analytical chemistry

Article

Data-driven rescoring of metabolite annotations significantly improves sensitivity

Ana S. C. Silva, Andrew Palmer, Vitaly Kovalev, Artem Tarasov, Theodore Alexandrov, Lennart Martens, and Sven Degroeve

Anal. Chem., **Just Accepted Manuscript** • DOI: 10.1021/acs.analchem.8b03224 • Publication Date (Web): 06 Sep 2018

ACS Publications

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Data-driven rescoring of metabolite annotations significantly improves sensitivity

Ana S. C. Silva,[*,†,‡,¶] Andrew Palmer,[§] Vitaly Kovalev,[§] Artem Tarasov,[§] Theodore Alexandrov,[§,‖] Lennart Martens,[*,†,‡,¶] and Sven Degroeve[*,†,‡,¶]

†*VIB-UGent Center for Medical Biotechnology, Ghent, 9000, Belgium*

‡*Department of Biochemistry, Faculty of Medicine, Ghent University, Ghent, 9000, Belgium*

¶*Bioinformatics Institute Ghent, Ghent University, Ghent, 9000, Belgium*

§*Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, 69117 Germany*

‖*Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, CA 92093, La Jolla, USA*

E-mail: anascsilva@vib-ugent.be; lennart.martens@vib-ugent.be; sven.degroeve@vib-ugent.be

Abstract

When analysing mass spectrometry imaging datasets, assigning a molecule to each of the thousands of generated images is a very complex task. Recent efforts have taken lessons from (tandem) mass spectrometry proteomics and applied them to imaging mass spectrometry metabolomics, with good results. Our goal is to go a step further in this direction and apply a well established, data-driven method to improve the results obtained from an annotation engine. By using a data-driven rescoring strategy, we are able to consistently improve the sensitivity of the annotation engine while maintaining control of statistics like estimated rate of false discoveries. All

the code necessary to run a search and extract the additional features can be found at `https://github.com/anasilviacs/sm-engine`, and to rescore the results from a search in `https://github.com/anasilviacs/rescore-metabolites`.

Introduction

Imaging Mass Spectrometry (IMS) is a technique that couples mass spectrometry's ability to resolve molecules of a complex sample with the ability to spatially localize them ([1]). It has been widely adopted for the molecular analysis of biological samples, as it allows to visualise the distribution of molecules across a sample without the need for antibodies or chemical labels ([2]). Recent technological developments have unlocked higher resolving powers for small molecule analysis ([3]) and so the field of metabolomics has shown great interest in the technology, as the addition of spatial localization of a compound can add new insights in the interpretation of the biological mechanisms it is involved in.

An IMS dataset is obtained by generating many mass spectra from a single biological sample where each spectrum is associated with one specific location in the sample (typically following a grid-like pattern). The distribution of a detected ion can be visualised by plotting the intensity of its mass spectral peak as an image, like the one in Figure 1. As these datasets are untargeted, many thousands of peaks are detected per pixel leading to thousands of potential images.

A critical step in the analysis of IMS data is assigning which molecule led to each of the thousands of images. It is clearly unfeasible to perform this task manually so automatic, accurate and sensitive annotation is required. Recently, automatic annotation systems (called annotation engines) able to annotate large IMS datasets have been proposed. One of them, *massPix* ([4]), annotates an IMS dataset by using a simple approach, already widely used in the field: exact mass matching the peaks from the dataset to a database. Specifically, *massPix* does so against a database of lipids and accounts for mass differences due to measurement-introduced fragmentation or biological modifications. This software

package allows for posterior analysis of the annotated mass images, but does not provide a statistical framework to validate the obtained annotations. Another one was proposed by Palmer *et al.*[5] that is capable of annotating a dataset given a database of known metabolites. This annotation engine is inspired by the standard proteomics pipeline for identification of tandem mass spectra. The annotation engine consists of three main components: (1) a database of metabolites which are searched for within a dataset; (2) a scoring function that ranks all the possible matches; and (3) a target-decoy strategy to allow for the estimation of the false discovery rate, thus providing statistical control of the quality of the matches. Our work expands upon this statistically controlled approach. Each of these parts is detailed below.

