De-Novo-Sequencing using Spectrum-Graphs, enabling Open Searches

Dominik Habermann

5. Juni 2023

1 Einleitung

1.1 Biomedizinische Fragestellung

Peptide sind organische Verbindungen von miteinander verknüpften Aminosäuren. Bei der Sequenzierung von Peptiden versucht man die Aminosäuresequenz – also die Abfolge an vorhandenen Aminosäuren – zu bestimmen. Das Wissen über die Aminosäuresequenz ist von großer Bedeutung für den Forschungsbereich der Proteomik. Die Proteomik beschäftigt sich mit der Erforschung von Proteinen. Dies beinhaltet unter anderem auch die Analyse von Enzymen.

Da es 20 verschiedene relevante Aminosäuren gibt [1, S. 377], die weitestgehend beliebig miteinander kombiniert werden können, existiert eine stark wachsende Anzahl an möglichen Kombinationen. Die Regeln der Kombinatorik liefert uns hierfür die Formel $f(x) = 20^x$ wobei x hier die Anzahl an Aminosäuren ist. Es ist direkt erkennbar, dass selbst bei einer geringen Peptidlänge die Anzahl an möglichen Sequenzen eine Größenordnung erreicht, die von Computersystemen nicht mehr verarbeitet werden kann. Zum Vergleich: Proteine können aus wenigen Hundert bis hin zu aus mehreren Zehntausend Aminosäuren bestehen. Die Frage, die sich hier stellt: *Ist es zumindest für kurze Peptide möglich diese sicher zu sequenzieren?*

1.2 Methoden der Aminosäuresequenzierung

Das Ziel der verschiedenen Sequenzierungsverfahren ist eine möglichst exakte Bestimmung der Aminosäuresequenz. Alle Sequenzierungsverfahren arbeiten mit der Massenspektrometrie (MS). Dabei handelt es sich um ein Verfahren, welches chemische Verbindungen identifizieren kann (eine genauere Erklärung folgt in Kapitel 2). Viele Analysen arbeiten mit dem Ansatz, dass die Ergebnisse einer MS – genannt wird es Massenspektrum – mit einer Datenbank verglichen werden. Wenn die chemische Verbindung bereits einmal identifiziert wurde, dann wird sich ein Eintrag in der Datenbank finden lassen.

Die hier vorgestellten Methoden *pNovo+* und *Open-pNovo* gehören zur Gruppe der De-Novo-Peptidsequenzierungen. Im Gegensatz zu anderen Verfahren werden hierbei keinerlei

Daten aus Datenbanken verwendet. Die De-Novo-Peptidsequenzierung hat den bedeutsamen Vorteil, dass auch Peptide sequenziert werden können zu denen es keine oder nur unvollständige Informationen gibt. Statt einer MS findet eine Tandem-Massenspektrometrie Anwendung. Bei dieser Form der MS werden zwei MS Durchgänge hintereinander durchgeführt, wobei nach dem ersten Vorgang ein Teil der Probe isoliert wird und vor der 2. MS „fragmentiert“ wird (hierzu eine Beschreibung in Kapitel 2.1 mit mehr Details).

2 Massenspektrometrie (MS)

Wie bereits in Kapitel 1 erwähnt, wird die MS verwendet, um chemische Strukturen zu identifizieren. Die ersten modernen Ansätze der MS wurden zu Beginn des 20. Jahrhunderts entwickelt [2, S. 5678]. Seitdem gab es etliche Erweiterungen; das Grundprinzip ist dennoch immer gleich geblieben. Grob vereinfacht besteht eine MS aus folgenden vier Schritten:

- **Ionisation**: Die Moleküle in der Probe bekommen eine positive oder negative Ladung
- **Überführung in die Gasphase**: Durch Energie wird die Probe in die Gasphase überführt
- **Anlegen eines elektrischen Feldes**: Die Ionen werden durch ein elektrisches Feld beschleunigt
- **Massenanalyse**: Ionen werden anhand des Masse-Ladungs-Verhältnisses „sortiert“

Für die Schritte gibt es verschiedene Verfahren, wobei die Unterschiede hier nicht relevant sind. Jedes dieser Verfahren nutzt die physikalische Eigenschaft aus, dass Ionen in einem Magnetfeld in Abhängigkeit ihres Verhältnisses zwischen ihrer Masse und ihrer Ladung (häufig abgekürzt mit m/z) unterschiedlich reagieren. So wird bei der MS nicht die Masse gemessen – auch wenn der Name es vermuten lässt – sondern die Ionenhäufigkeit bei einem bestimmten m/z Verhältnis [3, S. 140]. Diese Häufigkeit wird dann in einem Massenspektrum graphisch dargestellt. Abbildung 1 zeigt ein computergeneriertes Massenspektrum[1] von zwei ähnlichen Aminosäuren.

