Simon Dellicour[1]
Thomas Lecocq[2]

[1]Evolutionary Biology and Ecology, Université Libre de Bruxelles, Brussels, Belgium
[2]Laboratoire de Zoologie, University of Mons, Mons, Belgium

Received April 11, 2013
Revised June 21, 2013
Accepted July 16, 2013

Short Communication

GCALIGNER 1.0: An alignment program to compute a multiple sample comparison data matrix from large eco-chemical datasets obtained by GC

GCALIGNER 1.0 is a computer program designed to perform a preliminary data comparison matrix of chemical data obtained by GC without MS information. The alignment algorithm is based on the comparison between the retention times of each detected compound in a sample. In this paper, we test the GCALIGNER efficiency on three datasets of the chemical secretions of bumble bees. The algorithm performs the alignment with a low error rate (<3%). GCALIGNER 1.0 is a useful, simple and free program based on an algorithm that enables the alignment of table-type data from GC.

Keywords: Data processing / GC / Peak alignment / Retention time alignment
DOI 10.1002/jssc.201300388

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

The study of chemicals involved in biological interactions is one of the most studied and interesting frameworks in biology because all living organisms emit, detect, and respond to chemical cues. Chemical ecology deals with molecularly mediated biological interactions [1]. It covers a broad range of chemical interactions like chemical communication, mutualistic interactions, or chemical defences of organisms [2] and also includes synthesis, emission, transmission, and detection of chemicals. Chemical information is widely applied in ecology [3–6], ethology [7, 8], and taxonomy [9, 10].

Any modern chemo-ecological study needs the characterization of chemicals as a prerequisite and requires the comparison of many samples by chemical analyses. Advances in analytical chemistry are highly beneficial to this discipline [1]. The instruments currently used in analytical chemistry (e.g. GC) allow a fast and easy generation of large high-dimensional data. Within the framework of biology, and particularly in chemo-ecology, rapid comparison of many samples is needed in order to support hypothesis testing [1]. The huge amount of samples makes the computing of a sample comparison data matrix unworkable without automated electronic methods [11]. Several computer programs have been developed to build aligned matrices from chemical compounds [11–16]. Furthermore, many algorithms allowing the alignment of GC–MS/LC–MS data have been designed and published [11]; however, only some of them are easily accessible or included in publicly available toolkits [13]. Most of these algorithms use data produced by GC–MS (retention time space, m/z space and analytical run space [17, 18]). However, many similar samples can be repeatedly analyzed by chromatography methods to obtain a large and robust dataset without the need to additionally/necessarily perform MS (i.e. GC–FID).

Here, we propose GCALIGNER 1.0, a program allowing the alignment of samples analyzed by GC based on retention time (RT) or Kovats retention indices (I) without MS information. It has been specifically designed to align chemical compounds detected by GC in different samples, thereby allowing the comparison of the composition of a high numbers of samples. We have tested our alignment program on chemical datasets from cephalic labial gland secretions (CLGS) of three bumble bee species. In addition, we also present GCKOVATS 1.0, a program to compute I prior to GCALIGNER run.

Correspondence: Mr. Thomas Lecocq, Laboratoire de Zoologie, University of Mons, Place du Parc, 20, Mons 7000, Belgium
E-mail: thomas.lecocq@umons.ac.be
Tel: 003265373405

Abbreviations: CLGS, cephalic labial gland secretions; **FID**, flame ionization detector; ***I***, Kovats retention index; **RA**, relative amounts; **RT**, retention time

2 Materials and methods

2.1 Alignment algorithm

The algorithm implemented in GCALIGNER builds a new matrix by analyzing each RT row of the initial nonaligned dataset. The dataset is made up of the RT linked to the relevant data (e.g. area integration of each peak). The same procedure is applied to each row (i):

(a) Detection of the smallest RT $t_{S(i)}$ on this row.

(b) Detection of the nearest (not yet analyzed) RT $t_{N(i)}$ from $t_{S(i)}$ on this row.

(c) $t_{N(i)}$ is considered "homologous" to $t_{S(i)}$ if $\alpha * (t_{N(i)} - t_{S(i)}) < (t_{N(i+1)} - t_{N(i)})$, if $\alpha * (t_{N(i)} - t_{S(i)}) < (t_{S(i+1)} - t_{S(i)})$ and if $\alpha * (t_{N(i)} - t_{S(i)}) < (t_{S(i)} - t_{H(i-1)})$ in which α is a weight parameter allowing the systematic increase or decrease of the rate of acceptance of an RT as "homologous" to $t_{S(i)}$, $t_{S(i+1)}$ is the RT following $t_{S(i)}$ in row $(i + 1)$ of the same injection/sample, $t_{N(i+1)}$ is the RT following $t_{N(i)}$ in row (i+1) of the same injection/sample and $t_{H(i-1)}$ is the highest RT of the last aligned row in the final matrix.

