BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 7 2014, pages 988–995
doi:10.1093/bioinformatics/btt738

Systems biology

Advance Access publication December 20, 2013

BiPACE 2D—graph-based multiple alignment for comprehensive 2D gas chromatography-mass spectrometry

Nils Hoffmann[1,*], Mathias Wilhelm[1,†], Anja Doebbe[2,3], Karsten Niehaus[3,4] and Jens Stoye[1,*]

[1]Genome Informatics, Faculty of Technology and CeBiTec, [2]Algae Biotechnology & Bioenergy, [3]Faculty of Biology and CeBiTec, [4]Proteomics and Metabolomics Research, Bielefeld University, 33501 Bielefeld, Germany

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Comprehensive 2D gas chromatography-mass spectrometry is an established method for the analysis of complex mixtures in analytical chemistry and metabolomics. It produces large amounts of data that require semiautomatic, but preferably automatic handling. This involves the location of significant signals (peaks) and their matching and alignment across different measurements. To date, there exist only a few openly available algorithms for the retention time alignment of peaks originating from such experiments that scale well with increasing sample and peak numbers, while providing reliable alignment results.

Results: We describe BiPACE 2D, an automated algorithm for retention time alignment of peaks from 2D gas chromatography-mass spectrometry experiments and evaluate it on three previously published datasets against the mSPA, SWPA and Guineu algorithms. We also provide a fourth dataset from an experiment studying the H_2 production of two different strains of *Chlamydomonas reinhardtii* that is available from the MetaboLights database together with the experimental protocol, peak-detection results and manually curated multiple peak alignment for future comparability with newly developed algorithms.

Availability and implementation: BiPACE 2D is contained in the freely available Maltcms framework, version 1.3, hosted at http://maltcms.sf.net, under the terms of the L-GPL v3 or Eclipse Open Source licenses. The software used for the evaluation along with the underlying datasets is available at the same location. The *C.reinhardtii* dataset is freely available at http://www.ebi.ac.uk/metabolights/MTBLS37.

Contact: nils.hoffmann@cebitec.uni-bielefeld.de or jens.stoye@uni-bielefeld.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 8, 2013; revised on November 28, 2013; accepted on December 13, 2013

1 INTRODUCTION

Comprehensive 2D gas chromatography (GC × GC) has developed into a *de facto* standard analytical technique to resolve complex mixtures of volatile chemical compounds in various areas of research. It is often used for the separation and characterization of petrol and gasoline mixtures (Arey *et al*., 2005; Fraga *et al*., 2000; Johnson *et al*., 2004; Ventura *et al*., 2011; von Mühlen *et al*., 2006). It couples two chromatographic columns, with possibly different characteristics, to achieve an improved separation of complex mixtures over one-dimensional gas chromatography. However, identification of analytes in GC × GC with a one-dimensional detector such as a flame ionization detector is often limited to the use of retention index calculations (Arey *et al*., 2005) and can be improved by coupling a mass spectrometer to the chromatograph (GC × GC-MS). In the field of metabolomics, which studies the multitude of organic molecules produced, modified and consumed by living organisms (metabolites), both GC × GC and GC × GC-MS have been successfully used to characterize and quantify volatile metabolites in different organisms (Doebbe *et al*., 2010; Koek *et al*., 2011; Pierce *et al*., 2006; Porter *et al*., 2006; Vial *et al*., 2009).

The variability of retention times (RTs) in both GC × GC and GC × GC-MS requires sophisticated algorithms for automatic alignment of corresponding analyte signals between different samples. Additionally, the size and number of acquired sample data pose a significant challenge to automated methods and effectively prevent large-scale manual intervention by human experts (Hoffmann and Stoye, 2012). Preprocessing of GC × GC-MS data involves the filtering and noise reduction of the raw signal, the localization, deconvolution, integration and normalization of analyte signals of interest (peaks) (Amador-Muñoz and Marriott, 2008), as well as downstream matching of peaks to create a multiple alignment of related signals from different samples. Then, statistical methods like analysis of variance, principal components analysis (Ventura *et al*., 2011), partial least squares (Johnson *et al*., 2004) and variants thereof may be applied to compare samples for significant differences and similarities within and between experimental factor groups. A more complete overview of current data processing methods and programs for GC × GC and GC × GC-MS data may be found in Matos *et al*. (2012), Reichenbach *et al*. (2012) and Kallio *et al*. (2009).

In this article, a novel automated method for the multiple alignment of GC × GC-MS peaks is introduced: bidirectional best-hit peak assignment and clique extension for 2D chromatograms (BiPACE 2D), which is based on comparing peak mass spectra and RTs in two dimensions between a large number of samples.

2 METHODS

2.1 Peak alignment for GC × GC-MS

For GC × GC-MS peak alignment, the mass spectrum behind each time point on the 2D chromatographic plane can be used as an additional

*To whom correspondence should be addressed.
†Present address: Proteomics and Bioanalytics Research Group, TU Munich, 85354 Freising, Germany.

criterion for peak similarity or identity. The use of electron ionization in GC × GC-MS produces rich fragmentation spectra that are comparable with fingerprints of each analyte. However, electron ionization is known to lead to identical or similar spectra (due to detector noise) for certain classes of analytes, especially structural isomers. Thus, an additional criterion for their distinction is required, such as the RT information of the mass spectrum in the first and second dimensions of separation. A peak in GC × GC-MS may encompass many mass spectra, so that a reduction to a representative mass spectrum is advisable for improved signal-to-noise ratio and better spectral database search results (Oh *et al.*, 2008). If those results are reliable and return few false-positive identifications, one can subsequently use the assigned names to associate peaks across samples. However, the results of database searches may consistently associate a spectrum erroneously with an analyte that happens to be just above the identification threshold used by the database, whereas the true analyte is missing from the database.

Typically, peaks should be aligned between samples that were measured under identical (homogeneous) separation conditions. However, the algorithms of Jeong *et al.* (2012), Kim *et al.* (2011b) and Wang *et al.* (2010) also support alignment of peaks that were measured under different (heterogeneous) conditions (e.g. different temperature gradient) that lead to non-linear shifts especially in the first RT of the GC × GC chromatogram.