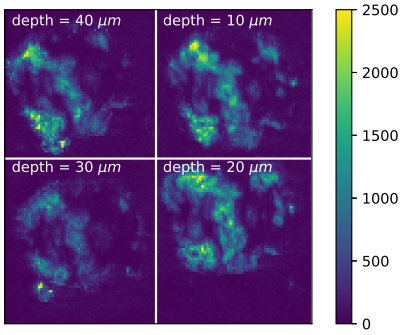

Figure 1: An m/z image from sections of a human colorectal adenocarcinoma, from m/z=253.2172. The intensity of each pixel corresponds to intensity of the spectrum which was acquired at that position, at the m/z value of 253.2172.

The annotation of metabolites starts from a database of metabolite molecular formulas. A set of candidate target ions is built by adding to each molecular formula a specific adduct (which depends on the instrument settings used). For each of these candidate ions a theoretical isotopic distribution is generated. The engine searches the data for images that correspond to the first four isotopes in the theoretical isotopic distribution. An empirical isotope pattern is calculated as the average image intensity for each isotope ion image.

The second step is the application of a scoring function that estimates the goodness of a match between the empirical and theoretical spectra and images. In Palmer *et al.*[5] this

scoring function is called the MSM (metabolite-signal match) score. This scoring function considers both spectral and spatial features (computed from the matched images). The calculation of the score is further detailed in the next section.

The third step concerns a statistical framework for deciding which ions can be annotated as present in the sample. All ions from every entry in the database receive a score but not all of them are actually in the dataset. A score threshold must be set that divides the entries into "present" and "absent". This raises the question of how to set such a score threshold. To answer this question, a null distribution of match scores is estimated and used to calculate the False Discovery Rate (FDR,[6]) for different match score thresholds. The null distribution is computed from a set of decoy ions built by adding to each metabolite molecular formula in the database a specific decoy adduct, from a predefined list of implausible adducts. The data is searched for these decoy metabolites in the same way as it was searched for target metabolites, and the null distribution is computed from the distribution of the decoys' MSM scores.

One issue identified by Palmer *et al.*[5] is that when different sets of decoy ions are used, inconsistent null distributions are obtained. The implication of this inconsistency is that the resulting set of annotated ions will be dependent on the selected decoy adducts. To overcome this, the authors propose to compute many null distributions from different sets of decoys, computing a score threshold for each of them and then taking the median of these score thresholds. The output of the annotation engine is, for each target metabolite ion, the MSM score and the FDR level it passes. This process is done separately for each subgroup of target ions, with each subgroup being defined by the adduct added to the metabolite.

In this manuscript we show that the accuracy of the metabolite match scores computed by this annotation engine can be significantly improved. This is done by exploiting more fine-graded annotation information and reformulating the construction of a scoring function as a data-driven multivariate classification task. This task is tackled by a semi-supervised machine learning approach as was proposed by[7] to increase the sensitivity of true peptide

matches in tandem mass spectrometry experiments. This approach consists of training a support vector machine (SVM) on the results of an annotation engine. The model is then trained to separate high confidence annotations from false positive annotations (i.e., decoys), and applied on the entire set of annotations, generating a new scoring that can more accurately separate the two. This results in more annotated spectra at the same level of FDR. A more detailed explanation of this method is presented in the following section.

Methods

Data-driven methodologies that aim to improve signal identification sensitivity have demonstrated their usefulness in proteomics as a tool to more accurately score candidate peptide-spectrum-matches (PSMs). One such tool is Percolator ([7]). It is a semi-supervised two-class machine learning approach that exploits additional information in the form of features that describe a candidate match to more accurately score PSMs, which results in significantly more spectra identified at a fixed FDR level. We propose to adapt and apply this tool to the output from the annotation engine previously described.

In short, the Percolator algorithm describes each candidate match – against both the target and decoy databases – as a fixed size vector of features, such that each feature represents some property of the match that can potentially be used to judge its correctness. Initially a training set is compiled from all matches against the decoy molecule database, labeled as the negative class, and the most confident matches against the target molecule database, labeled as the positive class. In this case these confident matches are obtained by selecting only the ones that pass a certain FDR threshold, computed from the MSM score assigned by the annotation engine. Next, a linear classification model is fitted and optimized on this training set. The optimal linear model is then applied to rescore all the matches in the dataset (including the non-confident matches which were not included in the training set). The process of constructing a training set is repeated, this time using the scores obtained

from the fitted linear model to select the matches that represent the positive class instead of the MSM score. At this stage there are typically more representatives of the positive class than in the previous iteration. A new linear classification model is fitted on this new training set and then used to rescore all the matches. The whole process is repeated until no more additional matches against the target database can be labeled as positive, resulting in a final set of rescored molecules that should contain more annotations at the same initial FDR threshold. A more detailed explanation of the algorithm can be found in.[7] As a data-driven approach, the quality of the data is determinant of the quality of the obtained model, that is: the initial set of matches against the target database at a certain FDR threshold (the matches labeled as positive) is crucial, as these represent what the model will learn is a true match.

