Extended abstract

SeqBIM

Evaluation of open search methods based on theoretical mass spectra comparison

Albane Lysiak[1,3], Guillaume Fertin[1]*, Géraldine Jean[1], Dominique Tessier[2,3]

[1] *Université de Nantes, CNRS, LS2N, F-44000, Nantes, France*
[2] *INRAE, BIBS facility, F-44316, Nantes, France*
[3] *INRAE, UR BIA, F-44316*
***Corresponding author**: guillaume.fertin@ls2n.fr

Abstract

Mass spectrometry remains the privileged method to characterize proteins. Nevertheless, most of the spectra generated by an experiment remain unidentified after their analysis, mostly because of the modifications they carry. Open Modification Search (OMS) methods offer a promising answer to this problem. However, assessing the quality of OMS identifications remains a difficult task.
Aiming at better understanding the relationship between (i) similarity of pairs of spectra provided by OMS methods and (ii) relevance of their corresponding peptide sequences, we used a dataset composed of theoretical spectra only, on which we applied two OMS strategies. We also introduced two appropriately defined measures for evaluating the above mentioned spectra/sequence relevance in this context: one is a color classification representing the level of difficulty to retrieve the peptide sequence that generated the identified spectrum ; the other, called LIPR, is the proportion of common masses, in a given Peptide Spectrum Match (PSM), that correspond to dissimilar sequences. These two measures were also considered in conjunction with the classical False Discovery Rate (FDR). The three above mentioned measures allowed us to clearly determine which of the two studied OMS strategies outperformed the other, both in terms of number and of accuracy of identifications. Even though quality evaluation of PSMs in OMS methods remains challenging, the study of theoretical spectra is a favorable framework for going further in this direction.

1. Introduction

Mass spectrometry in tandem MS mode (MS/MS) is the most powerful method to identify proteins and characterize their modifications on a large scale. However, most of the spectra are left unidentified after their analysis by a dedicated software. This is likely due to the large proportion of spectra generated from proteins carrying modifications [1]. Software usually infer the identification of an experimental spectrum from its similarity to reference spectra. When a peptide carries a modification, its mass is by nature modified, which prevents its identification by conventional methods that compare each experimental spectrum with only a *restricted* set of reference spectra, approximately sharing the same mass in order to avoid excessive runtime. On the other hand, Open Modification Search (OMS) methods compare each experimental spectrum to *all* the reference spectra representing a proteome. Thus, this comparison produces a list of Peptides to Spectrum Matches (PSMs) per experimental

spectrum, and a mass difference $\Delta m \neq 0$ between the experimental spectrum and its associated peptide is assumed to be due to one or several modifications that differentiate them. Many scores exist to evaluate the similarity between two spectra, which all take into account, at a certain level, the number of peaks (i.e., of masses) that are shared by the two spectra, a number called shared peaks count (SPC). Despite the scientific relevance of better spectra identifications, OMS methods are still underused, notably because their reliability remains debated. It is therefore important to better describe the advantages and limitations of these methods. We focused our study on a thorough understanding of two widely spread strategies to determine the best PSM for each experimental spectrum. In the first strategy *Strategy1*, the best PSM is chosen according to a score that does not take Δm into account. The second strategy *Strategy2* tries to improve the alignment – and thus the score – of all the PSMs according to Δm before the choice of the best PSM (see Figure 1). In order to determine the most efficient strategy, a prerequisite was to be able to implement both strategies using the same software, which implies the availability of very efficient spectra comparison and alignment algorithms. The SpecOMS software [2], which we have previously developed, fulfills these conditions.