The MSort algorithm (Oh *et al.*, 2008) sorts and associates peaks based on their absolute RT difference for each separation dimension and mass spectral similarity using Pearson's correlation coefficient. It successively builds a sorted peak table created from unassigned peak tables and matches peaks from a reference table, a search table with the highest number of merged peaks, against the remaining peaks using a sorting criterion until all searchable peaks have been processed.

The algorithm DISCO (Wang *et al.*, 2010; Wei *et al.* ,2013) uses landmark peaks in each sample that are mapped to landmark peaks in a reference sample using Euclidean distance to calculate RT similarity and Pearson's correlation coefficient to determine the similarity of mass spectra. Based on the landmark peaks, the method determines a local linear interpolation that is applied to non-landmark peaks, thereby correcting for non-linear RT distortion.

Kim *et al.* (2011a,b) have introduced two different algorithms to approach the peak alignment problem in GC × GC-MS. The Smith-Waterman peak alignment (SWPA) approach uses variants of dynamic programming to find a peak matching with maximal score for pairwise alignments. Their mixture similarity peak alignment (mSPA) method includes the optimization of a likelihood function based on a parameterized mixture similarity, which involves the dot product as mass spectral similarity and RT deviation calculation with different distance metrics. Both of their approaches extend the pairwise alignments transitively to a multiple peak alignment, based on a prior chosen reference peak list.

Guineu (Castillo *et al.*, 2011) uses the Score Alignment algorithm, which is based on a combined score, using predefined windows for first and second dimension RT deviations and retention index deviation. The method scores neighboring peaks against potential target peak groups, building candidate paths of related peaks. The weighted cosine product is used to avoid alignment of mass spectra with low pairwise scores, with a user-defined minimum threshold. Path generation and evaluation is performed in parallel and followed by a subsequent post-processing phase, where peaks that were assigned to multiple groups are reassigned to the peak group with highest score until all such conflicts are resolved.

Jeong *et al.* (2012) use a statistical model to align the peaks, based on pairwise peak scores calculated from mass spectral similarity (cosine score) and RT deviation score functions. Their approach uses landmark peaks with a high posterior probability according to their model to calculate an RT correction for the remaining peaks, which fall below a specific posterior probability threshold. They additionally calculate a corrected RT for aligned peaks.

The BiPACE and BiPACE RT algorithms were introduced in Hoffmann *et al.* (2012), showing their applicability for peak alignment of one-dimensional gas chromatography-mass spectrometry data. BiPACE 2D is a novel extension of BiPACE that uses the 2D RT information in addition to mass spectral similarity to align peaks across multiple chromatograms without requiring a user-defined reference chromatogram.

2.2 BiPACE 2D pairwise peak similarity function

Given a chromatogram $C = \{p_1, p_2, \ldots, p_\ell\}$ as an ordered set of peaks, we define a 2D peak $p = (\mathbf{m}, \mathbf{i}, t_1, t_2)$ as a tuple of a mass vector $\mathbf{m}$, an intensity vector $\mathbf{i}$, both with the same dimensions, a first column RT t_1 and a second column RT t_2. For two peaks p and q, represented by their binned mass spectral intensity vectors with first column RTs $t_{1,p}$, $t_{1,q}$, second column RTs $t_{2,p}$, $t_{2,q}$ and RT tolerances of D_1 and D_2, for the first and second column, respectively, we define a similarity function following Robinson *et al.* (2007) as follows:

$$f_{2d}(p,q) := \exp\left(-\frac{(t_{1,p}-t_{1,q})^2}{2D_1^2}\right)\cdot\exp\left(-\frac{(t_{2,p}-t_{2,q})^2}{2D_2^2}\right)\cdot s(p,q), \quad (1)$$

where $s(p,q)$ is an arbitrary similarity function between the mass spectral intensity vectors, such as the cosine, the weighted cosine (Stein and Scott, 1994), the dot product, Pearson's linear correlation coefficient or Spearman's rank correlation coefficient. The $f_{2d}(\cdot,\cdot)$ can be interpreted as a likelihood function that independently scores the proximity and mass spectral similarity of its arguments. It is maximized by peaks that have low deviation in RTs and a high mass spectral score. The impact of deviations in either RT dimension can be individually adjusted via the RT tolerance parameters D_1 and D_2 of the Gaussian RT penalty terms, where a higher value allows for larger RT deviations.

To prune the search space early during the pairwise all-against-all peak similarity calculation phase of our algorithm, each RT penalty term has an additional threshold parameter (T_1 and T_2, respectively) that allows to effectively stop any further evaluation of $f_{2d}(\cdot,\cdot)$ if the value of the threshold for that term is not attained or exceeded. Thus, the mass spectral score function may not need to be evaluated at all, resulting in a large speedup at the expense of reduced sensitivity toward peaks with larger RT deviations.

2.3 Peak pair matching

Given K chromatograms $C_1, \ldots, C_K$, let $S = (V, E)$ be a complete K-partite, weighted graph with vertex set $V = P_1 \cup \ldots \cup P_K$, where partition P_i consists of all peaks from chromatogram C_i, and edge set E represents the similarity values between all peak pairs from different chromatograms. Finding all peak groups with maximal pairwise similarity in S is equivalent to enumerating all cliques of S, a problem that relates to the classic *NP*-complete problem CLIQUE (Karp, 1972) with a runtime exponential in the size of the input, in this case the number of peaks.

Then, for each pair of peaks $p \in C$ and $q \in C'$ from distinct chromatograms C and C', the peak with highest similarity to p in C', denoted q', and the peak with highest similarity to q in C, denoted p', can be assigned as bidirectional best hits (BBHs) of each other, if $p = p'$ and $q = q'$ (Overbeek *et al.*, 1999). All peak similarities of p to other peaks in C' and of q to other peaks in C are set to a minimum similarity value, effectively removing the corresponding edge in S, whereas the similarities of the BBHs p and q are retained.