The goal of this work is to adapt this rescoring strategy to the metabolomics annotation engine proposed by Palmer *et al.*[5] This entails modifying the engine itself so that additional features can be extracted from each match to both target and decoy databases. Rescoring ions is done in batches and the results are aggregated in a manner similar to what is currently implemented by the annotation engine. As is standard in the field of proteomics we propose assigning statistical significance to each each match by computing q-values (i.e., the minimum FDR at which a match is deemed correct) ([8]).

Feature extraction

Throughout the annotation engine by Palmer *et al.*,[5] target and decoy hits are described by three metrics:

- ρ_{chaos}, measure of spatial chaos of the principal isotopic image. $\in [0, 1]$, where 1 signifies high spatial structure and 0 a lack of structure in the image (an initial version of this metric is detailed in[9]);

- $\rho_{spatial}$, the weighted average of the correlation between isotopic images. $\in [0, 1]$, where

1 describes high correlation between isotopic images and 0 low correlation between the images;

- $\rho_{spectral}$, the relationship between the theoretical spectrum and the empirical spectrum obtained from the data, $\in [0, 1]$, where 1 describes similar spectra and 0 different spectra.

The MSM score is calculated as $\rho_{chaos} \times \rho_{spatial} \times \rho_{spectral}$ and thus takes values also $\in [0, 1]$, where larger values are associated with a more likely match. These four values are recorded for each target metabolite ion.

We propose to improve the scoring function by means of applying a data-driven method. To do so, it is sensible to extract as much information as possible from each hit. This is done by including 30 features (shown in Table 1 and further detailed in *Supplementary Information 1*) calculated from the set of isotopic images, theoretical spectrum intensities and experimental spectrum intensities. These features contain information from each match that complement the 4 original metrics, resulting in a total set of 34 features per annotated ion.

Table 1: Additional features extracted from each ion match

feature	**description**
image_corr_xy	Pearson correlation between pair of isotopic images x and y. The correlation is calculated for every pair of images corresponding to the four matched peaks, for a total of 6 features.
snr	Signal-to-noise ratio of the principal isotopic image.
percent_0s	Percentage of zeros in the principal isotopic image.
peak_int_diff_x	Difference between theoretical and empirical intensity of each peak in the matched isotope spectrum. Total of 4 features.
percentile_x	From the distribution of pixel intensities for the principal isotopic image, the intensity value that corresponds to each 10th percentile. Total of 9 features.
quart_x	From the distribution of pixel intensities for the principal isotopic image, the intensity value that corresponds to each quartile. Total of 3 features.
ratio_peak_xy	Ratio between the intensities of the empirical isotopic peaks, where x and y correspond to the index of each peak. Total of 6 features.

Semi-supervised learning

After a dataset is annotated by the engine, the rescoring method can be applied. The iterative rescoring process is started by selecting from all of the matches to the target database the highest confidence ones. This is done by setting an FDR threshold. For the analysis that follows, we choose an FDR threshold of 5%. This value was chosen after inspecting the annotation engine's output, by trying to maximise the amount of training data without compromising its quality too much. Different thresholds may be needed for different datasets and as such an initial analysis of the annotation engine's results is always recommended.

The null distribution variability issue, which the authors identify and that leads to the decoy sampling strategy described in a previous section, is also relevant for the rescoring procedure. The null distribution represents whe the model will learn to represent the negative class, and variability in these representative examples can translate into different learned scoring functions. To minimise those effects we follow the same strategy as what is currently implemented by the annotation engine. We sample 20 groups of decoy ions for each group of target ions. The sampling process is done so that each molecular formula appears once in each set. This way only ions for which matching images were found in the data are included. With this we make sure that there is a non-empty 34-length feature vector for every point in the dataset. Percolator is executed 20 times for each target ion set, each time with a different decoy set. From these 20 iterations we get a list of 20 q-values for each target metabolite ion. Similarly to how the annotation engine aggregates the results, we take the median q-value over these 20 iterations as the definitive q-value for each ion.

