**ImsMetabolitesFinder**

***Abstract –*** *ImsMetabolitesFinder is an application that isolates the presence of a target molecule in an IMS result files and use the information to analyze the instrument independent characteristics such as the mobility(K0) and cross sectional area(A). The program uses various peak analysis techniques to maximize the reliability the measurement section result.*

**II. Methods**

Cross section extraction is a very well-studied biological technique and has been around since <year>[cite something]. **<**Explainthe techniques of running ims with different voltages, get drift time, Linear fit, etc>. <Insert reasons> requires the accurate cross section with high degree of confidence and the ability to reject various sources of noise and fake detection. First of all, it is important to note that the mapping between signal and target ion presence can sometimes be very subjective, especially for low intensity peaks. As a result, the ImsMetabolitesFinder algorithm is designed with the idea of optimizing final cross section calculation reliability in mind.

Goals

Extract cross section area and mobility from IMS results reliably if the input dataset provides enough information or declare inconclusive if otherwise.

< introduces frame accumulation and why, then define voltage group>

*Facts/Assumptions*

[1] Points agree with each other during the linear fit.

[2] More than enough points were provided for the linear fit.

[B] Each point represents a high confidence drift time measurement of the target ion at a given configuration of drift tube voltage, temperature and pressure.

[3] Mobility/Cross section/Drift time math model accurately models correctly how drift time relates to mobility.

———————————————————————

**∴** [A] We are confident that final mobility / cross section confidence is accurate and we can use it for library construction.

[C] If the ion is indeed present, at the target Mz, there will be a peak with reasonable intensity, good peak shape and isotopic distribution closest to the theoretical isotopic distribution.

[1] If at the target Mz there are non-target chemicals making a presence along with the true target presence, intensity and peak shapes are uncertain but the isotopic profile will not be as good as real target isotopic profile.

[D] If at the target Mz there are peaks resulted from noise, it will have a low peak shape or a low intensity score, and a low isotopic score.

.———————————————————————

**∴** [B] The feature at a given Mz needs to have **reasonable** intensity score, **reasonable** peak shape score and **high and also** **highest** isotopic distribution to strongly indicate the target presence. As a result the feature’s drift time measurement is very likely the target ion drift time measurement.

[1] Voltage, temperature and pressure are stable for a voltage group.

[2] The voltage group has enough accumulations

———————————————————————

**∴** [C] If the ion is indeed present, at the target Mz there would be peak with reasonable intensity, good peak shape and isotopic distribution closest to the theoretical isotopic distribution.

[1] If there is a peak has good peak shape and reasonable intensity, it cannot be noise.

[2] Noise’s isotopic profile is usually [1,0,0.] .

———————————————————————

**∴** [D] If at the target Mz there are peaks resulted from noise, it will have a low peak shape or a low intensity score, and a low isotopic score.

[1] The feature peak has a wide peak

[2] Wider peak with an obvious Gaussian distribution gives us higher drift time accuracy because we can extract drift time scan resolution higher than what the instrument can provide.

———————————————————————

**∴** [E] Wider peak means a peak is preferable for analysis but it does not tell anything in target selection.

**III. Algorithm Description**

Input

The input files are UIMF files created by direct-injection IMS. The following assumptions are made:

1. The datasets have unknown number of different drift tube voltages, each repeated for unknown number of frames.
2. There may or may not be other chemicals in the MZ tolerance range.
3. If the target is only a Mz number. The target has to be the most abundant species within the MZ tolerance range.

Empirical formula analysis

Calculate theoretical isotopic profiles and target Mzs from the input empirical formula.

*Voltage grouping*

Group similar voltages. While reading the input file frame by frame, the program groups frames of similar voltages together. The program then extracts partial XICs within a certain mass tolerance around the target Mz (ppm = 10) and sums XICs from frames in the voltage group. The program then normalizes the summed XIC by number of frames used to sum(Averaging).

The frame would be classified into the same voltage group if it satisfies one of the following criteria.

1. The new voltage is within 3 standard deviation of the average voltage.
2. The new voltage is within ±5V of the voltage group’s average voltage.
3. If the new voltage is not doubling the standard deviation by a factor of 2.

*Feature detection and preliminary filtering.*

Smooth XICs with the 2D version of the watershed algorithm so artificial gaps will be filled if not already filled by accumulation. Because scoring is relatively computationally expensive, so the program then filters out really small peaks using Kelvin’s peak filter(filter level = 0.25), which is fast compared to scoring. Compute each feature’s drift time among other properties. If a feature’s drift time is at the edge of the UIMF heatmap, throw away the feature.

*Feature scoring*

Register features as a M/z, drift time pair. Compute the feature’s intensity score, peak shape score and isotopic score as following

1. Intensity score: Sum the intensities of the feature XIC.
2. Isotopic profile score: For every M/Z specified by the theoretical isotopic Mz, get the XIC around that theoretical Mz at the feature scan number. Sum the XIC intensities across frames of a particular voltage group. Remove saturated isotope peaks(IP). Compare the distance between Theoretical isotopic profile with observed isotopic profile using:

* Angle method
* Euclidean Distance

Normalize both vectors

∥x−y∥2

* Pearson Correlation
* Bhattacharyya Distance

1. Peak shape score:

KolmogorovSmirnovTest(IP)

Jaque-BeraTest

*Selecting drift time for target*

For each voltage group, remove features with low intensity score(< 10% global max). Remove features with low peak shape score (Jaque-Berra p value = 0.01). Remove features with low isotopic score(0.5). Select the feature with the highest isotopic score(currently factor in intensity to mitigate uncertainties from the isotopic score accuracy, sorted using 2 \* IntensityScore + IsotopicScore) from the remaining features and assign it to the voltage group(See conclusion B). Compute the voltage group stability score as the product of variances of the voltage group’s voltage, temperature and pressure.

Remove voltage groups with low voltage stability score (0.00000000000000001) and voltage groups without a best feature left. If there are many 0 voltage groups left, conclude NEG(negative, no ion presence found). If there are more than 0 but less than 3 such voltage groups, report NEP(ion presence were found but not enough points to fit to ion mobilty, the suggested course of action in this case is to add more voltages to experiment and it will probably be a POS).

*Mobility computation ( drift time linear fit)*

Calculate the mobility fit line using linear least squares. Reject outlier whose cook’s distance > 3. Then calculate the fit line again without outliers. Compute R squared as a metric of how fit points agree with each other. If the R squared above a threshold, conclude POS(positive) to indicate that the program thinks the cross section and mobility calculation are reliable.

*Create QA data.*

For analyses marked as POS, The linear fit/ outlier detection result is written as a plot. 3 Feature scores and the voltage reliability scores are averaged cross fit points and plotted into a stack bar diagram for comparison with other datasets regardless of POS, NEP, or NEG. Visual weights are added to the scores for better interpretation of the data.

<batch processor – provide means for automated drift time library construction>