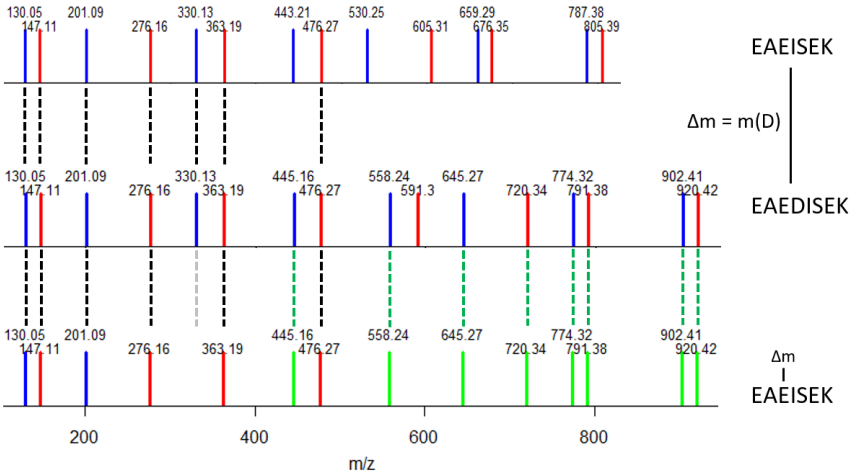

Figure 1. MS/MS spectra matches and their peptide sequences. *b*-ions (containing masses information from the prefix of the peptide) are displayed in blue, *y*-ions (containing masses information from the suffix of the peptide) in red, and matches between spectra in dashed lines. Intensities of all peaks are set to an arbitrary unit value. The middle EAEDISEK MS/MS spectrum shares 7 masses (black dashed lines) with the above native EAEISEK spectrum. After a shift of Δm at position 3 in EAEISEK (below), 8 new masses match with EAEDISEK (shown in green), and one match is removed (grey dashed line). The SPC is then improved from 7 (*raw SPC*) to 14 (*shift SPC*).

To compare in-depth the limits of each strategy, we decided to ground this study using the theoretical spectra derived from the human proteome, considering successively each theoretical spectrum in the role of an experimental spectrum. Doing so, we eliminate the inherent identification difficulties due to the imperfection of experimental spectra (e.g. noise, missing peaks) and concentrate on the benefits of each strategy. Hence, PSMs with $\Delta m \neq 0$ can only be explained by differences in terms of sequence, namely insertions, deletions and/or substitutions of one or several amino acids. Every PSM matching a peptide to itself was considered irrelevant and

consequently forbidden. Many OMS methods estimate the FDR of their results with a target/decoy approach [3]. We also used this approach to compare both strategies, even though it is still unclear whether or not this method underestimates the incorrect identifications [4, 5]. That is why we propose two additional measures of the PSM characteristics to evaluate their quality and compare the strategies.

2. Methods

Peptide identification using SpecOMS We implemented two strategies to find the best PSMs in SpecOMS. Next, we applied these two strategies to compare the set of theoretical spectra generated from the human proteome against themselves. A peptide that plays the role of an experimental spectrum in PSM is called the *bait*, whereas a peptide associated to a bait in a PSM is called a *hit*. Parameters were set in such a way that SpecOMS extracted from its data structure SpecTrees [6] all pairs of spectra of the form (bait,hit) whose SPC is greater than or equal to 7. Depending on the run, the parameter "shift" of SpecOMS was set to `false` (*Strategy1*) or `true` (*Strategy2*) (see Figure 2). More precisely, in *Strategy1*, for each bait, SpecOMS selects the best PSM based on the highest SPC, a score that we call *raw SPC*. In *Strategy2*, the best PSM for a given bait b is selected after the following two-step procedure is applied: first, for every candidate hit h for b such that $\Delta m \neq 0$, SpecOMS realigns h to b by shifting its masses (by Δm) at each possible relevant location in the spectrum, and retains the shift location in h that yields the best newly computed SPC. Second, SpecOMS chooses the best PSM among the candidate PSMs for b, based on the newly computed SPC, that we call *shift SPC*.

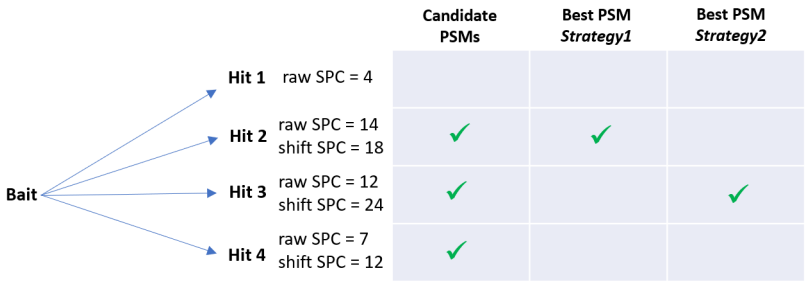

Figure 2. Determining the best PSM in each strategy. Suppose, fictitiously, that a given bait is to be compared to 4 peptides (called hits). Hit 1 is discarded as its *raw SPC* with bait is below the imposed threshold of 7. Hits 2, 3 and 4 are candidate PSMs for bait. Since $\Delta m \neq 0$ for Hits 2, 3 and 4, a shift may be applied, and in that case *shift SPC* is obtained (with *shift SPC* $\geq$ *raw SPC* by definition). In *Strategy1*, the best PSM for bait is Hit 2, as it is based on *raw SPC*. In *Strategy2*, the best PSM for bait is Hit 3, as it is based on *shift SPC*.