The output of this tool consists of a CSV file containing the list of all annotated target metabolite ions along with the aggregated q-value for each. Optionally, the decoys' q-values can be returned as well, as can the intermediate q-values resulting from the 20 null distribution samplings.

Results

A gold standard for proper validation of computational metabolite annotations is not currently available. However, the sensitivity and the reproducibility of these annotations can be evaluated using a set of technical replicates - i.e, multiple observations of the same sample done in the same experimental conditions. Building such a dataset constitutes a challenge in IMS since the same biological sample cannot be ionised multiple times. In this manuscript, we considered sequential sections of the same tissue processed with the same instrument and protocol as a sufficient approximation of a set of technical replicates. When analysing the results, it must be emphasized that some biological together with technical variability ([10]) always needs to be taken into account when addressing the reproducibility of a IMS experiment. As such, we do not expect an absolute overlap in the sets of annotations obtained for each section.

The analysis presented here was performed on a benchmarking dataset that consists of IMS measurements on 48 neighbouring sections from the same human colorectal adenocarcinoma (available from the MetaboLights repository,[11] **MTBLS415**). This dataset was generated using desorption electrospray ionisation (DESI) IMS according to the procedure outlined in the associated publication ([12]). Additional steps prior to the annotation of the sections are detailed in *Supplementary Information 2*

The 48 datasets were searched using the modified version of the annotation engine by Palmer *et al.*[5] against the Human Metabolome Database (HMDB,[13]). Since the data was obtained using the negative mode of the instrument, the target adducts are $-H^+$ and $+Cl^-$ as per the suggestion in Palmer *et al.*[5] The annotation engine calculates and stores five FDR thresholds: 1%, 5%, 10%, 20% and 50%. Each annotated ion is associated to the value that corresponds to the lowest of these thresholds that it passes. If it does not pass any threshold, it is assigned a FDR value of 100%. This is the output of the annotation engine, along with the 34 features that describe each match.

A possible measure for validating the annotations obtained is a quantification of the

overlap between them for all the datasets (i.e. sections) being studied. To achieve this across the 48 datasets, we count how many times each metabolite ion is annotated. If there is significant overlap, we expect most ions to be annotated in all 48 datasets; if there is not, they will be annotated in fewer. This count is done for the annotations obtained from the annotation engine and the rescoring pipeline at an estimated FDR of 5%. The values are shown in Figure 2.

Looking at the bars that correspond to engine annotations, we see not many ions are found in all 48 datasets, with the majority being found in 8 or less. As explained above, this imposes a limit on the results of the rescoring algorithm, as it implies that the initial set of annotations (a sort of seed for the rescoring process) is distinct across the datasets. Nevertheless, several metabolite ions are annotated by the rescoring method in 43 or more datasets. This means that despite the differences in the starting set of annotated ions, the learned scoring functions consistently pick up certain metabolite ions. A full list of which ions are annotated at 5% FDR with each method and in how many datasets can be found in *Supplementary Information 3*.

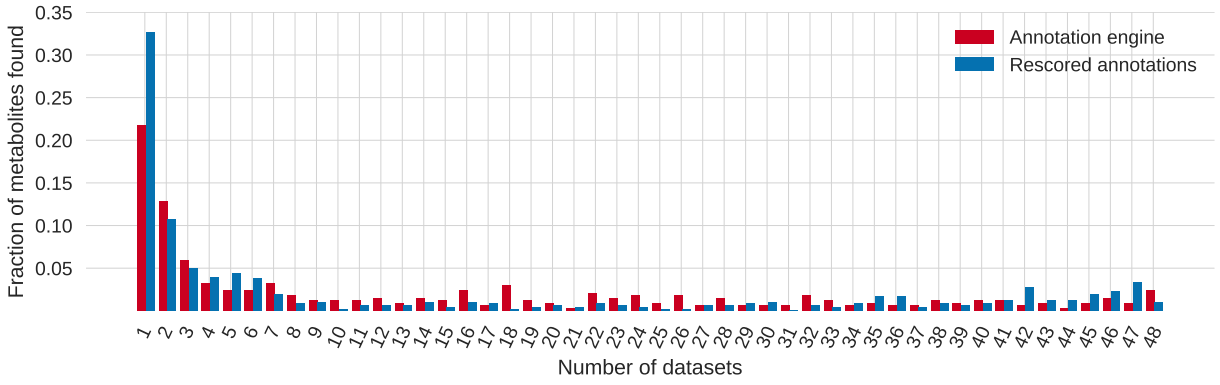

Figure 2: In the x-axis is the number of datasets in which a metabolite ion is annotated, and on the y-axis the fraction of ions annotated. Red and blue represent the engine's and the rescored sets of annotations, respectively. The height of the bar represents a normalized value such that for each method the sum of all columns equals 1.

