

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

Dominik Habermann

17. Mai 2023

## 1 Einleitung

Im ersten Kapitel findet zu Beginn eine Erklärung der wichtigsten Begriffe und Abkürzungen statt. Dazu wird eine Themenabgrenzung durchgeführt sowie die Ausgangssituation beschrieben.

## Begriffserklärungen

**De-Novo-Peptidsequenzierung** Eine Methode zur Peptidsequenzierung – also zur Bestimmung der Aminosäuresequenz von **Proteinen**. Mit „De-Novo“ wird ausgedrückt, dass diese Methode nicht auf Datenbanken angewiesen ist, sondern direkt mit den Daten einer **Tandem-Massenspektrometrie** arbeitet.

**ETD** Electron Transfer Dissociation

**HCD** Higher-energy Collisional Dissociation

**Ionisation** Prozess bei dem ein Atom oder ein **Molekül** eine **positive bzw. negative Ladung** annimmt.

**Massenspektrum** Graphische Darstellung der **Masse/Ladungs-Verhältnisse** von den in der Probe enthaltenen Ionen.

**Peak** Englischer Begriff, der mit *Spitze* oder *Höhepunkt* übersetzt werden kann. Er bezeichnet in Massenspektren ein relevantes lokales Maximum. Diese lokalen Maxima stellen ~~u.a.~~ die gesuchten **Moleküle** dar.

**Spektrum-Graph** Aufbereitete Massenspektrumdaten; dargestellt als gerichteter, azyklischer Graph.

**Stereoisomer** Bezeichnung für Chemikalien, die die gleiche Struktur besitzen, aber sich durch die räumliche Positionierung der Atome unterscheiden. Die physikalischen Eigenschaften verschiedener Stereoisomere eines gleichen Stoffes können sich teilweise massiv unterscheiden.

**Tandem-Massenspektrometrie** Analyseverfahren zur Aufklärung von unbekannten biochemischen Strukturen.

## 1.1 Themenabgrenzung

Folgende Aspekte sind Bestandteil dieser Ausarbeitung:

- Was ist die De-Novo-Peptidsequenzierung?
- Was erhofft man sich von dieser Technologie?
- Welche Probleme liegen vor, die von der Seite der Informatik gelöst / verbessert werden können?
- Inwiefern spielen die Spektrums-Graphen dabei eine Rolle?

## 1.2 Ausgangssituation

Mit Hilfe der De-Novo-Peptidsequenzierung ist grundsätzlich die Bestimmung von unbekannten Aminosäuresequenzen möglich. Das Verfahren arbeitet allerdings nicht in jeder Situation zuverlässig genug. Dadurch wird das Ermitteln von unbekannten Sequenzen erschwert. Auch bei bereits bekannten Sequenzen führt die nicht ausreichende Zuverlässigkeit dazu, dass bei Ergebnissen nicht sicher unterschieden werden kann, ob eine Änderung in der Aminosäuresequenz vorliegt oder ob fehlerhafte Daten bestimmt wurden. Das Ziel ist mit Unterstützung von Software eine Möglichkeit bereitzustellen, um die Zuverlässigkeit der De-Novo-Peptidsequenzierung zu erhöhen. Gleichzeitig soll die Implementierung ein effizienteres Werkzeug darstellen als die bereits verfügbaren Ansätze.

# 2 De-Novo-Peptidsequenzierung und Spektrums-Graphen im Detail

In diesem Abschnitt werden die relevanten Herangehensweisen sowohl für die Datengewinnung als auch für deren Auswertung erklärt.

## 2.1 Datengewinnung

Die De-Novo-Peptidsequenzierung nutzt die sogenannte Tandem-Massenspektrometrie für die Bestimmung der Peptidsequenz. Dabei wird die physikalische Eigenschaft ausgenutzt, dass jedes Atom bzw. jedes Molekül – wenn es einer Ionisation unterzogen wurde – ein charakteristisches Massenspektrum besitzt. Das Massenspektrum stellt also eine Art „Fingerabdruck“ eines Moleküls dar und macht dieses ermittelbar.