2.4 Merging and multiple alignment construction

Based on the reduced set of BBHs, let V' be the set of all vertices that are part of at least one BBH and $S' = (V', E')$ the reduced K-partite BBH graph with V' as its vertex set and E' as its unweighted BBH edge set. The S' has, by construction, a polynomial bound on the number of maximal cliques contained in it. These cliques can be enumerated in polynomial time (Rosgen and Stewart, 2007).

Then, starting from the initial bicliques (the BBHs), the cliques in S' are merged into larger cliques, if all peaks within the potential clique are BBHs of each other. Otherwise, the largest peak group that is also a clique is retained, whereas all peaks that are not BBHs of at least one of the retained peaks in the group are removed. Merging is continued until all cliques have been processed. After completion of the merging phase, cliques with at least k peaks, as controlled by the minimum clique size (MCS) parameter, are reported in a multiple alignment table, ordered by their median RT. Peaks that are not included in any of the final cliques are optionally reported in the automated mass spectal deconvolution and identification system (AMDIS)-compatible mass spectral program (MSP) format (Stein, 1999) with their mass spectrum, RT, originating file and a unique ID for manual inspection.

Following Hoffmann *et al.* (2012), the time and space complexity of BiPACE 2D and BiPACE are equivalent to $\mathcal{O}(K^2\ell^2)$ in time and $\mathcal{O}(K^2\ell)$ in space, where K is the number of chromatograms and ℓ the upper bound of the number of peaks in each chromatogram. The lower asymptotic space complexity results from the BBH selection process that stores for every peak in one chromatogram only its best corresponding peak in every other chromatogram.

2.5 Reference dataset generation

To evaluate their 2D RT alignment algorithm mSPA, Kim *et al.* (2011b) use the raw peak lists as created by the ChromaTOF software (LECO Corp, St Joseph, MI, USA) and create reference multiple alignments based on the assigned peak names. Because each peak list can contain multiple peaks with the same name, the authors use a method to resolve such potential conflicts by selecting the peak with the largest recorded area as the representative for a group of otherwise identically named peaks. This approach is referenced in the remainder of this article as 'grouping by maximum area' (GMA). As we have investigated, GMA may lead to arbitrary and spurious assignments of peaks to the same alignment row (with identical names across peak reports), hampering the clear definition of what *true* and *false* positives as well as negatives are. Examples of potentially problematic assignments for mSPA dataset I (see Section 3.1 for details) are the compounds (*naphthalene*), (*naphthalene, 2-methyl-*), (*anthracene*), (*benzo[ghi]perylene*), (*indeno[1,2,3-cd]pyrene*), (*phenol, pentachloro-*) and (*phenol, 2,3,5,6-tetrachloro-*), all of which appear to have been assigned the wrong name in a number of cases, as is shown in Supplementary Figure S1 of Supplementary File S1. Corresponding plots for the other datasets are available as Supplementary Figures S2–S4 in the same supplementary file. The complete data comparing the GMA reference creation approach to our proposed approach for all datasets used in this work are contained in Supplementary File S2 for each dataset.

2.5.1 The MGMA approach To address the issues with the approach used by Kim *et al.* (2011b), the reference generation method was modified to remove spurious assignments that relate back to potentially false assignments of peak names by the ChromaTOF software. As mentioned previously, additional problems may arise from the selection process that relies solely on picking the peak with the largest area as the representative for a group of identically named peaks within each report.

The new method 'modified grouping by maximum area' (MGMA) calculates, for all equally named peaks (peak groups) in all peak reports, the standard deviation of the RTs in the first and second dimension of separation.

For an arbitrary group of peaks P with the same name, $x := \sigma(t_1(P))$ and $y := \sigma(t_2(P))$ are defined as the standard deviations of the peak group RTs in the first (t_1) and second (t_2) dimension of separation. An elliptical function centered at $x_0 = 0$ and $y_0 = 0$ and with a major radius defined by a (maximum allowed standard deviation for t_1) and minor radius defined by b (maximum allowed standard deviation for t_2) is then used to calculate the decision criterion z:

$$z = \frac{(x - x_0)^2}{a^2} + \frac{(y - y_0)^2}{b^2} \tag{2}$$

If $z \leq 1$, the group is retained. Otherwise, if $z > 1$, the group at (x, y) is outside of the bounds of the ellipse defined by x_0, y_0, a, b and is marked as a potential outlier group. The parameters used for a and b for the different datasets examined in Section 3 are given in Table 1.

Peaks belonging to outlier groups are removed by MGMA without further consideration, as a large deviation in one or both RT dimensions may be a strong hint toward wrongly assigned peak names. An approach to further discriminate the members of such groups may lead to additional sources of false peak assignments and has thus not been considered at this stage.

Additionally, all peaks that occur only once throughout all peak reports are removed, as they cannot provide any reliable grouping information and may again have resulted from spurious identifications by the vendor software due to different sample quality and/or non-optimal parameter settings used during peak detection and putative peak identification.

The final reference multiple alignment is then created using GMA on the remaining peaks. MGMA thus reports a completely contained subset of the original peaks as reported by GMA. An example for this is given in Section 3.3.

2.6 Peak alignment performance evaluation

A reference alignment peak group defines whether a peak, represented by its index in the original peak list, is present in a sample or absent. Each column in the reference alignment corresponds to one sample's peak list, whereas each row represents an aligned peak group, spanning multiple samples. The results of each alignment algorithm are tested against each reference alignment group until either a match is found or the group is reported to be non-assignable to a counterpart in the reference alignment.

If a reported peak group can be positively assigned, all of the group's peaks that are present in the corresponding reference alignment group are counted as true positives (TP). Peaks that are absent in both the reference and reported alignment peak group are counted as true negatives (TN). A peak that is reported as absent in the reference alignment group, but as present in the alignment algorithm's reported group, is recorded as a false positive (FP). A false negative (FN) is counted if a peak is present in the reference alignment group but not reported by the alignment algorithm's corresponding peak group.

