Isotope Cluster-Based Compound Matching in Gas Chromatography/Mass Spectrometry for Non-Targeted Metabolomics

ARTICLE *in* ANALYTICAL CHEMISTRY · MARCH 2013

Impact Factor: 5.64 · DOI: 10.1021/ac303774z · Source: PubMed

4 AUTHORS, INCLUDING:

André Wegner
University of Luxembourg
11 PUBLICATIONS **89** CITATIONS
SEE PROFILE

Sean Sapcariu
University of Luxembourg
7 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Karsten Hiller
University of Luxembourg
45 PUBLICATIONS **1,821** CITATIONS
SEE PROFILE

analytical chemistry

Article

pubs.acs.org/ac

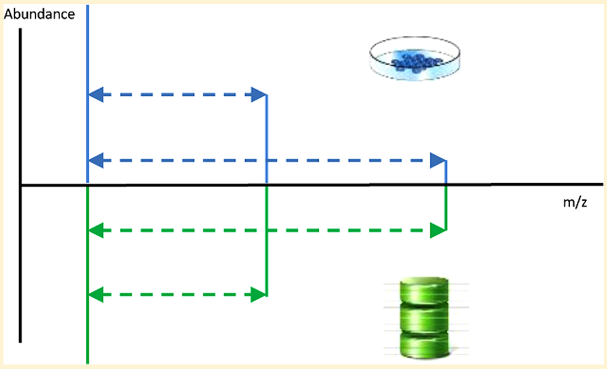
Isotope Cluster-Based Compound Matching in Gas Chromatography/Mass Spectrometry for Non-Targeted Metabolomics

André Wegner, Sean C. Sapcariu, Daniel Weindl, and Karsten Hiller*

Luxembourg Centre for Systems Biomedicine, 7, avenue des Hauts-Fourneaux, L-4362 Esch-Belval, Luxembourg

S Supporting Information

ABSTRACT: Gas chromatography coupled to mass spectrometry (GC/MS) has emerged as a powerful tool in metabolomics studies. A major bottleneck in current data analysis of GC/MS-based metabolomics studies is compound matching and identification, as current methods generate high rates of false positive and false -negative identifications. This is especially true for data sets containing a high amount of noise. In this work, a novel spectral similarity measure based on the specific fragmentation patterns of electron impact mass spectra is proposed. An important aspect of these algorithmic methods is the handling of noisy data. The performance of the proposed method compared to the dot product, the current gold standard, was evaluated on a complex biological data set. The analysis results showed significant improvements of the proposed method in compound matching and chromatogram alignment compared to the dot product.



During the last years, gas chromatography coupled to mass spectrometry (GC/MS) has proven to be a valuable tool for the analysis of sets of small molecules or metabolites. GC/MS has a large range of useful applications, including non-targeted metabolomics,[1] biomarker determination,[2] and metabolic flux analysis.[3] Electron impact ionization (EI) is the most commonly used technique to ionize samples, as this method creates reproducible fragmentation patterns. A typical comparative metabolomics analysis consists of three steps: First, compounds (chromatographic peaks) are detected in every measured chromatogram. Second, detected compounds are matched across all chromatograms and quantitative values are calculated. Third, matched quantitative values are statistically analyzed (e.g., principal component analysis, self-organizing maps, etc.). While in the beginning metabolomics studies mainly focused on the quantification of a targeted set of previously known metabolites, recent studies have tried to quantify all detectable metabolite peaks within a chromatogram. However, this non-targeted approach generates a bottleneck already at the first step of analysis. Metabolites of low concentration will be hard to distinguish from "noise peaks" (usually small peaks near the GC baseline). The more sensitive the peak detection step, the more erroneously detected compounds are present in the data set, making it more difficult to match "real" chromatographic peaks across different samples. On the other hand, "real" chromatographic peaks might be overlooked when less sensitive settings are applied. In addition, acquired mass spectra can be contaminated by extraneous mass spectral peaks, which can, for example, arise from co-eluting compounds or column bleeding.

Usually, a measured mass spectrum is assigned to a reference compound by selecting the most similar entry from a reference library using a spectrum matching algorithm. On the other hand, in the case of a non-targeted approach, component spectra present in many chromatograms need to be aligned. Sources of noise, as described above, can negatively influence correct compound matching, leading to higher rates of false positive and false negative compound identification and alignment. There are several spectrum-matching-based identification algorithms available,[4−6] and of these, dot product has proven to be the best performing in terms of accuracy.[7] While the dot product is accurate for clean spectra, we have discovered that it generates many false positives for spectra derived from complex biological samples, where many metabolites of low concentration are present.