### 2.1.1 Tandem-Massenspektrometrie bei größeren Molekülen

Bei größeren Molekülen (wie einem Protein) führt die Ionisation dazu, dass das Molekül in kleinere spezifische Ionen zerfällt (sog. Fragmentierung). Die Fragmentierungsinformationen einer De-Novo-Peptidsequenzierung sind meist unvollständig, da fehlende Daten bei einem Fragmentierungsschritt die Güte des Endergebnisses negativ beeinflusst. Dies wird insbesondere dann ein Problem, wenn unbekannte Änderungen in einer Peptidsequenz vorhanden sind.

Um dieses Problem zu verringern können unterschiedliche Techniken parallel eingesetzt werden, welche verschiedene Fragmente erzeugen und daher auch verschiedenartige Massenspektren zur Folge haben.<sup>1</sup>

## 2.2 Datenaufbereitung

Typischerweise betrachtet man die sog. „Peaks“ in den Massenspektren. Jeder Peak stellt ein unterschiedliches Ion dar. Dazu kommen Messungenauigkeiten sowie Hintergrundrauschen. Durch die hohe Anzahl an möglichen Ionen kann nicht ohne weiteres differenziert werden, welcher der Peaks von welchen Ionen erzeugt wurden und welche nicht.

Der Algorithmus für die Datenaufbereitung berechnet den natürlichen Logarithmus von den Intensitäten der Peaks, um Hintergrundrauschen und Messungenauigkeiten nicht überzupriorisieren. Zusätzlich dazu werden Peaks, die in einem Toleranzbereich nebeneinander liegen, zusammengefasst. Am Ende werden die Peaks entfernt, bei denen bekannt ist, dass es sich nicht um relevante Ionen handeln kann. (z.B. Peaks von Isotopen)

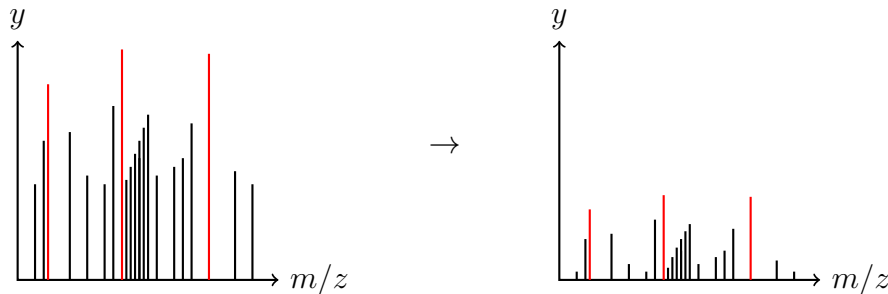


Abbildung 1: Anwendung des  $\ln$  auf Rohdaten. Rote Peaks stellen hier exemplarisch fehlerhafte Daten dar, die nach dem  $\ln$  reduziert wurden.

<sup>1</sup>Konkret: Es wird sowohl das Higher-energy Collisional Dissociation (HCD) als auch das Electron Transfer Dissociation (ETD) Verfahren angewendet.

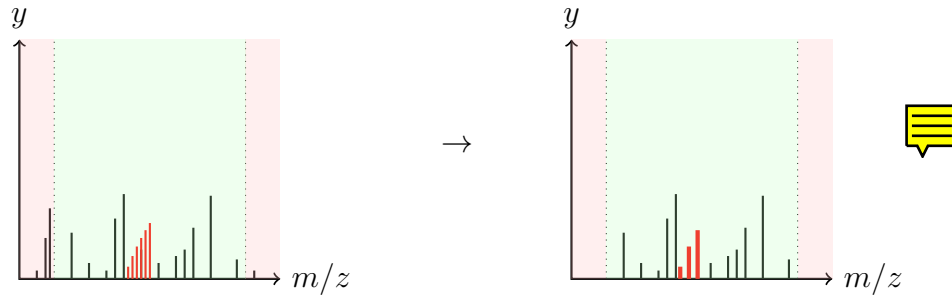


Abbildung 2: Entfernen von irrelevanten Peaks sowie zusammenfassen naheliegender Peaks. Hier symbolisieren die roten Peaks jene, die zusammengefasst werden.

