Haare. Sie helfen als Verdauungsenzyme bei der Zerlegung von Nahrungsbestandteilen oder sorgen für die Kontraktion von Muskeln um Bewegung zu ermöglichen. Der größte Teil der heute bekannten Proteine agiert als Biokatalysatoren oder Enzyme. Diese ermöglichen jeweils ganz bestimmte biochemische Reaktionen, die alle zusammengenommen den Stoffwechsel eines Lebewesens ausmachen. Hochspezialisierte Proteinformen sind die Voraussetzung für fast alle Formen der Zellfunktion [25].

Neben der Primärstruktur (siehe Abbildung 2.6, links) eines Proteins, der spezifischen Abfolge der Aminosäuren aus denen es besteht, sind noch ihre Sekundär-, Tertiär- und Quartärstruktur von Bedeutung. Sekundär, Tertiär- und Quartärstruktur beschreiben die räumliche Anordnung von Proteinabschnitten, dem Protein als solchen und von Proteinkomplexen. Die Primärstruktur lässt nur wenige Rückschlüsse auf räumliche Gestalt eines Proteins zu. Abschnitte einer Aminosäurekette eines Proteins können sich zu Schrauben (Singular Helix) aufwinden (siehe Abbildung 2.6, zweites Bild links) oder in parallele Stränge einer Mehrfach-Schlaufe anordnen, die zusammen ein so genanntes Beta-Faltblatt bilden (siehe Abbildung 2.6, zweites Bild rechts). Solche Proteinsubstrukturen charakterisieren die Sekundärstruktur eines Prote-

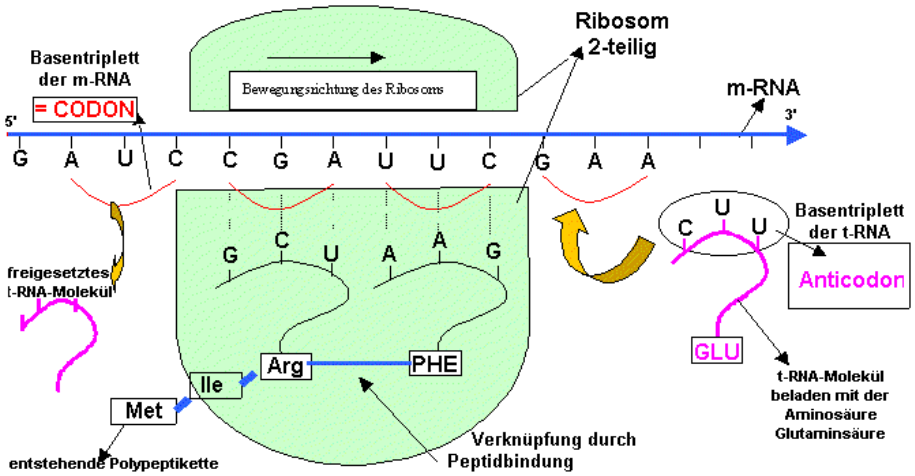

Abbildung 2.5: Schematische Darstellung der zweiten Phase der Proteinsynthese. Quelle: http://www.scheffel.og.bw.schule.de (Stand vom 21.09.04)

Proteityp	**Erklärung**
Strukturproteine	Erfüllen Stützfunktionen
Katalysatoren	Stimulieren bestimmte Stoffwechselreaktionen
Regulationsproteine	Steuern Stoffwechselprozesse
Transportproteine	Sauerstoff- oder Nährstoffransport
Kontraktile Proteine	Sorgen für die Bewegung in den Muskeln
Abwehrproteine	Teil des Immunsystems
Speicherproteine	Einlagerung von Nährstoffen und Vitaminen
Rezeptorproteine	Weiterleitung chemischer Reize

Tabelle 2.3: Zusammenstellung der wichtigsten Proteinfunktionen

ins. Die über Schlaufen verbundenen Sekundärstrukturen bilden schließlich die Struktur des kompletten Proteins, die so genannte Tertiärstruktur (siehe Abbildung 2.6, rechts). Oft hat man es mit zusammengesetzten Proteinen zu tun, die aus mehreren Untereinheiten bestehen. Die Struktur eines solchen Komplexes nennt man Quartärstruktur [25].

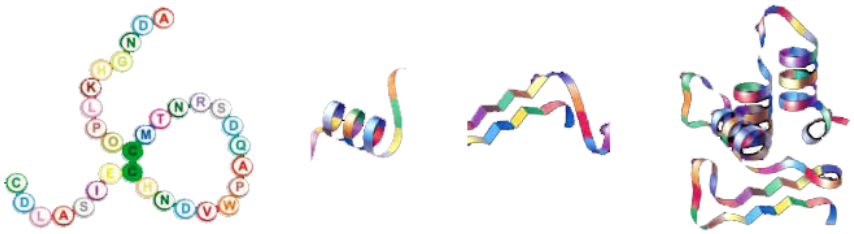

Abbildung 2.6: Die Abbildung ganz links stellt einen Teil der Primärstruktur des dargestellten Proteins dar. Als zweites von links folgt die Darstellung eines Proteinabschnittes, der die räumliche Struktur einer α-Helix besitzt. Die dritte Abbildung zeigt ebenfalls einen Teilabschnitt des dargestellten Proteins, dieser hat die Form eines so genannten β-Faltblatts. Die letzte Abbildung stellt die Tertiärstruktur des Gesamtproteins dar. Quelle: [25]

2.3 Das Proteom und die Proteomik

Das wohl überraschendste Ergebnis des Human Genome Projects war die Feststellung, dass das menschliche Genom weitaus weniger komplex ist, als bis dato angenommen. Ursprünglich war man von 80.000 bis 140.000 Genen ausgegangen und musste diese relativ hohe Zahl 2001 zunächst auf etwa 30.000 bis

40.000 [1] und 2004 ein weiteres Mal auf etwa 25.000 [26] senken. Damit haben Menschen nicht wesentlich viel mehr Gene als die Ackerschmalwand (*Arabidopsis thaliana*) — ein bescheidenes Unkraut — oder ein Fadenwurm (*Caenorhabditis elegans*).

Trotz dieser relativ geringen Menge an Genen fand man heraus, dass durch alternatives Spleißen und post-translationale Modifikationen bis zu einige hunderttausend verschiedene Proteine gleichzeitig in einer Zelle exprimiert sein können. Vorsichtige Schätzungen bzgl. der Anzahl der post-translationalen Modifikation an menschlichen Proteinen gehen davon aus, dass es pro Protein durchschnittlich 3 Modifikationen gibt [27]. Wenn man davon ausgeht, dass durchschnittlich etwa 10.000 verschiedene Gene pro Zellzustand exprimiert werden, kommt man schon alleine mit den post-translationalen Modifikationen auf etwa 30.000 verschiedene Proteine [28]. Solch eine Population von Proteinen, die alle zum selben Zeitpunkt und im selben Zellzustand expremiert wurden, nennt man Proteom.

Der Begriff des Proteoms geht auf den Australier Marc Wilkins zurück [29], der diesen Begriff während einer Konferenz in Italien prägte, um nicht ständig die Umschreibung „Alle Proteine, die von einem Genom, einer Zelle oder einem Gewebe exprimiert werden“, benutzen zu müssen. Diese Wortschöpfung und die von ihr abgeleitete Bezeichnung für die assoziierte wissenschaftliche Disziplin der Proteomik, fanden auf Grund ihrer lexikalischen Verwandtschaft zu dem bereits etablierten Begriffspaar Genome und Genomik schnell breiten Zuspruch.

Unter dem Begriff der Proteomanalyse oder auch Proteomik versteht man sämtliche Methoden zur qualitativen und quantitativen Analyse der zu einem bestimmten Zeitpunkt und unter exakt definierten Randbedingungen in einem Organismus, einer Zelle oder auch in einer Zellorganelle vorhandenen Proteine [30].

Der Begriff des Genoms wird häufig mit dem des Proteoms verglichen. Dieser Vergleich ist insofern irreführend, als dass das Genom die Gesamtheit der Gene, d. h. die Erbinformation einer Zelle bzw. eines Organismus darstellt und als solches statisch ist. Das Proteom repräsentiert hingegen einen bestimmten Zellzustand, der durch eine charakteristische Mischung von Proteinen zu einem bestimmten Zeitpunkt gekennzeichnet ist. Diese Zusammensetzung ist im Laufe des Zellzyklus oder des Lebens eines Organismus ständigen Änderungen unterworfen. Daraus folgt, dass das Proteom im Gegensatz zum Genom dynamisch ist. Ein gutes Beispiel, um dies zu verdeutlichen, sind die verschiedenen Entwicklungsstadien eines Schmetterlings, nämlich Ei, Raupe, Puppe und der Schmetterling selbst. Alle vier Entwicklungsstadien beruhen auf dem gleichem Genom, besitzen aber deutlich unterschiedliche Proteome. Mit dem Proteom besser vergleichbar ist die Gesamtheit der aktiven Gene eines bestimmten Zustands. Dieser wird als Transkriptom bezeichnet und ist ebenfalls dynamisch [25].

Das Transkriptom bestimmt welche Proteine hoch- bzw. herunterreguliert werden. Es wird durch eine Vielzahl von Einflüssen, inneren wie äußeren, in seiner Zusammensetzung beeinflusst (siehe Abbildung 2.7). Der Mechanismus der Genregulation ist für die Zusammensetzung der Proteinpopulation lebender Zellen von entscheidender Bedeutung. Er ermöglicht es, ein Protein nicht nur zu exprimieren oder dies zu verhindern, sondern erlaubt es darüber hinaus auch festzulegen, wie viele Proteine eines bestimmten Typs exprimiert werden sollen. Er bestimmt also auch die einzelnen Proteinkonzentrationen.

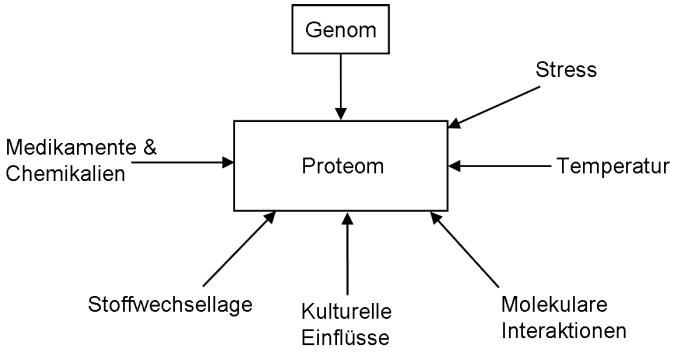

Abbildung 2.7: Zusammenstellung einiger auf die Proteinexpression Einfluss nehmender Faktoren. Quelle: [30]

Kapitel 3

Methoden der Proteinanalytik

Eine der Hauptaufgaben der Proteomforschung ist die Analyse der Gesamtheit der in einer Zelle oder einem Gewebe vorhandenen Proteine. Um das Proteom eines bestimmten Zelltyps, Gewebes oder Organismus zu einem bestimmten Zeitpunkt und zu definierten Bedingungen zu analysieren, müssen eine Reihe von Analyseschritten durchlaufen werden, bevor man letzten Endes die Primärstruktur der einzelnen Proteine kennt.

In der Einleitung dieses Dokuments ist bereits erwähnt worden, dass es Ziel dieser Diplomarbeit ist, einen *de novo*-Algorithmus für die Proteinidentifikation zu entwickeln. Dieser Algorithmus soll dazu in der Lage sein die Aminosäuresequenz eines Proteins ausgehend von vorher identifizierten Peptidsequenzen ohne Sequenzabgleiche mit Proteindatenbanken zu bestimmen. Da sowohl der datenbankgestützte Ansatz der Proteinidentifikation (siehe Kapitel 4) als auch der *de novo*-Ansatz Massenspektren als Datengrundlage nutzen, sollen in diesem Kapitel die Grundlagen der Massenspektrometrie vermittelt werden. Bevor man aber im Rahmen der Proteomforschung ein Protein oder Proteingemische einer massenspektrometrischen Analyse unterziehen kann, müssen in der Regel noch einige andere Analyseschritte vorausgehen. Da die Massenspektrometrie also nur ein Analyseschritt im Gesamtablauf der Proteinidentifikation ist, wird sie im Folgenden als Teil des Gesamtidentifikationsprozesses vorgestellt.

3.1 Exemplarisches Vorgehen bei der Proteinidentifikation

Typischerweise gliedert sich der Prozess der Proteinidentifikation in die folgenden Schritte (siehe Abbildung 3.1)

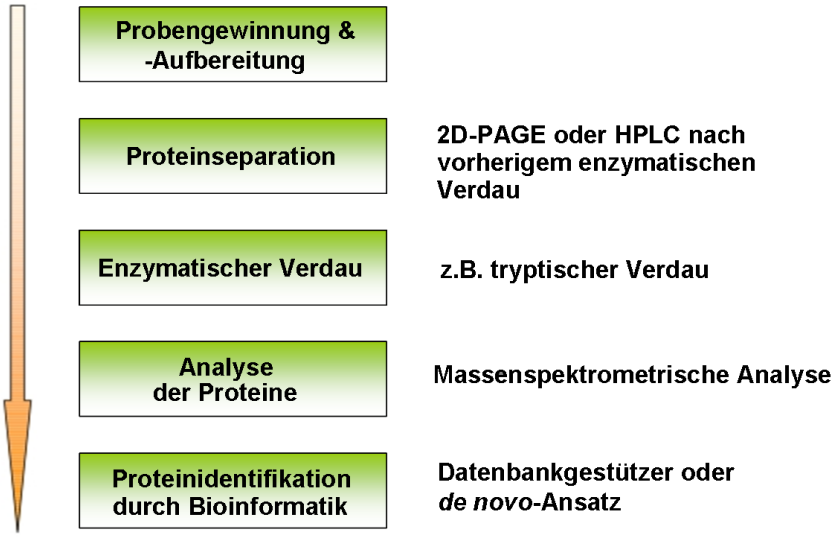

Abbildung 3.1: Zusammenstellung der Analysephasen der Proteinidentifikation.

3.1.1 Probengewinnung und -aufbereitung

Da die Proteinidentifikation häufig im Rahmen der Erforschung von Krankheiten, ihrer Symptome und Begleiterscheinungen stattfindet, werden für solche Analysen zwei verschiedene Zelltypen oder Zellstadien analysiert, die mit einer speziellen Erkrankung in Verbindung stehen. Im Zusammenhang mit der Erforschung von Krebserkrankungen z.B., werden für gewöhnlich bestimmte Zelltypen untersucht (z.B. im Kontext der Leberzirrhose, Leberzellen) und zwar vor und nach Ausbruch der Erkrankung. Durch solche so genannten differentiellen Analysen wird festgestellt, welche Proteine und in welcher Konzentration bestimmte Proteine von kranken Zellen exprimiert werden. Dies erlaubt es im Umkehrschluss, bestimmte Krankheiten schon frühzeitig zu erkennen.

Die zu untersuchenden Zellen oder Zellbestandteile müssen für die eigentliche Analyse entsprechend gewonnen und aufbereitet werden. Untersucht man z.B. eine bestimmte Krebserkrankung, so werden Proben eines entsprechenden Tumors (und seiner verschiedenen Stadien) aus erkranktem Gewebe entnommen und hinsichtlich interessanter Zellen und Zellbestandteile aufgearbeitet. Die in diesen biologischen Strukturen enthaltenen Proteine müssen anschließend extrahiert, getrennt und „sichtbar“ gemacht werden. Dazu werden die zu untersuchenden Zellen als Teil einer Probe zunächst einmal labortechnisch aufbereitet und die unerwünschten Zellbestandteile entfernt, dies kann z.B. durch Zentrifugation geschehen.

3.1.2 Proteinseparation

Da in der Proteinanalytik praktisch immer mit Proteinproben gearbeitet wird, die entweder viele verschiedene Proteine oder eine hohe Konzentration eines bestimmten Proteins enthalten, ist der erste eigentliche Analyseschritt eine Proteinseparationsmethode. Je nachdem wie komplex das zu untersuchende Proteingemisch und wie groß der Anteil der eigentlich interessanten Proteine an diesem Gemisch ist, können auch mehrere Proteinseparationsschritte notwendig werden.

Im Allgemeinen werden in der Proteinanalytik zwei relativ unterschiedliche Proteinseparationsmethoden eingesetzt. Beide wurden schon in den 70er Jahren des zwanzigsten Jahrhunderts entwickelt und seit dem kontinuierlich verbessert. Dies ist zum einen die so genannte zweidimensionale Gelelektrophorese (2D-PAGE), die bereits 1975 entwickelt wurde [31, 32] und zum anderen die so genannte High Performance Liquid Chromatography (HPLC) [33, 34], die eine spezielle Form der Flüssigchromatographie darstellt. Beide Verfahren unterliegen zwar gewissen Beschränkungen, jedes Verfahren hat spezifische Vor- und Nachteile, stellen aber nichtsdestotrotz Schlüsseltechnologien für die gesamte Proteinanalytik dar. Durch sie werden die weiteren Analyseschritte überhaupt erst möglich. Im Folgenden wird davon ausgegangen, dass die so genannte 2D-PAGE als Proteinseparationsmethode verwendet wird, da diese gegenüber der HPLC-Methode einige entscheidende Vorteile besitzt und in ihrer Anwendung anschaulicher ist.

Um ein so genanntes 2D-Gel zu erstellen, wird auf einem rechteckigen Elektrophorese-Gel zunächst ein Proteingemisch in einer Richtung entsprechend der Säure-Basen-Eigenschaften der in ihm enthaltenen Proteine getrennt (1. Dimension). Dies geschieht per so genannter isoelektrischer Fokussierung (IEF), bei der eluierte Proteine in einen Gelzylinder gegeben werden, an den anschließend ein elektrisches Feld angelegt wird. Dieses Feld trennt die basischen und sauren Proteine hinsichtlich ihres pH-Werts auf. Anschließend trennt man die so entstandenen Proteinfraktionen durch ein rechtwinklig zur ersten Trennung angelegtes elektrisches Feld (2. Dimension). Hierbei wandern die Proteine entsprechend ihrer Molekülgrößen unterschiedlich schnell in das Gel hinein und trennen sich dabei auf. Nach Beendigung der Elektrophorese legt man das Gel in eine Farbstofflösung, um die darin enthaltenen Proteine anzufärben und damit sichtbar zu machen. Als Ergebnis erhält man ein zweidimensionales Muster von Flecken (so genannte Spots), deren Positionen charakteristisch für die jeweiligen Proteine sind. Gute Trenngele können heute bereits bis zu 10.000 separate Proteinspots auflösen. Die vergleichende Auswertung dieser komplizierten Muster gelingt nur dank hoch auflösender elektronischer Kameras und hoch spezialisierter Experten, die durch leistungsfähige Bildanalysesoftware unterstützt werden [25].

Aus dem so entstandenen Protein-Gel lassen sich einige wertvolle Informationen über das aufgetrennte Proteingemisch gewinnen. Zunächst ist es möglich, die ungefähre Anzahl der in dem Gemisch enthaltenen voneinander trennbaren Proteine zu entnehmen. Zweitens sieht man deutlich, welche Proteine in besonders großen Mengen vorkommen (zugehörige Spots sind besonders ausgeprägt) und man lernt drittens, welche Molekülgrößen und Säure-Basen-Eigenschaften diese Proteine haben. Der wichtigste Vorteil, der sich aus der Erstellung des Gels ergibt, ist jedoch die Möglichkeit die Proteine einzelner Spots zu extrahieren und

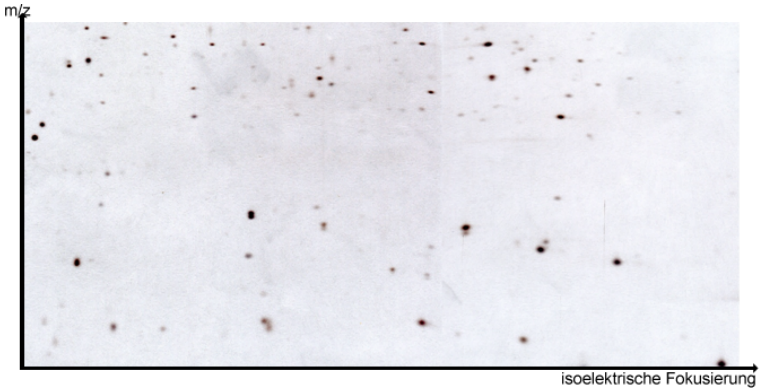

Abbildung 3.2: Beispiel für ein mit der 2D-Gelelektrophorese erzeugtes Proteingel. Die Trennung gemäß pH-Wert erfolgte von links nach rechts, die Trennung gemäß den Proteingrößen bzw. gemäß des Masse/Ladungsverhältnises (m/z) senkrecht dazu. Quelle: [35].

anschließend mittels Massenspektrometrie zu analysieren und zu identifizieren.

Die zweidimensionale Gelelektrophorese ist sehr leistungsfähig und ist seit ihrer Entdeckung zu einer der bedeutendsten Proteineseparationsmethoden avanciert. Leider hat sie aber auch Grenzen. Proteine mit sehr niedrigem (sauerer Bereich) oder sehr hohem pH-Wert (basischer Bereich) lassen sich nicht gut voneinander trennen. Besonders enttäuschend ist die bisher erreichte Trennleistung bei Membranproteinen. Diese über lange Abschnitte in den Membranen der Zellhülle gelösten Proteine haben einen stark hydrophoben (wasserabweisenden), lipophilen (fettliebenden) Charakter und sind deshalb im wässrigen Milieu der Trenngele nur schwer löslich. In den letzten Jahren haben sich aber gerade die Membranproteine für die Pharmaforschung als von großem Interesse erwiesen, da sie in der interzellulären Kommunikation, die beim Auftreten von Krankheiten oft gestört ist, eine zentrale Rolle spielen.

Ein weiteres Problem der 2D-PAGE ist das Anfärben von Proteinspots. Hier können sich Proteine sehr unterschiedlich verhalten. Die Verwendung einer gewissen Menge eines Farbstoffs führt nicht bei allen angefärbten Proteinen eines Gels zu einer vergleichbaren Farbintensität. Aus diesen Gründen investiert man gegenwärtig noch immer viel Entwicklungsarbeit in die Verbesserung der Elektrophoresetechnik und in neue zusätzliche Techniken, die die Begrenzungen der 2D-PAGE überwinden können.

Trotz einiger Limitationen wie der aufwendigen technischen Durchführung, einer begrenzten Reproduzierbarkeit und limitierten Dynamik ist die 2D-PAGE bis heute die einzige hoch auflösende Aufreinigungs- und Trennmethode, welche die Darstellung und Quantifizierung von bis zu 10.000 Proteinen aus komplexen Gemischen wie Zellen, Geweben oder Körperflüssigkeiten ermöglicht [30]. Die Proteinseparation gemäß HPLC stellt eine sinnvolle Ergänzung zur 2D-PAGE dar, da sie automatisierbar ist, keine aufwändige Erstellung eines Gels erfordert und die direkte mehrdimensionale massenspektrometrische Analyse von proteolytisch verdauten komplexen Proteingemischen erlaubt. Dies ist insbesondere bei der Analyse von post-translationalen Modifikationen ein großer Vorteil.

3.1.3 Proteolyse der zu untersuchenden Proteine

Nachdem die zu untersuchenden Proteine mittels 2D-PAGE auf dem Gel sichtbar gemacht wurden, ist es nun möglich die zugehörigen Gel-Spots mit Hilfe eines Robotersystems präzise aus dem Gel auszuschneiden und anschließend weiter zu analysieren, dafür genügen bereits schon geringste Mengen an Probenmaterial (ein Gel-Spot besitzt oft nur eine Masse von wenigen Nanogramm).

Um nun in weiteren Analyseschritten auf die Aminosäuresequenz des zu analysierenden Proteins schließen zu können, wird dieses mit Hilfe von speziellen Enzymen, so genannten Proteasen, in kleinere Bestandteile (Peptide) zerlegt, die man bzgl. ihres Aufbaus untersucht. Man bezeichnet diesen Prozess als proteolytischen Verdau. Proteasen sind spezielle Proteine, die auf die Spaltung der Peptidbindungsbrücken anderer Proteine spezialisiert sind. In der Proteinanalyse werden in der Regel so genannte spezifische Proteasen eingesetzt, diese trennen die Peptidbindungen zwischen den Aminosäuren des zu verdauenden Proteins an definierten, eben spezifischen, Stellen auf. Die Schnittstellen, an denen eine Protease ein Protein schnei-

det, werden dabei durch seine so genannte Substratspezifität festgelegt (siehe Tabelle 3.1). Die häufig für den proteolytischen Verdau verwendete Protease Trypsin schneidet ein Protein nach dem Auftreten der Aminosäuren Arginin und Lysin.

Protease	**spezifische Schnittstellen**
Trypsin (strict)	Arginin (R) & Lysin (K)
Trypsin	Arginin (R), Lysin (K), Leucin (L), Aspargin (N) und Histidin (H)
Chymotrypsin	Phenylalanin (F), Tryptophan (W) und Tyrosin (Y)
Glu-C	Asparaginsäure (D) & Glutaminsäure (E)
Lys-C	Lysin (K)
Elastase	Alanin (A), Valin (V), Leucin (L), Isoleucin (I)

Tabelle 3.1: Zusammenstellung der am häufigsten verwendeten Proteasen und ihrer spezifischen Schnittstellen.

Proteasen, wie das eben genannte Trypsin z.B., können Proteine auch unspezifisch oder unvollständig schneiden. Schneidet eine Protease ein Protein unspezifisch, so trennt sie das Protein nach dem Vorkommen einer substratunspezifischen Aminosäure auf. Die Eigenschaft einer Protease, unvollständig schneiden zu können, führt dazu, dass definierte Schnittstellen auch übersprungen werden können. Beide Seiteneffekte, das Schneiden an unspezifischen Stellen als auch das Überspringen von definierten Schnittstellen, treten in Abhängigkeit von den gewählten Reaktionsbedingungen unter denen der Verdau stattfindet mehr oder weniger häufig auf.

3.1.4 Grundlagen der Massenspektrometrie

Die im Folgenden dargestellten Grundlagen der Massenspektrometrie basieren auf einem Artikel der curricularen Chemie-Enzyklopädie ChemgaPedia (www.chemgapedia.de, Stand 31. März 2006).

Grundprinzip der Massenspektrometrie (MS) ist es, aus anorganischen oder organischen Substanzen in geeigneter Weise Ionen zu erzeugen, d.h. die Moleküle aus denen diese Substanzen bestehen elektrisch aufzuladen und diese Ionen anschließend mit Hilfe eines Registriersystems bzgl. ihrer Masse und Häufigkeit qualitativ und quantitativ zu erfassen. Die Ionisation der Substanzen kann thermisch, durch elektrische Felder oder durch Beschuss der Probe mit Elektronen, Ionen oder Photonen erfolgen. Im Allgemeinen sind die in der Proteinforschung entstehenden Ionen positiv geladen und können einzelne ionisierte Atome, ionisierte Moleküle, deren Bruchstücke oder Assoziate (Vereinigung von mehreren gleichartigen Molekülen zu größeren Komplexen) sein. Die Massenspektrometrie ist eine zerstörerische Analysemethode, bei der das Analyt verbraucht wird.

Massenspektrometer lassen sich aufgrund der von ihnen eingesetzten Ionisierungs- oder Ionenseperationstechnik unterscheiden, in Bezug auf die Proteinanalystik ist dabei die Unterscheidung bzgl. der Ionisierungsmethode die wichtigere der beiden. Da die in der Proteinanalytik zu untersuchenden Proteine und Proteingemische oft stark differierende chemische Eigenschaften besitzen, benötigt man verschiedene Typen von Ionisierungsmethoden. Im Kontext der Proteinanalystik sind dabei zwei Methoden besonders wichtig, die Matrix-assisted-Laser-Desorption-Ionisation (MALDI) [36, 37, 38, 39] (siehe Abschnitt 3.1.5) und die Elektrospray-Ionisation (ESI) [27] (siehe Abschnitt 3.1.5), beide wurden in den 80er Jahren des zwanzigsten Jahrhunderts entwickelt.

3.1.5 Aufbau eines Massenspektrometers

Unabhängig von der eingesetzten Ionisierungstechnik lässt sich der grundlegende Aufbau eines Massenspektrometers in fünf Teile gliedern: Das Einlasssystem, die Ionenquelle, den Analysator, den Detektor und das so genannte Datensystem (siehe Abbildung 3.3).

Das Einlasssystem

Über das Einlasssystem gelangt die zu analysierende Probe in den luftleeren Bereich des Massenspektrometers. Die hierfür verwendete Überführungsmethode hängt von den Eigenschaften des Analyten (Siedepunkt, thermische Stabilität, etc.) und der im Folgenden verwendeten Ionisationsart ab.

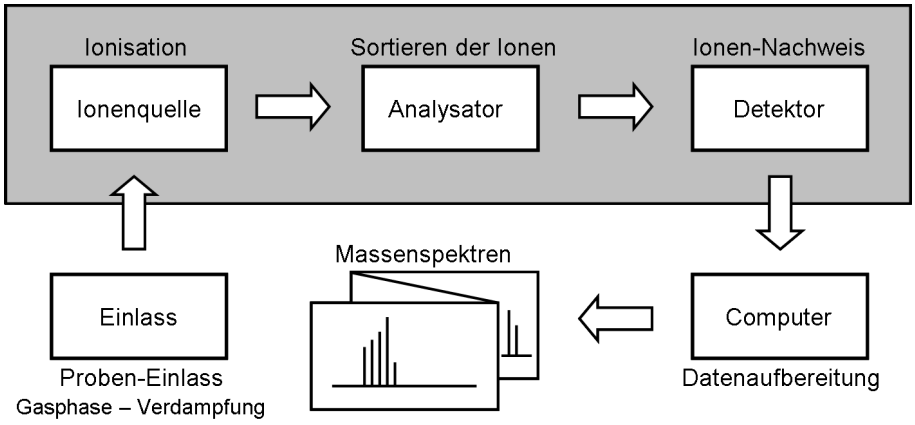

Abbildung 3.3: Schematischer Aufbau eines Massenspektrometers.

Die Ionenquelle

In der Ionenquelle wird die zu untersuchende Probe ionisiert. Dabei werden die Probemoleküle durch Zufuhr von Energie in gasförmige Ionen umgewandelt. Für diesen Prozess nutzt man die kinetische Energie von Elektronen, Ionen, Molekülen oder Photonen. Mit Hilfe dieser Methoden lassen sich nahezu alle bekannten Verbindungen ausreichend gut und reproduzierbar ionisieren. Bei der Auswahl der für eine bestimmte Probe zu verwendenden Methode richtet man sich nach dem physikalischen Zustand der Probe sowie nach ihrer thermischen Stabilität. In der Regel sind Massenspektrometer so konstruiert, dass mehrere Ionisationstechniken genutzt werden können.

Das Massenspektrum eines Moleküls hängt sehr stark von der verwendeten Ionisationstechnik ab. Grundsätzlich lassen sich sämtliche Ionisationsmethoden in „harte“ und „weiche“ Methoden einteilen:

- **Harte Ionisation**
  Die zugeführte Energie ist so hoch, dass zusätzlich zur Ionisation Fragmentierungsreaktionen ausgelöst werden. Diese Fragmentierungen sind von der chemischen Struktur abhängig, man kann sie also zur Strukturaufklärung verwenden. Diese Form der Ionisation wird aufgrund der bei den Molekülen auftretenden Fragmentierungserscheinungen nicht in der Proteinanalyse eingesetzt.

- **Weiche Ionisation**
  Die untersuchten Substanzen werden nicht oder nur geringfügig fragmentiert. Es werden Molekül- oder Quasi-Molekülionen gebildet. Quasi-Molekülionen sind ioniersierte Moleküle deren atomare Zusammensetzung während der Ionisation verändert wurde. Diese Veränderung manifestiert sich in einem Protonentransfer zwischen den Atomen aus denen das Quasi-Molekülion besteht. Die Umsetzung der weichen Ionisation gelang erst in den 80er Jahren des zwanzigsten Jahrhunderts durch die Entwicklung sanfter Desorptions-/Ionisationsmethoden (siehe die Abschnitt zu ESI und MALDI weiter unten), wodurch Proteine der massenspektrometrischen Analyse überhaupt erst zugänglich wurden.

  Je nach Art der Probenzufuhr lassen sich Ionenquellen noch in die folgenden Subtypen unterteilen:

  – **Gasphasen-Ionenquellen**
    Proben, die sich im Vakuum verdampfen lassen, können vor der Ionisierung in die Gasphase überführt werden. Die Zufuhr der Probe erfolgt über ein indirektes oder ein direktes Einlass-System oder mit Hilfe eines Gaschromatographen (LC und HPLC).
  – **Desorptions-Ionenquellen**
    Die Zufuhr der bereits kondensierten Probe in die Ionenquelle erfolgt über ein Direkteinlass-System. Mit Hilfe spezieller Ionisierungstechniken, z.B. MALDI, werden direkt aus der kondensierten Phase gasförmige Ionen gebildet. Es können also auch nichtflüchtige und thermisch labile Verbindungen untersucht werden.
  – **Spray-Ionenquellen**
    Flüssige oder eluierbare Proben lassen sich mit Hilfe einer Kapillare in die Ionenquelle einbringen und dort zu einem feinen Nebel versprühen (ESI). Aus den Nebeltröpfchen treten Ionen in

die Gasphase über und werden anschließend in das Vakuumsystem des Massenspektrometers überführt.

Die Elektrospray-Ionisation (ESI)

Grundprinzip der so genannten Elektrospray-Ionisation (ESI) ist es, eine Lösung, welche die zu untersuchenden Substanzen enthält, durch elektrische Kräfte in ein extrem feines Aerosol aus hochgeladenen Tröpfchen zu überführen. Dazu verwendet man eine metallene Kapillare, welche zugleich die Kathode eines starken elektrischen Feldes (üblicherweise wird eine Spannung 2 bis 5 kV zum Aufbau des Feldes verwendet) darstellt (siehe Abbildung 3.4). Die hohe Feldspannung sorgt dafür, dass die in der Ionenquelle vorherrschenden elektrostatischen Kräfte so groß werden, dass der von der Kapillare erzeugte Flüssigkeitsstrahl sich sehr schnell in eine Tröpfchenwolke verwandelt. Da man aber letztendlich die in den Tröpfchen enthaltenen einzelnen Molekülionen getrennt untersuchen möchte, werden die Tröpfchen auf ihrem Weg ins Vakuum des Massenspektrometers mit Hilfe eines heißen Trocknungsgases sukzessive verdampft. Die kontinuierliche Verkleinerung des Tröpfchendurchmessers führt zu einem stetig anwachsenden Ladungsdichteverhältnis auf der Oberfläche der Tröpfchen. Ab einem gewissen Tröpfchendurchmesser ist die Oberflächenspannung der Tröpfchen so niedrig geworden, dass sie nicht mehr länger dazu in der Lage ist die interagierenden Coulomb-Kräfte der einzelnen geladenen Molekülionen zu kompensieren (siehe Abbildung 3.5). Ab diesem Punkt ist das so genannte Rayleigh-Limit erreicht [40, 41] und die Masse der einzelnen Tröpfchen schrumpft rapide, da die gleichartiggeladenen Molekülionen sich aufgrund der verminderten Oberflächenspannung der Tröpfchen gegenseitig aus diesen heraus katapultieren. Es entsteht eine dichte Raumladungswolke, die aus unzähligen winzigen Tröpfchen besteht. Jedes dieser Tröpfchen besitzt zwar nur eine äußerst geringe Masse, betrachtet man aber die Gesamtmasse sämtlicher Tröpfchen, so macht diese einen hohen Anteil der gesamten Molekülionenmasse aus [42].

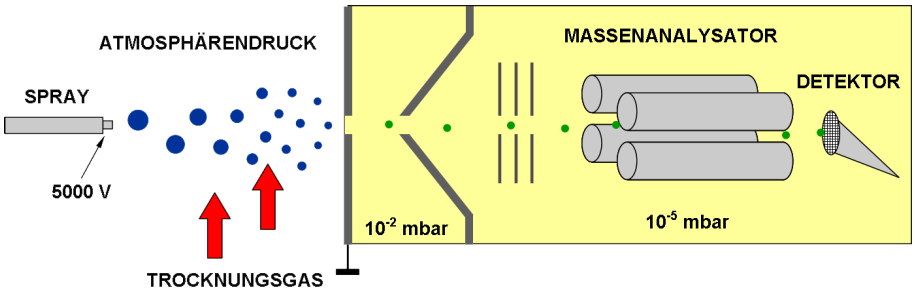

Abbildung 3.4: Schematische Darstellung der Ionenquelle eines ESI-MS. Eine Lösung mit den zu untersuchenden Molekülen wird über eine Kapillare, die zugleich die Kathode eines starken elektrischen Feldes ist, in die Ionenquellkammer gesprüht. Die so entstehenden Lösungströpfen werden mit Hilfe eines Trocknungsgases nach und nach soweit verkleinert, sodass nur noch die Molekülionen der eigentliche Probe detektiert werden. Quelle: [43].

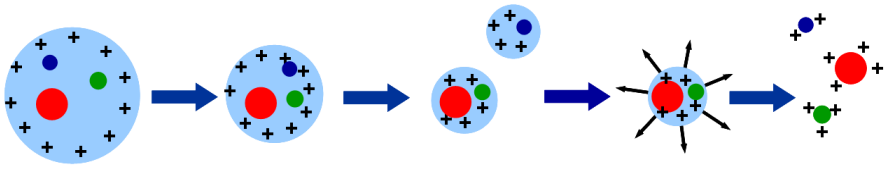

Abbildung 3.5: Darstellung des Schrumpfungsprozesses eines Aerosoltröpfchens, wie er in der ESI-Ionenquelle stattfindet. Der Tropfen schrumpft kontinuierlich, bis sein Durchmesser so klein geworden ist, dass seine Oberflächenspannung die Abstoßungskräfte der gleichartiggeladenen Molekülionen nicht mehr kompensieren kann. Die in dem Tropfen enthaltenen Molekülionen stoßen sich gegenseitig ab und verlassen so den schrumpfenden Tropfen. Zuletzt bleiben nur noch die freien Molekülionen übrig, die entlang der Feldlinien des elektrischen Feldes in Richtung des Detektors wandern. Quelle: [43].

