Masterarbeit

Bioinformatik

Identifikation und Lokalisation von cross-link Positionen im MS/MS Spektrum

Sophie Kolbe
Mat.-Nr.:4299134
Mail: kolbe.sophie@gmail.com

Betreuer:
Prof. Dr. Knut Reinert
Freie Universtiät Berlin,
Algorithmische Bioinformatik

Dr. Olvia Gräbner
Caprotec Bioanalytics GmbH

Berlin, der 28. April 2014

Abstract

The purpose of this thesis is to identify and localise cross link positions in peptids, including the development of a tool for analysis and validation of according MS/MS spectras.

Through this highly complex topic the focus will be to interpretate retrospectivly already accepted cross link data. Aiming for preferably detailled and differentiated investigation for MS/MS spectras, it shall be possible to make qualitative statments later on. An explanation about cross links in general and the resulting variable modifications is provided by the Introduction. In the second chapter Identification follows an analysis of the variable masses as well as a protoypic implementation for peptide identifikation. Checking significance of the results via multiple different spectras, the estimated error tolerance is located within the expectation and the intermediate data displayed in the appendix A. After the identification affiliates a scoring function which presents an assessed results list of matching peptids and will be described in chapter 3 verification. In a final step follows the localisation of the binding site in chapter 4. Here the contribution of the matches as well as their quality is evaluated and plotted over the whole peptide sequence. Incisive changes in the kurve will indicate positions of high binding affinity. Considering the dominant ion series within the data and their relative abundance according to the applied MS/MS methods computes the qualitative probability for a cross link for each position in the peptid.

A standard peptide provides the required data for the implementation of the protoyp , reducing the complexity of the problem to a miminum using known properties. Later on in chapter 5 results will be another analysis with real data of a second cross link peptide. These results will confirm the steps until now and yield sophisticated conclusions. Finally in chapter 6 conclusion and forecast the future enhancement will be discussed and possible applications of the tool within other frameworks advised.

Zusammenfassung

Die Intention der vorliegenden Arbeit ist die Identifikation und Lokalisation von Cross-Link Bindungsstellen in Peptiden sowie die Entwicklung eines Tools zur Analyse und Validierung der dazugehörigen MS/MS Spektren.

In diesem komplexen Themengebiet liegt der Fokus auf der rückwirkenden Interpretation der gemessenen Cross-Link-Daten. Ebenso besteht die Absicht, eine vorzugsweise detaillierte und differenzierte Untersuchung der MS/MS Spektren zu erlauben und qualitative Aussagen zu ermöglichen. Eine Erläuterung über Cross-Links und der daraus erzeugten variablen Modifikationen wird in der Einleitung gegeben. Im Kapitel 2 Identifizierung folgte eine Analyse der variablen Massen und eine prototypische Implementierung zur Identifikation des Peptids. Die Signifikanz der Resultate wird mit unterschiedlichen Spektren überprüft und die geschätzte Fehlertoleranz liegt innerhalb der Erwartungen, die Zwischenergebnisse finden sich im Anhang A. Nach der Identifizierung folgt eine angeschlossene Scoring-Funktion, die eine gewertete Ergebnisliste von passenden Peptiden präsentiert und in Kapitel 3 Verifizierung beschrieben wird. Zuletzt folgt die Lokalisierung der Bindungsstelle in Kapitel 4; hier wird die Verteilung der Matches sowie deren Qualität ausgewertet und über die gesamte Peptidsequenz hinweg abgebildet. Anhand von prägnanten Wechsel der Kurve lassen sich die möglichen Bindungsstellen ablesen. Mit Berücksichtigung der dominanten Ionenserien innerhalb den Daten und der relativen Häufigkeiten dieser bei den verwendeten MS/MS Methoden ergibt sich die qualitative Wahrscheinlichkeit eines Cross-Links für jede Position im Peptid.

Zur Implementierung des Prototyps werden die Daten eines Standardpeptids verwendet, um die Fragestellung auf die fundamentalen Eigenschaften zu reduzieren. Später in Kapitel 5 Ergebnisse wird eine erneute Analyse mit realen Daten eines zweiten Cross-Link-Peptids durchgeführt. Die Resultate bestätigen die bisherigen Zwischenschritte und liefern differenzierte Aussagen. Zum Abschluss der Arbeit enthält Kapitel 6 Fazit und Ausblick Vorschläge für eine zukünftige Weiterentwicklung des Tools sowie deren Anwendungsmöglichkeiten in anderen Frameworks.

Inhaltsverzeichnis

1 Einführung

1.1 Einleitung

1.1.1 Cross-Links

Medizinische Präparate, tierisches Leder und menschliche Haare haben etwas gemeinsam. Sie alle besitzen eine Form von chemischer Querverbindungen, den Protein Cross-Link (engl. Vernetzung). Ein Cross-Link verbindet lineare Polymerketten miteinander und fixieren sie, ähnlich wie eine Büroklammer mehrere Papiere zusammenhält. Die Bindung kann kovalent oder ionisch sein, erhöht die Stabilität eines Gewebes und macht es haltbarer gegenüber Umwelteinflüssen wie Hitze und UV-Strahlung. Der Schmelzpunkt sinkt und die Löslichkeit in allen Medien nimmt ab. Neben Polymeren aus Kohlenhydraten oder Polyester kommen Cross Links besonders häufig bei den Peptidketten von Proteinen vor.

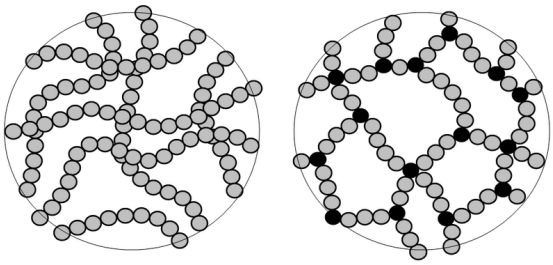


Abbildung 1: Polymere vor und nach einer Vernetzung [1]

Durch die Fähigkeit ein festes Gewebe zu erzeugen, werden Protein Cross-Links unter Anderem in der regenerativen Medizin eingesetzt. Kollagen-Quervernetzungen helfen bei Keratokonus oder anderen Veränderungen der Augen-Hornhaut und werden als Corneal-Cross-Links [2] bezeichnet. Pyridinium-Cross-Links wiederum dienen als pathologische Marker für den Knochen- und Nierenstoffwechsel [3].

In modernen Laboren sind Cross-Links ebenso weit verbreitet. Zum Einen sind sie bei der Herstellung von Polyacrylamid-Gelen im Einsatz. Zum Anderen stabilisieren sie Protein-Protein-Interaktionen und machen damit die reversiblen, flüchtigen Bindungszustände von Proteinen detektierbar. Ein Protein-Cross-Link erhöht die Thermostabilität, vermindert die Denaturierung und blockt die Proteolyse von Peptiden, sodass sich deren biologische Halbwertszeit erhöht. Einige Standard-Untersuchungsmethoden wie die SDS-Page oder der Westernblot sind sogar auf dieses Festigkeit angewiesen. Ein Stoff, der Quervernetzungen herstellt, wird Cross-Linker genannt. Die Anwendungsgebiete reichen über weite Felder der organischen und anorganischen Chemie.

1.1.2 Capture Compound Technologie

Cross-Linker können auch als Filter genutzt werden indem gezielt bestimmte Proteine kovalent binden. Das vereinfacht die Synthese und Analyse von Stoffgemischen erheblich, denn je sauberer ein Protein extrahiert werden kann, desto effizienter sind Verfahren und Produktionsabläufe. Das ist gerade in Anwendungsgebieten mit hohem Qualitätsanspruch wie etwas in der Medikamentenentwicklung wichtig.

Zum Filtern per Cross-Link dienen spezielle Fänger-Moleküle (engl. Capture Compounds, kurz CC). Sie bestehen aus mehreren Untereinheiten mit verschiedenen Fähigkeiten und Aufgaben. Das hier dargestellte Capture Compounds enthält vier funktionelle Einheiten und besitzt ein molekulares Gewicht von etwa 1115 Dalton (Da).

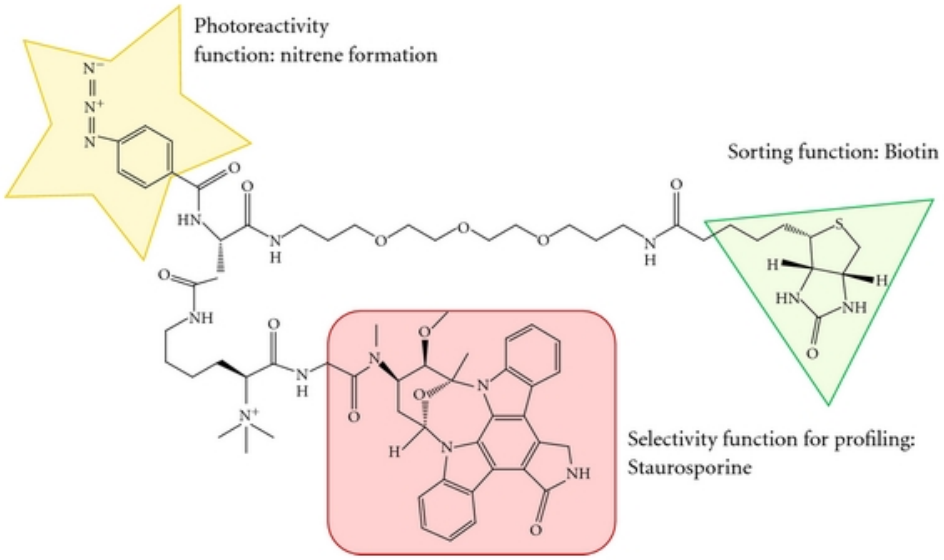


Abbildung 2: Capture Compound [4]

Der Arbeitsweise eines CC beinhaltet mehrere Stufen. Zunächst reagiert die rot-markierte Selektivitäts-Einheit (engl. Selectivity Function) und bindet reversibel an das Protein. Die Bindungsstelle wird über eine spezifische Peptidsequenz bestimmt und berücksichtigt die Stereometrie der Umgebung, sodass nur eine ganz bestimmte Art von Proteinen 'gecatcht' wird. Anschließend bildet die orange-markierte Reaktivitäts-Gruppe (engl. Reactivitiy Function) mit einer beliebigen Aminosäure in der Nähe ein irreversible, kovalente Bindung aus, der eigentlich Cross-Link entsteht. Die Reaktion wird durch Photonen induzierten und deshalb auch als 'blitzen' bezeichnet. Weil die Bindung zufällig in der Umgebung der Selektivitäts-Einheit entsteht, kann die exakte Position nur schwer bestimmt werden. Zum Markieren des Cross-Links in einer Lösung wird die gelb-markierte Sortierungs-Einheit (engl. Sorting Function) verwendet, eine Biotin Gruppe.

1.2 Grundlagen

1.2.1 Basiswissen

Proteinstruktur Als organische Moleküle bilden Proteine das chemische Rückgrat unseres Körpers. Sie übernehmen zahllose Funktionen, vom Stoffwechsel über Strukturbildung bis hin zur Immunreaktion. Die Fähigkeiten eines Proteins werden dabei über seine Form bestimmt, die in der Primärstruktur einer langen Perlenkette ähnelt. Deren Glieder bestehen aus einer von 20 möglichen Aminosäuren (AA) und sind jeweils über eine Peptid-Bindung miteinander verknüpft. Jedes Protein besitzt damit eine einzigartige Sequenz aus Aminosäuren, die wie ein chemischer Fingerabdruck eindeutig identifizierbar ist. Durch die unterschiedlichen physikalischen Eigenschaften der einzelnen Aminosäure wie Ionenstärke, Lipophilität und Stereometrie formt sich eine individuelle räumliche Struktur aus.

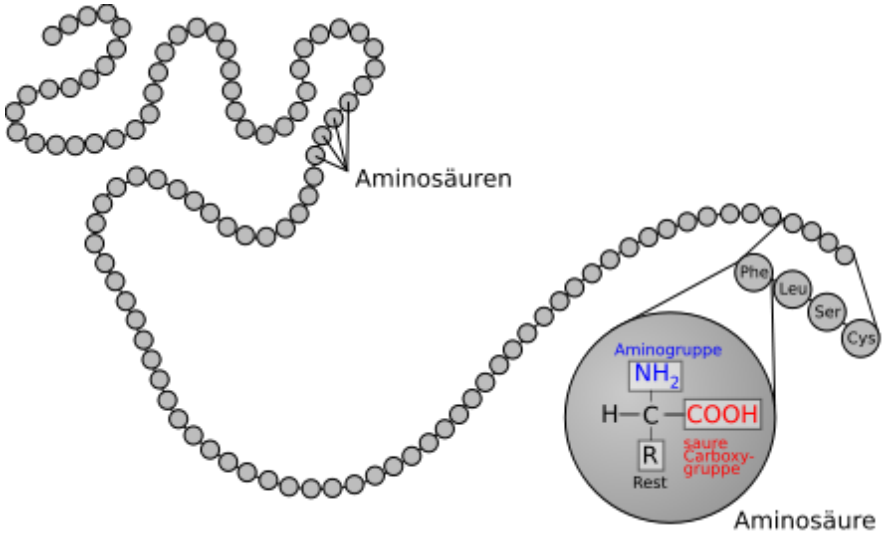


Abbildung 3: Primärstruktur eines Proteins [5]

Massenspektrometrie Moderne Protein-Identifizierung wird mittels Massenspektrometer (MS) durchgeführt. Dabei wird eine Probe in ihre Bestandteile aufgetrennt und dessen Masse ihrem Masse-Ladungs-Verhältnis m/z nach sortiert. Der Grundaufbau besteht aus einer Ionenquelle, einem Analysator und einem Detektor, die sich allesamt in einem Hochvakuum befinden. Überwacht wird der Aufbau über ein Datensystem, dass am Ende ein Massenspektrum der Probe liefert. Die einzelnen Elemente des MS können variieren ja nach Anforderungen an die Qualität und Eigenschaften der Probe. Verwendet für diese Abschlussarbeit wurde eine Orbitrap Maschine mit ESI und CID verwendet.

Abbildung 4: Aufbau eines Massenspektrometer [6]

Zunächst wird die gelöste Probe in der Ionenquelle mithilfe einer Kapillare und einer Elektrode positiv geladen, eine sogenannte Elektrospray-Ionisation (ESI). Der jeweilige Ladungszustand der Teilchen ist dabei zufällig, sodass einfach wie mehrfach geladene Kationen entstehen. Anschließend beschleunigt ein konstantes elektrisches Feld die Probenteilchen und leitet sie anschließend in ein folgendes wechselndes elektronisches Feld ein. Das senkrecht zur Flugbahn stehende Wechselfeld ist aus vier Stabelektroden aufgebaut, ein sogenannter Quadropol. Anders als zuvor erfolgt hier keine weitere Beschleunigung, sondern eine Ablenkung der Ionen nach der Lorentzkraft. Je stärker geladen und je kleiner ein Teilchen ist, desto stärker wird seine Flugbahn gekrümmt. Auf diese Weise werden die Teilchen im Analysator eines MS selektiert.

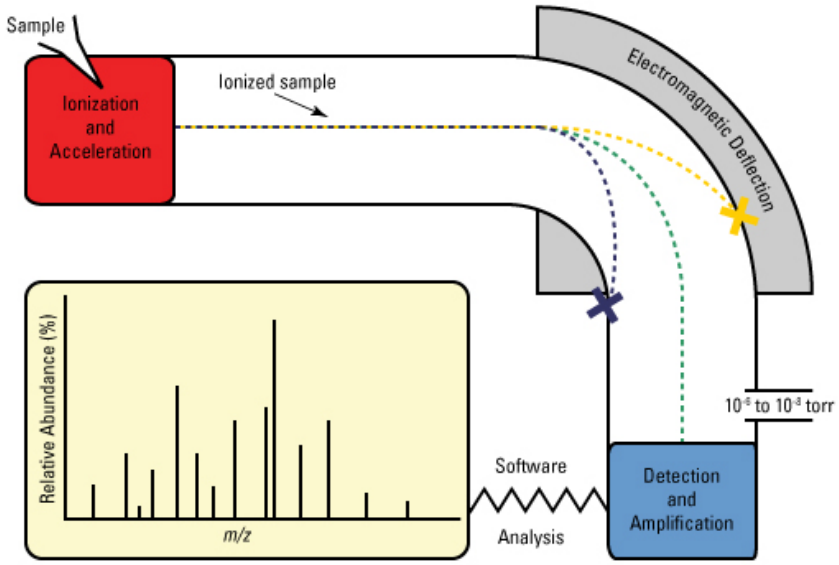


Abbildung 5: Vereinfachtes Prinzip eines Massenspektrometers. In der Realität sieht der Aufbau durch die modernen Bauteile (z.B. Quadropol, Ion Traps) deutlich anders aus. [6]

Alle Bestandteile einer Probe werden also ihrem Masse-Ladung-Verhältnis m/z nach aufgetrennt, ähnlich wie ein Prisma Licht nach der Wellenlänge aufspaltet. Der Detektor registriert die Flugbahn der Teilchen, zeichnet deren Häufigkeit auf und sortiert sie ihrem m/z-Wert nach aufsteigend. Die entstehende Verteilung wird als Massenspektrum

(MS Spektrum) bezeichnet und ist charakteristisch für die Bestandteile einer Probe. Die detektierten Ionen entsprechen dabei den Peaks im MS Spektrum und werden in der Peak Liste gespeichert.

Aufreinigung Bei der Analyse von Biomolekülen ist es sinnvoll, vorher die Probe zu bearbeiten um die entstehenden Datenmengen zu reduzieren und zu strukturieren. Proteine bilden hoch komplexe Strukturen und sind bis zu 100 Kilo-Dalton (kDa) schwer, daher ist einige Vorbereitungen nötig für eine eindeutige Proteinidentifizierung. Als Erstes werden die gelösten Proteine in der Probe enzymatisch verdaut und durch ein Enzym mit restriktiven und bekannten Schnittmuster gespalten. Oft wird Trypsin dafür verwendt. Die Proteine werden dabei an mehrfach an enzym-spezifischen Stellen ihrer AA-Sequenz hydrolysiert und die Peptidbindungen zwischen den Aminosäuren aufgelöst. Die entstehenden Bruchstücke - Peptide genannt - sind mit etwas 1-2 kDa deutlich kleiner als das Ausgangsprotein. Ist die AA-Sequenz des Proteins bekannt, entsteht ein eindeutiges Spaltungsmuster, dass je nach Protein unterschiedlich ist. Ein beliebiges Peptid kann daher jederzeit seinem Ursprungs-Protein zugeordnet werden.

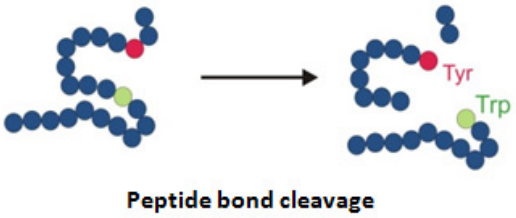


Abbildung 6: Protein Zersetzung [7]

Nach dem Proteinverdau schließt sich eine Chromatographie an. Diese Methode trennt Stoffgemische zeitlich auf und nutzt dafür die unterschiedlichen Eigenschaften der Bestandteile wie Löslichkeit, Größe oder Ionenstärke. In unserem Fall wird eine Hochleistungsflüssigkeitschromatographie (engl. high performance liquid chromatography, HPLC) verwendet. Ziel der HPLC ist es, die Peptide nach der Digestion zeitlich versetzt und ihren Eigenschaften nach geordnet zu eluieren. Damit erhält der folgende Massenspektrometer eine Zeitachse für alle Peaks und die Analyse der Probe wird strukturierter.

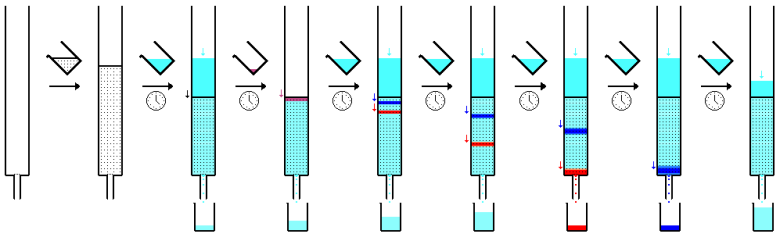


Abbildung 7: Schematische Arbeitsschritte einer Chromatographie (Schwerkraft). [8]

1.2.2 Tandem-Massenspektrometrie

Jedes eluierte Peptid wird nun per Massenspektromter gemessen und die zugehörige Masse bestimmt, auch Pepmasse genannt. Das allein reicht jedoch nicht aus, um ein Protein eindeutig zu Identifizieren, denn mehrere Peptide können die gleiche Masse besitzen. Im MS Spektrum sind diese Pepmassen dann nicht mehr differenzierbar. Die Ausweg besteht darin, die Auflösung der Spektren zu erhöhen und zwei MS hintereinander zu koppeln. Dieser Doppelaufbau wird Tandem-Massenspektrometer oder kurz MS/MS genannt. Zwischen den beiden Durchläufen im Tandemspektrometers erfolgt dabei eine Peptid-Fragmentierung, die aus einem Precursor Ion des ersten MS (MS1) durch collision-induced dissociation, kurz CID, mehrere Product Ionen erzeugt. Im anschließend zweiten Durchgang (MS2) werden die Peptid-Fragmente erneut ihrem m/z Wert nach selektiert und detektiert.