## 2.3 Konvertierung von Massenspektren

Das Ziel der Konvertierung ist das Erzeugen eines Spektrum-Graphen. Um von einem Massenspektrum zu einem Spektrum-Graphen zu kommen, werden die Peaks, die nach der Datenaufbereitung (Siehe ...) übrig bleiben, als **Knoten** gewertet. Dazu kommt ein **Start- und Endknoten**. Jeder Knoten bekommt eine Gewichtung; diese Gewichtung entspricht der Stärke des Peaks.

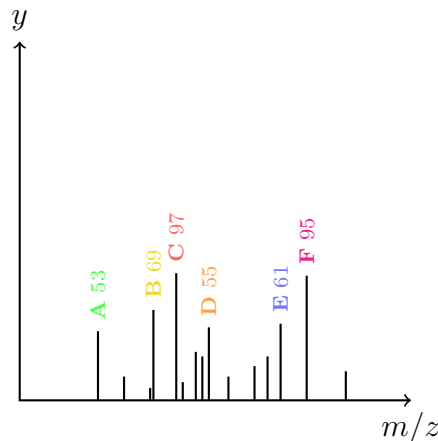


Abbildung 3: Ausgewählte Peaks mit einem exemplarischen **x** Wert.

Gerichtete Kanten zwischen den Knoten werden **ausgebildet**, wenn diese eine Differenz von **genau einer** oder zwei **Aminosäurereste**<sup>2</sup> besitzen. Der Einfachheit halber wird im folgenden eine Kante ausgebildet, wenn die Differenz genau **4** beträgt.

<sup>2</sup>Da eine Aminosäure **vielerlei an Reste besitzen kann**, ergeben sich mehr als 40 Differenzen, die diese Bedingung erfüllen.

$(u, v)$	$u$	$v$	$\Delta(u, v)$	$\Delta(u, v) \bmod 4$
A (A,B)	53	69	16	0 ✓
A (A,C)	53	97	44	0 ✓
B (A,D)	53	55	2	2 ✗
A (A,E)	53	61	8	0 ✓
B (A,F)	53	95	42	2 ✗
A (B,C)	69	97	28	0 ✓
B (B,D)	69	55	14	2 ✗
A (B,E)	69	61	8	0 ✓
B (B,F)	69	95	26	2 ✗
B (C,D)	97	55	42	2 ✗
A (C,E)	97	61	36	0 ✓
B (C,F)	97	95	2	2 ✗
B (D,E)	55	61	6	2 ✗
A (D,F)	55	95	40	0 ✓
B (E,F)	61	95	34	2 ✗

Tabelle 1: Bestimmung der Kanten

Darstellung der Daten **als** gewichteter, gerichteter azyklischer Graph. **Zusätzlich benötigt der Graph noch separate Start- und Zielknoten; diese sind für die späteren Berechnungen unerlässlich.**

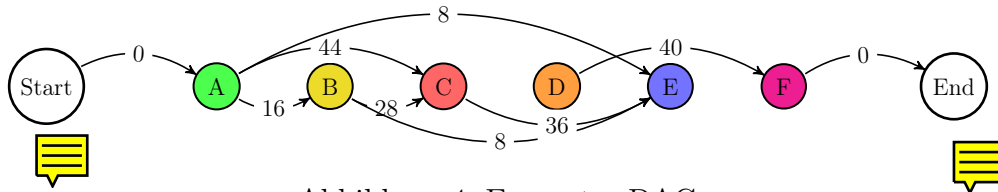



Abbildung 4: Erzeugter DAG

Bereits an diesem Minimalbeispiel ist zu erkennen, dass die gebildeten Knoten in einem Spektrums-Graphen nur wenige ausgehende Kanten besitzen. **Dies ist nicht dem Beispiel geschuldet sondern ist tatsächlich auch in der Praxis der Regelfall.** Dies ist eine hilfreiche Beobachtung für die Datenauswertung (siehe Abschnitt 2.4 „Datenauswertung“).

