

Merged Spectrum for Metabolite Identification in MassBank

Hisayuki Horai¹, Masanori Arita^{1,2,3}, Yoshito Nihei¹, and Takaaki Nishioka¹

¹ Institute for Advanced Biosciences, Keio University & JST-BIRD

² Graduate School of Frontier Sciences & PRESTO (JST), University of Tokyo

³ RIKEN Plant Science Center

Overview

- For each metabolite, we prepared a "merged spectrum" by overlaying the mass spectra of the same metabolite measured on ESI-MS/MS under the different CID conditions.
- Two types of merged spectra, normalized and post-normalized, were prepared. Normalized type was by merging normalized mass spectra, while post-normalized type was normalized after merging raw mass spectra.
- In the present study, we evaluated how well the merged spectral library works as the reference spectra for metabolite identification.

Conclusions

Concerning the normalization, pre-normalization before merging obtains better results than post-normalization after merging. It shows that the base peak of each raw spectrum is important even if its intensity is small in raw mass spectra.

MassBank recently opens the spectral search service against the normalized merged mass spectral library. We will expand merged spectra in accord with the increase of raw spectra of metabolites.

Introduction

MassBank is a database of mass spectra for life sciences. Currently 10,679 mass spectra of 1,384 metabolites measured on different types of mass spectrometry are available. The critical problem for metabolite identification (Figure 1) on MassBank is low in the similarity among the mass spectra of each metabolite measured under different analytical conditions.

In this study, we propose a method to overcome the problem by combining those measured on ESI-MS/MS for each metabolite into one artificial mass spectrum called "merged spectrum". Users are able to search their mass spectral measured under different analytical conditions against the library of the merged spectra in MassBank.

Methods

Merged Spectrum: A single artificial spectrum generated from raw spectra measured under different conditions. We use two simple automatic generation methods shown in Figure 2. In this study, we make a merged spectra for each metabolite from raw spectra measured in different of collision energy and same ionization mode (i.e. positive and negative) on ESI-MS/MS.

Evaluation of Metabolite Identification: We compare metabolite identification for a set of raw spectra and for a set of merged spectra generated from the raw spectra. For each set of spectra, we execute spectral search in MassBank and calculate the precision (true positive ratio to positive), the recall (TPR, true positive ratio to all correct answers), FPR (false positive ratio to all incorrect answers). For the comparison among different sets of spectra, we use the maximum F-value (harmonic mean between precision and recall) and maximum AUC (area under FPR-TPR curve).

Results

We use 4,431 QqTOF-MS/MS spectra and 4,205 QqQ-MS/MS spectra of 898 metabolites measured in 2~5 levels of collision energy (Table 1).

Evaluation Run 1: We evaluate the relevance for the set of integrated spectra generated from QqTOF raw spectra using QqQ raw spectra as a query set.

Evaluation Run 2: We evaluate the relevance for the set of QqTOF raw spectra by using same query set of evaluation 1.

Comparison: The obtained results is shown in Table 2. The F-value and AUC for both of the merged QqTOF spectra are significantly greater than for raw QqTOF spectra. Furthermore, the pre-normalized merged spectra contributes to the improvement more than the post-normalized merged spectra.

