False identification transfer rate in proteomic: how to estimate it?

Alicia Lionneton*†[1], Christophe Bruley[2], and Thomas Burger[3]

[1]Etude de la dynamique des protéomes – Laboratoire Biosciences et bioingénierie pour la santé – France
[2]CEA Tech Grenoble – CNRS : FR3425, Centre de Recherche Inserm, Université Grenoble Alpes, U1038 – 17 rue des Martyrs 38054 Grenoble Cedex 9, France
[3]CEA Grenoble (BIG, Biologie à grande Echelle, EDyP) – INSERM U1038, Université Grenoble Alpes – 17 rue des Martyrs 38054 Grenoble Cedex 9, France

Abstract

Liquid Chromatography coupled with tandem Mass-Spectrometry (LC-MS/MS) is a reference analysis method to characterize proteins in biological samples. In this work, we focus on estimating the rate of errors that result from the commonly used methods to mutualize identifications of protein fragments between samples of a cohort analyzed by LC-MS/MS.
In bottom-up LC-MS/MS analyses, proteins are first digested by enzymes into peptides before being separated by LC and detected by MS. Then, the MS signal of a peptide can be simplified into a set of three characteristics: a Retention Time (RT), a series of mass-to-charge ratios (m/z) and the associated intensities – which are used to identify the sequence and abundance of the peptide.

Nevertheless, these characteristics are not sufficient to systematically identify all detected peptides. This is why it is customary to mutualize peptides' identifications between the multiple samples of a cohort or study. Concretely, a peptide's identification is transferred to a non-identified peptide that has similar set of characteristics in another sample. However, peptides' characteristics fluctuate between samples, notably RTs which are poorly reproducible. Consequently, samples are re-aligned using a regression that predicts the characteristics of an unidentified peptide based on those in other samples. If its predictions fall nearby an unidentified signal, it makes sense to transfer the peptide's identity. However, doing so can lead to incorrect transfers. It is therefore necessary to estimate the rate of false identification transfer.

To do so, some authors have proposed methods that informally mimic the knockoff filters procedure to control the false discovery rate. In this context, knockoffs are essentially artefactual erroneous identification transfers which should occur as frequently as real false transfer. While accurately generated knockoffs are paramount to correct false transfer rate estimate, no method is available to date to provably do so. On the other hand, the regression setting makes the cross-validation framework natural to estimate the false transfer rate. Doing so would require dealing with multiple train/test splits, where training data would be used for regression and test data for transfer onto the nearest MS characteristics (necessarily unidentified in the test set). Unfortunately, cross-validation is not feasible operationally

*Speaker
†Corresponding author: alicia.lionneton@cea.fr

because of computational constraints.

To cope with this, we propose to conflate training and testing data during regression and transfer. Doing so will lead to regression overfitting and error underestimation, yet, it also increases the search space for transfer, as identified peptides are no longer considered as such –which should lead to error overestimation.

The underlying assumption of our proposal is that both trends balance in a conservative way. The false identification transfer rate underestimate resulting from overfitting is offset by the overestimate due to the search space size increase. Our poster will empirically establish the validity of this assumption, by comparing our proposed approach with a genuine cross-validation on various experimental data.

Keywords: proteomics, mass spectrometry, error rate estimation, cross validation