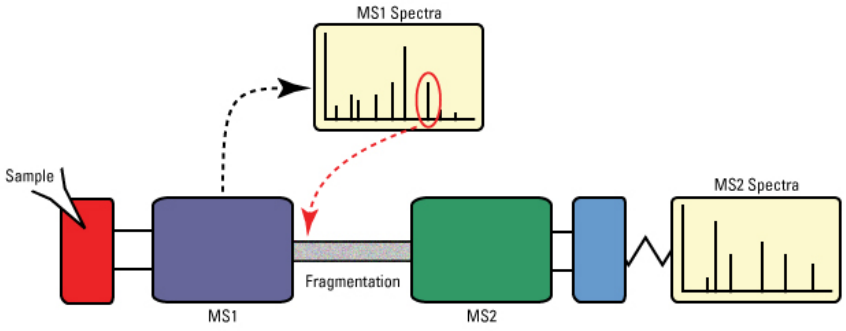


Abbildung 8: Funktionsprinzip eines Tandem-Massenspektrometers. [6]

Im MS1 Spektrum sind die Amplituden der Peptid-Massen mit jeweiliger Retentionszeit von der Chromatographiesäule aufgezeichnet. Etwa 1/10-tel Sekunde nach der Messung wird die Precursor-Masse fragmentiert und das MS/MS gemacht. Diese MS2 Spektrum enthält jeweils die Product Ionen des Precursors und ist je nach Peptid unterschiedlich. Dadurch kann eine eindeutige Identifizierung wieder stattfinden. Für die spätere Datenbanksuche bietet die Pepmasse und die Retentionszeit aus dem MS1 eine guten Richtwert und schränkt die Anzahl der Möglichkeiten ein. MS1 und MS2 sollten immer zusammen betrachtet werden sollten. Die größte gemessene Masse im MS2 ist nie größer als die Precursormasse.

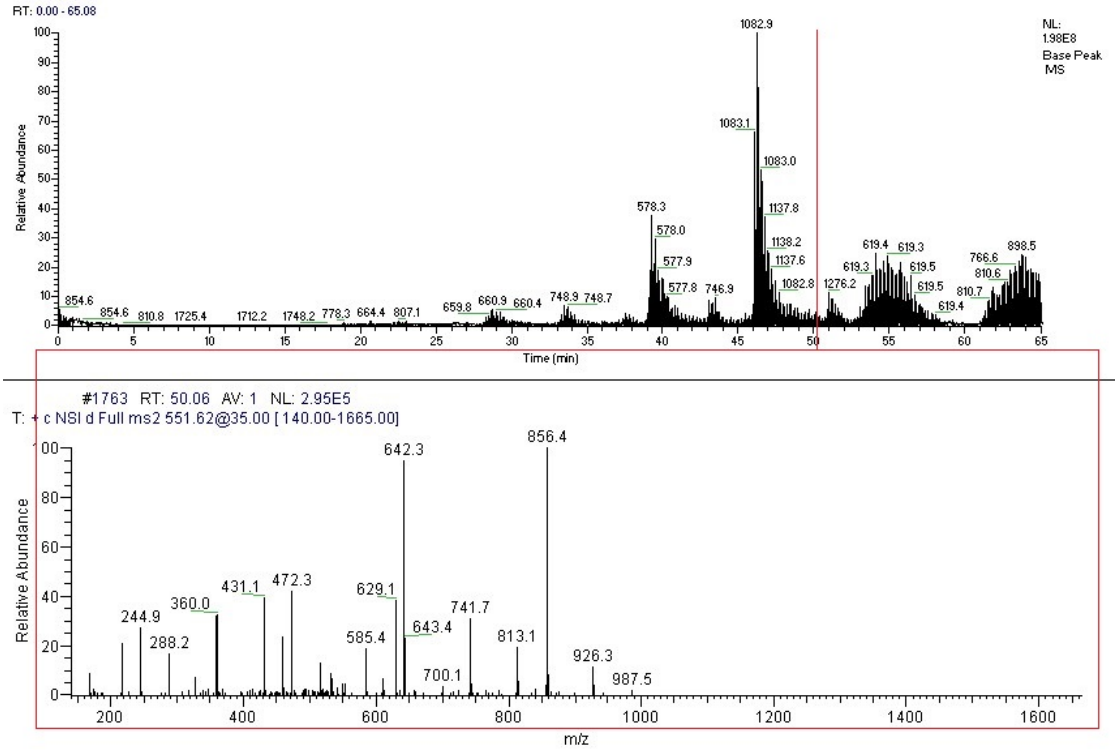


Abbildung 9: MS/MS-Spektrum vor der Fragmentierung (oben) und nach der Fragmentierung (unten) [9]. Der Precurosr Peak ist im MS rot markiert und liefert das Ausgangspeptid der folgenden Fragmentierung in Product Ionen für das MS/MS.

Peptid Fragmentierung Während der Peptid Fragmentierung werden neutrale Teilchen wie Neon oder Argon auf das bereits stark beschleunigte Peptid-Ion aus dem MS1 abgeschossen. Beim Zusammenstoß wandelt sich die hohe kinetische Energie E_{kin} des Precursors um in innere Energie U und löst die Bindungen zwischen den einzelnen Aminosäuren auf (vgl. CID). Je nach E_{kin} des Precursor Ions werden dabei anderen Typen von Product Ionen erzeugt, denn die Bindungen im Peptid werden ihrem Energie-Niveau nach gespalten. Die schwächeren Peptid-Bindungen zwischen den Aminosäuren sind als Erste betroffen, wobei die Bruchstelle jeweils kurz davor, mittendrin oder kurz dahinter entsteht. Das ganze Ähnelt einem Crashtest mit einigen Soll-Bruchstellen des Fahrzeugs.

Welche der Peptidbindungen zuerst dissoziiert, ist zufällig. Jede Bruchstelle erzeugt dabei zwei sich ergänzende Fragmente des Peptids, jeweils eines mit C-Terminus und eines mit N-Terminus. Die Ladung des Precursor Ions befindet zufällig nur auf einer der beiden Seiten und lediglich dieses Fragment ist detektierbar als ein Product Ion. War der Precursor mehrfach geladen, sind zwei geladene Product Ionen möglich (Precursor: $3^+ \rightarrow$ Product: $1^+ + 2^+$). Landet die Ladung auf dem Fragment mit N-Terminus, wird es a-,b- oder c-Ion genannt. Findet sie sich auf der C-Terminus Seite, heißt es x-,y- oder z-Ion

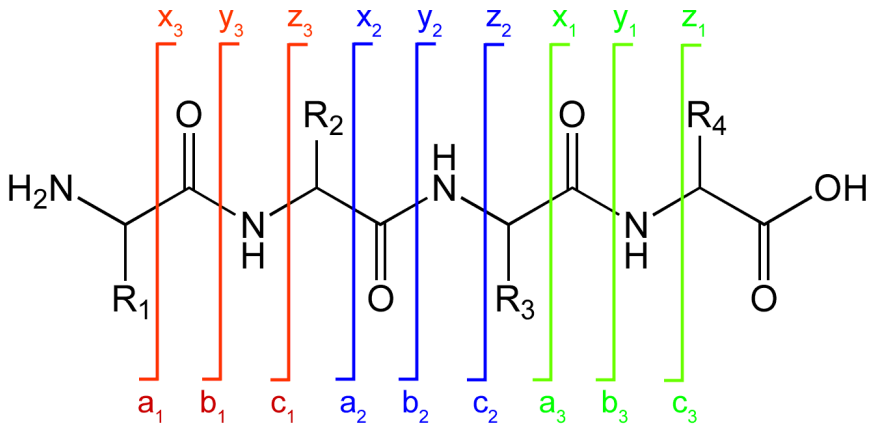


Abbildung 10: Peptid-Fragmentation nach Roepstorff und Fohlman (1984) [10]

Theoretisch sind alle Fragment-Typen möglich, vorherrschend sind aber y- und b-Ionen aufgrund der konstanten Beschleunigung im MS1. Wenn genug Precursor Ionen gespalten werden, existieren irgendwann für jede Peptidbindung die zugehörigen Product Ionen. Bei einem Peptid mit der AA-Sequenz $1, 2, 3, ..., n$ liegen dann also die Ionen Serien $b_1, b_2, b_3, ..., b_n$ und $y_n, y_{n-1}, y_{n_2}, ..., y_1$ vor. Sortiert nach ihrer Größe unterscheiden sich zwei adjazente Ionen einer Serie nur um die Masse der einen abgespaltene AA-Residue voneinander. Diese Massendifferenz ist charakteristisch für die jeweilige Aminosäure (siehe Tabelle 6 auf Seite 44) und rekonstruiert die komplette Peptidsequenz bei einer vollständigen Ionenserie, auch De-Novo-Sequenzierung genannt.

1.2.3 Auswertung von Tandem-Massenspektren

Molekular Masse und Theoretische Spektren Ist die AA-Sequenz eines Peptids bereits bekannt, kann leicht die Molekulare Masse M und der m/z-Wert des Teilchens berechnet werden. Dazu müssen die Massen die einzelnen Aminosäuren aufsummiert, mit einem Fixwert für den C- und N-Terminus sowie dem Ladungsteilchen addiert und durch dessen Ladung dividiert werden:

$$M = M_c + M_n + \sum N_i \cdot M_i \;\rightarrow\; \frac{m}{z} = \frac{M + z \cdot H}{z}$$

M = Molekulare Masse	M_c = Masse C-Terminus	M_n = Masse N-terminus
i = Aminosäure	N_i = Anazahl AA	M_i = Masse AA
$\frac{m}{z}$ = Masse-Ladungs-Verhätnis	H = Masse Proton	z = Ladung

Neben dem Molekulargewicht kann aus der Peptidsequenz ebenso die theoretische Fragmentierung abgeleitet werden. Denn zusammen mit einer angenommen Ladung lassen sich aus der molekularen Masse die theoretischen Ionenserien berechnen, indem vom

aktuellen m/z-Wert für jede Aminosäure der Sequenz die jeweils entsprechende Residuenmassen abgezogen wird. Da diese Berechnung relativ simpel ist, werden frei zugänglich Onlinetools [11] angeboten die theoretische Spektren berechnen. Zur Illustration dient dazu an dieser Stelle ein Beispielpeptid, das Glu-1-Fibrinopeptide B (GluFib) [12]. GluFib ist als Massen-Standard weit verbreitet wird zum Eichen von MS Anlagen verwendet; seine monoisotopische Masse beträgt 1569.67 Da und er besitzt die Peptidsequenz 'EGVNDNEEGFFSAR'.

Ladung	m/z
(M)	1569.66961
(M+H)+	1570.67743
(M+2H)2+	785.84265
(M+3H)3+	524.23106
(M+4H)4+	393.42526

Seq	Nr.	B Ions	Y Ions	Nr.
E	1	130.05046	1570.67743	14
G	2	187.07193	1441.63484	13
V	3	286.14034	1384.61337	12
N	4	400.18327	1285.54496	11
D	5	515.21021	1171.50203	10
N	6	629.25314	1056.47509	9
E	7	758.29573	942.43216	8
E	8	887.33832	813.38957	7
G	9	944.35979	684.34698	6
F	10	1091.42820	627.32551	5
F	11	1238.49661	480.25710	4
S	12	1325.52864	333.18869	3
A	13	1396.56576	246.15666	2
R	14	1552.66687	175.11955	1

Tabelle 1: Theoretische Pepmasse von GluFib bei unterschiedlichen Ladungen (links) und das theoretisches Spektrum von 'EGVNDNEEGFFSAR' bei Ladung +1 und y/b-Ionentyp (rechts) [11]

Peak Annotation Normalerweise ist die Peptidsequenz eines Precursors nicht gegeben und soll via MS/MS ermittelt werden. Dazu müssen Peaks im MS/MS gefunden werden, deren Massen sich um jeweils eine Aminosäure voneinander unterscheiden. Eine zusammenhängende Ionenserie rekonstruiert dann die Peptidsequenz des Precursors. Die Herausforderung bei der Auswertung von MS/MS Spektren besteht daher in der korrekten Annotation der Peaks zu einer Ionenserie. In der Regel wird dazu ein experimentelles MS/MS Spektrum gegen ein theoretischen Spektrum mit der gleichen Pepmasse verglichen. Stimmt ein Peak in beiden Spektren überein, wird dieser entsprechend gekennzeichnet und mit einem Label versehen. Denn wie oben in der Tabelle 1 abgebildet, kann jeder theoretischen Peak zu einer Ionenserie und dem Fragment-Ion mit entsprechender Ladung zugeordnet werden.

Eine Datenbank aus bekannten Peptid-Fragmentierungen liefert die theoretischen Spektren für den Vergleich. Anhand der Anzahl und Qualität der Matches erhält dann jeder Kandidat einen Score und je höher dieser, desto wahrscheinlicher handelt es sich um das zugehörige Peptid. Oft reicht schon ein Sequence Tag von vier bis fünf zusammenhän-

gende Peaks aus für eine eindeutige Identifizierung. Nach diesem Prinzip untersucht gängige Proteomics Analyse Software wie SEQUEST [13] oder MASCOT [14] ihrer MS/MS Spektren.

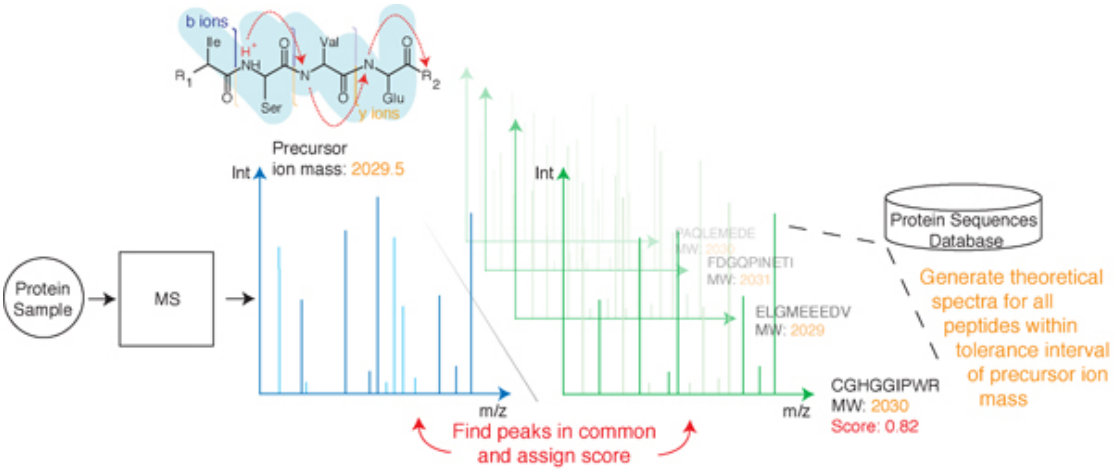


Abbildung 11: Protein Identifikation via MS/MS und Protein Sequenz Datenbank [15]

Je stärker geladen das Precursor Ion war, desto mehr Kombinationen an Ionenserien und Ladungszuständen sind möglich. Die Liste der experimentellen Peaks verlängert sich und das MS/MS Spektrum wird immer kontinuierlicher. Gleichermaßen passen jetzt komplett andere Peptide zufällig mit dem experimentellen Spektrum übereinstimme, denn je mehr Peaks im Spektrum, desto wahrscheinlicher willkürliche Matches. Die korrekte Annotation wird immer schwieriger und die Quote der False-Positives steigt.

Darüber hinaus sind die relativen Häufigkeiten der einzelnen Product Ionen nicht identisch, sondern poisson-verteilt [16]. Zusammen mit dem normalen Hintergrundrauschen, der Gaußverteilung der Intensitäten und Artefakten aus der Fragmentierung wird das Peptid Signal stark gedämpft. Ein Peak ist nicht mehr zweifelsfrei als solcher erkennbar, eine eindeutig Identifikation wird unmöglich. Zwar helfen spezielle Filter das Spektrum zu glätten und die Peaks klarer zeichnen, gleichzeitig gehen dadurch freilich Informationen verloren, die zusammen mit dem Background ausgefiltert werden. Das ist ein zentrales Problem vieler Suchalgorithmen.

Im Großen und Ganzen ist es aber möglich, Peptide zu identifiziert und ihrem Ursprungsprotein zuzuordnen, solange die MS/MS Spektren eine genügende Qualität besitzen. Einige Intensitäten lassen sich allerdings keiner Ionenserie zuordnen und bleiben ohne Label. Hier ein reales Beispiel für eine gelungene Annotation von GluFib.

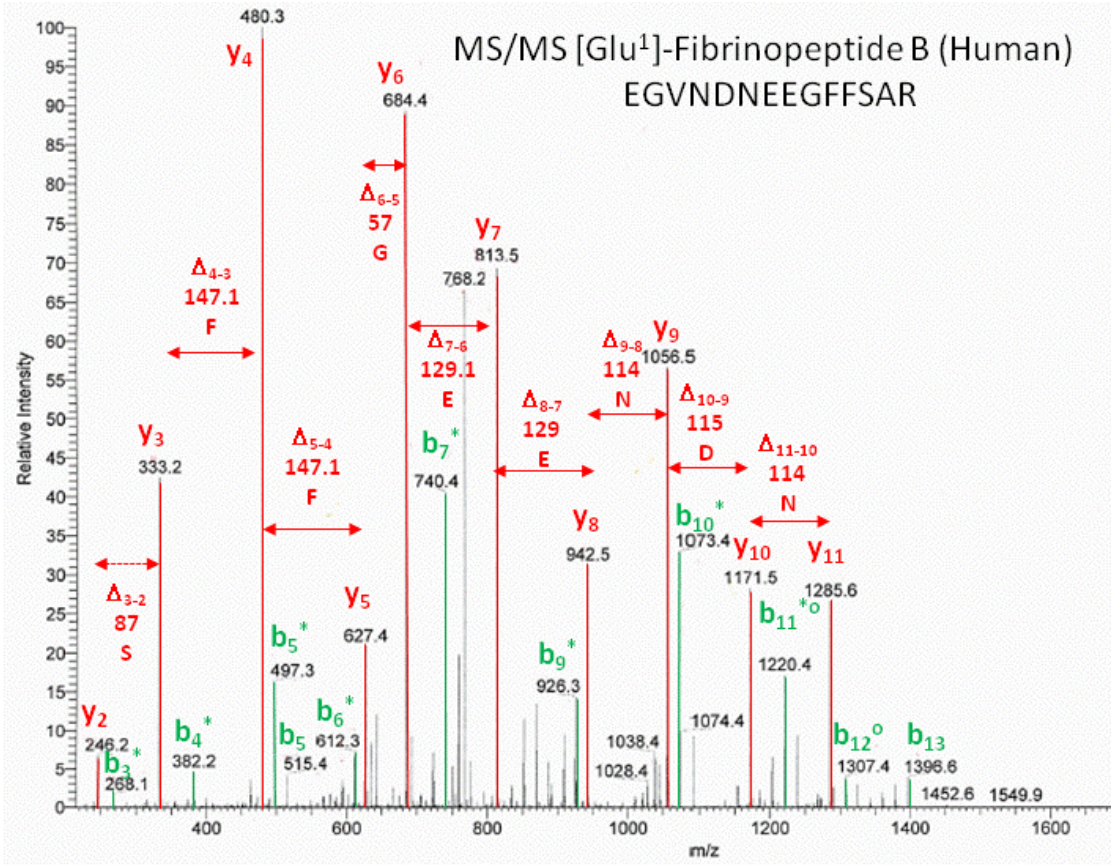


Abbildung 12: Peak Annotation von GluFib [17]

Indes existieren weitere inhaltliche Ursachen, die eine Annotation beeinträchtigen, wie im nächsten Absatz beschrieben wird.

