

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

14. Juli 2023

Dominik Habermann

Ruhr Universität Bochum

Aminosäure Sequenzierung

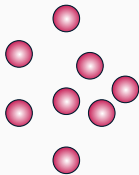
De-Novo-Sequenzierung

pNovo+ Algorithmus

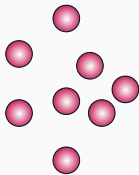
Open-pNovo Algorithmus

Zusammenfassung

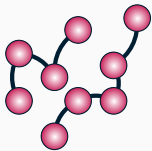
Aminosäure Sequenzierung



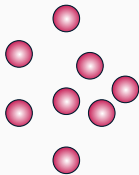
Aminosäure (AA)



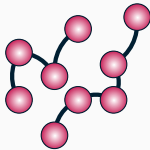
Aminosäure (AA)



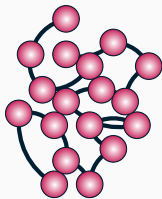
Peptid



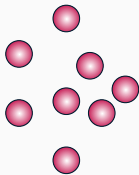
Aminosäure (AA)



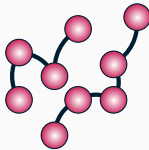
Peptid



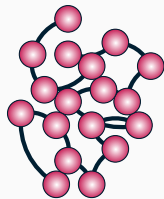
Protein



Aminosäure (AA)



Peptid



Protein

- Peptid $\hat{=}$ Kurze Ketten an AA
- Protein $\hat{=}$ Verkettung von Peptiden

- Ziel: Bestimmung der AA-Sequenz von Peptiden

- Ziel: Bestimmung der AA-Sequenz von Peptiden
- Warum ist die AA-Sequenz relevant?

- Reihenfolge der AA hat unter anderem Einfluss auf:

- Reihenfolge der AA hat unter anderem Einfluss auf:
 - 3D Aufbau eines Proteins
 - Funktionsweise
 - Fähigkeiten
 - Notwendigen Umgebungsbedingungen (Temperatur, pH-Wert, etc.)

- Reihenfolge der AA hat unter anderem Einfluss auf:
 - 3D Aufbau eines Proteins
 - Funktionsweise
 - Fähigkeiten
 - Notwendigen Umgebungsbedingungen (Temperatur, pH-Wert, etc.)
- \Rightarrow AA-Sequenz ist von wesentlicher Bedeutung

- Biomedizinische Relevanz:

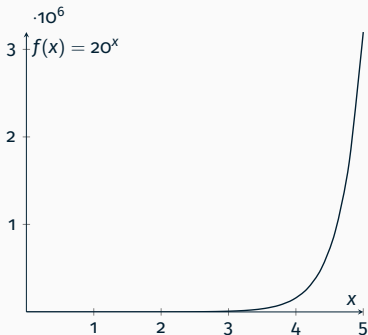
- Biomedizinische Relevanz:
 - Katalogisierung von Proteinen
 - Analyse von Enzymen
 - Toxikologie von Proteinen

- Biomedizinische Relevanz:
 - Katalogisierung von Proteinen
 - Analyse von Enzymen
 - Toxikologie von Proteinen
- Zuverlässige Sequenzierung möglich?

- 20 relevante AA
- Weitestgehend beliebig kombinierbar
- Bereits bei wenigen AA: Kaum händelbarer Suchraum

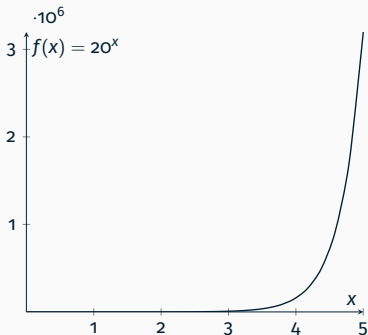
AA Sequenzierung – Suchraum

- 20 relevante AA
- Weitestgehend beliebig kombinierbar
- Bereits bei wenigen AA: Kaum händelbarer Suchraum



AA Sequenzierung – Suchraum

- 20 relevante AA
- Weitestgehend beliebig kombinierbar
- Bereits bei wenigen AA: Kaum händelbarer Suchraum