The following commonly applied measures are used to assess the quality of a multiple alignment:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

For each multiple alignment result obtained from a method, all unmatched peaks of the reference alignment, excluding absent ones, are

Table 1. Parameters used for alignment reference generation for the different datasets. *a* is the major and *b* the minor radius of the elliptical decision boundary function *z* (Equation 2)

	Number of peaks in reference				
Dataset	*a*	*b*	GMA	MGMA	Manual
mSPA I	50.0	0.5	752	592	—
mSPA II	55.0	0.5	1682	1081	—
SWPA I	800.0	0.5	1201	1090	—
CHLAMY I	250.0	0.5	2723	1629	436

added to the number of FN to normalize Recall and F1 score with respect to the size of the reference alignment.

We discuss the advantages and disadvantages of our row-wise multiple alignment evaluation over the pairwise alignment evaluation as used by Kim *et al.* (2011a,b) in Supplementary File S1. In the remainder of this work, we refer to the value of the average F1 score based on pairwise alignment evaluation for each algorithm parameterization as $F1_p$. Differences in the resulting numbers for $F1_p$ to the numbers published in Kim *et al.* (2011a) are due to the evaluation scheme that we use. We compare all possible unique pairwise combinations of alignment column pairs against the corresponding reference alignment columns. Thus, our absolute TP, FP, TN and FN numbers are higher, and the $F1_p$ tends to be generally lower.

3 RESULTS AND DISCUSSION

The GNU R scripts available with Kim *et al.* (2011a,b) were carefully adapted to be able to run them within an automated evaluation pipeline. The mSPA and SWPA dataset peak lists were used unaltered as input to all evaluated programs, keeping peaks that were split across multiple modulations, while removing peak artifacts with an identical area. To make the gap-less multiple alignment output of mSPA and SWPA comparable with the gapped multiple alignment of BiPACE 2D, we modified the corresponding R-code to not remove incomplete peak groups. Guineu was modified to parse the ChromaTOF peak file format with separate fields for first and second column RTs and was further adapted to run without a graphical user interface and to record the original row index of each peak in the original peak list for later evaluation.

The algorithms BiPACE, BiPACE RT and BiPACE 2D were evaluated against Guineu's score alignment (Castillo *et al.*, 2011), mSPA (Kim *et al.*, 2011a) and its variants PAD, PAS, DW-PAS, SW-PAD and PAM, as well as against SWPA (Kim *et al.*, 2011b) and its variants SWRM, SWRE, SWRME and SWRME2. The mSPA, SWPA and Guineu methods used the ChromaTOF peak lists as input directly. For the BiPACE methods, we converted the ChromaTOF peak lists to a backwards compatible extended netCDF format (Rew and Davis, 1990), supporting first and second column elution time. BiPACE also supports mzML input files (Martens *et al.*, 2011), containing the standardized spectrum attributes *first_column_elution_time* and *second_column_elution_time*. The parameterizations reported as optimal by both the OP-PAM and the likelihood-based parameter optimization for the SWPA methods were explicitly included in the evaluation for each of the respective methods. The results for the mSPA SW-PAD variant using Pearson's correlation between spectra and Euclidean distance for RT matching correspond to the results of the DISCO algorithm (Kim *et al.*, 2011a). A detailed overview of the best results for each dataset and variant is available in Supplementary Tables S1–S4 of Supplementary File S1. Each algorithm was run and evaluated for a range of different parameter values. The user-configurable parameters (penalty terms, mass spectral score function, RT distance function) for mSPA and SWPA were taken from the corresponding publications (Kim *et al.*, 2011a,b). We tested all viable combinations of score function (dot product, linear correlation) and RT distance functions (Manhattan, Euclidean, Canberra, Maximum). Kim and Zhang (2013) provide a recent comparison of mSPA using an additional set of similarity functions that were not evaluated here. For BiPACE and its variants, the varied parameters included the mass spectral score function, RT penalty terms (BiPACE RT and BiPACE 2D) and RT penalty threshold (BiPACE RT and BiPACE 2D). The parameter values for all methods are available for each dataset individually within Supplementary File S2. Plots of the runtime and memory usage of each parameterized method are included in Supplementary File S1. They reflect only the peak alignment phase, not the data import and filtering phases of the algorithms.

3.1 mSPA datasets

The authors of the mSPA publication (Kim *et al.*, 2011a) evaluated their algorithms on two different datasets. The first one, here termed mSPA dataset I, consists of 10 samples of 106 standard compound mixtures, measured throughout with the same temperature gradient. It contains 1672 peaks in total, of which 752 in 81 rows were used in the GMA reference alignment. These were further reduced to 592 peaks in 64 rows by MGMA.

The second dataset, mSPA dataset II, contains five samples of rat plasma with spiked-in 6–compound standards, also measured under identical temperature gradient conditions. The original peak reports contained 3575 peaks. These were reduced to 1682 peaks in 493 rows by GMA's reference alignment generation, and further reduced by MGMA to 1081 peaks in 320 rows.

Table 1 holds the parameters used to generate the MGMA reference alignments, for mSPA dataset I and mSPA dataset II, respectively.

3.1.1 Results for mSPA Dataset I Figure 1 shows the F1 score obtained for each of the methods under consideration, against both GMA and MGMA reference alignments. For both references, BiPACE 2D achieves the highest F1 scores (0.9296 for GMA and 0.9472 for MGMA reference), followed by BiPACE RT (0.9181 and 0.9322). The Guineu (0.9082 and 0.9165) and mSPA-PAM (0.905 and 0.9169) variants follow closely behind. BiPACE 2D also has a consistently better precision value (0.9551 and 0.9607) than any of the other methods. On both references, the best BiPACE 2D instance uses RT penalty parameters of $D_1 = 10$, $D_2 = 0.5$, $MCS = 2$, the dot product as mass spectral similarity and RT penalty thresholds of $T_1 = 0$ and $T_2 = 0.99$, effectively allowing only small differences in the second dimension RT. BiPACE 2D also achieves the best average pairwise $F1_p$