Die Maxima werden „Peaks“ genannt und haben für eine Aminosäure einen charakteristischen m/z Wert. Obwohl sich die beiden Aminosäuren in der Abbildung 1 nur durch ein Atom unterscheiden (das linke Sauerstoffatom wurde durch ein Schwefelatom ersetzt) sind deren Peaks weit voneinander entfernt und machen die beiden Aminosäuren dadurch sicher unterscheidbar.

Bei einzelnen Aminosäuren funktioniert die MS zuverlässig; bei Peptiden allerdings steht man vor dem Problem, dass das Massenspektrum unübersichtlicher wird und auch Peaks, die von Hintergrundrauschen stammen, schwerer herausgefiltert werden können. Abhilfe schafft hier die Tandem-Massenspektrometrie.

[1] Generiert von der Website: `https://www.protpi.ch/Calculator/MassSpecSimulator`

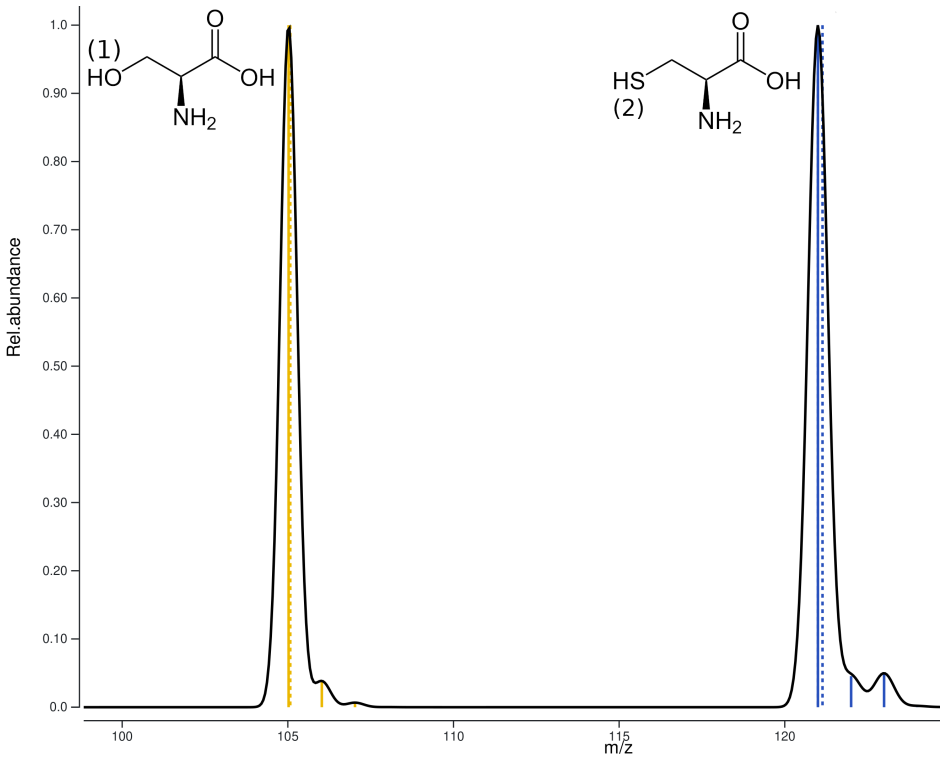

Abbildung 1: Computergeneriertes Massenspektrum von der Aminosäure *Serin* (1) und *Cystesin* (2). Peak von *Serin* liegt bei 105 m/z ; bei *Systesin* um 121 m/z .