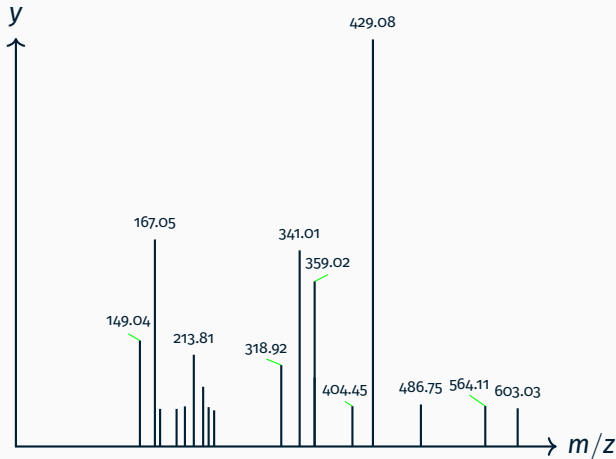
- Zum Vergleich: Proteine bis zu mehreren *zehntausend* AA

- \Rightarrow Intelligentes Sequenzierungsverfahren notwendig

- \Rightarrow Intelligentes Sequenzierungsverfahren notwendig
- Hilfsmittel: Massenspektrometrie (MS)
- MS kann chemische Strukturen bestimmen
- Rückschluss auf die AA-Sequenz möglich

- MS erzeugt MS-Spektren

- MS erzeugt MS-Spektren
- Beispiel: vereinfachte Darstellung von realen Messwerten



De-Novo-Sequenzierung

Sequenzierung und De-Novo-Sequenzierung

Suche in Sequenzdatenbank	De-Novo-Sequenzierung
Datenbanken als Hilfsmittel	Ohne weitere Hilfsmittel
Identifizierung von <i>bekannten</i> Sequenzen	Bestimmung <i>unbekannter</i> Sequenzen

Sequenzierung und De-Novo-Sequenzierung


Suche in Sequenzdatenbank	De-Novo-Sequenzierung
Datenbanken als Hilfsmittel	Ohne weitere Hilfsmittel
Identifizierung von <i>bekannten</i> Sequenzen	Bestimmung <i>unbekannter</i> Sequenzen

- De novo: lat. „Von neuem“

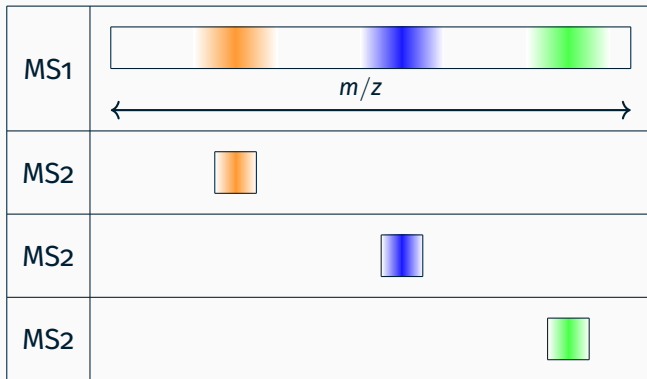
- Zusätzliche Informationen notwendig

- Zusätzliche Informationen notwendig
- Verwendung einer 2. MS
- Verfahren: Tandem-Massenspektrometrie MS2

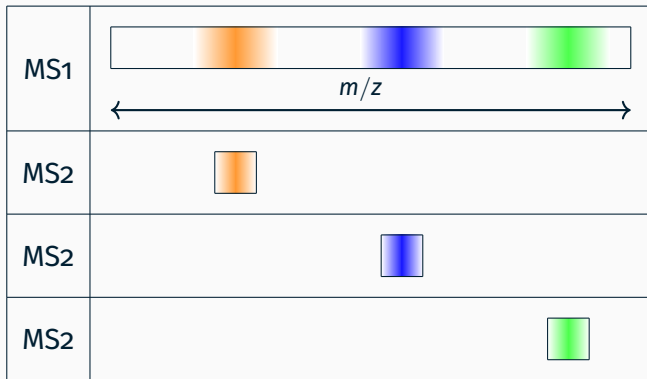
De-Novo-Sequenzierung – MS2

MS1	 <p>The diagram shows a horizontal rectangular box representing a mass spectrum. Inside the box, there are three distinct colored regions: an orange region on the left, a blue region in the middle, and a green region on the right. Below the box, a horizontal double-headed arrow spans the width of the box, with the label m/z centered above it.</p>
MS2	
MS2	
MS2	