Die Matrix-assisted-Laser-Desorption-Ionisation (MALDI)

Das Herzstück der MALDI-Technik ist ein Laser, der einen extrem kurzen (einige milliardstel Sekunden) und intensiven Blitz von ultraviolettem Licht erzeugt, mit dem die Proteinprobe in der Ionenquelle beschossen wird (siehe Abbildung 3.6). Bei direktem Laserbeschuss würde sich das häufig hitzeempfindliche Probenmaterial extrem schnell und stark aufheizen. Dieser Effekt ist bei typischen technischen Laseranwendungen erwünscht, würde empfindliche Substanzen wie Proteine allerdings zerstören. Deswegen wird ein physikalisch-chemischer Trick angewendet: Die hitzeempfindliche Probe wird durch einen Matrixkristall, auf dem die zu untersuchenden Proteinmoleküle isoliert und sehr verdünnt vorliegen, geschützt. Heutigen Modellvorstellungen zu Folge, geht man davon aus, dass die im Kristall regelmäßig angeordneten Matrixmoleküle einen Grossteil der Energie des Laserlichts absorbieren (siehe Abbildung 3.7). Das Laserlicht dringt nur oberflächlich in den Kristall ein und führt in einer dünnen Oberflächenschicht der Probe zu extremen strukturellen Veränderungen, in deren Folge es zu einer Mikroexplosion kommt. Hierdurch werden Teile der MALDI-Matrix und des Probenmaterials, welche durch den Laser ionisiert wurden und sich in Folge dessen zu einer Wolke aus winzigen Partikeln und Gasen zusammengeschlossen haben, von der Kristalloberfläche ins Vakuum geschleudert. Diesen Prozess nennt man Laserdesorption oder -ablation. Durch seine technische Einfachheit, die hohe Genauigkeit der Massenbestimmung sowie die Schnelligkeit und Automatisierbarkeit der Messung ist die MALDI-Technologie heute ein unverzichtbares Werkzeug in der Bioanalytik [44].

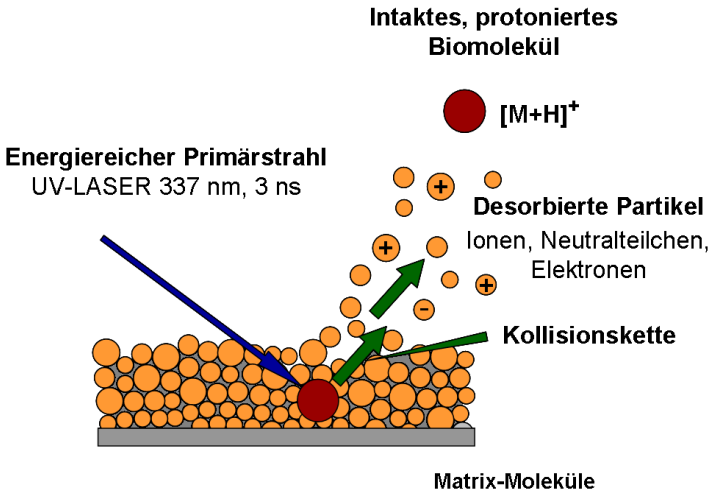

Abbildung 3.6: Schematische Darstellung des MALDI-Ionisierungsprozesses. Quelle: [43].

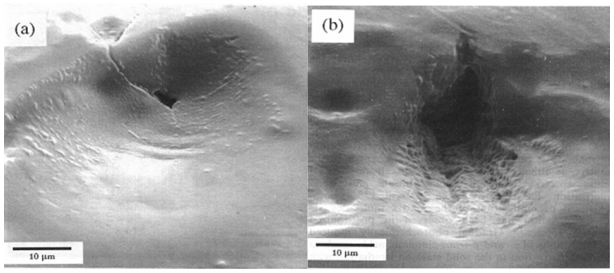

Abbildung 3.7: Voher-Nachher-Aufnahme einer MALDI-Matrixplatte. Links sieht man einen kleinen Ausschnitt der Matrixplatte mit der darauf aufgetragenen Probe. Rechts die gleiche Stelle auf der Matrixplatte nach Ionisation der Biomoleküle aus der Probe. Quelle: [43].

Der Massenanalysator

Aufgabe des Massenanalysators ist es die in der Ionenquelle erzeugten und beschleunigten Ionen von einander zu trennen und diese dann dem Detektor zu Massenbestimmung zu zuführen. Entscheidend für

die Trennung der Molekülionen ist ihr Masse/Ladungsverhältnis m/z. Die Trennung der Ionen beruht auf verschiedenen physikalischen Prinzipien:

- Ablenkung von Ionenstrahlen in elektrischen oder magnetischen Feldern (Sektorfeldgeräte)
- Filterung von Ionen unterschiedlicher Masse in elektrischen Wechselfeldern (Quadrupolmassenfilter, Ionenfalle, Zyklotronresonanz-Analysator)
- Auftrennung aufgrund der unterschiedlichen Flugzeit von Ionen im feldfreien Raum (TOF (engl.): Time of Flight)

Für spezielle Messungen kann man auch mehrere Analysatoren hintereinander schalten. Man erhält damit entweder hochauflösende Sektorfeldgeräte, mit denen man die Masse ausgewählter Ionen mit hoher Genauigkeit bestimmen kann, oder Tandem-Massenspektrometer (MS/MS), die bei der Strukturaufklärung der Primärstruktur von Proteinen eine wichtige Rolle spielen.

Der Detektor

Die in der Ionenquelle gebildeten und vom Massenanalysator entsprechend ihres Masse/Ladungsverhältnises getrennten Ionen werden von einem Detektor registriert. Dabei wird die Intensität des jeweils zugehörigen Ionenstroms ermittelt. Der Detektor erzeugt ein elektrisches Signal, einen so genannten Peak, welches nach seiner Digitalisierung zur Auswertung bereit steht.

Die Anfertigung eines Massenspektrums kann ortsabhängig oder zeitabhängig erfolgen. Man spricht von einer ortsabhängigen Detektion, wenn die Ionen vom Massenanalysator auf Bahnen mit unterschiedlichem Radius gelenkt und an verschiedenen Orten registriert werden. Zumeist verwendet man aber Massenspektrometer, die den Ionenstrom zeitabhängig registrieren, da hierfür lediglich ein elektrischer Verstärker benötigt wird. Die Trennung der Ionen muss daher so gestaltet werden, dass Ionen mit unterschiedlichem Masse/Ladungsverhältnis den Detektor nacheinander erreichen (TOF).

Die wichtigsten Kenngrößen eines Detektors, sind seine Genauigkeit und Empfindlichkeit (statische Grössen) sowie der von ihm abgedeckte Detektionsbereich und seine Ansprechzeit (dynamische Größen). Leider lassen sich diese Kenngrößen für einen bestimmten Detektor nicht alle gleichzeitig optimieren. Deshalb muss sich die Wahl des anzuwendenden Detektors nach den Anforderungen des jeweiligen Experiments richten.

Das Datensystem

Das Datensystem dient der Erfassung der von dem Detektor gemessen Daten. Diese Daten werden anschließend bearbeitet und gespeichert und stehen dann für weitere Auswertungen zur Verfügung. Zudem dient das Datensystem der Steuerung des gesamten Massenspektrometers. Um diese Aufgaben erfüllen zu können bedarf es des Einsatzes leistungsfähiger Computersysteme und entsprechend leistungsstarker Algorithmen. In diesem Zusammenhang erfüllt die Bioinformatik drei wesentliche Aufgaben:

1. **Datenerfassung**
   Im ersten Schritt müssen die vom Detektor registrierten analogen Signale in digitale Signale umgewandelt werden. Anschließend muss die Menge der gemessenen Daten durch die Einführung eines Intensitäts-Schwellwertes reduziert werden. Zu guter Letzt wird aus dem Peak-Zentrum, dem Bereich eines Massenspektrums in dem die Anzahl der gemessenen Signale am höchsten ist, mit Hilfe einer zuvor abgelegten Kalibrierfunktion der Wert des Masse/Ladungsverhältnisses und aus der Peakfläche die Intensität des gemessenen Signals ermittelt.
2. **Datenbearbeitung**
   Zur Datenbearbeitung gehören mathematische Operationen wie die Normierung auf den Basispeak (Peak mit der größten Intensität), die Subtraktion von Background-Spektren, die Spektrenaddition und die Rekonstruktion des zeitlichen Verlaufs der Intensität in der so genannten RIC-Funktion (RIC (engl.): reconstructed ion current). Anschließend lassen sich die ermittelten Daten interpretieren. Hierbei spielen Spektrenbibliotheken und Suchalgorithmen eine wichtige Rolle, da erst durch diese eine Vielzahl von Informationen zugänglich werden.

3. **Steuerung**
   Zur Steuerung eines Massenspektrometers gehört sowohl die Instrumentkontrolle als auch die Optimierung der Messbedingungen.

Nachdem die Funktionalität und Beschaffenheit der einzelnen Bestandteile eines Massenspektrometers erläutert wurden, soll nun das Vorgehen beim Einsatz der Massenspektrometrie in der Proteinidentifikation erläutert werden.

3.1.6 Peptidmassenspektren (PMF)

Nach der enzymatischen Spaltung der zu untersuchenden Proteine folgt eine Analyse der resultierenden Proteinbestandteile. Da proteinspaltende Enzyme Proteine in der Regel nur an ganz bestimmten Stellen durchtrennen, ergibt sich für jedes verdaute Protein ein charakteristisches Muster von Peptiden, der so genannte Peptide Mass Fingerprint (PMF). Die Idee, diesen Fingerabdruck für die Proteinidentifikation zu nutzen, wurde 1993 von fünf verschiedenen Arbeitsgruppen unabhängig von einander veröffentlicht [10, 13, 45, 46, 47].

Um einen solchen Fingerabdruck zu erhalten, werden die aus der spezifischen Proteolyse des zu untersuchenden Proteins entstandenen Peptide mittels MALDI-TOF MS (siehe Abbildung 3.8) oder ESI-MS analysiert. Das so entstandene Massenspektrum kann dann zur Suche in Proteindatenbanken (siehe Kapitel 4) verwendet werden und so das zu untersuchende Protein mittels Massenabgleich seiner Peptide mit den Peptiden aus den Datenbankeinträgen anderer Proteine identifiziert werden. Genügen die so gewonnenen Information nicht um das Protein zuverlässig identifizieren zu können oder möchte man nicht nur eine Proteinidentifikation durchführen, sondern zudem die Aminosäuresequenzen der einzelnen in der Probe enthaltenen Peptide bestimmen, so führt man eine weitere massenspektrometrische Analyse durch, welche die Primärstruktur der einzelnen Peptide bestimmt (Tandem-MS-Analyse oder MS/MS-Analyse).

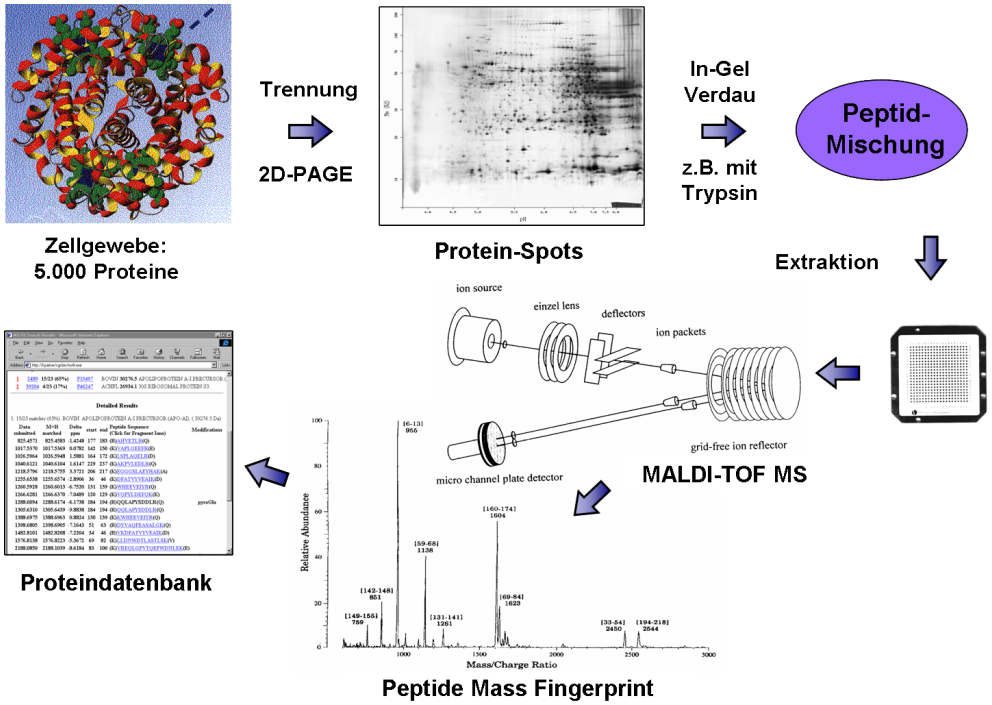

Abbildung 3.8: Schematische Darstellung des Ablaufs einer Proteinidentifikation gemäß MALDI-TOF MS. Quelle: [43].

Die PMF-Proteinidentifizierung wird hauptsächlich zur Identifikation von proteinreinen Proben verwendet (in der jeweiligen Probe ist nur ein bestimmtes Protein enthalten), sie kann aber auch für einfache Proteingemische angewendet werden [48]. Da die Proteinidentifikation per MS/MS-Analyse meist zuverlässiger als eine einfache MS-Analyse ist, und bei komplexen Proteingemischen die einzig Erfolg versprechende Analysemethode darstellt, ist sie heutzutage de facto Standard [30].

3.1.7 Peptidfragmentspektren (PFF)

Praktische alle heutzutage verwendeten Massenspektrometer erlauben die Selektion und Isolation von Peptiden anhand ihres Masse-/Ladungsverhältnisses. Nach der Isolation einzelner Peptide können diese mit verschiedenen Techniken wie PSD (Post Source Decay) [49] fragmentiert werden, so dass man ein Peptidfragmentspektrum erhält (siehe Abbildung 3.9). Da die Fragmentierung der Peptide hauptsächlich an den Peptidbindungen der Aminosäureketten geschieht, entsteht eine Art Leiter aus Peptidfragmentmassen, deren Abstände den Massen der Aminosäurereste entsprechen [50, 51]. Auf der Basis dieser Abstände lässt sich auf die Struktur des ursprünglichen Peptides schließen. Analog zu dem Verhältnis zwischen einem Protein und seinem Peptidmassenspektrum, gilt für Peptidfragmentspektren (PFF, Peptide Fragmentation Fingerprint) und Peptide, dass ein Peptidfragmentspektrum einem spezifischen Fingerabdruck des analysierten Peptides entspricht.

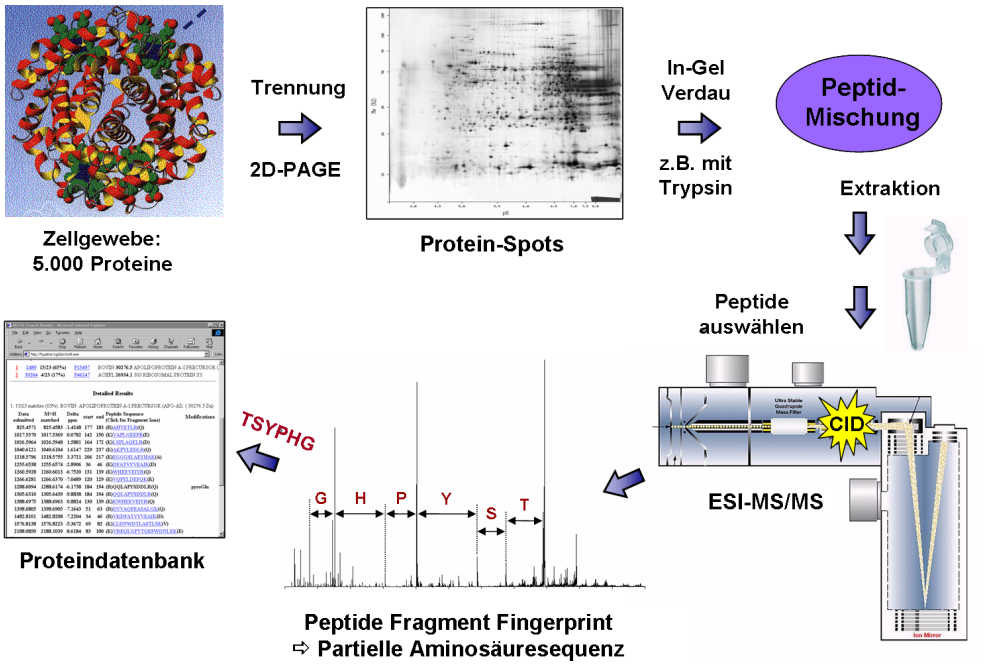

Abbildung 3.9: Schematische Darstellung des Ablaufs einer Proteinidentifikation gemäß ESI-MS/MS. Quelle: [43].

Sowohl Peptidmassenspektren als auch Peptidfragmentspektren werden heutzutage hauptsächlich mittels Proteinsequenzdatenbanken identifiziert. Falls die Aminosäuresequenzen der fragmentierten Peptide lang genug sind, kann eine eindeutige Proteinzuordnung gelingen. Da die Proteinidentifikation mittels Peptidfragmentspektren, im Gegensatz zur Identifikation per Peptidmassenspektren, auf Primärstrukturanalysen beruht, dürfen die verwendeten Datenbanken bezüglich der in ihnen enthaltenen genetischen Informationen unvollständig sein. Falls Peptide nicht in einer Datenbank enthalten sind, können Computeralgorithmen zur *de novo*-Sequenzierung heran gezogen werden. Da die Fragmentierung der Peptide allerdings oft unvollständig ist und teilweise nicht nur an den Peptidbindungen erfolgt, ist die Interpretation von Peptidfragmentspektren schwierig. Mehrdeutigkeiten bei der Analyse der Massenspektren, die auch mit erheblichem manuellem Aufwand nicht geklärt werden können, lassen sich nur selten vermeiden. Da Peptidfragmentspektren direkte Primärstrukturinformation enthalten, lassen sie sich im Gegensatz zu Peptidmassenspektren, die einen Überblick über das gesamte Protein geben, ausgezeichnet zur Aufklärung von post-translationalen Modifikationen, Aminosäuresubstitutionen und Sequenzfehlern heranziehen. Insbesondere im Hinblick auf die Analyse von komplexen Proteinmischungen, kann eine erfolgreiche Proteinidentifizierung nur mittels Peptidfragmentspektren gelingen. In der Regel werden Peptidmassenspektren und Peptidfragmentspektren nacheinander auf Basis der gleichen Probe ermittelt und gemeinsam genutzt, um eine eindeutige Proteinidentifizierung zu gewährleisten [30].

Kapitel 4

Die Rolle der Bioinformatik in der Proteomanalyse

Sämtliche in Kapitel Drei vorgestellten Analysemethoden der Proteomforschung haben eines gemeinsam: Sie erzeugen ein kaum überschaubares Datenaufkommen. Die Archivierung und Bildauswertung von Gelen, die Auswertung von Massenspektren und die Suche auf Genom- oder Proteom-Datenbanken wären allesamt ohne den Einsatz beträchtlicher Rechnerleistung, spezieller Software und Datenbanken mit entsprechenden Kapazitäten unmöglich.

Neben den ebengenannten Anwendungsgebieten beschäftigt sich die Bioinformatik noch mit weiteren Aufgabenstellungen aus der Proteomik. Diese werden in Abschnitt 4.1 überblickartig zusammengefasst. Im Anschluss an diesen Überblick richtet sich das Hauptaugenmerk dieses Kapitels auf eines der wichtigsten Betätigungsfelder der Bioinformatik innerhalb der Proteomik, der Interpretation massenspektrometrischer Daten (Abschnitt 4.2). Da jeder maschinelle Ansatz zur Interpretation massenspektrometrischer Daten mit einigen grundsätzlichen Probleme zu kämpfen hat, und diese von den bisher erarbeiteten datenbankorientierten Lösungsansätzen nur mehr oder weniger gut gelöst wurden, werden diese in Abschnitt 4.3 näher behandelt.

4.1 Die Aufgaben der Bioinformatik in der Proteomforschung

Schon seit Beginn der Genom- und Proteinforschung und den damit verbundenen Veröffentlichungen von sequenzierten Genomen und Proteomen, steigt die Menge der in Datenbanken gesammelten Sequenzinformation exponentiell an. Auch nach der Entschlüsselung des menschlichen Genoms verdoppelt sich die Menge der bekannten Sequenzen ca. jährlich (siehe Abbildung 4.1). Letztendlich lassen sich derartig große Datenmengen schon lange nicht mehr manuell handhaben und es werden Computersysteme benötigt, die diese Daten in eine Form bringen, die für Wissenschaftler effizient nutzbar ist.

Aus dieser Notwendigkeit heraus entstand die Bioinformatik als interdisziplinäre Wissenschaft zwischen Biologie und Informatik. Wichtige Aufgabenbereiche der Bioinformatik sind Datenarchivierung, Datensicherung, Bereitstellung des Zugangs zu archivierten Daten, Konsistenzhaltung, Erstellung von Querverweisen und Datenanalyse. Der Begriff der Bioinformatik ist bis heute nicht exakt definiert. Ursprünglich verstand man unter dem Begriff der Bioinformatik nur die Nutzung der angewandten Mathematik um experimentelle Protein- und Oligonukleotidsequenzen zu interpretieren. Typische Anwendungen aus der Bioinformatik sind Sequenzmustersuchen, die z.B. bei der Promotorerkennung [52] durchgeführt werden oder Homologiesuchen, wie sie das Programm Blast vornimmt [53]. Heutzutage umfasst die Bioinformatik ein sehr viel größeres Aufgabengebiet, das von der Vorhersage von Proteinstrukturen über statistische Analysen klinischer Studien bis zum Design von so genannten Bioinformatikplattformen reicht.

In der Proteomanalyse hat der technische Fortschritt, durch den die heutige Hochdurchsatzanalytik möglich wurde, zur Produktion von Datenmengen geführt, die manuell nicht mehr interpretierbar und in das bereits vorhandene Wissen nicht mehr manuell integrierbar sind. Kernaufgabengebiet der Bioinformatik in der Proteomanalyse ist derzeit vor allem die Interpretation von massenspektrometrischen Daten.

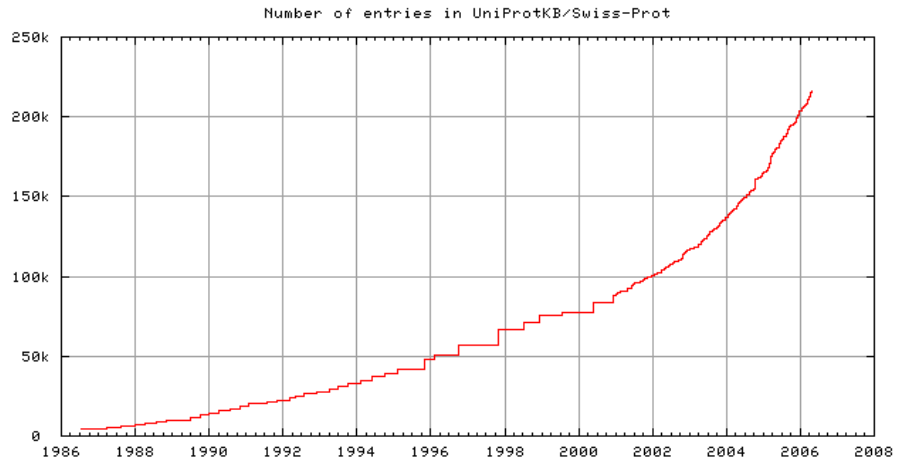

Abbildung 4.1: Statistik über die Entwicklung der Anzahl der Datenbankeinträge in der Proteindatenbank Swiss-Prot seit ihrer Entstehung (Stand vom 18. April 2006). Quelle: Swiss-Prot protein knowledgebase release 49.5 statistics

Bisher war der Erfolg der massenspektrometrischen Proteincharakterisierung abhängig vom manuellen und zeitintensiven Eingriff eines erfahrenen Benutzers. Seitdem pro Tag aber mehrere zehntausend Massenspektren pro Gerät erzeugt werden können, sind manuelle Analysemethoden nicht mehr adäquat. Es besteht daher ein großer Bedarf an Algorithmen zur Verbesserung der MS-Dateninterpretation, welche die Notwendigkeit einer manuellen Expertenanalyse ersetzen bzw. erleichtern und unterstützen [30].

Da Proteome ständigen Veränderungen unterliegen, ist es für Proteomstudien zwingend notwendig die einflussnehmenden Parameter so exakt wie möglich zu bestimmen, um so anhand eines möglichst genau definierten Proteomstatus die gefundenen Ergebnisse ihrer Kausalität zuordnen zu können. Daher ist es sinnvoll den Prozess der Proteomanalyse komplett mit den dazugehörigen Daten mit bioinformatischen Methoden zu erfassen. Hierfür eignen sich relationale Datenbanken, die es erlauben z.B. die Verbindung zwischen Probe, 2D-PAGE, Massenspektren und Sequenzdatenbankeinträgen abzubilden [54, 55, 56]. Dies stellt aufgrund der Heterogenität und Vielfalt der proteomischen Daten eine große Herausforderung dar. Alle relevanten Daten über Experimenthypothesen, Probendefinition, Protein/Peptid-Isolation und Fraktionierung, MS-Probenpräparation, massenspektrometrische Analysen und Interpretation der Massenspektren müssen gespeichert und Werkzeuge für die Datenanalyse und Visualisierung bereitgestellt werden. Die Entwicklung derartiger Bioinformatikplattformen ist trotz einiger Fortschritte noch immer in ihrem Anfangsstadium, und obwohl dringend benötigt, noch nicht allzu weit in der Proteomforschung verbreitet. Zurzeit gibt es mehrere kommerziell entwickelte Systeme, die sich hinsichtlich ihrer Eigenschaften und Merkmale deutlich unterscheiden. Zu den professionellen und kommerziell entwickelten Systemen zählen Proteinscape (Protagen AG, Bruker Daltonik GmbH), WorksBase (Bio-Rad Laboratories, Inc.) und ProteinLynx Global Server (Micromass), diese sind dazu in der Lage, den gesamten Ablauf einer Proteomanalyse von der Experimentplannung bis zur Primärstrukturaufklärung der Proteine relational abzubilden.

Der größte Teil der Erkenntnisse über identifizierte Proteine ist in Proteinsequenzdatenbanken gespeichert, welche als simpel strukturierte alphanumerische Textdateien, in der die Proteinsequenzen sequentiell gespeichert werden, verfügbar sind [57, 58, 59, 60]. Jeder Datenbankeintrag enthält mehrere Datenfelder, die spezielle vorgegebene Formate besitzen. In diesen Datenfeldern werden z.B. der Name des Proteins, Literaturverweise und Primärsequenzen gespeichert. Im Bereich der Proteomforschung dienen solche Sequenzdatenbanken der Proteinidentifizierung mittels Massenspektrometrie. Proteinsequenzdatenbanken werden aber auch häufig bei Homologie- oder Ähnlichkeitssuchen mit Algorithmen wie z.B. Blast verwendet. Proteinsequenzdatenbanken müssen in der Regel sehr hohen Ansprüchen genügen [61]. Sie sollen eine möglichst geringe Redundanz aufweisen, möglichst vollständig, aktuell, fehlerlos und kompatibel zu sämtlichen verfügbaren Bioinformatiksystemen sein. Zurzeit erfüllt keine Proteinsequenzdatenbank sämtliche der eben genannten Ansprüche vollständig. Die Proteindatenbank Swiss-Prot (216.380 Sequenzeinträge, Stand 18. April 2006) ist ein Beispiel für eine qualitativ hochwertige, gut annotierte und manuell editierte Datenbank (http://us.expasy.org/sprot). Allerdings enthält sie nicht immer die aktuellsten Sequenzinformationen, in Folge dessen enthält Swiss-Prot deutlich weniger Sequenzen als

beispielsweise die NCBI non-redundant (http://www.ncbi.nlm.nih.gov), welche Proteine aus sämtlichen bisher untersuchten Proteomen unterschiedlicher Organismen (Mensch, Maus, Ratte, usw.) enthält und mit insgesamt mehr als 3.4 Mio. Sequenzen (Stand Mitte April 2006) eine sehr umfassende Sammlung der Einträge aus mehreren anderen Datenbanken (GenBank, EMBL, DDBJ, PDB, Swiss-Prot, PIR, PRF) bereitstellt. Die NCBInr (abkürzende Schreibweise für NCBI non-redundant) wird beinahe täglich aktualisiert. Dafür sind die einzelnen Proteineinträge in der NCBInr weitaus weniger informativ, da sie neben der Aminosäuresequenz des jeweiligen Proteins lediglich eine NCBI-spezifische Accessionnummer, die wissenschaftliche Bezeichnung für das jeweilige Protein, eine Angabe bzgl. der Proteindatenbank, aus der das annotierte Protein stammt und eine zu dieser Proteindatenbank gehörige Accessionnummer enthalten. Zurzeit gibt es mehrere Ansätze neue, qualitativ hochwertige Proteindatenbanken zu entwickeln. Beispiele hierfür sind „Universal Protein Knowledgebase“ (UniProt, http://www.pir.uniprot.org), „International Protein Index“ (IPI, http://www.ebi.ac.uk/IPI) und „Human Protein Reference Database“ (HPRD, http://www.hprd.org).

Neben den Sequenzdatenbanken, die hauptsächlich die Primärstruktur von Proteinen enthalten, existieren einige weitere hoch spezialisierte Datenbanken. Beispiele hierfür sind metabolische Datenbanken, 2D-PAGE Datenbanken wie SWISS-2DPAGE (http://us.expasy.org/ch2d) [62] oder 3D-Strukturdatenbanken wie PDB (http://www.rcsb.org/pdb) [63]. Da erst vor nicht allzu langer Zeit einheitliche Standards für proteinspezifische Daten geschaffen worden sind [57, 64] ist der Austausch von Daten immer noch recht kompliziert. Immerhin ist aber ein deutlicher Trend festzustellen, die Daten im XML-Datenformaten zugänglich zu machen, was die computergestützte Erfassung und Bearbeitung der Daten deutlich vereinfacht [65].

4.2 Interpretation von Massenspektren durch die Bioinformatik

4.2.1 Präprozessierung von MS-Daten

Fast alle MS-basierten Suchmaschinen zur Proteinidentifikation akzeptieren die zu untersuchenden Massenspektren in der Form von so genannten Signallisten, dies sind Zusammenstellungen sämtlicher Signale eines Massenspektrums sowie der dazugehörigen Intensitäten und Ladungszahlen. Da Massenspektren heutzutage automatisiert und im Hochdurchsatz generiert werden, ist eine manuelle Signalerkennung und -extraktion eher selten geworden. Diese Aufgabe ist fast vollständig von Algorithmen übernommen worden [66, 67, 68]. Allerdings ist der erfahrene Benutzer den Algorithmen in komplizierten Datensituationen heutzutage noch immer überlegen, da Faktoren wie Rauschen, Signalüberlagerungen und sich unter bestimmten Bedingungen verändernde Verhältnisse zwischen dem monoisotopischen Signal und den anderen isotopisch aufgelösten Signalen eines zu untersuchenden Peptides die automatische Signalinterpretation erheblich erschweren. Im Falle der ESI siehe Abschnitt 3.1.5 werden die Proteine und Peptide gewöhnlich in höheren Ladungszahlen z (Ladungszahl z liegt für die Peptide bei ESI im Bereich von 1-4) detektiert, wodurch eine Dekonvolution (z = 1) notwendig wird. Dieser Prozess ist weitgehend automatisiert durch Algorithmen, die bei ausreichender Massengenauigkeit in der Lage sind, anhand des Isotopenmusters die Ladungszahl zu bestimmen, oder zumindest einzugrenzen [69, 70].

Sowohl bei der MALDI-TOF MS (siehe Abschnitt 3.1.5) als auch bei der ESI-MS (siehe Abschnitt 3.1.5) ergibt sich nach der Time-of-Flight-Analyse die Problematik der Kalibrierung. Sämtliche bisher entwickelten Ansätze zur automatischen Kalibrierung beruhen entweder auf der externen, statistischen oder internen Kalibrierung anhand von zugesetzten Standardpeptiden. Neben den Signalen der eigentlich zu untersuchenden Peptide enthalten die ermittelten Spektren oft eine Vielzahl weiterer Signale, die nicht auf das analysierte Protein zurückzuführen sind. Typischerweise sind dies Bestandteile der verwendeten MALDI-Matrix oder Farbreststoff aus der Färbeprozedur der 2D-PAGE. Zusätzlich zu diesen Kontaminationsquellen enthalten die Spektren häufig Signale, die auf das Protein Keratin zurückzuführen sind, welches dann für Gewöhnlich aus der Haut oder dem Haar eines Laboranten stammt. Solche Signale können die korrekte Identifizierung eines Proteins erheblich erschweren oder gar verhindern, wenn sie mit den Signalen der eigentlichen Zielpeptide überlappen, deren Ionisation unterdrücken oder zufällig bei der Proteinidentifizierung Datenbankpeptiden zugeordnet werden. Gleichzeitig stellen sie aber auch interessante Kandidaten für eine interne Kalibrierung dar. Dem Autor der vorliegenden Arbeit ist nur ein dokumentierter Ansatz bekannt (siehe [30], Stichwort „ScoreBooster“), der diese Signale systematisch in größeren Datensätzen erfasst, zur Kalibrierung benutzt, und anschließend aus der Signalliste streicht.

Dies gelingt, da die in [30] beschriebene Methode zur Spektrenkalibrierung dazu in der Lage ist sich bei der Datengewinnung im Hochdurchsatz dynamisch der jeweils vorliegenden Datensituation anzupassen. Manuelle Kalibrierungen durch einen erfahrenen Benutzer sind aber auch heutzutage immer erforderlich. Allerdings ist zu erwarten, dass die Proteinidentifizierungsraten auf Basis von PMF-Spektren bei automatischer Kalibrierung durch Verbesserung der heutigen Algorithmen erheblich gesteigert werden können.

4.2.2 Interpretation von Peptidmassenspektren

Um Proteine an Hand von Massenspektren, welche Ergebnis eines spezifischen proteolytischen Verdaus sind, zu identifizieren, werden Suchen in Proteinsequenzdatenbank durchgeführt. Hierfür verwendet man in der Praxis verschiedene Computeralgorithmen (PMF-Suchmaschinen), die letzten Endes alle auf dem gleichen Grundkonzept basieren [10, 13, 45, 46, 47]. Zunächst werden sämtliche in Frage kommenden Proteine einer Datenbank einem *in silico*-Verdau, auch theoretische Proteolyse genannt, gleicher Spezifität unterworfen. Aus den so entstandenen Peptiden wird für jeden Sequenzdatenbankeintrag ein theoretisches Massenspektrum erzeugt. Der Grad der Ähnlichkeit zwischen dem gemessenen Spektrum und den theoretischen Spektren wird bewertet und derjenige Datenbankeintrag, der die größte Ähnlichkeit zu dem gemessenen Spektrum besitzt, ist mit größter Wahrscheinlichkeit der korrekte Treffer (siehe Abbildung 4.2). Normalerweise erlauben PMF-Suchmaschinen das Treffen einer Vorauswahl bzgl. der Datenbankeinträge, die sich an dem Molekulargewicht, dem isoelektrischem Punkt oder taxonomischer Klassifizierungen orientiert.

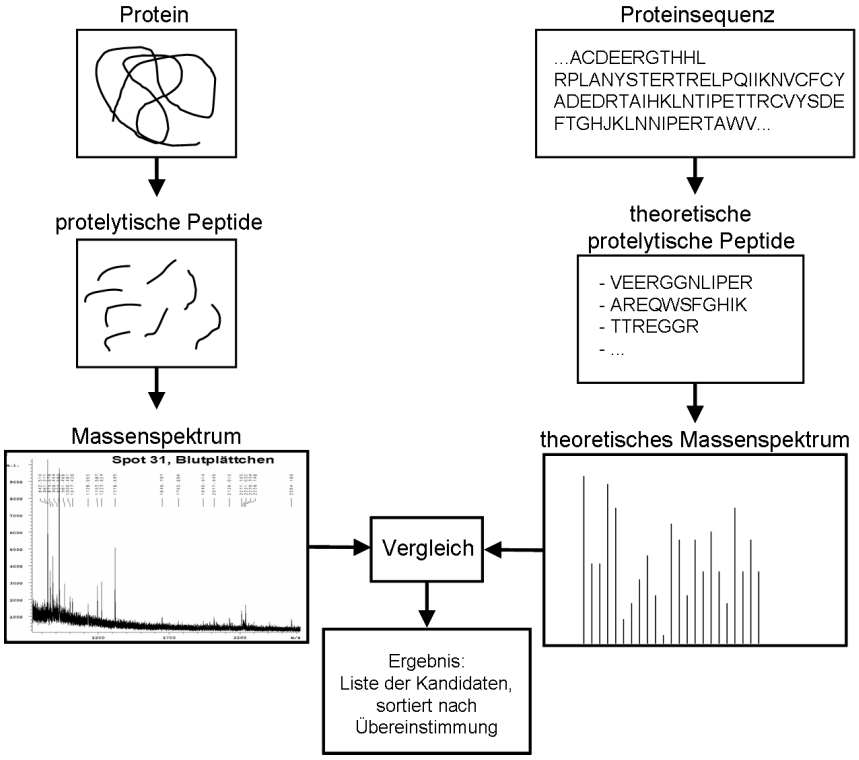

Abbildung 4.2: Schematische Darstellung der Arbeitsweise von Software zur massenspektrometrischen Proteinidentifizierung mittels Sequenzdatenbanken. Die Algorithmen generieren zu allen Proteineinträgen aus der Datenbank ein theoretisches Massenspektrum, das dann mit dem tatsächlich gemessenen Massenspektrum verglichen wird. Quelle: [30]

Um eindeutige Identifikationen erzielen zu können, wird eine gute Massengenauigkeit benötigt. Die Proteinidentifizierung konnte in den letzten Jahren durch technische Verbesserungen auf dem Gebiet der Massenspektrometrie deutlich verbessert werden. Nur mit Hilfe dieses Forschritts ist es heute noch möglich in den stark angewachsenen Sequenzdatenbanken signifikant Proteine identifizieren zu können. Durch die erhöhte Massengenauigkeit werden für eine eindeutige Proteinidentifikation insgesamt weniger gemessene Peptidmassen benötigt. Ein zweiter wichtiger Faktor ist der Grad der Spezifität der durchgeführten Pro-

teolyse. Proteasen wie das zumeist verwendete Trypsin schneiden auch unspezifisch oder nicht vollständig (so genannte übersprungene Schnittstellen), was die Zuordnung gemessener Spektren zu Datenbankeinträgen erschwert.