Data and theoretical spectra generation The human proteome was downloaded from Ensembl 99, release GrCh38. Proteins predicted with the annotation "protein coding" were added to 116 contaminant proteins downloaded from the cRAP contaminant database. The resulting set of proteins is referred to as the *target* database, *in silico* digested by trypsin. Each peptide is fragmented *in silico* by SpecOMS, so as to transform it into a *theoretical spectrum*. For this, ions from the b and y series are generated. For a given peptide, the set of generated masses represents its theoretical spectrum.

Measuring the quality of PSMs

False Discovery Rate (FDR). The first classical measure that we used is the number of PSMs we can validate at a given False Discovery Rate (FDR). We calculated the FDR as the proportion of best PSMs of the form (bait,hit) for which the hit is a decoy, over the total number of best PSMs. In this work, we were essentially interested by PSMs for which the FDR is less than 1%.

Color classification. Another parameter which we consider as informative for validating MS/MS results, notably in this context, is our ability to explain a PSM of the form (bait,hit) obtained by a given strategy ; by "explain", we mean unambiguously determine the transformation (in terms of amino acid sequence) that is required to retrieve the bait starting from the hit. Thus, the question we ask ourselves is the following: given a PSM (bait,hit) together with Δm, *shift SPC* and its corresponding best shift location, how difficult is it to precisely explain bait from hit ? For this, we introduce here a classification of PSMs into three colors (Green, Orange or Red), depending on this level of difficulty, from the easiest (Green) to the hardest (Red). In a nutshell, Green means that we are able to explain the link between hit and bait unambiguously, Orange contains some level of ambiguity, and Red means that further information and/or computational efforts are necessary to explain the relationship between bait and hit. See Figure 3 for examples of the color classification.

Bait	*Hit*	*Δm* (Da)	Shift location	Interpretation : from the hit to the bait
TSSDSISR	TSDSISR	87.0320	3	TSDSISR + S = TSSDSISR
YQEFQNR	EPPNPEYQEFQNR	-663.2864	6	EPPNPEYQEFQNR – EPPNPE = YQEFQNR
ETVHIPGAR	ETIPGAR	236.1273	2	ETIPGAR + Δm = ET-Δm-IPGAR
VISPEDGK	VIESPDGK	0	/	/
VCASIAQK	VTIQCQK	0	/	/
WFSIYDQR	FWSIQDYFR	-147.0684	/	/

Figure 3. Illustration of the Green/Orange/Red classification of PSMs. The first two rows present PSMs with a bait unambiguously deductible from the information given by the PSM, thus classified as Green. In the first example, Δm corresponds to the mass of S, which can thus be added in the hit at the given location to retrieve the bait. In the second example, the absolute value of Δm corresponds to the mass of EPPNPE, which can be deleted from the hit to retrieve the bait. In the third row, Δm can correspond to two possible amino acid sequences (VH or HV). Such PSM is thus classified as Orange. In the last three rows, transforming hit into bait is too ambiguous, although sequences may be close (e.g., first red row). In all cases, such PSMs are classified as Red.

Low Information Peaks Rate (LIPR). In an MS/MS experiment, spectra are considered similar to each other if they share a high number of masses. Ions from the same series (i.e., y-ions or b-ions in our case), which represent the same fragment, necessarily possess the same mass. Consequently, common masses represent relevant information concerning sequence similarity. However, the converse is not always true : identical masses may not represent identical sequences, for example when amino acids are permuted (e.g., AEAE and EEAA have the same mass) or in more complex situations when combinations of different amino acids turn out to have

the same total mass (e.g., KE and GVT have the same mass). Following the above discussion, we introduce here a new measure, that we call Low Information Peaks Rate (or LIPR). For a given PSM (bait,hit), LIPR(bait,hit) is the percentage of common masses between bait and hit that *do not* correspond to identical sequences. A LIPR close to 0 implies that the two amino acid sequences of bait and hit are very similar. In that case, one can argue that the PSM at hand is relevant, and that retrieving bait from hit may be feasible. On the other hand, when LIPR is close to 100, both sequences, although sharing a non negligible number of masses, represent very dissimilar sequences, and the PSM can thus be considered as debatable.

