Journal of Discrete Algorithms 13 (2012) 19–31

Contents lists available at SciVerse ScienceDirect

Journal of Discrete Algorithms

www.elsevier.com/locate/jda

JOURNAL OF
DISCRETE
ALGORITHMS

Non-metric similarity search of tandem mass spectra including posttranslational modifications

Jiří Novák*, Tomáš Skopal, David Hoksza, Jakub Lokoč

SIRET Research Group, Department of Software Engineering, Faculty of Mathematics and Physics, Charles University in Prague, Malostranské nám. 25, 118 00 Prague, Czech Republic

ARTICLE INFO

Article history:
Available online 29 October 2011

Keywords:
Similarity search
Tandem mass spectrometry
Metric access methods
Posttranslational modifications
Peptide identification

ABSTRACT

In biological applications, the tandem mass spectrometry is a widely used method for determining protein and peptide sequences from an "in vitro" sample. The sequences are not determined directly, but they must be interpreted from the mass spectra, which is the output of the mass spectrometer. This work is focused on a similarity-search approach to mass spectra interpretation, where the parameterized Hausdorff distance (d_{HP}) is used as the similarity. In order to provide an efficient similarity search under d_{HP}, the metric access methods and the TriGen algorithm (controlling the metricity of d_{HP}) are employed. Moreover, the search model based on the d_{HP} supports posttranslational modifications (PTMs) in the query mass spectra, what is typically a problem when an indexing approach is used. Our approach can be utilized as a coarse filter by any other database approach for mass spectra interpretation.

1. Introduction

Proteins, organic molecules made of amino acids, are the basis of all living organisms. They are essential for construction of cells and for their proper function [15]. For bioinformatics purposes, a protein can be understood as a linear sequence over 20-letter subset of the English alphabet,[1] where each letter corresponds to an amino acid. A protein sequence must be determined from an "in vitro" protein sample, while tandem mass spectrometry is a very fast and popular method for this task. The proteins in the sample are split by enzymes into shorter pieces called *peptides*, and these are subsequently analyzed by the tandem mass spectrometer [8]. However, instead of direct production of the desired peptide sequences, the spectrometer outputs a set of experimental mass spectra[2] that have to be *interpreted* in order to obtain the peptide sequences. In particular, the interpretation of an experimental spectrum may be accomplished by means of similarity search.

In order to interpret an experimental spectrum, a database D_P of known protein sequences (e.g., MSDB [11]) can be employed. The peptide sequences and their hypothetical spectra are generated from the database D_P, forming a virtual database D_S of mass spectra. Then, the experimental spectrum is used as a query object and the database D_S is searched for its nearest neighbor spectrum (the most similar spectrum from D_S). The experimental spectrum is then interpreted as a peptide sequence corresponding to the spectrum found as the nearest neighbor.

The interpretation of spectra is often complicated by posttranslational modifications (PTMs) occurring in the query. The PTMs are usually not supported in existing similarity approaches among which using of cosine distance is popular.

* Corresponding author.
E-mail address: novak@ksi.mff.cuni.cz (J. Novák).
URL: http://www.siret.cz/novak (J. Novák).

[1] The letters B, J, O, U, X and Z are omitted.

[2] Each spectrum in the set corresponds to one peptide.

1570-8667/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jda.2011.10.003

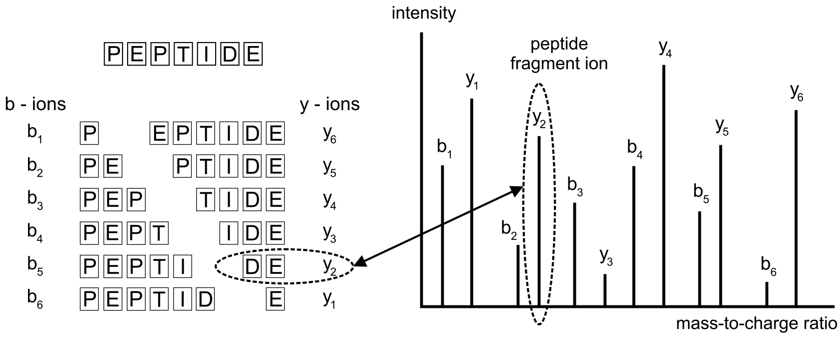

Fig. 1. An example of a mass spectrum.

1.1. *Paper contribution*

We present the non-metric parameterized Hausdorff distance d_{HP}, which exhibits better correctness of mass spectra interpretation than the cosine similarity does. Moreover, we propose a technique for efficient similarity search in a database of mass spectra indexed under d_{HP}, where for indexing we employ metric access methods (MAMs). In order to use MAMs efficiently, prior to indexing we utilize the TriGen algorithm to control the metricity of d_{HP}. The MAM, which we have chosen in our study, is the M-tree.

Due to the complexity of similarity search of mass spectra with PTMs, this problem is usually neglected in existing indexing approaches. Here, we extend the approach based on d_{HP} to support processing of spectra including PTMs. This extension can be also used in the approaches for mass spectra interpretation based on the cosine similarity.

2. Related work

We briefly describe the structure of data captured by the mass spectrometer and the common techniques employed for mass spectra interpretation using the database search approach.

2.1. *Mass spectrometry fundamentals*

The mass spectrum is a histogram of peaks corresponding to fragment ions (Fig. 1). A peak is represented by a pair $(\frac{m}{z}, I)$, where $\frac{m}{z}$ is the ratio of mass and charge, and I is the intensity of a fragment ion occurrence. For our purposes it is sufficient to consider $z = 1$ only, thus the ratios $\frac{m}{z}$ are equal to the mass m of fragment ions in Daltons.[3] The precursor mass m_p (the mass of a peptide before splitting) and charge z_p are also provided as an additional information for each peptide spectrum captured by the spectrometer.

In a mass spectrum, there are several types of fragment ions that are highly important for correct peptide sequence identification. The most frequent types of fragment ions with well predictable structure are y-ions and b-ions.[4] Each type of fragment ions forms a ion series, e.g., y-ions series or b-ions series (Fig. 1). The completeness of y-ions and b-ions series is crucial for correct spectra interpretation, because the mass difference between two neighboring peaks in one series, e.g., y_i and y_{i+1} corresponds to a mass of one amino acid.

Often, many of the y-ions or b-ions may never arise in the spectrometer and thus the number of missing y-ions and b-ions is too high to correct mass spectra interpretation. In fact, more than 85% of spectra captured by the spectrometer cannot be interpreted neither by an algorithm nor manually because the split process generates non-standard fragments. However, there are more factors making the interpretation complex. Up to 80% of peaks in each experimental spectrum may correspond to fragment ions with very complicated or unpredictable chemical structure and they make the recognition of y-ions and b-ions difficult. Such peaks are regarded as noise.

2.1.1. *Posttranslational modifications*

The interpretation of spectra is often complicated due to chemical modifications of amino acids, because masses of amino acids are changed in that case and thus peaks are shifted. This may happen during a sample preparation for the mass analysis, during the mass analysis in the spectrometer or after the translation of proteins in organisms. The last are so-called posttranslational modifications (PTMs; Fig. 2). Since it is not necessary to distinguish the modifications in our study, we use the term PTMs for all the modification types. The database UNIMOD [25] gathers discovered protein modifications for the mass spectrometry. At the time of writing this paper, there were about 660 known modifications.

[3] Dalton (Da) is a unit of the relative atom mass.

[4] In fact, more types of fragment ions with predictable structure may arise in the spectrometer, but many of them occur very rarely.