All-to-all spectra comparisons within minutes for peptides identification in tandem mass spectrometry

Matthieu David [1, 2]*, Guillaume Fertin [1], Hélène Rogniaux [2], Dominique Tessier [2]
[1] LS2N UMR CNRS 6004, Université de Nantes, F-44300, Nantes, France
[2] INRA UR1268 Biopolymères Interactions Assemblages, F-44316 Nantes, France
* To whom correspondence should be addressed: matthieu.david@univ-nantes.fr

1. INTRODUCTION

Proteins are omnipresent in living organisms and fulfill numerous functions. Frequently, *Post-Translational Modifications* (PTMs) affect proteins and modify those functions. Identifying proteins and their PTMs is most commonly performed through Tandem Mass Spectrometry (MS/MS). In MS/MS experiments, *peptides* are obtained from the digestion of proteins. From those peptides, the mass spectrometer produces *experimental spectra* that are then compared to a set of *theoretical spectra* extrapolated from the predicted fragmentation of peptides derived from a protein database. For each experimental spectrum s, a scoring function computes the similarity between s and the theoretical spectra. The best scored comparison corresponds to the inferred peptide identification. The set of peptides identified is then used to infer which proteins are present in the sample. However, for tractability reasons, the scoring function cannot be evaluated for *all* possible pairs of experimental vs theoretical spectra. To overcome this limitation, traditional peptide identification algorithms compare each experimental spectrum only against a *small subset* of theoretical spectra, filtered on the proximity of the *total mass* of the spectra. Since they modify the total masses of peptides, PTMs greatly limit this approach as the correct identification does not appear in this subset of theoretical spectra.. In order to avoid this drawback, one may include some PTMs in the search (e.g. through additional predicted theoretical spectra) but this approach quickly becomes unsatisfactory: first, it is limited to known PTMs only; second, the computational time drastically increases with the number of PTMs taken into account. *Open Modification Search* algorithms (1,2) attempt to loosen the total mass filter (without removing it entirely), but remain rather unsuccessful at tackling the computational time increase. Additionally, they generate more false positive identifications. We propose a new OMS algorithm, *SpecOMS*, able to compare hundreds of thousands of spectra for peptide identification, within minutes and without any preliminary mass filter.

2. METHODOLOGY

Let s_e be a spectrum from the set of experimental spectra S_E and s_t be a spectrum from the set of theoretical spectra S_T. We define $Sim(s_e,s_t)$, the scoring function used in our work, as the number of shared masses between s_e and s_t. *SpecOMS* relies on a compact data-structure, *SpecTrees* (3), that stores the necessary information for $Sim(s_e,s_t)$ to be efficiently computed -- such computation being achieved by an algorithm called *SpecXtract*. More precisely, for each s_e in *SpecTrees*, *SpecXtract* computes $Sim(s_e,s_t)$ for *all* s_t in S_T. A second algorithm, *SpecFit*, seeks to improve $Sim(s_e,s_t)$ for each pair (s_e,s_t) *SpecXtract* proposed, taking in account the mass difference between s_e and s_t. *SpecFit* then reports for each s_e the best pair obtained as the identification.

3. RESULTS

At the moment, *SpecOMS* is the fastest available OMS peptide identification algorithm. The analysis of 37,703 experimental spectra (PXD001468 dataset (4)) with a target database of 510,685 theoretical spectra (Homo Sapiens GRCh37) takes no more than 5 minutes using 3.5 GB of memory, on a standard workstation. For comparison, the recent OMS tools MODa (5) or PIPI (6) require respectively 10GB and 25GB of memory and 11 and 9 hours to complete, exploring a still limited search space. *SpecOMS* identifies 11,404 spectra in the dataset (i.e., around 30%) including missed-cleavage peptides (lack of action of the digestion enzyme), isotopic peptides (variants due to the presence of carbon 13) and known PTMs (carbamylation, deamidation, oxidation, formylation, dioxidation). Interestingly, *SpecOMS* also identifies peptide variants (amino-acids substitutions in the peptide sequence) and rarer modifications with a high score, some yet unrepresented in PTMs databanks. Finally, *SpecOMS* identifies peptides with labile PTMs (loss of neutral fragments that do not appear on the spectrum), for example glycosylated peptides that can be confirmed subsequently by marker ions. Although MODa and PIPI are able to identify more spectra within a short mass range they remain, unlike *SpecOMS*, unable to access identifications with high mass differences.

4. CONCLUSION

SpecOMS rapidity, coupled to a low memory usage, enables its routine use in tandem mass spectrometry laboratories. Without any kind of preliminary mass filter, *SpecOMS* quickly generates the profiles of chemical modifications and PTMs present in a sample at a global scale. Consequently, *SpecOMS* immediately highlights unwelcome artifacts, giving a chance to improve experimental protocols and to concentrate efforts on biological interesting PTMs. *SpecOMS* is therefore an excellent tool to direct follow-up biological analyses.

5. REFERENCES

1. Ahrne E., Muller M. and Lisacek F. 2010 Unrestricted identification of modified proteins using MS/MS. *Proteomics* 10, 671-686.

2. Na S. and Paek E. 2015 Software eyes for protein post-translational modifications. *Mass Spectrom Rev* 34, 133-147.

3. David M., Fertin G. and Tessier D. 2016 SpecTrees: An efficient without a Priori Data Structure for MS/MS Spectra Identification. In: Frith M., Storm Pedersen C. (eds) *Algorithms in Bioinformatics* WABI 2016. Lectures Notes in Computer Science, vol 9838, 75-76; Springer, Cham.

4. Chick J. M., Kolippakkam D., Nusinow D. P., Zhai B., Rad R., Huttlin E. L. and Gygi S.P. 2015 A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* 33, 743-749.

5. Na S., Bandeira N. and Paek E. 2012 Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell Proteomics* 11(4), M111.010199.

6. Yu F., Li N. and Yu W. 2016 PIPI: PTM-Invariant Peptide Identification Using Coding Method. *J. Proteome Res* 15(12), 4423-4435.