De-Novo-Sequenzierung – MS2



De-Novo-Sequenzierung – MS2

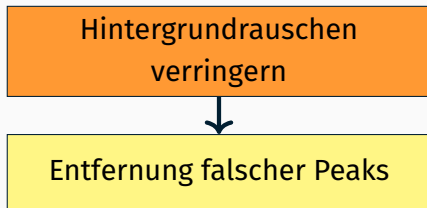


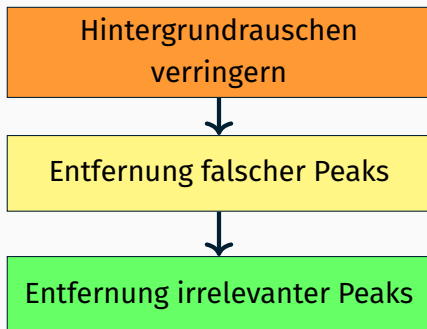
- MS1 quasi eine Filterung
- MS2 wird gezielt auf m/z Intervalle angewendet

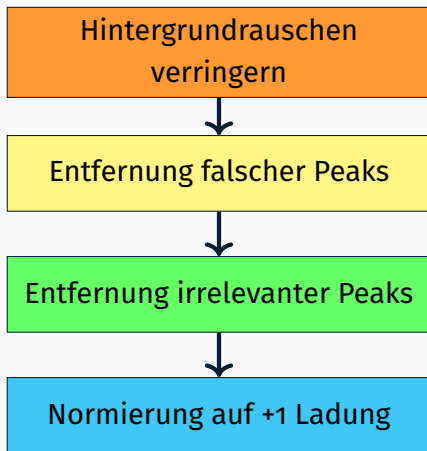
pNovo+ Algorithmus

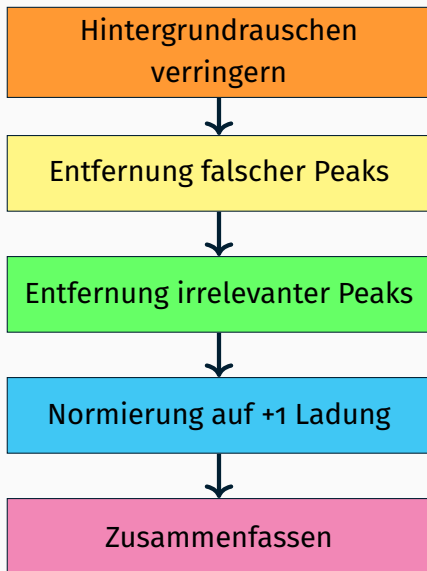
- Erweiterung von pNovo
- Algorithmus für die De-Novo-Sequenzierung
- Auswertung von MS2-Spektren
- Rekonstruktion der AA-Sequenz
- Hilfsmittel: Spektrum-Graph

Hintergrundrauschen
verringern

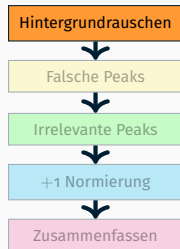




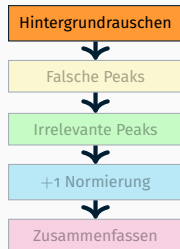
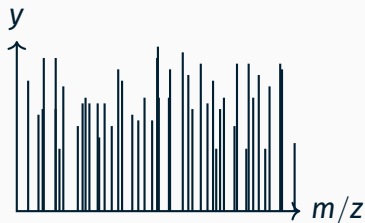




