Information Sciences 301 (2015) 262–270

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Peptide sequencing via graph path decomposition

Yinglei Song [a,*], Albert Y. Chi [b]

[a] School of Electronics and Information Science, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu 212003, China
[b] Department of Mathematics and Computer Science, University of Maryland Eastern Shore, Princess Anne, MD 21853, USA

ARTICLE INFO

Article history:
Received 31 March 2014
Received in revised form 10 November 2014
Accepted 3 January 2015
Available online 9 January 2015

Keywords:
Peptide sequencing
Graph path decomposition
Dynamic programming

ABSTRACT

In proteomics, an important problem is to determine the amino acids sequence of a short peptide solely from its tandem mass spectrometry spectrum. Previous work has shown that a spectrum can be modeled with a directed acyclic graph and the amino acids sequence of the peptide can be obtained by computing the longest antisymmetric path in the graph. In this paper, we study the longest antisymmetric path in a general directed acyclic graph and show that the problem can be solved in linear time when both the number of symmetric vertex pairs and the path width of the graph are bounded from above by constants. We have implemented this linear time algorithm and tested its performance on experimental spectrums. Our testing results suggest that the algorithm can efficiently process the MS/MS spectrum of a peptide and provide sequencing results of high accuracy. Since the algorithm does not impose any additional requirements on the structures of the underlying directed acyclic graph, it might be of independent interest and can potentially be applied to problems of similar nature in software testing and software validation.

1. Introduction

Tandem mass spectrometry (MS/MS) is an important experimental approach for analyzing protein sequences in proteomics [3,4,19]. In experiments, Enzymes are used to break a protein molecule into short peptide sequences and the sequences of amino acids in these peptides can be determined from their MS/MS spectra. The sequence of amino acids in the protein molecule can then be obtained by combining the sequencing results of these peptides.

In order to obtain the amino acids sequences of these peptides, peptides that have the same amino acids sequence are fragmented into charged suffix and prefix subsequences. These charged subsequences are called *ions*, and a mass spectrometer can be used to measure their mass/charge ratios. A mass peak in a spectrum corresponds to a particular mass/charge ratio captured by the mass spectrometer.

Theoretically, two types of ions are contained in a MS/MS spectrum. They are b-ions associated with the N-terminal of a peptide and c-ions associated with its C-terminal. Two mass peaks are *complementary* if they represent the b-ions and c-ions resulting from fragmentation that occurs at the same position in the backbone of a peptide. The mass/charge ratio of a mass peak is the *mass value* of the peak. In the ideal case, fragmentation occurs in each position along the sequence backbone of the peptide, the difference between the mass values of two mass peaks that are of the same ion type and consecutive in the spectrum is the mass of a single amino acids. The amino acids sequence of the peptide can thus be inferred by analyzing the mass peaks in the spectrum with mass values of single amino acid residues.

* Corresponding author.
E-mail addresses: yingleisong@gmail.com (Y. Song), aychi@umes.edu (A.Y. Chi).

http://dx.doi.org/10.1016/j.ins.2015.01.003
0020-0255/

However, sequencing a peptide solely from its experimental MS/MS spectrum is difficult in practice [7,8]. First, the ion type of a mass peak in the spectrum is unknown and cannot be easily determined. In addition, multiple fragmentation may occur in a peptide and a large number of noisy mass peaks thus may appear in an experimental spectrum. Some mass peaks that are crucial for sequencing may not appear in a spectrum due to experimental errors.

The goal of the *de novo* sequencing problem is to obtain the amino acids sequence of a peptide solely by analyzing its MS/MS spectrum. A large number of computer algorithms have been developed for the problem. In the early stage, approaches based on exhaustive enumeration are used to compare the experimental spectrum with the theoretical spectra of all possible candidates for the peptide and the one whose theoretical spectrum is the most similar to the experimental one is output as the sequencing result [23]. Approaches based on exhaustive enumeration are slow and computationally inefficient. Later, heuristic approaches based on prefix pruning were used to remove the peptides whose prefixes do not match the experimental spectrum well [24,32,33]. However, prefix pruning may introduce errors to the sequencing result and is not guaranteed to improve the computational efficiency for sequencing.