Da es für jedes gemessene Signal eine gewisse statistische Wahrscheinlichkeit gibt, mit der es zufälliger Weise mit einer aus einer Datenbank theoretisch berechneten Peptidmasse übereinstimmt, unterliegt der gesamte Prozess der Proteinidentifizierung auf der Basis von gemessenen Peptidmassen einer bestimmten Zufallswahrscheinlichkeit. Somit bleibt das stete Risiko einer falsch positiven Identifizierung. Wie leistungsfähig ein Proteinidentifizierungsalgorithmus ist, hängt somit also nicht nur von der Anzahl der gelungenen Proteinidentifizierungen ab, sondern auch davon, ob er dazu in der Lage ist falsch positive und richtig positive Treffer zu unterscheiden.

Die simpelsten und zugleich ältesten Algorithmen [10, 45, 46, 47] führen die Proteinidentifikation auf der Basis einer einfachen Sortierung der Sequenzdatenbankeinträge gemäß der Anzahl der übereinstimmenden Peptidmassen zwischen den theoretischen und dem gemessenen Massenspektrum durch. Der so genannte MOWSE-Algorithmus [13], wobei MOWSE für „Molecular Weight Search“ steht, benutzt zusätzlich dazu die Häufigkeitsverteilung der Peptidmassen in Sequenzdatenbanken, wodurch die Signifikanz der Ergebnisse erheblich gesteigert werden konnte. Der MOWSE-Algorithmus ist Bestanteil der Suchmaschinen MS-Fit [12] und Mascot [71]. Während der Entwicklung von Mascot wurde der MOWSE-Algorithmus zu einer wahrscheinlichkeitsbasierten Bewertung der Sequenzdatenbankeinträgen erweitert. ProFound [15] und Phenyx [16, 17, 18] sind die im Hinblick auf die ihnen zugrunde liegende Wahrscheinlichkeitstheorie die wohl am weitesten entwickelten Algorithmen zur Proteinidentifizierung anhand von Peptidmassenspektren. Sie wenden Bayesische Wahrscheinlichkeitstheorie für Berechnung der Wahrscheinlichkeit eines passenden Sequenzdatenbankeintrag an. Anhand der Bayesischen Wahrscheinlichkeitsberechnung lassen sich spezifische Eigenschaften der Peptidsequenzen bewerten. Zudem lassen sich experimentell ermittelte Zusatzinformationen in die Wahrscheinlichkeitsberechnungen mit einbeziehen [30]. Die bereits in Abschnitt 1.2 erwähnte Proteinidentifikationssoftware Peakardt [19, 20, 21] stellt in diesem Kontext einen Sonderfall dar, da sie die Generierung der theoretischen Massenspektren mit Hilfe eines evolutionären Algorithmus bewerkstelligt. Dieser erzeugt zu Beginn zufällig ausgewürfelte Peptidsequenzen, die mit Hilfe einer evolutionären Strategie über mehrere Generationen hinweg optimiert werden und vergleicht die resultierenden Lösungen anschließend mit den gemessenen Spektren. Die angewendete evolutionäre Optimierungstrategie greift dabei auf Standardoperationen wie Mutation, Rekombination und Selektion zurück, um theoretische Peptidsequenzen mit optimalem Fitnesswert zu erzeugen.

4.2.3 Interpretation von Peptidfragmentspektren

Analog zu Peptidmassenspektren lassen sich Peptidfragmentspektren ebenfalls mittels automatischer Datenbanksuchen identifizieren (PFF-Suchmaschinen). Das bereits 1994 der Öffentlichkeit vorgestellte Programm Sequest [11, 72], welches die erste PFF-Suchmaschine auf dem Markt war, ist auch heute noch eine der am weitesten verbreiteten Suchmaschinen zur automatischen Interpretation von Peptidfragmentspektren. Nach der Durchführung eines theoretischen Verdaus sämtlicher Proteinsequenzen aus der Datenbank, werden die erzeugten theoretischen Peptide, deren Masse zu der Masse des fragmentierten Peptides passt, diesem zugeordnet. Für diese ausgewählten Peptide wird ein theoretisches Fragmentspektrum generiert. Die hieraus resultierenden theoretischen Fragmentspektren werden mit dem gemessenen Spektrum verglichen, und anhand eines Punktesystems (Preliminary Score) bewertet. Die fünfhundert besten theoretischen Massenspektren, also die mit dem höchsten Preliminary Score, werden mittels Fourier-Transformation (Fast Fourier Transformation, FFT) mit dem gemessenen Massenspektrum per Kreuzkorrelation verglichen. Als Ergebnis nennt Sequest die Peptide und die zugehörigen Datenbankproteine mit den höchsten Kreuzkorrelationswerten. Auch Mascot ist dazu in der Lage Peptidfragmentspektren identifizieren zu können, hierzu verwendet Mascot die gleiche wahrscheinlichkeitsbasierte Bewertung wie für die Proteinidentifikation auf der Basis von Peptidmassenspektren.

Die Proteinidentifikation mittels Peptidfragmentspektren funktioniert nur, falls die fragmentierten Peptide in einer Sequenzdatenbank enthalten sind. Enthält das zu identifizierende Protein post-translationale Modifikationen, ist seine Primärstruktur durch alternatives Spleißen bedingt oder ist der zu dem Protein gehörige Organismus noch nicht vollständig sequenziert worden, verbleibt nur die *de novo*-Sequenzierung. Fast alle neueren *de novo*-Sequenzieralgorithmen greifen auf so genannte Spektrumgraphen, welche das gemessene Spektrum repräsentieren, zurück. In einem Spektrumgraph werden die in den Spektren enthaltenen Signale als Vektoren dargestellt. Die Massenunterschiede zwischen diesen Vektoren werden als

Knotenpunkte repräsentieren. Aufgrund der Bewertungen der einzelnen Knoten versucht der Algorithmus einen jeweils optimalen Pfad durch den Spektrumgraph zu finden. Dies wird allgemein hin als lokales Verfahren bezeichnet. Bei den so genannten globalen Verfahren werden alle theoretischen Spektren berechnet und bewertet. Diese Verfahren haben sich aufgrund der kombinatorischen Vielfalt der möglichen Lösungen als zu aufwendig erwiesen. Da die Fragmentierung der Peptide in der Regel nur unvollständig erfolgt, und die gemessenen Massenspektren Hintergrundrauschen enthalten können, ist die Anwendung eines automatischen Algorithmus für die *de novo*-Sequenzierung oft schwierig oder gar nicht möglich. Da zudem nur selten die gesamte Aminosäuresequenz eines zu analysierenden Peptides durch die im Massenspektrum enthalten Signale erklärt werden kann, müssen die berechneten Ergebnisse in fast allen Fällen einer aufwendigen manuellen Interpretation unterzogen werden. Dies führt dazu, dass die Anwendung von *de novo*-Algorithmen soweit es geht vermieden wird [30].

4.3 Probleme der datenbankgestützten Interpretation von MS- und MS/MS-Daten

Im Falle von co- und post-translationalen Modifikationen, dem Auftreten von unspezifischen Schnittstellen der Protease, übersprungenen Schnittstellen oder Sequenzfehlern wie z.B. Aminosäuresubstitutionen stellt die Auswertung von Peptidfragmentspektren noch immer eine nicht zu vernachlässigende Herausforderung dar, die oft nur mit Hilfe von erheblichem manuellen Aufwand zu meistern ist. Oft ist eine Identifizierung mittels Datenbanksuchen nicht möglich, da die Algorithmen in diesen Fällen nicht die richtigen theoretischen Fragmentmassenspektren aus dem entsprechenden Sequenzdatenbankeinträgen generieren. Suchmaschinen wie Sequest oder Mascot sind zwar in der Lage einige wenige Modifikationen oder Substratunspezifitäten von Enzymen bei der Generierung theoretischer Fragmentspektren zu berücksichtigen, jedoch führt dies in der Regel dazu, dass die Signifikanz der berechneten Ergebnisse aufgrund der Vielzahl an Kombinationsmöglichkeiten stark herabgesenkt ist. Zudem wächst die Anzahl der zu erzeugenden theoretischen Fragmentspektren quadratisch im Verhältnis zur Anzahl der gesuchten Modifikationen. Wie bereits oben erwähnt, besitzen große Sequenzdatenbanken wie die NCBI non-redundant derzeit mehr als 3.4 Mio. Einträge, was ca. einer Anzahl von 20^6 Peptiden bei einem gewöhnlichem tryptischen Verdau entspricht. Selbst wenn man im Durchschnitt nur drei Modifikationen pro Peptide zulässt, und nicht berücksichtigt, dass in der Literatur mehr als zweihundert unterschiedliche co- und post-translationale Modifikationen beschrieben werden [73], bedeutet die Berechnung der zugehörigen theoretischen Massenspektren bereits einen erheblichen Zeitaufwand. Um sämtliche möglichen Modifikationen und deren Kombinationen für ein einziges Peptid zu berücksichtigen, müssten 20^6 theoretische Fragmentspektren generiert werden. Wendet man dieses vorgehen auf sämtliche Sequenzeinträge einer Datenbank an, müssten damit zusammengenommen insgesamt 20^{12} theoretische Fragmentspektren berechnet werden. Lässt man den damit verbundenen Zeitaufwand einmal außer Acht, würde die sich allein aus der Statistik ergebende große Anzahl zufälliger Treffer das Ergebnis eines solchen Ansatzes nicht mehr interpretierbar machen. Zurzeit gibt es keine verfügbare Software, die dieses kombinatorische Problem löst, und dem Benutzer ein hochdurchsatzkompatibles System zur automatischen und globalen Identifikation von unerklärten Spektren zur Verfügung stellt. Sämtlichen bisher existierenden Ansätzen (FindPept [74], „Mutation tolerant search“ [75, 76] und „Mascot error tolerant search“ [77]) mangelt es an Hochdurchsatzkompatibilität und geeigneten Visualisierungen der komplexen Ergebnisse. Zudem kann nur eine bestimmte Auswahl unterschiedlicher Modifikationen erkannt werden.

Ein ganz grundsätzliches Problem der massenspektrometrischen Charakterisierung von Proteinen und Peptiden durch PMF- und PFF-Algorithmen ist es, dass die gemessenen Signale zufällig zu einer theoretischen Peptidsequenz passen können, was dazu führt, dass es generell eine gewisse Wahrscheinlichkeit für falsch positive Zuordnungen gibt. Faktoren wie die Größe der benutzten Datenbank, Enzymspezifität (der Grad zu dem ein Enzym, ein Protein ausschließlich gemäß seiner definierten Schnittstellen schneidet) Rauschen, Kontaminationen, Massengenauigkeit, Sequenzabdeckung im Spektrum oder Komplexität der Probe beeinflussen diese Wahrscheinlichkeit.

Darüber hinaus besteht die Problematik, dass die genaue Vorhersage eines Massenspektrums anhand von Peptid- oder Proteinsequenzen, wie es bei der Proteinidentifikation auf Basis von Proteindatenbankeinträgen geschieht, äußerst schwierig ist, da bis heute nicht sämtliche im Inneren eines Massenspektrometers ablaufenden physikalischen und chemischen Prozess vollständig aufgeklärt sind. Vor nicht all zu langer Zeit sind zwar viel versprechende Ansätze für die Generation theoretischer Spektren publiziert worden

[78, 79, 80, 81], aber sowohl der Zeitaufwand der notwendigen Berechnungen als auch die mangelnde Übertragbarkeit auf andere, als die von den Autoren genannten, experimentelle Bedingungen lassen eine routinemäßige Nutzung dieser Erkenntnisse noch nicht zu.

Aufgrund der eben geschilderten Probleme und Schwierigkeiten der datenbankgestützten Proteinidentifikation ergibt sich zwangsläufig die Notwendigkeit einen Proteinidentifikationsalgorithmus zu entwickeln, der dazu in der Lage ist, die Primärstruktur eines Proteins ohne Sequenzabgleiche mit den Einträgen einer Proteindatenbank zu ermitteln. Ein solcher *de novo*-Algorithmus benötigt als Datengrundlage hauptsächlich Sequenzinformationen, die durch MS- bzw. MS/MS-Analysen des zu identifizierenden Proteins gewonnen werden können, eben genau die Aminosäuresequenzen der Peptide aus denen das zu identifizierende Protein besteht. Auf der Basis dieser Informationen ist er dazu in der Lage, trotz co- und post-translationaler Modifikationen, dem Auftreten von unspezifischen Schnittstellen bei den verwendeten Proteasen, trotz übersprungener Schnittstellen oder Sequenzfehlern, die Aminosäuresequenz des zu identifizierenden Proteins zu bestimmen.

Kapitel 5

Anforderungsdefinition und -analyse

Dieses Kapitel beschreibt die funktionalen Anforderungen an einen *de novo*-Algorithmus für die Proteinidentifikation. Um die Anforderungen an einen solchen Algorithmus beschreiben zu können, muss zunächst das konzeptionelle Vorgehen unter Berücksichtigung der *de novo*-Eigenschaft beschrieben werden (siehe Abschnitt 5.1). In Abhängigkeit des gewählten Vorgehens, ergeben sich aus der zur Verfügung stehenden Datengrundlage (siehe Abschnitt 5.2) und denen im Allgemeinen mit der Identifikation von Proteinen verbundenen Problemen (siehe Abschnitt 5.3) funktionale Anforderungen an einen *de novo*-Algorithmus. Unter Berücksichtigung des allgemeinen Vorgehens bei der *de novo*-Proteinidentifikation, der zugehörigen Datengrundlage und sämtlicher damit verbundener Probleme, lassen sich die funktionalen Anforderung an einen *de novo*-Algorithmus in einer formalen mathematischen Problemdefinition zusammenfassen (siehe Abschnitt 5.4).

5.1 Vorgehen des *de novo*-Ansatzes

Die Idee des *de novo*-Ansatzes für die Proteinidentifikation ist es, ausschließlich Sequenzinformationen für die Identifikation von Proteinen zu verwenden. Diese Sequenzinformationen, entstammen unmittelbar der massenspektrometrischen Analyse der zu untersuchenden Biomoleküle und den dabei entstandenen Beobachtungen und Erkenntnissen. Aufbauend auf den am MPC gesammelten Erfahrungen, ergibt sich das durch Abbildung 5.1 beschriebene konzeptionelle Vorgehen.

Ausgangspunkt für den Identifikationsprozess ist ein zu identifizierendes Protein mit unbekannter Aminosäuresequenz. Dieses befindet sich im Idealfall zusammen mit anderen Proteinen identischer Primärstruktur in einer Probe. Diese Probe wird, wie bereits in Abschnitt 3.1.3 beschrieben, der spezifischen Proteolyse unterworfen. Allerdings erfordert der hier beschriebene *de novo*-Ansatz im Unterschied zur datenbankgestützten Proteinidentifikation, dass die in der Probe enthaltenen Proteine mit mehreren unterschiedlichen Proteasen verdaut werden. Da die unterschiedlichen Proteasen unterschiedliche Substratspezifitäten besitzen (siehe Tabelle 3.1), und damit Proteine bezüglich unterschiedlicher Aminosäuren schneiden, verfügen die so entstandenen Peptide über gemeinsame Subsequenzen. Diese gemeinsamen Subsequenzen sind der Proteinidentifikation dienlich, falls sie in der Form von N- und C-terminalen Überlappungen mit anderen Peptiden auftreten. Solche N- bzw. C-terminale Überlappungen zwischen unterschiedlichen Peptiden entsprechen auf der Darstellungsebene von Aminosäurensequenzen, gemeinsamen Präfixen bzw. Suffixen zwischen den unterschiedlichen Peptiden. Wurde zu Begin eine geeignete Auswahl an Proteasen getroffen — die Anzahl und konkrete Auswahl der zu verwendenden Proteasen hängt von der Aminosäuresequenz des zu identifizierenden Proteins ab und lässt sich, da die Proteinsequenz ja zu Beginn unbekannt ist, nur durch ausreichend Erfahrung und theoretische Durchschnittsanalysen abschätzen — und besitzen die ausgewählten Proteasen einen ausreichend hohen Grad an Enzymspezifizität, so lassen sich die Suffix-Präfix-Übereinstimmungen zwischen den unterschiedlichen Peptiden, nach Identifikation der Primärstruktur sämtlicher Peptide (siehe Abschnitte 3.1.6 und 3.1.7), effizient berechnen und für die Erzeugung eines so genannten Peptid-Layouts ausnutzen. An Hand eines solchen Layouts lässt sich dann auf die ursprüngliche Aminosäuresequenz des zu identifizierenden Proteins schließen.

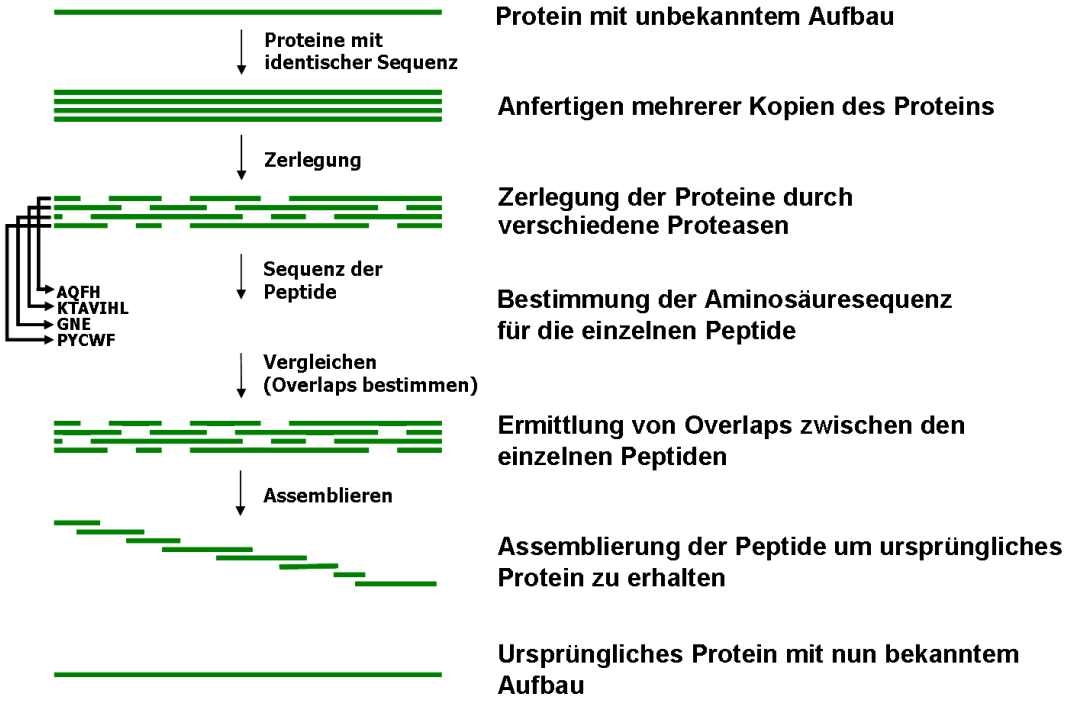

Abbildung 5.1: Schematische Darstellung des Ablaufs einer Proteinidentifikation gemäß des *de novo*-Ansatzes.

5.2 Nutzbare Datengrundlage

Praktisch alle der in Abschnitt 4.3 geschilderten Probleme bei der maschinellen Proteinidentifikation, sind auf den Einsatz von Proteindatenbanken zurück zu führen. Da der zu entwickelnde Algorithmus dazu in der Lage sein soll, die konzeptionellen Probleme der datenbankgestützten Proteinidentifikationsalgorithmen zu überwinden (siehe Abschnitt 4.3), bleiben als mögliche Eingabedatenquellen nur noch der enzymatische Verdau der Proteine und die anschließende massenspektrometrische Analyse, der durch den Verdau entstandenen Peptide. Beschränkt man sich bzgl. der Eingabe des Algorithmus auf die durch enzymatischen Verdau und Massenspektrometrie präzise bestimmbaren Eigenschaften des zu identifizierenden Proteins, so setzt sich die Eingabe aus denen in Abschnitt 5.2.1 bis Abschnitt 5.2.6 beschriebenen Kenngrößen zusammen.

5.2.1 Masse des zu identifizierenden Proteins

Die Masse des zu identifizierenden Proteins, im Folgenden als m_p bezeichnet, lässt sich mit Hilfe der Massenspektrometrie relativ genau bestimmten. Leider enthalten die zu analysierenden Proben nur selten ausschließlich das zu identifizierende Protein, sondern häufig mehr oder weniger komplexe Proteingemische, die noch andere eigentlich störende Proteine enthalten. Diese Proteine können bzgl. ihrer Primärstruktur identisch oder völlig verschieden zu dem zu untersuchenden Protein sein. Im Allgemeinen stellt dies aber in Bezug auf die massenspektrometrische Analyse der Probe kein großes Problem dar, da sich solche Proben mit vertretbarem Aufwand (siehe Abschnitt 3.1.2) bzgl. ihrer Bestandteile aufreinigen lassen.

Der Grad der Exaktheit mit der die Proteinmasse des zu identifizierenden Proteins bestimmt werden kann, hängt von dem physikalischen Auflösungsvermögen des verwendeten Massenspektrometers ab. Da eine allgemein gültige Obergrenze für die auftretende maximale Massenabweichungen nicht fest vorgegeben werden kann und die zudem durch technische Verbesserungen kontinuierlich weiter sinkt, muss die zu berücksichtigende Massentoleranz variabel gehalten werden. Im Folgenden beschreibt m_{diff}, den zu einem verwendeten Massenspektrometer gehörigen Wert der maximalen Massenabweichung.

5.2.2 Aminosäuresequenzen der identifizierten Peptide

Wie bereits in Kapitel Drei angedeutet (siehe Abschnitte 3.1.6 und 3.1.7) und in Kapitel Vier (siehe Abschnitt 4.2) ausführlich beschrieben wurde, lässt sich die Primärstruktur einzelner Peptide mit Hilfe der Massenspektrometrie bestimmen. Wie in Abschnitt 5.3.1 beschrieben, kann die Bestimmung der Aminosäuresequenz eines Peptides durch gewisse physikalische und chemische Prozesse erschwert, bzw. verfälscht werden. Dies muss bei der späteren Rekonstruktion des ursprünglichen Proteins berücksichtigt werden.

5.2.3 Massen der identifizierten Peptide

Die Massen der identifizierten Peptide lassen sich nach Bestimmung ihrer Aminosäuresequenz (siehe Abschnitte 3.1.6 und 3.1.7) aus den Massen (siehe Tabelle 5.1) ihrer Aminosäuren berechnen.

Aminosäure	**Einbuchstaben-Code**	**Monoisotopische Masse**	**Average-Masse**
Alanin	A	71.03711	71.0788
Arginin	R	156.10111	156.1875
Asparagin	N	114.04293	114.1039
Asparginsäure	D	115.02694	115.0886
Cystein	C	103.00919	103.1388
Glutaminesäure	E	129.04259	129.1155
Glutamin	Q	128.05858	128.1307
Glycin	G	57.02146	57.0519
Histidin	H	137.05891	137.1411
Isoleucin	I	113.08406	113.1594
Leucin	L	113.08406	113.1594
Lysin	K	128.09496	128.1741
Methionin	M	131.04049	131.1926
Phenylalanin	F	147.06841	147.1766
Prolin	P	97.05276	97.1167
Serin	S	87.03203	87.0782
Threonin	T	101.04768	101.1051
Tryptophan	W	186.07931	186.2132
Tyrosin	Y	163.06333	163.1760
Valin	V	99.06841	99.1326

Tabelle 5.1: Zusammenstellung der 20 proteinogenen Aminosäuren in Hinblick auf deren spezifische Massen. Zusätzlich zu dem Namen und dem Einbuchstaben-Code wird für jede Aminosäure auch ihre so genannte monoisotopische und ihre so genannte Average-Masse angegeben. Die monoisotopische Masse entspricht der Masse der Aminosäure, wenn sie Teil eines einfach geladenen Molekülions ist. Die Average-Masse leitet sich aus dem Durchschnittswert der Aminosäuremasse für jegliche bekannte Form von Molekülionbeteiligung ab.

5.2.4 Scores der identifizierten Peptide

Wie im Folgenden noch näher erläutert wird (siehe Abschnitt 5.3) unterliegt die Peptididentifikation einigen Beschränkungen und Problemen, dies hat zur Folge, dass es bzgl. der Identifikationsgüte einzelner Peptide qualitative Unterschiede gibt, die durch einen Score repräsentiert werden. In Bezug auf die identifizierten Peptide entspricht dieser Score einer reellen Zahl zwischen 0 und 1, die angibt wie exakt die Identifikation eines Peptides aufgrund der vorangegangenen massenspektrometrischen Untersuchungen durchgeführt werden konnte. Dieser Score wird den Peptiden während der Phase der Peptididentifikation zugeordnet (siehe Abschnitt 3.1.6).

5.2.5 Absolute Häufigkeiten der identifiziert Peptide

Bei der Identifikation der Primärstruktur der in der Probe enthaltenen Peptide kann es vorkommen, dass mehrere unterschiedliche Peptide die gleiche Aminosäuresequenz besitzen. Dies ist im Wesentlichen auf zwei Faktoren zurückzuführen.

Da die Gene höherer Lebewesen häufig fragmentiert sind (die proteinkodierenden Abschnitte eines Gens sind über gewisse Bereiche der DNS-Moleküle eines Lebewesens verteilt) oder eine sehr einfache repetitive Struktur besitzen, können Proteine bestimmte Aminosäuresequenzen mehrfach enthalten. Zudem beinhalten die zu untersuchenden Proben, wie in Abschnitt 5.2.1 bereits angedeutet, in der Regel mehrere Proteine was dazu führt, dass unterschiedliche identifizierte Peptide die gleiche Aminosäuresequenz besitzen können.

5.2.6 Überlappungen zwischen den Aminosäuresequenzen der identifizierten Peptide

Da der gesamte hier vorgestellte *de novo*-Proteinidentifikationansatz auf der Verwendung von mehreren Proteasen für den enzymatischen Verdau von Proteinen basiert, und jede der verwendeten Proteasen eine oder mehrere spezifische Schnittstellen besitzt, ergibt sich aus den verdauten Peptiden ein Überlappungsmuster, welches sich für die Identifikation der Primärstruktur des zu untersuchenden Proteins ausnutzen lässt.

5.3 Grundlegende Probleme der *de novo*-Proteinidentifikation

Die bereits zu Beginn (siehe Abschnitt 1.2) und in Abschnitt 4.3 beschriebenen Probleme der datenbankgestützten Proteinidentifikation lassen sich zwar durch den *de novo*-Ansatz umgehen, über diese Probleme hinaus ergeben sich aber noch andere Schwierigkeiten, die die korrekte Identifikation eines Proteins verhindern können. Diese Problemquellen ergeben sich aus der massenspektrometrischen Analyse von Biomolekülen und der dazu nötigen enzymatischen Spaltung dieser Biomoleküle und müssen bei der Entwicklung eines *de novo*-Proteinidentifikationsalgorithmus gelöst werden. Die sich direkt oder indirekt aus der massenspektrometrischen Proteinanalyse ergebenden Probleme werden in den Abschnitten 5.3.1 bis 5.3.6 charakterisiert.

5.3.1 Transpeptidierung

Unter dem Begriff der Transpeptidierung, auch als proteasekatalysierte Peptidsynthese bezeichnet (im Englischen „peptide rearrangement“ genannt), versteht man den Prozess der „zufälligen“ Peptid- oder Aminosäurekondensation nach proteolytischer Spaltung eines Proteins in Peptide. Dies bedeutet vereinfacht ausgedrückt, dass Peptide anschließend an die proteolytische Spaltung des Proteins, Peptidbindungen mit anderen Peptiden eingehen können und so Polypeptide entstehen, die auf Grund der Substratspezifizität der verwendeten Protease nicht vorhersagbar sind (siehe Abbildungen 5.2 und 5.3). Der Umfang in dem solche proteasekatalysierten Peptidsynthesen stattfinden, hängt von der Menge der für die enzymatischen Spaltung eingesetzten Protease und ihrer Einwirkzeit ab. Tendenziell gilt für den Verdau eines Proteins, dass die Verwendung einer großen Menge an Protease und/oder eine lange Einwirkzeit auf das Protein die Wahrscheinlichkeit für das Auftreten von Transpeptidierungen erhöhen.

MDVTIQHPWF KRALGPFYPS RLFDQFFGEG LFEYDLLPFL
SSTISPYYRQ SLFRTVLDSG ISEVRSDRDK FVIFLDVKHF
SPEDLTVKVL EDFVEIHGKH NERQDDHGYI SREFHRRYRL
PSNVDQSALS CSLSADGMLT FSGPKVQSGL DAGHSERAIP
VSREEKPSSA PSS

Abbildung 5.2: Aminosäuresequenz des Proteins Alpha-A-Crystallin aus der Augenlinse der Maus (*mus musculus*). Die rot markierte Subsequenz entspricht, dem in Abbildung 5.3 dargestellten Peptid.

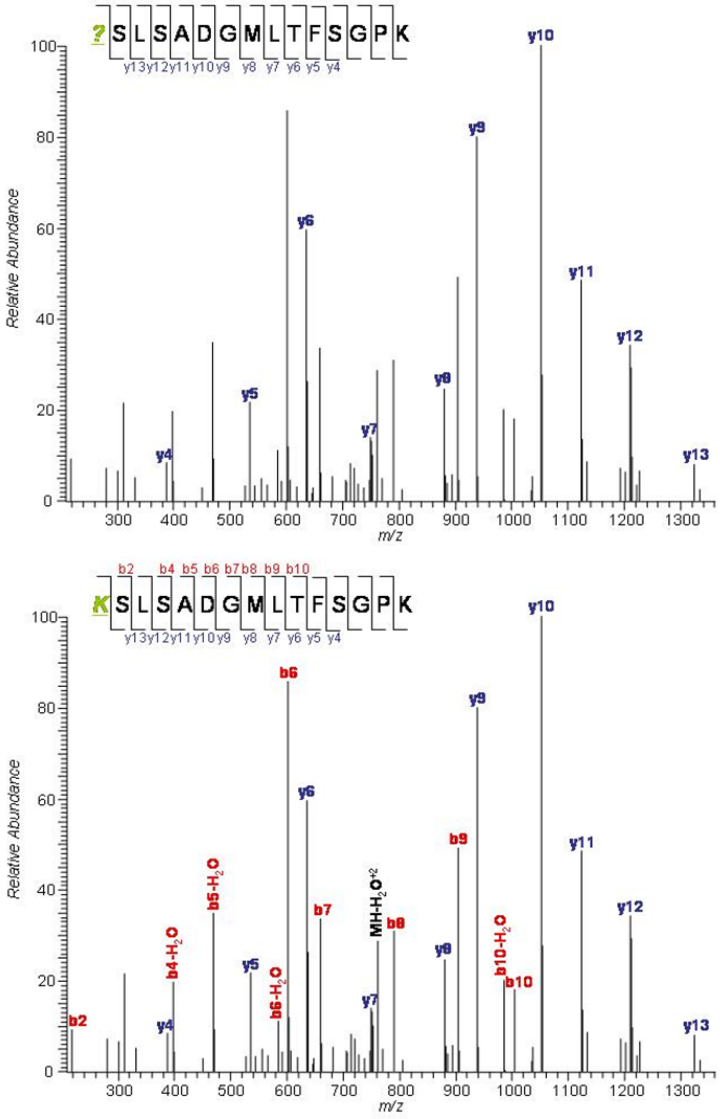

Abbildung 5.3: Fragmentmassenspektren und Sequenzen eines Peptides des Proteins Alpha-A-Crystallin ohne und anschließend mit Transpeptidierung durch die Aminosäure Lysin. Quelle: [35]

Das Phänomen der Transpeptidierung wurde ursprünglich bereits 1898 von van't Hoff beschrieben [82]. Er postulierte, dass Trypsin eine inherente Fähigkeit zur Proteinsynthese aus von ihr selbst gespalteten Segmenten haben muss. Vierzig Jahre später wurde die enzymatische Synthese mit dem Katalysator Chymotrypsin sowohl von Bergmann, als auch Fruton beschrieben [83, 84]. Bis in die siebziger Jahre des zwanzigsten Jahrhunderts war das Interesse für die reverse Proteolyse klein, dies änderte sich jedoch schlagartig als die Gruppen von Kullman [85, 86] und Isowa [87] mit Hilfe dieses Phänomens bioaktive Peptide synthetisiert hatten. Seitdem wird die synthetisierende Eigenschaft des Trypsins für die industrielle Konversion von Schweineinsulin in Humaninsulin benutzt [88].

Schon mehrfach wurde das Auftreten synthetischer Peptide bzw. von Aminosäuresequenzänderungen nach *in vitro*-Proteolyse mit Trypsin als Nebenprodukt beobachtet und auch massenspektrometrisch analysiert [35, 89, 90, 91, 92, 93]. Leider ist der Mechanismus hinter der proteasekatalysierte Peptidsynthese bis heute nur sehr unzureichend erforscht. Die im Zuge der Evolution der Proteinanalytik gewonnenen Erkenntnisse über dieses Phänomen beschränken sich im Wesentlichen darauf, dass man um das Auftreten von Peptidsynthesen in Verbindung mit bestimmten Proteasen weiß und diese auch durch aufwändige Einzelanalysen nachträglich nachweisen kann. Es gibt aber nach Kenntnis des Autors bis heute keine Publikation, die den der Transpeptidierung zugrunde liegenden Mechanismus erschöpfend beschreibt.

5.3.2 Mehrfachidentifikationen strukturell identischer Peptide

Da man in der Proteinanalytik fast ausschließlich mit Proteinproben arbeitet, die mehr als ein Protein enthalten, kann es zur mehrfachen Identifikation bestimmter Peptide mit identischer Primärstruktur kommen. Dies kann zum einen daran liegen, dass die zu untersuchende Probe ein Proteingemisch enthält in dem ein bestimmtes Protein mehrfach enthalten ist, oder zum anderen daran, dass die während des Verdaus entstehenden Peptide, aufgrund der Verwendeten Proteasen und der Primärstruktur der in der Probe enthaltenen Proteine, einfach tendenziell häufiger bestimmte Aminosäuresequenzen besitzen. Darüber kommt es auch vor, dass die zu untersuchenden Proteine repetitive Primärstrukturen aufweisen, sodass bestimmte Aminosäuresequenzen mehrfach innerhalb der Aminosäuresequenz eines Proteins auftreten können.

5.3.3 Sequenzüberdeckung durch identifizierte Peptide

Da der *de novo*-Ansatz nicht auf Proteindatenbanken und die darin verzeichneten identifizierten Proteine zurückgreift, benötigt er eine entsprechend hohe Ausbeute an massenspektrometrisch identifizierten Peptiden, um die Aminosäuresequenz des zu identifizierenden Proteins vollständig überdecken zu können. Nur wenn sämtliche Aminosäuren des ursprünglichen Proteins durch identifizierte Peptide erklärt und überdeckt werden können, lässt sich das ursprüngliche Protein rekonstruieren.

Die Peptidausbeute bei der Peptididentifikation per MS/MS-Analyse kann aber aus mehreren Gründen sehr gering ausfallen.

Technische Limitationen Die Proteine aus einem komplexen Proteingemisch zerfallen durch den proteolytischen Verdau nicht selten in mehrere hundert Peptide. Da die Aufnahme von Massenspektren während der MS/MS-Analyse einzelner Peptide aber nicht kontinuierlich, sondern zu diskreten Zeitpunkten stattfindet, werden mit unter auch viele Massenspektren erzeugt, die anschließend nicht zur Identifikation des untersuchten Peptides taugen. Zudem werden PFF-Spektren nur für die Peptide erzeugt, die eine ausreichend hohe Intensität aufweisen, sprich von denen das Datensystem des Massenspektrometers (siehe Abschnitt 3.1.5) ausgehen kann, dass das vorliegende Signal nicht auf Rauschen oder Zufall beruht.

Biochemische Limitation Je nach Gewebetyp und Zelllokation aus denen die zu identifizierenden Proteine stammen, sind die durch den Verdau entstehenden Peptide unterschiedlich gut durch Massenspektrometer zu analysieren. Proteine, die aus der Zellmembran einer Zelle stammen, sind lipophil (siehe Kapitel 3) und daher schlecht wasserlöslich. Dies hat Auswirkungen auf die Peptidausbeute bei der Analyse, da sich solche lipophilen Peptide oft nur schlecht, manchmal gar nicht ionisieren lassen und sich damit der eigentlichen Analyse entziehen.

Physikalische Limitationen Das Auflösungsvermögen der heute standardmäßig eingesetzten Massenspektrometer hat sich über die Jahre kontinuierlich verbessert. Nichtsdestotrotz ist es auch heute noch auf ein bestimmtes Massenfenster beschränkt. Ionisierte Peptide, deren Masse kleiner als 500 oder größer als 8000 Dalton ist, können im Allgemeinen nicht korrekt detektiert werden. Das Massenfenster in dem sich die analysierbaren Peptide bewegen, lässt sich zwar durch Anpassung der Geräteeinstellungen zu einem gewissen Grad nach oben hin verschieben, dies sorgt dann aber dafür, dass sich die untere Massengrenze der detektierbaren Peptide ebenfalls nach oben verschiebt. Nach unten hin ist das Auflösungsvermögen eines Massenspektrometers durch die physikalischen Grundlagen, auf denen sein Detektor beruht, beschränkt.

Je nachdem wie viele der oben genannten Beschränkungen gleichzeitig zum Tragen kommen, kann der Anteil der in einem einzigen Lauf einer MS-Analyse per Datenbankabgleich sinnvoll erklärbaren PMF-Spektren bei 30 bis 40% liegen. Dies hat in direkter Konsequenz Auswirkungen auf die Peptidausbeute und damit auf die Anzahl der unterschiedlichen Peptide, die anschließend an die Erzeugung der MS-Spektren durch eine MS/MS-Analyse identifiziert werden können. In letzter Konsequenz führt eine geringe Anzahl an identifizierbaren Peptiden zu Lücken in der Gesamtsequenz des zu rekonstruierenden Proteins.

5.3.4 Peptide mit geringem Score

Zusätzlich zu dem quantitativen Problem, der unter Umständen geringen Peptidausbeute während der Peptididentifikation, besteht noch das qualitative Problem der Güte mit der ein Peptid identifiziert wurde. In Abhängigkeit von der Qualität der PMF-Massenspektren, die den PFF-Massenspektren im Zuge der Peptideidentifikation vorausgingen, sowie prinzipiell sämtlicher voraus gegangener Analyseschritte des gesamten Identifikationsprozesses, erfolgt die Peptididentifikation mehr oder weniger verlässlich.

