Methods for Development of *De Novo* Sequencing Algorithms

Susan Williams*

Department of Biological Studies, Flinders University, Adelaide, Australia

ABSTRACT

De novo peptide sequencing is the method of determining a peptide amino acid sequence from tandem mass spectrometry in mass spectrometry. For researching the biological activity of a protein, knowing the amino acid sequence of peptides from a protein digest is critical. The Edman degradation process was used in the past to accomplish this. The use of a tandem mass spectrometer to solve peptide sequencing problems is now more widespread. In general, two methodologies are used: database search and *de novo* sequencing. The database search is a simplified form in which the unknown peptide's mass spectra data is submitted and processed to locate a match with a known peptide sequence, with the peptide with the highest matching score being chosen. Because it can only match to existing sequences in the database, this method fails to recognize novel peptides. The assignment of fragment ions from a mass spectrum is known as *de novo* sequencing. For interpretation, many algorithms are utilized, and most instruments come with *de novo* sequencing applications.

Keywords: Mass spectrometry; Peptide sequencing; Sub sequencing

DESCRIPTION

Development of *de novo* sequencing algorithms

Manual *de novo* sequencing is time-consuming and labor-intensive. The interpretation of spectra is usually done using algorithms or programmes that come with the mass spectrometer. An ancient method is to list all probable peptides for the precursor ion in mass spectrum and compare each candidate's mass spectrum to the experimental spectrum [1]. The peptide with the most comparable spectra has the greatest likelihood of being the correct sequence. However, there could be a great number of potential peptides. A precursor peptide with a molecular weight of 774, for example, has a total of 21,909,046 potential peptides. It takes a long time, even if it is done on a computer.

Another method is "subsequencing," which matches small sequences of peptides that reflect only a portion of the total peptide rather than listing the entire sequence of potential peptides [2]. When sequences that closely match the fragment ions in the experimental spectrum are discovered, they are extended one by one by residues until the best match is identified.

The third method employs a graphical representation of the data, in which fragment ions with similar mass differences of one amino acid residue are linked by lines. It is easier to obtain a clear image of ion series of the same kind in this manner [3]. This method may be useful for manual *de novo* peptide sequencing; however it is ineffective in a high-throughput environment.

The graph theory is the fourth strategy that is believed to be successful. Bartels was the first to mention using graph theory in *de novo* peptide sequencing. Peaks in the spectrum become vertices in a graph known as a "spectrum graph." A directed edge will be applied if two vertices have the same mass difference of one or more amino acids. This sort of algorithm includes the SeqMS, Lutefisk, and Sherenga algorithms.

Deep learning approaches have recently been used to overcome the difficulty of *de novo* peptide sequencing. DeepNovo was the first breakthrough, using a convolutional neural network topology to achieve significant increases in sequence accuracy and entire protein sequence assembly without the use of

databases [4]. In order to extract information from a raw spectrum, additional network architectures, such as PointNet, have been used. The sequence prediction problem is then applied to the *de novo* peptide sequencing problem. Neural-network-based *de novo* peptide sequencing models will create the most likely next amino acid until the predicted peptide's mass matches the precursor mass, given a previously predicted partial peptide sequence.

CONCLUSION

PepNovo is a high-throughput *de novo* peptide sequencing tool that scores peptides using a probabilistic network. One spectrum interpretation usually takes less than 0.2 seconds. PepNovo outperforms Sherenga, PEAKS, and Lutefisk, amongst other popular algorithms. PepNovo+, a new version, is now available. DeepNovo outperformed earlier approaches, such as PEAKS, Novor, and PepNovo, by a large margin, according to the benchmark analysis in the original publication. The Tensorflow framework is used to implement DeepNovo in Python.

REFERENCES

1. Webb-Robertson BJ, Cannon WR. Current trends in computational inference from mass spectrometry-based proteomics. Brief Bioinform. 2007;8(5):304-17.

2. Smith LM, Kelleher NL. Proteoform: A single term describing protein complexity. Nat Methods. 2013;10(3):186-7.

3. Capriotti AL, Cavaliere C, Foglia P, Samperi R, Laganà A. Intact protein separation by chromatographic and/or electrophoretic techniques for top-down proteomics. J Chromatogr A. 2011;1218(49):8760-76.

4. Roth MJ, Plymire DA, Chang AN, Kim J, Maresh EM, Larson SE, et al. Sensitive and reproducible intact mass analysis of complex protein mixtures with superficially porous capillary reversed-phase liquid chromatography mass spectrometry. Anal Chem. 2011;83(24):9586-92.