2.1 Tandem-Massenspektrometrie (MS/MS)

Bei der Tandem-Massenspektrometrie (MS/MS oder MS2) werden zwei MS Vorgänge hintereinander mit einer Probe durchgeführt. Die erste MS dient dazu Ionen aus einem bestimmten m/z Bereich auswählbar zu machen. Es entspricht also quasi einer Form der Filterung. Vor der 2. MS werden die ausgewählten Reste einer Fragmentierung unterzogen. Bei einer Fragmentierung führt man Energie zu mit dem Ziel, dass die Ionen zerfallen und Fragment-Ionen bilden. Diese Fragment-Ionen werden dann auf dem Massenspektrum nach der 2. MS sichtbar gemacht.

Fragment-Ionen sind kleiner als die ursprünglichen Ionen. So kann die 2. MS mit einer höheren Selektivität durchgeführt werden, welches Peaks durch Hintergrundrauschen verringert. Auch lassen sich Ionen besser identifizieren, die ein sehr ähnliches m/z -Verhältnis besitzen. Nach der 2. MS liegt eine Fülle an Fragment-Ionen-Peaks vor, aus denen sich die ursprünglichen chemischen Strukturen ableiten lassen, da Ionen in spezifische Fragmente zerfallen [4]. Zusammengefasst kann man sagen, dass das MS/MS Verfahren Ergebnisse höherer Güte erzeugt im Vergleich zur einfachen MS.

3 De-Novo-Peptidsequenzierung mit *pNovo+*

Die Sequenzierung nach pNovo+ ist eine De-Novo-Peptidsequenzierung, die mit einem Spektrums-Graphen für die Auswertung der MS2-Spektren arbeitet und eine Erweiterung des pNovo Verfahrens darstellt [5]. Der Hauptansatz ist, dass zwei MS/MS Durchläufe mit jeweils verschiedenen Fragmentierungsmethoden[2] durchgeführt werden (pNovo hat nur mit einem MS2-Spektrum gearbeitet). Durch die Wahl einer anderen Fragmentierungsmethode ändert sich auch das MS2-Spektrum. Wenn nun Fragmentierungsmethoden verwendet werden, die möglichst komplementäre Spektren erzeugen, dann lässt sich durch das Zusammenführen der beiden MS2-Spektren die Qualität der Ergebnisse verbessern. Zum Beispiel lassen sich dadurch viele Peaks, die vom Hintergrundrauschen stammen, entfernen.

Für die Ermittlung der Sequenz eines Peptids wird zunächst ein Spektrums-Graph gebildet – in Form eines DAG (directed acyclic graph). In diesem Graphen wird dann der längste Pfad bei gegebenen Start- und Endknoten berechnet. Die Reihenfolge der Knoten, die im längsten Pfad durchlaufen werden, stellt dann die Peptidsequenz dar.

3.1 Vorverarbeitung der MS2-Spektren

Bevor aus den MS2-Spektren der Spektrums-Graph gebildet werden kann, müssen die Daten vorverarbeitet werden. Für die Auswertung ist es von entscheidender Bedeutung, dass möglichst wenig Peaks verwendet werden, die vom Hintergrundrauschen stammen. Im weiteren Verlauf werden an einem exemplarischen MS2-Spektrum die Verarbeitungsschritte dargestellt.

Der erste Schritt ist das Verwenden des natürlichen Logarithmus der Intensitäten (entspricht hier der y-Achse). Die Idee dabei ist, dass Hintergrundrauschen nicht über-priorisiert wird.

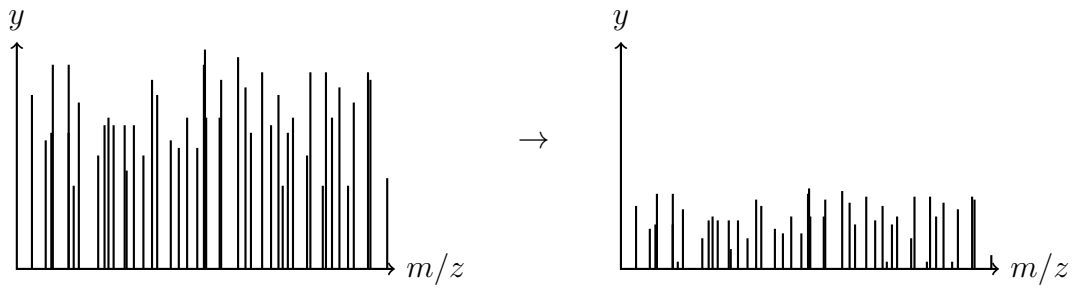

Abbildung 2: Anwendung des ln auf einem exemplarischen MS2-Spektrum.

