Poster Note 64421

Validating and Comparing Component Detection Algorithms for LC-MS Data Assignment

Jane Razumovskaya, Joseph Brown, David Wright, Richard Baran, Iman Mohtashemi
Thermo Fisher Scientific, San Jose, CA, USA

Overview

Purpose: One of the major problems in evaluating performance of chromatographic feature detection algorithms is lack of completely annotated datasets. Creating an annotated dataset requires creating a known mixture, analyzing it with a mass spectrometer and manually verifying the presence of each compound. Besides being a painstakingly long process, creating and verifying a truly complex mixture is almost impossible with a manual process. Additionally confirmation of existence of the mixture requires manually reviewing large number of scans that can become very tedious and time consuming. Further Component Detection has another layer of complexity compared to chromatographic peak detection, where component detection involves not only peak detection but also includes identifying isotopes and adducts and grouping the relevant information correctly. Visualizations describing various aspects of data can speed up the annotation process. Here we describe a semi-automated annotation method and a viewing tool, which we named TotalRecall, since it involves examining every ion. It is used in creating exhaustively annotated datasets for feature detection, by means of grouping all objects into True Positive (TP) and True Negative (TN) categories. This allows for evaluating and comparing feature (and can be extended into component) detection algorithms using various metrics including Precision-Recall or ROC curves.

Methods: Several datasets were annotated using the exhaustive TotalRecall approach. The monoisotopic ions from isotopic features become True Positives, the other isotopes – True Negatives . All ions that do not form isotopic-features are considered as True Negatives. Then we compare performances of feature detection algorithms using Precision-Recall curves against both the exhaustive annotations and the original manual annotations (targeted) which contain only a few True Positives.

Introduction

The evaluation of feature detection algorithms is typically done by analyzing a sample with known content, i.e. where there is a known answer key for confirmation of the detection or non-detection of every expected peak by multiple criteria. The criteria may involve m/z of each analyte, retention time, peak area/intensity, isotopic cluster, formula/peptide sequence, charge state, etc. The criteria depend on the goal and sophistication of the evaluated algorithm. The basic criteria are typically just the m/z and retention time which are the most easily observed by the annotator and calculated by the algorithms. These targeted answer keys for each sample are typically provided by the experimental scientists who provided the sample and performed the mass spectrometry experiment.

There are three types of samples that we can consider for annotation: a) a simple mixture of one/several pure analytes b) a complex mixture of pure analytes spiked into complex matrix and finally c) a mixture where some of the contents are known based on the biologically significant function.

In the first case, the presence of all the expected analytes has to be verified by hand and recorded into the answer key to be provided along with the collected spectra. This is a trustworthy method, which largely guarantees correct answer key and a good evaluation process. Using the method we can assess that the algorithm detects all the expected compounds and none others (assuming a very pure sample), providing both space for True Positive and True Negative identifications, where True Negatives are all the identifications that are not True Positives. However, it cannot be used to mimic the real life scenario where the analytes of interest are mixed in a complex biological matrix such as urine or blood. And while they provide useful assessment into how well algorithms detect expected analytes and none other, they fail to stress algorithms for complex cases. Essentially while we can expect both TP and FP rate assessment of algorithms with this analysis, majority of feature detection algorithms will behave very similarly under this test.

Second and third cases can be used as the stress test for the algorithms, due to the presence of complex background which can affect peak shapes, cause real-life interferences and test dynamic range. These approaches require even more stringent manual analysis to observe the evidence for the presence of the expected analytes. However, they are still only useful for providing a list of limited True Positives to be detected by each algorithm as only verified and recorded analytes can be monitored. Then, all the features that are present in the sample, but not expected by the annotator will not appear in the list of True Positives, however, they should not be included in True Negatives, as they are real features that should be detected by the algorithms. In this case some algorithms will detect the complex cases better than others, however, they may also detect a lot of True Negatives, which are not provided in the answer key, Thus, this approach does not present us with a good assessment on the FP rate of an algorithm.

Here we propose a method that stresses feature detection algorithms. Below we show an example of performance of an algorithm using Precision-Recall curve on the sparsely annotated dataset based on targeted annotation of spiked chemicals vs on the exhaustive annotations provided by the tool. The Precision-Recall curve was chosen over the ROC curve due to the high skew of the annotations towards number of True Negatives, as well as the absence of real biological significance in case of True Negatives in feature detection.

Methods

Sample Preparation

Several datasets were annotated. A buspirone dataset is a 0.5uM incubation of buspirone with rat hepatocytes which causes oxidative metabolism. For this example the 5 minute time point was chosen from a time-course study, and a selected time slice of 2.5 minutes, containing buspirone and it's oxidation products. Two Amino Acid datasets (positive and negative modes) generated from a sample of pure 17 amino acids from Amino acid STD H kit: http://www.piercenet.com/browse.cfm?fldID=CA420DFB-42E8-4333-B461-974BD14C1353

Mass Spectrometry

The data were collected on Thermo Scientific™ Q Exactive™ MS instruments.

