

# Construction of theoretical spectra for peptide tandem mass spectra identification through database search

Tema Fridman<sup>1,2,\*</sup>, Vladimir Protopopescu<sup>1</sup>, Greg Hurst<sup>3</sup>, Andrei Borziak<sup>1</sup>  
and Andrey Gorin<sup>1</sup>

*1 Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6164*

*2 Joint Institute for Computer Science, University of Tennessee /ORNL, ORNL, PO Box 2008, Oak Ridge, TN 37831-6164*

*3 Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6131*

*\*To whom correspondence should be addressed fridmant@ornl.gov*

ORNL is managed by UT-Batelle, LLC for the U.S. Department of Energy under contract DE-AC05-00OR22725

Most “peptide identification through database search” methods require a comparison between the experimental tandem mass spectrum and the theoretical spectra with the goal of finding the “best matching” theoretical spectrum in the database. To this end, one has to first construct the theoretical spectra and then design a scoring function that will assign the best score to the correct theoretical spectrum. All existing scoring functions share one basic element: their scores depend on the *number of matches* - the number of predicted (theoretical) peaks found in the experimental spectrum.

Here we suggest to consider each experimental spectrum as an informational signal generated by the peptide; the theoretical spectra are the predicted informational signals. Thus, when creating theoretical spectra, we have to insert an optimal amount of “information” such that the peptides could be best distinguished from each other. More precisely, on the one hand, we have to provide sufficient information to allow discrimination between different peptides. On the other hand, providing too much information may lead to an increase of the number of false identifications.

We derive the optimal number of peaks (i.e., the minimum amount that provides the required efficiency of spectra identification) in the theoretical spectra as a function of: (i) the experimental accuracy,  $\sigma$ , of the measured ratio  $m/z$ , (ii) experimental spectrum density, (iii) size of the database, and (iv) fragmentation efficiency. We show that if theoretical spectra are constructed including N- and C-terminus ions only, then for  $\sigma = 0.5$ , which is typical for high throughput data, peptide chains of 8 amino acids or longer can be identified based on the *number of matches alone*, at a rate of false identification below 1%. To discriminate between shorter peptides, additional (e.g., intensity-inferred) information is necessary. We derive the dependence of the probability of false identification on the number of peaks in the theoretical spectra and on the types of ions that the peaks represent. It is shown that inclusion of neutral loss ions into the theoretical pattern sharply raises the false identification rate. Our results suggest that the class of mass spectrum identification problems for which more elaborate development of fragmentation rules (such as intensity model, etc.) is required, can be reduced to the problems that involve homologous peptides.