Therefore, the objective of this work was to develop a novel spectrum similarity measure that allows performing a highly sensitive peak detection step in a non-targeted metabolomics study without compromising the specificity of the compound matching. We propose a novel method of spectrum matching that places a higher emphasis on the specific fragmentation pattern resulting from electron impact ionization. We tested our algorithm called ion cluster-based matching (ICBM) by comparing it to the dot product on complex metabolomic data sets generated in our lab. These comparisons were evaluated using receiver operating characteristic (ROC) analysis.[8]

Received: December 26, 2012
Accepted: March 20, 2013
Published: March 20, 2013

THEORETICAL BACKGROUND

Definition of Terms and Key Aspects of the Method. In the case of GC/MS, a detected compound is characterized by its mass spectrum and its retention time. A compound's mass spectrum S is defined as a set of pairs of masses and intensities

$$S = \{(m_1,i_1) = p_{1,} (m_2,i_2) = p_{2,..}, (m_n, i_n) = p_n\}$$
$$m_i < m_{i+1} i \in N \tag{1}$$

These pairs are also called peaks. The summed intensity I_S of a spectrum is calculated by summing up all peak intensities

$$I_s = \sum_{j=0}^{n} i_J \tag{2}$$

Within a spectrum, the term ion cluster refers to a unique group of multiple peaks adjacent to one another, which represent ions of the same elemental but different isotopic compositions.[9] Ion clusters with the same molecular weight but different chemical formulas will differ in the composition of their isotopic peaks. Therefore, a compound's mass spectral fingerprint is defined through the combination of fragmentation and isotopic peak composition. Mathematically, an ion cluster is then defined as a subset of S

$$f = \{p_k, ..., p_l\},\ 0 < k < l \le n, f \subset S \tag{3}$$

where p_k denotes the first peak and p_l the last peak of the ion cluster. A spectrum S can have multiple ion clusters, which are all disjointed subsets of S. As a reference point within an ion cluster, we use the peak with the highest intensity (I_M). In most cases, the highest peak within an ion cluster originates from a straight combination of the lightest isotopes of all elements, also called monoisotopic peak. In a given isotope cluster, all peaks are denoted in the relationship of masses relative to the mass of the monoisotopic peak. For example, the peak with one mass unit above I_M is denoted as I_{M+1}, and the peak with one mass unit below I_M as I_{M-1}. The term isotope cluster normalization refers to the ratio r_i of each isotope cluster's peak intensity in relation to the intensity of its monoisotopic peak

$$r_i = \frac{I_{M+i}}{I_M},\ k \le i \le l \tag{4}$$

The summed intensity of an isotope cluster is calculated by summing up all its intensities

$$I_f = \sum_{j=k}^{l} I_{M+j} \tag{5}$$

Throughout this paper, we will use the subscripts "mes" and "lib" to differentiate between the measured and the library spectrum.

Because the retention time is dependent on the used instrument, the GC-capillary, or the applied temperature program, etc. we will use the Kovats[10] retention index (RI) for all retention time-based similarity measures. Assuming that the determined retention indices for a certain compound are distributed in a Gaussian manner across different chromatograms, a Gaussian function is used for the RI based similarity index calculation

$$\text{Score}_{\text{RI}} = e^{-(ri_{\text{lib}} - ri_{\text{mes}})^2/2x^2} \tag{6}$$

where ri_{mes} and ri_{lib} are the retention indices of the measured and the library compound, and x is the maximum tolerated retention index difference. This calculation results in a score ranging from 0 (no similarity) to 1 (identical retention indices).

Dot Product. According to Stein and Scott,[7] each mass intensity pair p_i of both the measured and reference spectrum is usually weighted according to the following rule

$$p^w = [\text{mass}(m/z)]^a \times [\text{intensity}]^b \tag{7}$$

where a and b are the weighting factors that represent the contribution of the m/z value and the peak intensity, respectively. Stein and Scott reported that the optimal values are $b = 0.6$ for intensity scaling and $a = 3$ for mass weighting. This way, the relative influence of minor intensities at higher masses is increased. However, several different values have been proposed to be the optimal weighting factors over the last years. Recently, Kim et al. showed that weighting factors should be chosen individually based on the used reference library.[11] We denote a set of weighting factors as $w = (a,b)$. The dot product of the library ($S^w{}_{\text{lib}}$) and measured ($S^w{}_{\text{mes}}$) spectrum is then calculated as follows

$$\text{Score}_{\text{SpecDot}} = \frac{S^w_{\text{lib}} \times S^w_{\text{mes}}}{||S^w_{\text{lib}}||\ ||S^w_{\text{mes}}||} \tag{8}$$

As a result, the spectrum similarity score is located within the interval zero (no identity) to one (identical spectra).

Isotope Cluster-Based Matching. The isotope cluster based matching algorithm consists of four main steps (Figure 1).

1. All mass spectral isotope clusters are determined for the measured and the reference spectrum respectively.

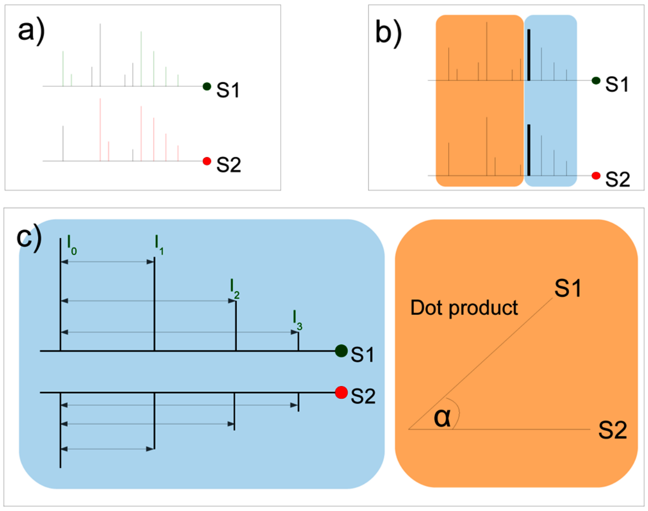


Figure 1. Overview of the ICBM algorithm. (A) Detected isotope clusters of the measured spectrum are colored in green and of the library in red. (B) Detected isotope clusters are aligned based on the masses of their monoisotopic peaks. Two isotope clusters are considered a match if the mass of their monoisotopic peak is identical (shown in bold). After, the alignment the peaks of both spectra are divided in two separate subsets. Peaks of identical isotope clusters are shown in blue, and peaks of non-matching isotope clusters or not grouped within a fragment in orange. (C) Two similarity scores based on the fragment alignment are calculated and combined.