Data Analysis

Consider all the m/z signals in a scan:

1. Group them into isotope clusters (every m/z signal becomes either as an A0 or an isotope to an existing A0 on a scan basis).
2. The True Positive and True Negative lists for feature detection are then compiled using that information in the following fashion:
3. Compile the True Positive list by assigning all A0s with isotopic clusters,
4. Compile the True Negative list by assigning all A0s with no isotopic clusters ('orphans'), The A1 through An isotopes are assigned to the True Negative list as they should not be considered as separate features by a feature detection algorithm.

TotalRecall (in-house semi-automated annotation tool)

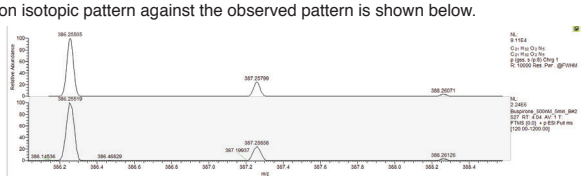
Using the TotalRecall visualization tool, the lists are examined manually to confirm the True Positive list by looking at the chromatographic peak shape of the A0 and its correlation to the corresponding isotopes. The well correlated A0s are retained as True Positives, in cases where the isotopes did not appear to be well correlated with the A0, the A0 is moved to True Negative list (illustrated in Figure 4).

Precision-Recall curve generation:

$\text{Precision} = \frac{TP}{TP+FP}$ $\text{Recall} = \frac{TP}{TP+FN}$, calculated for varied algorithm thresholds,

where TP stands for the number of True Positive features detected by the algorithm, FP stands for number of detected features that don't match to True Positives, and FN stands for number of True Positives not detected by the algorithm.

FIGURE 1. This is a buspirone isotopic pattern,(C21H31N5O2) the alignment between the simulation isotopic pattern against the observed pattern is shown below.



Thermo SCIENTIFIC

FIGURE 2. – Example of a True Positive. The pattern is automatically detected in each scan across the peak, then the chromatographic alignment of the A0 against it's isotopes is examined, in case of aligned isotopes.

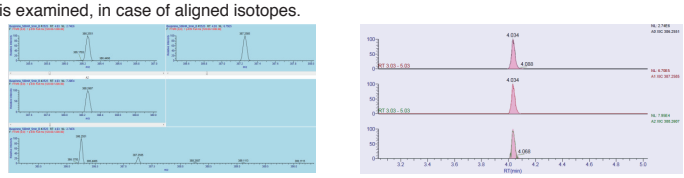
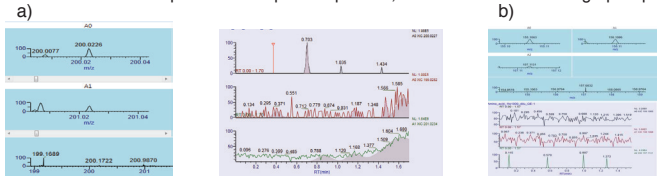


FIGURE 3. Example of True Negative. a) If the chromatographic peak is visible, however, all the potential isotopes are misaligned, the feature is annotated as True Negative. b) Another case is when potential isotopes are present, but there is no chromatographic peak.

a) b)



Results

FIGURE 4. The annotation procedure to determine what features are going into TP and TN lists. The Ignore list is utilized for storing features that will not be applicable to TP rate and FP rate calculations. The Unknown list is used for storage of ambiguous results, to be reviewed later.

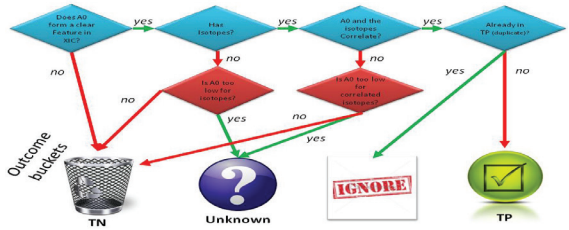


FIGURE 5. An example of visualization of a feature in TotalRecall- A0 with four aligned isotopes. The Clusters tab contains all the A0s with corresponding isotopes, The Isotopic Features True Positives tab contains all the annotations of True Positives, and the Isotopic Features True Negatives tab contains the True Negatives. The Ignore tab contains the list of features that are not True Positives, but which do not count against the detection algorithm. The Unknown tab is a placeholder for features which status is not yet resolved. By the end an annotation is fully completed, Unknown list is expected to be empty.