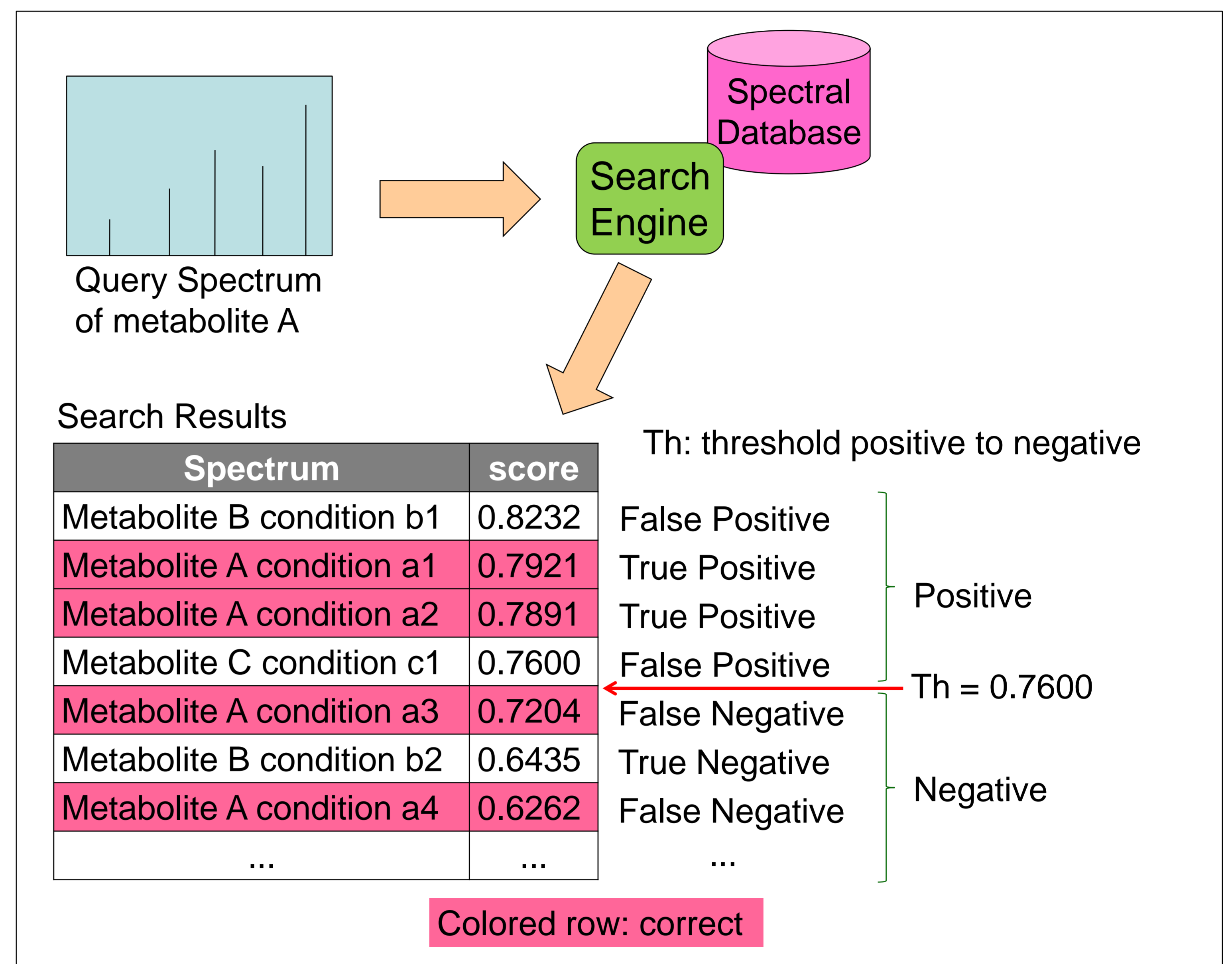


Figure 1: Metabolite Identification by Spectral Search

Figure 1 shows the overview of metabolite identification by spectral search. A searched database spectrum is correct if it is a spectrum of same metabolite. The difference of analytical conditions is not concerned.