Neutral Loss Bei der Fragmentierung ist die wahrscheinlichste Bruchstelle die Peptidbindung zwischen den Aminosäuren. Zusätzlich können in den Seitenketten der Aminosäuren aber weitere Bindungen dissoziieren und sehr kleine Fragmente erzeugen, wie CO, N_2, C_2H_4 oder CH_2N. Am Häufigsten entstehen hierbei NH_3 und OH, Amino- und Hydroxy-Bruchstücke. Zwar tragen diese keine Ladung, dennoch das beeinflussen sie das Masse-Ladung-Verhältnis ihres Ursprungs-Ions wegen der nun fehlenden Masse. Daher die Bezeichnung Neutral Loss. Die erzeugten Product-Ionen besitzen einen etwas geringeren m/z-Wert, sodass der dazugehörige Peak etwas nach links verschoben ist verglichen mit dem des normalen Product-Ion. Um das zu kompensieren, wird ein zusätzliches theoretisches Spektrum berechnet und um die Masse des Neutral Losses korrigiert. Weil die Zusatzfragmentierung aber weniger häufig passiert als die normale, kann ein Peptid in der Regel auch ohne deren Berücksichtigung identifiziert werden. Allerdings hebt eine Analyse der Neutral Loss Spektren den Score einer Annotation erheblich.

Modifikationen von Peptiden Eine andere Quelle für Abweichungen im MS/MS Spektrum sind Veränderungen am Protein selbst. Die Peptid-Datenbank beinhaltet nur die native Version eines Moleküls, in vivo sind Proteine aber phosphoryliert, ubiquitiniert oder anders post-translational modifiziert. Durch jede Form der Abwandlung nimmt die Peptid-Masse zu und die zugehörige Peak Liste wird im MS/MS ein Stück nach rechts verschoben, ähnlich wie der Verlust von Masse beim die Neutral Loss sie nach links rückt. Im Gegensatz zum Neutral Loss wird bei einer Modifikation allerdings die abnormale Masse irgendwann bei der Fragmentierung wieder abgespalten. Ab diesem Punkt sind die Massen der Product-Ionen wieder normal und die Peaks wieder auf den Positionen vom unmodifiziertem Spektrum. Es entsteht ein Komposit Spektrum aus modifizierten und unmodifizierten Ionen.

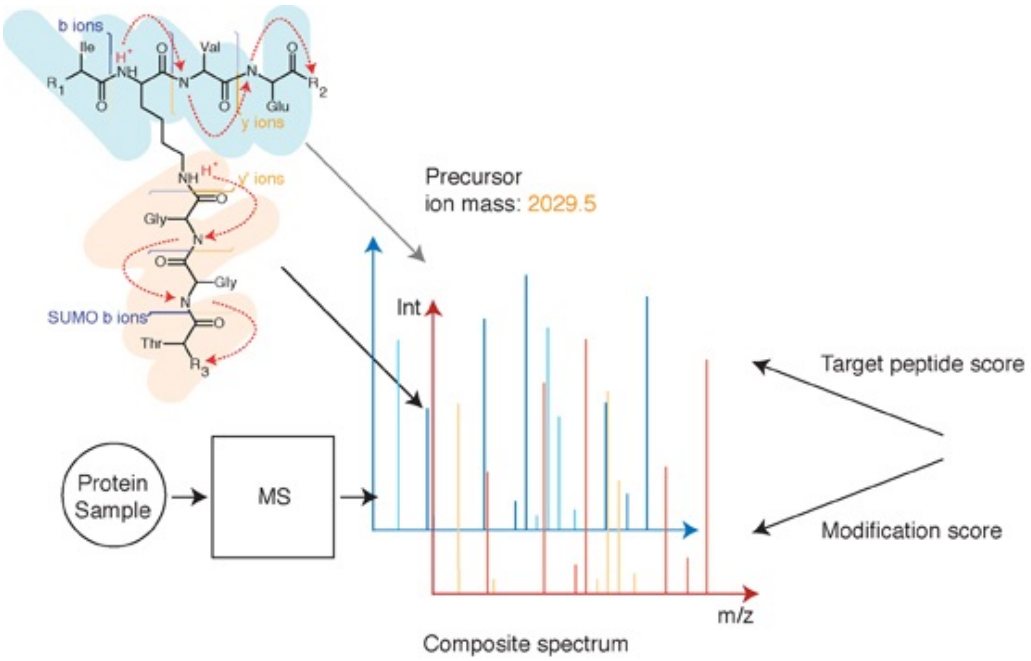


Abbildung 13: Massen-Shift durch Ubiquitinierung erzeugt das Komposit-Spektrum mit normalen (blau) und verschobenen Peaks. SUMO = Small ubiquitin-like modifier. [15]

Bei einer unbekannten Modifizierung bleibt nur eine De-Novo-Sequenzierung übrig. Ist die Art der Modifikation oder dessen Position allerdings bekannt, lässt sich die veränderte Pepmasse und ein modifiziertes theoretisches Fragmentierung noch berücksichtigen. Für einfache Variationen, wie Phosphorylierung oder Methylierung, existieren sogar Suchoptionen bei Analysesoftware für MS/MS. Für komplexen Modifikationen gibt es jedoch bisher wenige Lösungsansätze. Deren Problematik liegt darin, dass größere Strukturen im MS/MS mit fragmentieren und selbst zusätzliche Product Ionen erzeugen. Besonders wenn die Modifikation variabel ist und mehrere Fragmentationsmöglichkeiten innerhalb des Appendix bestehen, wird die Annotation stark beeinträchtigt. Eine Identifizierung des Peptids ist hier nicht mehr möglich.

1.3 Motivation

Das Unternehmen *Caprotec Bioanalytics GmbH* mit Sitz Berlin-Adlershof entwickelte die Capture Compund Mass Spectrometry (CCMS) Technologie. Die vorliegende Abschlussarbeit wird in Kooperation mit *Caprotec* und der Freie Universität Berlin verfasst. Bisher verursacht ein gebundenes Capture Compound starke Interferenzen im MS/MS des Zielpeptids und verhindert eine Identifizierung mit Standardmethoden wie SEQUEST oder MASCOT. Der Grund dafür liegt in der variable Fragmentierung der Modifikation, die zu komplex und zu individuelle ist für etwaige Suchoptionen. Die entstehenden Massenshifts sind unregelmäßig, schlecht vorherzusagen und die Bindungspostion nicht fest.

In dieser Arbeit soll deshalb untersucht werden, ob sich ein Tool entwickeln lässt, dass ein gebundenes Capture Compound und das dazugehörige Peptid im MS/MS Spektrum erkennen kann. Überprüft werden soll, welche Erkennungsmerkmale für die Bindung des Cross-Links im MS/MS Spektrum existieren und wie trotz Modifikation ein Peptid identifiziert werden kann. Desweiteren soll versucht werden, ob die Bindungspositionen des Capture Compounds lokalisiert werden kann und ob diese eindeutig ist oder aber variiert innerhalb einer Verteilung. Es handelt sich dabei um eine Machbarkeitsstudie, bei der Beweis der Durchführbarkeit im Vordergrund steht.

Als langfristiges Ziel ist die Entwicklung eines Workflows zur Analyse von Capture Compound Spektren geplant. Eine Erkennungsroutine soll zukünftig in MS Rohdaten die modifizierte Spektren erkennen und extrahieren. Die veränderte MS/MS Spektren werden dann mit einem Tool annotiert und die Peptide identifiziert sowie die Bindungstelle des CC grob lokalisiert. In ein passendes Framework für Massenspektrometrie eingebunden wie OpenMS [18] gibt es viele Anwendungsmöglichkeiten dafür. Die Entwicklung dieses Tools bildet diese Abschlussarbeit sowie ein Testlauf mit realen Daten.

1.3.1 Bisherige Arbeiten

Konstante Modifikationen von Peptiden können bereits gut erkannt werden mit gängigen Annotationstools. Einzelne Phosphorylierungen, Methylierungen oder modifizierte Aminosäuren lassen sich leicht miteinkalkulieren in das theoretische Spektren und bereiten kaum Probleme bei der Identifizierung. Die Komplikationen entstehen erst, wenn große Modifikationen verwendet werden, wie beispielsweise die Glykosylierung von Proteinen. In der Fragmentierungsphase des MS/MS bleiben diese umfangreichen Strukturen nicht intakt und splittern in viele Unterarten von Modifikationen auf. Das Gleiche gilt für den Capture Compound Cross-Link. Aufgrund der großen biologischen Relevanz gibt es bereits ein eigenes Fachgebiet für die Bestimmung von Glykosylierung, Glycomics genannt, sowie etliche Ansätze und Publikationen zur Identifikation [19, 20].
Was die Bindungsstelle eines Cross-Links angeht, so wurde hier ebenso bereits viel Material veröffentlicht. Im Fokus stehen allerdings eher allgemeingültige Aussagen sowie die Grundmechanismen hinter einer Reaktion von Biomolekülen für eine spätere Modellierung mit Vorhersage und Simulierung [21].

1.3.2 Eigener Beitrag

Der Fokus ist, die Erkennung und Verifizierung von variablen Modifikationen zu verbessern. Wenn eine Annotation gelingt, kann mit den erzeugten Ergebnissen auch die mögliche Bindungsstelle der Modifikation analysiert werden. Das dafür notwendige Tool wird in Python verfasst werden und heißt *'Identifikation und Lokalisation von Variablen Modifikationen'* oder kurz *'ilvamo'*. Zur Entwicklung von *ilvamo* wird vorerst ein Prototyp erstellt und anhand eines Minimalbeispiels erprobt. Das dazu verwendete Modell ist der Cross-Link zwischen dem Capture Compound B1SAH6 und dem Standardpeptid Glu-1-Fibrinopeptide B (GluFib) [12]. Ziel ist es, ein Tool mit möglichst gegliederten Worklfow zu schreiben, damit später differenzierte Ergebnisse generiert werden können. Je detaillierter Bindungsverhalten untersuchbar ist, desto größer der zukünftige Nutzen des Tools.

Zunächst werden alle theoretischen Grundlagen analysiert wie etwa das exakte Fragmentationsmuster der Cross-Link-Modifikation und die daraus entstehenden Massenshifts. Neben der Qualität der x-Link Spektren wird ebenso die Bindungsrate des CC untersucht werden. Fehlertoleranz, Genauigkeit und Verhalten bei unterschiedlichen Ladungszuständen werden evaluiert und mithilfe von verschiedenen Scoring-Methoden gewertet. Die Ergebnisse und Beobachtungen diesbezüglich bilden zusammen mit der jeweils beschriebenen Implementierung die Kapitel 2 Identifizierung, 3 Verifizierung und 4 Lokalisierung. Desweiteren wird *ilvamo* in Kapitel 5 Ergebnisse mit weiteren, etwas realistischeren Daten von der modifizierten Methyltransferase des Thermus Aquaticus, kurz M.Taql [22], getestet. Auch hier werden die Ergebnisse, Verhalten und unterschieden zum Testfall mit Glufib analysiert. Anschließend wird in Kapitel 5 Fazit und Ausblick die Kapazität von *ilvamo* definiert und zukünftige Optimierungen vorgeschlagen.

ilvamo besteht am Ende aus vier Python Skripten: *main.py, identify.py, plot_specs.py* und *score.py*. Alle zusammen ergeben das Tool und agieren zunächst als Stand Alone, sollen später aber in ein Framework eingepasst werden. Der Zugriff ist via Shell und jedem beliebigen Python-Editor möglich, alle Ergebnisse werden als Plot in png. Format oder als Liste bzw. Array an den User ausgegeben.

2 Identifizierung

Grundgedanke Wie jede Protein-Modifikation ist auch die Bindung eines Capture Compound mit einem Massenshift im MS/MS verbunden. Da vorher allerdings weder die exakte Position bekannt ist, noch die Modifikation stabil mit fragmentiert, gestaltet sich das Annotieren der Ionenserien schwierig. Wäre beides gegeben, ist die Peptid-Identifikation schnell erledigt mit einer theoretische Fragmentierung wie in Tabelle 2.

Ladung	m/z	m/z **+500**
(M)	1519.818	2019.818
(M+H)+	1520.826	2020.826
(M+2H)2+	760.917	1010.917
(M+3H)3+	507.614	674.281
(M+4H)4+	380.962	505.962

Seq	Nr.	B Ions	Y Ions	Nr.
M	1	132.048	2020.826	15
G	2	189.070	1889.786	14
L	3	302.154	1832.764	13
P	4	399.207	1719.680	12
P	5	496.260	1622.627	11
L	**6**	**1109.343**	**1525.574**	**10**
S	7	1196.376	912.490	9
L	8	1309.460	825.458	8
P	9	1406.512	712.374	7
S	10	1493.544	615.322	6
N	11	1607.587	528.290	5
A	12	1678.624	414.247	4
A	13	1749.662	343.209	3
P	14	1846.714	272.172	2
R	15	2002.815	175.120	1

Tabelle 2: Theoretische Pepmasse (links) und Spektrum (rechts) vom Peptid 'MGLPPLSLPSNAAPR' bei Ladung +1, y/b-Ionentyp und mit einer angenommenen Modifizierung an der sechsten Aminosäure L von 500. Die daraus entstehenden modifizierten Ionenserien sind farbig gekennzeichnet [11]

In diesem Fall kann die normale Datenbanksuche schon die gewünschten Ergebnisse liefern. Weil aber beide dieser Voraussetzungen nicht erfüllt sind, muss für die CC Modifikation eine Alternative her. Der hier in dieser Arbeit vorgestellter Ansatz trennt die Problematik in zwei Teilaufgaben auf: Erstens die Identifikation des Peptids und zweitens die Lokalisation der Bindungsstelle aus dem MS/MS heraus. Das erste Problem wird in dem aktuellen Abschnitt sowie in 3 Verifizierung ab Seite 24 erläutert. Für das zweite Problem muss die Sequenz zuerst bekannt sein, daher bezieht sich das Kapitel 4 Lokalisierung auf Seite 28 auf die Ergebnisse der ersten Teillösung.

Zur Identifizierung der Peptidsequenz muss die Masse der Modifikation ermittelt werden. Nur so kann via Peak Vergleich das Spektrum annotiert werden. Zwar ist die exakte Masse des Capture Compounds bekannt, allerdings bleibt nach der Fragmentation im MS2 nicht intakt erhalten. Der Cross-Link dissoziiert zusammen mit dem Peptid und bildet mehrere neue, variable CC-Fragmente aus.

2.1 Variable Modifikation

Für die unbekannten Massen des fragmentierten Capture Compound lautet der Ausweg, den CC B1SAH6 separat in ein MS/MS zu schicken. Auf diese Weise ergibt sich ein konkretes Spektrum mit Fragment Ionen, dass Aufschluss gibt über die Anzahl, Größe und Häufigkeit der variablen Modifikationen. Sie dienen später als Reporter Ionen im Cross-Link-Spektrum und markieren dort quantitativ die Bindung des CC am Peptid. Je häufiger die Reporter Ionen auftreten, desto mehr Capture Compound hat gebunden und desto stärker sind die modifizierten Ionenserien im Cross-Link-Spektrum vertreten. Ebenso kann der Massenshift durch die CC-Bindung durch die Reporten abgeschätzt werden. Zieht man deren Masse von der Gesamtmasse des CC ab, ergibt sich die Restmasse der Modifikation am Peptid.

CC Masse: 1116 Da − Reporter (+1): 284 = Modifikation: 832 Da

Der Massen-Shift im MS/MS entsteht also durch die Restmasse des Cross-Linkers am Peptid. Ist dessen Umfang bekannt, kann die Masse mit dem theoretisches Massenspektrum verrechnet werden und der Peak Listen Vergleich kann starten. Natürlich sind noch weitere Fragmente im MS/MS des Capture Compounds vorhanden und damit weitere Dissoziationen möglich. Für jede davon müsste also ein eigenes theoretisches Spektrum her. Allerdings kann aus der relativen Häufigkeit der Reporter MS/MS die wahrscheinlichsten Massenverschiebungen bestimmt werden und so für jedes Cross Link Spektrum die stärksten Modi-Massen ermittelt werden. Mit dieser Information lässt sich die Anzahl der benötigten theoretischen Spektren begrenzen.

Name	Reporter	Modifikation
cc1	284	832
cc2	447	669
cc3	1115	0

Tabelle 3: MS/MS Fragmente vom Capture Compound B1SAH6

Damit ist die unbekannte Masse der variablen Modifikation erfasst. Falls der Capture Compound nicht fragmentiert (cc3), hängt die gesamte Masse des Cross-Linkers am Peptid. Der jeweilige Ladungszustand ist nach der Fragmentierung nicht abschätzbar und ein Teil der Ladung bleibt am Reporter Ion. Aufgrund dessen wird angenommen, dass der Massenshift nicht bei allen Ladungsebene gleichtzeitig auftritt. Bei der folgenden Annotation des Cross-Link-Spektrums müssen anders als sonst auch höhere Ladungszahlen berücksichtigt werden. In dem Fall von B1SAH6 wird die maximale Ladung auf 4 geschätzt, sodass die alle Reporter Ionen und Massenshifts darunter oder spätestens exakt bei $z = 4$ auftreten.

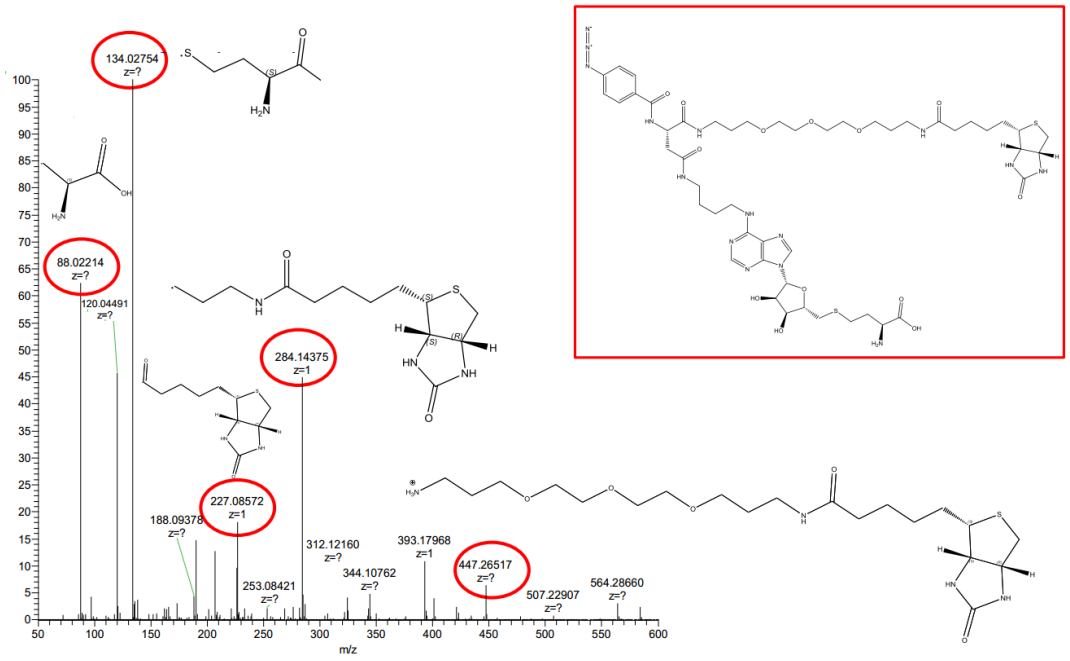


Abbildung 14: MS/MS eines ungebundenen B1SAH, HCD-Fragmentierung. Die markierten Peaks gehören zu dem jeweiligen abgebildeten Teilstücken des CC; die messbaren CC-Fragmente in den x-link Daten sind 284 und 447 Peaks.

2.2 Implementierung

Ausgangssituation Als Ausgangspunkt für *ilvamo* dienen die experimentellen Spektren vom GluFib Peptid, dass bereits oben in der Einleitung als Massen-Standard vorgestellt wurde. Mit diesem Modell-Fall soll das neue Tool Stück für Stück entwickelt werden. Die Messdaten des GluFibs vom MS/MS sind im .xzml File gegeben, jeweils einmal mit und ohne Cross-Link. Um daraus einzelne Spektren zu erhalten, wird vom .xmzML File via OpenMS TOPPView [18] per Hand der zugehörige Scan extrahiert. Anschließend folgt mit Konvertierung in das mgf-Format mittels eine einfachen Pipeline in OpenMS TOPPAS [18]. Dieser Schritt ist notwendig, weil es in Python bereits einen detaillierter Framework basierend auf .mgf Files vorhanden ist. Mit Pyteomics [23] wird eine Infrastruktur zur Analyse von Proteomics Daten gestellt mit File Access & Output, Berechnung physikalisch-chemischer Eigenschaften und theoretischer Ionenserien. Mithilfe dieser Bibliothek wird *ilvamo* angesetzt.

Im Gegensatz zum nativen, unmodifizierten Glufib ist die Cross-Link(x-Link) Version deutlich verrauschter, hat schwächere Intensitäten und weniger Kontrast zwischen den Peaks. Filter und Normalisierung könnten das ausgleichen, allerdings verschwinden viele Informationen zusammen mit dem Rauschen. Auf jede Aufbereitung des modifizierten Spektrums wird daher verzichtet und für eine späteren Version des Tools angesetzt.