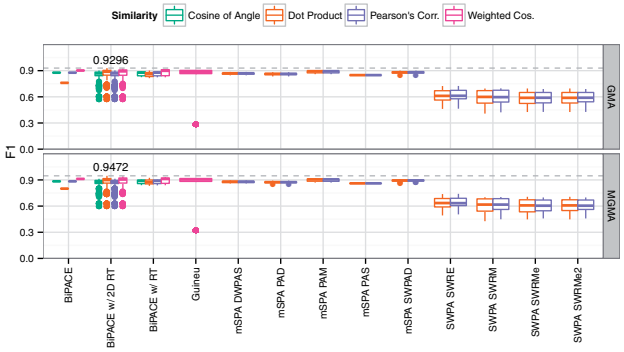

Fig. 1. F1 score for all parameterizations of the evaluated algorithms for mSPA dataset I. BiPACE 2D using the weighted cosine as similarity function between mass spectra outperforms all other methods

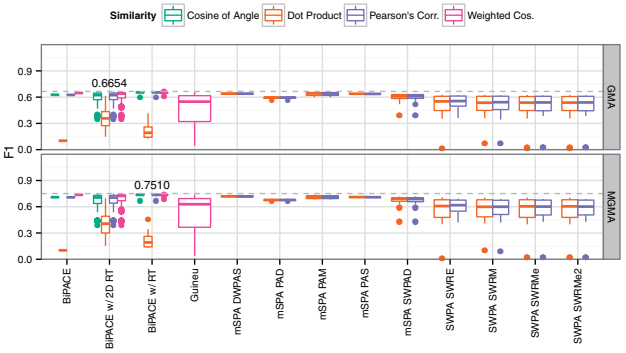

Fig. 2. F1 score for all parameterizations of the evaluated algorithms for mSPA dataset II. BiPACE 2D and BiPACE RT perform better than any other method using the weighted cosine and Pearson's linear correlation as similarity functions between mass spectra

scores, 0.9203 ± 0.022 with $D_1 = 10$, $D_2 = 0.25$, $MCS = 2$, $T_1 = 0$, $T_2 = 0.25$ and 0.9374 ± 0.021 with $D_1 = 10$, $D_2 = 0.5$, $MCS = 2$, $T_1 = 0$, $T_2 = 0.99$, each time using the dot product as mass spectral similarity. More details may be found in Supplementary File S1, Section 3.

3.1.2 Results for mSPA Dataset II Comparing the F1 score for mSPA dataset II, BiPACE 2D (0.6654 on GMA reference) and BiPACE RT (0.751 on MGMA with $D = 25$, $T = 0.25$) perform better than any Guineu, mSPA or SWPA variant (Figure 2). The best instances use the weighted cosine or cosine mass spectral score function, or Pearson's linear correlation, and not the dot product, in comparison with the results in mSPA dataset I, where the dot product was more competitive. The best BiPACE 2D instance on the GMA reference also achieves the highest $F1_p$ value (0.6857 ± 0.0264) with parameters $D_1 = 25$, $D_2 = 0.5$, $MCS = 2$ and RT penalty thresholds of $T_1 = 0.75$ and $T_2 = 0$, effectively allowing only small differences in the first dimension RT, whereas allowing larger differences in the second dimension RT. BiPACE RT scores the highest $F1_p$ value of 0.8231 ± 0.0201 on the MGMA reference with $D = 30$, $T = 0.9$ and $MCS = 2$. The F1 and $F1_p$ values for Guineu, mSPA and the SWPA variants do not fall far behind in this case on either reference alignment in comparison with the other datasets (see Supplementary File S1, Section 4 for more details).

3.2 SWPA dataset

The SWPA publication (Kim *et al.*, 2011b) used two different datasets for evaluation purposes. However, SWPA dataset II (spiked-in) was excluded from this evaluation because it was identical to mSPA dataset II. SWPA dataset I is a combination of 16 samples, measured using three different temperature gradients. Therefore, it should be a significant challenge for RT-based algorithms. The dataset originally contained 2499 peaks, which were reduced to 1201 peaks in 83 alignment rows by GMA, and to 1090 peaks in 75 rows by MGMA. The parameters for the MGMA reference alignment are given in Table 1.

3.2.1 Results for SWPA dataset I For this dataset, the mSPA-PAM variant using dot product as pairwise spectral similarity and maximum distance for RT difference performs best when

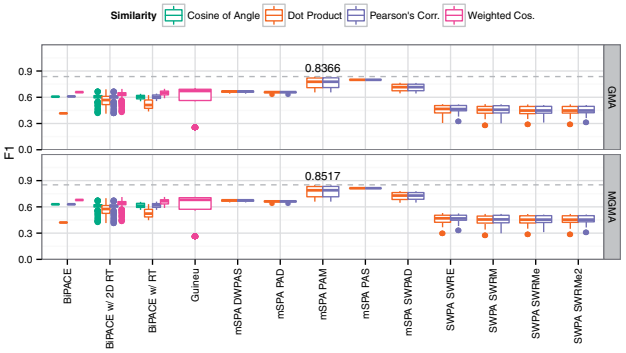

Fig. 3. F1 score for all parameterizations of the evaluated algorithms for SWPA dataset I. mSPA-PAM, mSPA-PAS and mSPA-SW-PAD perform better than the BiPACE, SWPA and Guineu variants

considering F1 score, with values of 0.8366 (GMA reference) and 0.8517 (MGMA reference). The best Guineu instance achieves values of 0.7061 and 0.7148, whereas the best BiPACE instance achieves F1 values of 0.6966 and 0.7136, respectively. More details are shown in Figure 3. Considering the $F1_p$ score, the order is unchanged, with mSPA-PAM achieving 0.7942 ± 0.0914 and 0.8101 ± 0.0882, Guineu scoring 0.5142 ± 0.314 and 0.529 ± 0.309 and BiPACE trailing with 0.4772 ± 0.248 and 0.4995 ± 0.2479, on GMA and MGMA references.