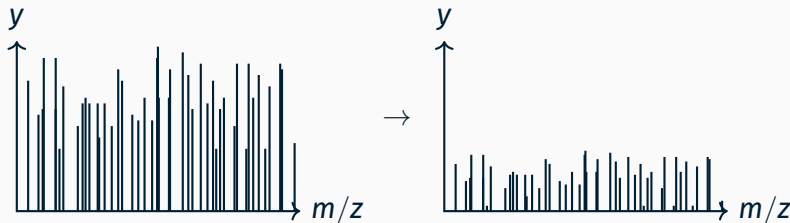
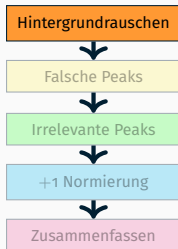
- Überpriorisierung fehlerhafter Daten vermeiden
- Tool: $\ln()$



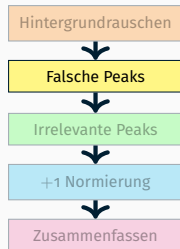
- Überpriorisierung fehlerhafter Daten vermeiden
- Tool: $\ln()$



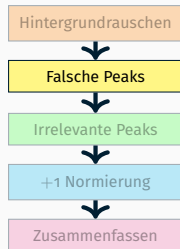
- Überpriorisierung fehlerhafter Daten vermeiden
- Tool: $\ln()$



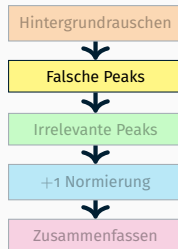
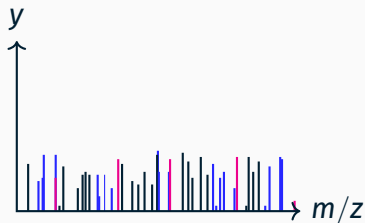
- Monoisotopische Peaks auswählen



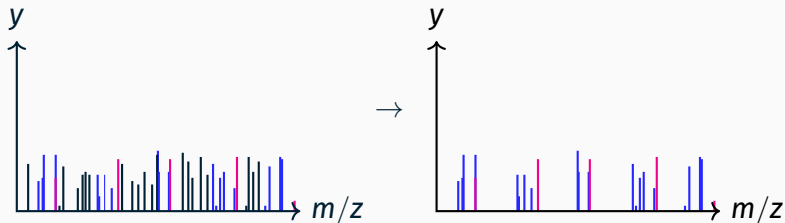
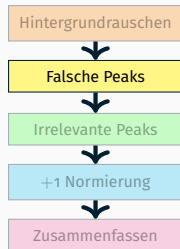
- Monoisotopische Peaks auswählen
- Peaks mit definierten Abstand auswählen



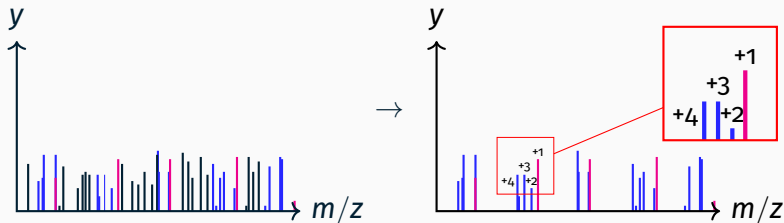
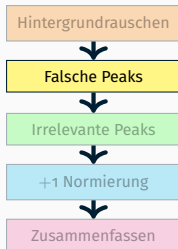
- **Monoisotopische Peaks** auswählen
- Peaks mit definierten **Abstand** auswählen



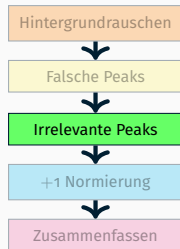
- **Monoisotopische Peaks** auswählen
- Peaks mit definierten **Abstand** auswählen



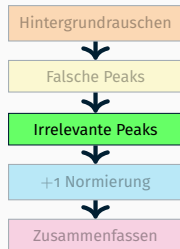
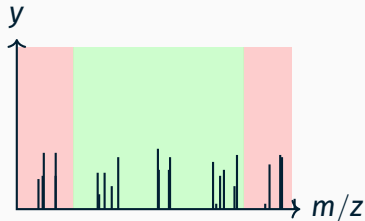
- **Monoisotopische Peaks** auswählen
- Peaks mit definierten **Abstand** auswählen
- „Charge state“ Zuweisung:
 - **Monoisotopischer Peak**: +1
 - **Abstand**: steigend mit +1 pro Schritt