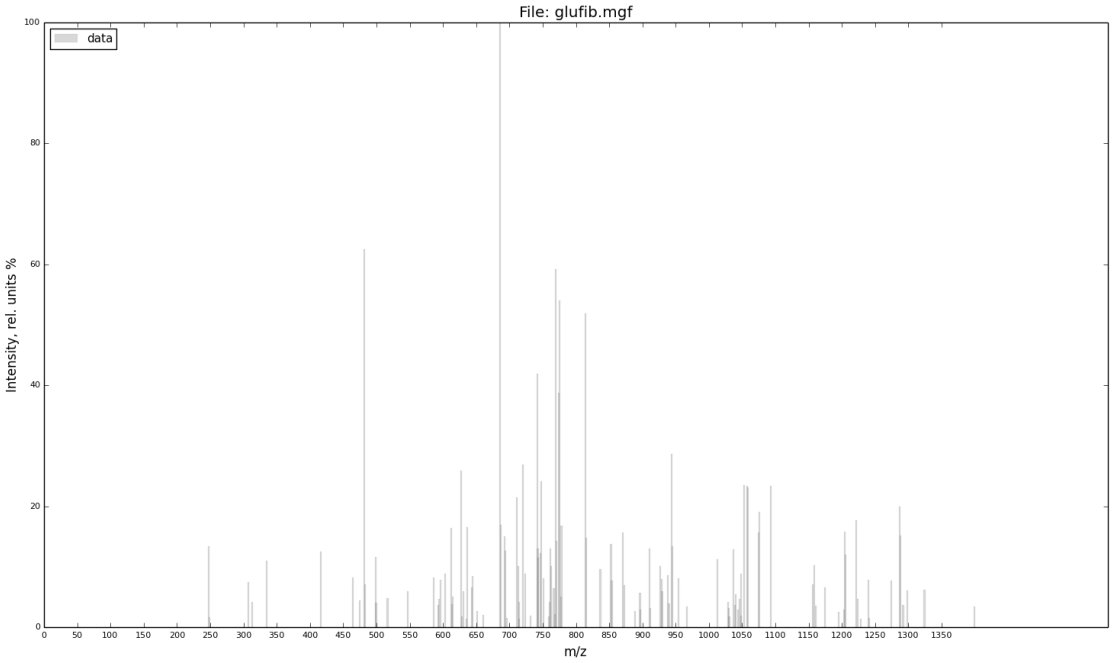


Abbildung 15: Gemessenes GluFib Spektrum (ohne Cross Link)

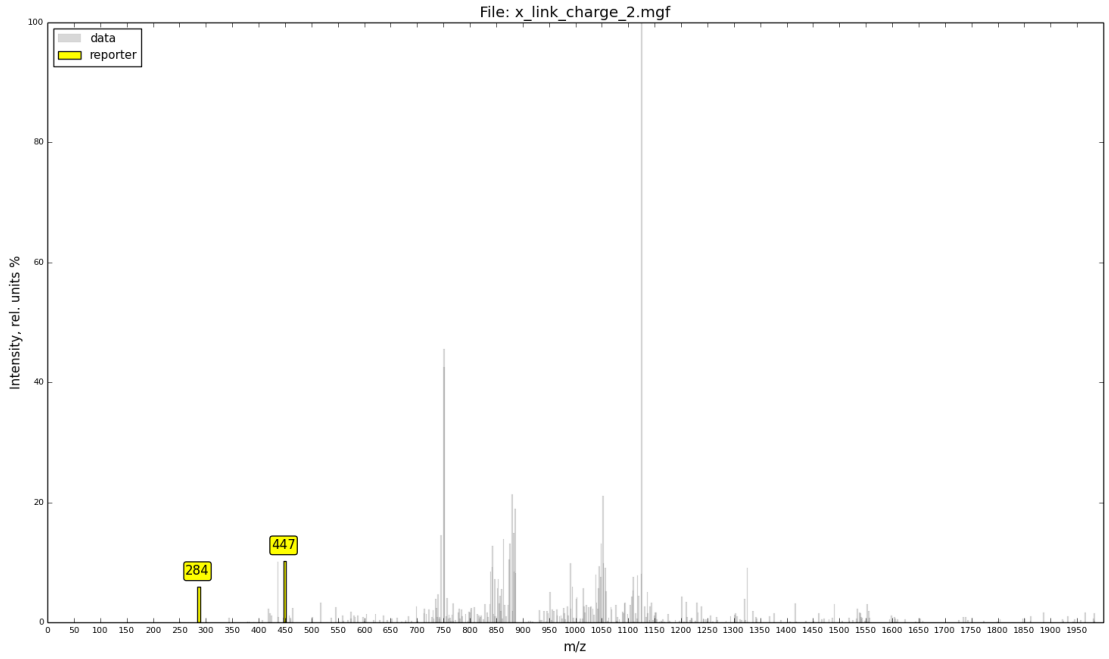


Abbildung 16: X-link GluFib mit markierten Reporter Ionen 284 und 447. An der m/z-Achse ist die deutlich größere Massenspannweite erkennbar.

Theoretisches Spektrum Das theoretischen Spektren für den Vergleich stammt bei *ilvamo* aus der Funktion *fragment_y()* bzw. *fragment_b()*, einer leicht abgewandelten Version der *fragments(y,b)* Routine aus dem Pyteomics Package. Anschließend werden den CC-Massen (siehe Tabelle 3) addiert und daraus die modifizierte Serie erzeugt. Zusammengefasst entstehen hier vier theoretische Spektren, eine unmodifizierte und jeweils eine modifizierte Serien für cc1,cc2 und cc3. Die modifizierten Peak Listen sind dann jeweils um die Residualmasse des CC nach rechts verschoben im Spektrum. Im gleichen Schritt können bei ebenso andere Massenverschiebungen wie Neutral Losses mit verrechnet werden. Bei dem entwickeltem Tool sind differenzierte theoretisches Spektren möglich je nach Intention des Anwenders.

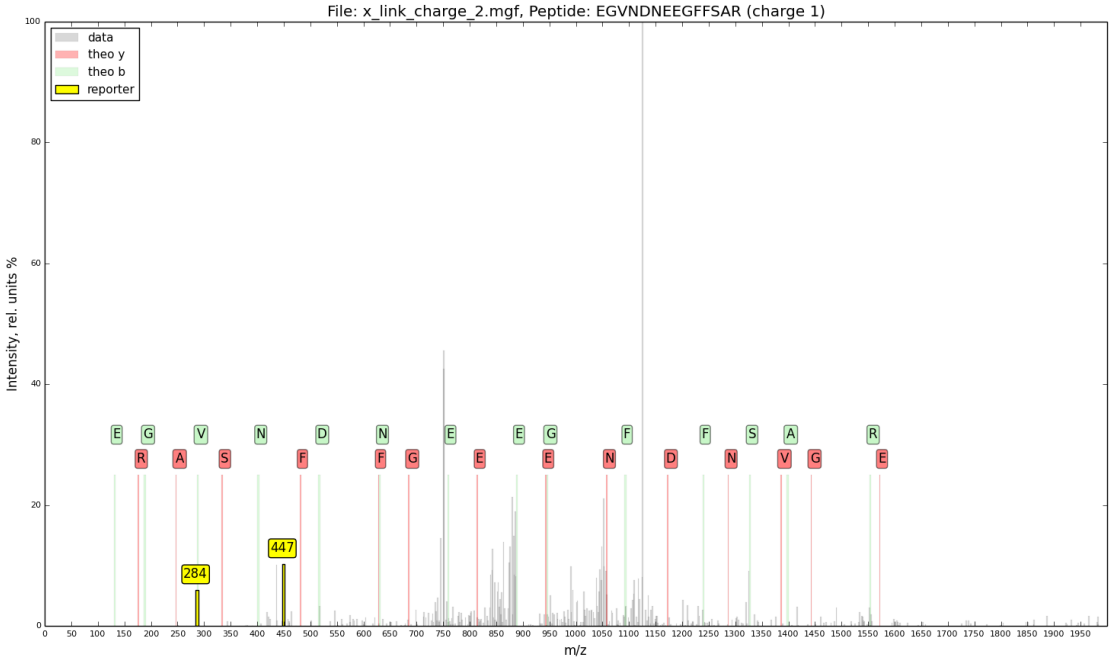


Abbildung 17: X-link GluFib mit dem eingezeichneten, unmodifizierten theoretischen Spektrum aus b- und y-Ionen der Peptidsequenz 'EGVNDNEEGFSAR' bei +1.

Bei steigender Ladung z verschiebt sich das Masse-Ladung-Verhältnis m/z des theoretischen Spektrum wieder nach links aufgrund des größeren Teilers z . Da die Messwerte am unteren linken Ende der m/z-Achse erst bei etwa 250 m/z beginnen, können nur Ladungszustände bis etwa $z = 4$ erkannt werden. Das ist der obere Grenzwert für die Ladung, um wenigstens die kleinste mögliche Modifikation von Glufib noch zu detektieren und ist das Maximum für weitere Berechnungen.

$$z = \frac{m}{\frac{m}{z}} = \frac{\text{Masse Glufib: } \approx 1570 + \text{kleinste Modifikation (cc2): } \approx 669 = 2239}{\text{kleinster messbarer Wert: } \approx 250} = \lfloor 4 \rfloor$$

Sequence Tags Nachdem das theoretische Ionenserien ermittelt ist, folgt bei *ilvamo* der Spektren-Vergleich via der Funktion *compare2theo()*. Stimmt ein Peak in beiden Listen innerhalb einer Fehlertoleranz überein, markiert das Tool den Match und speichert die zugehörige Aminosäure, den exakten m/z-Wert und die Intensität des Peaks mit der Funktion *identify()*. Mehrere Matches bilden einen Sequence Tag und stellen die Grundlage für den späteren Score dar. Je länger und zusammenhängender ein Sequenz Tag, desto wahrscheinlicher, dass es sich um das angegeben Peptid handelt. Im Beispiel-Spektrum von Glufib ist natürlich bekannt, welches Peptid vorliegt. Die Modifizierten Serien sind dennoch neu, sodass mit diesem Minimalbeispiels das Tool erprobt wird. Die Ladung der Modifikationen ist nicht konstant, es muss in allen möglichen Ladungszuständen nach Massenshifts gesucht werden. Je nach Ladung sind andere Sequence Tags zu erwarten.

Der Grenzwert für den Fehler kann vom Anwender wählt werden und ist default auf $\pm 0.5 m/z$ gesetzt, skaliert mit der Ladung. Je kleiner, desto signifikanter die Ergebnisse. Zur Überprüfung der Sensitivität wird ein zufälliges Spektrum mit ähnlicher Masse wie Glufib annotiert. Die Ergebnisse sind im Anhang auf Seite 45. Aufgrund der variablen Modifikationen entsteht starkes Rauschen und viele Artefakte in den x-Link Spektren, sodass die sehr viel geringere Fehlertoleranz von unmodifizierten Spektren bei wenigen parts per million (ppm) keine Matches liefert. Darüber muss sich der Anwnder bewusst sein. Hier zunächst die Ergebnisse der Glufib-Annotation.

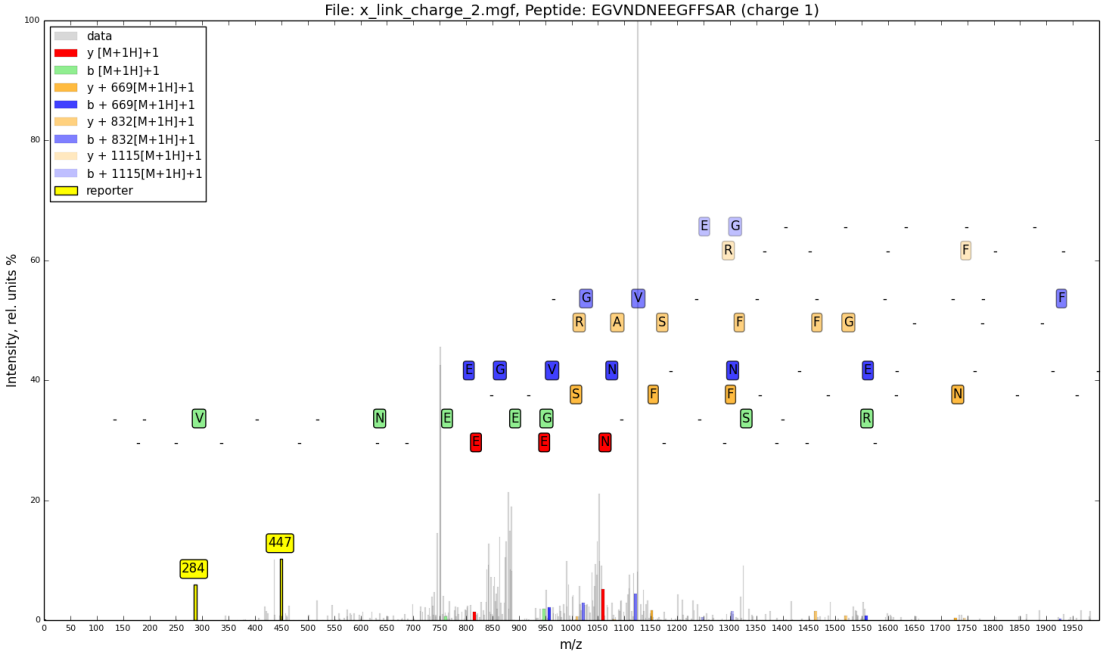


Abbildung 18: x-link GluFib bei $z = 1$. Gleiche Modifikation liegen direkt übereinander; rot (nativ)/orange (modifziert) für y- und grün (nativ)/ blau(modifiziert) für b-Ionen.

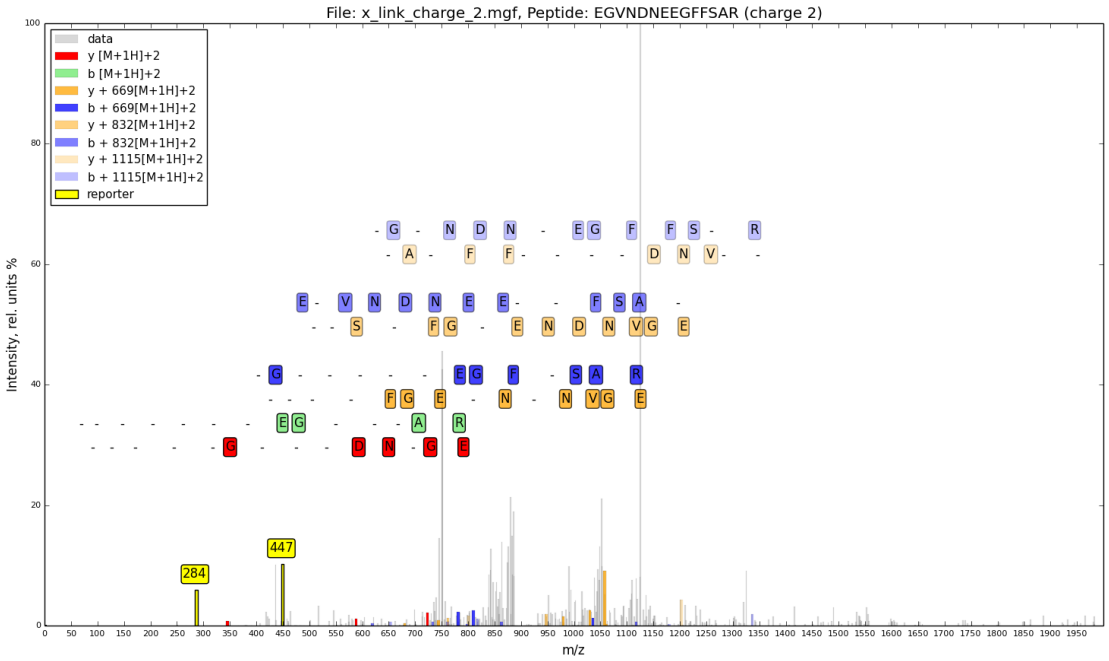


Abbildung 19: x-link GluFib Spektrum bei $z = 2$

Was beim Betrachten der annotierten x-link Spektren sofort auffällt ist, dass die markierten Peaks oft sehr schwache Intensitäten haben. Die Reporter Ionen sind stärker als die meisten modifizierten Peaks, ein grundsätzliches Problem in X-Link Spektren. Beim Filtern wären diese schwachen Peaks wahrscheinlich eliminiert worden. Als nächstes fällt auf, dass die Sequence Tags einen gewissen Schwerpunkt besitzen. Während bei $z = 1$ die nativen Serien mittig in der Sequenz Hits erzielen, befinden sich die modifizierten Serien eher hinten und vorne. Diese Disjunktion entspricht dem Abbrechen der verschobenen Serien durch den Verlust der modifizierten Aminosäure, wie bei Tabelle 2 auf Seite 15 vorhergesagt. Bindet das CC in der Mitte des Peptids, wären bei modifizierten b-Serien eher vorne Hits und bei y-Serien tendenzielle hinten.

Beim Spektrum von $z = 2$ allerdings ist das Gegenteil der Fall. Viele Sequence Tags überlappen sich hier und teilweise ergeben sich fast durchgängig modifizierte Serien. Das steht im Gegensatz zur Theorie der Disjunktion und dem Abbruch der x-link Serien. Eine Erklärung dafür könnte sein, dass mehrere Bindungspositionen existieren und die parallel im MS/MS Spektrum auftreten. Da Glufib aufgrund seiner relativ kurzen Peptidkette keine sterische Faltung besitzt, könnte das B1SAH6 durchaus an mehreren Stellen reagieren. Es scheint widersprüchliche Hinweise auf die Bindungsstelle des Capture Compound zu geben. Im Abschnitt 4 Lokalisierung auf Seite 28 wird auf diese Thematik weiter eingegangen und die jeweiligen Zusammenhänge diskutiert.

Output Die die oben abgebildeten Plots werden von *identify_plot()* erzeugt und ruft direkt *identify()* auf. Die Funktion *identify()* selbst liefert drei Parameter im Array-Format zurück: Die gefunden Sequenz mit Gaps bei nicht-Match, die Intensitäten der Matches und deren exakter m/z-Wert im experimentellen Spektrum. Standardmäßig werden b- und y- Ionenserien verwendet, insgesamt entstehen also sechs Rückgabewerte je Modifikation, die als Variablen *b_seq,b_mz,b_amp* bzw. *y_seq,y_mz,y_amp* zurückgegeben werden. Die Ergebnisse können zusammengefasst als Print mit einem eingebauten Kommando '-o' ausgegeben wir, ein Beispiel-Output von *identify('-o')* befindet sich im Anhang auf Seite 45. Die Amplituden sind als prozentuale Intensitäten angegeben und immer relativ zum maximalen Peak der Messung. Zum optischen Vergleich an dieser Stelle eine kürzere Version nur mit den Hits in der Peptid Sequenzen 'EGVNDNEEGFFSAR' von Glufib.

