Abstract

An abstract abstracts the thesis...

Zusammenfassung

Eine Zusammenfassung fasst die Arbeit zusammen...

Contents

1 Background

The proteome is the entire set of proteins that is, or can be, expressed by a cell, tissue, or organism at a particular time and under a specific set of conditions. Unlike the genome, which is largely static, the proteome is highly dynamic, reflecting the cell's current functional state and its response to internal and external cues **Aebersold2016**. Proteins are the primary molecular machines that execute virtually all biological processes, including enzyme catalysis, cell signaling, and structural support **Alberts2017**.To significantly increase the functional diversity beyond the genetic code, proteins undergo Post-Translational Modifications (PTMs), which are chemical alterations that occur after a protein has been synthesized. PTMs are critical regulatory mechanisms, with over 400 types known, including phosphorylation, acetylation, and ubiquitination **Khoury2011**. These modifications act as molecular switches, profoundly influencing cellular processes by regulating a protein's activity (e.g., activating or deactivating enzymes), subcellular localization, stability, and interaction with other molecules **Hunter2012**; **Walsh2006**. Consequently, the accurate characterization of PTMs is essential for understanding cell biology and disease states.The standard method for large-scale protein analysis is Mass Spectrometry (MS)-based proteomics, specifically the bottom-up approach. In this workflow, the complex protein mixture is first enzymatically cleaved into smaller fragments called peptides. These peptides are then separated by liquid chromatography (LC) and introduced into the mass spectrometer. The MS instrument performs two main steps: in MS1, it measures the mass-to-charge ratio (m/z) of the intact peptide (the precursor ion); then, in MS/MS (MS2), it isolates a precursor ion, fragments it (e.g., via HCD), and measures the m/z of the resulting fragment ions. These fragment ion masses contain the sequence information of the peptide **Domon2015**.The most common computational method to identify the peptide sequence from the MS/MS spectrum is database searching **Yates1995**. A search algorithm compares the experimentally measured fragment ion spectrum against a comprehensive sequence database (like a FASTA file of all known proteins for an organism). The algorithm calculates the theoretical fragment masses for every possible peptide in the database, including predefined PTMs, and assigns the best-matching sequence (Peptide-Spectrum Match, PSM) by scoring the match quality **Mann2001**.However, this databasedependent strategy faces a critical limitation when dealing with PTMs. To identify a PTM, the modification must be specified in advance as a variable modification in the search. Adding even a small number of variable PTMs dramatically increases the size of the search space, leading to a combinatorial explosion of possible peptide sequences that must be tested **Tanner2005**. This greatly increases the computation time and, more critically, raises the False Discovery Rate (FDR), making confident identification difficult **Chalkley2019**. Furthermore, the method is inherently incapable of identifying novel or unexpected PTMs that are not included in the predefined search parameters, thereby limiting the scope of biological discovery.

De novo peptide sequencing (DNPS) refers to the approach of inferring the amino-acid sequence of a peptide directly from its tandem mass spectrum, without relying exclusively on a reference protein database. The advantage of DNPS is that it can detect peptides that are not present in the database — for example from sequence variants, novel splice forms, or unexpected post-translational modifications (PTMs). Because typical database search methods assume a predefined set of possible peptides, they may miss those outside that set; DNPS fills that gap. For instance, the model Casanovo uses a transformer architecture to translate spectral peak sequences into peptide sequences and thereby improves on standard DNPS accuracy [Yilmaz22] [CasanoDoc].

When we turn to "state-of-the-art" (SOTA) methods for MS-based proteomics, there are a few widely used software tools worth mentioning. First, MaxQuant (which integrates the search engine Andromeda) performs database-search based identification and quantification of peptides and proteins from high-resolution MS data; it supports labeling (e.g., TMT, SILAC) and has advanced calibration and scoring algorithms [MaxQuantGuide] [Andromeda2011]. Second, MSFragger offers ultra-fast database searches and open modification searches (i.e., tolerance for large mass shifts) so it is particularly powerful for identifying modified peptides under complex search settings [MSFraggerSite] [MSFraggerMod]. Third, while not strictly a DNPS tool, Casanovo (as already noted) represents the cutting edge in DNPS and shows how the field is moving towards machine learning approaches that complement or even go beyond classical database searches [Yilmaz22].

Together, these tools illustrate the two complementary paradigms in proteomics: database-search methods (e.g., MaxQuant/Andromeda, MSFragger) that match spectra to known peptide sequences, and de-novo methods (e.g., Casanovo) that infer sequences without prior database constraints.

In proteomics, isobaric labelling such as Tandem Mass Tag (TMT) allows multiple samples to be combined and analysed in a single run, increasing throughput by enabling parallel identification and quantification of peptides and pro-

teins across conditions [Lee$_{T}hermo23][Chen21].TMTreagentsconsistofareactivegroupthatlabelspept$
$terminusandlysinesidechains), amass-normaliserregiontokeepalltaggedpeptidesisobaric, andarepor$
$plex)[ThermoTMTSystems][QuantGuide21].Becausealllabelledpeptideshavethesamenominalmass, t$
$eluteandareselectedtogetherinMS1, andquantificationcomesfromthereporterionsinMS/MS.$

However, using TMT introduces implications for the mass spectrometric workflow and peptide identification. For example, the prominent reporter-ion region in the low m/z range (around 126–131) competes with fragment ions from the peptide backbone and may suppress certain b- or y-ions, particularly at low m/z, which can reduce sequence information [QuantGuide21]. In addition, labelling the N-terminus and lysine residues introduces systematic mass shifts and altered fragmentation behaviour (e.g., heavier fragments, attenuated high-mass ions or modified cleavage efficiency) that may reduce identification rates or depth in database searches [Chen21] [SabaViner16]. Moreover, different multiplex levels (e.g., TMTpro 16-plex or 18-plex) often require optimised collision energies and may yield altered intensity patterns or ion coverage compared to unlabelled or single-label workflows [ThermoTMTSystems].

Because most existing de novo peptide sequencing (DNPS) or database search pipelines have been developed and trained on unlabeled or simpler label systems, they may not fully account for the systematic biases introduced by TMT labelling. To maximise the discovery of modified peptides (especially those with post-translational modifications) in complex TMT experiments, adapting models to account for reporter-ion regions, mass shifts and altered fragmentation becomes important. In the work proposed here, we aim to bridge that gap by expanding DNPS and search strategies to TMT workflows and PTM-rich data.

2 Materials and Methods

3 Results and Discussion

4 Conclusion