1 Introduction

Increasing the sensitivity and discriminatory power of protein mass-spectrometry (MS) is a key goal. This goal is particularly important for analyzing limited samples since low peptide levels pose challenges to peptide sequence identification. This challenge is extreme for the quantification of proteomes from single cells that we recently made possible by Single Cell ProtEomics by Mass Spectrometry (SCoPE-MS). To help overcome this challenge, we employ retention times (RT) to boost peptide identification.

The retention time (RT) of a peptide is an informative feature of its sequence. It has long been used in targeted proteomics. In shotgun proteomics, however, it has been used mostly for label free quantification by matching ions between runs based on their MS1 M/z and an RTs window

We sought to extend the use of RTs to ions with MS2 spectra within a rigorous Bayesian framework. This extension is particularity relevant

-

We estimate the confidence in peptide-spectrum-matches (PSM) based on comparing observed mass-spectra with theoretical predication for the spectrum of each peptide sequence in the database and in a reversed sequence database, the latter providing a null distribution. These results are used to estimate posterior error probability (PEP) for each PSM based on the MS spectra. For many PSMs, spectra alone are enough for confident identification, and thus result in very small PEP, i.e., the spectra provide strong evidence for the match and thus confidence in the identified peptide sequence. However, for many PSMs, the spectra alone are not sufficient evidence for confident assignment of the spectra to the associated sequence. In such cases, we would like to use knowledge about the retention time (RT) of the peptides as addition piece of evidence, independent from the spectra, to boost the confidence in correct PSMs and decrease the confidence for incorrect PSMs. To this end, we suggest the following framework of Bayesian inference:

- $P(\text{ PSM = Correct } | \, RT)$ – the posterior probability that a PSM is correct given its observed retention time (RT)
- $P(RT \;\, | \text{ PSM = Correct })$ – 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.
- $P(\text{ PSM = Correct })$ – the prior probability for the PSM estimated from the spectra, i.e., $1 -$ PEP, where PEP is the posterior error probability estimated only from the spectra.
- $P(RT)$ – 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.
- $P(RT \, | \text{ PSM = Incorrect })$ – 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 once assigned to the PSM. It is estimated from the empirical distribution of RTs for all PSMs in the experiment.

$$P(\text{ PSM = Correct } \mid RT) = \frac{P(RT \mid \text{PSM = Correct })P(\text{ PSM = Correct })}{P(RT)}$$

$$P(RT) = P(RT \mid \text{PSM = Correct })P(\text{ PSM = Correct }) + P(RT \mid \text{PSM = Incorrect })P(\text{ PSM = Incorrect })$$