Für das Verständnis des nächsten Schrittes muss man sich in Erinnerung rufen, dass eine gleiche Aminosäure keineswegs immer die gleiche Masse hat. Durch Isotope existiert eine gewisse „Massenbandbreite“ für ein und dieselbe Aminosäure. MS Systeme sind heute so genau, dass sie diese Differenzen erkennen. Dies hat den ungewollten Effekt, dass mehrere

[2] pNovo+ verwendet die higher energy collisional dissociation (HCD) und die electron transfer dissociation (ETD) Fragmentierungsmethoden.

Peaks zu einer Aminosäure gehören können [6]. Gleichzeitig können die „Massenbandbreiten“ zweier Aminosäuren sich überschneiden, sodass im ungünstigen Fall zwei Peaks kaum unterscheidbar nebeneinander liegen.

Eine Möglichkeit mit dieser Problematik umzugehen ist die Verwendung der monoisotopischen Masse. Die monoisotopische Masse ist die „[...] exact mass of the most abundant naturally occurring stable isotope determined relative to the mass of ^{12}C, which is assigned the exact value of 12.0000.“ [7]. Ohne dabei jetzt tiefer ins Detail zu gehen kann man sagen, dass alle Peaks, deren Intensitäten mit einer möglichen monoisotopischen Masse übereinstimmen, auf jeden Fall einer Aminosäure entsprechen und (höchstwahrscheinlich)[3] kein Hintergrundrauschen sind [8]. Diese Peaks bekommen eine sogenannte *charge state*.

Der Algorithmus verwendet die *charge state* Peaks als Ausganspunkte für weitere Berechnungen. Wenn die m/z Differenz zu einem anderen Peak einem m/z eines Peptidfragmentes entspricht, dann stammt dieser Peak höchstwahrscheinlich von einem Fragment. Insgesamt werden damit die relevanten Peptidfragmente herausgeholt. Abbildung 3 zeigt das Ergebnis nach den beiden zuvor genannten Schritten.

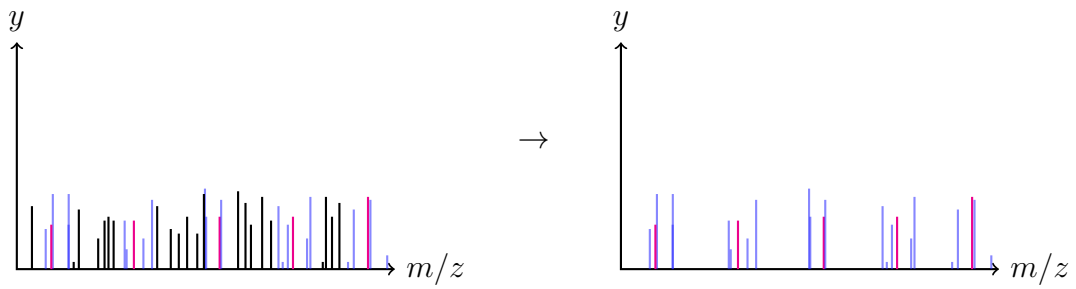

Abbildung 3: Entfernen von Peaks, die keiner monoisotopischen Masse entsprechen oder benachbart mit einer Differenz von einem Fragment-Ion sind. Magenta-Peaks stellen monoisotopische Massen dar; blaue-Peaks besitzen eine passende Differenz zu einem Magenta-Peak

Tatsächlich ist die Verarbeitung an dieser Stelle noch etwas komplexer. So existieren auch noch sogenannte *isotopic cluster*[4], die gesondert verarbeitet werden. Für das grundsätzliche Prinzip ist dieses Detail allerdings weniger relevant.

Im letzten Vorberarbeitungsschritt werden Peaks aus einem irrelevanten m/z Bereich entfernt und naheliegende Peaks werden zusammengefasst, indem der Mittelwert sowohl des m/z Wertes als auch der der Intensität bestimmt wird. Üblicherweise liegt der Bereich für das Zusammenfassen bei $\pm 20ppm$.

