Journal of Scientific and Practical Computing
Vol.2, No.1 (2008) 13–20

THE IDENTIFICATION OF ION TYPES IN TANDEM MASS SPECTRA BASED ON A GRAPH ALGORITHM

CHUNMEI LIU, AJONI BLAKE, LEGAND BURGE

Department of Systems and Computer Science, Howard University, 2300 Sixth Street, NW Washington, District Columbia 20059, USA
chunmei@scs.howard.edu, ajonib@gmail.com, blegand@scs.howard.edu

Abstract: It is important to identify the ion types of mass peaks in a tandem mass spectrum in order to interpret the spectrum and derive its peptide sequence. In this paper, we propose a graph algorithm to partition the mass peaks in a tandem mass spectrum into three sets: b-ion set, y-ion set and noise set. In particular, we construct a spectrum graph with each vertex representing a mass peak in a spectrum and connect vertices with three types of edges: two vertices are connected with a type one edge if the mass difference of the two corresponding mass values is the mass of an amino acid or sum of multiple amino acids; if two vertices are not connected with a type one edge, they will be connected with a type two edge if the sum of their mass values is not the parent mass of the spectrum; otherwise, two vertices are connected with a type three edge. We then develop a dynamic programming algorithm to partition the vertex set of the spectrum graph into three sets and thus identify the ion types of the mass peaks. Our experiment results on 17 experimental spectra show that the algorithm can achieve an average accuracy of $\sim 95\%$.

*keywords:*Ion type, Tandem Mass Spectrometry, Peptide sequencing.

1. Introduction

Tandem Mass Spectrometry (MS/MS) has been developed into a dominant technique in proteomics in protein identification[12]. There are two common types of ions present in a MS/MS spectrum: b-ions and y-ions. Through MS/MS spectra, techniques have been extensively developed to sequence the peptides of the spectra. Among them database search[6,11] and de novo sequencing[4,5,7,8,9,10,14] are two most popular ones. For each query spectrum to be sequenced, database search techniques compare the query spectrum against each theoretical spectrum derived from each peptide of a peptide sequence database and the few best peptides will be filtered out as matches. Database search techniques have been used extensively. However, if the peptide sequence of the query spectrum does not appear in the search database, the techniques will fail. De novo sequencing technique derives peptide sequences directly from spectra. In particular, it usually constructs a spectrum graph from a query MS/MS spectrum to be sequenced and then finds the longest path in the spectrum graph to get the peptide sequence of the MS/MS spectrum[5,8,9]. It is highly dependent on the quality of spectra and a MS/MS spectrum usually is incomplete