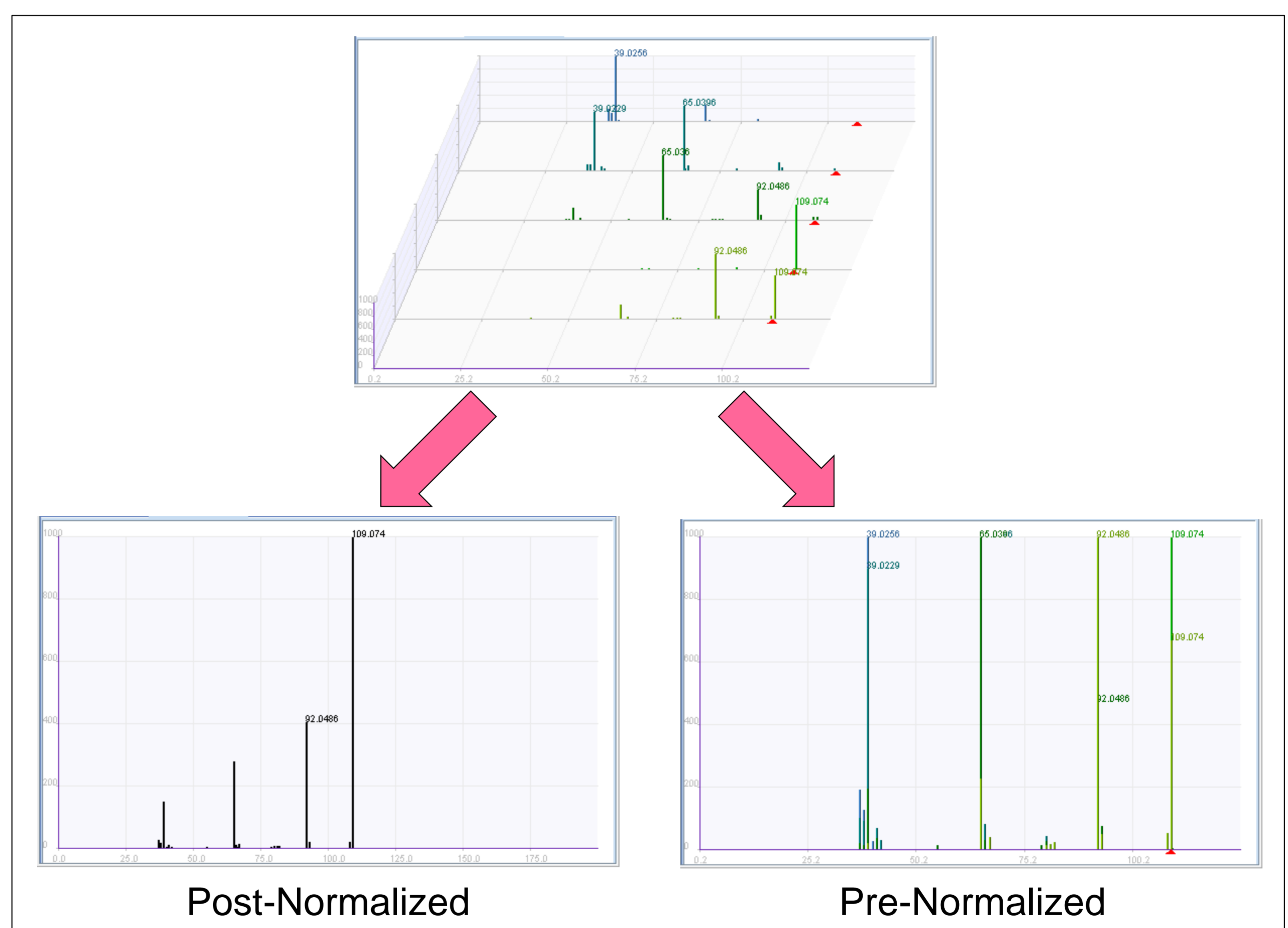


Figure 2: Merged Spectrum

Method to generate a pre-normalized merged spectrum:

1. For each raw spectrum, normalize the values of intensity.
2. Make the set union of (m/z , relative intensity) pairs of the normalized raw spectra.

Method to generate a post-normalized merged spectrum:

1. Make the set union of (m/z , intensity) pairs of raw spectra.
2. Normalize the spectra represented by the set union.

In making a set union of both generation, when there are peaks of same m/z in different raw spectra, then the largest value of intensity is selected.