5.3.5 Probenkontamination

Da die zu untersuchenden Proben in der Regel Proteingemische sind, können diese auch potentielle Verunreinigungen enthalten. Dazu zählen Bestandteile von Proteinen, wie Keratin aus der Haut und den Haaren eines Laboranten oder Restbestandteile des verwendeten proteolytischen Verdauungsenzyms (z.B. Trypsin, Glu-C, Lys-C, usw.).

5.3.6 Eindeutigkeit der berechneten Peptid-Layouts

Wie bereits in Abschnitt 5.1 beschrieben, muss für die Identifikation eines Proteins gemäß *de novo*-Ansatzes ein Peptid-Layout berechnet werden, anhand dessen die Aminosäuresequenz des ursprünglichen Proteins rekonstruiert werden kann. In Abhängigkeit davon, wie schwerwiegend die Problemfaktoren aus den Abschnitten 5.3.1 bis 5.3.5 bei der Rekonstruktion des ursprünglichen Proteins zum Tragen kommen, wird die Bestimmung eines solchen Layouts erschwert.

Die Berechnung eines korrekten Peptid-Layouts wird zusätzlich dadurch erschwert, dass es aufgrund der oben genannten Problemfaktoren mehr als ein mögliches Peptid-Layout zu jedem zu rekonstruierenden Protein geben kann. Transpeptidierungen (siehe Abschnitt 5.3.1) sorgen dafür, dass Peptide nicht mit ihrer eigentlich zu erwartenden Primärstruktur identifiziert werden oder das zwei oder mehr Formen ein und des selben Peptides identifiziert werden und diese anschließend bei der Rekonstruktion des ursprünglichen Proteins berücksichtigt werden müssen. Mehrfachidentifikationen strukturell identischer Peptide (siehe Abschnitt 5.3.3) und Probenkontaminationen (siehe Abschnitt 5.3.5) erzeugen recht ähnliche Probleme.

Konnte während der Peptididentifikation eine nur geringe Ausbeute an identifizierten Peptiden erzielt werden (siehe Abschnitt 5.3.3) oder konnte man nur wenige Peptide mit hoher Wahrscheinlichkeit korrekt identifizieren (siehe Abschnitt 5.3.4), so erhält man unter Umständen nicht genügend Sequenzinformation, um ein vollständiges Peptid-Layout zu berechnen und kann in Folge dessen auch nicht die komplette Aminosäuresequenz des ursprünglichen Proteins rekonstruieren.

5.4 Problemdefinition

Nachdem das grundlegende Vorgehen bei der *de novo*-Proteinidentifikation (siehe Abschnitt 5.1), die dazugehörige Datengrundlage (siehe Abschnitt 5.2) und die damit verbundenen Probleme (siehe Abschnitt 5.3) beschrieben wurden, kann nun das eigentliche der *de novo*-Proteinidentifikation zugrunde liegende Problem formal definiert werden. Diese formale Problemdefinition fasst sämtliche funktionalen Anforderungen an das Peptide-Assembly-Problem zusammen. Ein *de novo*-Algorithmus, der dieser Problemdefinition entspricht, wird auch sämtliche formalen Anforderungen an den hier beschriebenen Ansatz der *de novo*-Proteinidentifikation erfüllen. Um dies zu erreichen, wird das Peptide-Assembly-Problem als Maximum-Likelihood-Problem formuliert. Diese Formulierung des vorliegenden Problems basiert auf der Arbeit von Eugene W. Myers [94].

5.4.1 Das Peptide-Assembly-Problem

Definition 5.4.1 *Gegeben seien F, die Multimenge der identifizierten Peptide und die Abweichungsrate $0 \leq \epsilon < 1$. Finde eine Rekonstruktion R und ein gültiges ϵ-Layout dessen beobachtete Peptide-Startpunkt-Verteilung D_{obs} eine minimale Abweichung δ zu der tatsächlichen Peptidstartpunktverteilung D_{src} aufweist.*

Der erste Teil der Definition 5.4.1 beschreibt eine Lösung des Peptide-Assembly-Problems als eine Kombination aus einem Rekonstruktionsstring R, der die Aminosäuresequenz des rekonstruierten Proteins repräsentiert und dem dazugehörigen so genannten ϵ-Layout (siehe Abbildung 5.4). Dieses Layout wird durch eine Menge von $|F|$ vielen Paaren von positiven ganzen Zahlen $(s_i, e_i)_{i\in[1,|F|]}$, mit $1 \leq s_i, e_i \leq |R|$ beschrieben, welche die Start- und Endposition der Peptide $p_i \in F$ in R angeben [94]. Damit beschreibt ein solches Layout die Verteilung der einzelnen, für die Rekonstruktion des zu identifizierenden Proteins verwendeten, Peptide in dem zugehörigen Peptidüberlappungsmuster.

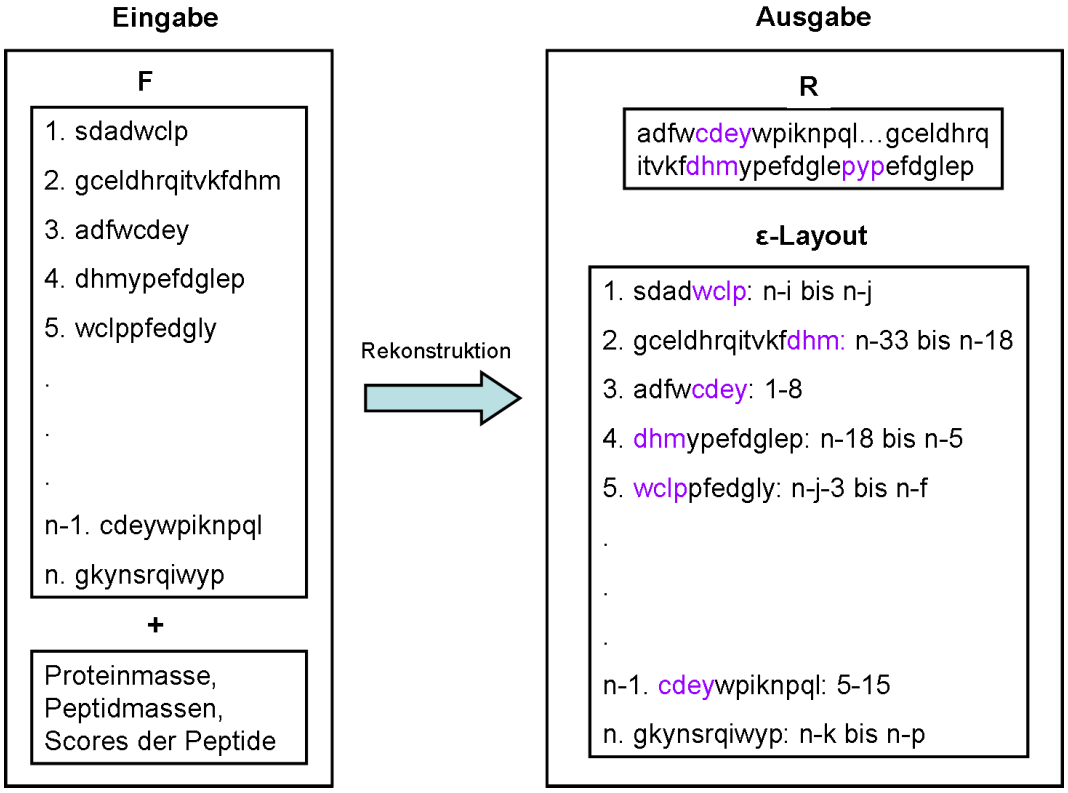

Abbildung 5.4: Schematische Gegenüberstellung von Ein- und Ausgabe eines Algorithmus für das Peptide-Assembly-Problem. Die Eingabe besteht aus F, der Menge sämtlicher identifizierter Peptide, deren Massen und Scores, sowie der Masse des zu rekonstruierenden Proteins. Diese Informationen werden zusammen mit den während des Rekonstruktionsprozesses ermittelten Überlappungen zwischen den einzelnen Peptiden für die Ermittlung einer Lösung für das Peptide-Assembly-Problem benutzt. Die Lösung wird durch den Rekonstruktionsstring R und das zugehörige ϵ-Layout repräsentiert. Das Layout gibt für jedes der Peptide aus F an, welche Position es in dem Rekonstruktionsstring R einnimmt. Dazu wir für jedes Peptid eine Start- und Endposition angegeben.

Ein wichtiges Merkmal eines Peptid-Layouts ist die Eigenschaft ϵ-gültig zu sein [94].

Definition 5.4.2 *Ein Layout heißt ϵ-gültig, falls es die folgenden beiden Bedingungen erfüllt:*

1. *Die Anzahl der Unterschiede zwischen der Aminosäurensequenz eines Peptides p_i und des ihm zugewiesenen Substrings aus R ist durch $\epsilon|p_i|$ beschränkt;*
2. *Die Masse des rekonstruierten Proteins m_{cur} darf die Masse des ursprünglichen Proteins m_p nicht um mehr als den Wert von m_{diff} übersteigen.*

Wie bereits in Abschnitt 5.2.1 definiert, beschreibt m_{diff} die maximale Massenabweichung, die bei der Bestimmung der Masse des zu identifizierenden Proteins auftritt. Der tatsächliche Wert von m_{diff} hängt dabei von dem verwendeten Massenspektrometer ab. Moderne Massenspektrometer erreichen bei geeigneter Gerätekonfiguratione eine maximale Massenabweichung $m_{diff} \leq 0.3$ Dalton.

Der zweite Teil der Definition 5.4.1 greift die in Abschnitt 5.3.6 bereits angedeutete Problematik auf, dass nicht notwendiger Weise immer eine Eins-zu-eins-Beziehung zwischen dem zu identifizierendem Protein und den hierfür zur Verfügung stehenden Ausgangsdaten besteht. Der Lösungsraum des Peptide-Assembly-Problems besteht daher aus der Menge aller gültigen ϵ-Layouts und das zu lösende Problem liegt in der Auswahl eines besten Layouts. Um später zwischen mehreren möglichen Layouts entscheiden zu können, wird für die konzeptionelle Realisierung des zweiten Teils der obigen Definition eine Fitnessfunktion auf Grundlage der so genannten Kolmogorov-Smirnov-Teststatistik [94, 95] entwickelt.

Um dies zu ermöglichen lässt sich zu jedem berechneten ϵ-Layout eine so genannte beobachtete Peptidstartpunktverteilung

$$D_{obs}(x) = |\{p_i : s_i = x\}|/|F|$$

ermitteln [94]. Der Definitionsbereich der zugehörigen Verteilungsfunktion $D_{obs}(x)$ entspricht dabei $1 \leq x \leq |R|$, wobei $|R|$ die Länge der rekonstruierten Proteinsequenz angibt [94].

Die beobachtete Peptidstartpunktverteilung D_{obs} eines berechneten gültige ϵ-Layouts, lässt sich bei bekannter sequentieller Anordnung der identifizierten Peptide leicht berechnen. Damit man anschließend ermitteln kann welches der im Lösungsraum enthaltenen ϵ-Layouts nun das bzw. eines der besten ist, vergleicht man die berechneten beobachteten Peptidstartpunktverteilungen mit der so genannten tatsächlichen Peptidstartpunktverteilung D_{src}. Dazu berechnet man die Abweichung zwischen beobachteter und tatsächlicher Peptidstartpunktverteilung [94]:

$$\delta = max_{1 \leq x \leq |R|} |D_{obs}(x) - D_{src}(x)|.$$

Die tatsächliche Peptidstartpunktverteilung lässt sich leider nicht so direkt und so einfach wie die beobachtete Peptidstartpunktverteilung berechnen. Um diese zu bestimmen muss auf biologisches Hintergrundwissen über die Peptidstartpunktverteilungen von bereits identifizierten Proteinen zurückgriffen werden. Dieses Wissen lässt sich durch den Einsatz eines so genannten theoretischen Verdaus und anschließender Proteinrekonstruktionen erschließen. Eine präzise Beschreibung der Berechnung von D_{src} erfolgt in Kapitel Sechs.

Kapitel 6

Implementierung

Nachdem im vorherigen Kapitel die Anforderungen an einen *de novo*-Proteinidentifikationsalgorithmus bestimmt und in einer formalen Problemdefinition zusammengefasst wurden (siehe Abschnit 5.4.1), erfolgt in diesem Kapitel die Beschreibung der zugehörigen Implementierung.

Die Struktur dieses Kapitels ergibt sich aus der Beschreibung der einzelnen Bestandteile des Gesamtalgorithmus, der hier für die *de novo*-Proteinidentifikation entwickelt werden soll, und der für diese Bestandteile zu entwickelnden Teillösungen. Um unnötigen Berechnungs-Overhead zu vermeiden, werden die Eingabedaten zu Beginn bzgl. redundanter Informationen, gefiltert (siehe Abschnitte 6.1, 6.2 und 6.3). Nach Filterung der Eingabe werden sämtliche für die weiteren Schritte essentiell wichtigen Überlappungen zwischen den identifizierten Peptiden bestimmt (siehe Abschnitt 6.4). Die Berechnung dieser Überlappungen kann wahlweise approximativ oder nicht-approximativ erfolgen. Auf Basis der berechneten Überlappungen wird der für die weiteren Rekonstruktionsschritte unverzichtbare Overlap-Graph G erstellt (siehe Abschnitt 6.5). Dieser dient bei den nachfolgenden Berechnungen als zentrale Datenstruktur. Nach der Beendigung der Overlap-Berechnungen werden die durch G repräsentierten peptidischen Überlappungsinformationen zunächst einmal aufbereitet (siehe Abschnitt 6.6) und anschließend in einem Rekonstruktionszwischenschritt so genannte Polypeptide, dies sind Substrukturen des eigentlich zu identifizierenden Proteins, rekonstruiert (siehe Abschnitt 6.7). In einem letzten Schritt werden sämtliche Rekonstruktionsmöglichkeiten für das zu identifizierende Protein ermittelt und für den Fall, dass es mehr als eine verbliebene Rekonstruktionsmöglichkeit gibt, bezüglich ihrer Lösungsgüte bewertet (siehe Abschnitt 6.8).

Der gesamte Rekonstruktionsprozess setzt sich daher also aus den folgenden Rekonstruktionsoperationen zusammen:

1. Filtern von Kontaminationen (siehe Abschnitt 6.1)
2. Filtern von Infixen (siehe Abschitt 6.2)
3. Behandlung von Transpeptidierungseffekten (siehe Abschnitt 6.3)
4. Overlap-Berechnung (siehe Abschnitt 6.4)
5. Generierung des Overlap-Graphen (siehe Abschnitt 6.5)
6. Aufbereitung des Overlap-Graphen (siehe Abschnitt 6.6)
7. Rekonstruktion der Polypeptide (siehe Abschnitt 6.7)
8. Ermittlung einer optimalen Rekonstruktion (siehe Abschnitt 6.8)

6.1 Filtern von Kontaminationen

Wie bereits in Abschnitt 5.3 bei der Auflistung der grundlegenden Probleme der *de novo*-Proteinidentifikation erwähnt wurde, können massenspektrometrisch untersuchte Proteinproben Verunreinigungen

enthalten. Solche Kontaminationen lassen sich mit Hilfe eines Sequenzabgleichs zwischen den Aminosäuresequenzen der identifizierten Peptide und einer so genannten Kontaminantenliste mit hoher Genauigkeit identifizieren und aus der Eingabe des Rekonstruktionsalgorithmus entfernen. Da der hier zu entwickelnde Algorithmus später Teil der Proteinidentifikationssoftware Peakardt werden soll und Peakardt bereits einen Mechanismus zum Filtern solcher Kontaminationen in Linearzeit bereitstellt, kann die Überprüfung der Eingabe mit Hilfe dieses Mechanismus erfolgen.

Peakardt bietet die Möglichkeit Kontaminationen aus Peptidmassenspektren an Hand von charakteristischen Peptidmassen herauszufiltern. Häufig auftretende Kontaminationsquellen, wie Keratin oder Restbestandteile von verwendeten Verdauenzymen, besitzen aufgrund ihrer spezifischen Primärstruktur und des jeweils verwendeten Verdauenzyms ein charakteristisches Peptidmuster. Dieses Peptidmuster entspricht, wie bereits in Abschnitt 3.1.6 diskutiert wurde, einem Fingerabdruck des als Kontamination enthaltenen Proteins. Die Peptidmassen eines solchen spezifischen Peptidmusters lassen sich daher als Anhaltspunkt für den Nachweis einer Probenkontamination verwenden. Da je nach Versuchsaufbau und Auswahl der verwendeten Chemikalien, mit denen eine zu untersuchende Probe in Berührung kommt, neue Arten von Probenkontaminationen auftreten können, ist der in Peakardt implementierte Mechanismus zum Filtern von Kontaminationen erweiterbar. Um Probenkontaminationen filtern zu können, verwaltet Peakardt eine Liste von Peptidmassen, die charakteristisch für bestimmte Kontaminationen sind. Diese Liste lässt sich durch neue Peptidmassen erweitern (siehe Abbildung 6.1).

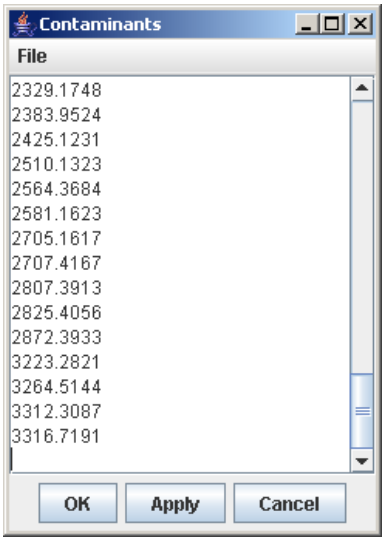

Abbildung 6.1: Screenshot des Dialogs zur Anpassung der in Peakardt enthaltenen Kontaminantenliste. Zusätzlich zu den bereits spezifizierten Kontaminanten lassen sich weitere durch Angabe ihrer spezifischen Peptidmassen angeben.

6.2 Filtern von Infixen

Um die Größe der Eingabe von vornherein auf ein absolutes Minimum zu reduzieren, lassen sich sämtliche Peptide, die Infix eines anderen Peptides sind, aus der Menge der identifizierten Peptide herausfiltern. Der Begriff Infix bezeichnet im Zusammenhang mit der in Kapitel 5 definierten Problemstellung ein Peptid, dass bzgl. seiner Aminosäuresequenz vollständig in der Aminosäuresequenz eines anderen Peptides als Subsequenz enthalten ist. Diese Maßnahme sorgt in der Regel, wie die in Kapitel Sieben zusammengefassten Testergebnisse zeigen werden, für eine durchaus bedeutsame Verkleinerung der Eingabegröße.

Beim Filtern der Infixe werden die Aminosäuresequenzen der identifizierten Peptide paarweise miteinander verglichen. Bei diesen paarweisen Vergleichen zweier Peptide p_i und p_j werden zwei Eigenschaften, die charakteristisch für Infixe sind, überprüft. Ein Peptid p_i ist genau dann Infix eines anderen Peptides p_j, falls die Länge der Aminosäuresequenz von p_i kleiner der Länge der Aminosäuresequenz von p_j ist und p_i Subsequenz der Aminosäuresequenz von p_j ist. Sind beide Bedingungen erfüllt, kann das jeweilige Peptide p_i aus der Eingabe entfernt werden. Um zu vermeiden, dass Peptide die eine identische Primärstruktur besitzen und mehrfach identifiziert wurden, aus der Eingabe herausgefiltert werden, wird

eine Überprüfung der Sequenzlängen vorgenommen. Ansonsten würde eine der formalen Anforderungen an die Problemdefinition aus Kapitel 5 verletz werden (siehe Abschnitt 5.3.2).

Durch das Filtern von Infixen werden lediglich redundante Informationen aus der Eingabe gelöscht, da die Sequenzabgleiche bei der Infix-Bestimmung grundsätzlich nicht-approximativ erfolgen und damit nur Peptide herausgefiltert werden, deren biologisch relevanten Sequenzinformationen bereits in mindestens einem weiteren Peptid enthalten sind. Hierdurch wird sichergestellt, dass im Hinblick auf die in den nächsten Schritten erfolgenden weiteren Berechnungen keine wichtigen Informationen aus der Eingabe verloren gehen. Die Sequenzinformationen der herausgefilterten Peptide bleiben in den Aminosäuresequenzen der identifizierten Peptide, die Superstrings der gefilterten Peptide sind, erhalten (siehe Abbildung 6.2).

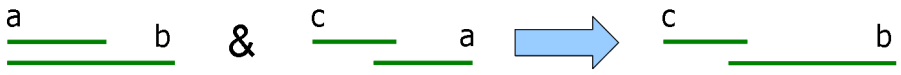

Abbildung 6.2: Schematische Darstellung des Vorgehens bei der Infix-Filterung. Es werden insgesamt drei Peptide bzgl. der Unterschiede in ihren Primärstrukturen miteinander verglichen. Wie ganz links dargestellt, besitzen Peptid a und Peptid b eine gemeinsame Subsequenz der Länge $|a|$. Daher ist a ein Infix von b. Zusätzlich hierzu steht Peptid a in Beziehung zu Peptid c, wobei keines der beiden Peptide a und c Infix des jeweils anderen ist. Aus der Beziehung zwischen Peptide a und b und der Transitivität der Überlappungsbeziehung folgt, dass Peptid a entfernt werden kann, ohne dass biologisch relevante Sequenzinformationen oder Informationen über die Beziehungen zwischen den in der Eingabe verbleibenden Peptiden verloren gehen.

Wie man sich leicht überlegen kann, ist das Filtern von Infixen auf der Basis paarweiser Sequenzvergleiche, der in der Eingabe enthaltenen Peptide, in quadratischer Zeit möglich. Dies ist, wie die Testergebnisse in Kapitel Sieben zeigen werden, für die praktische Anwendung des entwickelten Algorithmus ausreichend effizient.

6.3 Behandlung von Transpeptidierungseffekten

Dieses in Abschnitt 5.3.1 definierte grundlegende Problem der Proteinidentifikation bedarf einer separaten Prozessierung der Eingabe. Nachdem sämtliche Infixe aus der Eingabe herausgefiltert wurden, lassen sich auf Transpeptidierungseffekte zurückzuführende Veränderungen an den Aminosäuresequenzen der identifizierten Peptide mit Hilfe einer Liste von bekannten Transpeptidierungseffekten entfernen. Hierzu werden Sequenzvergleiche zwischen den Aminosäuresequenzen der einzelnen identifizierten Peptide und den bekannten Aminosäure- oder Peptidkondensaten durchgeführt. Enthält die Primärstruktur eines identifizierten Peptides ein solches Kondensat, so wird dieses entfernt und die Masse des identifizierten Proteins neu berechnet. Der zur Behandlung solcher Transpeptidierungen entworfene Mechanismus wurde von vornherein flexibel angelegt, sodass es jederzeit möglich ist neu entdeckte Transpeptidierungseffekte in die Überprüfung der Eingabe mit aufzunehmen. Die Überprüfung sämtlicher in einer Eingabe enthaltenen Peptide lässt sich in linearer Zeit durchführen.

Bedauerlicherweise gibt es derzeit nur sehr wenige gesicherte Erkenntnisse zu dem Problemkomplex der Transpeptidierung, daher muss die Behandlung von Transpeptidierungseffekten bei der Proteinidentifikation auf die aktuell vorliegenden, leider sehr überschaubaren, gesicherten Erkenntnisse beschränkt bleiben [35, 89, 90, 91, 92, 93]. Da die Berücksichtigung von Transpeptidierungseffekten oft nicht von vornherein erwünscht ist und Seiteneffekte wie dieser zudem nur unter bestimmten Analysebedingungen zu erwarten sind (Verdau findet z.B. in einem besonders saueren Milieu statt oder es wird eine übermäßig große Menge an Protease für den Proteinverdau verwendet), bleibt die Anwendung des implementierten Mechanismus optional.

6.4 Overlap-Berechnung

Um das ursprüngliche Protein P aus den identifizierten Peptiden aus F rekonstruieren zu können, müssen die durch den proteolytischen Verdau mit mehreren unterschiedlichen Enzymen entstandenen

Überlappungen zwischen den Peptiden ausgenutzt werden. Da die Überlappungen zwischen den einzelnen Peptiden nicht explizit in der Eingabe enthalten sind, müssen zunächst einmal sämtliche paarweisen Überlappungen zwischen den Peptiden aus F bestimmt werden. Der im Folgenden beschriebene Algorithmus zur Bestimmung solcher Überlappungen basiert auf den Arbeiten von Wu und Manber [96, 97].

6.4.1 Ermittlung der Overlaps

Da die Peptididentifikation, wie in Abschnitt 4.3 und 5.3 bereits erwähnt, fehlerbehaftet ist, können die Aminosäuresequenzen der identifizierten Peptide Abweichungen von den eigentlich korrekten Aminosäuresequenzen, wie sie in dem ursprünglichen Protein enthalten sind, aufweisen. Dies hat Konsequenzen für die Ermittlung der Überlappungen zwischen den identifizierten Peptiden. Peptide, die eigentlich gemeinsame Subsequenzen besitzen sollten, scheinen nicht miteinander in Beziehung zu stehen, oder es existieren Überlappungen zwischen Peptiden, die eigentlich keine gemeinsamen Subsequenzen aufweisen. Zudem gibt es für zwei Peptide, die unabhängig von der Problematik der Identifikationsgüte gemeinsame Aminosäuren besitzen, im Allgemeinen mehr als nur eine Möglichkeit sich zu überlappen.

Im Folgenden wird der Begriff Overlap als abkürzende Bezeichnung für so genannte Suffix-Präfix-Überlappungen zwischen den Aminosäuresequenzen zweier sich überlappender identifizierter Peptide verwendet. Solche Overlaps zeichnen sich dadurch aus, dass sie zwei identifizierte Peptide durch eine gemeinsame Subsequenz miteinander in Beziehung setzen, ohne dass eines dieser beiden Peptide Infix des anderen ist (siehe Abbildung 6.3).

Abbildung 6.3: Darstellung der beiden grundsätzlich möglichen Konstellationen für einen Overlap zwischen zwei Peptiden a und b.

Für jede der beiden in Abbildung 6.3 dargestellten Konstellationen eines gemeinsamen Overlaps zwischen zwei Peptiden a und b, gibt es in Abhängigkeit von der Länge Überlappung zwischen a und b, mehrere Möglichkeiten für einen gemeinsamen Overlap. Die Länge des gemeinsamen Overlaps kann sich auf das Minimum (der Overlap zwischen zwei Peptiden beruht auf lediglich einer gemeinsamen Aminosäure), Maximum (der gemeinsame Overlap hat die Länge der kürzeren der zwei Aminosäuresequenzen) oder einen Wert dazwischen belaufen. Deshalb muss die Wahl des im Weiteren zu verwendenden Overlaps zwischen zwei Peptiden auf der Basis von statistischen Erwägungen getroffen werden. Der ausgewählte Overlap sollte statistisch gesehen nicht auf einer relativ zufälligen Übereinstimmung von Aminosäuren basieren. Um dies sicherzustellen wird für jedes Paar sich überlappender Peptide der so genannte Least-Random-Overlap berechnet.

Der Least-Random-Overlap zweier Peptide p_i und $p_j \in F$ ist per Definition der längste Overlap zwischen p_i und p_j, der die zu spezifizierende, von der Güte der Peptididentifikation abhängige, Mindestlänge mol nicht unter- und die maximale ebenfalls anzugebende Fehlerschwelle dis nicht überschreitet. Die Berechnung eines solchen Least-Random-Overlaps lässt sich mit einem approximativen bzw. nicht-approximativen Pattern-Matching-Algorithmus, wie er von Wu und Manber in [96] beschrieben wird, effizient bewerkstelligen.

Bezüglich der beiden eben genannten Parameter mol und dis ist noch anzumerken, dass diese beim derzeitigen Stand der Entwicklung, von einem erfahrenen Benutzer angegeben werden müssen. Es ist aber durchaus vorstellbar und erwünscht, dass diese Parameter zukünftig aufgrund der eingelesenen Eingabedaten automatisch bestimmt werden.

6.4.2 Approximatives und nicht-approximatives Pattern-Matching

Ausgangsproblem des approximativen Pattern-Matchings ist es zu einem vorgegebenen Text T sämtliche exakten oder eventuell abweichenden Vorkommen eines Suchmuster P zu finden. Übertragen auf das Problem des Least-Random-Overlaps bedeutet dies, dass es für ein vorgegebenes Peptid p_i festzustellen gilt, ob p_i einen gemeinsamen Substring in der Form einer Suffix-Präfix-Übereinstimmung mit einem zweiten

vorgegebenen Peptid p_j besitzt. Die Berechnung einer solchen Suffix-Präfix-Übereinstimmung kann dabei wahlweise approximativ oder nicht-approximativ erfolgen. Wie ähnlich bzw. wie unähnlich sich dabei die gemeinsamen Subsequenzen zweier Peptide p_i und p_j sehen dürfen, um trotz allem noch als identisch aufgefasst zu werden, wird durch die Parameter *mol* und *dis* aus dem vorherigen Unterabschnitt bestimmt. Der Parameter *dis* entspricht dabei der maximalen Levenshtein-Distanz, um welche sich die gemeinsamen Subsequenzen der beiden Peptide unterscheiden dürfen. Unter der Leveshtein-Distanz zweier Strings versteht man im Allgemeinen die maximale Anzahl an Einfüge-, Lösch- oder Vertauschungsoperationen die notwendig sind, um jeweils einen der beiden Strings in den anderen umzuwandeln.

Um nun den Least-Random-Overlap zwischen den Aminosäuresequenz zweier gegebener Peptide p_i und $p_j \in F$ zu berechnen, muss zunächst das Vorgehen von Wu und Manber [96] auf die vorliegende Problemstellung übertragen werden. Dies gelingt ohne größeren Aufwand, da der in [96] beschriebene Algorithmus für beliebige Alphabete Σ anwendbar ist. Damit die Funktionsweise des Gesamtalgorithmus einfacher zu verstehen ist, wird zunächst der Algorithmus zur Berechnung von nicht-approximativen Matchings erläutert. Die Erweiterung für approximative Matchings wird sich anschließend kanonisch zu der Funktionsweise für nicht-approximative Matchings verhalten und daher anschließend leicht nachzuvollziehen sein.

Berechnung nicht-approximativer Matchings

Für die Berechung sämtlicher nicht-approximativen Matchings zweier Strings P und T wird ein Array von Bitvektoren R, mittels dynamischer Programmierung, schrittweise berechnet. Die Größe des Bitvektorarrays R beträgt $m = |T| + 1$ und jeder der m einzelnen Bitvektoren R_0 bis R_m besitzt die Größe $n = |P|$. Die einzelnen Einträge des Bitvektorarrays R besitzen die folgende Bedeutung:

Definition 6.4.1 *Geben seinen zwei Strings P und $T \in \Sigma^*$. Des Weiteren gelte $R_0[i] = 0 \ \forall i, 1 \leq i \leq n-1$; $R_0[0] = 1$.*

$$R_{j+1}[i] = \begin{cases} 1, & \text{falls } R_j[i-1] = 1 \wedge P[i] = T[j+1], \\ 0, & \text{sonst.} \end{cases}$$

Natürlich sprachlich ausgedrückt besagt Definition 6.4.1, dass das $i-te$ Bit eines Bitvektors R_j, daher $R_j[i]$, genau dann den Wert eins annimmt, falls die ersten i Buchstaben von P mit den letzten i Buchstaben der ersten j gelesenen Buchstaben von T übereinstimmen. Durch die schrittweise Berechnung sämtlicher Bitvektoren R_1 bis R_m, erhält man Angaben zu sämtlichen Übereinstimmungen zwischen P und T (siehe Abbildung 6.4).

T = aabaacaabacab, P = aabac

	R0	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13			
		a	a	b	a	a	c	a	a	b	a	c	a	b	a	b	c
a	1	1	1	0	1	1	0	1	1	0	1	0	1	0	1	0	0
a	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0
b	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0
a	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0
c	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1

Abbildung 6.4: Beispiel für ein auf Basis von nicht-approximativen Matchings berechneten Bitvektorarrays. Das hier dargestellte Bitvektorarray wurde schrittweise berechnet. Dabei wird der jeweils aktuell zu berechnende Bitvektor R_{j+1} auf Basis des unmittelbar vorher berechneten Bitvektors R_j per dynamischer Programmierung bestimmt. Die drei Bitvektoren am rechten Rand sind die Bitmasken der in Σ enthaltenen Buchstaben. Sie dienen der effizienten Berechnung des jeweils aktuellen R_{j+1}.

Da der gesamte Ansatz auf dynamischer Programmierung beruht, muss die Bestimmung des jeweils aktuellen R_{j+1} in konstanter Zeit zu bewerkstelligen sein. Dazu wird vor der eigentlichen Berechnung

von R für jeden Buchstaben aus dem zugrunde liegenden Alphabet Σ eine Bitmaske erzeugt. Diese Bitmasken besitzen die Länge $n = |P|$ und haben die folgende Eigenschaft:

Definition 6.4.2 *Für $x \in \Sigma$ und $0 \leq i \leq n-1$ gilt:*

$$x[i] = \begin{cases} 1, & \text{falls } p_i = x, \\ 0, & \text{sonst.} \end{cases}$$

Die zu einem Buchstaben aus Σ gehörige Bitmaske ist also an genau den Position gleich eins, an denen P den entsprechenden Buchstaben aufweist. Mit Hilfe dieser in $O(|\Sigma|n)$ durchführbaren Präprozessierung lässt sich ein Bitvektor durch zwei simple Berechnungsschritte ermitteln. Für die Berechnung des Bitverktors R_{j+1} wird R_j zunächst arithmetisch um eine Stelle nach rechts verschoben. Anschließend wird überprüft, ob der zuletzt gelesene Buchstabe von T mit dem zuletzt gelesenen Buchstaben von P übereinstimmt ($P[i] = T[j+1]$). Diese beiden Schritte lassen sich verallgemeinert so zusammenfassen:

Definition 6.4.3 *Für zwei gegebene Peptide p_i und p_j, wobei in Bezug auf die Berechnung nicht-approximativer Matchings p_i dem Suchpattern P und p_j dem Text T entspricht, lassen sich sämtliche Bitvektorarrays $R_j + 1$ durch folgende Formeln bestimmen. $R_0 = 1000\ldots000$ ist der initiale Bitvektor mit $|P| = n$ Stellen.*

$$R_{j+1} = Rshift[R_j] \text{ AND } x$$

Wobei x die Bitmaske des als $j+1$-ten gelesenen Buchstabens ist.

Da das arithmetische Schieben eines Bitvektors der maximalen Länge n und die Bildung der Konjunktion zweier Bitvektoren mit maximaler Länge n in $O(n)$ durchführbar ist, bleibt die Gesamtrechenzeit für die Berechnung der nicht-approximativen Matchings durch $O(|\Sigma|n)$ beschränkt. Diese pseudo-polynomielle Rechenzeit ist für kleine Alphabete, wie das der proteinogenen Aminosäuren, unproblematisch.

Um nun festzustellen, ob zwei identifizierte Peptide p_i und p_j einen gemeinsamen Overlap besitzen, der für die Rekonstruktion des ursprünglichen Proteins nützlich ist, muss festgestellt werden, ob p_i einen mindestens *mol* Buchstaben langen Suffix besitzt, der Präfix von p_j ist oder ob p_j einen ebenfalls mindestens *mol* Buchstaben langen Suffix besitzt, der Präfix von p_i ist. *mol* gibt die Minimale Überlappungslänge an, die ein Overlap zwischen zwei Peptiden besitzen muss, um für die Rekonstruktion des ursprünglichen Proteins berücksichtigt zu werden.

Kehrt man zu dem Beispiel für T und P aus Abbildung 6.4 zurück, dann wäre es für den Fall, dass p_i P und p_j T entspricht, unnötig Bitvektoren mit einem Index größer als $|P| = 5$ zu berechnen, da es keinen längsten Suffix von P geben kann, der länger als P selbst ist. Im umgekehrten Fall, also $P = p_j$ und $T = p_i$ müssten ebenfalls nur die ersten 5 Bitvektoren berechnet werden, da es keinen längsten Präfix von T geben kann, der Suffix von P ist und länger als 5 ist.

Gilt $R_n[n] = 0$ bedeutet dies lediglich, dass der längste Suffix von p_i, der Präfix von p_j ist, nicht Länge n haben kann. Daher muss die Überlappungsberechnung für den nächst kürzeren Suffix von p_i wiederholt werden. Der ganze Prozess wiederholt sich also für p_j und die $n-1$ letzten Buchstaben von p_i. Was die maximale Anzahl der durchzuführenden Berechnungen angeht, so gilt wieder, dass die Suche nach einem längsten Suffix von p_i, der Präfix von p_j ist, abgebrochen werden kann, sobald p_i kürzer als *mol* wird.

Betrachtet man nun wieder die worst-case-Rechenzeit, so werden maximal $n - mol + 1$ viele Iterationen des Gesamtberechnungsprozesses durchgeführt, um den längsten Suffix von p_i zu bestimmen, der Präfix von p_j ist. Im worst-case ist $n - mol = n$. Daher ergibt sich als Gesamtabschätzung $O(|\Sigma|n^2)$.

Berechnung approximativer Matchings

Sollen nun für zwei identifizierte Peptide p_i und p_j, nicht nur sämtliche nicht-approximativen Matchings berechnet werden, sondern möchte man zudem sämtliche approximativen Matchings berechnen, so müssen eventuell durchzuführende Einfüge-, Lösch- und Vertauschungsoperationen auf den Sequenzen der beiden

Peptide berücksichtigt werden. Die Anzahl der Abweichungen, die durch solche Operationen maximal ausgeglichen werden dürfen, hängt von dem durch dis definierten Wert für die maximale Levenshtein-Distanz zwischen den gemeinsamen Overlaps der Peptide ab.

