Beyond Shared Peak Count: Finding and Using Correlations in MS/MS Data for Peptide Identification

Yan Fu[1], Ruixiang Sun[1], Rong Zeng[2], Simin He[1] and Wen Gao[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P. R. China
[2]Research Center for Proteome Analysis, Key Lab of Proteomics, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, P. R. China

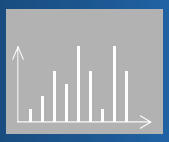
Contact: yfu@ict.ac.cn

I. Common Scoring Method: SPC

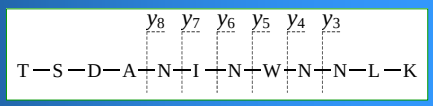
Shared Peak Count (SPC), i.e. the number of matched peaks between the experimental MS/MS spectrum and the theoretical one predicted from a peptide sequence in the database, is a widely used scoring method for peptide identification via MS/MS combined with database searching. Under the SPC, a false positive occurs after a search when less fragment ions predicted from the correct peptide get matched with the observed peaks than those predicted from some false peptide in the database (Figure 1 (c) is an example of false positive under the SPC scoring method).

II. Motivation

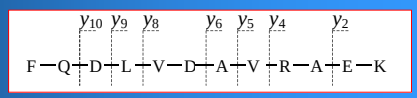
Our insight is that when positively correlated fragment ions are matched together, the matches are of a higher reliability as a whole than as individuals. Therefore, they should be emphasized to some extent for identifying peptides. Consecutive fragment ions, complementary fragment ions, and those fragment ions that differ in the charge state or the neutral loss of water/ammonia, are all potential positively correlated fragment ions (see Figure 1 for an example for the case of consecutive fragment ions).



(a) Experimental spectra



(b) The correct peptide



(c) A false peptide

Figure 1. The matched fragments of the correct peptide (b) are more consecutive than those of the false peptide (c).

III. Correlation Analysis of Matched Consecutive ions

We quantitatively analyzed the correlations on a dataset of ion trap tandem mass spectra. The matches of predicted fragment ions to the observed peaks were regarded as random events and the correlation coefficients between the matches of correlated fragment ions were computed for correct and false peptides respectively. By looking beyond the simple count of matched fragment ions or the SPC, and into the subtle correlations of matched fragment ions, we found interesting matching patterns that could be used to distinguish correct peptides from false positives. The consecutiveness of matched ions turns out to be one of the most important indicators for the correct peptides. In the case when the correct peptide has less matched fragment ions than false peptides, the correlation coefficient between matches of consecutive fragment ions that differ by up to four amino acid residues in correct peptides is significantly higher than that in false peptides (Figure 2).

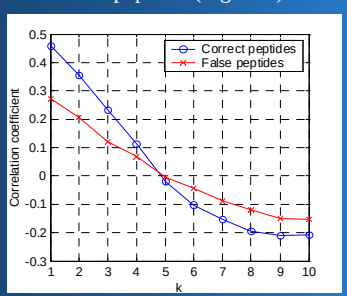


Figure 2. Correlation coefficients of matches of k-AAs consecutive fragment pairs. The k-AAs consecutive fragment pair is defined as two consecutive fragments that differ by k amino acid residues. For example, the y_3 and y_6 fragments in Figure 1 (b) constitute a 3-AAs consecutive fragment pair. It is shown that more correlated consecutive fragments in MS/MS data is a strong indicator for the correct peptide, even though the correct peptide may have less matched fragments than the false peptides (b)

IV. New Scoring Methods

We incorporated this kind of information into our peptide-scoring algorithm using the kernel technique popular in machine learning. As a result, a decrease of more than 10% in false positive rate was obtained compared to the scoring function SPC.

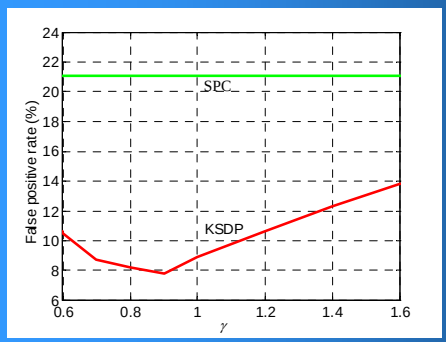


Figure 3. Compared to the common peptide-scoring method SPC, our kernel-based scoring method KSDP decreases the false positive rate by more 10%. Gamma is the parameter in the Radial Basis Function kernel used in KSDP. In this experiment, up to five consecutive fragment ions are considered by the KSDP.

V. Software

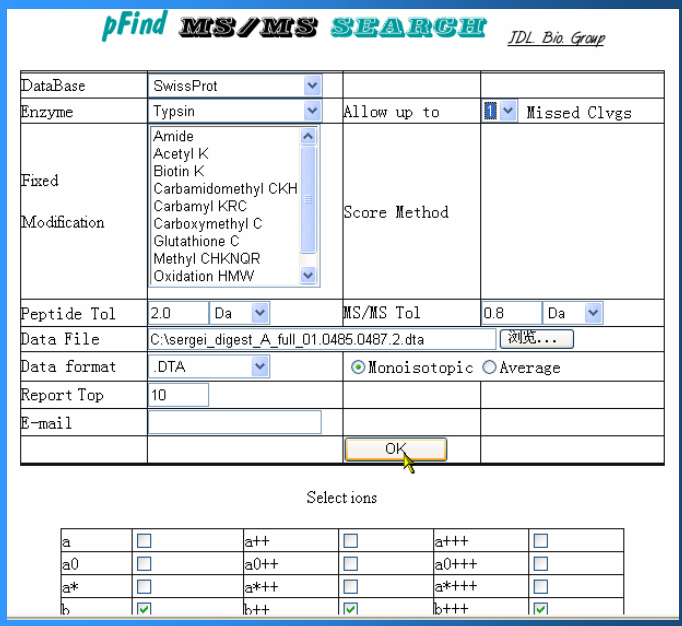


Figure 4. The interface of the pFind software. pFind is a database search engine for peptide and protein identification via MS/MS. It support most of the common formats of MS/MS data and can run in batch-mode. One of the characteristics of pFind is to allow users to manually select fragment ion types to be considered in searching. pFind is accessible on the Internet: http://pfind.jdl.ac.cn.

VI. Reference

Yan Fu, Qiang Yang, Ruixiang Sun, Dequan Li, Rong Zeng, Charles X. Ling, and Wen Gao. Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. Bioinformatics, 20: 1948-1954, 2004.