Listing 1: Output *identify()* für x-link Glufib bei Ladung $z = 1 - 4$

```
file: glufib_xlink.mgf

        charge: 1
unmodified
y: ['-', 'G', '-', '-', '-', 'N', 'E', 'E', 'G', '-', 'F', 'S', '-', '-']
b: ['-', '-', 'V', 'N', 'D', '-', 'E', 'E', 'G', 'F', 'F', 'S', 'A', '-']
cc fragment:669
y: ['-', '-', '-', '-', '-', '-', '-', '-', '-', '-', 'F', 'S', 'A', '-']
b: ['E', 'G', 'V', 'N', '-', 'N', 'E', 'E', '-', '-', '-', '-', '-', '-']
cc fragment:832
y: ['-', '-', '-', '-', '-', '-', '-', 'E', '-', '-', '-', 'S', 'A', 'R']
b: ['E', 'G', 'V', '-', 'D', 'N', '-', 'E', '-', 'F', '-', '-', '-', '-']
cc fragment:1115
y: ['-', '-', '-', '-', '-', '-', '-', 'E', 'G', 'F', 'F', 'S', '-', '-']
b: ['-', '-', '-', '-', '-', '-', '-', '-', '-', '-', '-', '-', '-', '-']

        charge: 2
unmodified
y: ['E', 'G', 'V', 'N', '-', 'N', 'E', '-', 'G', 'F', '-', '-', '-', '-']
b: ['-', '-', '-', '-', '-', 'N', 'E', 'E', 'G', '-', 'F', 'S', 'A', 'R']
cc fragment:669
y: ['-', 'G', 'V', 'N', 'D', '-', 'E', 'E', 'G', '-', '-', 'S', 'A', 'R']
b: ['E', 'G', '-', 'N', 'D', 'N', 'E', 'E', 'G', '-', 'F', 'S', 'A', 'R']
cc fragment:832
y: ['E', 'G', 'V', '-', 'D', '-', '-', 'E', 'G', 'F', 'F', 'S', 'A', 'R']
b: ['E', 'G', 'V', '-', '-', 'N', 'E', 'E', 'G', 'F', '-', '-', 'A', '-']
cc fragment:1115
y: ['-', 'G', 'V', 'N', 'D', 'N', 'E', 'E', 'G', 'F', 'F', 'S', '-', 'R']
b: ['E', 'G', 'V', 'N', 'D', '-', 'E', '-', '-', 'F', 'F', 'S', '-', 'R']
```

```
        charge:3
unmodified
y: ['-', 'G', 'V', '-', '-', '-', 'E', 'E', '-', '-', '-', '-', '-', '-']
b: ['-', '-', '-', '-', '-', '-', 'E', '-', 'G', '-', '-', 'S', '-', 'R']
cc fragment:669
y: ['E', 'G', 'V', 'N', 'D', 'N', 'E', 'E', 'G', '-', 'F', '-', '-', '-']
b: ['-', 'G', '-', '-', '-', 'N', '-', 'E', 'G', 'F', 'F', 'S', 'A', 'R']
cc fragment:832
y: ['-', 'G', 'V', '-', 'D', 'N', 'E', 'E', '-', '-', '-', 'S', 'A', 'R']
b: ['-', 'G', '-', 'N', '-', 'N', 'E', 'E', 'G', 'F', 'F', '-', 'A', 'R']
cc fragment:1115
y: ['-', 'G', 'V', 'N', 'D', 'N', 'E', 'E', 'G', 'F', 'F', 'S', 'A', '-']]
b: ['E', 'G', 'V', '-', 'D', 'N', 'E', 'E', 'G', 'F', '-', 'S', '-', 'R']

        charge:4
unmodified
y: ['E', 'G', '-', '-', '-', '-', '-', '-', '-', '-', '-', '-', '-', '-']
b: ['-', '-', '-', '-', '-', '-', '-', '-', 'G', 'F', '-', '-', '-', 'R']
cc fragment:669
y: ['-', 'G', '-', '-', 'D', 'N', '-', 'E', '-', 'F', 'F', 'S', '-', '-']
b: ['-', '-', '-', 'N', 'D', 'N', '-', 'E', '-', 'F', '-', 'S', '-', '-']
cc fragment:832
y: ['-', 'G', 'V', 'N', 'D', 'N', 'E', 'E', 'G', '-', '-', '-', '-', 'R']
b: ['-', '-', 'V', 'N', '-', '-', '-', 'E', 'G', 'F', 'F', 'S', 'A', '-']
cc fragment:1115
y: ['E', 'G', 'V', '-', 'D', 'N', 'E', '-', 'G', 'F', 'F', '-', 'A', 'R']
b: ['-', '-', 'V', 'N', '-', 'N', 'E', 'E', 'G', 'F', 'F', 'S', 'A', '-']
```

Auffällig ist, dass um die Mitte der Sequenz bei den Aminosäuren '-EE-' eine Häufung der Hits entsteht, viele der annotierten Sequence Tags überschneiden sich dort. Das könnte an der grundsätzlichen Normalverteilung der Messung liegen oder aber auf eine erhöhte Wahrscheinlichkeit für diese Bindungsstelle hinweisen. Im Abschnitt 4 Lokalisierung auf Seite 28 findet eine detaillierte Erläuterung statt. Unabhängig davon variiert deutlich die Länge der modifizierten Sequence Tags abhängig von der Ladung; je höher z desto mehr modifizierten Hits und desto weniger native. Das steht sicherlich im Zusammenhang mit der Reichweite der m/z-Achse, denn ab einer Ladung von $z = 4$ sind nativen Ionen kleiner als die untere Grenze (250 m/z) und vor einer Ladung von $z = 2$ ist die schwerste Modifizierung cc3 größer als die obere Grenze ($2000 m/z$).

Allgemein wirken die Ergebnisse auf den ersten Blick und ohne die dazugehörigen Amplituden zu gut, um realistisch zu sein. Beim späteren Scoring auf Seite 26 wird aber schnell deutlich, dass es große Unterschiede in der Qualität der Matches gibt. Je nach zugelassenen Fehler kann schon vorher ein stärkere Selektion vom Anwender erfolgen.

3 Verifizierung

Im Falle von Glufib ist die Verifizierung des Peptids redundant, weil die Sequenz vorher bekannt ist. Normalerweise wäre aber eine Liste von Peptiden, die zur Precursor-Masse passen, gegeben, aus der die am Besten passende Sequenz ermittelt werden soll. Jedem geeignetem Peptid wird anhand der übereinstimmenden Peaks eine Bewertung, Score genannt, zugeteilt. Neben Faktoren wie der Sequenzlänge und Anzahl der Matches beeinflussen den Score ebenso die Intensitäten der Hits, die Gewichtung der konsekutiven Matches und Non-Matches oder die Anwesenheit von bestimmten Ionen. Beim Scoring gibt es diverse Ansätze und Kriterien mit jeweils unterschiedlicher Sensitivität [24]. Jeder davon soll die Qualität der Peptid-Matches bewerten.

Um eine umfassende Bewertung der ambivalenten x-link Spektren zu erhalten, werden zwei Scores kombiniert verwendet. Der erste Score, genannt Peptide-Score, soll die Intensitäten eines Matches validieren. Ein zweiter Score, der xcorr, korreliert die m/z-Wert von Hit und synthetischen Peak miteinander, um die Abweichungen zu validieren. Diese Aufgabe wird bei *ilvamo* von den beiden Scoring-Funktionen *peptide_score()* und *xcorr()* übernommen.

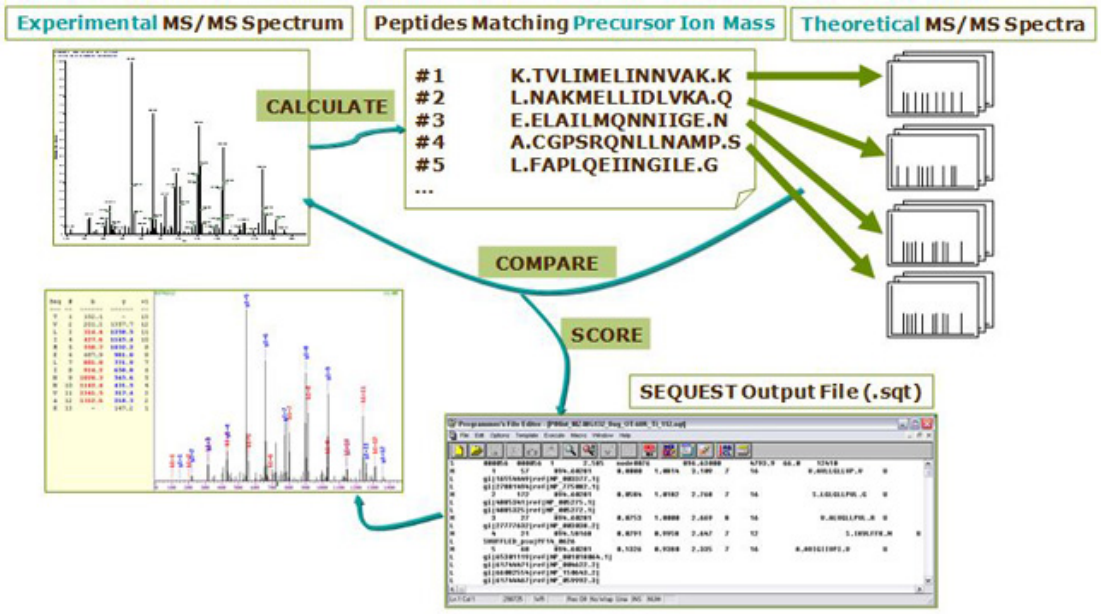


Abbildung 20: Schema einer Datenbank gestützten Analyse inklusive Scoring [25]

3.1 Summe der Intensitäten

Der Peptid-Score beschäftigt sich mit den Intensitäten der annotierten Peaks, der Anzahl an konsekutiven Hits und der Gesamtlänge der Serie. Dieser Score ähnelt dem Preliminary Score von SEQUEST Algorithmus der ersten Generationen [13], belohnt allerdings nicht wie dort die Anwesenheit von Ammonium-Ionen. Alle gefundenen Hits, native oder modifiziert, werden hier aufaddiert und über die Länge der Peptidsequenz normalisiert.

$$S_p = \frac{\left(\sum_{j=1}^{m} I_j\right) \cdot n_m(1 + n_c)}{l}$$

S_p = Peptid Score I_j = Intensität Match j n_m = Anzahl Matches
n_c = Anzahl konsekutive Matches l = Länge Peptide

Dieser Score beschränkt sich im Wesentlichen auf Intensitäten der Matches. Probleme treten auf, sobald die Peaks sehr kleine prozentuale Amplituden besitzen, wie es bei x-link Spektren häufig der Fall ist. Um dem gegenzusteuern, werden die drei Modifikationen cc1,cc2 und cc3 sowie die nativen Ionenserien zusammen aufsummiert. Das erhöht die Sensitivität des Peptids Scores, weil nur die Peptide, die über alle acht Ionenserien hinweg gute Hits liefern, einen hohen Wert erhalten. Durch die bedingten Ereignisse sinkt die Wahrscheinlichkeit für zufällige passende Matches.

Aufgerufen wird der erste Score durch die Funktion *peptide_score()*. Durch die Normalisierung über die Länge des Peptids sind ebenso verschieden große Sequenzen vergleichbar. Wenn nach höheren Ladungszuständen gesucht wird, summiert die Funktion alle Serien einer Ladung auf und dividiert sofort nach Ende der Schleife. Dadurch kann später nachvollzogen werden, bei welcher Ladung der Score am Stärksten variiert. Diese Punkte bilden den Fokus späterer Analysen. Hier ein Beispiel-Output von *peptide_score()*:

Listing 2: Peptide-Score von x-link Glufib

```
file: glufib_xlink.mgf
        charge:1
peptide: EGVNDNEEGFFSAR -> score: 109
best score: [['EGVNDNEEGFFSAR'], [109.35017151434189]]
        charge:2
peptide: EGVNDNEEGFFSAR -> score: 479
best score: [['EGVNDNEEGFFSAR'], [479.4414601447528]]
        charge:3
peptide: EGVNDNEEGFFSAR -> score: 166
best score: [['EGVNDNEEGFFSAR'], [166.88940865905778]]
        charge:4
```

```
peptide: EGVNDNEEGFFSAR -> score: 7
best score: [['EGVNDNEEGFFSAR'], [7.562192662900005]]
```

3.2 Korrelation der m/z-Werte

Der zweite Score ist der Korrelationskoeffizient nach Pearson für die exakten Positionen der Matches im Spektrum. Zur Berechnung wird die Methode aus dem numpy-Package verwendet. Der xcorr bewegt sich inklusiv zwischen -1 und +1 und liefert eine Korrelationsmatrix zurück, von dem nur der Wert XY.

$$P_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} * C_{jj}}}$$

P_{ij} = Pearson Koeffizient C_{ij} = Kovarianz ij C_{jj}/C_{ii} = Auto-Kovarianz jj/ii

Die synthetischen Spektren für den Vergleich bei *ilvamo* werden extra berechnet, nicht wie normalerweise aus einer Datenbank extrahiert. Somit ist die theoretische Peak Liste sehr viel kürzer (≈ 20) als die experimentelle Peak Liste aus dem x-link Spektrum (≈ 1200). Für die Korrelationsberechnung müssen beide Spektren gleich mächtig sein. Um das zu erreichen, werden die Ergebnisse von *identify()* verwendet und mit den theoretischen Peaks korreliert. Je weniger Gaps in der Sequenz und je geringer die Abweichung zum synthetischen Fragment, desto größer der Korrelationskoeffizient der Peak Serie.

Dieser zweite Score wird von der Funktion *xcorr()* berechnet und ist ähnlich wie der Peptid-Score nach Ladungszuständen unterteilt. Es kann differenziert dargestellt werden, wo welche modifizierte Serie besonders gute Matches erzeugt, das fokussiert die spätere Analyse auf bestimmte Ionenserien und Ladungszustände. Beispielsweise wird beim x-link Glufib schnell klar, dass bei $z = 2$ deutlich bessere Übereinstimmungen zwischen theoretischen und experimentellen Peaks besteht, denn im Gegensatz zu $z = 1$ sind die Korrelationskoeffizienten sehr viel größer. Diese Information ist grundsätzlich so auch in dem Output von *peptide_score()* oben zu erkennen; hier ist der Score für $z = 2$ mehr als vier mal so mächtig wie der für $z = 1$.

Listing 3: Xcorr für x-link Glufib

```
file: glufib_xlink.mgf
      charge: 1
mass: 0
seq: EGVNDNEEGFFSAR ==> xcorr y: 0.088 , xcorr b: 0.511
mass: 669
seq: EGVNDNEEGFFSAR ==> xcorr y: -0.263 , xcorr b: -0.595
mass: 832
seq: EGVNDNEEGFFSAR ==> xcorr y: -0.798 , xcorr b: -0.183
mass: 1115
seq: EGVNDNEEGFFSAR ==> xcorr y: -0.395 , xcorr b: -0.588
```

```
        charge: 2
mass: 0
seq:  EGVNDNEEGFFSAR ==> xcorr y: 0.692 , xcorr b: 0.61
mass: 669
seq:  EGVNDNEEGFFSAR ==> xcorr y: 0.747 , xcorr b: 0.725
mass: 832
seq:  EGVNDNEEGFFSAR ==> xcorr y: 0.841 , xcorr b: 0.187
mass: 1115
seq:  EGVNDNEEGFFSAR ==> xcorr y: 0.165 , xcorr b: 0.501

        charge: 3
mass: 0
seq:  EGVNDNEEGFFSAR ==> xcorr y: 0.452 , xcorr b: nan
mass: 669
seq:  EGVNDNEEGFFSAR ==> xcorr y: 0.114 , xcorr b: 0.295
mass: 832
seq:  EGVNDNEEGFFSAR ==> xcorr y: 0.52 , xcorr b: 0.432
mass: 1115
seq:  EGVNDNEEGFFSAR ==> xcorr y: 0.441 , xcorr b: 0.453

        charge: 4
mass: 0
seq:  EGVNDNEEGFFSAR ==> xcorr y: nan , xcorr b: nan
mass: 669
seq:  EGVNDNEEGFFSAR ==> xcorr y: 0.279 , xcorr b: 0.047
mass: 832
seq:  EGVNDNEEGFFSAR ==> xcorr y: 0.556 , xcorr b: 0.643
mass: 1115
seq:  EGVNDNEEGFFSAR ==> xcorr y: 0.32 , xcorr b: 0.373
```

4 Lokalisierung

Das zweite Problem bei der Analyse von x-Link Spektren ist die Lokalisation der Bindungsstelle. Ist die Peptide Seqeuenz erstmal bekannt, kann theoretisch anhand der Ionenserien die Position das CC ermittelt werden. Das liegt an dem Aufbau des Komposit-Spektrums. Hier sind modifizierte wie native Ionen enthalten und im Vergleich werden beide theoretischen Serien Hits erzielen. Die Frage ist nur, wie viele und wie lang die gefunden Sequence Tags daraus werden.

Wie in Tabelle 2 auf Seite 15 zu sehen ist, sind nativen und modifizierten Ionenserie disjunkt zueinander. Bis zur Bindungsstelle p_m sind die Ionen also verschoben und sobald die modifizierte AA Residue abgespalten ist, rutschen die Massen der Fragmente wieder auf ihre nativen m/z-Werte zurück. Dieser Sprung in den Massen markiert die Bindungsposition. Bei einer Annotation würden die Übereinstimmungen mit der modifizierten Ionenserie ab einer bestimmten Stelle aufhören und die zur nativen Serie parallel beginnen sowie vice versa. Anhand dieses Bruches in den Sequence Tags kann die Bindungsposition angegeben werden und dessen Detektierung ist Ziel dieses Abschnittes.

Sind die Massen der variablen Modifikation und die Peptid Sequenz ermittelt, kann die Lokalisation starten. Letztere beruht demnach auf den Ergebnissen der vorangehenden Identifikation. Um die fehlende Information über die Position des CC auszugleichen, wurde dort jeweils ein theoretisches Spektrum mit und ohne Modifikation berechnet. Damit ist sichergestellt, dass solange B1SAH6 gebunden ist, die synthetischen Serien an irgendeiner Stelle matchen und sich modifizierte und native gegenseitig ergänzen. Diesen Umstand wird weiter bei der Lokalisierung genutzt.

Die Grundidee ist, die genaue Verteilung der Hit Peaks aus der Identifizierung zu analysieren. Natürlich wird nicht in jeder Annotation der Bruch der Sequence Tags sauber zu sehen sein. Zufall oder Fehler verschieben die Ionenserien, außerdem kann besonders im Falle von Glufib von mehreren Bindungsstellen ausgegangen werden. Wenn jedoch die Ergebnisse mehrerer Durchgänge, z.B. bei verschiedenen Ladungen oder mehrere Datensätzen, in einer Superposition abgetragen sind, kann die Position des Massenshifts immer weiter eingrenzt werden. Durch diese Überlagerung der Daten'wellen' können Meta-Informationen über die Verteilung der Matches gesammelt werden.

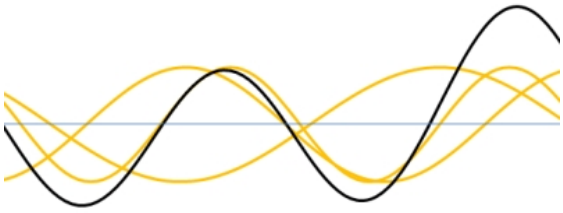


Abbildung 21: Superposition in der Wellentheorie [26]

4.1 Verteilung der Matches

Wenn genügend Informationen generiert werden, sollten sich Schwerpunkte oder zumindest Cluster von Hit Peaks bilden. Um diese Abzubilden wird ein Plot erstellt, der auf seiner X-Achse die Postionen der Peptidsequenz enthält und auf der y-Achse die prozentualen Amplituden der Matches abträgt. Dieser Sequence-Plot enthält also für jede Position p_i im Peptid $p_1, p_2, ..., p_n$ die zugehörigen Peaks aus dem experimentellen Spektrum. Jeder gefunden Sequence Tag wird eingezeichnet und je mehr Daten hinzukommen, desto kontinuierlicher wird das erzeugte Meta-Spektrum.

Über mehrere Ladungen aufsummiert ergibt sich somit ein Superposition der Hit Peaks. In der Theorie ergibt sich hierbei ein Zug aus Peaks, jeweils einer für modifizierte und einer für unmodifizierte Matches. Das entstehende Plateau hält an, bis der Modifikationspunkt p_m im Peptid erreicht ist. Dort bricht der Peak-Train der modifizierten Matches ab und die Intensitäten fallen auf null. Gleichzeitig bildet sich ein neues Plateau aus für die unmodifizierten Matches. Der Schnittpunkt der beiden Graphen wäre dann die Bindungsstelle der Modifikation. Falls es davon mehrere gibt, wären das im Plot ebenfalls durch eine Wechsel der Peak-Trains zu sehen. Nach diesem Ansatz kann die Bindungsstelle des Capture Compounds bestimmt werden.

In der Praxis sind die Intensitäten der x-link Matches allerdings so gering, dass sich kein Plateau ausbildet. Das Ergebnis der Superposition ist eher ein Sägezahnblatt. Das passiert weil die Hit Peaks eine extrem große Varianz in ihren Amplituden besitzen, sodass die Meta-Daten sehr unregelmäßig und verrauscht sind. Um dem entgegen zu wirken, sollte die Differenz zwischen modifizierten und unmodifizierten Matches gebildet werden. Das reicht jedoch noch nicht aus, denn auch hier sorgt die große Varianz der Peaks für eine ambivalente Ergebniskurve. Das eigentliche Signal des Massenshifts geht einfach unter im Rauschen.

Die Lösung besteht darin, die gebildeten Differenzen nicht einzeln zu betrachten, sonder aufzusummieren zu einer Gesamtdifferenz und dann den Verlauf dieser Kurve zu analysieren. Natürlich wird die kummulierte Differenz langfristig nur steigen. Interessant ist aber nicht der Endwert der Summe, sonder der Verlauf dahin. Der Umschlagspunkt vom Massenshift sollte eine sprunghafte Veränderung erzeugen, entweder nach unten oder nach oben, je nach dem ob modifizierte oder native Serie an der jeweiligen Peptidposition überwiegt. Hier ein Beispiel für die Lokalisation im x-Link Glufib.