Zusätzlich zu dem Bitvektorarray R, welches sämtliche nicht-approximativen Matchings zwischen den Subsequenzen zweier Peptide charakterisiert, beschreibt R^d sämtliche Matchings zwischen den Aminosäuresequenzen zweier Peptide, die maximal $0 \leq d \leq dis$ Unterschiede in der Form von Einfügungen, Löschungen und Vertauschungen aufweisen. Da sich die Berechnung der Bitvektoren von R nicht verändert, muss nur noch das Prinzip, nach dem die Bitvektoren von R^d entstehen, beschrieben werden:

Definition 6.4.4 *Geben seinen zwei Strings P und $T \in \Sigma^*$. Des Weiteren gelte $R_0^d[i] = 0$ $\forall i$, $d+1 \leq i \leq n-1$; $R_0^d[0]$ bis $R_0^d[d] = 1$. $R_{j+1}^d[i] = 1$, falls:*

1. *die ersten $i-1$ Buchstaben von P mit den $i-1$ letzten Buchstaben von T bis auf maximal d Unterschiede übereinstimmen und $P[i] = T[j+1]$ gilt (Übereinstimmung von $P[i]$ und $T[j+1]$),*
2. *die ersten $i-1$ Buchstaben von P mit den $i-1$ letzten Buchstaben von T bis auf maximal $d-1$ Unterschiede übereinstimmen und $P[i] \neq T[j]$ gilt (Substitution an der Position $T[j+1]$),*
3. *die ersten $i-1$ Buchstaben von P mit den $i-1$ letzten Buchstaben von T bis auf maximal $d-1$ Unterschiede übereinstimmen und $P[i] \neq T[j+1]$ gilt (Löschung an der Position $P[i]$),*
4. *die ersten i Buchstaben von P und die letzten $i-1$ Buchstaben von T bis auf maximal $d-1$ Unterschiede übereinstimmen und $P[i] \neq T[j]$ gilt (Einfügung an der Position $T[j+1]$).*

Aus diesem Prinzip lässt sich die folgende Verallgemeinerung für die Berechnung von R^d ableiten:

Definition 6.4.5 *Es gelte $R_0^d = 1 \ldots 1000 \ldots 000$ ist der initiale Bitvektor mit $|P| = n$ Stellen und d Einsen. Für zwei gegebene Peptide p_i und p_j, wobei in Bezug auf die Berechnung approximativer Matchings p_i dem Suchpattern P und p_j dem Text T entspricht, lassen sich sämtliche Bitvektorarrays $R_j^d + 1$, mit $0 \leq d \leq dis$, durch folgende Formeln bestimmen.*

$$\begin{aligned} R_{j+1}^d &= Rshift[R_j] \text{ AND } x \text{ OR } Rshift[R_j^{d-1}] \text{ OR } Rshift[R_{j+1}^{d-1}] \text{ OR } R_j^{d-1} \\ &= Rshift[R_j^d] \text{ AND } x \text{ OR } Rshift[R_j^{d-1} \text{ OR } R_{j+1}^{d-1}] \text{ OR } R_j^{d-1}. \end{aligned}$$

Der Bitvektor x entspricht hierbei wieder der Bitmaske des $j+1$-ten gelesen Buchstabens von T. Die im Vergleich zu Definition 6.4.3 hinzugekommenen Disjunktionen werden für die Berechnung der approximativen Overlaps benötigt. Wird zu einem Bitvektorarray R^d mit $d \geq 1$ ein Bitvektor R_{j+1}^d berechnet, so werden durch die drei Terme $Rshift[R_j^{d-1}]$, $Rshift[R_{j+1}^{d-1}]$ und R_j^{d-1} mögliche Substitutionen, Löschungen und Einfügungen von einzelnen Buchstaben berücksichtigt.

Die im Vergleich zu der Berechnung der nicht-approximativen Matchings notwendigen zusätzlichen arithmetischen Schiebe- und logischen Vergleichsoperationen erzeugen asymptotisch betrachtet keinen zusätzlichen Mehraufwand. Damit verursacht die Bestimmung sämtlicher approximativer Matchings zweier gegebener Peptide p_i und p_j asymptotisch gesehen die gleiche Zeitkomplexität wie die Berechnung sämtlicher nicht-approximativer Matchings zwischen diesen beiden Peptiden. Die Gesamtrechenzeit für die Berechnung der approximativen Matchings bleibt daher durch $O(|\Sigma|n)$ beschränkt.

Analog zu der Argumentation bzgl. des zusätzlichen Berechnungsaufwands zur Bestimmung des längsten Suffixes von p_i, der Präfix von p_j ist, aus dem vorherigen Abschnitt, ergibt sich für die Bestimmung sämtlicher approximativer Overlaps insgesamt eine worst-case-Rechenzeit von $O(|\Sigma|n^2)$. Bezüglich des Speicherplatzverbrauchs gilt, dass für die Berechnung der Bitvektoren R_{j+1} eines Bitvektorarrays R_{j+1}^d maximal ein zusätzliches Bitvektorarray der Größe nm im Speicher gehalten werden muss, da für die Berechnung von R_{j+1} die Bitvektoren R_j, R_j^{d-1}, R_{j+1}^{d-1} und R_j^{d-1} benötigt werden und diese entweder aus dem Bitvektorarray R^d oder R^{d-1} stammen.

6.5 Der Overlap-Graph

Die bei der Berechnung der Overlaps gewonnenen Informationen über die Überlappungsbeziehungen zwischen den identifizierten Peptiden untereinander müssen im Hinblick auf die noch folgenden Rekonstruktionsschritte auf geeignete Art und Weise persistent gemacht werden. Die hierfür verwendete Datenstruktur sollte aber nicht nur einen guten Kompromiss zwischen Speicherplatzverbrauch und mittlerer Zugriffszeit auf die gespeicherten Daten darstellen, sondern zudem die Berechnung einer Lösung, des anschließend zu behandelnden Peptide-Assembly-Problem, möglichst gut unterstützen. Da es sich bei dem Peptide-Assembly-Problem (siehe Abschnitt 5.4) um ein kombinatorisches Problem handelt, dessen Lösung in der Berechnung einer geeigneten Permutation sämtlicher identifizierter Peptide besteht, gilt es von vorn herein möglichst viele der Permutationen, die keine korrekte Lösung ergeben, auszuschließen und so die Anzahl der potentiell korrekten Permutationen auf ein Minimum zu beschränken. Die hierfür erforderlichen Eigenschaften vereinen sich in einem so genannten gewichteten Overlap-Graphen.

6.5.1 Definition des Overlap-Graph

Ein gewichteter Overlap-Graph lässt sich als gerichteter Graph $G = (V, E, w)$ definieren. Die Knotenmenge V ordnet jedem massenspektrometrisch identifizierten Peptid einen Knoten zu. Die Kantemenge E enthält die gerichteten Kanten des Graphen. Eine gerichtete Kante zwischen zwei Knoten von i und $j \in V$ entspricht einem Overlap zwischen den Peptiden, die durch die beiden Knoten repräsentiert werden.

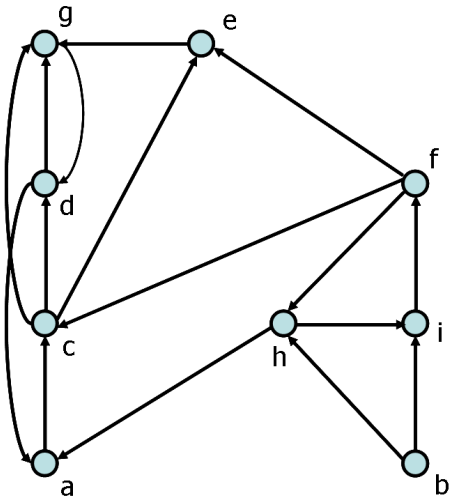

Abbildung 6.5: Beispiel für einen Overlap-Graphen, der aus neuen Peptiden besteht.

Was die Ausrichtung der gewichteten Kanten angeht, so hängt diese von der Art der Überlappung zwischen den jeweils betrachteten Peptiden p_i und p_j ab. Da es nach dem Herausfiltern sämtlicher in F enthaltenen Infixe keine Überlappungen zwischen zwei Peptiden p_i und p_j mehr geben kann für die gilt, dass eines der beiden Peptide komplett in der Aminosäuresequenz des anderen als Präfix oder Suffix enthalten ist, können nur die folgenden drei Overlap-Konstellationen auftreten:

i. Ein Präfix von p_i ist Suffix von p_j: E enthält die gerichtete Kante $e(p_j, p_i)$,

ii. Ein Präfix von p_j ist Suffix von p_i bzw. ein Suffix von p_i ist Präfix von p_j: E enthält analog zu i. die gerichtete Kante $e(p_i, p_j)$,

iii. Ein Suffix von p_j ist Präfix von p_i: E enthält analog zu i. die gerichtete Kante $e(p_j, p_i)$.

Die Kanten des Graphen sind gemäß der Kantengewichtsfunktion w gewichtet. Die Kantengewichtung spielt bei der späteren Rekonstruktion der Primärstruktur eines zu identifizierenden Proteins eine entscheidende Rolle, da sie es ermöglicht, bei der Ermittlung eines Rekonstruktionspfades auf dem Overlap-Graphen, die Fortsetzung dieses Rekonstruktionspfades von den Kantengewichten der von dem aktuellen

Knoten ausgehenden Kanten, abhängig zu machen. Diese Entscheidung lässt sich bei geeigneter Definition der Kantengewichtsfunktion w von der biologischen Signifikanz der zu betrachtenden Overlaps abhängig machen.

Um eine biologisch sinnvolle Gewichtung, für die in G enthaltenen gerichteten Kanten, berechnen zu können, müssen die folgenden Kenngrößen betrachtet werden:

- $|Overlap(p_i, p_j)|$: Länge der Overlaps zwischen zwei Peptiden p_i und p_j,
- $f_{id}(p_i)$, $f_{id}(p_j)$: Identifikationsscores der an der Kante beteiligten Peptide p_i und p_j (siehe Abschnitt 5.2.4),
- $|p_j|$: Länge der Aminosäuresequenz des zu der Rekonstruktion R hinzukommenden Peptides p_j,
- $\overline{|Overlap(p_i, p_j)|} = min(|p_i|, |p_j|)$: obere Schranke für die Länge des Overlaps zwischen p_i und p_j, die sich aus der Länge des kürzeren der beiden Peptide ergibt,
- $|diff(Overlap(p_i, p_j))|$: Anzahl der Abweichungen, die bei der Bestimmung des Overlaps zwischen p_i und p_j auftraten.

Setzt man diese Kenngrößen in Bezug auf die biologische Signifikanz der Overlaps sinnvoll in Beziehung zueinander, erhält man für die Kantengewichtungsfunktion w die folgende Definition.

Definition 6.5.1 *Gegeben seien die eben aufgezählten Kenngrößen zweier, durch einen Overlap miteinander in Beziehung stehender, Peptide p_i und p_j. Basierend auf diesen Kenngrößen ergibt sich das Kantengewicht der in G enthaltenen zugehörigen Kante wie folgt:*

$$w(p_i, p_j) = \frac{|Overlap(p_i, p_j)| * f_{id}(p_i) * f_{id}(p_j) * |p_j|}{\overline{|Overlap(p_i, p_j)|} * (1 + |diff(Overlap(p_i, p_j))|)}.$$

Diese Definition der Kantengewichtsfunktion w ist vom Standpunkt der Biologie aus gesehen sinnvoll, da sie mehrere entscheidende Eigenschaften besitzt:

1. Overlaps zwischen Peptiden, die zwar sehr lang sind, aber auch sehr viele divergierende Aminosäuren enthalten und solche bei denen die Überlappung auf nur sehr wenigen gemeinsamen Aminosäuren basiert, werden entsprechend schlecht bewertet.
2. Overlaps, die im Vergleich zu ihrer maximal möglichen Gesamtlänge verhältnismäßig kurz sind, werden tendenziell schlechter bewertet, als Overlaps, die kürzer aber insgesamt näher an ihrer theoretisch möglichen Maximallänge sind.
3. Ein Overlap zwischen zwei Peptiden p_i und p_j, der von seiner maximal möglichen und tatsächlichen Gesamtlänge vergleichbar zu dem Overlap zwischen zwei anderen Peptiden p_k und p_l ist, wird schlechter als der Overlap zwischen p_k und p_l bewertet, falls die Identifikationsscores der beiden Peptide p_i und p_j niedriger als die von p_k und p_l sind.

Da die Auswertung der Funktion $w(p_i, p_j)$ lediglich konstante Rechenzeit benötigt, ergibt sich für die Erzeugung von G eine zeitliche Gesamtkomplexität von $O(n^2)$, wobei n der Anzahl der in F enthaltenen Peptide entspricht.

6.5.2 Repräsentation des Overlap-Graphen im Speicher

Der Graph lässt sich auf algorithmischer Ebene als Adjazenzmatrix M_G repräsentieren. Die Zeilen und Spalten der quadratischen Matrix M_G werden mit den aufsteigend durchnummerierten Indizes der identifizierten Peptiden indiziert.

Ein Eintrag der Form $M_G(i, j) = w(p_i, p_j)$ bedeutet, dass G eine gerichtete Kante von dem Knoten i zu dem Knoten j enthält und die zugehörige Kante das Kantengewicht $w(p_i, p_j)$ besitzt. Existiert zwischen

zwei Peptiden p_i und p_j keine Überlappungsbeziehung, so enthält M_G in der i-ten Zeile und j-ten Spalte eine Null als Kantengewicht.

Da Einträge auf der Hauptdiagonalen den Kantengewichten von Schlaufen im Graphen entsprechen, also von Kanten, die von einem Peptid p_i zu p_i selbst verlaufen und diese hier nicht von Interesse sind, werden die Einträgen von M_G entlang der Hauptdiagonalen auf -1000.0 gesetzt. Im Prinzip könnte es auch jeder andere negative Wert sein, der Wert -1000.0 ist daher lediglich eine implementationstechnische Konvention. Diese Konvention stellt sicher, dass Einträge, die zur Hauptdiagonalen von M_G gehören, bei der Rekonstruktion des ursprünglichen Proteins nicht als wählbare Kante interpretiert werden.

Wird M_G zeilenweise gelesen, so lassen sich die Einträge der jeweils aktuell betrachteten Spalten als mögliche Nachfolger für das aktuell betrachtete Peptid interpretieren. Befindet man sich während der Proteinrekonstruktion z.B. in der i-ten Zeile, so lässt sich das Nachfolgerpeptid per Vergleich sämtlicher Kantengewichte in der i-ten Zeile ermitteln. Wird M_G dagegen spaltenweise gelesen, so lassen sich die Einträge in den einzelnen Zeilen der aktuell betrachteten Spalten als die möglichen Vorgänger des aktuell betrachteten Peptides interpretieren. Da sich der Overlap-Graph mit Hilfe beider Lesearten traversieren lässt, sind beide Lesearten für die Rekonstruktion nützlich.

6.6 Aufbereitung des Overlap-Graphen

Da der Overlap-Graph trotz initial durchgeführter Filterungen (siehe Abschnitte 6.1, 6.2 und 6.3) für Proteine, die während des enzymatischen Verdaus in sehr viele Peptide zerfallen, noch immer sehr groß werden kann — dies ist auf die Durchführung der für die *de novo*-Proteinidentifikation notwendigen Mehrfachverdauungen mit unterschiedlichen Proteasen zurückzuführen — muss die Anzahl der insgesamt zu betrachtenden Proteinrekonstruktionen auf andere Art und Weise gesenkt werden.

Eine Möglichkeit die Anzahl der zu betrachtenden Rekonstruktionsmöglichkeiten zu senken, liegt in der Zerlegung des Overlap-Graphen in seine starken Zusammenhangskomponenten (im Folgenden mit SCC für *strong connecting component* abgekürzt).

6.6.1 Bestimmung der SCCs des Overlap-Graphen

Der Algorithmus, mit dem die SCCs des Overlap-Graphen bestimmt werden, basiert auf dem von Tarjan 1972 veröffentlichten Algorithmus zur Tiefensuche auf gerichteten Graphen [98]. Die Identifikation der SCCs von G lässt sich durch die folgenden vier Schritte bewerkstelligen:

(a) In einem ersten Tiefensuchdurchlauf durch G werden alle Depth-First-Spannbäume von G ermittelt. Dabei erhält ein besuchter Knoten seine DFS-Nummer erst nach Beendigung des zugehörigen rekursiven DFS-Aufrufs (siehe Abbildung 6.6);

(b) Konstruiere G_r, den zu G inversen Overlap-Graphen. G_r ist zu G insofern invers, als dass die Kantenrichtungen in G_r genau umgekehrt zu denen in G sind (siehe Abbildung 6.7 links);

(c) In einem zweiten Tiefensuchdurchlauf auf G_r, werden die zu G_r gehörigen Depth-First-Spannbäume konstruiert. Die Abarbeitung der Knoten orientiert sich dabei an den in (a) vergebenen DFS-Nummern. Es wird stets der Knoten mit der höchsten noch verbliebenen DFS-Nummer zuerst abgearbeitet (siehe Abbildung 6.7);

(d) Die Knotenmengen der in (c) ermittelten DFS-Spannbäume bilden die starken Zusammenhangskomponenten von G.

(e) Anschließend an die eigentliche Bestimmung der SCCs des Overlap-Graphen, erfolgt ein zusätzlicher klassischer Depth-First-Search-Durchlauf, der eine Einteilung der Kantenmenge von G in Tree-, Back-, Forward- und Cross-Kanten ermittelt.

Da jeder der fünf aufgeführten Berechnungsschritte in $O(n+m)$ durchgeführt werden kann, wobei n der Anzahl der Kanten und m der Anzahl der Knoten in G entspricht, liegt die asymptotisches Gesamtrechenzeit bei $O(n+m)$.

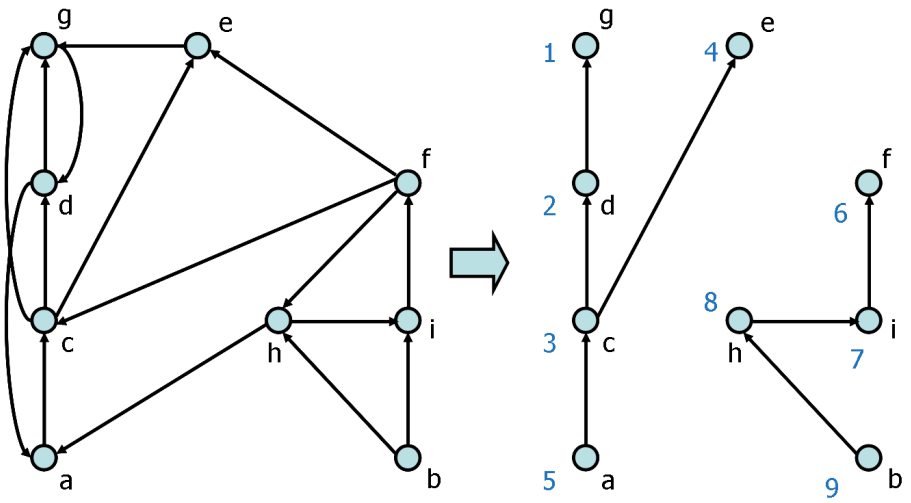

Abbildung 6.6: Erster Schritt der SCC-Bestimmung. Die Zahlen an den Knoten des DFS-Spannbaums geben die bei Abschluss des Rekursiven DFS-Aufrufs vergebenen DFS-Nummern an.

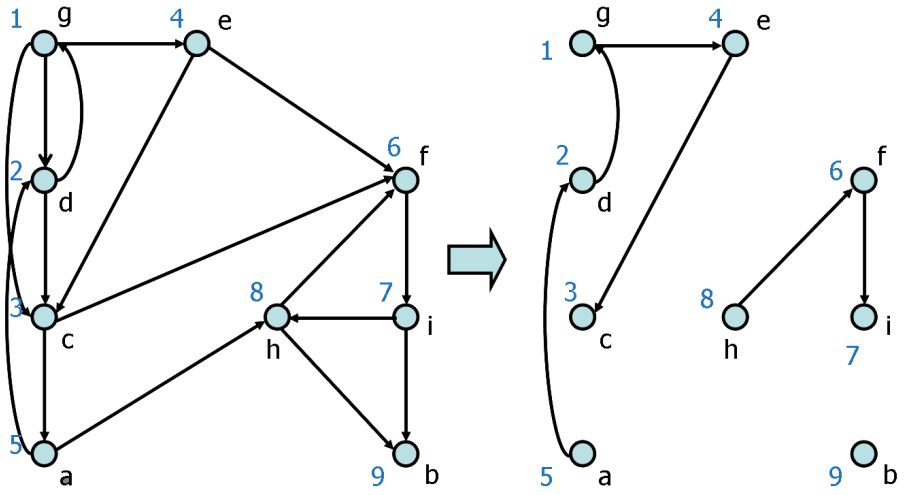

Abbildung 6.7: Zweiter und Dritter Schritt der SCC-Bestimmung. Die Nummern an den Knoten des Overlap-Graphen links entsprechen den DFS-Nummern aus dem ersten DFS-Durchlauf, die Nummern an den Knoten auf der rechten Seite den DFS-Nummern aus dem zweiten DFS-Durchlauf.

6.6.2 Nutzen der Aufbereitung des Overlap-Graphen

Die eben beschriebenen Maßnahmen haben im Hinblick auf die Minimierung der insgesamt zu betrachtenden Proteinrekonstruktionen die folgenden nützlichen Eigenschaften:

1. Die Rekonstruktion des ursprünglichen Proteins lässt sich nach Bestimmung der starken Zusammenhangskomponenten leichter bewerkstelligen. Die Aminosäuresequenzen der zu den starken Zusammenhangskomponenten gehörigen Polypeptide stellen Teilsequenzen des ursprünglichen Proteins dar. Rekonstruiert man zunächst diese Polypeptide und fügt sie anschließend auf geeignete Art und Weise zu einer Aminosäuresequenz zusammen, erhält man eine Rekonstruktion des gesamten ursprünglichen Proteins. Da durch den Zwischenschritt der Polypeptidrekonstruktion einige der identifizierten Peptide bereits in den Polypeptiden enthalten sind, müssen anschließend insgesamt weniger unterschiedliche Kombinationsmöglichkeiten für die identifizierten Peptide und damit auch weniger Proteinrekonstruktionen und ϵ-Layouts betrachtet werden.

2. Durch die Bestimmung der Menge der Back-Kanten wird festgestellt, ob der Overlap-Graph kreisfrei ist. Ist er es nicht, so wird die Struktur der in G enthaltenen Kreise unabhängig von dem gewählten

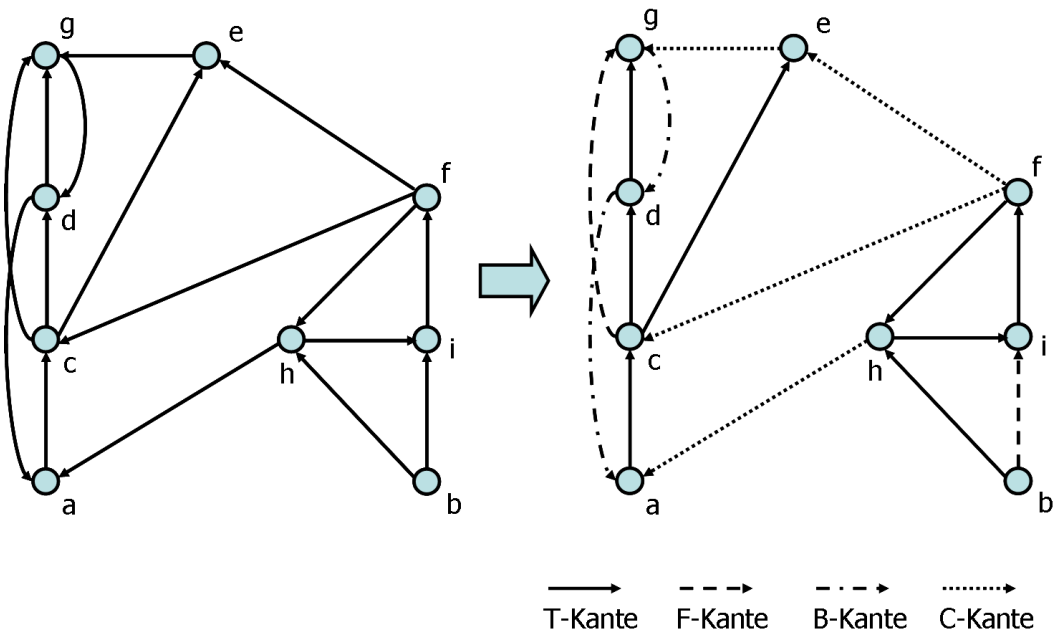

Abbildung 6.8: Bestimmung einer Partitionierung der Kantenmenge von G. Die Kantenmenge von G wird in die vier disjunkten Tree-, Forward-, Back- und Cross-Kantenmengen zerlegt.

Startpunkt der Tiefensuche eindeutig bestimmt. Damit Rekonstruktionen nicht in Endlosschleifen geraten, müssen die identifizierten Kreise bei der Bestimmung der Struktur der SCCs entsprechend behandelt werden (siehe die Abschnitte 6.7.1, 6.7.2 und 6.7.3).

3. Die Einteilung der Kantenmenge E in Tree-, Back-, Forward- und Cross-Kanten ermöglicht eine potentielle Minimierung der vorliegenden Kantenmenge E. Grundsätzlich werden alle vier Kantensorten zur Rekonstruktion der SCCs und des eigentlichen Proteins auf Basis des Overlap-Graphen gebraucht. Allerdings lassen sich so genannte SCC-externe Kanten zuweilen aus dem Graphen herausfiltern.

   An dieser Stelle muss, was die Menge der Tree- und Cross-Kanten angeht, zwischen zwei Ausprägungen von Kanten differenziert werden, dies sind die so genannten SCC-internen und SCC-externen Kanten. Wie die Benennung dieser beiden Ausprägungen bereits andeutet, verlaufen SCC-interne Tree- oder Cross-Kanten innerhalb der starken Zusammenhangskomponenten eines Overlap-Graphen und verbinden damit Knoten, die zu der Selben starken Zusammenhangskomponente gehören. SCC-externe Tree- oder Cross-Kanten verbinden wiederum Knoten, die zu unterschiedlichen SCCs gehören (siehe Abbildung 6.9).

   Was nun den Nutzen dieser beiden Ausprägungen von Tree- und Cross-Kanten bzgl. der Proteinrekonstruktion angeht, so sind SCC-internen Tree- und Cross-Kanten für den Rekonstruktionsprozess der Polypeptide unabdingbar, da sie Knoten aus verschiedenen Teilen ein und derselben SCC miteinander verbinden. Die Rekonstruktion der Polypeptide fußt also auf der Verwendung von SCC-internen Tree- und Cross-Kanten, sowie Back- und Forward-Kanten, die alle zusammengenommen die Struktur der SCCs ausmachen.

   Die durch SCC-externe Tree- und Cross-Kanten repräsentierten Informationen über den Überlappungsgrad zweier Peptide, die durch Knoten aus zwei unterschiedlichen SCCs repräsentiert werden, haben für den Rekonstruktionsprozess der Polypeptide keine Bedeutung. Vielmehr kommt ihr Nutzen bei der Rekonstruktion des Gesamtproteins zum Tragen, da sie Überlappungsbeziehungen zwischen den Peptiden der einzelnen Polypeptide repräsentieren. Durch Ausnutzung dieser Überlappungsinformationen lässt sich die Anzahl der Permutationen, die insgesamt bei der Bestimmung der Primärstruktur des ursprünglichen Proteins zu betrachtenden sind, senken.

   Da SCC-externe Kante aber nur dann für die letzte Phase des Rekonstruktionsprozesses nützlich sind, wenn der Grad der biologischen Variabilität der Aminosäuresequenz des zu untersuchenden Proteins und die Kantengewichte der betrachteten SCC-externen Tree- und Cross-Kanten hoch

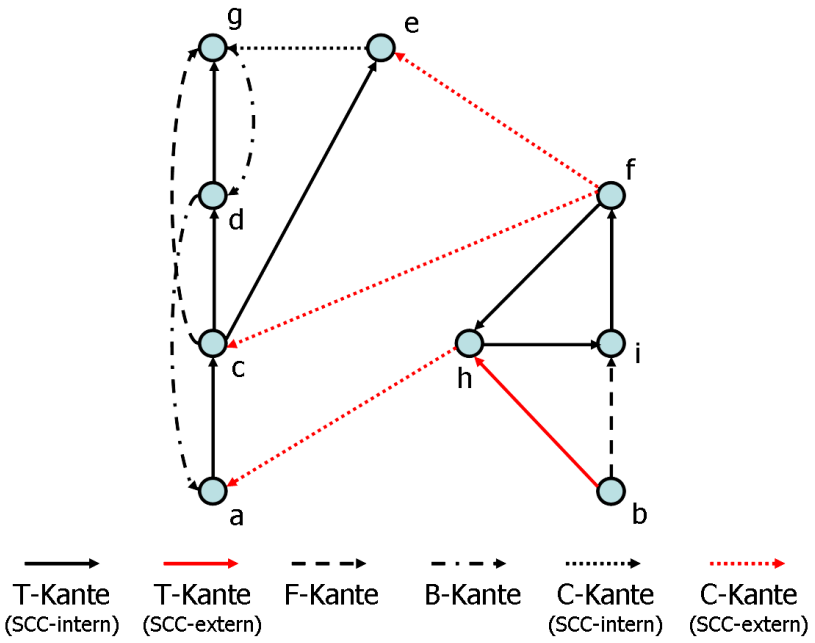

Abbildung 6.9: Bestimmung einer Partitionierung der Kantenmenge von G inklusive einer Differenzierung zwischen SCC-in- und SCC-externer Tree- und Cross-Kanten. Die Kantenmenge von G wird zusätzlich zu der disjunkten Zerlegung in Tree-, Forward-, Back- und Cross-Kanten noch bzgl. SCC-interner und SCC-externer Tree- und Cross-Kanten unterteilt.

genug sind, können sich SCC-externe Kanten aber auch kontraproduktiv auf den Gesamtrekonstruktionsprozess auswirken. Ist der Grad der biologischen Variabilität der Aminosäuresequenz des ursprünglichen Proteins gering, treten daher also bestimmte Aminosäuresequenzen extrem häufig in der Primärstruktur des ursprünglichen Proteins auf, so entstehen während des Aufbaus des Overlap-Graphen zwischen den verschiedenen SCCs eine Vielzahl von SCC-externen Tree- und Cross-Kanten. Dies führt dazu, dass die Ermittlung der korrekten Reihenfolge in der die Sequenzen der rekonstruierten Polypeptide aneinandergefügt werden müssen, durch SCC-externe Tree- und Cross-Kanten eher erschwert als erleichtert wird.

Sollte sich daher bei der Bestimmung der SCCs per Tiefensuche herausstellen, dass es eine Vielzahl von SCC-externen Kanten gibt, die auf eine geringe biologische Variabilität der Peptide zurückzuführen sind (die Kanten konstruieren eng miteinander verknüpfte Kreise) und sollten diese Kanten zusätzlich ein geringes Kantengewicht besitzen, so werden diese aus dem Overlap-Graphen entfernt. Mit geringem Kantengewicht ist in diesem Zusammenhang ein unter dem Median der Kantengewichte sämtlicher SCCs liegendes Kantengewicht gemeint. Dieser Wert lässt sich während der Ermittlung der SCCs leicht in $O(n)$ berechnen, wobei n der Anzahl der in G enthaltenen Kanten entspricht. Ist also das Kantengewicht einer SCC-externen Tree- oder Cross-Kante in Relation zu den Kantengewichten sämtlicher anderer Kanten des Overlap-Graphen überdurchschnittlich niedrig und gehört sie zu einem Geflecht von eng miteinander verwobenen Kreisen, so wird sie aus E entfernt.

6.7 Rekonstruktion der Polypeptide

Für die Rekonstruktion der einzelnen Polypeptide müssen mehrere Kenngrößen verwaltet werden:

- m_p: Masse des zu rekonstruierenden Proteins.
- m_{diff}: Betrag der Massenabweichung, um den m_p maximal unter- oder überschritten werden darf (durch beschränkte Messgenauigkeit des zur Analyse verwendeten Massenspektrometers bedingt).
- m_{poly}: Masse des Polypeptides, welches gerade rekonstruiert wird. Der Wert von m_{poly} entspricht der Masse der Peptide, welche für die Rekonstruktion der aktuellen SCC herangezogen werden.

- m_{cur}: Masse der bisher rekonstruierten Teillösung, die sich aus den Massen der bisher rekonstruierten Polypeptide ergibt.
- m_{p_i}: Masse des Peptides, das zu dem aktuell betrachteten Knoten gehört,
- *averageEdgeWeight*[]: Array, welches für jede SCC den Median der in ihr vorhandenen Kantengewichte enthält.
- *numberOfTraversels*[]: Array, das zu jeder Kante des Overlap-Graphen die Anzahl der Traversierungen verwaltet.
- *strongComponents*: Liste, welche sämtliche starken Zusammenhangskomponenten in der Form von Peptidlisten enthält.
- *visitedNodes*: Menge der Knoten, die während der Rekonstruktion des aktuellen Polypeptides bereits besucht wurden (enthält keine Mehrfachnennungen).
- *peptideOrder*: Reihenfolge in der die einzelnen zu den Knoten gehörigen Peptide in dem rekonstruierten Polypeptid auftauchen; Mehrfachnennungen sind möglich; spiegelt den innerhalb einer SCC abgeschrittenen Rekonstruktionspfad wieder.
- *parentNodes*: Liste sämtlicher Knoten, von denen aus der aktuell betrachtete Knoten p_i direkt bzw. indirekt über eine Folge von Kanten erreichbar ist; wird für den Backtracking-Mechanismus benötigt.
- *childNodes*: Liste sämtlicher Knoten, die von dem aktuell betrachteten Knoten p_i aus direkt erreicht werden können.
- *backtrackingStartingPoints*: Liste sämtlicher Knoten, von denen aus eine Backtracking-Phase begonnen wurde; dient der Begrenzung der im worst-case insgesamt durchzuführenden Backtracking-Phasen.
- *nextEdge*: Zufallsvariable für die Auswahl einer von mehreren ausgehenden Kanten per Tournierselektion; wird für Overlap-Graphen auf Basis approximativer Overlaps benötigt.

Der hier angegebene Algorithmus für die Polypeptidrekonstruktion (siehe Algorithmus 1) arbeitet in Abhängigkeit von der in Abschnitt 6.4 für die Berechnung der Overlaps verwendeten maximalen Levenshtein-Distanz *dis* unterschiedlich.

6.7.1 Rekonstruktion der Polypeptide unter Verwendung nicht-approximativer Overlaps

Zunächst wird das Vorgehen bei der Proteinrekonstruktion auf Basis nicht-approximativ berechneter Overlaps beschrieben.

Die Rekonstruktion jedes Polypeptids beginnt mit der Suche eines geeigneten Startknotens. Dazu sucht man sich aus der Knotenmenge der aktuell betrachteten SCC den ersten Knoten heraus, für den $m_{cur} + m_{p_i} \leq m_p + m_{diff}$ gilt. Dieses Vorgehen ist legitim, da es innerhalb einer SCC keine Knoten mit ausschließlich einer Art von inzidenten Kanten (eingehende oder ausgehende Kanten) gibt. Hieraus folgt, dass es keine prädestinierten Start- oder Endknoten gibt, von denen aus die Rekonstruktion des aktuellen Polypeptides gestartet werden sollte.

Wurde ein geeigneter Startknoten gefunden, so werden m_{cur}, m_{poly}, *visitedNodes* und *peptideOrder* entsprechend aktualisiert. Gibt es keinen solchen Knoten, so wird die Rekonstruktion des nächsten Polypeptides begonnen bzw. der Prozess der Rekonstruktion der Polypeptide mit der Behandlung der letzten SCC beendet.

Innerhalb des eigentlichen Rekonstruktionsalgorithmus ist bei der Ermittlung eines Nachfolgerknotens zu unterscheiden, ob der aktuelle Knoten eine oder mehrere ausgehende Kanten besitzt. Verfügt der aktuelle Knoten p_i über lediglich eine ausgehende Kante (*childNodes.size*() == 1), so wird für den zu ihm adjazenten Knoten p_j überprüft, ob die Bedingung $m_{cur} + m_{p_j} \leq m_p + m_{diff}$ gilt, also ob das zu dem Knoten p_j gehörige Peptid zu der bisher berechneten Rekonstruktion der aktuellen SCC hinzugenommen

Algorithmus 1 Rekonstruktion sämtlicher zu den SCCs gehöriger Polypeptide

Require: $strongComponents.size() > 0$

if $(dis == 0)$ **then**
  for $(i = 0; i < strongComponents.size(); i{+}{+})$ **do**
    Initialisiere m_p, m_{diff}, m_{poly}, m_{cur}, $averageEdgeWeight[]$, $numberOfTraversels[]$, $visitedNodes$, $peptideOrder$ und $backtrackingStartingPoints$.
    Wähle geeigneten Startknoten mit Eigenschaft $m_{cur} + m_{p_j} \leq m_p + m_{diff}$.
    Passe Werte von m_{poly}, m_{cur}, $visitedNodes$ und $peptideOrder$ entsprechend an. Bestimme $parentNodes$ für p_i.
    while $(visitedNodes.size() < strongComponents.get(i).size()$ && $m_{cur} \leq m_p + m_{diff})$ **do**
      if $(childNodes.size() == 0$ && $parentNodes.size() > 0)$ **then**
        Leite Backtracking ein, da der aktuelle Rekonstruktionspfad nicht fortgeführt werden kann. Vermerke den aktuellen Knoten in $backtrackingStartingPoints$ (siehe Abschnitt 6.7.3).
      end if
      if $(childNodes.size() == 1)$ **then**
        Gehe zum Nachfolgerknoten p_j, falls $m_{cur} + m_{p_j} \leq m_p + m_{diff}$ gilt (siehe Abschnitt 6.7.1).
      end if
      if $(childNodes.size() \geq 2)$ **then**
        Wähle den Nachfolgerknoten p_j in Abhängigkeit von den Kantengewichten der von p_i ausgehenden Kanten aus (siehe Abschnitt 6.7.1).
      end if
    end while
  end for
else if $(dis > 0)$ **then**
  for $(i = 0; i < strongComponents.size(); i{+}{+})$ **do**
    Initialisiere m_p, m_{diff}, m_{poly}, m_{cur} $averageEdgeWeight[]$, $numberOfTraversels[]$, $visitedNodes$, $peptideOrder$, $backtrackingStartingPoints$ und $nextEdge$.
    Wähle geeigneten Startknoten mit Eigenschaft $m_{cur} + m_{p_j} \leq m_p + m_{diff}$.
    Passe Werte von m_{poly}, m_{cur}, $visitedNodes$ und $peptideOrder$ entsprechend an. Bestimme $parentNodes$ für p_i.
    while $(visitedNodes.size() < strongComponents.get(i).size()$ && $m_{cur} \leq m_p + m_{diff})$ **do**
      if $(childNodes.size() == 0$ && $parentNodes.size() > 0)$ **then**
        Leite Backtracking ein, da der aktuelle Rekonstruktionspfad nicht fortgeführt werden kann. Vermerke den aktuellen Knoten in $backtrackingStartingPoints$ (siehe Abschnitt 6.7.3).
      end if
      if $(childNodes.size() == 1)$ **then**
        Gehe zum Nachfolgerknoten p_j, falls $m_{cur} + m_{p_j} \leq m_p + m_{diff}$ gilt (siehe Abschnitt 6.7.1).
      end if
      if $(childNodes.size() \geq 2)$ **then**
        Würfele aktuellen Wert der Zufallsvariable $nextEdge$ aus.
        Bestimme den Nachfolgerknoten p_j per linear skalierter Tournierselektion unter den von p_i ausgehenden Kanten. Verwende hierfür den aktuellen Wert von $nextEdge$ (siehe Abschnitt 6.7.2).
      end if
    end while
  end for
end if

werden darf. Ist dies der Fall, so werden die fünf Kenngrößen m_{cur}, m_{poly}, $visitedNodes$, $peptideOrder$ und $numberOfTraversels[e(p_i, p_j)]$ aktualisiert und die Rekonstruktion kann nach Senken des Kantengewichts der Kante $e(p_i, p_j)$ fortgesetzt werden. Das Senken des Kantengewichts der auf dem Rekonstruktionspfad abgeschrittenen Kanten dient der Vermeidung von Endlosschleifen während der Rekonstruktion. Wird ein Pfad wiederholt abgeschritten, so wird das Kantengewicht der abgeschrittenen Kanten jedes Mal gesenkt. Geschieht dies häufig genug, so werden Kanten von G bei der Rekonstruktion nicht länger berücksichtigt und es muss unter Umständen ein anderer Rekonstruktionspfad ermittelt werden (siehe Abschnitt 6.7.3). Dieser soll dann idealer Weise zu bisher noch nicht besuchten Knoten der aktuellen SCC führen. Der Betrag um den das Kantengewicht einer abgeschrittenen Kante gesenkt wird, setzt sich aus dem Produkt von $averageEdgeWeight[i]$ und der Anzahl der bisherigen Traversierungen der betrachteten Kante ($numberOfTraversels[e(p_i, p_j)]$) zusammen.