3. Results and Discussion

We successively implemented *Strategy1* and *Strategy2* to compare all the theoretical spectra generated from the human proteome (572 063 spectra) against a database merging the target and decoy human proteins (1 148 608 spectra). About 80% of the 572 063 tryptic peptides from the human proteome share at least 7 peaks with any other peptide, and about 23% of them share at least 10 peaks (target or decoy). Respectively to an FDR less than 1%, *Strategy1* validates 17 160 PSMs with a minimal *raw SPC* of 17 (i.e. considering best PSMs for which *raw SPC* $\geq$ 17), while *Strategy2* validates 57 784 PSMs with a minimal *shift SPC* of 21. *Strategy2* recruits more than three times more PSMs than *Strategy1*, thus we can conclude that *Strategy2* behaves better than *Strategy1* according to the number of validated PSMs. But one may wonder to what extent the information given by these PSMs is enough to restore the correct amino acid sequence of the baits.

We could get the distributions of the sets PSM_1 and PSM_2 obtained respectively by *Strategy1* and *Strategy2* among the 3 color classes, as well as the evaluation of the LIPR feature. Both strategies behave in a similar fashion, but at less than 1% FDR, *Strategy2* validates roughly three times more Green PSMs than *Strategy1* (27 211 vs 9 153). Thus, at first glance, the number of additional identifications obtained by *Strategy2* (compared to *Strategy1*) does not come at the cost of a deterioration of the quality of the results. In terms of LIPR, it can be noted that its average value is higher for PSM_1 (38.5% for PSM_1 vs 22.97% for PSM_2).

The performances obtained by *Strategy2* also lead us to conclude that, among the two, *Strategy2* is the one that should be implemented in OMS software. We saw that although *Strategy2* recruits more Green and Orange PSMs than *Strategy1*, it contains proportionally more Red PSMs. However, the average LIPR from the Red class obtained by *Strategy2* is much lower than for *Strategy1*. This leads us to think that a proportion of the Red PSMs from *Strategy2*, that share enough peaks corresponding to common subsequences, could be considered as "almost valid" PSMs. More precisely, we believe that with additional methodological and computational effort, some of these Red PSMs could be transfered to the Orange or Green category, an effort that methods implementing *Strategy2* should pursue.

By comparing two OMS strategies with theoretical peptides and new indicators, we also developed an environment which allowed us to see and understand elements that are more difficult to comprehend in an experimental context. This protocol could be used to understand principles that are at the heart of other (OMS) MS identification tools in order to configure and calibrate them.

References

[1] J. Griss, Y. Perez-Riverol, S. Lewis, D. L. Tabb, J. A. Dianes, N. Del-Toro, M. Rurik, M. W. Walzer, O. Kohlbacher, H. Hermjakob, R. Wang, and J. A. Vizcaíno. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nature Methods*, 13(8):651–656, 2016.

[2] M. David, G. Fertin, H. Rogniaux, and D. Tessier. SpecOMS: A Full Open Modification Search Method Performing All-to-All Spectra Comparisons within Minutes. *Journal of Proteome Research*, 16(8):3030–3038, 2017.

[3] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, 2007.

[4] W. S. Noble. Mass spectrometrists should search only for peptides they care about. *Nature Methods*, 12(7):605–608, 2015.

[5] A. Sticker, L. Martens, and L. Clement. Mass spectrometrists should search for all peptides, but assess only the ones they care about. *Nature Methods*, 14(7):643–644, 2017.

[6] M. David, G. Fertin, and D. Tessier. SpecTrees: An Efficient Without a Priori Data Structure for MS/MS Spectra Identification. In *Algorithms in Bioinformatics (WABI)*, volume 9838 of *Lecture Notes in BioInformatics*, pages 65–76, 2016.