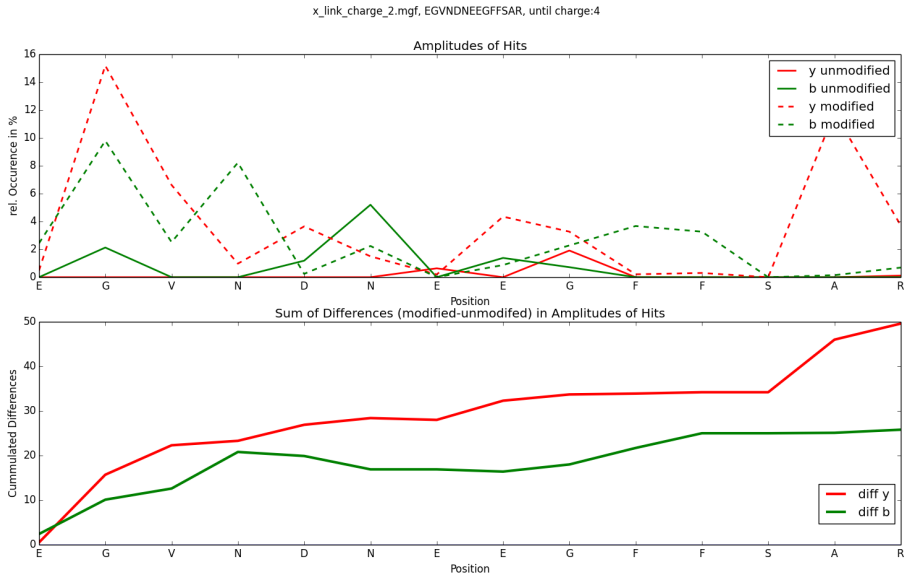


Abbildung 22: Sequence Plot der Intensitäten von annotierten Peaks aus dem x-link Glufib Spektrum bis zur Ladung $z = 4$ aufsummiert.

Im Sequence Plot oben ist die Verteilung der Matches zwischen modifizierten und unmodifizierten Tags dargestellt. Unten ist der Verlauf der summierten Differenzen zu sehen. Dabei fallen zwei markante Knicke auf: Bei der ersten N (p_{m1}) und der hinteren S Residue (p_{m3}) ist jeweils ein Sprung zu sehen. Eine weitere kleine Erhebung ist zwischen den beiden Doppel E Aminosäuren (p_{m2}) in der Mitte zu erkennen. Alle drei sind Kandidaten für eine Modifikations-Bindungsstelle. Aufgrund der Richtung der Differenz heißt ein Absinken der Kurve eine Stelle mit vielen unmodifizierten Peaks und ein Anstieg nach oben eine Position mit deutlich mehr modifizierten Matches. Je nach Steilheit der Änderung kann auf die relative Häufigkeit und Wahrscheinlichkeit einer Bindungsstelle geschlossen werden. Demnach sind die Stellen p_{m1} und p_{m3} deutlich öfter modifiziert als die mittlere Position p_{m2}. Da es bisher keine Bewertungskriterien gibt, kann p_{m2} durchaus eine zufällige Erscheinung sein.

4.2 Implementierung

Im *ilvamo*-Tool wird für die Lokalisation zunächst die Funktion *score_distribution()* aufgerufen. Sie verwendet die Ergebnisse von *identify()* und überträgt die Amplituden der Matches in eine Matrix. Jede Spalten steht für eine Position im Peptid, jede Zeile entspricht einer Ionenserie; in der ersten Zeile die y Ionen und in der zweiten Zeile die b Ionen.

	E	G	V	N	D	N	E	E	G	F	F	S	A	R
Y	0	0.12	0	0	0	0.57	0.98	0.65	1.66	0	0.35	0.04	0	0
B	0	0	0.10	0.29	0.33	0	3.35	1.05	0.22	1.39	0.04	0.04	0.11	0

Tabelle 4: Amplituden Matrix mit prozentualen Intensitäten der Machts von x-link Glufib via *score_distribution()*

Danach folgt *sequence_score()*, das Herzstück der Lokalisierung. Via *score_distribution()* wird je eine Matrix für modifizierte ($matrix_{cc}$) und eine für unmodifizierte ($matrix_{zero}$) Matches erzeugt und diese über alle Ladungen aufsummiert. Die in $matrix_{cc}$ enthaltenen Intensitäten werden dabei durch die Anzahl der Modifikationen geteilt und normalisiert, sodass eine Art Durchschnitts-Modifikation gebildet wird. Das dient dazu, dass später Minuend und Subtrahend gleich mächtig sind. Sonst wäre die Differenz immer zugunsten der Modifikation verzerrt. Anschließend wird die Differenzmatrix $matrix_{diff}$ gebildet, welche die Diskrepanz zwischen modifizierten und nativen Matches enthält. Hier werden per Matrixsubtraktion alle unmodifizierten Hits von den modifizierten abgezogen.
Als Letztes ruft das Kommando '-p' den *sequence_plot()* auf, eine Darstellung der Ergebnisse inklusive der kummulierten Summe der Differenzen. Wie oben abgebildet enthält der Sequence Plot die exakten Verteilungen der Peaks oben sowie die aufsummierten Diskrepanzen unten. Damit ist eine Analyse möglicher Bindungsstellen der Modifikation möglich.

Die y-Achse im Sequence Plot ist aufgrund der hohen Varianz der Intensitäten nicht fest gesetzt, sondern skaliert mit den jeweiligen Amplituden mit. Beim Vergleich mehrere Plots sollte deshalb immer die jeweilige Skalierung beachtet werden. Bei späteren Ergebnissen wird klar, dass analog zum Peptidscore die Qualität der Matches daran abgelesen werden kann. Beim selben Cross-Link Spektrum erzeugen verschiedenen Peptide unterschiedlich Kurven und Differenzen. Wird jede Ladung einzeln betrachtet, kann in den Sequence Plots wie in einem Daumenkino geblättert werden. Es lässt sich beobachten, wann welche Peaks gefunden werden und wie sie den Verlauf der summierten Differenzen ändern.
Einige der eingetragenen Intensitäten im Sequence Plot können durch die Aufsummierung der Modifikationen als $matrix_{cc}$ stellenweise über 100% liegen, auch trotz Normalisierung. Dieser Effekt tritt allerdings selten auf.

5 Ergebnisse

Bisher wurde anhand des Standard-Peptids Glufib demonstriert, wie *ilvamo* Schritt für Schritt ein x-Link Spektrum annotiert, bewertet und die Bindungsstelle der Modifikation ermittelt. Um das Tool nun unter reellen Bedingungen zu testen, wird ein Cross-Link zwischen B1SAH6 und der Methyltransferase des Thermus Aquaticus, kurz M.Taql [22], analysiert. Die Sequenz von M.Taql besteht aus 421 Aminosäuren, besitzt ein Gewicht von 47 862 Da und trägt die Uniprot ID P14385.Nach dem Proteinverdau mit Trypsin bleiben von mtaq 10 Peptide erhalten. Gegeben sind außerdem vier Scans aus dem MS/MS, die von M.Taql stammen, und jeweils ein Peptid enthalten. Aufgabe ist, das korrekte Peptid dem Scan zuzuordnen und die Modifikationsstelle zu Lokalisieren. Die folgenden Ergebnisse beziehen sich auf diese Untersuchung.

Listing 4: AA Sequence M.Taql

```
          10         20         30         40         50         60
MGLPPLLSLP SNSAPRSLGR VETPPEVVDF MVSLAEAPRG GRVLEPACAH GPFLRAFREA

          70         80         90        100        110        120
HGTAYRFVGV EIDPKALDLP PWAEGILADF LLWEPGEAFD LILGNPPYGI VGEASKYPIH

         130        140        150        160        170        180
VFKAVKDLYK KAFSTWKGKY NLYGAFLEKA VRLLKPGGVL VFVVPATWLV LEDFALLREF

         190        200        210        220        230        240
LAREGKTSVY YLGEVFPQKK VSAVVIRFQK SGKGLSLWDT QESESGFTPI LWAEYPHWEG

         250        260        270        280        290        300
EIIRFETEET RKLEISGMPL GDLFHIRFAA RSPEFKKHPA VRKEPGPGLV PVLTGRNLKP

         310        320        330        340        350        360
GWVDYEKNHS GLWMPKERAK ELRDFYATPH LVVAHTKGTR VVAAWDERAY PWREEFHLLP

         370        380        390        400        410        420
KEGVRLDPSS LVQWLNSEAM QKHVRTLYRD FVPHLTLRML ERLPVRREYG FHTSPESARN


F
```

Listing 5: Peptide von M.Taql nach dem Verdau via Trypsin

```
peptide_sequence=['VETPPEVVDFMVSLAEAPR', 'LPDSSLVQWLNSEAMQK',
'LEISGMPLGDLFHIR', 'MGLPPLLSLPSNAAPR', 'DFYATPHLVVAHTK',
'TSVYYLGEVFPQK', 'VLEPACAHGPFLR', 'EYGFHTSPESAR', 'NLKPGWVDYEK',
'EPGPGLVPVLTGR']
```

Folgende vier MS/MS wurden aus den Pepmassen des MS von M.Taql erzeugt, jeweils mit eingezeichneten Reporter Ionen:

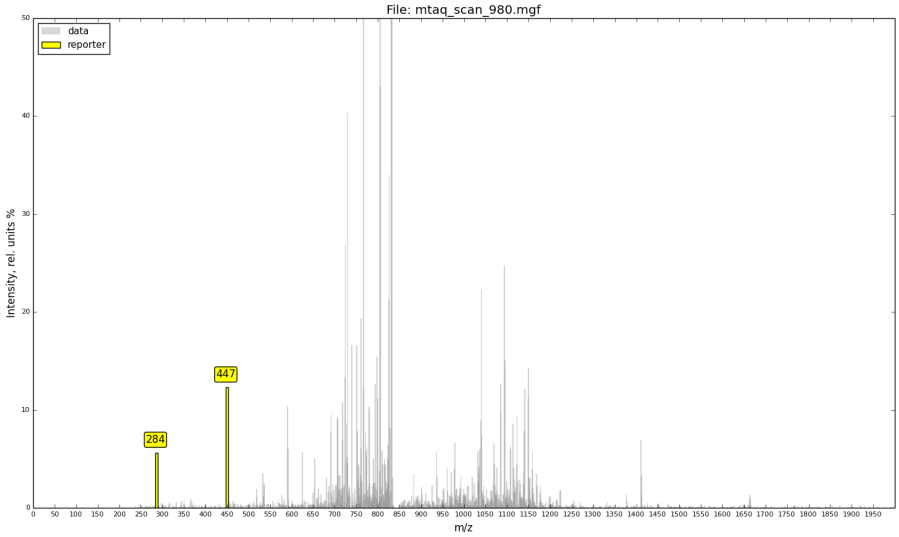


Abbildung 23: Spektrum von M.Taql Scan 980

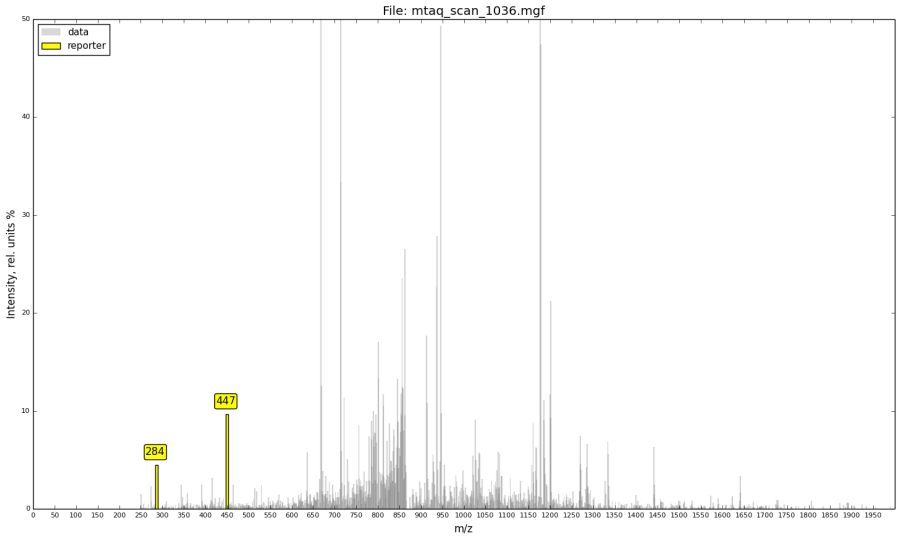


Abbildung 24: Spektrum von M.Taql Scan 1036

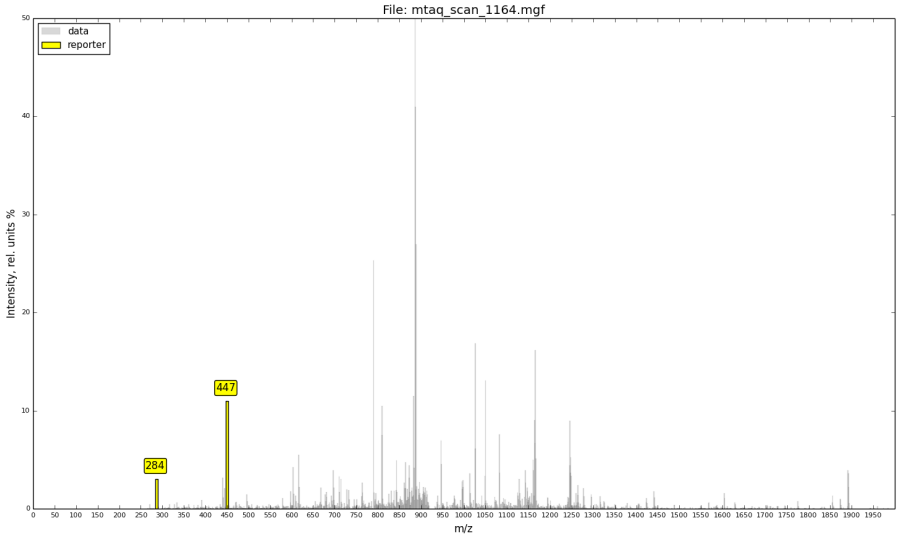


Abbildung 25: Spektrum von M.Taql Scan 1164

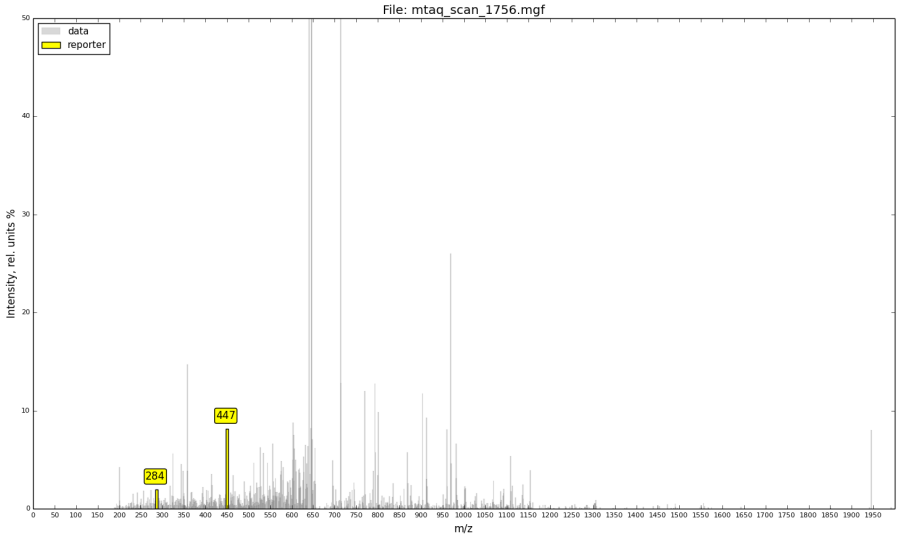


Abbildung 26: Spektrum von M.Taql Scan 1756

5.1 Peptiderkennung

Zunächst muss analysiert werden, welche Peptide aus der Liste zum jeweiligen Scan passen. Starten wird dabei die Analyse vom Scan 980 und die Funktion *petide_score()* mit $z = 1$. Normalerweise erhält der User nur eine Liste von Scores und den Besten Score als Rückgabe; bei diesem Output hier wird etwas detaillierter aufgelistet was passiert als normalerweise. Angegeben wird der jeweilige aktuelle aufsummierten Score zum Peptid und Anzahl zusammenhängender Hits über alle cc-Fragmente (cc1 = 669,cc2=832 & cc3=1115). Anschließend kommt die Peptid Liste mit dem endgültigem Score, der über die Länge des Peptids normalisiert ist und zusammenhängende Matches belohnt.

Listing 6: M.Taql 980 Output von *petide_score()*. '0' bedeutet native, unmodifizierte Serie

```
file   mtaq_scan_980.mgf

VETPPEVVDFMVSLAEAPR at 0: 139 and consecutive hits:11
VETPPEVVDFMVSLAEAPR at 669: 141 and consecutive hits:1
VETPPEVVDFMVSLAEAPR at 832: 181 and consecutive hits:8
VETPPEVVDFMVSLAEAPR at 1115: 181 and consecutive hits:0

LPDSSLVQWLNSEAMQK at 0: 384 and consecutive hits:12
LPDSSLVQWLNSEAMQK at 669: 446 and consecutive hits:7
LPDSSLVQWLNSEAMQK at 832: 468 and consecutive hits:4
LPDSSLVQWLNSEAMQK at 1115: 469 and consecutive hits:3

LEISGMPLGDLFHIR at 0: 255 and consecutive hits:12
LEISGMPLGDLFHIR at 669: 302 and consecutive hits:6
LEISGMPLGDLFHIR at 832: 307 and consecutive hits:1
LEISGMPLGDLFHIR at 1115: 307 and consecutive hits:1

MGLPPLLSLPSNAAPR at 0: 1397 and consecutive hits:15
MGLPPLLSLPSNAAPR at 669: 1567 and consecutive hits:8
MGLPPLLSLPSNAAPR at 832: 1586 and consecutive hits:4
MGLPPLLSLPSNAAPR at 1115: 1587 and consecutive hits:1

DFYATPHLVVAHTK at 0: 268 and consecutive hits:14
DFYATPHLVVAHTK at 669: 332 and consecutive hits:8
DFYATPHLVVAHTK at 832: 353 and consecutive hits:6
DFYATPHLVVAHTK at 1115: 353 and consecutive hits:3

TSVYYLGEVFPQK at 0: 487 and consecutive hits:9
TSVYYLGEVFPQK at 669: 546 and consecutive hits:5
TSVYYLGEVFPQK at 832: 558 and consecutive hits:2
TSVYYLGEVFPQK at 1115: 559 and consecutive hits:1
```

```
VLEPACAHGPFLR at 0: 1685 and consecutive hits:15
VLEPACAHGPFLR at 669: 1743 and consecutive hits:7
VLEPACAHGPFLR at 832: 1756 and consecutive hits:5
VLEPACAHGPFLR at 1115: 1757 and consecutive hits:1

EYGFHTSPESAR at 0: 248 and consecutive hits:11
EYGFHTSPESAR at 669: 286 and consecutive hits:7
EYGFHTSPESAR at 832: 296 and consecutive hits:5
EYGFHTSPESAR at 1115: 296 and consecutive hits:2

NLKPGWVDYEK at 0: 83 and consecutive hits:7
NLKPGWVDYEK at 669: 303 and consecutive hits:9
NLKPGWVDYEK at 832: 327 and consecutive hits:6
NLKPGWVDYEK at 1115: 328 and consecutive hits:1

EPGPGLVPVLTGR at 0: 233 and consecutive hits:11
EPGPGLVPVLTGR at 669: 465 and consecutive hits:11
EPGPGLVPVLTGR at 832: 482 and consecutive hits:4
EPGPGLVPVLTGR at 1115: 483 and consecutive hits:2

peptide: VETPPEVVDFMVSLAEAPR -> score: 181
peptide: LPDSSLVQWLNSEAMQK -> score: 469
peptide: LEISGMPLGDLFHIR -> score: 308
peptide: MGLPPLLSLPSNAAPR -> score: 1587
peptide: DFYATPHLVVAHTK -> score: 356
peptide: TSVYYLGEVFPQK -> score: 560
peptide: VLEPACAHGPFLR -> score: 1758
peptide: EYGFHTSPESAR -> score: 297
peptide: NLKPGWVDYEK -> score: 328
peptide: EPGPGLVPVLTGR -> score: 495

best score: ['VLEPACAHGPFLR', 1758.5142926678252]
```

Auffällig ist, dass die Hits der modifizierten Serien kaum den Score nach oben treiben. Die später höchsten Peptide sind bereits nach der nativen Serie gute benotet, ohne dass die modifizierten Serien viel daran ändern. Das könnte an den stärkeren Intensitäten der nativen Fragmente liegen oder daran, dass hier die Maximalpeaks enthalten sind. Der Intensitätsscore verzerrt sich dadurch zugunsten der Peptide mit weniger Matches, solange diese eine möglichst starke Amplitude haben.
Am Ende wird 'VLEPACAHGPFLR' wird beim Peptid Score als bester Match für den Scan 980 ausgegeben. Allerdings befindet sich 'MGLPPLLSLPSNAAPR' nur knapp dahinter. Um die Ergebnisse genauer zu differenzieren, wird der zweite Score mit *xcorr()* verwendet und als Referenz verwendet.