Besitzt der aktuell betrachtete Knoten p_i mehrere ausgehende Kanten ($childNodes.size() > 1$), so wird der Knoten zum Nachfolger von p_i, der mit dem aktuellen Knoten über die bzw. eine der Kanten mit

maximalem Kantengewicht e_{max} verbunden ist. Gibt es mehr als eine Kante mit Kantengewicht e_{max}, werden diese in einer separaten Liste verwaltet und erhalten einen Index, welcher ihrer Position in der Liste entspricht. Der Nachfolger von p_i wird per linear skalierter Tournierselektion unter diesen Kanten ausgewählt. Hierfür wird der Wert der Zufallsvariable *nextEdge* ausgewürfelt, die auf Grund des verwendeten Zufallszahlengenerator stets einen Betrag kleiner oder gleich der Anzahl der Kanten mit dem Kantengewicht e_{max} annimmt. Eine Kante $e(p_i, p_j)$ wird per Tournierselektion ausgewählt, falls der Wert der ausgewürfelten Zufallsvariable *nextEdge* dem Index der Kante $e(p_i, p_j)$ entspricht und $m_{cur} + m_{p_j} \leq m_p + m_{diff}$ gilt. Wird die Bedingung $m_{cur} + m_{p_j} \leq m_p + m_{diff}$ nicht erfüllt, so wird eine weitere Tournierselektion unter den Kanten mit Kantengewicht e_{max} durchgeführt. Sollte keine dieser Kanten wählbar sein, da keiner der von p_i aus über eine Kante mit Kantengewicht e_{max} erreichbaren Knoten mehr zu der Rekonstruktion hinzugefügt werden kann, so wird eine Backtracking-Phase eingeleitet (siehe Abschnitt 6.7.3). Konnte jedoch ein Nachfolgerknoten p_j ermittelt werden, so wird das Kantengewicht der gewählten Kante $e(p_i, p_j)$ soweit heruntergesetzt, dass bei der nächsten Überprüfung der Kantengewichte an dem Knoten p_i, der Nachfolger von p_i entweder wieder per Tournierselektion oder über die Kante mit dem bis dato zweithöchsten Kantengewicht bestimmt werden wird. Diese Maßnahme verhindert, dass Kreise auf dem Rekonstruktionspfad beliebig oft abgeschritten werden können. Das Senken der Kantengewichte führt aber nicht dazu, dass Kantengewichte negativ werden können. Fortgesetztes Senken des Kantengewichts einer Kante führt lediglich dazu, dass das Kantengewicht gegen Null geht und die Kanten beim Erreichen des Wertes Null aus den Adjazenzlisten der entsprechenden Knoten entfernt werden. Wurde das Kantengewicht der ausgewählten Kante $e(p_i, p_j)$ entsprechend verändert, so werden die Kenngrößen m_{cur}, m_{poly}, *visitedNodes*, *peptideOrder* und *numberOfTraversels*$[e(p_i, p_j)]$ aktualisiert.

6.7.2 Rekonstruktion der Polypeptide unter Verwendung approximativer Overlaps

Wurde für die Berechnung der Peptidüberlappungen in Abschnitt 6.4 eine Levenshtein-Distanz größer Null verwendet, wurde die Berechnung der Peptidüberlappungen daher approximativ durchgeführt, so arbeitet der Rekonstruktionsalgorithmus wie folgt.

Während des Testens mit synthetischen Testdaten und approximativ berechneten Overlaps stellte sich heraus, dass die Wahl eines Rekonstruktionspfades nach dem Greedy-Prinzip zu suboptimalen Rekonstruktionen führt. Analysen der zugehörigen berechneten Overlap-Graphen ergaben, dass bei Proteinrekonstruktionen auf der Basis approximativer Overlaps Rekonstruktionspfade entstehen, die von den Rekonstruktionspfaden der Proteinrekonstruktionen mittels nicht-approximativer Overlaps abweichen. Wurden die Overlaps nicht-approximativ berechnet, so glückte die Rekonstruktion des ursprünglichen Proteins in jedem der Tests, was auf eine geeignete Verkleinerung und Vereinfachung des zugehörigen Lösungsraums zurückzuführen war. Wurden die Overlaps aber approximativ berechnet, so führte dies oft dazu, dass der zugehörige Overlap-Graph zusätzliche Kanten enthielt, die ein höheres Kantengewicht als die Kanten des ursprünglichen Rekonstruktionspfades besaßen. Damit wurden diese neu hinzugekommenen Kanten gemäß des Greedy-Ansatzes den Kanten des ursprünglichen Rekonstruktionspfades vorgezogen.

Um nun die Rekonstruktion des ursprünglichen Proteins in Verbindung mit approximativ berechneten Overlaps effizienter zu gestallten, wurde der Rekonstruktionsalgorithmus wie folgt angepasst. Werden die Overlaps zwischen den identifizierten Peptiden approximativ berechnet, so gilt für Knoten mit mehreren ausgehenden Kanten, dass nun nicht mehr automatisch die Kante mit dem höchsten Kantengewicht zur Fortsetzung des weiteren Rekonstruktionspfades gewählt wird, sondern die als nächstes zu traversierende Kante mittels einer linear skalierten Tournierselektion bestimmt wird. Hierfür wird wieder der Wert der Zufallsvariable *nextEdge* ausgewürfelt, die aufgrund des verwendeten Zufallszahlengenerators stets einen Betrag kleiner oder gleich e_{max} annimmt. Der Betrag von e_{max} entspricht in diesem Fall dem maximalen Kantengewicht, der von dem aktuellen Knoten p_i ausgehenden Kanten. Eine vom aktuellen Knoten p_i ausgehende Kante $e(p_i, p_j)$ wird per Tournierselektion ausgewählt, falls $w(p_i, p_j) \geq nextEdge > w(p_i, p_k)$ und $m_{cur} + m_{p_j} \leq m_p + m_{diff}$ gilt. Da die von p_i ausgehenden Kanten gemäß ihres Kantengewichts absteigend sortiert sind, gilt für die Kante $e(p_i, p_k)$, dass sie die Kante mit dem zu $e(p_i, p_j)$ nächst kleineren Kantengewicht ist. Da die Generierung des Wertes der Zufallsvariable *nextEdge* abgesehen von dem Wert von e_{max} unabhängig von den Gewichten der vom dem aktuellen Knoten ausgehenden Kanten ist, besteht kein direkter Zusammenhang zwischen den Kantengewichten am aktuellen Knoten und der als nächstes zu traversierenden Kante. Wurde eine von p_i ausgehende Kante $e(p_i, p_j)$ ausgewählt, so werden auch hier wieder

die Kenngrößen m_{cur}, m_{poly}, *visitedNodes*, *peptideOrder* und *numberOfTraversels*$[e(p_i, p_j)]$ aktualisiert und das Kantegewicht der ausgewählten Kante gemäß des Produkts aus *numberOfTraversels*$[e(p_i, p_j)]$ und *averageEdgeWeight*$[i]$ gesenkt.

6.7.3 Backtracking-Mechanismus

Durch das Senken der Kantengewichte entlang des Rekonstruktionspfades kann es passieren, dass der aktuell betrachtete Knoten de facto keine ausgehenden Kanten mehr besitzt (siehe Abbildung 6.10). Gilt in einer solchen Situation $m_{cur} \leq m_p + m_{diff}$ und gibt es noch unbesuchte Knoten in der aktuellen SCC, so setzt ein mehrstufiger Backtracking-Mechanismus ein, dessen Aufgabe es ist, den Verlauf des bisherigen Rekonstruktionspfades zu ändern. Ansonsten würde die Rekonstruktion der aktuellen SCC an dieser Stelle beendet werden.

Damit der Aufwand für die Suche nach einem neuen Rekonstruktionspfad auf ein Minimum beschränkt bleibt, verläuft das Backtracking in mehreren Phasen. Ziel der Suche ist die Ermittlung eines direkten bzw. indirekten Vorgängers des aktuellen Knotens auf dem bisherigen Rekonstruktionspfad, von dem aus noch nicht besuchte Knoten erreicht werden können. Wird während der Ausführung einer Suchphase ein geeigneter Vorgänger gefunden, so endet mit dem Abschluss dieser Suchphase auch das Backtracking und die Rekonstruktion des aktuellen Polypeptides kann fortgesetzt werden.

In der ersten Phase des Backtrackings werden nur die unmittelbar erreichbaren Vorgänger des aktuellen Knotens untersucht. Ein direkter Vorgänger des aktuellen Knotens p_i wird genau dann in *parentNodes* vermerkt, falls er adjazent zu einem Knoten p_j mit den folgenden Eigenschaften ist:

- $p_j \neq p_i$
- $p_j \notin visitedPeptides$
- $m_{cur} + m_{p_j} \leq m_p + m_{diff}$

Gibt es mehrere direkte Vorgänger von denen aus die Rekonstruktion fortgesetzt werden könnte, hängt die Wahl des als nächsten zu besuchenden Knoten von den Kantengewichten zwischen Vorgänger und potentiellem Nachfolgerknoten ab. Wurde ein geeigneter Nachfolger ermittelt, so muss der in *peptideOrder* dokumentierte Rekonstruktionspfad angepasst werden. Dies bedeutet, dass ein Stück des vermerkten Rekonstruktionspfades, nämlich ab der letzten Nennung des gewählten direkten Vorgängers des aktuellen Knotens bis zu dem letzten als besucht vermerkten Knoten, entfernt werden muss. Der per Backtracking ermittelte bisher noch nicht besuchte Nachfolger wird als zuletzt besuchter Knoten vermerkt. Als Konsequenz der Veränderung des Rekonstruktionspfades müssen auch m_{cur}, m_{poly}, *visitedNodes* und die Einträge des Arrays *numberOfTraversels*[], die von der Veränderung des Rekonstruktionspfades betroffen sind, entsprechend angepasst werden (siehe Abbildung 6.11).

Sollte nach Beendigung der ersten Suchphase kein neuer Nachfolger feststehen, so wird in weiteren Suchphasen nach einem indirekten Vorgänger des aktuellen Knotens gesucht, der die oben genannten Bedingungen erfüllt (siehe Abbildungen 6.10 und 6.11). Der Backtracking-Mechanismus wird beendet sobald innerhalb einer der Suchphasen ein geeigneter Nachfolger ermittelt wurde oder keine neuen direkten bzw. indirekten Vorgänger mehr ermittelt werden konnten. Konnten weder ein direkter noch ein indirekter Vorgänger ermittelt werden, von welchem aus die Rekonstruktion fortgesetzt werden kann, so endet die Rekonstruktion der aktuellen SCC mit dem bisher rekonstruierten Polypeptid bzw. Peptid.

Wurde ein direkter oder indirekter Vorgänger ermittelt, von dem aus der Rekonstruktionspfad so verändert werden kann, dass dieser zu noch unbesuchten Knoten führt, so werden während der Anpassung des Rekonstruktionspfades auch die Kantengewichte der Kanten, die während des Backtrackings rückwärts abgeschritten wurden, auf den Wert vor der letzten Kantentraversierung zurückgesetzt. Zudem werden für diese Kanten die zugehörigen Einträge in dem Array *numberOfTraversels*[] entsprechend dekrementiert. Diese Maßnahmen sorgen dafür, dass die ursprüngliche Struktur der SCC, welche durch das Senken von Kantengewichten und das daraus resultierenden eventuellen Wegfallen von Kanten während der Rekonstruktion verändert wurde, wieder hergestellt wird (siehe Abbildung 6.11).

Die Rekonstruktion eines Polypeptides endet, falls sämtliche Knoten der aktuellen SCC bereits mindestens einmal besucht wurden oder falls das Hinzunehmen eines weiteren erreichbaren Peptides dazu führen

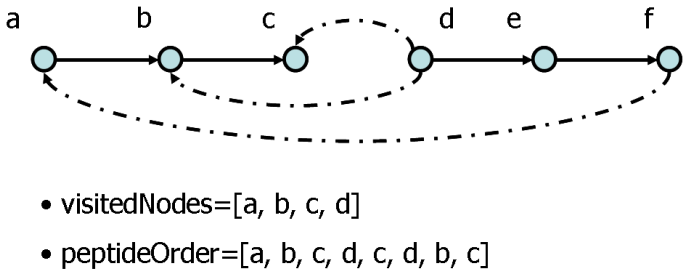

Abbildung 6.10: Beispiel eines Overlap-Graphen für den Backtracking-Mechanismus. Während der Rekonstruktion des dargestellten Proteins wurden vorhandene Kreise, an denen die Knoten b, c und d beteiligt sind, mehrmals abgeschritten. Die Kante zwischen den Knoten d und e sei dabei bisher noch nicht gewählt worden. Die Kante $e(c, d)$ wurde nach dem letzten Abschreiten aus G entfernt, da ihr Kantengewicht den Betrag Null annahm.

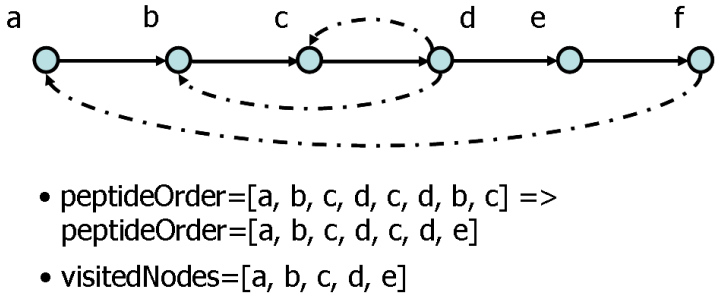

Abbildung 6.11: Darstellung des Ergebnisses des Backtracking-Mechanismus. Ausgehend von dem aktuellen Knoten, dem Knoten c, wird ein direkter oder indirekter Vorgänger von c ermittelt, von dem aus die Rekonstruktion des vorliegenden Proteins fortgesetzt werden kann. Dazu wird zunächst der direkte Vorgänger von c auf den bisherigen Rekonstruktionspfad untersucht: Dies ist der Knoten b. Von b aus lassen sich aber keine bisher noch nicht besuchten Knoten erreichen. Daher wird die Suche mit dem direkten Vorgänger von b fortgesetzt: Dies ist der Knoten d. Von d aus lassen sich die beiden einzigen bisher noch nicht besuchten Knoten e und f erreichen. Daher werden die Listen *peptideOrder* und *visitedNodes*, die Werte von m_{cur} und m_{poly} und die von der Veränderung der Rekonstruktionspfades betroffenen Einträge von *numberOfTraversels*[] entsprechend aktualisiert. Die Kantengewichte, der bei dem Backtracking rückwärts abgeschrittenen Kanten (in diesem Beispiel sind dies die Kanten $e(b, c)$ und $e(d, b)$), werden auf ihren Wert vor der zuletzt erfolgten Traversierung zurückgesetzt.

würde, dass $m_{cur} > m_p + m_{diff}$ gilt. Da der Rekonstruktionsalgorithmus auf der Basis approximativer als auch nicht-approximativer Overlaps das wiederholte Abschreiten von Kreis in G erlaubt, lässt sich die worst-case-Rechenzeit nicht durch $O(n + m)$ abschätzen. Vielmehr tritt der worst-case für den Fall ein, dass während einer Rekonstruktion ein Kreis in G wiederholt abgeschritten wird, der aus $m - 1$ Knoten von G besteht und deren Kanten das Kantengewicht e_{max} besitzen. e_{max} entspricht dabei dem Maximum sämtlicher in G enthaltener Kantengewichte. Im worst-case wird ein solcher Kreis e_{max} Mal abgeschritten bevor keine weiteren Knoten mehr zu der Rekonstruktion hinzugefügt werden können und der letzte noch fehlende Knoten per Backtracking entdeckt wird. Da die für eine Tournierselektion und eine Stufe des Backtrackings vorzunehmenden Operationen in $O(n + m)$ durchführbar sind, setzt sich die worst-case-Rechenzeit für die Proteinrekonstruktion aus $O(e_{max}\ (n + m))$ und einem Faktor B für die Anzahl der im worst-case maximal durchzuführenden Backtracking-Phasen zusammen. Daraus ergibt sich insgesamt die worst-case-Rechenzeit $O(B\ e_{max}\ (n + m)) = O(e_{max}\ (nm + m^2))$, da B im worst-case m der Anzahl der Knoten von G entspricht. Die Anzahl der im worst-case durchzuführenden Backtracking-Phasen ist durch m beschränkt, da während der Rekonstruktion einer SCC jeder Knoten der Ausgangspunkt einer Backtracking-Phase ist in *backtrackingStartingPoints* vermerkt wird und dieser, falls er während der Rekonstruktion einer SCC auf einem aus einer Backtracking-Phase resultierenden Rekonstruktionspfad erneut erreicht werden kann, aus der Auswahl der erreichbaren Nachfolgerknoten ausgeschlossen wird.

6.7.4 Zusammenfassen von Polypeptiden aufgrund von SCC-externen Tree- und Cross-Kanten

Wie in Abschnitt 6.6.1 dargestellt, enthält der Overlap-Graph SCC-externe Tree- und Cross-Kanten. Wurden, aus denen in Abschnitt 6.6.1 geschilderten Gründen, sämtliche SCC-externen Tree- und Cross-Kanten aus G entfernt, so hängt die Bestimmung eines optimalen ϵ-Layouts und der damit verbundenen rekonstruierten Proteinsequenz R, gänzlich von der Ermittlung einer geeigneten Permutation der rekonstruierten Polypeptide ab. Eine solche Permutation besitzt nach Definition des zugrunde liegenden Problems (siehe Abschnitt 5.4) eine minimale Abweichung δ für die beobachtete und tatsächliche Peptidstartpunktverteilungen der berechneten Lösung.

Besaßen die SCC-externen Tree- und Cross-Kanten jedoch ein ausreichend hohes Kantengewicht und waren sie nicht auf eine geringe biologische Variabilität des zu untersuchenden Proteins zurückzuführen, so wurden sie nicht aus G entfernt, sondern in einer separaten Liste verwaltet. Die in dieser Liste enthaltenen SCC-externen Kanten können im nun folgenden Schritt zur weiteren Vereinfachung des Gesamtrekonstruktionsprozesses verwendet werden.

Die verbliebenen SCC-externen Tree- und Cross-Kanten werden zunächst gemäß ihrer Kantengewichte absteigend sortiert. Anschließend wird die Liste gemäß dieser absteigenden Sortierung durchlaufen. Dies stellt sicher, dass das Zusammenfassen von Polypeptiden aufgrund der biologischen Signifikanz des gemeinsamen Overlaps zwischen diesen Polypeptiden geschieht.

Für jede der in der Liste enthaltenen Kanten wird überprüft, ob sich die an der jeweiligen Kante beteiligten Peptide an geeigneten Stellen innerhalb ihrer Polypeptide befinden. Dies bedeutet, dass sich das Peptid, von dem die betrachtete Kante ausgeht, an letzter Position in seinem Polypeptid befindet und das Zielpeptid, auf das die Kante verweist, entsprechend an erster Position in seinem Polypeptid befinden muss. Ist dies der Fall, so lassen sich die beiden zugehörigen Polypeptide zu einem insgesamt längeren Polypeptid zusammenfassen. Ansonsten wird mit der nächsten Kante in der Liste weitergemacht (siehe Abbildung 6.12). Durch diesen Zwischenschritt wird die Anzahl der insgesamt noch zu betrachtenden Polypeptidepermutationen weiter gesenkt. Der Lösungsraum wird weiter verkleinert.

Da sich die Liste der SCC-externen Kanten bzgl. der Kantengewichte der in ihr enthaltenen Kanten in $O(n\ log\ n)$ sortieren lässt und die Überprüfung, ob sich zwei starke Zusammenhangskomponenten mit Hilfe der jeweils aktuell betrachteten SCC-externen Kante zusammenfassen lassen, in linearer Zeit durchführen lässt, ergibt sich eine worst-case-Gesamtrechenzeit von $O(n\ log\ n)$, wobei n die Länge der Liste der SCC-externen Kanten angibt.

6.8 Ermittlung einer optimalen Rekonstruktion

Nachdem in den vorhergegangenen Berechnungsschritten die Anzahl der insgesamt zu betrachtenden Proteinrekonstruktionen durch das Filtern von Kontaminantionen (siehe Abschnitt 6.1), das Filtern von Infixen (siehe Abschnitt 6.2), die Identifikation und Rekonstruktion von Proteinsubstrukturen (siehe Abschnitte 6.6 und 6.7) und das Zusammenfassen solcher Substrukturen zu größeren Polypeptiden (siehe Abschnitt 6.7.4) systematisch verringert wurden, muss nun unter den verbliebenen potentiell korrekten Rekonstruktionen des zu identifizierenden Proteins ein optimales ϵ-Layout und der dazugehörige Rekonstruktionsstring R bestimmt werden.

Da die tatsächliche Aminosäuresequenz des ursprünglichen Proteins unbekannt ist und es diese zu ermitteln gilt, lässt sich die Güte einer berechneten Rekonstruktion nicht durch einen Sequenzvergleich zwischen berechneter und tatsächlicher Primärstruktur des ursprünglichen Proteins ermitteln. Die Güte einer ermittelten Rekonstruktion muss daher auf anderem Wege bestimmt werden. Der hierfür zu verwendende Mechanismus wurde bereits in Abschnitt 5.4 vorgestellt. Kernstück der dort definierten Fitnessfunktion δ sind die zwei Peptidstartpunktverteilungen D_{obs} und D_{src} deren Abweichung von δ berechnet wird. Um diese Abweichung für eine berechnete Proteinrekonstruktion bestimmen zu können, sind die folgenden drei Schritte notwendig.

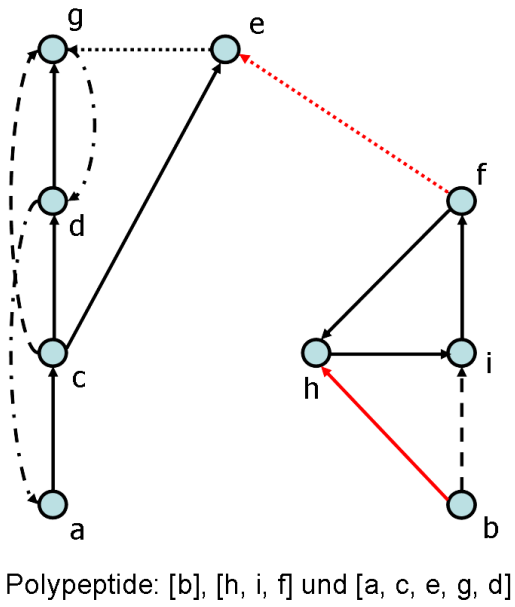

Abbildung 6.12: Zusammenfassen von Polypeptiden auf Grund von SCC-externen Tree- und Cross-Kanten. Nach der Rekonstruktion der zu den SCCs gehörigen Polypeptide wird überprüft, ob die beiden berechneten SCCs aufgrund der vorhandenen SCC-externen Tree- und Cross-Kanten zusammengefasst werden können. Nach der Bestimmung der SCCs gemäß Abschnitt 6.6.1, wurden zwei der vier SCC-externen Kanten entfernt; die Kanten $e(f, c)$ und $e(h, a)$. Mit Hilfe der beiden noch verbliebenen SCC-externen Kanten lassen sich die beiden Polypeptide $[b]$ und $[h, i, f]$ zusammenfassen. Die aus dieser Operation resultierenden Polypeptide $[b, h, i, f]$ und $[a, c, e, g, d]$ lassen sich allerdings nicht mehr weiter zusammenfassen, da die noch verbliebene SCC-externe Kante zwischen den Knoten f und e verläuft. Der Knoten e ist zwar Endknoten der von f ausgehenden Kante, er ist aber nicht an erster Stelle in der Peptidreihenfolge des Polypeptides, zu welchem der Knoten e gehört.

6.8.1 Bestimmung der beobachteten Peptidstartpunktverteilungen

Für jedes der aus dem Zusammenfassen der starken Zusammenhangskomponenten entstandene Paar aus ϵ-Layout und Rekonstruktionsstring R wird die jeweils zugehörige beobachtete Peptidstartpunktverteilung D_{obs} berechnet. Die beobachtete Peptidstartpunktverteilung D_{obs} einer Proteinrekonstruktion gibt für jedes der Peptide aus F, der Menge der identifizierten Peptide, eine Wahrscheinlichkeit an, gemäß derer das jeweilige Peptid ab einer vorgegebenen Position in seinem zugehörigen ϵ-Layout beginnt. Diese Wahrscheinlichkeit ergibt sich für jede der potentiellen Peptidstartpositionen aus dem Verhältnis zwischen der Menge der Peptide p_i, deren Aminosäuresequenz an einer vorgegebenen Startposition im Layout beginnt, und der Gesamtanzahl der Peptide aus F.

$$D_{obs}(x) = |\{p_i : s_i = x\}|/|F|.$$

Der Definitionsbereich von $D_{obs}(x)$ entspricht dabei $1 \leq x \leq |R|$, wobei $|R|$ die Länge der rekonstruierten Proteinsequenz angibt.

Die Bestimmung der einzelnen Peptidmengen einer Proteinrekonstruktion und der hieraus abgeleiteten beobachteten Peptidstartpunktverteilung D_{obs} kann in $O(n)$ geschehen, wobei n die Anzahl der potentiellen Peptidstartpunkte der jeweiligen Proteinrekonstruktion angibt. Für die Gesamtrechendauer ergibt sich $O(nm)$, wobei m die Anzahl der zu betrachtenden Proteinrekonstruktionen angibt.

6.8.2 Bestimmung der tatsächlichen Peptidstartpunktverteilung

Die Peptidstartpunktverteilung D_{src} kann, falls die Aminosäuresequenz des zu untersuchenden Proteins nicht bekannt ist, nicht direkt für dieses Protein berechnet werden. In Ermangelung einer umfangreicheren Datengrundlage und präziserer massenspektrometrischer Analysemethoden muss D_{src} approximiert

werden. Für diese Approximation nutzt man Peptidstartpunktverteilungen bereits identifizierter Proteine, diese stammen aus Proteinsequenzdatenbanken wie der NCBInr (www.ncbi.nlm.nih.gov), spezifischer IPI-Datenbanken (www.ebi.ac.uk/IPI/) oder einer Swiss-Prot-Datenbank (www.expasy.ch/sprot/sprot-top.html). Um die Peptidstartverteilung eines bereits identifizierten Proteins nachträglich bestimmen zu können unterwirft man es einem so genannten theoretischen Verdau, auch *in silico*-Verdau genannt.

Da die Approximation der tatsächlichen Peptidstartpunktverteilung für möglichst viele verschiedene Proteine korrekte Aussagen bzgl. der Startpunkte der in ihnen enthaltenen Peptide treffen soll, muss die Datengrundlage, auf Basis derer D_{src} berechnet wird, möglichst breit gewählt werden. Um dies zu gewährleisten wurde im Rahmen dieser Diplomarbeit die folgende Proteindatenbank für die Ableitung von D_{src} verwendet:

- **Bezeichnung**: NCBInr (non-redundant)
- **Stand**: 15.02.2006
- **Datenaufkommen**: 1,665 GB
- **Anzahl unterschiedlicher Proteine**: 3.292.317 Proteine

Die Wahl fiel auf eine NCBInr-Datenbank, da diese die Obermenge einer Vielzahl unterschiedlicher Proteindatenbanken (GenBank, EMBL, DDBJ, PDB, Swiss- Prot, PIR, PRF) darstellt. Zudem sind in der NCBInr Proteine aus den unterschiedlichsten Organismen vertreten (*Arabidopsis Thaliana*, *Bos Taurus*, *Neurospora Crassa*, usw.).

Um nun die Peptidstartpunktverteilung D_{src} zu ermitteln, führt man zunächst einen theoretischen Proteinverdau sämtlicher in der gewählten Proteindatenbank enthaltener Proteine durch. Bei einem theoretischen Verdau wird ein Protein, wie bei einem enzymatischen Verdau auch, in Peptide gespalten. Allerdings geschieht dies *in silico* und nicht wie bei Biomolekülen *in vitro*. Man verdaut also keine tatsächlichen Proteine, sondern zerlegt die Aminosäuresequenzen bereits identifizierter Proteine in Subsequenzen, welche die dabei entstehenden Peptide repräsentieren. Der theoretischen Verdau geschieht über einen Algorithmus, der Teil der Software Peakardt ist. Dieser liefert nach Angabe der Primärstruktur des zu verdauenden Proteins und nach Auswahl des für den Verdau zu verwendenden „Enzyms“ eine Liste mit den Aminosäuresequenzen und Massen der verdauten Peptide zurück (siehe Abbildung 6.13).

Anschließend lässt sich für jedes der theoretisch verdauten Proteine seine tatsächliche Peptidstartpunktverteilung berechnen, indem die aus dem theoretischen Verdau entstandenen Peptide dazu genutzt werden, um die verdauten Proteine zu rekonstruieren. Die Berechnung der tatsächlichen Peptidstartpunktverteilungen erfolgt dabei analog zu der Beschreibung im vorherigen Abschnitt.

Geht man davon aus, dass die beobachteten Peptidstartpunktverteilungen D_{obs_i}, sämtlicher aus einer Proteindatenbank stammender verdauter Proteine in einer Menge Dis der Mächtigkeit n enthalten sind, so lässt sich D_{src} wie folgt berechnen:

$$D_{src}(x) = \left(\sum_{i=1}^{n} D_{obs_i}(x) \right) * 1/n.$$

Der Definitionsbereich von $D_{src}(x)$ entspricht dabei wieder $1 \leq x \leq |R|$, wobei $|R|$ die Länge der oder einer der längsten rekonstruierten Proteinsequenzen angibt. Wendet man das beschriebene Vorgehen auf die Peptidstartpunkverteilungen der Proteine einer Proteindatenbank an, so erhält man als Ergebnis eine listenartige Repräsentation von D_{src}, die für jeden der potentiellen Startpunkte eines Peptides aus der Proteindatenbank eine Peptidstartpunktwahrscheinlichkeit angibt. Die Länge dieser Liste orientiert sich an der Anzahl der potentiellen Peptidstartpunkte des längsten in der Datenbank enthaltenen Proteins und stellt für Proteine dieser oder geringerer Länge eine entsprechende Approximation von D_{src} dar. Für die eben erwähnte Version der NCBInr vom 15. Februar 2006 ergibt sich aus dem beschriebenen Vorgehen eine 5208 Einträge umfassende Liste von Peptidstartpunktwahrscheinlichkeiten, welche als Approximation von D_{src} für die in Kapitel Sieben beschriebene Evaluierung des entwickelten *de novo*-Algorithmus verwendet wird.

Die Gesamtrechendauer für die Approximation von D_{src} entspricht der asymptotisch relevanten Berechnungsdauer für die in den Abschnitten 6.1 bis 6.8.1 angegebenen Algorithmen.

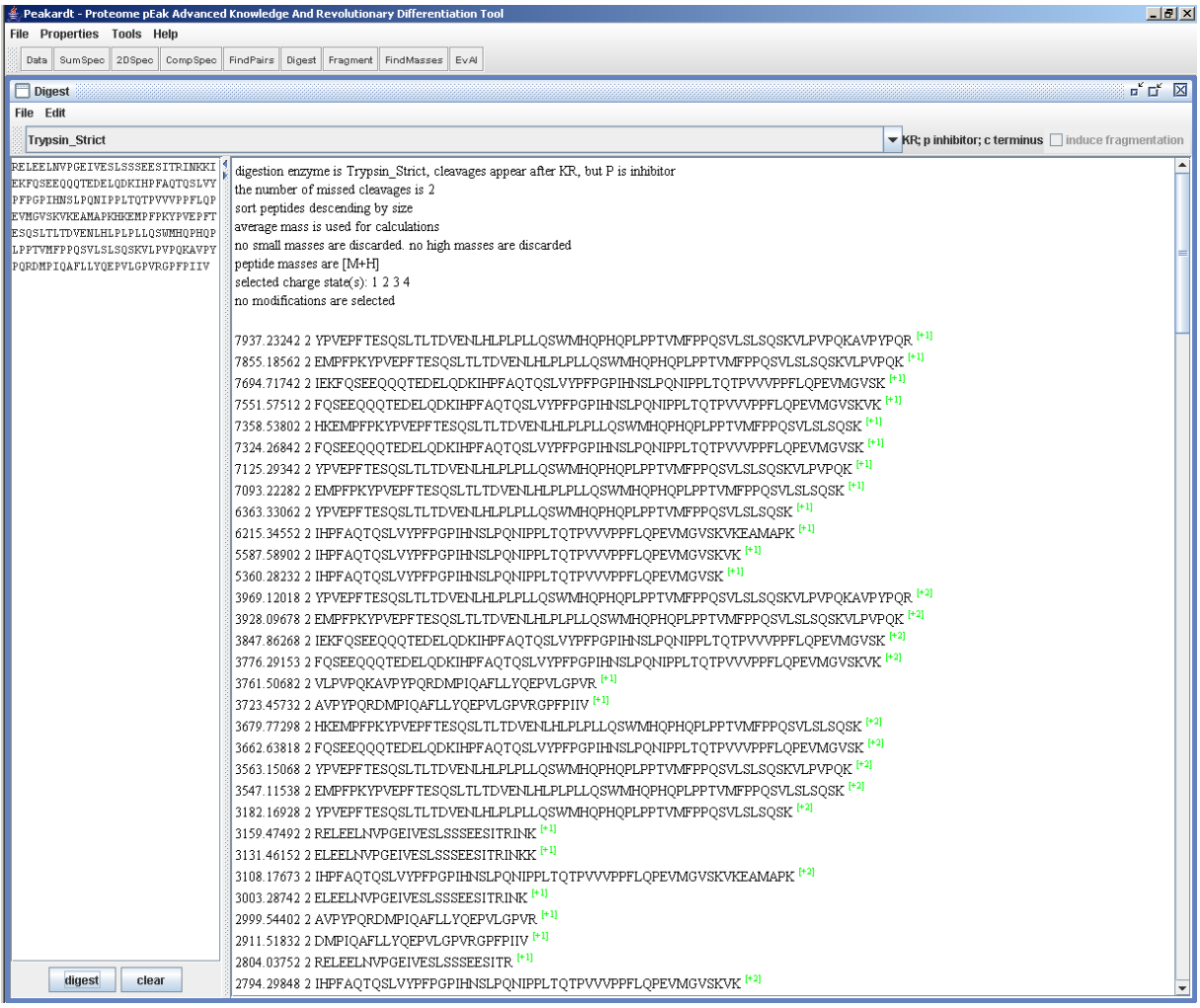

Abbildung 6.13: Screenshot eines theoretischen Verdaus durch die Software Peakardt. Im linken schmalen Teilfenster wird die Aminosäuresequenz des zu verdauenden Proteins angegeben. Mit Hilfe des Drop-Down-Menüs darüber lässt sich das zu verwendende Verdauungsenzym auswählen. Rechts neben dem Drop-Down-Menü werden die Substratspezifizitäten der ausgewählten Protease angegeben. Des Weiteren ist dort zu entnehmen, ob die Spaltung des Proteins an den entsprechenden C-Termini der spezifizierten Schnittstellen durch das Vorhandensein eines unmittelbar vorhergehenden Prolin-Moleküls gehemmt wird (p inhibitor oder p not inhibitor). Im rechten unteren Hauptteil des Digest-Dialogs werden die bei dem theoretischen Verdau entstandenen Peptide aufgelistet. Zu jedem Peptid werden seine Masse, seine Aminosäuresequenz und sein Ladungszustand angegeben. Direkt über dieser Auflistung werden die bei dem Verdau verwendeten Einstellungen des Verdau-Algorithmus angezeigt, dazu gehört die Angabe der verwendeten Protease inklusive ihrer Substratspezifizitäten und Inhibitoren, ob und wie viele definierte Sprungstellen übersprungen werden dürfen, gemäß welcher Kriterien die Einträge in der ausgegebenen Peptideliste sortiert wurden, ob die angegebene Masse die monoisotopische oder so genannte Average-Masse ist, ob Peptide unter- und oberhalb gewisser Massen bei der Erstellung der Ausgabe ignoriert wurden, welche möglichen Ladungszustände für die entstandenen Peptide berücksichtigt wurden und ob und vor allem welche post-translationalen Modifikationen bei der Durchführung des Verdaus berücksichtigt wurden.

