Semantic Ion Vectors - deep learning applied to mass spectrometry

Starcevic A.[1], Zucko J.[1], Damir O.[1], Diminic J.[1] and Cindric M.[2]

[1] *Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva 6, 10000 Zagreb, Croatia*
[2] *Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia*

Abstract:
Background: Protein mass spectrometry is the dominant method used for protein characterization. Peptide mass fingerprinting (PMF) is a phrase given to application of tandem mass spectrometry (MS2) to protein identification as the final goal of the method. In case of PMF, peptides do not fragment sequentially. To make things more complex, the process is not entirely random, with some fragmentations being more preferred over others. The resulting fragmentation spectrum captures fragment ions, which we observe as separate "peaks". Each peak is determined by a tuple of values: mass/charge (m/z) ratio (X-axis) and ion "intensity" (Y-axis), an underutilized and not fully understood value.

Results: A novel approach which relies on Deep learning techniques to capture distributed representations of peaks into Paragraph Vectors using unsupervised algorithm was employed to predict informative b and y ions and distinguish them from the "noise" peaks. Unlike Word embedding, a collective name used to describe the process of turning words and phrases into vectors of real numbers; this approach has taken a different turn. Numerical data (m/z and intensity couples) are turned into words (tokens), which were grouped into sentences and afterwards embedded using Paragraph Vectors. Several tokenization schemes have been implemented and performance of the resulting Ion Vectors have been tested under different parameters. Sole fitness criteria used was the performance of simple binary classification of peaks into "ions" (b and y) and "noise" (the rest). The best performing combination produced Semantic Ion Vectors, which were fed into classifier of choice and served as a model to make valuable predictions of "b-y ions" vs. "noise.

Conclusions: Resulting Semantic Ion Vectors can be used in variety of classification tools and provide accurate predictions of b-y ions. Tokenization method developed, efficiently reduced complexity of ion recognition task, by relying on constant "delta's" - distances between neighboring peaks rather than using m/z values directly. Intensity patterns and other information, commonly used as a set of rules in a synthetic manner (more suited to organic brain) were replaced by simple tokens and embedded into Paragraph Vectors using the vast number of spectral data in NIST repositories and neural networks.

Keywords:
bioinformatics, fragment spectrum, proteomics, b and y ions, Doc2Vec, deep learning, binary classification, big data, language processing, tokenization

This work has been funded by Croatian Science Foundation - Research project IP-06-2016 \"Exploring Gut Microbiome Equilibrium - MicroEquilibrium\".

*Corresponding author, e-mail: astar@pbf.hr