To make the comparison between the results obtained with the annotation engine by itself and after the application of the rescoring pipeline clearer, we take the set of all metabolite

ions identified by each technique (for all the 48 datasets), and compare the overlap between these two sets. The overlap in obtained annotations is shown in Figure 3

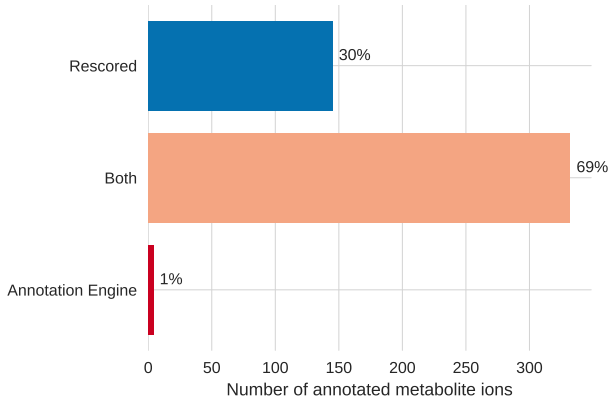

Figure 3: Number of annotated metabolite ions split into the ones found exclusively by the METASPACE engine (red), by the rescoring pipeline (blue) and the ones found by both methods (orange).

In the annotation engine results, a lot of ions have an MSM score of 0. This is due to one of two situations: either no match was found for that candidate ion, or one of the three MSM features has a value of 0. In the case when no match is found, we discard the candidate ions from the results, as it is not possible to calculate features in these situations. The second case is more interesting: due to the nature of the MSM score, it is enough that one of the three features that compose it to be zero for the match score to be zero for the score to be zero too. Such behaviour restricts what is considered a good match – all three metrics must have high values for the MSM score to be high. The converse side of this is that in some cases, even if two of the metrics have high values, it's enough for the third to be zero (or close to it) for the match to get a very low MSM score. This can lead to situations where potentially interesting metabolite ions are discarded from the results. When rescoring the matches, a linear combination of features is used. The nature of this type of model makes it less likely that a single feature would be enough to cause that match's score to be the minimum value in the range of allowed scores. Because of this, most of the matches that have an MSM score of zero (and thus don't pass any of the FDR thresholds) tend to be

distributed along a wider range of q-values after being rescored.

The annotation engine by Palmer *et al.*[5] returns to a user five sets of annotated ions, corresponding to five possible FDR levels. Having continuous q-values, we can analyse the rate at which the number of annotations increases as we allow more false discoveries into the results. Figure 4 represents this relationship for the aggregated annotations over the 48 datasets obtained with each method.

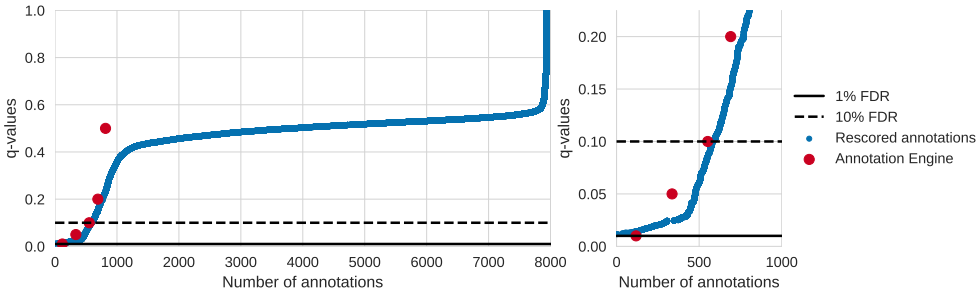

Figure 4: Trade-off between the amount of false discoveries and the number of identified metabolites. Ideally there would be a high number of annotations (rightmost side of the plot) with a low estimated FDR (lower part of plot area). The plot on the right is an enlarged view of the low FDR region.

