A Novel Data Mining Approach for Detecting the Mutated Peptides

Hang He[1] and Huangdong Meng[1,2]

[1]Department of Computer Science and Engineering,
University of Connecticut, Storrs, CT 06269, U.S.A.
[2]Joint Lab for Pervasive Healthcare Systems,
University of Connecticut - Xi'an Jiaotong University
{hang.he,huangdong.meng}@engr.uconn.edu

Abstract. Detecting the mutated positions in peptide sequences is a major challenge in peptide sequence reconstructing studies, as the accuracies of current algorithms always decrease by various biochemical reactions. In this article, we propose a novel data mining approach to improve the accuracy, in which the database consists of all the original peptide sequences is provided as references. For each peptide dataset, we firstly estimate the real mutated sequences by the direct computation algorithm (e.g. "De Novo-based Algorithm") and from the database references. And we then combine those two approximations as following: it starts with several peptide sequences that scored by the given mass spectra data and the database of all possible sub-sequences without mutations. After we divide all the sequences into shorter pieces (sub-sequences), we search the possible reference sequences in database by using each piece as a weighted tag, where the weight is computed by a preset function of sequence score. Secondly, a two-step approximation process is designed to identify the candidate sequences. Unlike the previous approach, we define the concept of "close" as measured by longest common sequence analysis, in which the different strategies are applied to create corresponding affine scores, instead of fully mapping the sequence tags. We therefore focus on each matched pair, in which we test the hypothesis that whether there are mismatches between the computation result and the reference and whether those mismatches are derived from the mutations in peptide sequences or not. As most systematic mistakes are expected to replace by mapped true value in database, we conduct retest the sequences until find the most convinced one. In the simulation studies, our approach achieves good performance on NIST09 mouse protein data that has about 19% higher accuracy compared with current method.

Keywords: Peptide reconstructing, Mutated detection, Mathematical approximation

Acknowledgments. Research is supported by Joint Lab Foundation for Undergraduate Innovation [110118].