[3]Natürlich ist es möglich, dass das Rauschen zufällig einer monoisotopischen Masse entspricht. Die Wahrscheinlichkeit dafür ist allerdings sehr gering.

[4]Definition eines *isotopic cluster* nach IUPAC: „Group of peaks representing ions of the same elemental composition, but different isotopic compositions.“ [9, S. 1556]

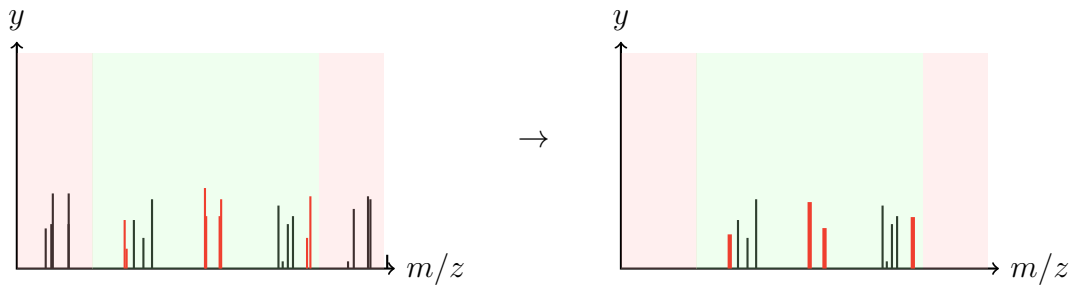

Abbildung 4: Entfernen von Peaks aus einem irrelevanten m/z Bereich und zusammenfassen naheliegender Peaks. Rot markierte Peaks sind jene, die zusammengefasst werden.

3.2 Bildung eines Spektrums-Graphen

Der Spektrums-Graph wird aus einem vorverarbeiteten MS2-Spektrum (siehe Kapitel: 3.1) gebildet. Im initialen Zustand werden die Peaks als Knoten interpretiert. Dazu kommt ein Start- und Endknoten. Jedem Knoten wird eine Masse zugeordnet; im initialen Zustand bekommt der Startknoten die Masse 0 und der Endknoten die Masse des vorherigen Knotens minus der Masse des Wassers $(18,02)$. Die Masse der übrigen Knoten entsprechen ihren jeweils korrespondierenden m/z Wert. Die gerichteten Kanten werden zwischen einem Knotenpaar hinzugefügt, wenn die Differenz deren Masse gleich ist mit der Masse von *ein* oder *zwei* Aminosäuren.

3.3 Identifikation der Aminosäuresequenz

Der gebildete DAG kann mit klassischen Algorithmen, die den längsten Pfad suchen, durchlaufen werden. Bezogen auf die Graphentheorie entspricht die Ermittlung der Aminosäurensequenz dem Suchen eines bestimmten Pfades – und nicht nach irgendeinem Pfad. Daher muss der Algorithmus mittels einer Breitensuche arbeiten, um alle möglichen Pfade zu bestimmen.

In aller Regel wird es mehrere Pfade geben. Bestimmte Sequenzen sind wahrscheinlicher als andere. So sind Pfade mit Kanten, die wegen der Massendifferenz von genau einer Aminosäure gebildet wurden, wahrscheinlicher [10]. Alle Pfade bekommen mittels einer Scoring-Funktion einen Wert zugewiesen. Der Pfad mit dem höchsten Scoring-Wert ist wahrscheinlich das richtige Ergebnis. Die Scoring-Funktion berücksichtigt unter anderem wie viele Fragmente, die einer bestimmten Aminosäure zugeordnet werden können, im MS2-Spektrum vorhanden sind [5]. Die Sequenz mit dem höchsten Scoring-Wert ist das Endergebnis.

3.4 Evaluierung von pNovo+

Mit echten Testdaten wurden die HCD und die ETD MS2-Spektren erstellt. Insgesamt bestand der Datensatz aus 8677 Elementen.

pNovo+			***PEAKS***		
HCD ∪ ETD	**HCD**	**ETD**	**HCD ∪ ETD**	**HCD**	**ETD**
$81,2\%$	$73,9\%$	$55,7\%$	$71,8\%$	$68,4\%$	30.6%

Tabelle 1: Gegenüberstellung der erfolgreichen De-Novo-Peptidsequenzierung von pNovo+ und PEAKS bei MS2-Spektren [10, S. 620].