6.8.3 Berechnung der Abweichung δ

Um nun festzustellen, welches der berechneten Paare aus ϵ-Layout und Rekonstruktionsstring R, das ursprüngliche Protein am präzisesten beschreibt, wird die in Abschnitt 5.4 definierte Fitnessfunktion δ angewendet. Um die maximale Abweichung zwischen D_{src} und der beobachteten Peptidstartpunktverteilung D_{obs}, der zu bewertenden Proteinrekonstruktion, zu berechnen, wird die Definition von δ angewendet:

$$\delta = max_{1 \leq x \leq |R|} |D_{obs}(x) - D_{src}(x)|$$

Da für jede berechnete Abweichung δ vermerkt wird, wie groß die maximale Abweichung zwischen den jeweiligen Peptidstartpunktverteilungen D_{obs} und D_{src} ist, kann nach Abschluss sämtlicher Berechnungen festgestellt werden, welche Proteinrekonstruktion dem ursprünglichen Protein am ähnlichsten bzw. identisch zu dem ursprünglichen Protein ist. Sollte keine der untersuchten Rekonstruktionen eine zu D_{src} identische Peptidstartpunktverteilung aufweisen, wird die Rekonstruktion bzw. werden sämtliche Rekonstruktionen mit der geringsten Abweichung zu D_{src} als Lösung ausgegeben.

Nachdem sämtliche vorherigen Rekonstruktionsschritte erfolgt sind, lässt sich dieser finale Schritt, bezogen auf die Anzahl der insgesamt berechneten Rekonstruktionen, in linearer Zeit durchführen.

An dieser Stelle sollte der Ehrlichkeit halber erwähnt werden, dass für den hier entwickelten Lösungsansatz für die *de novo*-Proteinidentifikation im Hinblick auf Proteine von bislang nicht sequenzierten Organismen, deren Genome sich im Prinzip sehr stark von denen bereits untersuchter Organismen unterscheiden können, keinerlei Garantien bzgl. der Qualität der berechneten Lösungen geben werden können. Dies ist auf die in Abschnitt 6.8.2 beschriebene Methode zur Berechnung der tatsächlichen Peptidstartpunktverteilung D_{src} zurückzuführen.

Kapitel 7

Evaluierung

Nachdem in vorherigen Kapiteln die Konzeption (siehe Kapitel Fünf) und Realisierung (siehe Kapitel Sechs) des entwickelten Algorithmus erfolgte, beschreibt dieses Kapitel dessen Evaluierung. Die Evaluierung des entwickelten *de novo*-Proteinidentifikationsalgorithmus findet in zwei Testphasen statt. Um den Algorithmus zunächst unter möglichst praxisnahen gut kontrollierbaren Testbedingungen mit einer möglichst großen heterogenen Testdatenmenge testen zu können, werden in der ersten Phase theoretische Testdaten als Eingabe verwendet. Um die hierfür notwendigen Testdatensätze zu erzeugen, wird auf bereits identifizierte Proteine und den *in silico*-Verdau zurückgegriffen (siehe Abschnitt 7.1). Die hierfür verwendeten Proteine stammen aus Proteindatenbanken wie der NCBInr oder IPI. Nach Abschluss der ersten Testphase, kommen in der zweiten Testphase reale Testdatensätze, wie sie auch bei der datenbankgestützten Proteinidentifikation verwendet werden, bei der Evaluierung zum Einsatz (siehe Abschnitt 7.2). Da die Erzeugung eines solchen Testdatensatzes einen ungleich aufwändigeren *in vitro*-Verdau eines realen Biomoleküls erfordert, ist die in der zweiten Testphase verwendete Testdatenmenge erheblich kleiner als in Testphase eins.

Um die im Folgenden zu präsentierenden Testergebnisse zu ermitteln, wurde der implementierte *de novo*-Proteinidentifikationsalgorithmus auf einem Dell Dimension 8400 gestestet. Dieser Testrechner ist mit einem Intel Pentium 4 (3.2 GHz) und einem ein Gigabyte großen Hauptspeicher ausgerüstet. Da der zu testende Algorithmus in Java implementiert ist, wurde für dessen Evaluierung die aktuellste Version der Java Vitual Machine der Firma Sun (JRE 1.5_06) verwendet.

7.1 Testläufe auf der Basis *in silico*-verdauter Proteine

Da die Erzeugung von Testdatensätzen auf der Basis von realen Proteinen zeit- und kostenaufwändig ist, hierfür ein Massenspektrometer und ein Laborant mit entsprechender Erfahrung benötigt wird und der *de novo*-Ansatz hohe Qualitätsanforderungen an die zu erzeugenden Testdaten stellt (die Masse des zu identifizierenden Proteins muss präzise bestimmt werden, die Aminosäuresequenz des zu identifizierenden Proteins muss vollständig durch Peptide überdeckt werden), wird in der ersten Testphase auf der Basis von theoretischen Testdatensätzen getestet.

Hierfür werden 45 bereits identifizierte Proteine verschiedenen Ursprungs ausgewählt. Von diesen 45 Proteinen stammen fünf aus einem so genannten internen Standard des MPC. Dieser stellt ein Proteingemisch dar, dessen Inhalt wohldefiniert ist und für die Kalibrierung von Massenspektrometern oder für Vergleichsvermessungen mit Proteingemischen unbekannten Inhalts verwendet wird (siehe Tabelle 7.1). Zwanzig weitere Proteine stammen aus der Human-Proteindatenbank des internationalen Proteinindex (häufig als IPI.human bezeichnet). Die hierfür verwendete Version 3.14 dieser Datenbank stammt vom 24.01.06 (siehe Tabelle 7.2). Die letzten zwanzig Proteine stammen aus der NCBInr-Datenbank vom 15.02.2006 (siehe Tabelle 7.3). Da das NCBI für seine Proteindatenbanken keine Versionsnummern vergibt, werden diese hinsichtlich ihres Erscheinungsdatums voneinander unterschieden.

Zum Zwecke der Testdatenerzeugung werden die zufällig ausgewählten Proteine einem spezifischen *in silico*-Verdau durch die in Tabelle 7.4 angegebenen Proteasen unterworfen. Die hieraus resultierenden Peptidmengen werden anschließend bzgl. der Massen ihrer Peptide gefiltert. Peptide mit einer Masse

Index	Accession	Masse (Da)	Sequenzlänge (AS)	mit Infixen	ohne Infixe	Quelle
1	>gi|115698	57.585,63	506	617	266	interner Standard
2	>gi|229351	11.458,33	102	141	73	interner Standard
3	>gi|1351907	69.293,42	607	759	468	interner Standard
4	>gi|1942750	16.979,49	153	179	103	interner Standard
5	>gi|2194089	18.309,26	162	171	85	interner Standard

Tabelle 7.1: Zusammenstellung der fünf Testproteine aus einem der internen Standards des MPC, die für die erste Testphase der Evaluierung verwendet werden. Die angegebenen Datenbank-Accessions entsprechen den Einträgen dieser Proteine in der hier verwendeten Version der NCBInr (Stand vom 15.02.2006). In der Spalte *Sequenzlänge* werden die Längen der Aminosäuresequenzen der Proteine in Aminosäuren (AS) angegeben. Die Einträge in der Spalte *mit Infixen* geben die Anzahl der Peptide an, die nach dem Filtern von Peptiden mit einer Masse kleiner als 500 oder größer als 7000 Dalton noch für die Rekonstruktion eines Proteins zu Verfügung stehen. In der Spalte *ohne Infixe* wird dagegen für jedes der Testprotein angegeben, wie viele nichtredundante Peptide in der Eingabe vorhanden sind. Die Abfolge der Tabellenzeilen entspricht der Ordnung der Protein-Accessions.

Index	Accession	Masse (Da)	Sequenzlänge (AS)	mit Infixen	ohne Infixe	Quelle
6	>IPI:IPI00002878.1	43.453,02	390	270	173	IPI.human
7	>IPI:IPI00002894.2	123.631,13	1107	983	606	IPI.human
8	>IPI:IPI00002957.1	72.654,20	648	519	235	IPI.human
9	>IPI:IPI00003021.1	112.265,44	1020	951	425	IPI.human
10	>IPI:IPI00003081.3	41.801,28	402	160	52	IPI.human
11	>IPI:IPI00003176.1	51.286,96	480	374	227	IPI.human
12	>IPI:IPI00003293.1	63.927,05	567	542	302	IPI.human
13	>IPI:IPI00017202.2	30.430,71	273	212	79	IPI.human
14	>IPI:IPI00145107.3	34.262,92	285	362	149	IPI.human
15	>IPI:IPI00146077.4	93.547,68	834	763	536	IPI.human
16	>IPI:IPI00147874.1	40.307,51	359	375	184	IPI.human
17	>IPI:IPI00151121.5	157.972,51	1380	1254	789	IPI.human
18	>IPI:IPI00151141.1	134.739,27	1243	874	535	IPI.human
19	>IPI:IPI00291005.7	36.294,93	333	344	203	IPI.human
20	>IPI:IPI00291076.5	90.734,87	823	704	267	IPI.human
21	>IPI:IPI00291136.3	108.547,51	1028	930	497	IPI.human
22	>IPI:IPI00291215.4	184.342,38	1638	1759	748	IPI.human
23	>IPI:IPI00448673.3	75.738,73	678	467	146	IPI.human
24	>IPI:IPI00479313.1	147.788,11	1373	944	320	IPI.human
25	>IPI:IPI00654646.1	206.025,82	1880	1428	867	IPI.human

Tabelle 7.2: Zusammenstellung der zwanzig Testproteine aus der IPI.human, die für die erste Phase der Evaluierung verwendet werden. Die angegebenen Testproteine stammen aus der Version 3.14 der Datenbank. Die Nachkommastelle bei IPI-Accessions gibt die Versionsnummer des, durch den vorderen Teil der Accession, identifizierten Proteins an. Die Abfolge der Tabellenzeilen entspricht der Ordnung der Protein-Accessions.

kleiner als 500 oder größer als 7000 Dalton (7 kDa) werden aus der Eingabe des zu testenden *de novo*-Rekonstruktionsalgorithmus entfernt. Anschließend wird die Peptidmenge hinsichtlich vorhandener Infixe gefiltert (siehe Abschnitt 6.2). Da bei einem *in silico*-Verdau eines Proteins keine Probenkontaminationen auftreten können, müssen die Peptidemengen nicht bzgl. des Auftretens von Probenkontaminationen gefiltert werden. Die Tabellen 7.1, 7.2 und 7.3 geben für jedes der *in silico* verdauten Proteine die Anzahl der bei seinem spezifischen Verdau entstehen Peptide, sowie die Anzahl der Peptide, die nach dem Filtern bzgl. der Peptidemassen und eventueller Infixe noch übrig bleiben, getrennt nach Ursprung des Proteins an. Die so entstehenden Peptidemengen, sowie die Massen der darin enthaltenen Peptide und die Masse des zu identifizierenden Proteins dienen als Eingaben für den Rekonstruktionsalgorithmus, wobei die Rekonstruktion der Proteine sowohl auf der Basis nicht-approximativer (siehe Abschnitt 7.1.1), als auch auf der Basis approximativer Overlaps (siehe Abschnitt 7.1.2) stattfindet.

Index	Accession	Masse (Da)	Sequenzlänge (AS)	mit Infixen	ohne Infixe	Quelle
26	>gi|225472	133.062,74	1176	1053	341	NCBInr
27	>gi|384312	92.893,60	838	563	278	NCBInr
28	>gi|1586823	171.258,95	1558	1311	566	NCBInr
29	>gi|1588659	67.129,25	628	401	162	NCBInr
30	>gi|7108333	140.744,82	1237	1245	535	NCBInr
31	>gi|18676480	169.028,01	1512	1365	521	NCBInr
32	>gi|18676488	145.539,25	1326	1061	723	NCBInr
33	>gi|34329249	106.233,42	983	645	272	NCBInr
34	>gi|38566905	155.444,45	1391	1149	623	NCBInr
35	>gi|38570346	33.738,93	303	162	111	NCBInr
36	>gi|45646096	25.778,42	235	174	123	NCBInr
37	>gi|49525773	100.222,19	883	827	213	NCBInr
38	>gi|50759309	284.427,86	2500	2105	1833	NCBInr
39	>gi|55773132	37.438,96	337	172	165	NCBInr
40	>gi|67539156	80.237,20	708	745	412	NCBInr
41	>gi|67986958	129.769,46	1180	1100	478	NCBInr
42	>gi|68245710	153.112,42	1361	1016	496	NCBInr
43	>gi|78364360	32.171,33	295	208	132	NCBInr
44	>gi|78773889	117.467,38	1174	1120	899	NCBInr
45	>gi|78883544	77.206,11	695	640	256	NCBInr

Tabelle 7.3: Zusammenstellung der zwanzig Testproteine aus der NCBInr (Stand vom 15.02.2006), die für die erste Phase der Evaluierung verwendet werden. Die Abfolge der Tabellenzeilen entspricht der Ordnung der Protein-Accessions.

Protease	spezifische Schnittstellen
Trypsin (strict)	Arginin (R) & Lysin (K)
Chymotrypsin	Phenylalanin (F), Tryptophan (W) und Tyrosin (Y)
Glu-C	Asparaginsäure (D) & Glutaminsäure (E)
Lys-C	Lysin (K)

Tabelle 7.4: Zusammenstellung der bei den Tests mit *in silico* und *in vitro* verdauten Proteinen verwendeten Proteasen und ihrer spezifischen Schnittstellen. Die Auswahl der Proteasen wurde aufgrund der Aminosäuresequenzen der verwendeten Testproteine getroffen und stellt sicher, dass bei dem Verdau der Testproteine eine vollständige Sequenzabdeckung der zu identifizierenden Proteine erreicht wird.

7.1.1 Rekonstruktion mittels nicht-approximativer Overlaps

Um eine Proteinrekonstruktion auf der Basis nicht-approximativer Overlaps durchführen zu können, werden zusätzlich zu den identifizierten Peptiden, deren Massen, Scores und der Masse des zu rekonstruierenden Proteins, noch zwei weitere Parameter benötigt: Die minimale Overlap-Länge *mol* und die bei der Rekonstruktion des zu identifizierenden Proteins maximal zugelassene Massentoleranz m_{diff}. Für die ausgewählten Testproteine ergeben sich, bei einer minimalen Overlap-Länge von zwei und einer maximalen Massentoleranz von 1,0 Dalton, die in den Tabellen 7.5, 7.6 und 7.7 zusammengefassten Ergebnisse. Die Rekonstruktion der 45 ausgewählten Testproteine gelang auf Basis nicht-approximativer Overlaps in jedem der 45 Testläufe. Die durchschnittliche Rechendauer belief sich auf 22,07 Sekunden.

Weitere Tests mit einer minimalen Overlap-Länge *mol* größer oder gleich drei ergaben, dass die Anzahl der korrekt rekonstruierten Proteine mit wachsendem *mol* kontinuierlich abnimmt. Nahm die minimale Overlap-Länge einen Wert größer gleich fünf an, so konnte keines der 45 Protein korrekt rekonstruiert werden (siehe Abbildung 7.1).

Index	Overlaps (%)	SCCs (%)	Polypeptide (%)	Layouts (%)	Gesamt (sec.)	Identifiziert
1	70	11	12	7	21,39	ja
2	89	5	4	1	0,93	ja
3	72	11	10	7	33,83	ja
4	86	5	5	4	1,33	ja
5	86	6	5	3	1,02	ja

Tabelle 7.5: Zusammenstellung der Testergebnisse für die Rekonstruktion der ersten fünf Testproteine auf der Basis nicht-approximativer Overlaps. Die Werte in den Spalten *Overlaps*, *SCCs*, *Polypeptide* und *Layouts* geben für jede der Rekonstruktionsphasen an, wie groß ihr Anteil an der Gesamtrechendauer ist. Die in der Spalte *Gesamt* angegebenen Werte entsprechen den Rechenzeiten für den gesamten Rekonstruktionsprozess der einzelnen Proteine. In der Spalte mit der Bezeichnung *Identifiziert* wird angegeben, ob die Proteinrekonstruktion erfolgreich war.

Index	Overlaps (%)	SCCs (%)	Polypeptide (%)	Layouts (%)	Gesamt (sec.)	Identifiziert
6	71	10	11	8	21,77	ja
7	68	8	12	12	61,81	ja
8	72	13	10	5	36,18	ja
9	73	15	11	4	56,95	ja
10	70	12	8	10	22,44	ja
11	72	19	7	2	26,80	ja
12	67	15	12	6	31,66	ja
13	84	8	6	2	15,24	ja
14	83	12	3	2	15,91	ja
15	72	13	10	5	46,56	ja
16	73	15	10	2	20,04	ja
17	68	10	13	8	77,05	ja
18	67	12	13	8	69,40	ja
19	88	8	3	1	18,59	ja
20	72	12	11	5	45,95	ja
21	69	14	13	4	57,40	ja
22	71	14	11	4	91,45	ja
23	72	13	9	6	37,85	ja
24	69	13	12	6	76,66	ja
25	75	9	5	11	104,96	ja

Tabelle 7.6: Zusammenstellung der Testergebnisse für die Rekonstruktion der zwanzig Testproteine aus der IPI-Proteindatenbank auf der Basis nicht-approximativer Overlaps.

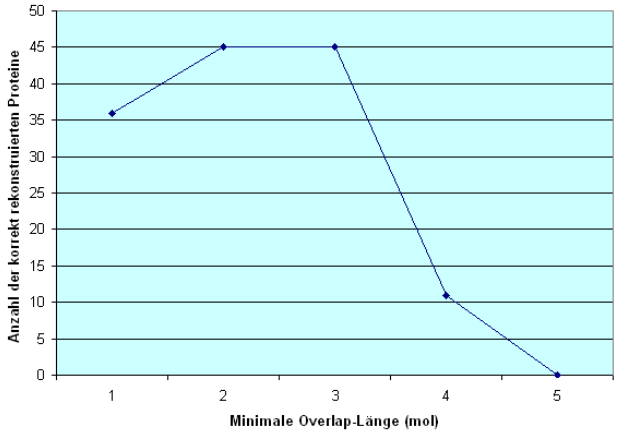

Abbildung 7.1: Graphische Darstellung des Zusammenhangs zwischen dem gewählten Wert für die minimale Overlap-Länge *mol* und der Anzahl der insgesamt korrekt rekonstruierten Proteine.

Index	Overlaps (%)	SCCs (%)	Polypeptide (%)	Layouts (%)	Gesamt (sec.)	Identifiziert
26	73	15	5	7	65,66	ja
27	71	10	9	10	46,79	ja
28	68	18	10	4	86,99	ja
29	82	10	6	2	35,06	ja
30	69	12	13	6	69,06	ja
31	72	15	12	1	84,42	ja
32	65	19	10	6	74,03	ja
33	75	8	11	6	54,89	ja
34	70	19	8	3	77,66	ja
35	89	4	5	2	16,92	ja
36	88	6	3	3	13,12	ja
37	74	12	10	4	46,30	ja
38	65	17	3	15	139,58	ja
39	83	10	4	3	18,82	ja
40	69	13	12	5	39,53	ja
41	75	11	10	4	65,88	ja
42	72	11	13	4	75,99	ja
43	87	10	2	1	16,47	ja
44	73	12	12	3	65,55	ja
45	69	15	11	5	38,80	ja

Tabelle 7.7: Zusammenstellung der Testergebnisse für die Rekonstruktion der zwanzig Testproteine aus NCBInr-Proteindatenbank auf der Basis nicht-approximativer Overlaps.

7.1.2 Rekonstruktion mittels approximativer Overlaps

Um Proteinrekonstruktionen auf der Basis approximativer Overlaps durchführen zu können, wird neben den im letzten Abschnitt aufgeführten Eingabedaten, noch eine Angabe bzgl. der maximal erlaubten Editierdistanz *dis* benötigt. Diese entspricht der maximal erlaubten Levenshtein-Distanz, um welche sich die Aminosäuresequenzen zweier Peptide hinsichtlich einer gemeinsamen Subsequenz unterscheiden dürfen (siehe Abschnitt 6.4.2).

Da die Polypeptidrekonstruktion bei der Proteinidentifikation auf der Basis approximativer Overlaps für die Ermittlung der Rekonstruktionspfade auf Tournierselektionen zurückgreift (siehe Abschnitt 6.7), müssen Testläufe zu einer konkreten Eingabe mehrfach wiederholt werden. Um die Eigenschaften der randomisierten Polypeptiderekonstruktion objektiv testen zu können, werden Rekonstruktionen bzgl. konkreter Eingaben jeweils einhundert Mal ausgeführt. Nach Abschluss einer solchen Rekonstruktion wird festgestellt, ob das Protein korrekt rekonstruiert wurde und wie lange die einzelnen Rekonstruktionsphasen jeweils gedauert haben. Für die ausgewählten Testproteine ergaben sich, bei einer minimalen Overlap-Länge von drei, einer maximalen Massentoleranz von 1,0 Dalton und einer maximalen Levenshtein-Distanz von eins, die in den Tabellen 7.8, 7.9 und 7.10 zusammengefassten Ergebnisse. Die Rekonstruktion der 45 ausgewählten Testproteine gelang auf der Basis approximativer Overlaps in durchschnittlich 84% der durchgeführten Testläufe. Die durchschnittliche Rechendauer belief sich auf 117,99 Sekunden.

Index	Overlaps (%)	SCCs (%)	Polypeptide (%)	Layouts (%)	Gesamt (sec.)	Identifiziert
1	72	9	10	9	62,26	88
2	89	2	2	6	1,80	86
3	75	13	7	7	94,95	100
4	85	3	4	8	2,69	83
5	86	3	4	7	2,85	89

Tabelle 7.8: Zusammenstellung der Testergebnisse für die Rekonstruktion der ersten fünf Testproteine auf der Basis approximativer Overlaps. Die Werte in den Spalten *Overlaps*, *SCCs*, *Polypeptide* und *Layouts* geben für jede der Rekonstruktionsphasen an, wie groß ihr Anteil an der Gesamtrechendauer ist. Die in der Spalte *Gesamt* angegebenen Werte entsprechen den Medianen der pro Protein ermittelten einhundert Rechenzeiten für die Proteinrekonstruktion als solche. In der Spalte mit der Bezeichnung *Identifiziert* wird für jedes der Testproteine die Anzahl der gelungenen Proteinrekonstruktionen angegeben.

Index	Overlaps (%)	SCCs (%)	Polypeptide (%)	Layouts (%)	Gesamt (sec.)	Identifiziert
6	72	8	8	12	61,01	79
7	71	5	12	12	173,16	83
8	75	10	5	10	11,41	86
9	78	10	4	11	17,96	89
10	84	3	3	10	7,08	90
11	76	10	4	10	8,45	78
12	79	3	6	13	9,99	76
13	84	2	3	11	4,81	88
14	82	5	4	9	4,98	89
15	82	5	2	11	130,46	83
16	76	5	4	15	6,32	82
17	81	4	3	12	215,87	78
18	79	4	3	14	194,44	84
19	88	3	3	6	5,87	88
20	76	5	2	17	128,74	81
21	72	6	4	18	160,80	73
22	81	7	3	9	256,22	75
23	75	3	2	20	106,06	86
24	73	5	4	18	214,77	79
25	82	3	1	14	294,08	76

Tabelle 7.9: Zusammenstellung der Testergebnisse für die Rekonstruktion der zwanzig Testproteine aus der IPI-Proteindatenbank auf der Basis approximativer Overlaps.

Index	Overlaps (%)	SCCs (%)	Polypeptide (%)	Layouts (%)	Gesamt (sec.)	Identifiziert
26	80	2	1	17	183,96	85
27	79	4	2	15	131,08	88
28	76	8	4	12	243,71	89
29	83	1	1	15	98,23	78
30	79	3	3	17	193,50	89
31	72	6	4	18	236,51	81
32	74	7	3	16	207,42	84
33	83	3	2	12	153,76	83
34	79	8	2	11	217,59	82
35	86	4	3	7	47,41	89
36	82	4	2	12	36,76	86
37	76	5	3	16	138,12	79
38	71	6	2	11	391,06	84
39	82	5	2	11	52,72	89
40	75	7	2	16	110,75	91
41	80	3	1	16	184,58	85
42	83	4	1	12	212,89	92
43	85	3	1	11	46,15	80
44	78	4	2	16	183,64	85
45	81	3	1	15	108,72	86

Tabelle 7.10: Zusammenstellung der Testergebnisse für die Rekonstruktion der zwanzig Testproteine aus NCBInr-Proteindatenbank auf der Basis approximativer Overlaps.

Werden die Testläufe für die Selben Proteine mit anderen Werten für die maximale Levenshtein-Distanz dis und/oder die minimale Overlap-Länge mol wiederholt, so ergeben sich in Bezug auf die Anzahl der korrekt rekonstruierten Proteine die in den Abbildungen 7.2, 7.3 und 7.4 dargestellten Mittelwerte (Median). Die Anzahl der insgesamt korrekt rekonstruierten Proteine nimmt mit wachsendem dis und mol kontinuierlich ab. Bei einem Wert von $dis = 1$ konnten noch 42 der 45 Testproteine korrekt rekonstruiert werden ($mol = 3$). Für $dis = 3$ können nur noch maximal 15 der 45 Proteine korrekt rekonstruiert werden ($mol = 3$).

Die in Abbildungen 7.2, 7.3 und 7.4 dargestellten Ergebnisse zeigen, dass die Erhöhung der minimalen

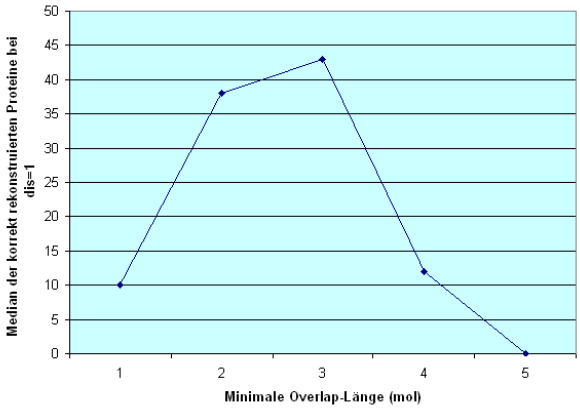

Abbildung 7.2: Graphische Darstellung des Zusammenhangs zwischen den gewählten Werten für die minimale Overlap-Länge *mol* und der Anzahl der durchschnittlich korrekt rekonstruierten Proteine bei einer maximalen Levenshtein-Distanz $dis = 1$.

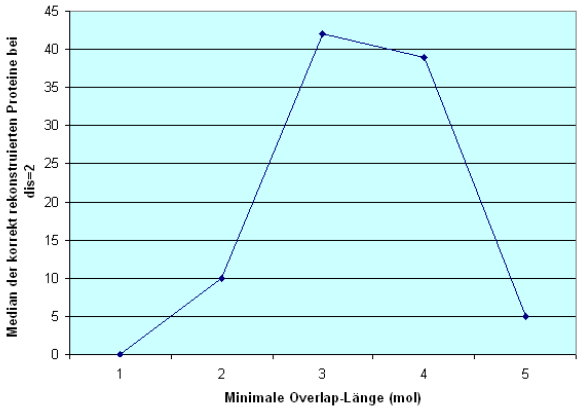

Abbildung 7.3: Graphische Darstellung des Zusammenhangs zwischen den gewählten Werten für die minimale Overlap-Länge *mol* und der Anzahl der durchschnittlich korrekt rekonstruierten Proteine bei einer maximalen Levenshtein-Distanz $dis = 2$.

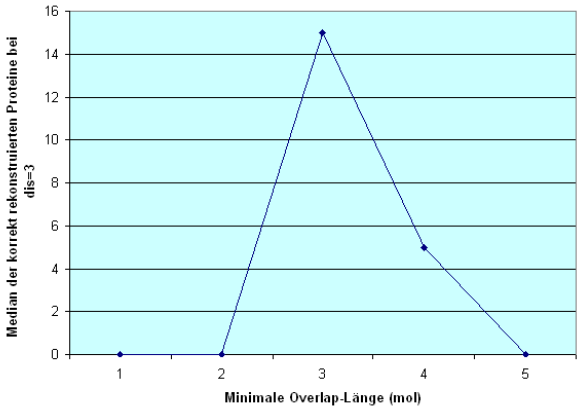

Abbildung 7.4: Graphische Darstellung des Zusammenhangs zwischen den gewählten Werten für die minimale Overlap-Länge *mol* und der Anzahl der durchschnittlich korrekt rekonstruierten Proteine bei einer maximalen Levenshtein-Distanz $dis = 3$.

Overlap-Länge, bis zu einem gewissen Grad, dazu in der Lage ist, die durch die Erhöhung der maximalen Levenshtein-Distanz bedingte anwachsende Anzahl an zu betrachtenden Proteinrekonstruktionen

zu senken. Nimmt die minimale Overlap-Länge aber einen verhältnismäßig hohen Wert an (für die hier verwendeten Testdaten einen Wert größer vier), kehrt sich dieser Selektionsprozess ins Gegenteil um. Statt falsche Proteinrekonstruktionen von vornherein auszuschließen, werden korrekte Rekonstruktionen aus dem Lösungsraum entfernt.

Da die oben angegebenen Testergebnisse andeuten, dass die Rekonstruktion eines Proteins auf der Basis approximativer Overlaps nicht in jedem Fall zu der Berechnung einer optimalen Rekonstruktion führt, muss zur Bewertung des Gesamtalgorithmus und insbesondere zur Bewertung der Leistungsfähigkeit der Scoring-Funktion δ ermittelt werden, wie groß die durchschnittliche und maximale strukturelle Abweichung zwischen einer berechneten und der optimalen Proteinrekonstruktion werden kann. Mit struktureller Abweichung ist hier die Anzahl an Aminosäuren gemeint, um die sich zwei Proteinrekonstruktionen unterscheiden. Die Tests auf Basis der *in silico* verdauten Proteine zeigen, dass die durchschnittliche strukturelle Abweichung zwischen der Primärstruktur einer berechneten Proteinrekonstruktion und der Aminosäuresequenz des zu identifizierenden Proteins für den Fall, dass neben der korrekten Rekonstruktion noch andere suboptimale Rekonstruktionen berechnet werden, bei fünf Prozent liegt. Die maximale strukturelle Abweichung liegt bei zehn Prozent.

Um diese Werte zu ermitteln wurde zunächst einmal die mittlere Sequenzlänge der verwendeten Testproteine berechnet, diese ergab sich aus dem Median der Sequenzlängen der Testproteine. Anschließend wurde nach Durchführung eines Testlaufs für jede der berechneten Rekonstruktionen die strukturelle Abweichung zwischen ihrer Aminosäuresequenz und der Primärstruktur des ursprünglichen Proteins ermittelt. Diese Werte wurden für sämtliche durchgeführten Testläufe bestimmt. Nachdem die gesammelten Werte aufsteigend sortiert und der Median dieser Messwerte bestimmt worden war, ergab sich die durchschnittliche sequentielle Abweichung als prozentualer Anteil des berechneten Medians an der durchschnittlichen Sequenzlänge der Testdaten. Die maximale strukturelle Abweichung ergab sich aus dem prozentualen Anteil der größten ermittelten strukturellen Abweichung an der durchschnittlichen Sequenzlänge der Testdaten.

7.2 Testläufe auf der Basis *in vitro*-verdauter Proteine

Wie bereits zu Beginn von Abschnitt 7.1 angedeutet, ist die Gewinnung von Testdatensätzen für den vorliegenden *de novo*-Proteinidentifikationsalgorithmus für heutige massenspektrometrische Analysemethoden alles andere als eine triviale Aufgabe. Zunächst einmal muss das zu identifizierende Protein mehrfach mit unterschiedlichen Proteasen verdaut werden. Von der hierfür verwendeten Protease darf weder zu viel noch zu wenig aufgetragen werden, da sonst die hieraus entstehenden Peptide entweder nur noch aus einigen wenigen Aminosäuren bestehen oder extrem lang werden. In beiden Fällen besteht das Problem, dass die heutigen Massenspektrometer nicht dazu in der Lage sind solche Peptide identifizieren zu können (siehe Abschnitt 5.3.3). Da genügend der aus den spezifischen Proteolysen entstandenen Peptide identifiziert werden müssen, damit die Aminosäuresequenz des zu identifizierenden Proteins vollständig überdeckt werden kann, müssen einzelne spezifische Proteolysen nicht selten mehrfach wiederholt werden, um eine ausreichende Sequenzabdeckung zu erzielen. Ein weiteres Problem ist die exakte Bestimmung der Masse des zu identifizierenden Proteins. Diese lässt sich mit Hilfe eines Massenspektrometers nur durch ausreichend viele MS/MS-Analysen mit einer hohen Sequenzabdeckung ermitteln. Nur wenn man über mehrere MS/MS-Analysen hinweg, bei einer ausreichend hohen Sequenzabdeckung, die Massen der verdauten Peptide exakt ermitteln kann, lässt sich letzten Endes auch auf die Masse des gesamten Proteins schließen.

Da die Testdatengewinnung auf der Basis *in vitro* verdauter Proteine momentan noch sehr aufwändig ist, ließen sich in der für diese Diplomarbeit veranschlagten Bearbeitungszeit leider insgesamt nur sechs Testdatensätze auf der Basis *in vitro* verdauter Proteine erzeugen. Die hierfür verwendeten Proteine wurden durch die in Tabelle 7.4 angegebenen Proteasen verdaut und anschließend mit Hilfe einer so genannten LCQ DECA XP der Firma Thermo Electron massenspektrometrisch analysiert. Zu der leider sehr geringen Größe der Testdatenmenge kommt noch hinzu, dass es trotz mehrfacher Wiederholung der massenspektrometrischen Analysen für keines der sechs verwendeten Testproteine gelang eine vollständige Sequenzabdeckung zu erzielen. Aufgrund der mehrfach durchgeführten MS/MS-Analysen der Proteine, ließen sich jedoch zwischen vierundsiebzig und achtundachtzig Prozent der Primärstruktur der Proteine überdecken, sodass die hieraus entstehenden proteinogenen Subsequenzen ausreichend lang sind, um als Testobjekte dienen zu können (siehe Tabelle 7.12). Die für die Erzeugung dieser Testdatenmenge

ausgewählten Proteine sind in Tabelle 7.11 angegeben.

Index	Accession	Masse (Da)	Sequenzlänge (AS)	Quelle
1	>gi|226030	23.623,31	209	NCBInr
2	>gi|229351	11.458,33	102	NCBInr
3	>gi|476486	26.018,70	222	NCBInr
4	>gi|1351907	69.293,42	607	NCBInr
5	>gi|1942750	16.979,49	153	NCBInr
6	>gi|4699636	63.273,40	583	NCBInr

Tabelle 7.11: Zusammenstellung der sechs Proteine, auf Basis derer die Erzeugung der eigentlichen Testproteine stattfand. Die angegebenen Datenbank-Accessions entsprechen den Einträgen dieser Proteine in der hier verwendeten Version der NCBInr (Stand vom 15.02.2006). In der Spalte *Sequenzlänge* werden die Längen der Aminosäuresequenzen der Proteine in Aminosäuren (AS) angegeben. Die Abfolge der Tabellenzeilen entspricht der Ordnung der Protein-Accessions.

Index	Masse (Da)	Sequenzlänge (AS)	mit Infixen	ohne Infixe
1	20.611,03	182	51	36
2	9.968,75	89	120	85
3	20.321,01	182	37	32
4	58.829,27	514	121	92
5	14.313,50	129	30	26
6	49.920,41	456	107	89

Tabelle 7.12: Zusammenstellung der sechs Testdatensätze, die auf Basis von *in vitro* verdauten Proteinen erzeugt wurden. Die Einträge in der Spalte *Index* ordnen dem jeweiligen Testdatensatz, dass Protein aus Tabelle 7.11 zu, aus welchem er erzeugt wurde. In der Spalte *Sequenzlänge* werden die Längen der Aminosäuresequenzen der erzeugten Testproteine in Aminosäuren (AS) angegeben. Die Einträge in der Spalte *mit Infixen* geben an, wie viele Peptide, nach dem Filtern von Peptiden mit einer Masse kleiner als 500 oder größer als 7000 Dalton, noch für die Rekonstruktion eines Proteins zu Verfügung stehen. In der Spalte *ohne Infixe* wird dagegen für jedes der Testproteine angegeben, wie viele nichtredundante Peptide in der Eingabe vorhanden sind.

7.2.1 Rekonstruktion mittels nicht-approximativer Overlaps

Die Testergebnisse der für die sechs Testdatensätze durchgeführten Rekonstruktionen auf der Basis nicht-approximativer Overlaps sind mit den in Abschnitt 7.1.1 angegebenen Ergebnissen für die *in silico* verdauten Proteine vergleichbar. Für die sechs Testdatensätze ergeben sich, bei einer minimalen Overlap-Länge von zwei und einer maximalen Massentoleranz von 1,0 Dalton, die in Tabelle 7.13 zusammengefassten Ergebnisse. Jedes der sechs *in vitro* verdauten Proteine ließ sich erfolgreich auf Basis nicht-approximativer Overlaps rekonstruieren. Die durchschnittliche Rechenzeit belief sich auf 5.94 Sekunden.

7.2.2 Rekonstruktion mittels approximativer Overlaps

Um die Eigenschaften der randomisierten Polypeptiderekonstruktion unter Verwendung realer Testdaten objektiv testen zu können, werden für jeden der sechs Testdatensätze einhundert Rekonstruktionen ausgeführt. Nach Abschluss einer solchen Rekonstruktion wird festgestellt, ob das Protein korrekt rekonstruiert wurde und wie lange die einzelnen Rekonstruktionsphasen jeweils gedauert haben. Für die sechs Testdatensätze ergeben sich, bei einer minimalen Overlap-Länge von drei, einer maximalen Massentoleranz von 1,0 Dalton und einer maximalen Levenshtein-Distanz von eins, die in der Tabelle 7.14 angegebenen Ergenisse. Die Rekonstruktion der sechs ausgewählten Testproteine gelang auf der Basis approximativer Overlaps in durchschnittlich 83% der durchgeführten Rekonstruktionen. Die durchschnittliche Rechenzeit belief sich auf $6,89$ Sekunden.

