A principled Bayesian framework increases confident peptide identifications from LC-MS/MS experiments

Albert Chen,[1] Alexander Franks,[2] Nikolai Slavov[1,3]
[1]Department of Bioengineering, Northeastern University, Boston, MA 02115, USA
[2]Department of Statistics and Applied Probability, UC Santa Barbara, CA 93106, USA
[3]Department of Biology, Northeastern University, Boston, MA 02115, USA

Abstract....

1 Introduction

Recent advancements in the sensitivity and discriminatory power of protein mass-spectrometry (MS) have enabled the analysis of increasingly limited amounts of samples. Most recently, we have achieved the quantification of single cell proteomes using the method Single Cell Proteomics by Mass Spectrometry (SCOPE-MS). The challenge, however, is identifying peptide sequences on extremely low levels of samples, where noise and interference can severely diminish identification rates. To help overcome this challenge, we developed the RTLib method to boost peptide identification rates from existing

The retention time (RT) of a peptide is an informative feature of its sequence. The predictive retention time of peptides, computed by software packages such as Skyline or ELUDE, is commonly used in Data Independent Acquisition (DIA), and in targeted MS/MS experiments where acquisition time is limited, i.e., multiple reaction monitoring (MRM). In shotgun proteomics and Data Dependent Acquisition (DDA), the retention time is only used for label-free quantification, and does not use the additional information in the MS2 spectra – which means it cannot be applied to tandem-mass-tag (TMT) data. The Percolator program can apply the retention time in a semi-supervised support vector machine (SVM), but as described in section ... [1].

We sought to extend the use of retention times to ions with MS2 spectra, within a rigorous Bayesian framework. The confidence in individual observations (peptide-spectrum-matches, or PSM) is based on comparing observed mass-spectra with 1) theoretical predication for the spectrum of each peptide sequence in the database and 2) in a reversed sequence database, which provides a null distribution. These results are used to estimate posterior error probability (PEP) for each PSM based on the MS spectra.

For some PSMs, the spectra alone provides strong evidence for the match and thus confidence in the identified peptide sequence. For others, however, the spectra alone are not sufficient evidence for confident assignment of the spectra to an associated sequence. In such cases, we would like to use the peptide retention time as an additional piece of evidence, independent from the spectra, to boost the confidence in correct observations and decrease the confidence for incorrect observations. To this end, we suggest the following framework of Bayesian inference:

- $P($ PSM = Correct $\mid RT)$ – the posterior probability that an observation is correct given its observed retention time (RT)

- $P($ PSM = Correct $)$ – the prior probability for the PSM estimated from the spectra, i.e., $1 -$ PEP, where PEP is the posterior error probability estimated from the spectra alone.

- $P(RT \mid$ PSM = Correct $)$ – the conditional likelihood of the RT for the peptide to which the PSM is matched. This probability is estimated from a RT library built from multiple experiments, described in section II.

- $P(RT \mid$ PSM = Incorrect $)$ – The probability of observing the RT of the PSM if it is incorrect, i.e., the probability that a measured spectrum will have the observed RT if it corresponds to a peptide sequence different from the one assigned to the PSM. It is estimated from the empirical distribution of RTs for all PSMs in the experiment.

- $P(RT)$ – The marginal likelihood for the RT, which we estimate as a sum of the probabilities that the PSM is correct and that the PSM is incorrect.

2 Results

2.1 Validation

2.2 Implementation

References

1. Kall, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4,** 923–925 (2007).