2. Isotope clusters are aligned based on the molecular mass of their monoisotopic peaks.
3. Two similarity scores are calculated. For isotope clusters with matching monoisotopic peaks, a score based on the isotope cluster's peak ratios is calculated. For non-matching isotope clusters and non-grouped peaks, a score based on the dot product is calculated.
4. The two similarity scores are combined to a final score, where 1 represents a perfect match and 0 a mismatch between the query and the reference spectrum.

Isotope Cluster Determination. The algorithm for the isotope cluster detection iterates through S once. All consecutive peaks with a mass difference of one unit and decreasing intensities are grouped together into separate isotope clusters

$$m_j - m_{j+1} = -1 \wedge I_{M_j} > \mathrm{I}_{M_{j+1}} \tag{9}$$

In case of overlapping isotope clusters or an isotope cluster containing elements with a high abundance of natural stable isotopes such as chlorine or bromine, the algorithm splits them into two separate isotope clusters.

Isotope Cluster Alignment. Isotope clusters of the measured spectrum are aligned to the isotope cluster of the library spectrum based on the mass of their monoisotopic peaks (Figure 1b). Two isotope clusters are considered a match if the masses of their monoisotopic peaks are identical. In case of a peak at mass m_i present in one isotope cluster but not in its counterpart, a peak of mass m_i and intensity 0 is added to the corresponding isotope cluster. This way, aligned isotope clusters always have the same number of peaks. The measured and the library spectra can then be divided into two subsets of peaks. One set contains all peaks from the matching isotope clusters (F), and the other contains the remaining peaks of the spectrum (R).

$$F \cup R = S \tag{10}$$

In the illustrated example, peaks of the set F are highlighted in blue and peaks of the set R in orange (Figure 1b).

Similarity Score Calculation. On the basis of the alignment, one score is calculated for set F (matched isotope clusters) and one score for R (non-matching isotope clusters and peaks not grouped within an isotope cluster) (Figure 1c). First, for each matched isotope cluster in F, all peak intensities are normalized by the respective monoisotopic peak. Second, the distance d between two isotope clusters is calculated by summing the absolute values between the differences of corresponding normalized peak intensities.

$$d = \sum_{i=0}^{n} |r_{\mathrm{mes}_i} - r_{\mathrm{lib}_i}| \tag{11}$$

To keep the score within the interval [0,1], the contribution of one isotope cluster pair to the total similarity score is weighted by the number of isotope cluster peaks and an intensity scale. Therefore, the isotope cluster's summed intensity is divided by the total intensity of the mass spectrum to obtain the intensity fraction of the isotope cluster in S

$$\begin{aligned} x &= \text{number of peaks within } f \\ \text{intensity scale} &= \frac{I_{f_{\mathrm{mes}}}}{I_{S_{\mathrm{mes}}}} + \frac{I_{f_{\mathrm{lib}}}}{I_{S_{\mathrm{lib}}}} \\ \text{scale} &= x \times \text{intensity scale} \end{aligned} \tag{12}$$

The total isotope cluster-based distance of the two mass spectra is then calculated as follows

$$\mathrm{Score}_F = \frac{1}{n} \times \sum_{i=0}^{n} d_i \times \mathrm{scale}_i \tag{13}$$

where n is the number of matched isotope cluster. This calculation transforms the distance in the interval between zero (identical spectra) to one (no identity). To make this score comparable to the score of the dot product, Score_F is inverted within the interval [0,1]

$$\mathrm{Score}_F = 1 - \mathrm{Score}_F \tag{14}$$

For the remaining peak set R, a similarity score based on the dot product as described above is calculated. These two similarity measures are then combined to form a composite spectrum similarity score. To reduce the bias to favor one of the two scores, a weighting factor w_F based on the summed intensities of all matched isotope clusters is calculated.

$$w_F = \left(\frac{I_{F_{\mathrm{lib}}}}{I_{S_{\mathrm{lib}}}} + \frac{I_{F_{\mathrm{mes}}}}{I_{S_{\mathrm{mes}}}} \right) \times \frac{1}{2} \tag{15}$$

When for example the summed intensity of all matched isotope clusters encompasses 70% of the total intensity of the measured and the library spectrum, the isotope cluster-based score is weighted with 0.7 and the dot product score with 0.3.

$$\mathrm{Score}_{\mathrm{SpecIC}} = (1 - w_F) \times \mathrm{ScoreDot} + w_F \times \mathrm{Score}_F \tag{16}$$

Compound Matching. We will use the term compound matching to refer to the following process: the spectrum of a measured compound is compared to spectra within a reference library. On the basis of a spectrum similarity score calculated with the dot product ($\mathrm{Score}_{\mathrm{SpecDot}}$ in eq 8) or the ICBM algorithm ($\mathrm{Score}_{\mathrm{SpecIC}}$ in eq 16), the best-matched entry from the reference library is assigned to the measured compound. Figure S-1 of the Supporting Information illustrates the process of compound matching for the dot product and the ICBM algorithm. In the case of the dot product, the measured spectrum is compared to all spectra within the reference library, whereas in the case of the ICBM algorithm, the measured spectrum is only compared to a subset of the reference library.