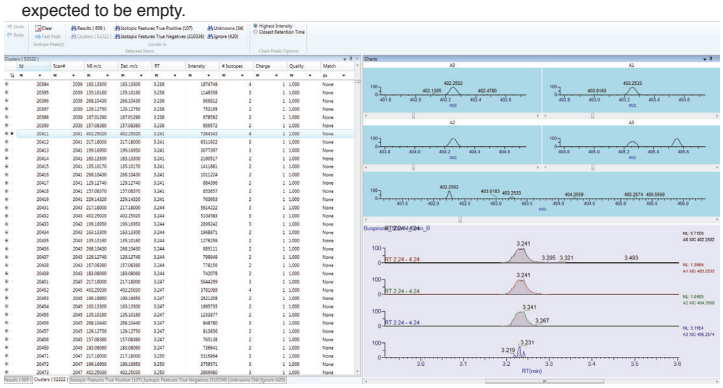


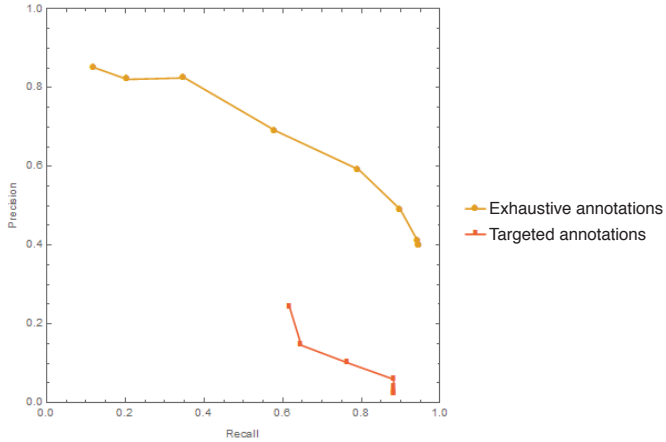
TABLE 1. Summary of numbers of targeted annotations made manually from the comparing the m/zs of spiked compounds (TP manual annotations), the number of annotations found by TotalRecall with manual curation (TP exhaustive annotations) and the number of true negatives found by TotalRecall (TN exhaustive annotations) within the selected time slices of each dataset. In case of Diclofenac, the targeted annotations were not available.

Datasets	RT-Range (min)	# of TP manual annotations	# of TP in exhaustive annotations	# of TN in exhaustive annotations
Buspirone	2.5-4.5	9	107	310336
AminoAcid+	0.5-1	27	127	84914
AminoAcid-	0.5-2.5	12	605	243053
Diclofenac	3-4	-	710	298067

Buspirone and Diclofenac datasets are complex samples with a high number of background compounds. AminoAcid+ is a slice of a file where the majority of peaks are related to the spiked standards, AminoAcid-, while done on the same sample, includes the solvent front. The number of manual annotations for even simple datasets, is a small fraction of all the features observed by TotalRecall assisted, semi-automated annotation method.

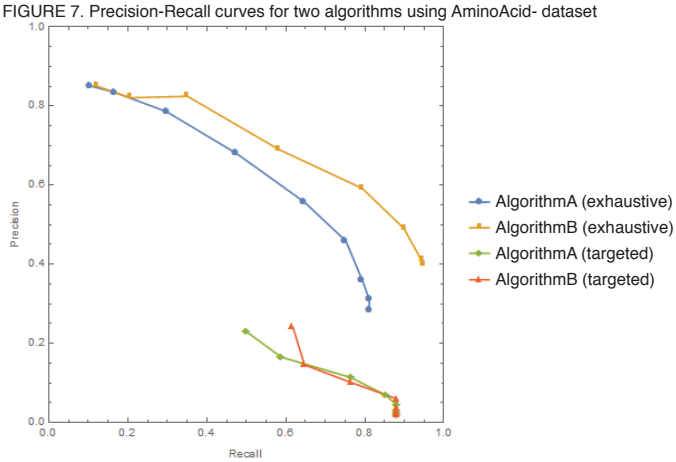
Comparison of the performance of a feature detection algorithm against an AminoAcid-targeted annotation (red, 12 peaks) and exhaustive annotation created using TotalRecall (yellow, 605 peaks). Individual data points on each curve correspond to different signal intensity thresholds for the feature detection algorithm. Using different signal intensity thresholds artificially decreases the recall of true positive peaks from annotations (as signal intensities of some annotated peaks fall below the threshold) but provide a view of detection of putative false positive peaks. The algorithm achieved recall close to 1 against both annotations with an apparently high false positive rate against a limited targeted manual annotation (orange). A large proportion of putative false positive peaks when comparing against manual targeted annotation (orange) are legitimate peaks which are absent from the annotation.

FIGURE 6. Precision-Recall curve for AminoAcid- dataset.