## 2.4 Datenauswertung

Um ~~mit~~ aus dem Graphen die Peptidsequenz zu gewinnen müssen **alle längsten Pfade** im DAG gefunden werden. Da die Kanten gewichtet sind, kann es durchaus mehrere längste Pfade geben. Gleichwohl es Algorithmen für das Problem des längsten Pfades in einem Graphen gibt, handelt es sich **hierbei um ein NP-schweres Problem.** Es existiert also (wahrscheinlich) **kein effizienter Algorithmus.** Erschwerend kommt hinzu, **dass der Graph nicht zwingend ein zusammenhängender Graph sein muss – auch wenn dies meist der Fall ist.** Der Graph muss daher vor Berechnungsbeginn auf diese Eigenschaft hin überprüft werden.

Im Falle der Spektrums-Graphen existiert die Eigenschaft, dass solche Graphen meist eine geringe Dichte an Kanten aufweisen. Dies hat den positiven Effekt, dass die Anzahl an überhaupt möglichen längsten Pfaden recht gering ist. Zusätzlich dazu kann die Warteschlange, die in den longest Path DAG Algorithmen verwendet werden, angepasst werden. Da die Gewichtung der Kanten als eine Art „Wahrscheinlichkeit“, dass die nächste Kante die reale Peptidsequenz darstellt, interpretiert werden kann, kann eine priorisierte Warteschlange verwendet werden, die die Laufzeit ebenfalls verbessert. In Summe führen diese Eigenschaften der Spektrums-Graphen dazu, dass das längste Pfade Problem in solchen Fällen auf die Laufzeit  $\mathcal{O}(abs(E) + \log(d))$  reduziert werden kann.

Zusammengefasst: Es wird versucht die speziellen Eigenschaften der Graphen auszunutzen, um die Laufzeit zu verbessern. 

## 3 Ergebnisse/Evaluierung

Im folgenden Kapitel werden die Probleme, die in der Praxis bei der Verwendung des Verfahrens auftreten, erläutert und mögliche Lösungsansätze aufgezeigt.

### 3.1 Probleme in der Praxis

#### 3.1.1 Qualität der Messwerte

Obwohl eine Datenaufbereitung stattfindet, ist das Verfahren bei der Verwendung von Spektrums-Graphen stark auf die Genauigkeit der Messwerte angewiesen. Zwar sind durch technische Fortschritte bei der Tandem-Massenspektrometrie die Daten hochwertiger geworden; dennoch gestaltet sich das Sequenzieren von ~~unbekannten Peptidsequenzen als~~ schwierig. Mit heutigen Gerätschaften lassen sich bei der Verwendung des genannten Verfahrens bis zu 13 Peptide mit einer durchschnittlichen Genauigkeit von 94% ermitteln. Danach nimmt diese sprunghaft ab. Für brauchbare Ergebnisse wird – je nach Literatur – eine Trefferquote von 90-95% vorausgesetzt.

#### 3.1.2 Fehlende Betrachtung der Stereoisomerie

Das komplette Verfahren basiert auf das Masse-Ladungs-Verhältnis, sodass Stereoinformationen schlicht nicht ermittelt werden können. Es kann zwar mithilfe einer energetischen Betrachtung bestimmt werden welche Stereoisomere in welchen Verhältnis auftreten (müssen). Dabei handelt es sich allerdings lediglich um eine grobe Abschätzung.