This subset is generated by selecting only those spectra from the library that have at least one matching isotope cluster. For very short spectra (less than 20 peaks), the spectrum similarity score is always calculated with the dot product.

Non-Targeted Chromatogram Alignment. If a number of GC/MS chromatograms are to be analyzed comparatively, it is necessary to align similar compounds among the different chromatograms. To additionally account for the retention time of a compound, a combined similarity score $\mathrm{Score}_{\mathrm{total}}$ based on the spectrum and retention index similarity is calculated. Because the spectral profile of a compound contains more information than the retention index, it is weighted stronger

$$\mathrm{Score}_{\mathrm{total}} = \sqrt[3]{\mathrm{Score}_{\mathrm{Spec}}^2 \times \mathrm{Score}_{\mathrm{RI}}} \tag{17}$$

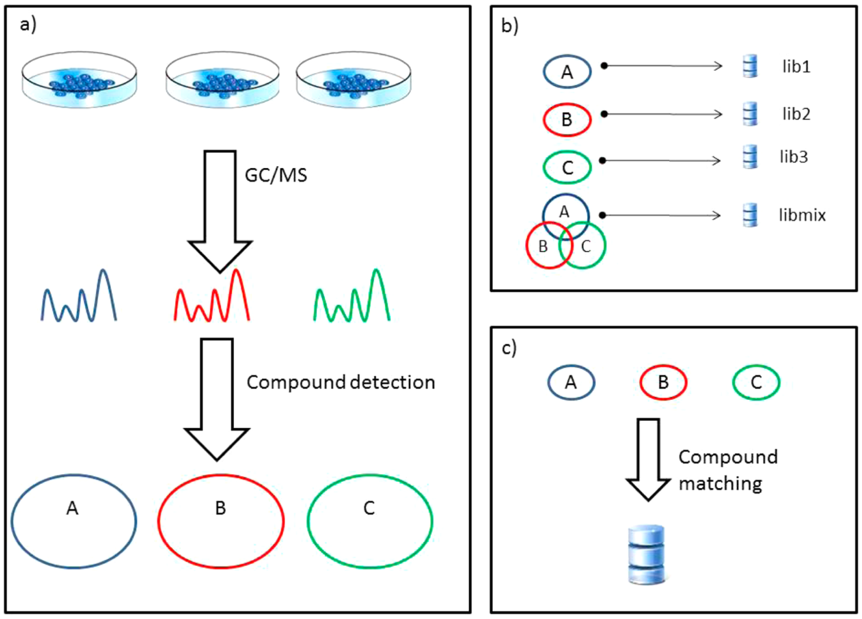


Figure 2. Overview of experiment 1a. (a) Samples were measured with GC/MS, and all compounds detected in each chromatogram were stored in the sets A, B, and C. (b) Reference sets were generated based on the sets A, B, and C. First, the mass profiles and retention indices of all compounds in A, B, and C were stored in lib1, lib2, and lib3, respectively. Second, the intersection of A, B, and C was stored in libmix. (c) All compounds in A, B, and C were matched against the reference sets lib1 to lib3 and libmix.

All detected compounds of the given chromatograms are then matched according to the following strategy: Starting with an empty reference set, the mass profiles and RIs of all detected compounds of the first chromatogram are added. Afterward, all remaining chromatograms are consecutively matched against this set. If a compound has been found in this set, the reference compound is assigned to this compound. However, if a compound is not found, this compound is added as a new reference compound to the reference set.

■ MATERIAL AND METHODS

Overview of Experiments. To validate the performance of our algorithm, we performed three different experiments. The first two experiments evaluated our method's usability in compound matching using the NIST library and a complex biological data set experimentally obtained in our lab. The third experiment shows the practical application of compound matching in a non-targeted chromatogram alignment.

GC/MS measurements for all samples were performed using an Agilent 6890 GC equipped with a 30 m DB-35MS capillary column. The GC was connected to an Agilent 5975C MS operating under electron impact (EI) ionization at 70 eV. The MS source was held at 230 °C and the quadrupole at 150 °C. The detector was operated in scan mode, and 1 μL of derivatized sample was injected in splitless mode.

Helium was used as carrier gas at a flow rate of 1 mL/min. The GC oven temperature was held on 80 °C for 6 min and increased to 300 °C at 6 °C/min. After 10 min, the temperature was increased to 325 °C at 10 °C/min for 4 min. The run time of one sample was 59 min.