Table 1: Spectral Set for Evaluation

Name	Description	Number of Metabolites	Number of Spectra
QUERY	QqQ raw spectra	860	4205
RAW	QqTOF raw spectra	898	4431
PRE	QqTOF pre-normalized merged spectra	898	898
POST	QqTOF post-generated merged spectra	898	898

Table 2: Summary of Evaluation Runs

No.	Database	Maximum F-value	Maximum AUC	Precision (Recall=0.5)	Recall (Precision=0.5)
Run 1a	PRE	0.416	0.924	0.34	0.29
Run 1b	POST	0.305	0.864	0.15	0.15
Run 2	RAW	0.288	0.735	0.02	0.10

Merged spectra get better results than raw spectra. Comparing the pre- and post-normalizations, the improvement rate of precision is larger than the rate of recall. It means that the weighting to base peaks contributes to precision more than to recall.

Evaluation Method

As shown in Figure 1, at first, scores for all spectra in spectral database are obtained from a search engine and counts the number of false/true positive/negative. Positive and negative spectra depends on Th, so then the numbers of true/false positive/negative also depend on Th.

TP(Th) : the number of true positive spectra

FP(Th) : the number of false positive spectra

FN(Th) : the number of false negative spectra

TN(Th) : the number of true negative spectra

In the next, the precision, recall, F-value, TPR, FPR and AUC are calculated for each Th as follows. The denominators of precision and recall are the number of positive spectra and all correct spectra respectively. AUC is the area of the tetragon whose vertices are (0, 0), (1, 0), (1, 1) and (FPR(Th), TPR(Th)).

$$\text{Precision(Th)} = \frac{\text{TP(Th)}}{\text{TP(Th)} + \text{FP(Th)}}$$

$$\text{Recall(Th)} = \text{TPR(Th)} = \frac{\text{TP(Th)}}{\text{TP(Th)} + \text{FN(Th)}}$$

$$\text{FPR(Th)} = \frac{\text{FP(Th)}}{\text{FP(Th)} + \text{TN(Th)}}$$

$$\text{F-value(Th)} = \frac{2 \cdot \text{Precision(Th)} \cdot \text{Recall(Th)}}{\text{Precision(Th)} + \text{Recall(Th)}}$$

$$\text{AUC(Th)} = \frac{1}{2} \cdot \text{FPR(Th)} \cdot \text{TPR(Th)} + \frac{1}{2} \cdot (1 + \text{TPR(Th)}) \cdot (1 - \text{FPR(Th)})$$

The evaluation for a query set is based on all scores for all spectra in the query set. The maximum F-value and the maximum AUC are the maximum value of F-value and AUC, respectively, for all Th between 0.0 to 1.0 as follows. The maximum F-value and the maximum AUC is independent from Th and depends only on the pair of the database and the query set.

$$\text{maximum F-value} = \max_{0.0 \leq \text{Th} \leq 1.0} (\text{F-value(Th)})$$

$$\text{maximum AUC} = \max_{0.0 \leq \text{Th} \leq 1.0} (\text{AUC(Th)})$$

MassBank is aimed to establish a public free mass spectral database based on researchers' voluntary. We would appreciate it very much if you could contribute spectra of metabolites and natural products to MassBank.