The comparatively low F1 score for all BiPACE variants can be explained by many peaks that are reported as being absent, while they are present in either reference alignment. Those absent peaks are counted as FNs and thus lead to a low Recall value (Supplementary File S1, Section 5). However, the BiPACE variants perform better when considering TN and FP values. They report fewer peaks per peak group, resulting in a more conservative alignment when compared with either the mSPA or SWPA variants, at the expense of more TPs.

BiPACE achieves high precision values for either reference (0.9049 and 0.9146), but lacks in Recall (0.5174 and 0.5397), leading to the comparatively low F1 and $F1_p$ scores, whereas BiPACE 2D has slightly lower precision values (0.8588 and 0.8631) but higher Recall (0.5841 and 0.607). Detailed numbers

for the best instances are given in Supplementary File S1, Table S3. A more detailed table including individual parameterizations of the best instances is contained in Supplementary File S2.

3.3 *Chlamydomonas reinhardtii* dataset

The *C.reinhardtii* dataset (CHLAMY dataset I) was originally analyzed in Doebbe *et al.* (2010). The experiment explored the difference in H_2 production yield between the *C.reinhardtii* wild-type strain *cc406* (WT) and the high H_2-producing strain *Stm6Glc4* (MUT) at two different time points, namely, before (T1) and during (T2) the H_2 production phase, with three replicates for each of the factor combinations WT-T1, WT-T2, MUT-T1, MUT-T2, yielding a total of 12 samples. For this article, the stored original samples of that experiment were prepared according to the protocol in Doebbe *et al.* (2010) and then reanalyzed using a LECO Pegasus 4D time-of-flight mass spectrometer (LECO, St Joseph, MI, USA). The Pegasus 4D system was equipped with an Agilent 6890 gas chromatograph (Agilent, Santa Clara, CA, USA).

3.3.1 Sample acquisition Splitless injection of 1 µl sample volume was conducted at 275°C injector temperature. The gas chromatograph was equipped with a 30 m × 0.25 mm × 0.25 µm film thickness, Rtx-5 ms (Restek Corp., Bellefonte, PA, USA) capillary column used as the primary column and a BPX-50 (SGE Incorporated, Austin, TX, USA) 2 m × 0.1 mm × 0.1 µm capillary column used as the secondary column. The temperature program of the primary oven was set to the following conditions: 70°C for 2 min, 4°C/min to 180°C, 2°C/min to 230°C and 4°C/min to 325°C hold 3 min. The temperature program of the secondary oven was set with an offset of 15°C to the primary oven temperature. The thermal modulator was set 30°C relative to the primary oven and used a modulation time of 5 s with a hot pulse time of 0.4 s. The mass spectrometer ion source temperature was set to 200°C, and the ionization was performed at −70 eV. The detector voltage was set to 1600 V and mass spectra were recorded at 200 scans/second using a scanning range of 50–750 m/z.

3.3.2 Sample processing The samples were processed automatically by the LECO ChromaTOF software v.4.22 at a signal-to-noise ratio of 100. The baseline offset was 0.8 and the two peak widths were set to 0.2 s (as measured from baseline to baseline) and 15 s (first dimension). By using the classification feature of the software, background peaks originating from column bleed or solvent tailing were removed.

Analytes were putatively identified by database searches using the Golm Metabolome Database, version 20100614 (Hummel *et al.*, 2007). The minimum required similarity threshold for assignment of a compound name was set to 600. The original ChromaTOF peak lists contained a total of 31 695 peaks for the 12 samples. All peaks with best matching library spectrum similarity <600, flagged as 'Unknown', were removed from further consideration. The original peak lists were exported from ChromaTOF using one field [R.T. (s)] for the first and second column elution time. Therefore, we introduced two separate columns for first and second column elution time ('1st Dimension Time (s)' and '2nd Dimension Time (s)') to make them suitable as input to both MSPA and SWPA.

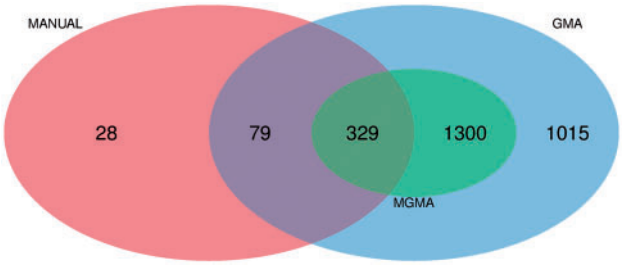

Fig. 4. Venn diagram of the peak set overlap for CHLAMY dataset I for GMA, MGMA and manual multiple alignment reference generation

The resulting peak lists for each sample were further rectified by removing all peaks with unclear Golm Metabolome Database identifications containing an 'NA'. These steps were required to make the peak lists compatible to MSPA's and SWPA's peak merging preprocessing step, which was needed for the generation of the evaluation reference alignments with GMA and MGMA. These unknowns would otherwise have lead to false peak group assignments based on the peaks' non-unique names. The removal of 'Unknown' peaks and rectification of 'NA's reduced the number of peaks to a total of 4860. The final GMA reference alignment contained 2723 peaks in 369 rows, whereas the MGMA reference, using the parameters given in Table 1, contained 1629 peaks in 224 rows.

3.3.3 Manual reference To define the manual reference alignment, the reduced peak lists without 'Unknown's and 'NA's were inspected and only peaks were kept that could be positively confirmed by assigned name and RTs within two of the three samples within each factor combination of the experiment. For a number of unclear cases, we additionally compared the mass spectra of the questionable peaks manually to check for common characteristic mass fragments. The final manual reference alignment contained 436 peaks grouped into 68 distinct rows.

The overlap of the three reference multiple alignments is visualized in Figure 4. As expected, the MGMA reference is a perfect subset of the GMA reference. More interestingly, there is a large overlap of the manual reference with the automated methods' reference alignments, supporting the claim that these automated methods for reference generation based on the assigned compound names capture most of the peak alignments contained in the manual reference alignment. A small proportion of 28 peaks (6.4% of peaks in the manual reference) occurs exclusively within the manual alignment and not within any of the automated methods. Of these, 24 peaks were differently assigned in the automated methods versus the manual reference, whereas four peaks were not reported at all by those methods. In all, 149 peaks were found missing from the manual reference in comparison with the GMA reference generation method, due to the stricter selection criteria that were used to exclude potential FP peak assignments.