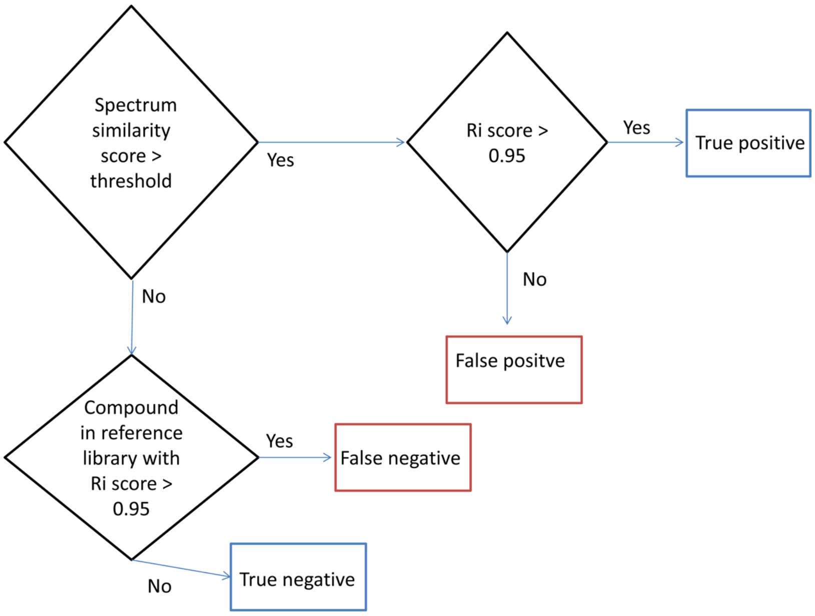
Metabolite derivatization was performed using an Agilent autosampler. Dried polar metabolites were dissolved in 15 μL of 2% methoxyamine hydrochloride in pyridine at 40 °C. After 30 min, an equal volume of MSTFA (2,2,2-trifluoro- N-methyl-N-trimethylsilyl-acetamide) + 1% TMCS (chloro-trimethyl-silane) were added and held for 30 min at 40 °C.

Mammalian cells for both experiments were grown on 6-well plates. Metabolite extraction was performed after washing each well with 1 mL 0.7% saline solution and quenching with 0.4 mL −20 °C methanol. After adding an equal volume of 4 °C cold water, cells were collected with a cell scraper and transferred to tubes containing 0.4 mL −20 °C chloroform. The extracts were vortexed at 1400 rpm for 20 min at 4 °C and centrifuged at 16000 *g* for 5 min at 4 °C. 0.3 mL of the upper aqueous phase was collected in specific GC glass vials and evaporated under vacuum to dryness at −4 °C using a refrigerated CentriVap Concentrator (Labconco).

All data were processed using the MetaboliteDetector software package.[12] The compound detection was performed with the following deconvolution settings: peak thresholds 5, minimal peak height 5, bins 10, deconvolution width 5.

Experiment 1a: Compound Matching Using a Complex Biological Data Set. We measured metabolites in six replicates of the mouse macrophage cell line RAW264.7 by GC/MS. Figure 2 illustrates the experimental setup. On average each replicate contained 1042 detected compounds (1061, 1050, 1024, 1047, 1035, and 1035). To demonstrate our method's ability in compound matching, we compared each sample's compound mixture to a set of reference sets. These reference sets were newly generated based on the six measured replicates (Figure 2b). For this purpose, the mass profiles and retention indices of all extracted compounds of all six replicates were collected in separate reference sets (lib1−lib6). In addition, a reference set was generated that only included compounds found in all six replicates (libmix). Mathematically, libmix is the intersection of the six reference sets generated previously. This process required an algorithm for compound matching. In order to eliminate bias, both the dot product and ICBM algorithm were used to match compounds present in all sample chromatograms, and only compounds found using both algorithms were considered for analysis. Finally, libmix contained 201 compounds. Compounds were matched based

Scheme 1. Process of Evaluating Correct Matches


Spectrum similarity score > threshold
Yes
Ri score > 0.95
Yes
True positive
No
No
False positve
Compound in reference library with Ri score > 0.95
Yes
False negative
No
True negative

on their mass spectral identity as described in the Theoretical Background section using the dot product (eq 8) and the ICBM algorithm (eq 16).

Evaluation of Correct Matches. The evaluation of a correct match is challenging because there is no additional information available in an experimental and non-targeted metabolite set besides the mass spectrum and the retention index. Additionally, it is not guaranteed that the queried spectrum is present in the reference set at all. Nevertheless, it can be expected that true matching compounds elute at similar retention times. In contrast, if two algorithmically matched compounds elute at distinct retention times, the match can be considered as incorrect. For that reason, we exploit the retention time to evaluate correct matches by calculating a retention index score (eq 6). We set the maximum retention index difference to 10. That guarantees that only compounds eluting at almost identical retention times obtain a score close to 1. Scheme 1 depicts the process of evaluating correct matches for this experiment. To illustrate the results of this process, we used receiving operator characteristic (ROC) curves, which are frequently used to evaluate the performance of binary classification and prediction models. By plotting the true positive rate (TPR) against the false positive rate (FPR), a model's ability to separate positive from negative cases is illustrated.

Compounds with a spectrum similarity score higher than the threshold are classified as positive matches and compounds with a lower score as negative matches. To further discriminate false positives (FP) and false negatives (FN), we calculated a retention index score. Positively identified compounds with a retention index score higher than 0.95 are categorized as true positives (TP), otherwise as false positive (FP). To check whether negatively identified compounds are truly absent from the library (TN) or were not detected by the algorithm (FN), retention index scores for all library compounds were calculated. A compound is counted as TN if there is no compound in the library with a retention index score higher than 0.95, otherwise the compound is counted as FN.