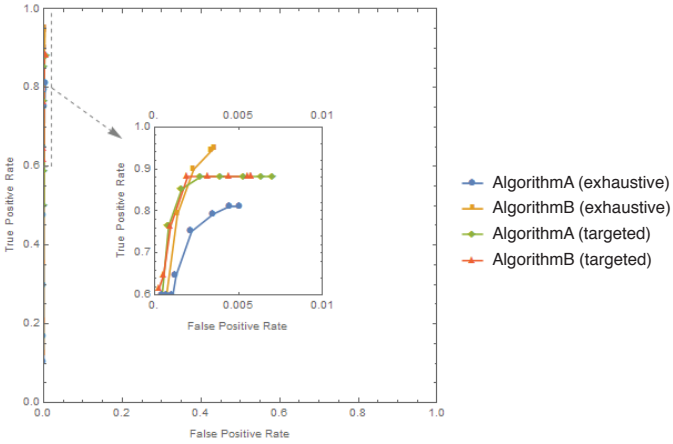
Comparison of the performance of two feature finding algorithms against a targeted annotation (red/green, 12 peaks) and exhaustive annotation created using TotalRecall (yellow/blue, 605 peaks). Individual data points on each curve correspond to different signal intensity thresholds for the feature detection algorithm. The algorithms show similar performance against the limited targeted manual annotation (red/green). The algorithms also show similar performance against the exhaustive annotation, but only when using high signal intensity thresholds for peak detection (yellow/blue, upper left corner of the plot). Clear differences between the performance of the two algorithms are visible against the exhaustive annotation when using lower signal intensity thresholds and thus detecting more peaks. It is the exhaustive annotation facilitated by TotalRecall that truly stresses the feature detection algorithms and allows the comparison of their performance.

FIGURE 7. Precision-Recall curves for two algorithms using AminoAcid- dataset



Receiver operating characteristic (ROC curve) is a popular alternative visualization of the performance of classification algorithms. Here, true positive rate (Recall) is plotted against false positive rate (FP/(FP+TN)). For peak finding algorithms, in an extreme case, any signal outside of detected peaks may be classified as a TN signal. Given a large number of such signals in typical LC-MS datasets, the false positive rates of peak picking algorithms become very small. For this reason, differences among the performance of alternative algorithms are highlighted more on Precision-Recall curves than on ROC curves (shown below).

FIGURE 8. ROC curves for two algorithms using AminoAcid- dataset



Conclusion

- The combination of the exhaustive annotations with the Precision vs Recall curves allows us to effectively assess and compare algorithm performances.
- The exhaustive annotations contain significantly higher number of True Positives as comparing to manually observing the presence of spiked compounds into the sample, allowing for a broad stress test of algorithm performance.
- Here we show a novel way to examine large-scale complex datasets to create an exhaustive annotation for feature detection with both True Positive features and True Negatives.
- TotalRecall supports visualizing the chromatographic alignment of isotopes along with the basic chromatographic peaks in order to confirm a True Positive feature.
- With a high skew towards number of True Negatives, Precision vs Recall curves are a better comparison matrix than the standard ROC curves.

Future Work

- Annotation of additional datasets
- Improvement of the TotalRecall user interface to improve the efficiency of manual annotation
- Compare multiple internal and external algorithms run under different conditions using the created annotations.
- Extend a single annotation sample to replicates and time-course studies, which requires a rigorous matching scheme between the observed features and the True Positives, both for intensity changes and retention time shifts.
- Extend to other areas beyond feature detection, such as pure isotope detection, adduct grouping and finally component detection.

Acknowledgements

We would like to thank Ralf Tautenhahn, Junhua Wang and Mark Sanders for providing samples and manual annotations for the work and Paul Gazis for invaluable discussion.

www.thermoscientific.com

Africa +43 1 333 50 34 0
Australia +61 3 9757 4300
Austria +43 810 282 206
Belgium +32 53 73 42 41
Canada +1 800 530 8447
China 800 810 5118 (free call domestic)
400 650 5118
Denmark +45 70 23 62 60
Europe-Other +43 1 333 50 34 0
Finland +358 10 3292 200
France +33 1 60 92 48 00
Germany +49 6103 408 1014
India +91 22 6742 9494
Italy +39 02 950 591
Japan +81 45 453 9100
Korea +82 2 3420 8600
Latin America +1 561 688 8700
Middle East +43 1 333 50 34 0
Netherlands +31 76 579 55 55
New Zealand +64 9 980 6700
Norway +46 8 556 468 00
Russia/CIS +43 1 333 50 34 0
Singapore +65 6289 1190
Spain +34 914 845 965
Sweden +46 8 556 468 00
Switzerland +41 61 716 77 00
UK +44 1442 233555
USA +1 800 532 4752

Thermo SCIENTIFIC
A Thermo Fisher Scientific Brand

PN64421-EN 0615S