Listing 7: M.Taql 980 Output von *xcorr()*

```
file    mtaq_scan_980.mgf

VETPPEVVDFMVSLAEAPR xcorr y: 0.043 , xcorr b: 0.083
LPDSSLVQWLNSEAMQK xcorr y: 0.265 , xcorr b: 0.434
LEISGMPLGDLFHIR xcorr y: 0.669 , xcorr b: 0.784
MGLPPLLSLPSNAAPR xcorr y: 0.349 , xcorr b: 0.049
DFYATPHLVVAHTK xcorr y: 0.692 , xcorr b: 0.434
TSVYYLGEVFPQK xcorr y: 0.207 , xcorr b: -0.28
VLEPACAHGPFLR xcorr y: 0.617 , xcorr b: 0.876
EYGFHTSPESAR xcorr y: 0.929 , xcorr b: 0.571
NLKPGWVDYEK xcorr y: 0.855 , xcorr b: 0.336
EPGPGLVPVLTGR xcorr y: 0.608 , xcorr b: 0.707
```

Auch hier sind mehrere Peptide gleichauf: 'VLEPACAHGPFLR', 'LEISGMPLGDLFHIR' und 'EYGFHTSPESAR' mit jeweils hohen Korrelationskoeffizienten in y und b Serien. Weil 'VLEPACAHGPFLR' allerdings im Peptid Score ebenso gut abschneidet, bekommt dieses Peptid in der Gesamtbewertung die höchste Wahrscheinlichkeit zugeordnet. Der Scan 980 wird also als 'VLEPACAHGPFLR' identifiziert. Anschließend wird das Spektrum mit dem identifizierten Peptid via *identify_plot()* geplottet und die einzelnen Matches in b- und y-Serien mit *identify()* ausgegeben.

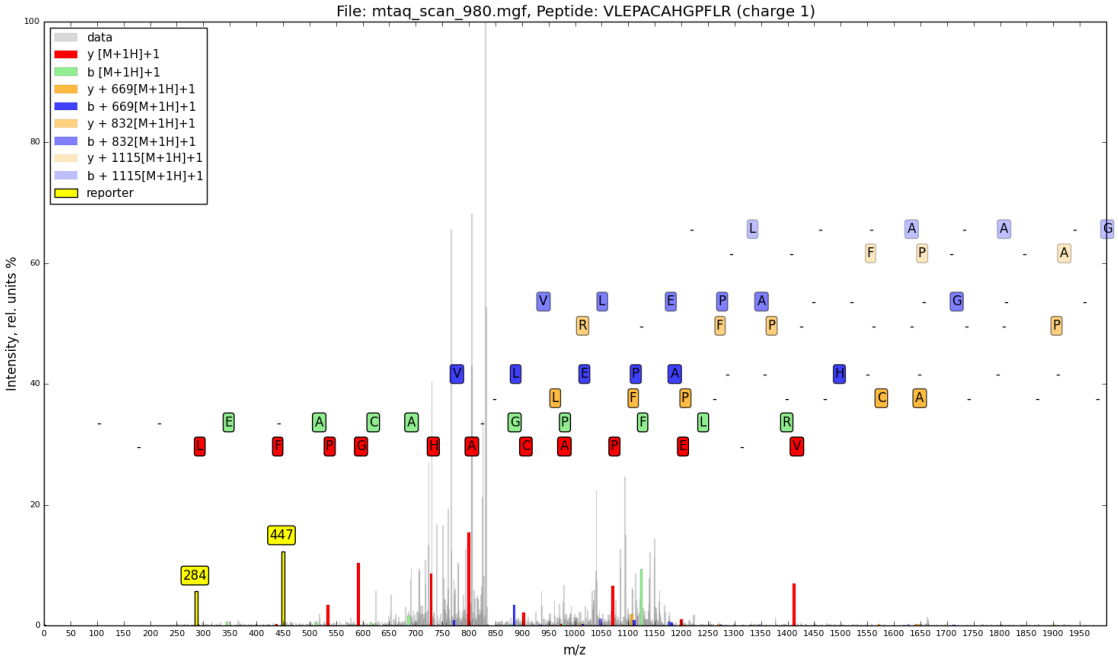


Abbildung 27: Annotiertes Spektra von mtaq 980 mit 'VLEPACAHGPFLR'

Bei den restlichen Scans sieht das Verfahren gleich aus. Erst wird das beste Peptid aus der Liste ermittelt, dann wird das Spektrum annotiert und der Plot erstellt. Hier die Ergebnisse zusammengefasst, die detaillierten Scores befinden sich im Anhang auf Seite 46. Die Entscheidung fürs beste Peptid wird jeweils unter den Top 3 des Peptide Scores getroffen in Verbindung mit dem xcorr.

Das Peptid 'MGLPPLLSLPSNAAPR' hat in allen Scans sehr gute Ergebnisse bei den Intensitäten und erzielt hohe Peptid Scores. Spätestens beim Korrelationskoeffizienten wird aber klar, dass dieses Peptid nur einmal in den Scans wirklich passt. Wieso trotzdem jedes Mal 'MGLPPLLSLPSNAAPR' so hoch scored, ist nicht ersichtlich. Möglicherweise deshalb, weil eines der Fragmente genau zu einem der Maximalpeaks passt. Wie oben angedeutet, würde das den gesamten Score stark verzerren.

Scan	Best Peptide	Peptide Score	Xcorr Score
980	VLEPACAHGPFLR	1759	y: 0.617 , b: 0.876
1036	MGLPPLLSLPSNAAPR	5241	y: 0.135 , b: 0.514
1164	NLKPGWVDYEK	424	y: 0.588 , b: 0.548
1765	EPGPGLVPVLTGR	519	y: 0.269 , b: 0.215

Tabelle 5: Ergebnisse Identifizierung M.Taql

5.2 Analyse zur Position

Nach dem die Scans von M.Taql identifiziert wurden, folgt nun die Lokalisierung der Bindungsstelle vom Capture Compound im Peptid. Dazu wird bei *ilvamo* die Funktion *sequence_score('-p')* verwendet. Prinzipielle enthält jeder unserer Beispiel-Scans eine Modifikation, denn alle besitzen Reporter Ionen. Tabelle 5 zeigt, dass die beste Identifizierung bei Scan 1036 von M.Taql erreicht wurde. Daher dient dieses Spektrum als Ausgangspunkt für eine Lokalisierung. Zur Veranschaulichung hier noch einmal der annotierte Plot vom Spektrum von Scan 1036 mit dem Peptid 'MGLPPLLSLPSNAAPR' bei $z = 1$.

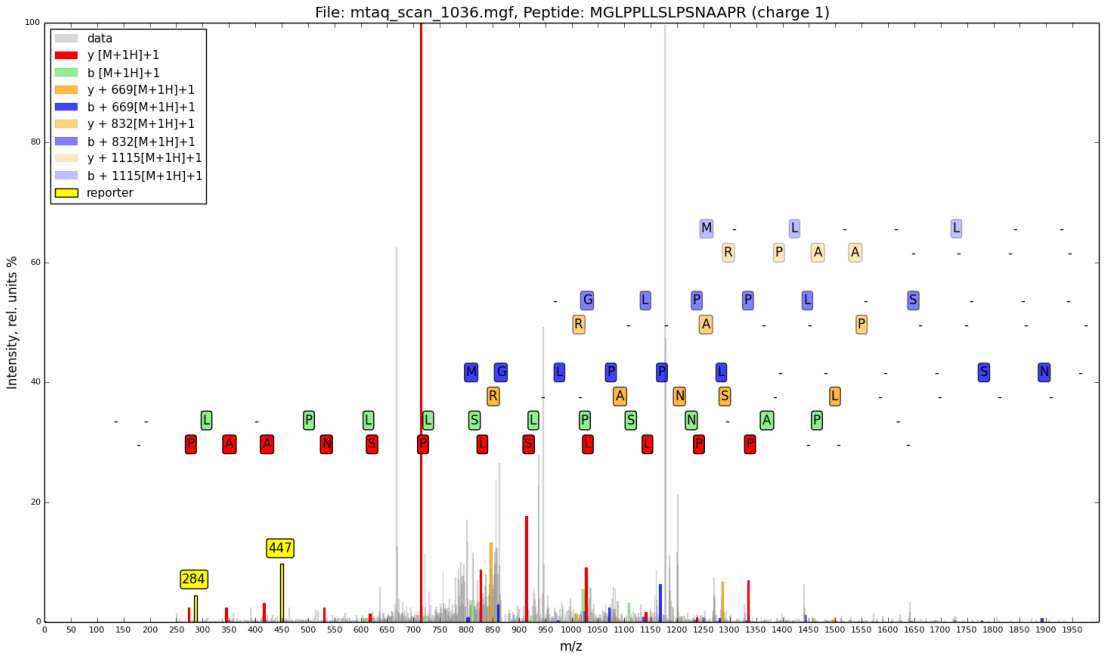


Abbildung 28: Annotiertes M.Taql Scan 1036 mit 'MGLPPLLSLPSNAAPR'

Die modifizierten Matches sind deutlich stärker in der Amplitude als die unmodifizierten. Demnach wird das Signal, dass den Übergang zwischen den beiden Phasen markiert, vermutlich sehr schwach sein. Deshalb werden alle vier Ladungszustände nacheinander überprüft und einzeln betrachtet. Ähnlich wie bei einem Daumenkino kann dann die Aufsummierung der kumulativen Differenz direkt beobachtet werden. Die relative Häufigkeit können dabei etwas größer als 100 % sein, dass liegt in der Aufsummierung der Peaks. Die Ergebnisse zeigen die außerdem, dass es mehr als einen Wechsel der Plateaus gibt. Grundsätzlich zeigt sich eine Häufung der Peaks bzw. ein Sprung in den Summen beim zweiten und dritten Prolin-Rest in der Sequenz. Das ist verwunderlich, denn Prolin ist durch seine ringförmige Struktur recht klein und bietet kaum Platz für eine Bindung. Während die b-Serien bei beide Prolinen einen Phasenwechsel zeigen, sind die y-Serien nur bei dem dritten Prolin signifikant modifiziert.

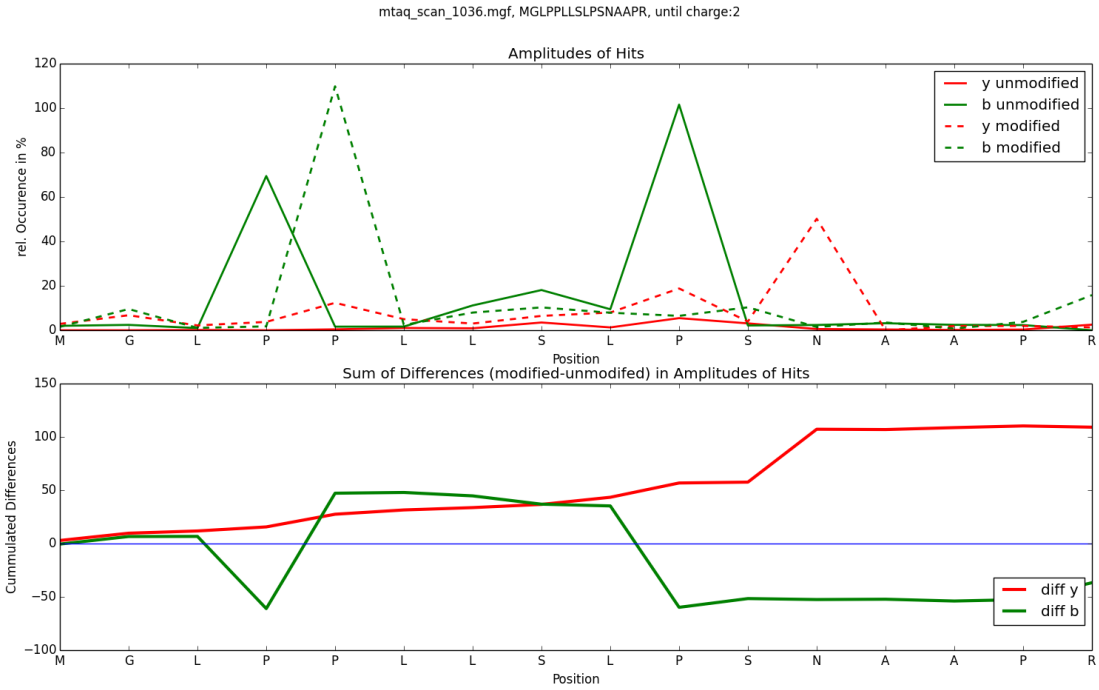


Abbildung 29: Sequence Plot M.Taql Scan 1036 mit 'MGLPPLLSLPSNAAPR'

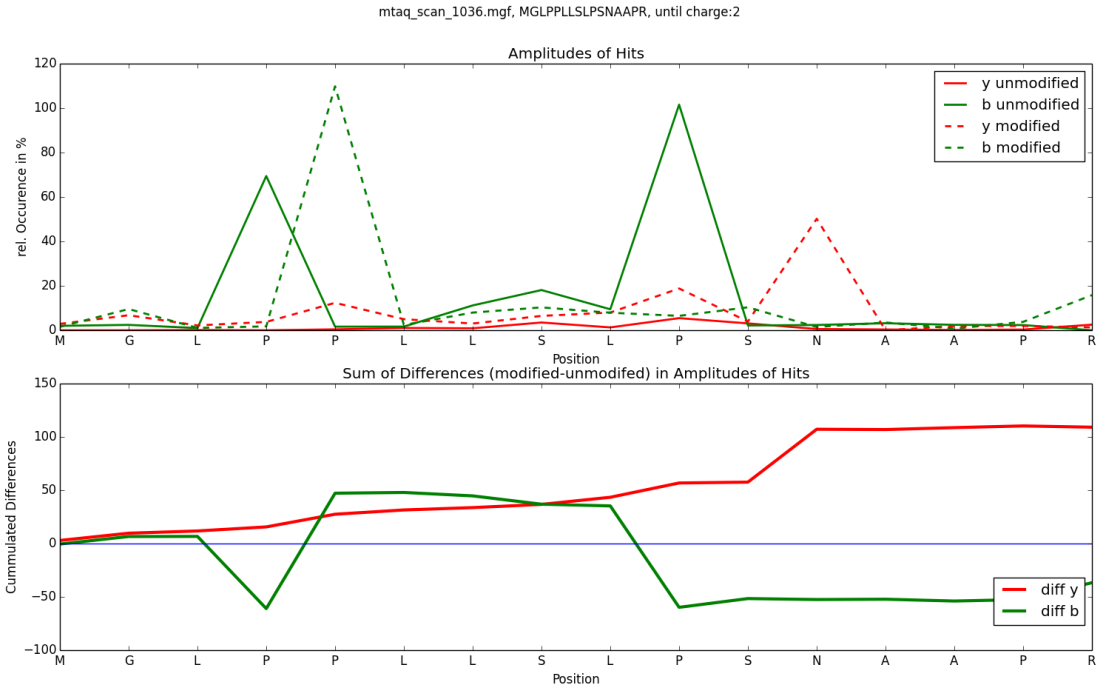


Abbildung 30: Sequence Plot von M.Taql Scan 1036 mit 'MGLPPLLSLPSNAAPR'

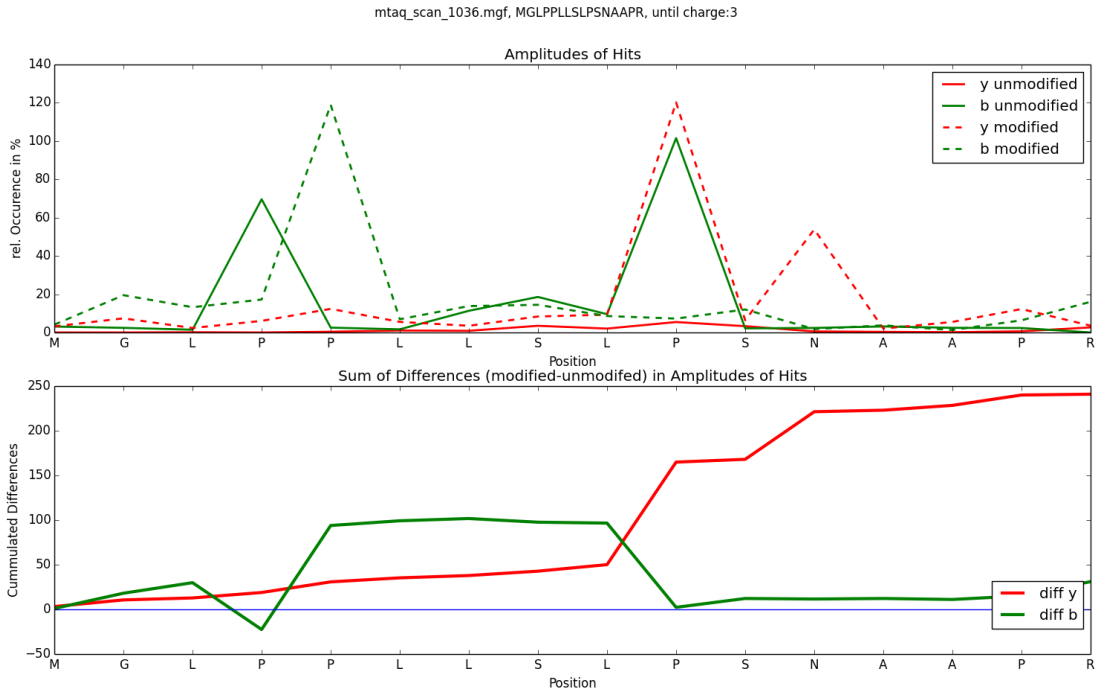


Abbildung 31: Sequence Plot M.Taql Scan 1036 mit 'MGLPPLLSLPSNAAPR'

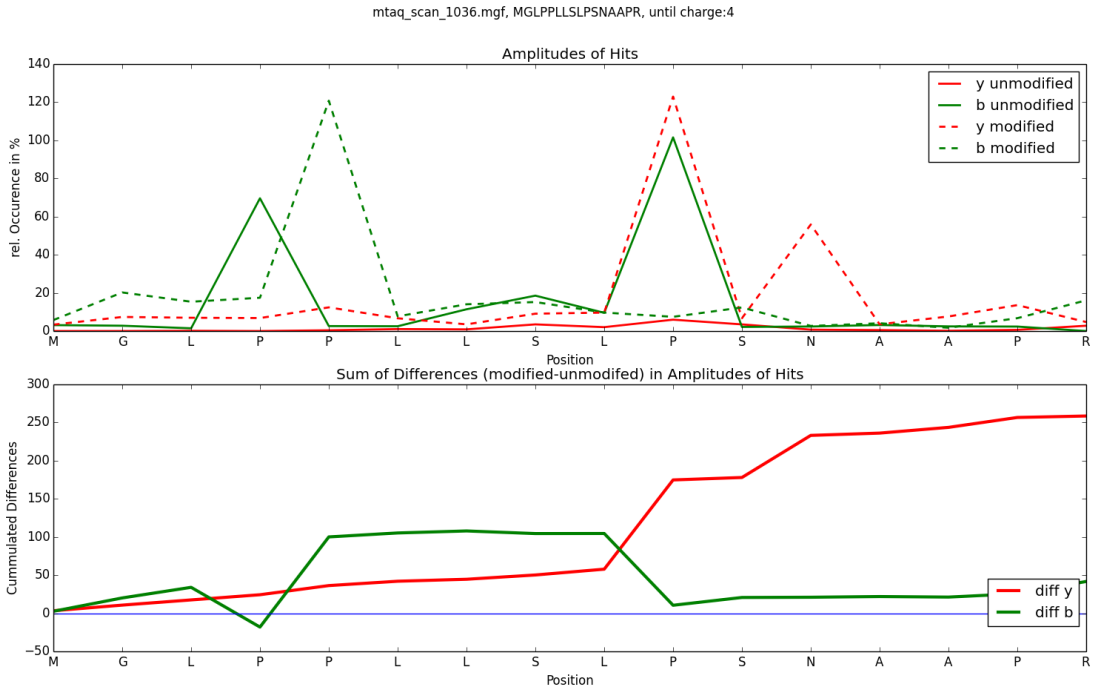


Abbildung 32: Sequence Plot M.Taql Scan 1036 mit 'MGLPPLLSLPSNAAPR'

Zusammenfassend kann also das dritte Prolin als häufigste Bindungsstelle des Capture Compounds angegeben werden. Sicherlich bindet die Modifikation auch am zweiten Prolin-Rest, allerdings fehlt hier die Bestätigung durch die y-Serie. Eine Fragmentierung via CID erzeugt häufiger y-Ionen[27] als b-Ionen, daher bekommt das dritte Prolin eine höhere Wahrscheinlichkeit in der Gesamtwertung.

