Abstract

An abstract abstracts the thesis...

Zusammenfassung

Eine Zusammenfassung fasst die Arbeit zusammen...

Contents

1 Background

In this chapter, the fundamental principles of mass spectrometry-based proteomics and the computational strategies for peptide identification are discussed. Particular focus is placed on the challenges introduced by chemical labeling and the emergence of deep learning models in *de novo* sequencing.

1.1 Mass Spectrometry-Based Proteomics

1.1.1 Bottom-Up Proteomics Workflow

Mass spectrometry (MS)-based proteomics is the gold standard for large-scale protein analysis. The "bottom-up" approach is the most widely adopted strategy, where proteins are extracted and enzymatically digested—typically using trypsin—into smaller peptides before analysis **Gevaert2003**. This is essential as peptides are more easily fractionated and ionized than intact proteins. The resulting mixture is separated via liquid chromatography (LC) and ionized using Electrospray Ionization (ESI) **Aebersold2016**.

1.1.2 Tandem Mass Spectrometry (MS/MS) and Peptide Fragment Ion Theory

Peptide sequences are identified using Tandem Mass Spectrometry (MS/MS). A precursor ion is isolated by its mass-to-charge ratio (m/z) and fragmented, often using Higher-energy Collisional Dissociation (HCD) **Steen2004**.

According to fragment ion theory, the peptide backbone fragments primarily at amide bonds, resulting in b-ions (N-terminal) and y-ions (C-terminal) **Roepstorff2010**. The mass difference between consecutive ions in a series corresponds to specific amino acids. However, post-translational modifications (PTMs) or chemical labels like Tandem Mass Tags (TMT) shift these masses, requiring advanced computational identification.

1.2 Peptide Identification Strategies

1.2.1 Database Search Engines (DBIS)

The most common identification method is database searching. Engines like MaxQuant or SEQUEST compare experimental MS/MS spectra against *in silico* digested sequences from databases like UniProt **Cox2008**; **Eng1994**. While robust, DBIS is limited by the "search space" problem: it cannot identify modifications not explicitly included in the database, leading to missed novel PTMs **Nesvizhskii2010**.

1.2.2 Principles of De Novo Peptide Sequencing

In contrast, *de novo* sequencing reconstructs sequences directly from fragment ion peaks without a reference database **Taylor1997**. While historically limited by noise, modern Transformer-based models now capture long-range dependencies between ions, making this approach ideal for discovering novel PTMs and variants in the "dark proteome" **Tran2019**.

1.3 Tandem Mass Tag (TMT) Labeling

1.3.1 Isobaric Labeling Chemistry

Tandem Mass Tag (TMT) labeling is used for high-throughput multiplexed quantification. TMT tags are isobaric, meaning labeled peptides from different samples appear as a single peak in MS1 scans, reducing instrument time and missing values **Thompson2003**; **Werner2014**.

1.3.2 Impact on Fragmentation Patterns

TMT tags introduce systematic mass shifts to the N-terminus and lysine side chains. Upon fragmentation, they release reporter ions (m/z 126–135) for quantification **McAlister2014**. For *de novo* sequencing, these tags are challenging because they alter fragmentation efficiency and shift b- and y-ion series significantly **Hogrebe2018**.

1.4 Deep Learning and Transformer Models

The identification of peptides is a sequence-to-sequence (Seq2Seq) task. While early models used LSTMs **Tran2019**, the Transformer architecture revolutionized the field with the Self-Attention mechanism **Vaswani2017**.

In a proteomic context, the Transformer's encoder extracts structural features from continuous m/z and intensity values through point-based encoding **Yilmaz2023**. The decoder then uses Beam Search to maintain a set of the k most likely sequences, ensuring the final result is globally consistent with the precursor mass **Qiao2021**.

2 Materials and Methods

3 Results and Discussion

4 Conclusion