TPR and FPR are then calculated as follows

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

$$\text{specificity (TRN)} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{18}$$

Because the reference sets were newly generated, there were no optimal weighting factors available. For that reason, we tested three different sets of weighting factors for this experiment: $w(3,0.5)$, $w(2,0.6)$, and $w(1.3,0.53)$.

Experiment 2: Non-Targeted Chromatogram Alignment. In a second experiment, we analyzed metabolites of the human neuronal cell line LUHMES cultured under four different conditions. LUHMES cells were cultivated and differentiated to post-mitotic neurons according to ref 10. The first two conditions were undifferentiated LUHMES cells cultivated at 5% and 20% oxygen. The last two conditions were LUHMES cells differentiated to post-mitotic neurons at 5% and 20% oxygen. For each condition, we measured three replicates by GC/MS. After compound detection a non-targeted chromatogram alignment of all 12 chromatograms was performed (using $\text{Score}_{\text{Total}}$ in eq 17). To demonstrate the performance of the ICBM algorithm compared to the dot product, compounds found in all 12 samples were quantified. The quality of the chromatogram alignment of the respective algorithm was measured by calculating the relative standard error of the mean for every detected metabolite within a replicate group. It is obvious that wrongly matched compounds result in higher standard errors of the mean compared to correctly matched metabolites.

■ RESULTS AND DISCUSSION

Experiment 1a: Compound Matching Using a Complex Biological Data Set. The performance of the ICBM algorithm was compared to that of the dot product by calculating the FPR and TPR as discussed in the Material and Methods section. For this purpose, we performed two sets of validation runs. In the first run, all compounds within the six replicates were matched against the reference set libmix. This experiment intends to simulate a targeted approach of analysis where a reference library with high quality spectra is given. Because libmix was restricted exclusively to compounds that were previously identified by both algorithms in all six replicates, both algorithms are expected to identify these compounds with a high specificity.

Figure 3 depicts the ROC curve of this run. As expected, both algorithms performed equally well as the ROC curve shows no significant difference between the two algorithms. The area under the curve only differs 0.4% for the best performances of both algorithms. The threshold to differentiate between a positive and negative hit was varied from 0.45 to 0.95. On average, each replicate contained 1042 compounds, but only 201 compounds were found across all six replicates. That means there is a high amount of analytical noise present in each measurement.

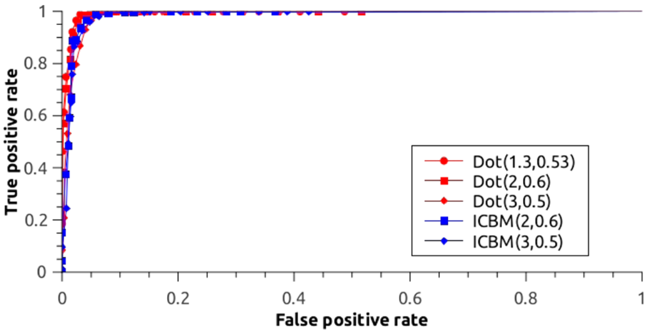


Figure 3. Simulation of a targeted analysis in experiment 1a. Receiver operating characteristic curves for the dot product are in red and ICBM algorithm in blue. Applied mass weighting factor a and intensity weighting factor b are denoted in parentheses. The area under the curve for the different runs is: Dot(1.3,0.53) = 0.9948, Dot(2,0.6) = 0.9934, Dot(3,0.5) = 0.9862, ICBM(1.3,0.53) = 0.9911, ICMB(2,0.6) = 0.9873, and ICMB(3,0.5) = 0.985.

To show the capability of our algorithm to handle noisy data, we did a second validation run. In contrast to the first run, we now matched each replicate's total compound mixture against the reference sets created from all of the other replicates (Figure 2). For example, all compounds detected in replicate 1 were compared subsequently against set2 through set6. All these sets include in addition the low quality spectra thus reflecting a non-targeted approach. Compound matching of this data set by our algorithm resulted in a significantly lower ratio of FPR to TPR, which is represented by a higher area under the curve (AUC) score (Figure 4). The ICMB algorithm outperforms the dot product for all sets of weighting factors. Both algorithms performed best for the weighting factor set w = (1.3,0.53). This run clearly demonstrates the power of the ICMB algorithm. It overcomes the bottleneck created through a sensitive peak detection step by keeping a high specificity of the compound matching.

The dot product AUC dropped from 0.994 to 0.883, while the ICBM algorithm only reduced from 0.9901 to 0.982. With a 10 percent difference between the AUC's of these two methods, this result clearly points to the advantage of our algorithm compared to the dot product. This result can be explained by looking at the input for both algorithms. The inputs for the dot product are the weighted intensities and m/z values; these values increase exponentially as m/z values increase. This allows characteristic peaks (at higher masses) to have a higher impact on the result; however, this sometimes leads to high scores for nonmatching spectra, where only a few peaks at high masses are similar. In contrast to the dot product, ICBM calculates scores in a more context-based approach. This has two effects: it reduces the effect of peaks with high m/z values masking the influence of other relevant peaks in the spectrum and increases the specificity of spectrum matching by removing low quality spectra from analysis. The specificity of ICBM can be controlled by setting a threshold for the minimum number of aligned fragments. In our analysis, an optimal threshold turned out to be one similar fragment; reference spectra with zero similar fragments to the measured spectrum were removed from the result set.