#### 3.1.3 Identifikation der Aminosäuren über Massendifferenz

Die Grundidee bei der Identifikation von Aminosäuren ist die Betrachtung der Massendifferenzen zwischen zwei Peaks. Zwar liefert dieser Ansatz häufig passende Ergebnisse. Dennoch ist solch eine Differenz nicht in der Lage jede Aminosäure immer eindeutig zu identifizieren, da bestimmte Kombinationen (fast) gleiche Differenzen besitzen. Der Algorithmus, der die

Gewichtungen bestimmt, arbeitet nur mit ganzzahligen Werten. Dadurch gehen leichte Unterschiede, die durch die Isotope (insb. die des Kohlenstoffes) begründet sind, meist durch die Float Integer Konvertierung verloren.

## 3.2 Lösungsansätze

### 3.2.1 Verbesserung der Ergebnisse durch Machine Learning

Bei der Sequenzierung werden ab einer gewissen Länge unweigerlich Fehler eintreten.[1, S.621, Figure 5] Dadurch, dass nicht jede Peptidsequenz gleich wahrscheinlich ist<sup>3</sup>, können mittels Machine Learning grundsätzlich die Ergebnisse verbessert werden. insbesondere dann, wenn die ermittelte Differenz keinen eindeutigen Rückschluss auf die Aminosäure zulässt.

## 4 Zusammenfassung

Im letzten Kapitel werden die ungelösten Probleme genannt und erklärt warum diese eine Relevanz für die Praxis haben. Am Ende findet eine kritische Betrachtung des Verfahrens im allgemeinen statt.

### 4.1 Ungelöste Probleme

Wie bereits in 3.1.2 erwähnt, kann das Verfahren designbedingt keine Stereoinformationen ermitteln. Daher ist es in diesem Fall besonders wichtig abzuschätzen, ob das Fehlen dieser Informationen tatsächlich eine Relevanz hat. Wenn nur die Peptidsequenz betrachtet werden soll, dann stellt dies kein Problem dar. Aber sobald jedweige Abschätzungen anhand der ermittelten Sequenz stattfinden soll, dann kann das Fehlen jener Informationen zu massiven Fehlern führen.

Wenn für die Verbesserung der Ergebnisse Machine Learning in Betracht kommt, dann muss dabei berücksichtigt werden, dass dadurch unter Umständen einer der großen Vorteile der De-Novo-Peptidsequenzierung verloren geht – und zwar dass keine Vorinformationen für die Sequenzierung notwendig sind. Hierbei kommt es auf den konkreten Anwendungsfall an, ob das Verlieren dieser Eigenschaft eine Bedeutung besitzt.

### 4.2 Kritische Betrachtung

Die De-Novo-Peptidsequenzierung mit der Unterstützung von Spektrums-Graphen stellt eine Möglichkeit dar ~~Poly~~peptide mit bis zu einer Länge von etwa 12 Peptiden ausreichend zuverlässig zu bestimmen. Die Autoren des Papers [2] haben die Software frei zur Verfügung gestellt, sodass sie in jedem Fall ein Blick wert ist. Gegenüber anderen Ansätzen ist das Verfahren zwar konkurrenzfähig, allerdings nicht immer die beste Wahl [2, S. 650]. Die

---

<sup>3</sup>Dies ist u.a. dadurch begründet, dass die Reste der Aminosäuren sich gegenseitig beeinflussen (können), sodass bestimmte Sequenzen energetisch ungünstig sind und lediglich vermindert auftreten.

Grundidee mittels der Massendifferenz auf die Aminosäuren zu schließen wird nie fehlerfrei sein, sodass dieses Verfahren weniger die bereits vorhandenen Systeme ersetzen kann, sondern eher ein weiteres Werkzeug für die De-Novo-Peptidsequenzierung darstellt.

## Literatur

- [1] 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.
- [2] 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.