In [8], a graph model is proposed to model an experimental spectrum. Such a graph is a *spectrum graph*, two vertices in a spectrum are *symmetric* if they represent complementary mass peaks. A path is *antisymmetric* if it does not contain two vertices that are symmetric. It is clear that an experimental spectrum can be modeled with a directed acyclic graph. In addition, the longest antisymmetric path in a spectrum graph corresponds to the amino acids sequence of the peptide. The *de novo* sequencing problem thus can be solved by computing the longest antisymmetric path in the spectrum graph [3,8,10,11,29,31]. In [7], a linear time dynamic programming algorithm is developed to compute the longest antisymmetric path in the spectrum graph of an ideal spectrum. However, the algorithm needs a large amount of time to process noisy mass peaks and thus cannot be directly used in practice.

Another type of approaches generate sequencing results by searching a database of experimental spectra [9,18,16,19,20], these approaches find the spectrum that has the highest similarity to the one that needs to be sequenced in the database and returns the peptide sequence that corresponds to the spectrum as the sequencing result. The sequencing accuracy is high for most of the approaches of this type. However, approaches based on database search are unable to sequence the peptides whose spectra are not in the database.

Our previous work exploits the notion of spectrum graphs and introduces non-directed edges to join symmetric vertices [15]. The resulting graph is an *extended spectrum graph*. We show that, based on a graph tree decomposition of the extended spectrum graph, the longest antisymmetric path can be computed in time $O(6^t n)$, where t is the tree width of the tree decomposition and n is the number of mass peaks in the spectrum [15]. The algorithm is efficient when the tree width is a small integer. Our experiments on extended spectrum graphs also show that the tree widths of the majority of extended spectrum graphs are small positive integers. However, the computational efficiency of the algorithm drops significantly when the tree width of the spectrum graph is larger than 10. Recently, an algorithm based on integer linear programming is developed in [1] to efficiently compute the longest antisymmetric path in a spectrum graph. However, the computation efficiency of this algorithm is not guaranteed.

In [25], we develop a parameterized algorithm that can compute the longest antisymmetric path in an extended spectrum graph in time $O(p^2 2^p n)$ based on a path decomposition of the graph, where p is the path width of the path decomposition and n is the number of vertices in the graph. However, the algorithm requires that any pair of vertices joined by a directed edge are not connected by an alternative directed path in the graph. It thus does not solve the problem in any given directed acyclic graph. In addition to proteomics, the problem also has important applications in a few other fields, such as software testing [30] and software validation [12]. Although the property holds for an extended spectrum graph, it may not hold for instances of the problem in these fields. It is thus highly desirable to develop algorithms that can efficiently solve the problem in any given directed acyclic graph.

In this paper, we study parameterized algorithms that can compute a longest antisymmetric path in any given directed acyclic graph. We choose the number of symmetric vertex pairs and the path width of the directed acyclic graph as parameters. We show that when the number of symmetric vertex pairs is at most m, the longest antisymmetric path can be computed based on a path decomposition of the graph in time $O(4^{p+m} n)$, where p is the path width of the path decomposition and n is the number of vertices in the graph. The problem can thus be efficiently solved when both p and m are small integers.

As an application of the algorithm, we test its performance on the *de novo* sequencing problem. The algorithm is tested on a large number of experimental spectra downloaded from the Open Proteomics Database(OPD) [21]. We tested the accuracy and computational efficiency of our algorithm and compare it with a few other sequencing tools, including PepNovo [18], NovoHMM [17], TDS [15], and PDS [25]. Our testing results show that, for majority of the tested spectra, the path width of the extended spectrum graph is less than 5 and the algorithm is able to efficiently determine the amino acids sequence of the tested spectra with high accuracy.

2. Models and algorithms

2.1. Spectrum graph models

We use $S = \{m_1, m_2, \cdots, m_{2k}\}$ to denote the set of mass peaks in the MS/MS spectrum of a peptide. Ideally, for each mass peak m_t $(1 \leqslant t \leqslant 2k)$, m_{2k+1-t} is complementary to it. The sum of the mass values of the two complementary mass peaks is the