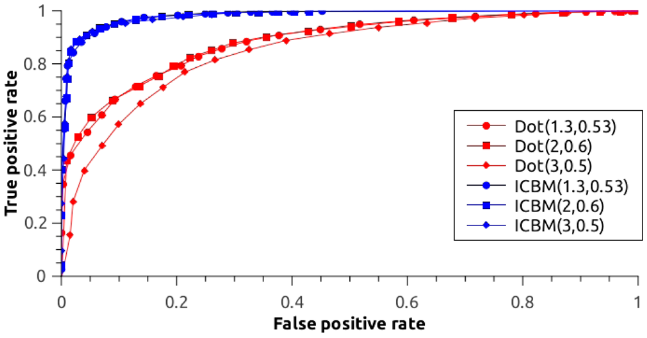


Figure 4. Simulation of a non-targeted analysis in experiment 1a. Receiver operating characteristic curves for the dot product are in red and ICBM algorithm in blue. Applied mass weighting factor a and intensity weighting factor b are denoted in parentheses. The ICBM algorithm outperforms the dot product in all cases. The area under the curve for the different runs is: Dot(1.3,0.53) = 0.883, Dot(2,0.6) = 0.885, Dot(3,0.5) = 0.848, ICBM(1.3,0.53) = 0.983, ICMB(2,0.6) = 0.9825, and ICMB(3,0.5) = 0.9815.

Two examples of spectrum comparison pairs are shown in Figures S3 and S4 of the Supporting Information. It is apparent by looking at each measured spectra that their counterparts are matched poorly. However, the dot product generates consistently much higher scores than the ICBM algorithm. In the case of Figure S3 of the Supporting Information, when the same weighting factors are applied, dot product produces scores of 0.970515, 0.915002, and 0.722992, while ICBM produces scores of 0.347901, 0.394415, and 0.434143, with the weighting factors (3,0.5), (2.0,6), and (1.3,0.6), respectively. These differences in matching scores can be a major source of mismatched spectra, causing the divergence in the performance of the two algorithms for the analysis of noisy data sets. It should be noted that the dot product is much more dependent on weighting factor values, as evidenced by the larger variation of matching scores in these examples, as well as the ROC curve of the non-targeted analysis of experiment 1a.

Additionally, we tested our algorithm on spectra obtainend from the NIST library (experiment 1b, Supporting Information) to have an example of the performance of our algorithm when used with a clean (manually curated) data set, although our algorithm was designed for noisy data sets. The accuracy of compound detection for the dot product was 82.2% compared to 83.3% for the ICBM algorithm. This result shows that the ICBM algorithm performs at least as well as the dot product for clean curated data sets. However, the number of false positive identifications is much higher for the dot product (Table 1, Supporting Information) when target compounds were removed from the library.

Something important to consider is the large number of artificial compounds arising from a highly sensitive peak detection step, and this problem may be is marginalized by the increased detection of genuine metabolites. For this reason, we tested whether the sensitivity of the compound detection has a significant effect on the number of detected compounds that can be matched across all chromatograms. We analyzed four in-house data sets including the two already used in experiment one and two, applying two different levels of sensitivity for the compound detection and counted the number of compounds that could be matched across all chromatograms. We controlled the sensitivity by increasing the peak threshold and minimal peak height from 5 for the high sensitivity setting to 15 for the low sensitivity settings. Because we already showed that the ICMB algorithm is superior compared to the dot product for the high sensitivity settings, we compared the results for the high sensitivity settings for the ICMB to the results of the lower sensitivity settings for the dot product (Figure S2, Supporting Information). As sensitivity was reduced, a significant decrease in detectable compounds was found. These results show that it is indeed beneficial to apply highly sensitive compound detection settings.

Experiment 2: Chromatogram Alignment. To show the impact of our algorithm in a more practical example, we did a non-targeted chromatogram alignment of a data set containing four different replicate groups. This analysis intends to demonstrate misalignments and, therefore, false positive identifications by the spectrum matching algorithm. As misaligned compounds will highly differ in intensity, the relative standard error for these compounds will be high within a replicate group. The experimental setup of this experiment is typical for a non-targeted metabolomics study such as biomarker determination, where metabolite levels between several different conditions are comparatively analyzed. On average, the four replicate groups contained 926, 928, 879, and 869 detected compounds, respectively. The dot product found 263, 280, 246, and 240 compounds that are present in all replicates within the four replicate groups, whereas the ICBM algorithm only found 242, 248, 230, and 227 compounds, respectively.

However, the averages of relative standard errors of each replicate group (Table 1) were higher in the dot product based analysis compared to the ICBM analysis. Because high standard errors indicate possible mismatches within the chromatogram alignment, this result further underlines the fact that the dot product leads to a higher false positive rate than the ICBM algorithm. Overall, there is a shift to lower standard errors when using the ICBM algorithm as illustrated in Figure 5.