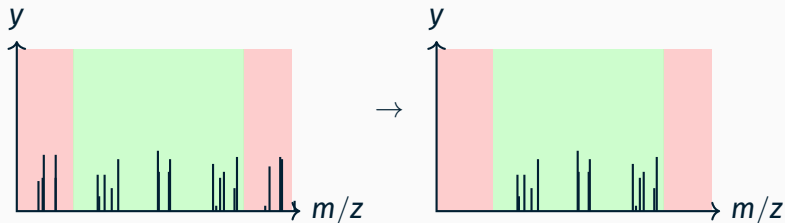
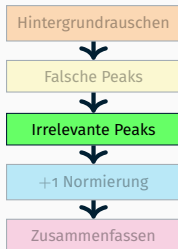
- Peaks aus irrelevantem Intervall entfernen
- m/z Bereiche, die garantiert unwichtig sind



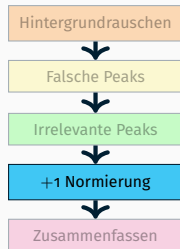
- Peaks aus irrelevantem Intervall entfernen
- m/z Bereiche, die garantiert unwichtig sind



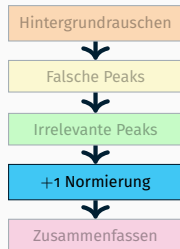
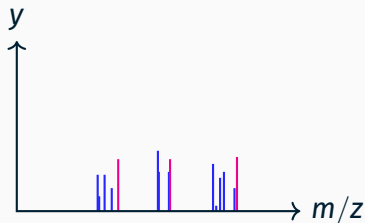
- Peaks aus irrelevantem Intervall entfernen
- m/z Bereiche, die garantiert unwichtig sind



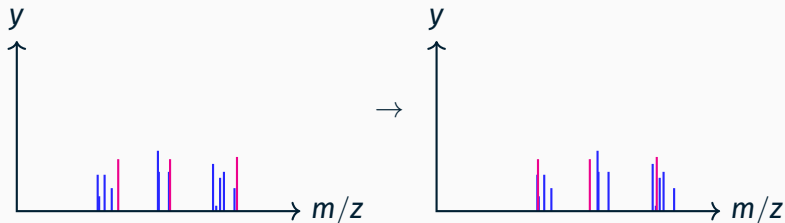
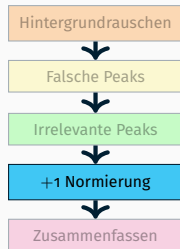
- Peaks auf Charge state +1 normieren
- \Rightarrow Verschiebung nach rechts auf m/z Achse



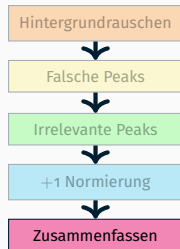
- Peaks auf Charge state +1 normieren
- \Rightarrow Verschiebung nach rechts auf m/z Achse



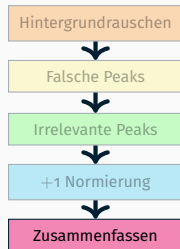
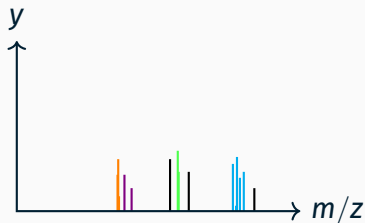
- Peaks auf Charge state +1 normieren
- \Rightarrow Verschiebung nach rechts auf m/z Achse