The manual reference alignment, the peak reports for each sample as exported from ChromaTOF and the raw data files in netCDF format are available as dataset MTBLS37 from the MetaboLights database (Haug *et al.*, 2013).

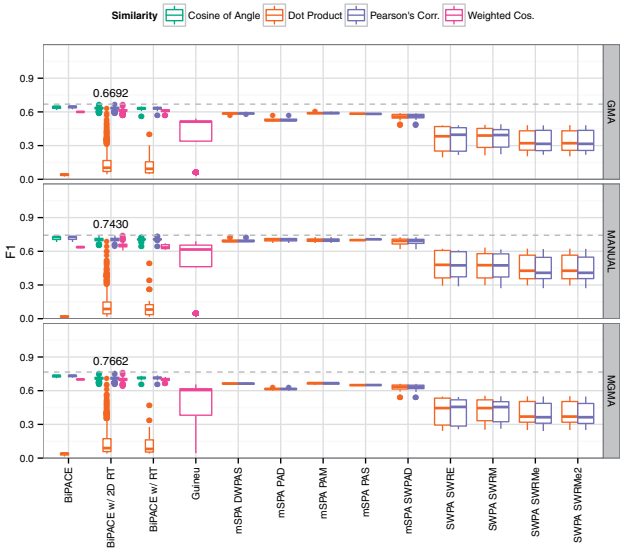

Fig. 5. F1 score for all parameterizations of the evaluated algorithms for CHLAMY dataset I. BIPACE 2D performs clearly better than any of the other methods using the dot product as pairwise similarity between mass spectra

3.3.4 Results for CHLAMY dataset I All three BIPACE variants using either the cosine or Pearson's linear correlation as similarity functions between mass spectra perform better than either GUINEU, MSPA or SWPA variant (Figure 5). BIPACE 2D achieves F1 scores of 0.6692 (GMA reference), 0.7662 (MGMA reference) and 0.7429 (MANUAL reference), with $D_1 = 100$, $D_2 = 0.5$, thresholds $T_1 = 0.99$ and $T_2 = 0.99$, together with an *MCS* value of 2. The F1 values are visualized in Figure 5. BIPACE 2D also achieves the highest Recall values of 0.5752 (GMA), 0.7079 (MGMA) and 0.7596 (MANUAL), whereas still maintaining reasonable values for Precision between 0.72 and 0.835 (Supplementary File S1, Section 6). The considerably low values for Recall achieved by the different methods may be due to the complexity of the biological samples and the large number of closely related peaks and associated peak areas. The best average pairwise $F1_p$ scores are also achieved by BIPACE 2D (GMA: 0.7198 ± 0.0335, MGMA: 0.8293 ± 0.0247, MANUAL: 0.864 ± 0.0606), with GUINEU (GMA: 0.6434 ± 0.0419, MGMA: 0.7766 ± 0.0338, MANUAL: 0.8481 ± 0.0528) placing second and different MSPA variants placing third (GMA, MSPA-PAM: 0.5976 ± 0.1283, MGMA, MSPA-DWPAS: 0.6907 ± 0.1296, MANUAL, MSPA-SW-PAD: 0.7475 ± 0.1741).

Supplementary File S1, Supplementary Table S4 holds a more detailed table of the best results for CHLAMY dataset I. Individual parameterizations of the best instances are contained in Supplementary File S2.

4 CONCLUDING REMARKS

We have introduced BIPACE 2D, an algorithm and software for RT alignment of peak data from GC × GC–MS experiments. It is implemented in the platform-independent Java programming language and aligns peak data in vendor-independent formats, such as mzML (Martens *et al.*, 2011) or netCDF. ChromaTOF peak lists can be converted to netCDF files using the software Maui (also available at http://maltcms.sf.net).

We have shown that BIPACE 2D is a competitive algorithm that is able to achieve better precision and recall, as well as F1 and average pairwise $F1_p$ score values in comparison with MSPA, SWPA and GUINEU on three of the four evaluated datasets. These three datasets were all acquired under homogeneous conditions, whereas the dataset where BIPACE 2D did not outperform MSPA and GUINEU was acquired under heterogeneous conditions. Thus, BIPACE 2D should ideally be applied to data acquired under the same conditions, but due to its low FP rate, it may still be a valid alternative for data acquired under heterogeneous conditions as well. Concerning the parameters for BIPACE 2D, the weighted cosine appears to be the most sensitive mass spectral similarity and should therefore be used as the default. The *MCS* parameter was set to the minimum size of two in all evaluated parameterizations, thus leading to all cliques being reported by BIPACE and its variants. The $D1$ and $D2$ parameters should be set according to the expected RT standard deviation of the samples under comparison, in separation dimensions one and two, respectively. Finally, the threshold parameters $T1$ and $T2$ allow for fine-tuning of the sensitivity of the algorithm, where higher values exclude potential matches earlier during the pairwise similarity calculation phase of BIPACE. It is further notable that BIPACE 2D was on average 3–10 times faster than any of the MSPA or SWPA variants for the larger and more complex datasets (SWPA dataset I and CHLAMY dataset I, see Supplementary File S1 for details), while consuming less memory. GUINEU achieved comparable speed but required more memory. We have demonstrated the applicability of BIPACE 2D to small datasets with a few compounds, as well as to larger datasets with hundreds to suspected thousands of different compounds. BIPACE's pairwise similarity calculation can be run in parallel using multiple CPU cores to speed up its runtime. It has been successfully tested on 250 files containing 100 000 peaks on commodity hardware within 9 GB of random access memory. This qualifies BIPACE 2D as a good candidate for automated medium to high-throughput applications in the field of metabolomics and analytical chemistry.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their valuable comments and suggestions. They would also like to thank Matthias Keck for help on sample acquisition and setup of the ChromaTOF processing method. They additionally would like to acknowledge Pablo Conesa Mingo and Reza Salek on behalf of the whole MetaboLights development and curation team for their help during the *C.reinhardtii* study submission process.