The trade-off between number of annotations and rate of false discoveries is seen as the number of annotations obtained at higher FDRs increases. The comparison between the two methods shows that for FDRs above 2% the number of annotated metabolites is larger in the rescored method. The number of annotated ions the engine returns eventually stabilizes at around 1000 annotated ions; this is because there are a lot of matches that do not pass any of the 5 FDR thresholds. These correspond mostly to matches that have an MSM score of 0. In comparison, the number of annotations obtained with the rescoring method keeps increasing with the increase of allowed FDRs, until about 8000 annotations.

Since the proposed method relies on the random sampling of decoy sets and posterior aggregation of the results obtained, a study of the robustness of this approach is necessary. To do so, we select one of the tissue sections (the one that corresponds to a depth of $120\mu m$) and repeat the rescoring process ten times, each time allowing for different sets of decoys to be sampled. This test tells us if the aggregation scheme proposed is able to offset the null

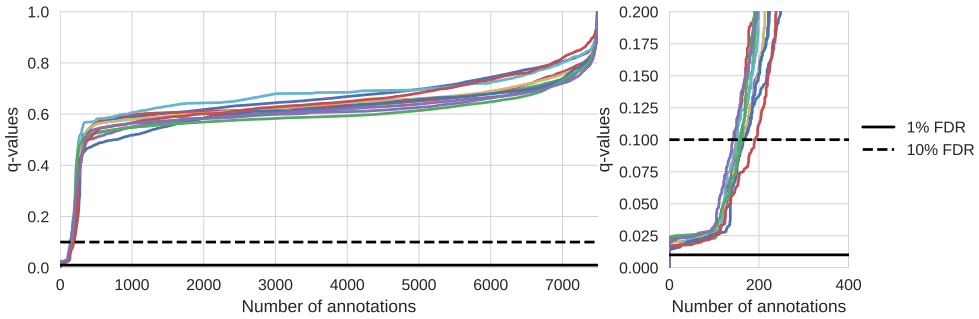

Figure 5: Trade-off between the FDR and number of annotated metabolites. Each line is one of ten repetitions of the rescoring pipeline.

distribution variability discussed previously. In Figure 5, a plot of the trade-off between the FDR and the number of annotations is shown. In Figure 6 the number of annotated ions at an FDR of 5% in each of these repetitions is shown.

The results are consistent particularly at the low FDR region – the region of interest, shown on the right side plot of Figure 5. For larger estimated FDR levels, as more false positives are included, it isn't surprising to see more variation. Figure 6 shows that there is consistency in the number of annotated metabolites in each repetition at the same FDR level – 124 ± 7.

More relevant than the number of annotated metabolites in each repetition is the overlap in the annotations themselves. This can be quantified by counting in how many repetitions each metabolite is annotated, as seen in Figure 7, which shows that most ions are consistently annotated in the 10 repetitions of this experiment.

During sample preparation, the tumour was sectioned at different depths; following that, groups of four adjacent sections were prepared for the IMS experiment. Given that there is a three-dimensional structure to the data, it can be surmised that similar metabolites will be clustered around similar locations, that is: adjacent sections will return similar sets of annotations. This can be verified by quantifying the overlap, in percentage, between pairs of sections that are close together and pairs of sections which are far apart. However, from this experiment, it is not possible to separate the effects from proximity between sections from