Spectral Search of MassBank

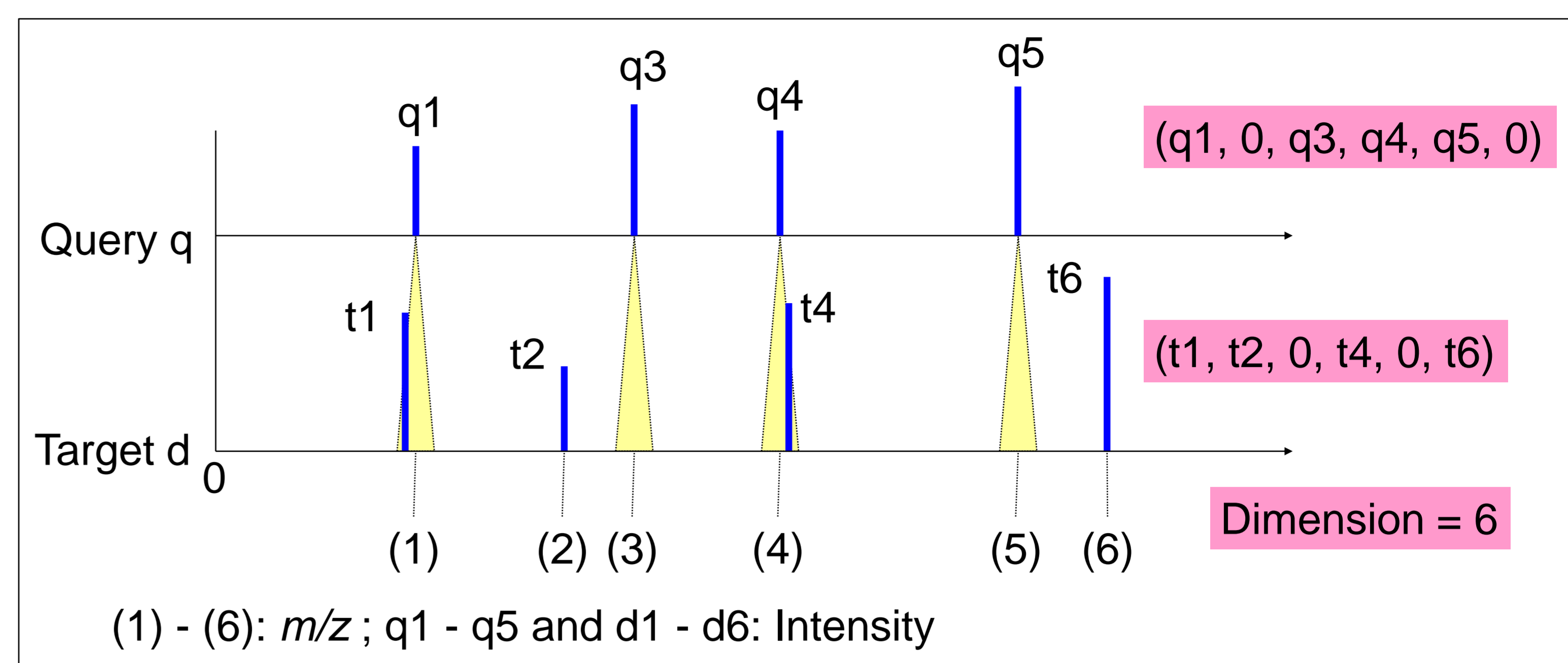


Figure 3: Vector Space of Real Number m/z

Spectral search of MassBank is based on cosine correlation. In order to use cosine correlation for real number m/z , it is critical to align m/z vectors of a query and a target by pruning an ignorable difference. We introduce a tolerance of m/z and ignore the difference of m/z with in the tolerance. In Figure 3, a yellow triangle denotes the range of the tolerance. The tolerance is user-specifiable in MassBank.

Scoring function using in MassBank is as follows. q and t are m/z vectors of a query and a target respectively. A relative intensity is represented in permillage. The exponents of a relative intensity and a m/z in normalizing function have been decided based on statistics of spectra in MassBank.