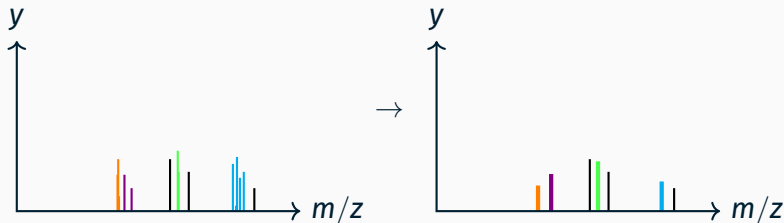
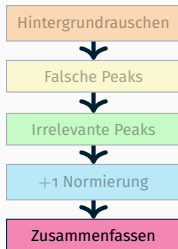
- Zusammenfassen von Peaks
- Abstand einen **Schwellwert** unterschreitet
- $y = \text{Median}$ aus zusammengefassten Peaks

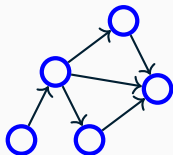
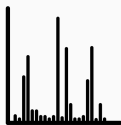


- Zusammenfassen von Peaks
- Abstand einen **Schwellwert** unterschreitet
- $y = \text{Median}$ aus zusammengefassten Peaks

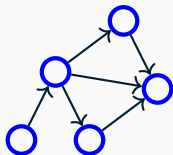
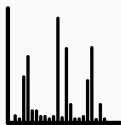


- Zusammenfassen von Peaks
- Abstand einen **Schwellwert** unterschreitet
- $y = \text{Median}$ aus zusammengefassten Peaks

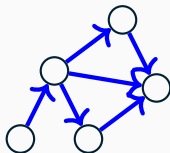
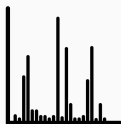




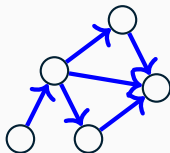
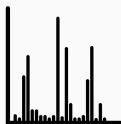
- Verwendung vorverarbeiteter MS2 Spektren
- Peaks \rightarrow Knoten:



- Verwendung vorverarbeiteter MS2 Spektren
- Peaks \rightarrow Knoten:
 - Peaks $\hat{=}$ Knotenpaar
 - Knoten bekommen eine „Masse“
 - Masse $\hat{=}$ m/z Wert
 - Startknoten (Masse = 0)
 - Endknoten (Masse = Hauptpeak - $M(\text{H}_2\text{O})$)

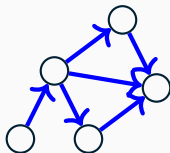
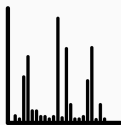


- Gerichtete Kanten zwischen Knotenpaar, wenn:



■ Gerichtete Kanten zwischen Knotenpaar, wenn:

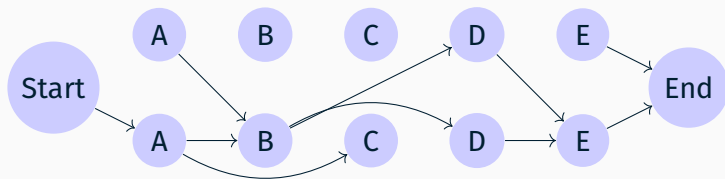
- Massendifferenz genau Masse **einer** AA entspricht
- Massendifferenz genau Masse **zwei** AA entsprechen



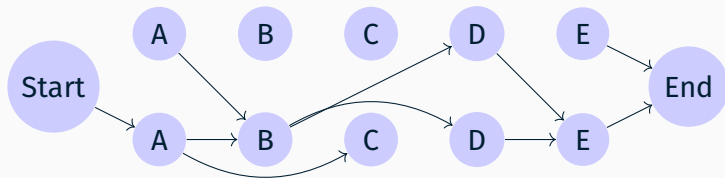
- Gerichtete Kanten zwischen Knotenpaar, wenn:
 - Massendifferenz genau Masse **einer** AA entspricht
 - Massendifferenz genau Masse **zwei** AA entsprechen
- $N + \binom{n+N-1}{N-1}$ Differenzen ($n = 2, N = 20$)
- **230** Differenzen

- Ergebnis: Directed acyclic graph (DAG)

- Ergebnis: Directed acyclic graph (DAG)
- Beispiel eines Spektrumsgraphen:



- Ergebnis: Directed acyclic graph (DAG)
- Beispiel eines Spektrumsgraphen:



- Alle möglichen Pfade von Start nach End ermitteln
- Scoring Funktion „bewertet“ jeden Pfad
- Pfad mit dem höchsten Scoring Wert ist das Ergebnis



- 8677 Datensätze
- Erfolgreiche Sequenzierungen: 81,2%
- Alternativalgorithmus (PEAKS): 71.8%



- 8677 Datensätze
- Erfolgreiche Sequenzierungen: 81,2%
- Alternativalgorithmus (PEAKS): 71.8%
- pNovo+ besser als die Konkurrenz!
- Side Note: pNovo+ ist frei verfügbar

Open-pNovo Algorithmus

- Peptide sind nicht zwingend stabil
- Wechselwirkungen können die Sequenz abändern
- Posttranslationale Proteinmodifikationen (PTM)

- Peptide sind nicht zwingend stabil
- Wechselwirkungen können die Sequenz abändern
- Posttranslationale Proteinmodifikationen (PTM)
- Mit De-Novo-Algorithmen an sich kein Problem ...

- Peptide sind nicht zwingend stabil
- Wechselwirkungen können die Sequenz abändern
- Posttranslationale Proteinmodifikationen (PTM)
- Mit De-Novo-Algorithmen an sich kein Problem ...
- ... wenn nach der Änderung eine AA zurückbleibt

- Bildung von nicht proteinogenen AA möglich
- AA, die normalerweise nicht in Peptiden vorkommen

- Bildung von nicht proteinogenen AA möglich
- AA, die normalerweise nicht in Peptiden vorkommen
- Spektrogramm zeigt solche AA
- Änderungen können von pNovo+ nicht erkannt werden

- Bildung von nicht proteinogenen AA möglich
- AA, die normalerweise nicht in Peptiden vorkommen
- Spektrogramm zeigt solche AA
- Änderungen können von pNovo+ nicht erkannt werden
- \Rightarrow pNovo+ erzeugt zwangsweise Fehler

- Neue Scoring Funktion: RankBoost
- Machine Learning Algorithmus aus 2003
- Erweiterung des AdaBoost Algorithmus
- Präferenzen in Datensätzen zu erkennen

- Neue Scoring Funktion: RankBoost
- Machine Learning Algorithmus aus 2003
- Erweiterung des AdaBoost Algorithmus
- Präferenzen in Datensätzen zu erkennen
- Filterung der nicht gültigen AA

- 45450 Datensätze





- 45450 Datensätze
- Erfolgreiche Sequenzierungen: 76,3%
- pNovo+: 74,5%
- Alternativalgorithmus PEAKS: 73,1%
- 2. Alternativalgorithmus Novor: 39,9%



- 45450 Datensätze
- Erfolgreiche Sequenzierungen: 76,3%
- pNovo+: 74,5%
- Alternativalgorithmus PEAKS: 73,1%
- 2. Alternativalgorithmus Novor: 39,9%
- Verbesserung im Vergleich zu pNovo+



- 45450 Datensätze
- Erfolgreiche Sequenzierungen: 76,3%
- pNovo+: 74,5%
- Alternativalgorithmus PEAKS: 73,1%
- 2. Alternativalgorithmus Novor: 39,9%

- Verbesserung im Vergleich zu pNovo+
- Allerdings: Weniger als 2% Punkte

Zusammenfassung

- De-Novo-Sequenzierung leichter durchführbar
- Beide Algorithmen liefern **sehr gute** Ergebnisse
- pNovo+ Ansatz mit Spektrumsgraphen ist wirkungsvoll
- Open-pNovo erkennt zuverlässig Proben mit PTMs

- De-Novo-Sequenzierung leichter durchführbar
- Beide Algorithmen liefern **sehr gute** Ergebnisse
- pNovo+ Ansatz mit Spektrumsgraphen ist wirkungsvoll
- Open-pNovo erkennt zuverlässig Proben mit PTMs
- Dennoch: hoher Optimierungsbedarf besteht weiterhin

Danke für die Aufmerksamkeit :)

Fragen?

Danke für die Aufmerksamkeit :)

Fragen?