Conflict of Interest: none declared.

REFERENCES

Amador-Muñoz,O. and Marriott,P.J. (2008) Quantification in comprehensive two-dimensional gas chromatography and a model of quantification based on selected summed modulated peaks. *J. Chromatogr. A*, **1184**, 323–340.

Arey,J.S. *et al.* (2005) Using comprehensive two-dimensional gas chromatography retention indices to estimate environmental partitioning properties for a complete set of diesel fuel hydrocarbons. *Anal. Chem.*, **77**, 7172–7182.

Castillo,S. *et al.* (2011) Data analysis tool for comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry. *Anal. Chem.*, **83**, 3058–3067.

Doebbe,A. *et al.* (2010) The interplay of proton, electron, and metabolite supply for photosynthetic h_2 production in *Chlamydomonas reinhardtii*. *J. Biol. Chem.*, **285**, 30247–30260.

Fraga,C.G. *et al.* (2000) Comprehensive two-dimensional gas chromatography and chemometrics for the high-speed quantitative analysis of aromatic isomers in a jet fuel using the standard addition method and an objective retention time alignment algorithm. *Anal. Chem.*, **72**, 4154–4162.

Haug,K. *et al.* (2013) MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.*, **41**, D781–D786.

Hoffmann,N. and Stoye,J. (2012) Generic software frameworks for GC-MS based metabolomics. In: Roessner,U. (ed.) *Metabolomics*. InTech, Rijeka, pp. 73–98.

Hoffmann,N. *et al.* (2012) Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets. *BMC Bioinformatics*, **13**, 214.

Hummel,J. *et al.* (2007) The Golm metabolome database: a database for GC-MS based metabolite profiling. In: Nielsen,J. and Jewett,M.C. (eds) *Metabolomics, number 18 in Topics in Current Genetics*. Springer Berlin, Heidelberg, pp. 75–95.

Jeong,J. *et al.* (2012) Model-based peak alignment of metabolomic profiling from comprehensive two-dimensional gas chromatography mass spectrometry. *BMC Bioinformatics*, **13**, 27.

Johnson,K.J. *et al.* (2004) Quantification of naphthalenes in jet fuel with GC×GC/Tri-PLS and windowed rank minimization retention time alignment. *J. Sep. Sci.*, **27**, 410–416.

Kallio,M. *et al.* (2009) Data analysis programs for comprehensive two-dimensional chromatography. *J. Chromatogr. A*, **1216**, 2923–2927.

Karp,R.M. (1972) Reducibility among combinatorial problems. In: Miller,R.E. and Thatcher,J.W. (eds) *Complexity of Computer Computations*. Plenum Press, New York, London, pp. 85–103.

Kim,S. and Zhang,X. (2013) Comparative analysis of mass spectral similarity measures on peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry. *Comput. Math. Methods Med.*, **2013**.

Kim,S. *et al.* (2011a) An optimal peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry using mixture similarity measure. *Bioinformatics*, **27**, 1660–1666.

Kim,S. *et al.* (2011b) Smith-Waterman peak alignment for comprehensive two-dimensional gas chromatography-mass spectrometry. *BMC Bioinformatics*, **12**, 235.

Koek,M.M. *et al.* (2011) Semi-automated non-target processing in GC × GC–MS metabolomics analysis: applicability for biomedical studies. *Metabolomics*, **7**, 1–14.

Martens,L. *et al.* (2011) mzml—a community standard for mass spectrometry data. *Mol. Cell. Proteomics*, **10**, R110.000133.

Matos,J.T. *et al.* (2012) Trends in data processing of comprehensive two-dimensional chromatography: state of the art. *J. Chromatogr. B*, **910**, 31–45.

Oh,C. *et al.* (2008) Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry peak sorting algorithm. *J. Chromatogr. A*, **1179**, 205–215.

Overbeek,R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.

Pierce,K.M. *et al.* (2006) Fisher ratio method applied to third-order separation data to identify significant chemical components of metabolite extracts. *Anal. Chem.*, **78**, 5068–5075.

Porter,S.E.G. *et al.* (2006) Analysis of four-way two-dimensional liquid chromatography-diode array data: application to metabolomics. *Anal. Chem.*, **78**, 5559–5569.

Reichenbach,S.E. *et al.* (2012) Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography. *J. Chromatogr. A*, **1226**, 140–148.

Rew,R.K. and Davis,G.P. (1990) NetCDF: an interface for scientific data access. *IEEE Comput. Graph. Appl.*, **10**, 76–82.

Robinson,M. *et al.* (2007) A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments. *BMC Bioinformatics*, **8**, 419.

Rosgen,B. and Stewart,L. (2007) Complexity results on graphs with few cliques. *Discrete Math. Theor. Comput. Sci.*, **9**, 127–136.

Stein,S. (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.*, **10**, 770–781.

Stein,S.E. and Scott,D.R. (1994) Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass. Spectrom.*, **5**, 859–866.

Ventura,G.T. *et al.* (2011) Analysis of petroleum compositional similarity using multiway principal components analysis (MPCA) with comprehensive two-dimensional gas chromatographic data. *J. Chromatogr. A*, **1218**, 2584–2592.

Vial,J. *et al.* (2009) Combination of dynamic time warping and multivariate analysis for the comparison of comprehensive two-dimensional gas chromatograms: application to plant extracts. *J. Chromatogr. A*, **1216**, 2866–2872.

von Mühlen,C. *et al.* (2006) Applications of comprehensive two-dimensional gas chromatography to the characterization of petrochemical and related samples. *J. Chromatogr. A*, **1105**, 39–50.

Wang,B. *et al.* (2010) DISCO: distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. *Anal. Chem.*, **82**, 5069–5081.

Wei,X. *et al.* (2013) MetPP: a computational platform for comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. *Bioinformatics*, **29**, 1786–1792.