$$\text{Score}(q, t) = \cos(\theta) = \frac{q' \cdot t'}{|q'| \cdot |t'|}$$

where $q' = \text{norm}(q)$, $t' = \text{norm}(t)$

$$\text{norm}(x) = \left(\sum_i \text{in}_i^{1/2} \cdot \text{mz}_i^2 \right)$$

where $x = (x_i)$,

in_i = relative intensity of x_i , $\text{mz}_i = m/z$ of x_i



Figure 4: Spectral Search in MassBank

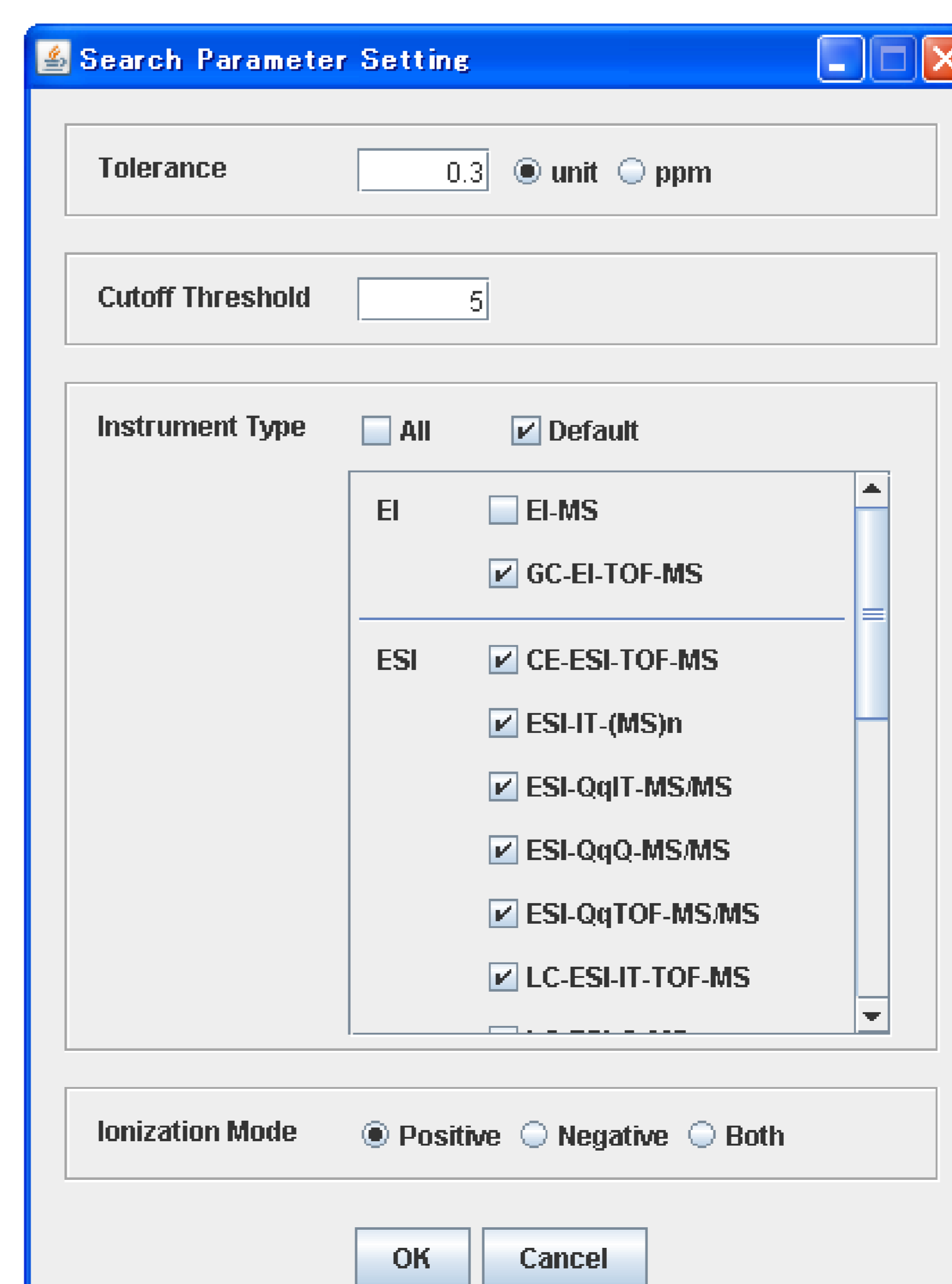


Figure 5: Default Parameters

Figure 4 shows the spectral search in MassBank. Search results are restricted by the number of matched peaks to a query. If the number of peaks in a query is less than 3, then all 3 peaks must be matched. Otherwise, at least 3 peaks must be matched.

Figure 5 shows the default parameters of the spectral search. The default value of the tolerance for matching peaks is 0.3 unit of m/z . Peaks whose relative intensity are less than or equal to the cutoff threshold (default value is 5) are ignored. A user is able to narrow the target spectra by selecting preferable instrument type and ionization mode.