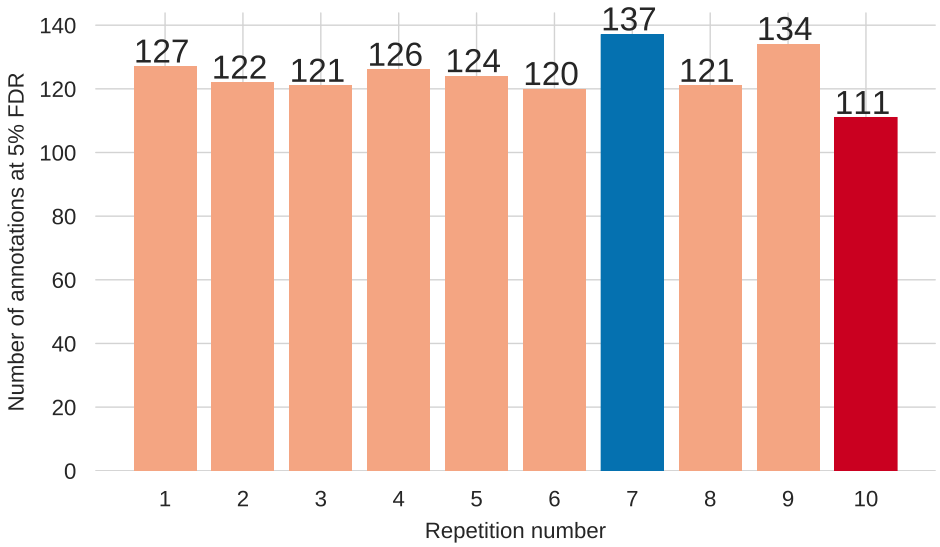

Figure 6: Number of annotations at 5% FDR obtained in ten repetitions of the rescoring pipeline. The number of annotations at 5% FDR is written on top of each bar; the repetition that yielded the more annotations bar is coloured in blue, and the one that yielded less is coloured in red.

the effects of the sections being obtained from the same mass spectrometry experiment. So the overlap seen can be attributed to two things: the biological similarity of the tissue, or the technical similarities from having been imaged during the same run of the mass spectrometer. This experiment is performed for all pairs of sections from the top-most set of sections (10, 20, 30 and 40 μm) and bottom-most (490, 500, 510, 520 μm), and the overlap is calculated for each pair, with each method. "Far" corresponds to the overlap between a section from the first set and one from the second, and "close" to sections from the same set. A boxplot of these values is shown in Figure 8.

We consistently observe a bigger overlap in annotations in the rescored set of results. Despite some exceptions, a trend is noticeable hinting at a bigger overlap between sections which are closed together compared to sections which are farther apart.

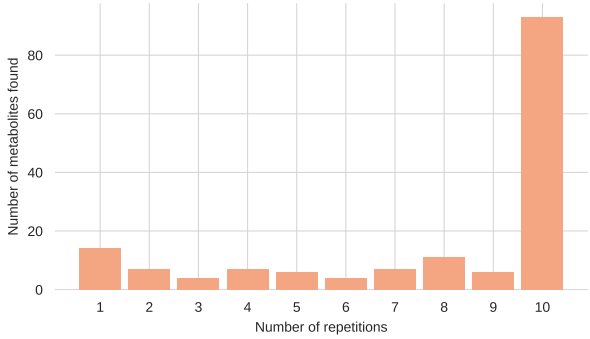

Figure 7: Number of repetitions of the rescoring pipeline where each metabolite is annotated at 5% FDR.

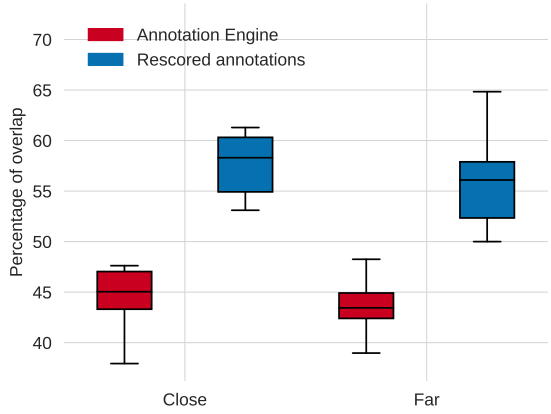

Figure 8: Overlap in annotations at 5% FDR (in percentage) between pairs of sections, split by method and by inter-section distance.

Conclusion

The method here presented takes from the more mature field of proteomics mass spectrometry and successfully applies it to the field of spatial metabolomics. Rescoring metabolite annotations in a data-driven manner, specific to each dataset under study, leads to an improvement in the amount of identified compounds at a certain level of FDR. Furthermore, by applying statistical tools that quantify an identified spectrum's quality to IMS data, we provide more information to the user, allowing for more nuanced interpretation of results. Furthermore, we provide this method as a post-processing step rather than as part of the

annotation engine's scoring scheme. This approach has the advantage of making it automatically adaptable to future iterations of the METASPACE annotation engine, which is undergoing active development cycles.

The scoring functions which are learned from the data improve the discrimination between annotated ions. The MSM score is very strict: if only one out of the three metrics that compose it has a value close to 0, then the MSM score will tend towards 0 as well. This results in a lower ability to differentiate between matches. The proposed approach leverages more out of the data by using more (and more diverse) features to describe each of the ions matched by the engine. This increase is complemented by the data-driven approach used to rescore the annotations. It would be a complex task to analytically choose how to weigh and combine each feature, and it would be an extremely complex task to guarantee that such a function would generalize well on different datasets, instruments and types of samples. That is not necessary by using the data-driven framework we propose. The weights and combinations of features are adapted to each of the datasets processed in an unbiased way. This type of technique opens up a lot of possibilities: in the future, additional features that explore different aspects of the data can be added to this framework, and their importance will be weighed in a data-driven manner. Although the task of annotating a IMS data is still a complicated matter, the method here presented can consistently improve on the results obtained by the annotation engine which was proposed byPalmer *et al.*[5] by itself.