Tabelle 1 zeigt, dass beider-maßen von der Verwendung zweier MS2-Spektren profitiert wird. Das könnte darauf hindeuten, dass De-Novo-Peptidsequenzierung Algorithmen im allgemeinen dadurch bessere Ergebnisse liefern können. Auffällig ist, dass pNovo+ mit jedem Datensatz besser abschneidet als PEAKS.

3.5 Ergebnisse und Diskussion

Das Ziel basierend auf pNovo ein De-Novo-Peptidsequenzierung Tool zu entwerfen, welches genauer und effizienter ist, wurde mit einer Verbesserung von $9,4$ Prozentpunkten definitiv erreicht. Die Grundidee, dass zwei komplementäre MS2-Spektren verwendet werden, scheint generell einen positiven Einfluss auf De-Novo-Peptidsequenzierungs-Algorithmen zu haben. Es bleibt allerdings offen, ob dieser Vorteil ebenfalls gegeben ist, wenn andere Fragmentierungsmethoden verwendet werden.

4 De-Novo-Peptidsequenzierung mit *Open-pNovo*

Bei Proteinen können posttranslationale Proteinmodifikationen (PTM) auftreten. PTMs sind Ereignisse, bei denen sich Änderungen im Protein einstellen; teilweise sind die Änderungen von einer Zelle erwünscht – teilweise stammen sie aber auch zum Beispiel von unerwünschten Wechselwirkungen nebeneinanderliegenden Aminosäuren [11, S. 256]. Ein Teil dieser PTMs führen zu einer Änderung der Aminosäuresequenz. Dies ist für die De-Novo-Peptidsequenzierung nicht weiter problematisch, da sowieso ohne eine Datenbank gearbeitet wird, sodass solche PTMs nicht einmal auffallen würden. Andere PTMs hingegen haben die Auswirkung, dass Stoffe gebildet werden, die nicht mehr zu der Gruppe der proteinogenen Aminosäuren gehören. Proteinogene Aminosäuren sind jene Aminosäuren, die für den Bau von Proteinen verwendet werden. Der Effekt ist also, dass Stoffe (oder deren Fragmente) bei einem Massenspektrum angezeigt werden, die kein Teil eines Peptids sein können. Bei der Sequenzierung von Peptidfragmenten muss dies daher berücksichtigt werden. Wenn im weiteren Verlauf von PTMs gesprochen wird, dann sind solche gemeint, die für die De-Novo-Peptidsequenzierung relevant sind.

Open-pNovo ist ein De-Novo-Peptidsequenzierungsverfahren, welches auf pNovo+ aufbaut und versucht die Problematik mit den PTMs zu lösen.

4.1 PTMs im konstruierten DAG

Die Konvertierung eines MS2-Spektrums läuft bis zum DAG analog ab wie in den Kapiteln 3.1 und 3.2 für pNovo+. Der Unterschied ist nun, dass es zwei Arten von Kanten gibt:

- **„Normale“ Kanten:** Kanten, die gebildet werden, wie es bereits für *pNovo+* gezeigt wurde.
- **„Modifizierte“ Kanten:** Kanten, die zum Graphen hinzugefügt werden, wenn die Massendifferenz zweier Knoten der Masse *genau einer* Aminosäure plus der Masse einer möglichen PTM-Änderung entspricht.

Eine Liste aller PTMs (sowohl relevante als auch nicht relevante) in der Datenbank Unimod beinhaltet aktuell 1510 Einträge[5] (Stand: 18.04.2022). Für die modifizierten Kanten gibt es daher mehr mögliche Differenzen als für die normalen Kanten.[6] Die hohe Anzahl an Differenzen für modifizierte Kanten hat die Konsequenz, dass viele Knoten zufällig verbunden werden und dass dadurch die Genauigkeit der Ergebnisse abnimmt. Dieses Problem kann man durch eine geringere Liste an möglichen PTMs abfedern, allerdings mit einem Verlust der Genauigkeit auf Seiten der PTMs. Es ist hier also eine Abwägung zu treffen.