Da die Testergebnisse für die *in vitro* verdauten Testproteine ebenfalls andeuten, dass die Rekonstruktion auf der Basis approximativer Overlaps nicht in jedem Fall zu der Berechnung einer optimalen Rekonstruktion führt, wurde für diese die durchschnittliche und die maximale strukturelle Abweichung zwischen

Index	Overlaps (%)	SCCs (%)	Polypeptide (%)	Layouts (%)	Gesamt (sec.)	Identifiziert
1	88	6	4	2	0,97	ja
2	70	9	12	9	0,92	ja
3	86	6	5	3	0,96	ja
4	74	12	10	4	7,91	ja
5	83	7	6	4	1,13	ja
6	81	11	7	1	10,51	ja

Tabelle 7.13: Zusammenstellung der Testergebnisse für die Rekonstruktion der sechs Testproteine auf der Basis nicht-approximativer Overlaps. Die Werte in den Spalten *Overlaps*, *SCCs*, *Polypeptide* und *Layouts* geben für jede der Rekonstruktionsphasen an, wie groß ihr Anteil an der Gesamtrechendauer ist. Die in der Spalte *Gesamt* angegebenen Werte entsprechen den Rechenzeiten für den gesamten Rekonstruktionsprozess eines Proteins. In der Spalte mit der Bezeichnung *Identifiziert* wird angegeben, ob die Proteinrekonstruktion erfolgreich war.

Index	Overlaps (%)	SCCs (%)	Polypeptide (%)	Layouts (%)	Gesamt (sec.)	Identifiziert
1	88	5	2	5	2,77	79
2	71	8	9	12	1,39	83
3	78	9	4	9	2,53	92
4	82	5	3	10	11,21	78
5	81	5	5	9	3,60	82
6	89	4	1	6	10,84	87

Tabelle 7.14: Zusammenstellung der Testergebnisse für die Rekonstruktion der sechs Testproteine auf der Basis approximativer Overlaps. Die Werte in den Spalten *Overlaps*, *SCCs*, *Polypeptide* und *Layouts* geben für jede der Rekonstruktionsphasen an, wie groß ihr Anteil an der Gesamtrechendauer ist. Die in der Spalte *Gesamt* angegebenen Werte entsprechen den Rechenzeiten für den gesamten Rekonstruktionsprozess der einzelnen Proteine (Mediane der Rechenzeiten der pro Testdatensatz durchgeführten einhundert Rekonstruktionen). In der Spalte mit der Bezeichnung *Identifiziert* wird für jedes der Testproteine die Anzahl der gelungenen Proteinrekonstruktion angegeben.

einer berechneten und der optimalen Proteinrekonstruktion ermittelt. Die Tests auf der Basis der *in vitro* verdauten Proteine zeigen, dass für den Fall, dass neben der korrekten Rekonstruktion noch andere suboptimale Rekonstruktionen berechnet werden, die durchschnittliche strukturelle Abweichung zwischen der Primärstruktur einer berechneten Proteinrekonstruktion und der Aminosäuresequenz des zu identifizierenden Proteins bei sechs Prozent liegt. Die maximale strukturelle Abweichung liegt bei zwölf Prozent.

Analog zu den Werten aus Abschnitt 7.1.2 wurden auch diese Werte ermittelt, indem zunächst einmal die durchschnittliche Sequenzlänge der verwendeten Testproteine bestimmt wurde, diese ergab sich aus dem Median der Sequenzlängen der Testproteine. Anschließend wurde nach Durchführung eines Testlaufs für jede der berechneten Rekonstruktionen die strukturelle Abweichung zwischen ihrer Aminosäuresequenz und der Primärstruktur des ursprünglichen Proteins ermittelt. Diese Werte wurden für sämtliche durchgeführten Testläufe bestimmt. Nachdem die gesammelten Werte aufsteigend sortiert und der Median dieser Messwerte bestimmt worden war, ergab sich die durchschnittliche sequentielle Abweichung als prozentualer Anteil des berechneten Medians an der durchschnittlichen Sequenzlänge der Testdaten. Die maximale strukturelle Abweichung ergab sich aus dem prozentualen Anteil der größten ermittelten strukturellen Abweichung an der durchschnittlichen Sequenzlänge der Testdaten.

7.3 Zusammenfassung der Evaluierung

Fasst man die Ergebnisse sämtlicher Testläufe aus den Abschnitten 7.1 und 7.2 zusammen, so stellt sich heraus, dass in 4346 (84%) der insgesamt 5151 durchgeführten Testläufe eine korrekte Rekonstruktion berechnet wurde. Nur in 805 (16%) aller Testläufe wurden ausschließlich falsche Proteinrekonstruktionen berechnet.

Als ein weiteres Ergebnis der Evaluierung mittels *in silico* und *in vitro* verdauter Proteine stellte sich heraus, dass die in Abschnitt 6.8 hergeleitete Scoring-Funktion δ unterschiedlichen Proteinrekonstruktionen nur dann unterschiedliche Scores zuweisen kann, falls die strukturelle Abweichung zwischen den

Proteinrekonstruktionen und der Primärstruktur des ursprünglichen Proteins größer oder gleich sechs Prozent ist. Da sich das Auflösungsvermögen der Scoring-Funktion δ aufgrund der aktuell zu Verfügung stehenden Datengrundlage (siehe Abschnitt 5.2) und des heutigen Erkenntnistands auf dem Gebiet der Massenspektrometrie leider nicht verbessern lässt, wird es in vielen Fällen nach Durchführung einer Proteinrekonstruktion leider nicht möglich sein unter sämtlichen berechneten Rekonstruktionen die eigentlich gesuchte Rekonstruktion hervorzuheben. Betrachtet man jedoch die oben erwähnten 4346 Testläufe, in denen eine korrekte Proteinrekonstruktion berechnet wurde genauer und untersucht, wie viele der berechneten inkorrekten Proteinrekonstruktionen einen von der korrekten Rekonstruktion verschiedenen δ-Score besitzen, so stellt man fest, dass dies auf etwa 78% sämtlicher suboptimalen Rekonstruktionen zutrifft. Die verbleibenden 22% müssen daher nach Abschluss des Rekonstruktionsprozesses zusammen mit der korrekten Lösung in einem gemäß δ-Score sortiertem Ranking ausgegeben werden.

Stellt man die Anzahl der geglückten Proteinrekonstruktionen auf der Basis nicht-approximativer Overlaps der Anzahl der durchschnittlich geglückten Rekonstruktionen auf der Basis approximativer Overlaps gegenüber, so ergibt sich aus den Testresultaten der durchgeführten Evaluierung eine klare Präferenz bzgl. Proteinidentifikationen mittels nicht-approximativer Peptid-Overlaps. Die Berechnung einer Proteinrekonstruktion unter Verwendung approximativer Peptid-Overlaps sollte nur dann erfolgen, falls umfangreiche massenspektrometrische Analysen eines Proteins mangels einer ausreichenden Menge an Probematerial nicht durchführbar oder zu arbeitsintensiv sind. Ansonsten empfiehlt es sich die Güte der Eingabedaten des hier vorgestellten Rekonstruktionsalgorithmus durch die vorhergehenden Analyseschritte (siehe Abschnitt 5.1) auf einem möglichst hohen Qualitätsniveau zu halten und die Rekonstruktion des zu identifizierenden Proteins mittels nicht-approximativer Peptid-Overlaps durchzuführen. Wie die Abbildungen 7.1, 7.2, 7.3 und 7.4 andeuten hängt der Erfolg einer Proteinrekonstruktion maßgeblich von den für die Rekonstruktionsparameter gewählten Werten ab. Unabhängig davon, ob eine Proteinrekonstruktion mittels approximativer oder nicht-approximativer Peptid-Overlaps durchgeführt wird, sollte die minimale Peptid-Overlap-Länge *mol* weder zu klein noch zu groß gewählt werden, da sonst in beiden Fällen Peptide, die für die Rekonstruktion essentiell wichtig sind, bei der Rekonstruktion außer Acht gelassen werden können. Im Hinblick auf Proteinrekonstruktionen auf der Basis von approximativen Peptid-Overlaps ist zudem zu beachten, dass eine Vielzahl suboptimaler Rekonstruktionspfade auf dem Overlap-Graphen von vornherein ausgeschlossen werden können, falls die minimale Peptid-Overlap-Länge *mol* angemessen hoch gewählt wird. Durch die Bestimmung und Berücksichtigung sämtlicher approximativer Matchings zwischen den Aminosäuresequenzen der identifizierten Peptide wächst die Anzahl der durch den Overlap-Graphen repräsentierten Überlappungsbeziehungen zum Teil dramatisch an. Eine Vielzahl der hierdurch zusätzlich hinzukommenden Überlappungsbeziehungen ist aber auf zufällige Übereinstimmungen zwischen den Aminosäuresequenzen der Peptide aus der Eingabe zurückzuführen, die erst durch die Betrachtung potentieller Einfügungen, Löschungen oder Aminosäurenvertauschungen aufgedeckt werden. Diese zufälligen Übereinstimmungen sind in der Regel sehr kurz (in gut drei-viertel sämtlicher während der Evaluierung beobachteter Fälle nur eine oder zwei Aminosäuren lang) und lassen sich durch die Wahl eines geeigneten Werts für *mol* aus dem Overlap-Graphen entfernen.

Kapitel 8

Zusammenfassung und Ausblick

8.1 Zusammenfassung

Im Rahmen der vorliegenden Diplomarbeit wurde ein Algorithmus für die so genannte *de novo*-Proteinidentifikation entwickelt. Aufgrund der in Kapitel Fünf und Sechs erfolgten Konzeption und Realisierung ist dieser Algorithmus dazu in der Lage, die theoretischen und praktischen Limitationen der automatischen Hochdurchsatz-Proteinidentifikation auf der Basis von Proteindatenbanken, wie sie in Kapitel Vier vorgestellt wurden, zu überwinden. Darüber hinaus realisiert er eine Lösung für einige weitaus grundlegendere Problemstellungen der Proteinidentifikation, zu diesen gehören die Erkennung und korrekte Behandlung von Transpeptidierungseffekten, von Mehrfachidentifikationen strukturell identischer Peptide, von Probenkontaminationen, sowie die Durchführung von Proteinidentifikationen auf der Basis fehlerhaft identifizierter Peptide. Diese grundlegenden Problemstellungen wurden im Rahmen der Anforderungsdefinition und -Analyse in Kapitel Fünf definiert und erläutert. Anders als die meisten der derzeitig standardmäßig eingesetzten Proteinidentifikationsalgorithmen ist der hier entwickelte Algorithmus für die Identifikation der Primärstruktur eines unbekannten Proteins nicht auf die Existenz einer zu diesem Protein genetisch kompatiblen Proteindatenbank angewiesen, sondern dazu in der Lage die Aminosäuresequenz eines zu identifizierenden Proteins auf Grund von experimentell ermittelten Daten aus der Massenspektrometrie zu bestimmen. Hierdurch stellt er eine interessante Alternative bzw. Ergänzung zu den derzeitig in der Proteinanalytik zu Verfügung stehenden Analysemethoden dar.

8.2 Ausblick

Während der Konzeption, Entwicklung und Evaluierung des vorliegenden Algorithmus ergaben sich weitergehende Fragestellungen aus den Bereichen der Bioinformatik und Proteinanalytik, deren Beantwortung weiterer Forschung bedarf:

1. **Durchführung der Proteinidentifikation gemäß *de novo*-Ansatz ohne vollständige Sequenzüberdeckung des zu identifizierenden Proteins durch Peptide mit bekannter Primärstruktur**. Aus Sicht der heutigen, hauptsächlich auf der Verwendung von Datenbanken basierenden Proteinanalytik, stellt sich die funktionale Anforderung des *de novo*-Ansatzes, Proteinidentifikationen ausschließlich auf der Basis einer vollständigen Sequenzabdeckung durch identifizierte Peptide durchführen zu können, als eine gravierende Einschränkung dar. Für die datenbankgestützte Identifikation eines Proteins genügen im Prinzip schon einige wenige identifizierte Peptide, wodurch im direkten Vergleich zur *de novo*-Methode insgesamt sehr viel weniger Massenspektren erzeugt werden müssen. Da der massenspektrometrischen Analyse eines zu identifizierenden Proteins aber in der Regel noch einige andere Analyseschritte vorausgehen, wie z.B. die Probengewinnung und -aufbereitung (siehe Anschnitt 3.1.1), die Proteinseparation (siehe Abschnitt 3.1.2) und die spezifische Proteolyse (siehe Abschnitt 3.1.3), sinkt mit der Anzahl der zu erzeugenden Massenspektren auch der für die Proteinidentifikation insgesamt zu betreibende Analyseaufwand. Obwohl die Qualität und Zuverlässigkeit einer Proteinidentifikation auf der Basis einiger weniger korrekt iden-

tifizierter Peptide im Hinblick auf die heutzutage verwendeten Proteindatenbanken schnell dazu führt, dass die Wahrscheinlichkeit für eine falsche positive Identifikation dramatisch wächst, ist die Veröffentlichung von Forschungsergebnissen, die auf einem solchen Vorgehen basieren derzeit noch Gang und Gebe.

Um nun die Anwendung des *de novo*-Ansatzes aus Sicht der Proteinanalytik attraktiver und weniger arbeitsintensiv zu gestalten, wäre es von großer Bedeutung, *de novo*-Proteinidentifikationen auf der Basis unvollständiger Sequenzabdeckungen durchführen zu können. Hierfür müssten Lücken in der Sequenzabdeckung des zu identifizierenden Proteins mit Hilfe von hypothetischen Peptiden geschlossen werden. Um die Primärstruktur solcher hypothetischen Peptide postulieren zu können, müssten umfangreiche auf einzelne Organismen bezogene Untersuchungen bzgl. der Aminosäureverteilungen bereits identifizierter Proteine durchgeführt werden. Aufgrund der hierbei gewonnen Erkenntnisse, wäre es mit Hilfe von bedingten Wahrscheinlichkeiten und in Abhängigkeit von der genetischen Zugehörigkeit des zu identifizierenden Proteins möglich hypothetische Peptide zu erzeugen, deren Primärstruktur die Lücken in der Sequenzabdeckung des zu identifizierenden Proteins schließen.

2. **Datenakquisition aufgrund eines enzymatischen Proteinverdaus mit lediglich einer spezifischen Protease**. Da die Proteinidentifikation auf der Basis des *de novo*-Ansatzes im Wesentlichen auf der Erzeugung eines peptidischen Überlappungsmusters beruht (siehe Abschnitt 5.1), muss das zu identifizierende Protein mit mehreren Enzymen unterschiedlicher Enzymspezifizität verdaut werden (siehe Abschnitt 3.1.3). Eine weitere Möglichkeit den Einsatz von *de novo*-Proteinidentifikationsalgorithmen weniger aufwändig zu gestalten, liegt in der Durchführung der spezifischen Proteolyse unter Verwendung lediglich einer spezifischen Protease. Verwendet man für den enzymatischen Verdau z.B. lediglich Trypsin und verdaut mit diesem Enzym mehrere Proben eines Proteins, wobei man unterschiedliche Mengen Trypsin unterschiedlich lange auf die jeweilige Proteinprobe einwirken lässt, so erhält man ein Überlappungsmuster, das für die Rekonstruktion eines Proteins ebenfalls geeignet scheint. Um die *de novo*-Proteinidentifikation zukünftig mit Hilfe lediglich einer Protease durchführen zu können, müssten unfangreiche Untersuchungen bzgl. der Entstehung und der Struktur solcher monoenzymatischen Überlappungsmuster durchgeführt werden.

3. **Berücksichtigung der Isoformen eines Proteins**. Unter der Isoform eines Proteins versteht man in der Molekularbiologie eine Variation eines Proteins mit leichten bis größeren strukturellen Unterschieden. Diese Unterschiede sind oftmals auf alternatives Spleißen oder co- und posttranslationelle Modifikationen (z.B. das Anhängen von speziellen Zuckermolekülen, was als Glykosierung bezeichnet wird) zurückzuführen. Die Entdeckung proteinogener Isoformen beim Menschen scheint eine weitere Konsequenz der relativ geringen Anzahl an unterschiedlichen Genen zu sein, welche im Human Genome Project gefunden wurden. Ein Organismus besitzt durch diesen Mechanismus die Möglichkeit trotz einer relativ geringen Anzahl an unterschiedlichen Genen eine Vielzahl katalytisch unterschiedlicher Proteine herzustellen. Hierdurch erweitert sich die Diversität eines Genoms beträchtlich. Im Hinblick auf die Weiterentwicklung des hier beschriebenen *de novo*-Proteinidentifikationsalgorithmus, müsste bei der Rekonstruktion eines zu identifizierenden Proteins auch das Auftreten von Isoformen dieses Proteins berücksichtigt und behandelt werden. Um dies zu ermöglichen müssten biochemische Studien angefertigt werden, in denen untersucht wird bei welchen Gattungen bzw. Spezies proteinogene Isoformen zu beobachten sind, welche Struktur diese im Einzelnen besitzen, wie groß die strukturellen Unterschiede zwischen einem Protein und einer seiner Isoformen maximal werden kann und wie häufig bestimmte Isoformen eines Proteins statistisch gesehen auftreten.

4. **Verbesserung des Auflösungsvermögens der Scoring-Funktion** δ. Wie bereits in Kapitel Sieben diskutiert (siehe Abschnitt 7.3), ist die Scoring-Funktion δ leider nicht in jedem Fall dazu in der Lage eine berechnete optimale Proteinrekonstruktion von strukturell ähnlichen, suboptimalen Rekonstruktionen zu unterscheiden. Daher sollte in Zusammenarbeit mit Molekularbiologen, Chemikern und Statistikern eine adäquatere statistische Methode zur Bestimmung der tatsächlichen Peptidstartpunktverteilung D_{src} entwickelt werden, mit deren Hilfe selbst geringe strukturelle Abweichungen zwischen einer berechneten Proteinrekonstruktion und der Primärstruktur eines zu untersuchenden Proteins festgestellt werden können.

5. **Automatisches bzw. evolutionäres Erlernen von Parametersätzen für die Proteinrekonstruktion gemäß des *de novo*-Ansatzes**. Wie in den Abbildungen 7.1, 7.2, 7.3 und 7.4 zusammenfassend dargestellt wird, hängt der Erfolg einer Proteinidentifikation gemäß des *de novo*-Ansatz maßgeblich von der Wahl geeigneter Werte für die minimale Overlap-Länge *mol*, die

maximale Levenshtein-Distanz dis und die maximale Massentoleranz m_{diff} ab. Da die Wahl eines Wertes für einen dieser drei Parameter aber auch Auswirkungen auf die Werte der beiden anderen Parameter haben kann, wäre für den alltäglichen Einsatz des hier entwickelten Algorithmus ein adaptives und lernfähiges System zur Wahl geeigneter Rekonstruktionsparameter wünschenswert. Um ein solches System umsetzen zu können, wären aber noch weitaus umfangreichere Rekonstruktionstests auf der Basis bereits identifizierter Proteine notwendig.

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

2D-PAGE	2-dimensionale Polyacrylamid-Gelelektrophorese
DDBJ	DNA Data Bank of Japan
DNS	Desoxyribonukleinsäure
EMBL	European Molecular Biology Laboratory
ESI	Elektrospray-Ionisation
FFT	Fast Fourier Transformation
HGP	Human Genome Project
HPLC	High Performance Liquid Chromatography
HRPD	Human Protein Reference Database
IEF	isoelektrische Fokussierung
IPI	International Protein Index
MALDI	Matrix-assisted-Laser-Desorption-Ionisation
MPC	Medizinisches Proteom-Center
mRNS	Messenger-Ribonukleinsäure
MS	Massenspektrometrie
MS/MS	Tandem-Massenspektrometrie
NCBI	National Center of Biotechnology Information
PDB	Protein Data Bank
PFF	Peptide Fragmentation Fingerprint
PIR	Protein Information Recource
PMF	Peptide Mass Fingerprint
PRF	Protein Research Foundation
PSD	Post Source Decay
PTM	post-translationale Modifikation
RIC	Reconstructed Ion Current
RNS	Ribonukleinsäure
TOF	Time of Flight
XML	Extensible Markup Language

Literaturverzeichnis

[1] International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. *Nature*, 409, 860-921.

[2] Strohman, R.C. 1997. The coming Kuhnian revolution in biology. *Nature Biotechnology*, 15, 194-200.

[3] Jasny, B. R. & Roberts, L. 2003. Building on the DNA Revolution. *Science*, 11, 277.

[4] Collins, F.S., Morgan, M. & Patrinos, A. 2003. The Human Genome Project: Lessons from Large-Scale Biology. *Science*, 11, 286.

[5] Frazier, M.E., Johnson, G.M., Thomassen, D.G., Oliver, C.E., Patrinos, A. 2003. Realizing the Potential of the Genome Revolution: The Genomes to Life Program. *Science*, 11, 290.

[6] Collins, F.S., Green, E.D., Guttmacher, A.E. & Guyer, M.S. 2003. A Vision for the Future of Genomics Research. *Nature*, 24, 835.

[7] Carroll, S.B. 2003. Genetics and the Making of Homo sapiens. *Nature*, 24, 849.

[8] Arnold, J. & Hilton, N. 2003. Genome Sequencing: Revelations from a Bread Mould. *Nature*, 24, 821.

[9] Hillier, L.W. et al. 2005. Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature*, 434, 724-731.

[10] Yates, J.R., Speicher, S., Griffin, P.R. & Hunkapiller, T. 1993. Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.*, 214, 397-408.

[11] Yates, J.R., Eng, J.K. & McCormack, A.L. 1995. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.*, 67, 3202-3210.

[12] Clauser, K.R., Baker, P. & Burlingame, A.L. 1999. Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.*, 71, 2871-2882.

[13] Pappin, D.J.C., Hojrup, P. & Bleasby, A.J. 1993. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.*, 3(6), 327-32.

[14] Pappin, D.J.C., Rahman, D., Hansen, H.F., Bartlet-Jones, M., Jeffery, W. & Bleasby, A.J. 1996. Chemistry, mass spectrometry and peptide-mass databases: Evolution of methods for the rapid identification and mapping of cellular proteins. *Mass. Spectrom. Biol. Sci.*, Humana Press, 135-150.

[15] Zhang, W. & Chait, B.T. 2000. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.*, 72, 2482-2489.

[16] Colinge, J., Masselot, A., Giron, M., Dessingy, T. & Magnin, J. 2003. OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics*, 3(8), 1454-1463.

[17] Colinge, J., Magnin, J. & Masselot, A. 2003. A systematic statistical analysis of ion trap tandem mass spectra in view of peptide scoring. *Proceeding of the Workshop on Algorithms in Bioinformatics (WABI)*, Page, R. & Benson, G., LNBI 2812, Springer, 25-38.

[18] Colinge, J., Chiappe, D., Lagache, S., Moniatte, M. & Bougueleret, L. 2005. Differential Proteomics via probabilistic peptide identification scores. *Anal. Chem.*, 77(2), 596-606.

[19] Reidegeld, K.A., Meyer, H.E. & Warscheid, B. 2004. In Silico Protein Digestion considering Post-Translational Modifications. *Poster* German Conference on Bioinformatics.

[20] Reidegeld, K.A. 2005. Peakardt.FindPairs - software for automatic quantitative evaluation of stable isotope-coded peptide mass spectra. *Poster* ASMS.

[21] Reidegeld, K.A., Linsenmann, G., Hebeler, R., Wiese, S., Oeljeklaus, S., Lakhal, B. & Meyer, H.E. 2005. Peakardt.FindPairs - A Univeral Software for Protein Quantitation via Stable Isotope-Labeling through Mass Spectrometry. *Poster*HUPO World Congress.

[22] Merkl, R. & Waack, S. 2003. Bioinformatik Interaktiv: Algorithmen und Praxis. *Wiley-VCH.*

[23] Lesk, A.M. 2002. Bioinformatik. Eine Einführung. *Spektrum Akademischer Verlag.*

[24] Cynthia, G. & Jambeck, P. 2001. Einführung in die Praktische Bioinformatik. Grundlagen, Anwendungen, Techniken und Tools. *O'Reilly.*

[25] Schürrle, K. 2003. Proteomforschung, die Werkzeuge des Lebens nutzen. *Technical report*, Bundesministerium für Bildung und Forschung (BMBF).

[26] Stein, L.D. 2004. Human genome: End of the beginning. *Nature*, 431, 915 – 916.

[27] Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. & Whitehouse, C.M. 1989. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246, 64-71.

[28] Kellner, R. 2000. Proteomics. Concepts and perspectives. *Anal. Chem.*, 366, 517-524.

[29] Wilkins, M.R., Pasquali, C., Appel, R.D., Ou, K., Golaz, O., Sanchez, J.C., Yan, J.X., Gooley, A.A., Hughes, G., Humphery-Smith, I., Williams, K.L. & Hochstrasser, D.F. 1996. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N.Y.)*, 14, 61-65.

[30] Chamrad, D. 2004. Bioinformatische Verfahren zur Analyse von Primärstrukturinformation mittels massenspektrometrischer Daten in der Proteomanalyse. *Dissertation*, Ruhr-Universität Bochum.

[31] Klose, J. 1975. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik*, 26, 231-243.

[32] O'Farrell, P.H. 1975. High resolution two-dimensional electrophoresis of proteins. *Biol. Chem.*, 250, 4007-4021.

[33] Lawrence, J.F. & Frei, R.W. 1976. Chemical derivatization in liquid chromatography. *New York: Elsevier Scientific Pub. Co.*

[34] Huber, J.F.K. 1978. Instrumentation for High-Performance Liquid Chromatography, *Journal of Chromatography*, 13, 115-226.

[35] Schaefer, H., Marcus, K., Sickmann, A., Herrmann, M., Klose, J. & Meyer, H.E. 2003. Identification of phosphorylation and acetylation sites in alphaA-crystallin of the eye lens (mus musculus) after two-dimensional gel electrophoresis. *Anal. Bioanal. Chem.*, 376, 966-972.

[36] Barber, M., Bordoli, R.S., Sedgwick, R.D. & Tyler, A.N. 1981. Fast atom bombardment of solids as an ion source in mass spectroscopy. *Nature*, 293, 270-275.

[37] Liu, L.K., Busch, K.L. & Cooks, R.G. 1981. Matrix-assisted secondary ion mass spectra of biological compounds. *Analytical Chemistry*, 53, 109.

[38] Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y. & Yoshida, T. 1988. Protein and polymer analysis up to m/z 100,000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.*, 2, 151.

[39] Karas, M. & Hillenkamp, F. 1988. Laser desorption ionization of proteins with molecular mass exceeding 10,000 Daltons. *Analytical Chemistry*, 60, 2299-2301.

[40] Dulcks, T. & Juraschek, R. 1999. Electrospray as an ionization method for mass spectrometry. *Aerosol Sci.*, 30, 927-943.

[41] Duft, D., Achtzehn, T., Müller, R., Huber, B.A. & Leisner, T. 2003. Rayleigh jets from levitated microdroplets, *Nature*, 421, 128.

[42] Brutschy, B. & Karas, M. 2004. Der mikroskopische Blick auf die Moleküle des Lebens. Massenspektrometrie: Wäge- und Analysetechnik in einem. *Forschung Frankfurt*, Johann Wolfgang Goethe Universität Frankfurt am Main.

[43] Schürch, S. 2004. Massenspektrometrie. Gestern - Heute - Morgen. *Presentation*, Lehrstuhl für Chemie und Biochemie, Universität Bern.

[44] Karas, M. & Brutschy, B. 2004. Der mikroskopische Blick auf die Moleküle des Lebens. *Forschung Frankfurt*, 1, 12-15.

[45] Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C. & Watanabe, C. 1993. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proceedings Natl. Acad. Sci. USA*, 90, 5011-5015.

[46] James, P., Quadroni, M., Carafoli, E. & Gonnet, G. 1993. Protein identification by mass profile fingerprinting. *Biochem Biophys. Res. Commun.*, 195, 58-64.

[47] Mann, M., Hojrup, P. & Roepstorff, P. 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass. Spectrom.*, 22, 338-345.

[48] Jensen, O.N., Podtelejnikov, A.V. & Mann, M. 1997. Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. *Anal. Chem.*, 69, 4741-4750.

[49] Spengler, B., Kirsch, D., Kaufmann, R. & Jaeger, E. 1992. Peptide sequencing by matrix-assisted laser-desorption mass spectrometry. *Rapid. Commun. Mass. Spectrom.*, 6, 105-108.

[50] Hunt, D.F., Buko, A.M., Ballard, J.M., Shabanowitz, J. & Giordani, A.B. 1981. Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer. *Biomed. Mass. Spectrom.*, 8, 397-408.

[51] Hunt, D.F., Yates, J.R., Shabanowitz, J., Winston, S. & Hauer, C.R. 1986. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. USA*, 83, 6233-6237.

[52] Down, T.A. & Hubbard, T.J. 2002. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, 12, 458-461.

[53] Altschul, S.F. & Gish, W. 1996. Local alignment statistics. *Methods Enzymol*, 266, 460-480.

[54] Chamrad, D.C., Koerting, G., Gobom, J., Thiele, H., Klose, J., Meyer, H.E. & Blueggel, M. 2003. Interpretation of mass spectrometry data for high-throughput proteomics. *Anal. Bioanal. Chem.*, 376, 1014-1022.

[55] Wilke, A., Ruckert, C., Bartels, D., Dondrup, M., Goesmann, A., Huser, A.T., Kespohl, S., Linke, B., Mahne, M., McHardy, A., Puhler, A. & Meyer, F. 2003. Bioinformatics support for high-throughput proteomics. *Biotechnol.*, 106, 147-156.

[56] Goh, C.S., Lan, N., Echols, N., Douglas, S.M., Milburn, D., Bertone, P., Xiao, R., Ma, L.C., Zheng, D., Wunderlich, Z., Acton, T., Montelione, G.T. & Gerstein, M. 2003. SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res.*, 31, 2833-2838.

[57] Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. & Yeh, L.S. 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, 32, 115-119.

[58] Barker, W.C., Garavelli, J.S., McGarvey, P.B., Marzec, C.R., Orcutt, B.C., Srinivasarao, G.Y., Yeh, L.S., Ledley, R.S., Mewes, H.W., Pfeiffer, F., Tsugita, A. & Wu, C. 1999. The PIR-International Protein Sequence Database. *Nucleic Acids Res.*, 27, 39-43.

[59] O'Donovan, C., Martin, M.J., Gattiker, A., Gasteiger, E., Bairoch, A. & Apweiler, R. 2002. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.*, 3, 275-284.

[60] Peri, S., Navarro, J.D., Kristiansen, T.Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T.K., Chandrika, K.N., Deshpande, N., Suresh, S., Rashmi, B.P., Shanker, K., Padma, N., Niranjan, V., Harsha, H.C., Talreja, N., Vrushabendra, B.M., Ramya, M.A., Yatish, A.J., Joy, M., Shivashankar, H.N., Kavitha, M.P., Menezes, M., Choudhury, D.R., Ghosh, N., Saravana, R., Chandran, S., Mohan, S., Jonnalagadda, C.K., Prasad, C.K., Kumar-Sinha, C., Deshpande, K.S. & Pandey, A. 2004. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.*, 32, 497-501.

[61] Bleasby, A.J. & Wootton, J.C. 1990. Construction of validated, non-redundant composite protein sequence databases. *Protein Eng.*, 3, 153-159.

[62] Hoogland, C., Sanchez, J.C., Tonella, L., Binz, P.A., Bairoch, A., Hochstrasser, D.F. & Appel, R.D. 2000. The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res.*, 28, 286-288.

[63] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.*, 28, 235-242.

[64] Orchard, S., Zhu, W., Julian, R.K. Jr., Hermjakob, H. & Apweiler, R. 2003. Further advances in the development of a data interchange standard for proteomics data. *Proteomics*, 3, 2065-2066.

[65] Vaysseix, G. & Barillot, E. 2001. XML, bioinformatics and data integration. *Bioinformatics*, 17, 115-125.

[66] Gras, R., Müller, M., Gasteiger, E., Gay, S., Binz, P.A., Bienvenut, W., Hoogland, C., Sanchez, J.C., Bairoch, A., Hochstrasser, D.F. & Appel, R.D. 1999. Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis*, 20, 3535-3550.

[67] Breen, E.J., Hopwood, F.G., Williams, K.L. & Wilkins, M.R. 2000. Automatic poisson peak harvesting for high throughput protein identification. *Electrophoresis*, 21, 2243-2251.

[68] Coombes, K.R., Fritsche, H.A. Jr., Clarke, C., Chen, J.N., Baggerly, K.A., Morris, J.S., Xiao, L.C., Hung, M.C. & Kuerer, H.M. 2003. Quality Control and Peak Finding for Proteomics Data Collected from Nipple Aspirate Fluid by Surface-Enhanced Laser Desorption and Ionization. *Clin. Chem*, 49, 1615-1623.

[69] Zhang, Z. & Marshall, A.G. 1998. A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *Am. Soc. Mass. Spectrom.*, 9, 225-233.

[70] Zheng, H., Ojha, P.C., McClean, S., Black, N.D., Hughes, J.G. & Shaw, C. 2003. Heuristic charge assignment for deconvolution of electrospray ionization mass spectra. *Rapid Commun. Mass. Spectrom.*, 17, 429-436.

[71] Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20, 3551-3567.

[72] Eng, J.K., McCormack, A.L. & Yates, J.R. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Am. Soc. Mass. Spec.*, 5, 976-989.

[73] Krishna, R.G. & Wold, F. 1993. Post-translational modification of proteins. *Adv. Enzymol. Relat. Areas Mol. Biol.*, 67, 265-298.

[74] Gattiker, A., Bienvenut, W.V., Bairoch, A. & Gasteiger, E. 2002. FindPept, a tool to identify unmatched masses in peptide mass fingerprinting protein identification. *Proteomics*, 2, 1435-1444.

[75] Pevzner, P.A., Dancik, V. & Tang, C.L. 2000. Mutation-tolerant proteinidentification by mass spectrometry. *Comput. Biol.*, 7, 777-787.

[76] Pevzner, P.A., Mulyukov, Z., Dancik, V. & Tang, C.L. 2001. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res.*, 11, 290-299.

[77] Creasy, D.M. & Cottrell, J.S. 2002. Error tolerant searching of interpreted tandem mass spectrometry data. *Proteomics*, 2, 1426-1434.

[78] Gay, S., Binz, P.A., Hochstrasser, D.F. & Appel, R.D. 1999. Modeling peptide mass fingerprinting data using the atomic composition of peptides. *Electrophoresis*, 20, 3527-3534.

[79] Kapp, E.A., Schutz, F., Reid, G.E., Eddes, J.S., Moritz, R.L., O'Hair, R.A., Speed, T.P. & Simpson, R.J. 2003. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem.*, 75, 6251-6264.

[80] Schutz, F., Kapp, E.A., Simpson, R.J. & Speed, T.P. 2003. Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochem. Soc. Trans.*, 31, 1479-1483.

[81] Elias, J.E., Gibbons, F.D., King, O.D., Roth, F.P. & Gygi, S.P. 2004. Intensitybased protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.*, 22, 214-219.

[82] van't Hoff, J.H. 1898. Studien zur chemischen Dynamik. *Anorg. Chem.*, 18, 1-13.

[83] Bergmann, M., Zervas, L. & Fruton, J.S. 1935. On Proteolytic Enzymes. VI. On the Specificity of Papain. *Biol. Chem.*, 111, 225-244.

[84] Bergmann, M. & Fruton, J.S. 1937. The Role of Specificity in the enzymatic synthesis of Proteins. Syntheses with intercellular Enzymes. *Biol. Chem.*, 118, 707-720.

[85] Kullmann, W.J. 1982. Protease-catalyzed peptide bond formation: application to synthesis of the COOH-terminal octapeptide of cholecystokinin. *Proc. Natl. Acad. Sci. USA*, 79, 2840-2844.

[86] Kullmann, W.J. 1984. Kinetics of chymotrypsin- and papain-catalysed synthesis of [leucine]enkephalin and [methionine]enkephalin. *Biochem.*, 220(2), 405–416.

[87] Takai, H., Sakato, K., Nakamizo, K. & Isowa, Y. 1981. Protease-catalyzed synthesis of oligopeptides in heterogenous substrate mixtures. *Peptide Chemistry*, Protein Research Foundation, Osaka, 213-214.

[88] Markussen, J. & Volund, A. 1985. Kinetics of trypsin catalysis in the industrial conversion of porcine insulin to human insulin. *Ciba Found. Symp.*, 111, 188-203.

[89] Rose, K., Gladstone, J. & Offord, R.E. 1984. A mass-spectrometric investigation of the mechanism of the semisynthetic transformation of pig insulin into an ester of insulin of human sequence. *Biochem.*, 220, 189-196.

[90] Canova-Davis, E., Kessler, T.J. & Ling,V.T. 1991. Transpeptidation during the analytical proteolysis of proteins. *Anal. Biochem.*, 196, 39-45.

[91] Goepfert, A., Lorenzen, P.C. & Schlimme, E. 1999. Peptide synthesis during in vitro proteolysis–transpeptidation or condensation? *Nahrung*, 43, 211-212.

[92] Lorenzen, P., Goepfert, A., Schieber, A. & Bruckner, H. 1997. Evidence for peptide synthesis in the course of in vitro proteolysis. *Nahrung*, 41, 87-90.

[93] Schaefer, H., Chamrad, D.C., Marcus, K., Reidegeld, K.A., Bluggel, M. & Meyer, H.E. 2005. Tryptic transpeptidation products observed in proteome analysis by liquid chromatography-tandem mass spectrometry. *Proteomics*, 5(4), 846-52.

[94] Myers, E.W. 1995. Toward simplifying and accurately formulating fragment assembly. *Comp. Biol.*, 2(2), 275-290.

[95] Chakravarti, I.M., Laha, R.G. & Roy, J. 1967. Handbook of Methods of Applied Statistics, Volume I, *John Wiley and Sons*, 392-394.

[96] Wu, S. & Manber, U. 1992. Fast text searching allowing errors. *Comm. ACM.*, 35, 83-91.

[97] Wu, S. & Manber, U. 1992. Agrep - a fast approximative pattern-matching tool. *Usenix Technical Conference*, 153-162.

[98] Tarjan, R. 1972. Depth first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2), 146-160.

Erklärung

Die vorliegende Diplomarbeit entstand im dem Zeitraum von November 2005 bis Mai 2006 auf Grund einer Kooperation zwischen dem Lehrstuhl 11 für Algorithm Engineering des Fachbereichs Informatik an der Universität Dortmund und des Medizinischen Proteom-Centers an der Ruhr-Universität Bochum.

Hiermit versichere ich, die vorliegende Arbeit selbstständig und unter ausschließlicher Verwendung der angegebenen Literatur und Hilfsmittel erstellt zu haben. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Dortmund, den 03.05.2006

(Unterschrift)