

A Procedures for Measuring Retention Time Deviations and Improving the Accuracy of Peptide Identifications for LC-MS/MS

Ruihua Fang¹, Pei Wang², Jimmy Eng¹, Petra Mannova, Laura M Beretta, and Martin McIntosh¹

Laboratory of Computational Proteomics Laboratory¹, Cancer Prevention Program², Fred Hutchinson Cancer Research Center, Seattle, WA

ABSTRACT

We present a statistical procedure for improving the accuracy of peptide identification of tandem mass (MS/MS) spectra using the peptides' LC retention time information. The LC retention time deviance score (Di), which measures the deviance between the observed and the predicted retention time of the peptide, was derived using a two-step procedure, which is independent of the LC configuration. We derived a statistical model that employs the Expectation Maximization (EM) algorithm for estimating the accuracy of the peptide assignment and that combines database search scores as well as other factors used by PeptideProphet (1) and the Di value. We found that Di value increased the classification power of database search scores such as XCorr score and the scores of PeptideProphet-like model. The algorithm and approach is freely available through the Computational Proteomics Analysis System (CPAS) (2).

INTRODUCTION

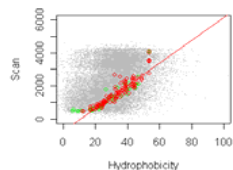
• False positive peptide identifications are often encountered in database searching of tandem mass spectrometry (MS/MS) in large scale proteomics analysis where tens of thousands of spectra are collected. In practice, an arbitrary score threshold was often used to separate "correct" from "incorrect" peptide identifications for a given analysis or for a search engine in general, while large numbers of peptide identifications often fall within a "gray area" where the score cannot distinguish correct from incorrect peptide identifications.

• Various studies have shown that the liquid chromatography (LC) retention time of a peptide can be predicted based on its amino acid sequence. It has also been shown that a neural network trained using peptides identified in LC-MS/MS from trypsin digestion of complex biological samples can predict the retention time of a peptide within a 5% error rate and this information has been used to improve the confidence of peptide identifications (3).

• In this study, we use the retention time information to improve the identification confidence in data base searching of tandem mass spectrometry. We introduce a retention time deviance score (Di) for each peptide. The predicted hydrophobicity index of the peptide was calculated using an open source prediction model (4) and Di was then defined as the normalized distance between the predicted hydrophobicity index and the observed retention time for each LC-MS measurement. A probability model (similar as the one used in PeptideProphet) was then derived based on Di value, the database searching scores, and other factors, which improved the discrimination power. The Di value automatically adapts to different LC-gradients and retention variabilities and is easily adaptable to different LC conditions and search algorithms.

RESULTS

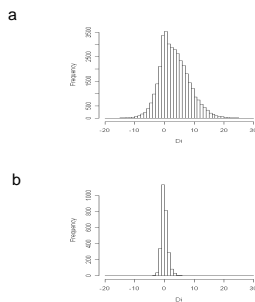
Fig 1. Correlation Between Elution Time of Peptides and Their Predicted Hydrophobicity



*Circle in red represent those peptides with PeptideProphet score exceeding 0.99; circle in green represent those peptides with positive ID; dots in gray represent those peptides with PeptideProphet score below 0.99

1. The hydrophobicity index of peptides with correct IDs or Peptide Prophet scores above 0.99 are linearly correlated with their observed retention time.

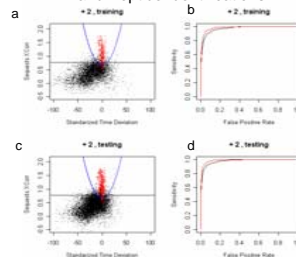
Fig 2. Distribution of Di



a. Distribution of Di of all the peptides; b. distribution of peptides with PeptideProphet score above 0.99

2. The Di values of all the peptides scattered in the range of -10 to 20. The Di values of the peptides with PeptideProphet scores above 0.99 are in the range of [-2.5, 2.5].

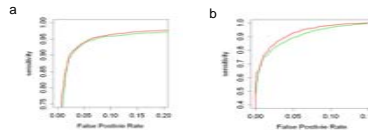
Fig 3. Correlation Between LC Retention Time Deviance and Peptide Identifications*



*a, c, training dataset, dot in red are peptides with peptide prophet score above 0.99, and dot in black are peptides with peptide prophet score below 0.99; b, d, testing dataset, the ROC curve in black is from XCorr score and the line in red is from the composite score of XCorr and Di using a quadratic fit.

3. Incorporation of Di increased the classification power of XCorr score.

Fig 4. Effects of Di on Classification Power of PeptideProphet-like Model



a. ROC curve for CP* (red), CPD** (green) on 22 LC-MS/MS analysis of protein standard mixture
b. same as in a except these ROC curve are from 5 LC-MS/MS analysis of yeast cell lysates.
*CP=Prob(identification is correct | L_value, charge, NTT, NMC, Mass error)
**CPD=Prob(identification is correct | L_value, charge, NTT, NMC, Mass error, Di)

4. Incorporation of Di into the probability model increased the classification power of the model in a yeast dataset as well as the proteins standard dataset. It also increased the number of positive peptide IDs in both datasets at an error rate of 5%.

Table 1. Effects of Di on Number of Correct Peptide Identifications at FDR control level of 0.05

Model	Number of Peptide IDs	
	Protein Standard	Yeast
CP*	2170	1270
CPD**	2296	1362

METHODS

Twenty-two LC-MS/MS measurement of an 18-protein standard mixture (1) and 5 LC-MS/MS measurement of a yeast cell lysate were analyzed. The protein standard dataset was searched against the sequences of these 18 proteins appended with the reversed human protein database (1). The yeast dataset was searched against a database of yeast protein sequences appended with the reversed UniProt database. The hydrophobicity index of each peptide was calculated using an open source algorithm (4). In each LC-MS/MS run, those peptides with PeptideProphet scores above 0.99 were used to map their hydrophobicity and observed retention time. The Di for each peptide in the entire run was then calculated according to Wasserman et al (5). Conditional probability of correct identification with and without Di were then calculated using the expectation maximization (EM) algorithm.

CONCLUSIONS

Incorporation of LC Retention time can increase the classification power of scoring system of XCorr and PeptideProphet-like model and increase the number of positive peptide IDs.

REFERENCES

- Keller A, Nesvizhskii AI, Kolker E, and Aebersold R. Anal. Chem. 2002, 74, 5383-5392.
- Rauch A, Bellow M, Eng J, Fitzgibbon M, Holzman T, Hussey P, Igra M, Maclean B, Lin CW, Detter A, Fang R, Faca V, Galken P, Zhang H, Whitaker J, States D, Hanash S, Paulovich A, McIntosh MW, J Proteome Res. 2006, 5(1), 112-21.
- Petritis K, Kangas LJ, Ferguson PL, Anderson GA, Pasa-Tolic L, Lipton MS, Aubery KJ, Strittmatter EF, Shen Y, Zhao R, Smith RD, Anal Chem. 2003 75(5), 1039-48
- Krokhin OV, Craig R, Spicer V, Ens W, Standing KG, Beavis RC, Wilkins JA, Mol Cell Proteomics. 2004, 3(9), 908-19.
- Neter J, Kutner MH, and Wasserman W, Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs 1985. R.D. Irwin

This work was funded by NCI subcontract 23XS144A.