4.2 Identifikation der Aminosäuresequenz

Auch bei Open-pNovo findet ein Algorithmus aus der Graphentheorie Anwendung, um alle möglichen Pfade durch den DAG zu bestimmen. Der Unterschied im Vergleich zu pNovo+ ist die Scoring-Funktion. Bei Open-pNovo wird ein Algorithmus namens RankBoost für diese Aufgabe verwendet. RankBoost ist ein Machine-Learning Algorithmus aus dem Jahr 2003, der versucht Präferenzen aus Datensätzen zu bestimmen [12, S. 933]. Ursprünglich wurde RankBoost geschaffen, um Filme für einen Nutzer vorzusortieren basierend auf Bewertungen anderer Nutzer. Open-pNovo überträgt den Anwendungszweck von RankBoost auf die Scoring-Problematik und nutzt diesen Algorithmus, trainiert durch reale Testdaten, um einen möglichst passenden Score zu bestimmen. Am Ende ist auch hier die Sequenz mit dem höchsten Score das Endergebnis.

4.3 Evaluierung von Open-pNovo

Open-pNovo wurde sowohl auf drei realen als auch auf drei generierten Testdaten getestet. Tabelle 2 zeigt die Ergebnisse im Vergleich zu pNovo+ und zwei anderen Algorithmen. Die generierten Datensätze enthielten die am häufigsten vorkommenden PTMs.

Testdatensätze	**Open-pNovo**	**pNovo+**	**PEAKS**	**Novor**
Real (20259)	76,3%	68,5%	65,8%	39,9%
Generiert (17877)	77,8%	0,6%	0,5%	0,2%

Tabelle 2: Vergleich der durchschnittlichen richtigen De-Novo-Peptidsequenzierungen von Open-pNovo und anderen Algorithmen [13, S. 650].

Die enorm schlechten Ergebnisse der anderen Algorithmen bei den generierten Testdaten ist ein Nebeneffekt des Ziels bei der Testdatengenerierung. Denn die Generierung wurde

[5] Siehe: `https://www.ebi.ac.uk/ols/ontologies/unimod`

[6] $1510 \cdot 20 = 30200$ für die modifizierten Kanten. $20^2 = 400$ für die normalen Kanten.

so ausgelegt, um die Grenzen von Open-pNovo zu ermitteln [13, S. 649]. Eine Aussagekraft haben diese Ergebnisse bezüglich der Trefferquote also nicht. Allerdings auch bei realen Testdaten zeigt sich Open-pNovo als voll konkurrenzfähig gegenüber den anderen Algorithmen.

Noch besser zeigt sich Open-pNovo, wenn der Recall Wert betrachtet wird – also die Anzahl an verschiedenen PTMs, die erkannt werden. In diesem Fall ist der Abstand zum nächstbesten Algorithmus deutlich größer geworden. ($7,8\% \rightarrow 30,3\%$)

Testdatensätze	**Open-pNovo**	**pNovo+**	**PEAKS**	**Novor**
Real (5034)	$61,6\%$	$31,3\%$	$32,0\%$	$13,7\%$

Tabelle 3: Vergleich der durchschnittlichen Recall Werte einer De-Novo-Peptidsequenzierung von Open-pNovo und anderen Algorithmen [13, S. 650].

4.4 Ergebnisse und Diskussion

Die Trefferquote von Open-pNovo liegt bei etwa 76% und ist damit voll konkurrenzfähig zu anderen Algorithmen und liefert meist die besten Ergebnisse. Es zeigt sich, dass die Wahl von RankBoost die Qualität der Ergebnisse entscheidend positiv beeinflusst. So steigt die durchschnittliche Trefferquote im Vergleich zu pNovo+ um knapp $7,8$ Prozentpunkte. Interessant dabei ist, dass bereits pNovo+ mit den Konkurrenzalgorithmen (allen voran PEAKS) mithalten kann.

Die Wahl eines Machine-Learning Verfahrens hat allerdings auch einen negativen Aspekt. Es werden viele Daten benötigt, um RankBoost ausreichend gut zu trainieren. Zwar kann Open-pNovo immer noch als De-Novo-Peptidsequenzierungs-Tool dienen, da für das Sequenzieren an sich keine Datenbank im Hintergrund notwendig ist, dennoch existiert eine gewisse „Abhängigkeit“ zu vorher sequenzierten Daten. Inwiefern dies tatsächlich ein Problem in der Praxis darstellt, kann an dieser Stelle nicht beurteilt werden.