Programming Language and Computational Tools

The methods described were programmed in Python 3.6 ([14]), resorting to the NumPy and SciPy ([15]) libraries for scientific computing, and the Pandas ([16]) data analysis library. Percolator version 3 ([17]) is used to rescore the obtained annotations.

Acknowledgement

The authors acknowledge funding from the European Union's Horizon 2020 Programme under Grant Agreement 634402 (PHC32–2014)

Supporting Information Available

Additional features calculated to better describe each match between a metabolite ion and a spectrum/set of images, data preparation steps before using the METASPACE annotation engine and a complete list of annotated metabolite ions along with a count of how many experiments it is found in with both the annotations engine by itself and the rescoring method.

References

(1) Dueñas, M. E.; Klein, A. T.; Alexander, L. E.; Yandeau-Nelson, M. D.; Nikolau, B. J.; Lee, Y. J. High spatial resolution mass spectrometry imaging reveals the genetically programmed, developmental modification of the distribution of thylakoid membrane lipids among individual cells of maize leaf. *Plant J* **2017**, *89*, 825–838.

(2) Watrous, J. D.; Alexandrov, T.; Dorrestein, P. C. The evolving field of imaging mass spectrometry and its impact on future biological research. *Journal of Mass Spectrometry* **2011**, *46*, 209–222.

(3) Römpp, A.; Guenther, S.; Takats, Z.; Spengler, B. Mass spectrometry imaging with high resolution in mass and space (HR2 MSI) for reliable investigation of drug compound distributions on the cellular level. *Analytical and Bioanalytical Chemistry* **2011**, *401*, 65–73.

(4) Bond, N. J.; Koulman, A.; Griffin, J. L.; Hall, Z. massPix: an R package for annotations

and interpretation of mass spectrometry imaging data for lipidomics. *Metabolomics* **2017**, *13*, 128.

(5) Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandrov, T. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods* **2016**, *14*, 57–60.

(6) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B* **1995**, *57*, 289–300.

(7) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **2007**, *4*, 923–925.

(8) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. *Journal of Proteome Research* **2008**, *7*, 40–44.

(9) Alexandrov, T.; Bartels, A. Testing for presence of known and unknown molecules in imaging mass spectrometry. *Bioinformatics* **2013**, *29*, 2335–2342.

(10) Ellis, S. R.; Bruinen, A. L.; Heeren, R. M. A. A critical evaluation of the current state-of-the-art in quantitative imaging mass spectrometry. *Analytical and Bioanalytical Chemistry* **2013**, *406*, 1275–1289.

(11) McKenzie, J.; Takats, Z. MTBLS415. 2016; `https://www.ebi.ac.uk/metabolights/MTBLS415`.

(12) Oetjen, J.; Veselkov, K.; Watrous, J.; McKenzie, J. S.; Becker, M.; Hauberg-Lotte, L.; Kobarg, J. H.; Strittmatter, N.; Mrz, A. K.; Hoffmann, F.; Trede, D.; Palmer, A.; Schiffler, S.; Steinhorst, K.; Aichler, M.; Goldin, R.; Guntinas-Lichius, O.; von Eggeling, F.;

Thiele, H.; Maedler, K.; Walch, A.; Maass, P.; Dorrestein, P. C.; Takats, Z.; Alexandrov, T. Benchmark datasets for 3D MALDI- and DESI-imaging mass spectrometry. *GigaScience* **2015**, *4*, 20.

(13) D. S. Wishart, Human Metabolome Database: completing the "human parts list". *Pharmacogenomics* **2007**, *8*, 683–686.

(14) Python Software Foundation, Python 3.6.4. `www.python.org`.

(15) Walt, S. v. d. The Numpy Array: A Structure for Efficient Numerical Computing. *Computing in Science and Engineering* **2011**, *13*, 22–30.

(16) McKinney, W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. 2010; pp 51–56.

(17) The, M.; MacCoss, M. J.; Noble, W. S.; Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *Journal of The American Society for Mass Spectrometry* **2016**, *27*, 1719–1727.

Graphical TOC Entry