6 Fazit und Ausblick

ilvamo kann ein MS/MS Spektrum zusammen mit einer gegebenen Peptid Liste identifizieren, annotierten und die möglichen Bindungsstellen einer variablen Modifikation mit bekannter, variabler Fragmentierung untersuchen.

Das Identifizieren wird dabei mithilfe von zwei unterschiedlichen Bewertungsmethoden absolviert; einmal die Summierung der Intensitäten und dazu der Korrelationskoeffizient der m/z-Werte. Jeder dieser Techniken betrachtet unterschiedliche Eigenschaften der Übereinstimmungen zwischen MS/MS und theoretischen Spektrum. Das identifizierte Peptid ergibt sich jeweils aus dem Vergleich der Top drei Ergebnissen beider Scores. Nach dem Identifizieren kann das MS/MS Spektrum annotiert und in einem Plot abgebildet werden mit beliebigen Ladungszustand, etwaigen Neutral Losses oder alternativen Modifikationen.

Will der Anwender die exakte Position der Modifikation im Peptid ermitteln, kann dafür sowohl der annotierte Plot als auch eine eigene Funktion verwendet werden. Letzteres markiert solche Stellen in der Peptidsequenz, an denen Sprünge in den annotierten Serien auftreten und die durch einen Massenshift bedingt sein können. Unter Einbeziehung der relativen Häufigkeiten der Peaks ergeben sich daraus grobe Wahrscheinlichkeiten für Bindungsstellen im Peptid. Zusammenfassend ist es mit dem Tool *ilvamo* also möglich, die ungefähre Position einer variablen Modifikation via MS/MS abzuleiten. Das Ziel dieser Projektstudie ist somit erreicht; weitere Ergänzungen und Verbesserungen sind möglich.

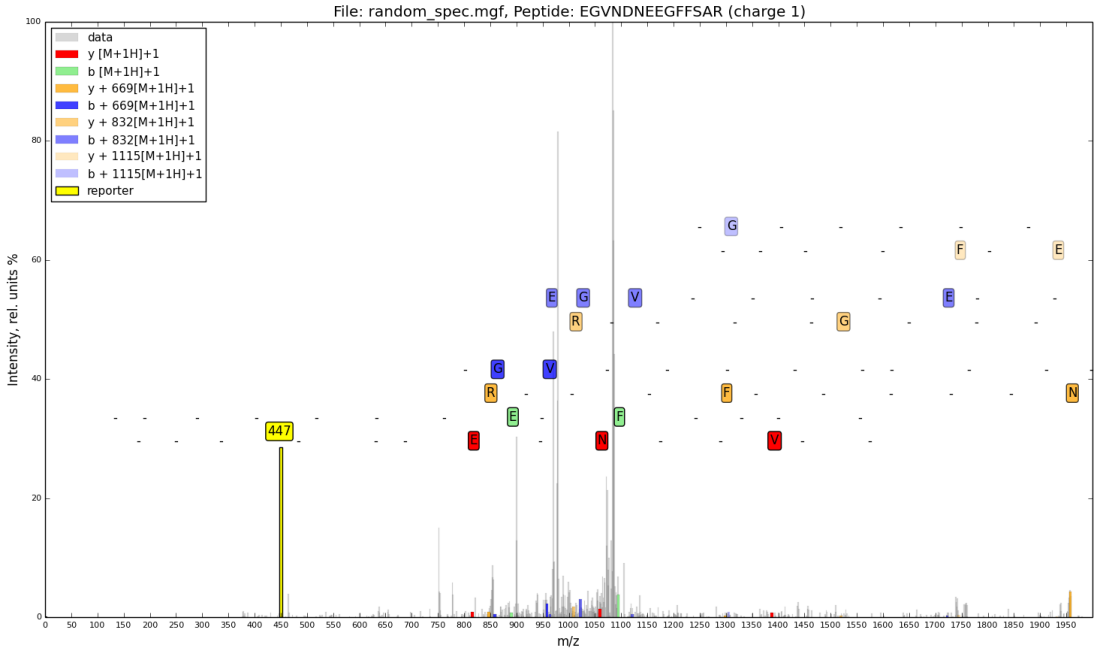
Eine Einbettung als Tool für OpenMS Toppas ist angestrebt. Außerdem sollten die Score Mechanismen zur Identifizierung noch verbessert werden. Ansonsten könnte die Entscheidungsphase bei Identifizierung und Bindungspositionen noch automatisiert werden. Dazu sind aber definitiv noch weiter Daten und Untersuchungen über das Verhalten und die Fragmentations-Muster der Capture Compound notwendig.

A Anhang

Tabelle 6: Monoisotopische Massen von Aminosäure-Residuen im Protein [17]

Residue	Code	Masse
Ala	A	71.03714
Arg	R	156.101111
Asn	N	114.042927
Asp	D	115.026943
Cys	C	103.009185
Glu	E	129.042593
Gln	Q	128.058578
Gly	G	57.021464
His	H	137.058912
Ile	I	113.084064
Leu	L	113.084064
Lys	K	128.092963
Met	M	131.040485
Phe	F	147.068414
Pro	P	97.052764
Ser	S	87.032028
Thr	T	101.047679
Trp	W	186.079313
Tyr	Y	163.06332
Val	V	99.068414

Abbildung 33: Ein beliebiges Random Spektrum und ähnlicher Masse wie Glufib im Vergleich mit dessen Peptid Sequenz 'EGVNDNEEGFFSAR' bei $z = 1$. Das Spektrum matcht zwar an einzelnen Stellen, ist aber im Ganzen deutlich schlechter annotiert als die Glufib Daten. Der default Fehler reicht also aus, um eine korrekte Identifizierung zu ermöglichen.



Listing 8: Beispiel Output von *identify()* für x-link Glufib bei Ladung $z = 1$, gerundet, sonst Genauigkeit bei 16 Stellen nach dem Komma.

```
file: glufib_xlink.mgf
        charge: 1
unmodified
y: ['-', 'G', '-', '-', '-', 'N', 'E', 'E', 'G', '-', 'F', 'S', '-', '-']
amp: [0, 0.12, 0, 0, 0, 0.57, 0.98, 0.65, 1.66, 0, 0.35,0.041, 0, 0]
m/z: [0, 1441.75, 0, 0, 0, 1056.55, 942.27, 812.94, 684.10, 0, 480.39,
333.22, 0, 0]
b: ['-', '-', 'V', 'N', 'D', '-', 'E', 'E', 'G', 'F', 'F', 'S', 'A', '-']
amp: [0, 0, 0.09, 0.29, 0.33, 0, 3.35, 1.05, 0.22, 1.39, 0.04, 0.04, 0.11,
 0]
m/z: [0, 0, 286.17, 400.19, 515.35, 0, 758.52, 887.37, 944.09,
 1091.24, 1238.80, 1325.63, 1396.21, 0]

cc fragment:669
```

```
y: ['-', '-', '-', '-', '-', '-', '-', '-', '-', '-', 'F', 'S', 'A', '-']
amp: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3.08, 1.07, 0.37, 0]
m/z: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1149.66, 1002.56, 915.06, 0]
b: ['E', 'G', 'V', 'N', '-', 'N', 'E', 'E', '-', '-', '-', '-', '-', '-']
amp: [1.17, 0.61, 0.48, 4.40, 0, 0.04, 0.05, 0.22, 0, 0, 0, 0, 0, 0]
m/z: [799.52, 856.00, 955.28, 1069.59, 0, 1298.87, 1427.97, 1556.65, 0, 0,
 0, 0, 0, 0]

cc fragment:832
y: ['-', '-', '-', '-', '-', '-', '-', 'E', '-', '-', '-', 'S', 'A', 'R']
amp: [0, 0, 0, 0, 0, 0, 0, 0.17, 0, 0, 0, 0.09, 0.88, 0.32]
m/z: [0, 0, 0, 0, 0, 0, 0, 1645.65, 0, 0, 0, 1165.43, 1078.79, 1007.43]
b: ['E', 'G', 'V', '-', 'D', 'N', '-', 'E', '-', 'F', '-', '-', '-', '-']
amp: [0.21, 1.29, 2.07, 0, 0.09, 0.06, 0, 0.10, 0, 0.07, 0, 0, 0, 0]
m/z: [962.04, 1019.56, 1118.75, 0, 1347.08, 1461.86, 0, 1719.74, 0,
1923.33, 0, 0, 0, 0]

cc fragment:1115
y: ['-', '-', '-', '-', '-', '-', '-', 'E', 'G', 'F', 'F', 'S', '-', '-']
amp: [0, 0, 0, 0, 0, 0, 0, 0.12, 0.07, 0.04, 0.06, 0.07, 0, 0]
m/z: [0, 0, 0, 0, 0, 0, 0, 1928.67, 1799.55, 1742.73, 1595.50, 1448.81,
0, 0]
b: ['-', '-', '-', '-', '-', '-', '-', '-', '-', '-', '-', '-', '-', '-']
amp: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
m/z: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

Listing 9: Scores für Scans 1036,1164 und 1765

```
file  mtaq_scan_1036.mgf
peptide: VETPPEVVDFMVSLAEAPR -> score: 662
peptide: LPDSSLVQWLNSEAMQK -> score: 2181
peptide: LEISGMPLGDLFHIR -> score: 2235
peptide: MGLPPLLSLPSNAAPR -> score: 5241
peptide: DFYATPHLVVAHTK -> score: 1453
peptide: TSVYYLGEVFPQK -> score: 711
peptide: VLEPACAHGPFLR -> score: 356
peptide: EYGFHTSPESAR -> score: 436
peptide: NLKPGWVDYEK -> score: 879
peptide: EPGPGLVPVLTGR -> score: 1332

VETPPEVVDFMVSLAEAPR xcorr y: 0.068 , xcorr b: 0.147
LPDSSLVQWLNSEAMQK xcorr y: 0.278 , xcorr b: -0.013
LEISGMPLGDLFHIR xcorr y: 0.348 , xcorr b: 0.773
MGLPPLLSLPSNAAPR xcorr y: 0.135 , xcorr b: 0.514
```

```
DFYATPHLVVAHTK xcorr y: 0.606 , xcorr b: 0.056
TSVYYLGEVFPQK xcorr y: 0.56 , xcorr b: 0.4
VLEPACAHGPFLR xcorr y: 0.778 , xcorr b: 0.431
EYGFHTSPESAR xcorr y: 0.434 , xcorr b: 0.764
NLKPGWVDYEK xcorr y: 0.586 , xcorr b: 0.599
EPGPGLVPVLTGR xcorr y: 0.608 , xcorr b: 0.793


file mtaq_scan_1164.mgf
peptide: VETPPEVVDFMVSLAEAPR -> score: 493
peptide: LPDSSLVQWLNSEAMQK -> score: 131
peptide: LEISGMPLGDLFHIR -> score: 354
peptide: MGLPPLLSLPSNAAPR -> score: 663
peptide: DFYATPHLVVAHTK -> score: 306
peptide: TSVYYLGEVFPQK -> score: 343
peptide: VLEPACAHGPFLR -> score: 364
peptide: EYGFHTSPESAR -> score: 174
peptide: NLKPGWVDYEK -> score: 424
peptide: EPGPGLVPVLTGR -> score: 387

VETPPEVVDFMVSLAEAPR xcorr y: -0.067 , xcorr b: -0.06
LPDSSLVQWLNSEAMQK xcorr y: 0.07 , xcorr b: 0.483
LEISGMPLGDLFHIR xcorr y: 0.437 , xcorr b: 0.306
MGLPPLLSLPSNAAPR xcorr y: 0.179 , xcorr b: 0.499
DFYATPHLVVAHTK xcorr y: 0.493 , xcorr b: 0.517
TSVYYLGEVFPQK xcorr y: 0.286 , xcorr b: 0.4
VLEPACAHGPFLR xcorr y: 0.693 , xcorr b: 0.94
EYGFHTSPESAR xcorr y: 0.989 , xcorr b: 0.43
NLKPGWVDYEK xcorr y: 0.588 , xcorr b: 0.548
EPGPGLVPVLTGR xcorr y: 0.695 , xcorr b: 0.914


file mtaq_scan_1756.mgf
peptide: VETPPEVVDFMVSLAEAPR -> score: 64
peptide: LPDSSLVQWLNSEAMQK -> score: 450
peptide: LEISGMPLGDLFHIR -> score: 292
peptide: MGLPPLLSLPSNAAPR -> score: 2034
peptide: DFYATPHLVVAHTK -> score: 338
peptide: TSVYYLGEVFPQK -> score: 78
peptide: VLEPACAHGPFLR -> score: 153
peptide: EYGFHTSPESAR -> score: 430
peptide: NLKPGWVDYEK -> score: 249
peptide: EPGPGLVPVLTGR -> score: 519
```

VETPPEVVDFMVSLAEAPR xcorr y: −0.249 , xcorr b: −0.219
LPDSSLVQWLNSEAMQK xcorr y: −0.217 , xcorr b: −0.304
LEISGMPLGDLFHIR xcorr y: −0.097 , xcorr b: −0.166
MGLPPLLSLPSNAAPR xcorr y: −0.052 , xcorr b: 0.102
DFYATPHLVVAHTK xcorr y: −0.143 , xcorr b: 0.103
TSVYYLGEVFPQK xcorr y: −0.128 , xcorr b: −0.145
VLEPACAHGPFLR xcorr y: 0.169 , xcorr b: 0.294
EYGFHTSPESAR xcorr y: 0.175 , xcorr b: 0.23
NLKPGWVDYEK xcorr y: 0.508 , xcorr b: −0.292
EPGPGLVPVLTGR xcorr y: 0.269 , xcorr b: 0.215

Literatur

[1] Vernetzung (Chemie). Vernetzung – Wikipedia, die freie Enzyklopaedie. http://de.wikipedia.org/wiki/Vernetzung_%28Chemie%29, 2013. [Online; Zugriff 8-Maerz 2014].

[2] Universiteatsklinkum Jena, Klinik fuer Augenheilkunde. Cornea Cross Linking (Hornhautvernetzung) bei Keratokonus . http://www.augenklinik.uniklinikum-jena.de/Patienteninformation/Krankheitsbilder+_+OP/Cornea+Cross+Linking.html, 2014. [Online; Zugriff 12-Maerz 2014].

[3] Immuchrom, Immunologie- und Chromatographiespezialisten aus Deutschland. Hydroxypyridinium Crosslinks . http://www.immuchrom.de/de/produkte/calciumknochenstoffwechsel-/hydroxypyridinium-crosslinks.html, 2014. [Online; Zugriff 12-Maerz 2014].

[4] Caprotec Bioanalytics GmbH. Capture Compound Masse Spectrometry (CCMS) Technology , 2013. [Authorisiert von Dr. Graebner].

[5] National Human Genome Research Institute. File:Primärstruktur von Proteinen.svg – Wikimedia, das freie Medienarchiv. http://commons.wikimedia.org/wiki/File:Prim%C3%A4rstruktur_von_Proteinen.svg?uselang=de, 2009. [Online; Zugriff 12-Maerz 2014].

[6] Thermo Scientific – Jared Snider, Ph.D. . Overview of Mass Spectrometry. http://www.piercenet.com/method/overview-mass-spectrometry, 2014. [Online; Zugriff 25-Maerz 2014].

[7] Antec . Protein/Peptide Cleavage (Digestion) . http://www.myantec.com/markets/electrochemistry-with-ms-detection-ec-ms/proteomics,-protein-chemistry,-biopharmaceuticals/protein-peptide-cleavage, 2014. [Online; Zugriff 25-Maerz 2014].

[8] quantockgoblin. Datei:Column chromatography sequence.png– Wikibooks, die freie Bibliothek. http://de.wikibooks.org/wiki/Datei:Column_chromatography_sequence.png, 2008. [Online; Zugriff 25-Maerz 2014].

[9] De-Novo-Peptidsequenzierung. De-Novo-Peptidsequenzierung – Wikipedia, die freie Enzyklopaedie. http://de.wikipedia.org/wiki/De-Novo-Peptidsequenzierung, 2013. [Online; Zugriff 8-Maerz 2014].

[10] Tandem mass spectrometry. Tandem mass spectrometry – Wikipedia, the free Enzyclopedia. http://en.wikipedia.org/wiki/Tandem_mass_spectrometry, 2014. [Online; Zugriff 9-Maerz 2014].

[11] Institute for Systems Biology (USA, Seattle) . Fragment Ion Calculator. http://db.systemsbiology.net:8080/proteomicsToolkit/FragIonServlet.html, 2014. [Online; Zugriff 11-Maerz 2014].

[12] protea. Glu-1-Fibrinopeptide B (Glu-Fib) Peptide Mass Standard. `https://proteabio.com/products/PS-165`, 2014. [Online; Zugriff 13-Maerz 2014].

[13] J K Eng, A L McCormack, and J R Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–89, November 1994.

[14] D N Perkins, D J Pappin, D M Creasy, and J S Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–67, December 1999.

[15] Dr. Patrick Pedrioli. Proteomics of Ubiquitin-Like Modifiers. `http://www.ppu.mrc.ac.uk/research/?pid=1005&sub1=research`, 2014. [Online; Zugriff 14-Maerz 2014].

[16] Jian Liu, Alexander W Bell, John J M Bergeron, Corey M Yanofsky, Brian Carrillo, Christian E H Beaudrie, and Robert E Kearney. Methods for peptide identification by spectral comparison. *Proteome science*, 5:3, January 2007.

[17] Prof. Dr. Henry Jakubowski. Biochemistry Online – College of Saint Benedict and Saint John's University (USA, Minnesota). `http://employees.csbsju.edu/hjakubowski/classes/ch331/protstructure/olcompseqconform.html`, 2014. [Online; Zugriff 10-Maerz 2014].

[18] Marc Sturm, Andreas Bertsch, Clemens Gröpl, Andreas Hildebrandt, Rene Hussong, Eva Lange, Nico Pfeifer, Ole Schulz-Trieglaff, Alexandra Zerck, Knut Reinert, and Oliver Kohlbacher. OpenMS - an open-source software framework for mass spectrometry. *BMC bioinformatics*, 9:163, January 2008.

[19] Joseph Zaia. Mass spectrometry and glycomics. *Omics : a journal of integrative biology*, 14(4):401–18, August 2010.

[20] Manfred Wuhrer. Glycomics using mass spectrometry. *Glycoconjugate journal*, 30(1):11–22, January 2013.

[21] Juri Rappsilber. The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *Journal of structural biology*, 173(3):530–40, March 2011.

[22] UniProtKB/Swiss-Prot. P14385. `http://www.uniprot.org/uniprot/P14385`, 2013. [Online; Zugriff 24-Maerz 2014].

[23] Anton Goloborodko, Lev Levitsky. pyteomics 2.3.0. `https://pypi.python.org/pypi/pyteomics`, 2014. [Online; Zugriff 17-Maerz 2014].

[24] Witold E Wolski, Maciej Lalowski, Peter Martus, Ralf Herwig, Patrick Giavalisco, Johan Gobom, Albert Sickmann, Hans Lehrach, and Knut Reinert. Transformation and other factors of the peptide mass spectrometry pairwise peak-list comparison process. *BMC bioinformatics*, 6:285, January 2005.

[25] Stowers Institute for Mediacal Research (USA, Kansas) . Database Matching. `http://research.stowers.org/proteomics/DataMatch.html`, 2013. [Online; Zugriff 24-Maerz 2014].

[26] Kevin Bertman . Rogue Waves. `http://kevinbertman.edublogs.org/2011/06/14/rogue-waves/`, 2011. [Online; Zugriff 25-Maerz 2014].

[27] Jainab Khatun, Kevin Ramkissoon, and Morgan C Giddings. Fragmentation characteristics of collision-induced dissociation in MALDI TOF/TOF mass spectrometry. *Analytical chemistry*, 79(8):3032–40, April 2007.

Abbildungsverzeichnis

Tabellenverzeichnis

Eidesstattliche Erklärung

Ich versichere hiermit, dass diese Arbeit von niemand anderem als meiner Person verfasst worden ist. Alle verwendeten Hilfsmittel sind im Literaturverzeichnis angegeben, Zitate aus fremden Arbeiten sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

Berlin, der 28. April 2014
Sophie Kolbe