Literatur

[1] Jens Rudat und Ulrike Engel. „Alanins Wunderlampe: Vorkommen, Nutzung und Produktion nicht-kanonischer Aminosäuren“. In: *Biologie in unserer Zeit* 51.4 (2021), S. 376–386.

[2] Jennifer Griffiths. „A brief history of mass spectrometry“. In: *Anal. Chem* 80.15 (2008), S. 5678–5683.

[3] Gary L. Glish und Richard W. Vachet. „The basics of mass spectrometry in the twenty-first century“. In: *Nature Reviews Drug Discovery* 2.2 (Feb. 2003), S. 140–150. ISSN: 1474-1784. DOI: 10.1038/nrd1011. URL: https://doi.org/10.1038/nrd1011.

[4] Jürgen H. Gross. „Tandem-Massenspektrometrie“. In: *Massenspektrometrie: Ein Lehrbuch*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, S. 447–514. ISBN: 978-3-8274-2981-0. DOI: 10.1007/978-3-8274-2981-0_9. URL: https://doi.org/10.1007/978-3-8274-2981-0_9.

[5] Hao Chi u. a. „pNovo: De novo Peptide Sequencing and Identification Using HCD Spectra“. In: *Journal of Proteome Research* 9.5 (2010). PMID: 20329752, S. 2713–2724. DOI: 10.1021/pr100182k. eprint: https://doi.org/10.1021/pr100182k. URL: https://doi.org/10.1021/pr100182k.

[6] James. Yergey u. a. „Isotopic distributions in mass spectra of large molecules“. In: *Analytical Chemistry* 55.2 (1983), S. 353–356. DOI: 10.1021/ac00253a037. eprint: https://doi.org/10.1021/ac00253a037. URL: https://doi.org/10.1021/ac00253a037.

[7] Athula B. Attygalle, Julius Pavlov und Josef Ruzicka. „Monoisotopic Mass?“ In: *Journal of the American Society for Mass Spectrometry* 33.1 (2022). PMID: 34870996, S. 5–10. DOI: 10.1021/jasms.1c00110. eprint: https://doi.org/10.1021/jasms.1c00110. URL: https://doi.org/10.1021/jasms.1c00110.

[8] Lekha Sleno. „The use of mass defect in modern mass spectrometry“. In: *Journal of Mass Spectrometry* 47.2 (2012), S. 226–236. DOI: https://doi.org/10.1002/jms.2953. eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/jms.2953. URL: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/jms.2953.

[9] Kermit K. Murray u. a. „Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013)“. In: *Pure and Applied Chemistry* 85.7 (2013), S. 1515–1609. DOI: doi:10.1351/PAC-REC-06-04-06. URL: https://doi.org/10.1351/PAC-REC-06-04-06.

[10] Hao Chi; Haifeng Chen; Kun He; Long Wu; Bing Yang; Rui-Xiang Sun; Jianyun Liu; Wen-Feng Zeng; Chun-Qing Song; Si-Min He; Meng-Qiu Dong. „pNovo+: De Novo Peptide Sequencing Using Complementary HCD and ETD Tandem Mass Spectra“. In: *Journal of proteome research* (2012), S. 615–625.

[11] Matthias Mann und Ole N. Jensen. „Proteomic analysis of post-translational modifications“. In: *Nature Biotechnology* 21.3 (März 2003), S. 255–261. ISSN: 1546-1696. DOI: 10.1038/nbt0303-255. URL: https://doi.org/10.1038/nbt0303-255.

[12] Yoav Freund u. a. „An efficient boosting algorithm for combining preferences“. In: *Journal of machine learning research* 4.Nov (2003), S. 933–969.

[13] Hao Yang; Hao Chi; Wen-Jing Zhou; Wen-Feng Zeng; Kun He; Chao Liu; Rui-Xiang Sun; Si-Min He. „Open-pNovo: De Novo Peptide Sequencing with Thousands of Protein Modifications“. In: *Journal of proteome research* (2016), S. 645–654.