Table 1. Mean of Standard Errors for Biological Replicates

replicate group	dot product	ICBM
1	0.182	0.138
2	0.205	0.176
3	0.208	0.187
4	0.142	0.116

ICBM results in higher frequency of smaller standard error values and lower frequency of higher standard error values compared to the dot product. The higher number of identified compounds by the dot product correlates roughly with the number of compounds having unusually high standard errors over 30%. In principal, a similar effect can be achieved by changing the sensitivity of the compound detection step (e.g., deconvolution settings or minimum number of peaks). This way, low quality spectra are removed before analysis. For a targeted approach, this might be suitable because compounds of interest are known in advance, and optimal compound-specific settings with the best trade-off between sensitivity and specificity can be determined by evaluating the results for these compounds. However, in a non-targeted methodology, such evaluation is not possible, and compounds of interest might be inadvertently removed. Therefore, settings with a high sensitivity should be applied in these cases.

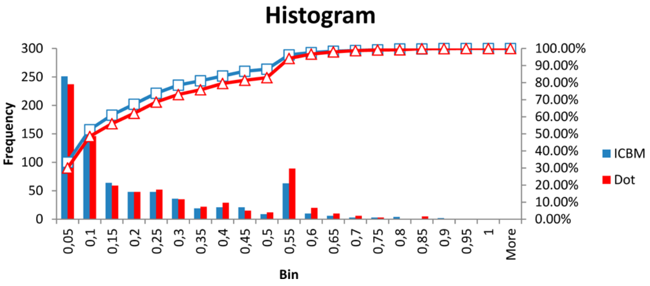


Figure 5. Histogram of relative standard errors of experiment 2. Distribution of the relative standard errors for the dot product is shown in red and for ICBM algorithm in blue. Absolute numbers are shown as a bar plot, and the cumulative percentage distribution as a line plot.

CONCLUSION

Herein, we describe a novel spectrum matching algorithm that can be used for chromatogram alignment and compound matching, techniques that are routinely used in metabolomics data analysis. We showed that the ICBM algorithm outperforms the dot product not only on noisy data sets but also on the well-curated NIST library. The ability of the ICBM algorithm to cope with noisy data makes it especially suited for the analysis in the context of non-targeted metabolomics studies. For example, studies involving biomarker detections are heavily dependent on the accuracy of the chromatogram alignment; investigators of these studies could benefit greatly by using the ICBM algorithm. In addition to compound identification and chromatogram alignment, the ICBM algorithm can also be applied to speed up metabolite identification when using large reference libraries. As described earlier, a measured compound is identified by comparing its spectrum to all spectra within a reference library. At best, a huge reference library, such as the NIST, is used to identify the maximum number of metabolites. However, the computation time for one identification increases linearly with the number of spectra. A strategy to reduce the number of comparisons is to compare the measured spectrum only to a subset of the reference library. This subset can be generated by precalculating isotope clusters of all library spectra using the ICBM algorithm. When stored in an appropriate way, for example, in a SQL database, a list of possible hits can be calculated based on the mass of the monoisotopic peaks of the most abundant fragments of the measured spectrum.

ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional information as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tel.:+352 46 66 44 6136. Fax: +352 46 66 44 6949. E-mail: karsten.hiller@uni.lu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge financial support from the Fonds National de la Recherche (FNR). Specifically, K.H., D.W., and S.C.S. are funded by the ATTRACT program "Metabolomics Junior Group", and A.W. is supported by the AFR Grant 1328318. The authors gratefully thank Prof. Rudi Balling for his support.

REFERENCES

(1) Huang, M.; Joseph, J. W. *Islets* **2012**, *4*, 210−222.

(2) Abbiss, H.; Maker, G. L.; Gummer, J.; Sharman, M. J.; Phillips, J. K.; Boyce, M.; Trengove, R. D. *Nephrology (Carlton)* **2012**, *17*, 104−110.

(3) Metallo, C. M.; Gameiro, P. A.; Bell, E. L.; Mattaini, K. R.; Yang, J.; Hiller, K.; Jewell, C. M.; Johnson, Z. R.; Irvine, D. J.; Guarente, L.; Kelleher, J. K.; Vander Heiden, M. G.; Iliopoulos, O.; Stephanopoulos, G. *Nature* **2012**, *481*, 380−384.

(4) Skolow, S.; Karnofsky, J.; Gustafson, P. *Finnigan Application Report 2*; Finnigan Corp.: San Jose, CA, 1978.

(5) Hertz, H. S.; Hites, R. A.; Biemann, K. *Anal. Chem.* **1971**, *43*, 681−691.

(6) Stauffer, D. B.; McLafferty, F. W.; Ellis, R. D.; Peterson, D. W. *Anal. Chem.* **1985**, *57*, 1056−1060.

(7) Stein, S.; Scott, D J *Am. Soc. Mass Spectrom.* **1994**, 859−866.

(8) Lasko, T. A.; Bhagwat, J. G.; Zou, K. H.; Ohno-Machado, L. *J. Biomed. Inform.* **2005**, *38*, 404−415.

(9) Vetter, W. *Biol. Mass Spectrom.* **1994**, *23*, 379−379.

(10) Kovats, E. *Helv. Chim. Acta* **1958**, *41*, 1915−1932.

(11) Kim, S.; Koo, I.; Wei, X.; Zhang, X. *Bioinformatics* **2012**, *28*, 1158−1163.

(12) Hiller, K.; Hangebrauk, J.; Jäger, C.; Spura, J.; Schreiber, K.; Schomburg, D. *Anal. Chem.* **2009**, *81*, 3429−3439.