BeNeLux Bioinformatics Conference – Antwerp, December 7-8 2015
Abstract ID: **000** Category: Abstract template

OPTIMAL DESIGN OF SRM ASSAYS USING MODULAR EMPIRICAL MODELS

Jérôme Renaux[1,], Alexandros Sarafianos[1], Kurt De Grave[1] & Jan Ramon[1].*
Department of Computer Science, KU Leuven.[1] [*]Jerome.renaux@cs.kuleuven.be

Targeted proteomics techniques such as Selected Reaction Monitoring (SRM) have become very popular for protein quantification due to their high sensitivity and reproducibility. However, these rely on the selection of optimal transitions, which are not always known in advance and may require expensive and time-consuming discovery experiments to identify. We propose a computer program for the automated identification of optimal transitions using machine learning and show encouraging results when compared to a widely used spectral library.

INTRODUCTION

A major issue with both SRM is to know which transitions to monitor in order to maximally detect a specific protein, these being different from one protein to another. Good candidates are transitions whose chemical properties will make them likely to occur and easy to detect by the mass spectrometer, while being sufficiently specific indicators of their parent protein.

Traditionally, targeted proteomics assays, which consist of lists of ions or transitions to monitor, are designed through costly exploratory experiments. Recently, attempts have been made to produce software to help design optimal assays. These efforts rely on some extent on collaborative databases of mass spectra which are mined to identify the best possible peptides to include in the assays. While successful, these approaches still depend on past exploratory analyses and on the coverage of the exploited databases. Therefore, their performance decrease in cases where such databases cannot be leveraged, such as when dealing with little-studied organisms or rare, low-abundance proteins.

We propose an approach called SIMPOPE (Sequence of Inductive Models for the Prediction and Optimization of Proteomics Experiments) that models all the steps of the typical tandem mass spectrometry (MS/MS) workflow in order to accurately predict the properties of peptide and fragment ions within a given proteome, and subsequently identify optimal assays among them.

METHODS

SIMPOPE consists of a sequential suite of predictive models for each step of the MS/MS workflow. It exploits knowledge from public databases and combines it with the generalizing power of machine learning models to compensate for noisy or missing data. All models are probabilistic, allowing to keep track of the inherent uncertainty of the successive predictions and to weight the results accordingly for the assay prediction.

Enzymatic cleavage is modelled using CP-DT(Fannes et al., 2013), which models the behaviour of the trypsin enzyme using random forests. Retention time prediction is achieved using the Elude tool from the Percolator suite (Moruz et al., 2010). The charge distribution of electrospray precursor ions is also modelled using random forests trained on experimental data mined from PRIDE (Vizcaino et al., 2013). Fragmentation patterns and product ion intensity are predicted with the help of random forest models trained on MS-LIMS data (Degroeve & Martens 2013; De Grave et al., 2014). Finally, prior knowledge about the abundance of proteins within a given proteome is incorporated as prior probabilities, obtained when available from PaxDB.

On the human proteome, these steps yield a total of 321 000 000 transitions together with their relevant chemical properties. We then compute a score for every single transition, based on these properties and on their aliasing with other transitions in terms of Q1 and Q3 m/z.

RESULTS & DISCUSSION

We validated our approach by computing scores for 2000 reference transitions from the SRMAtlas database (Picotti et al., 2014). Based on these scores, we can rank the reference transitions among all possible transitions. Intuitively, reference transitions should rank high, and therefore have a low rank (ideally, in the top five). Based on the average number of transitions per protein in our reference set, a perfect median rank would be 3.2, while a totally random scoring system should yield a median rank of 151. The approach we propose achieved a median rank of 15, signifying that using our scoring method, 50% of the reference transitions are ranked in the top 15. This result is encouraging as it shows that the scores predicted by SIMPOPE do correlate with the quality of the transitions. We can subsequently use that score as a feature to train an additional model on top of the ones described here to refine the assay prediction process (further results on the poster).

REFERENCES

Degroeve, S. & Martens, L. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics*, **29**, pp.3199–203 (2013).

Fannes, T. et al. *Journal of Proteome Research*, **12(5)**, pp.2253–2259 (2013).

De Grave, K. De et al. Prediction of peptide fragment ion intensity : a priori partitioning reconsidered. *International Mass Spectrometry Conference 2014*, (2014).

Moruz, L., Tomazela, D. & Käll, L. Training, selection, and robust calibration of retention time models for targeted proteomics. *Journal of Proteome Research*, **9(10)**, pp.5209–5216 (2010).

Picotti, P. et al. A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature*, **494(7436)**, pp.266–270 (2014).

Vizcaino, J. a. et al. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Research*, **41(D1)**, pp.D1063–D1069 (2013).