(d) If $t_{N(i+1)} < t_{S(i+1)}$, $t_{S(i+1)}$ is replaced by $t_{N(i+1)}$ until the last row of the analysis is reached.

Steps (b)–(d) are repeated until all the RTs of row (i) have been analyzed. At each repetition of step (2), $t_{N(i)}$ is determined as the nearest RT from row (i) that has not yet been analyzed in one of the preceding loops (2)–(4). Once all the RTs of row (i) have been analyzed, data corresponding to the set of "homologous" RTs (i.e. area, relative area) are copied into the new final matrix. Data corresponding to "nonhomologous" RTs are simply replaced by empty spaces in this final matrix and reported in the next row $(i + 1)$ of the subsequent row analysis.

This iterative process is performed until all the data of the initial matrix have been copied into the final aligned matrix. For all rows, the algorithm is based on the comparison between the RT of one specific row (i) and the RT of the following row $(i + 1)$, except for the last row in which data are simply copied at the end of the final matrix without any alignment and the first line, $t_{H(i-1)}$, which is simply equal to 0.

Users are invited to start with the default value of $\alpha = 1$. They can then decrease (i.e. $\alpha = 0.5, 0.25, 0.125$) or increase (i.e. $\alpha = 2, 3, 4$) this parameter in order to check whether better preliminary alignment results can be obtained by a different starting value for α. The choice of the initial value for α will depend on the average difference between compound RTs.

2.2 Computation of Kovats retention indices

The GCKOVATS algorithm converts RT into system-independent constants using the Kovats retention index method [19, 20] for temperature-programed chromatography, also called linear retention index. For the computation, *n*-alkanes serve as standards and interpolations are made. The *I* of a chemical compound is its RT normalized to the RT of adjacently eluting *n*-alkanes:

$$I_{(x)} = 100 * \left[n + (N - n) \frac{tr_{(x)} - tr_{(n)}}{tr_{(N)} - tr_{(n)}} \right] \quad (1)$$

in which $I_{(x)}$ is the Kovats retention index of the target compound *x*, *n* is the number of carbon atoms in the *n*-alkane directly eluting before *x*, *N* is the number of carbon atoms in the *n*-alkane larger than *x*, $tr_{(x)}$ is the RT of the target compound *x*, $tr_{(n)}$ is the RT of the target compound *x* of the *n*-alkane eluting before *x* and $tr_{(N)}$ is the RT of the target compound *x* of the *n*-alkane eluting after *x*.

Therefore, the computation of *I* of samples requires calibration information (RT of *n*-alkanes analyzed under exactly the same chromatographic conditions of those of the other samples). The range of employed *n*-alkanes has to cover the expected RT range of all possible target compounds.

2.3 Test case

We tested the GCALIGNER efficiency by performing an alignment of three chemical datasets of CLGS from the following North American bumble bee species: *Bombus* (*Pyrobombus*) *bimaculatus* Cresson 1863, *Bombus* (*Pyrobombus*) *ephippiatus* Say 1837, and *Bombus* (*Pyrobombus*) *flavifrons* Cresson 1863. CLGS are species-specific complex mixtures of compounds, mainly aliphatic, produced by males in order to attract conspecific virgin females.

We sampled a total of 55 individuals, 24 of *B. bimaculatus* in Montreal (Canada, WGS84: 45.3025°N 73.3339°W), 20 of *B. ephippiatus* in Chiapas (commercial bee breeders colony, Biobest bvba, Westerlo, Belgium) and 11 of *B. flavifrons* in Vancouver (Canada WGS84: 49.2072°N 123.1491667°W). Males were frozen at –20°C and the CLGS were extracted from their heads in 200 μL *n*-hexane as described in Ref. [21].

Samples were analyzed by using a Shimadzu GC-2010 with a SLB-5ms nonpolar capillary column (5% diphenyl/95% dimethyl siloxane; 30 m × 0.25 mm × 0.25 μm) and a flame ionization detector (GC–FID). We used a splitless injection mode (220°C) and He as carrier gas (50 cm/s). The temperature program of the column was set to 70°C for 2 min and then increased at a rate of 10°C/min to 320°C. The temperature was then held at 320°C for 5 min. All samples from the same species were successively analyzed by GC–FID. The peak areas of the compounds were quantified by using GCsolution Postrun (Shimadzu Corporation) with automatic peak detection and noise measurement (Width = 1 s, Slope = 10 000 uV/min, Drift = 0 uV/min, T.DBL = 1000 min, Min.Area/Height = 1000 counts). Relative amounts (RA,%) of each compound were calculated by dividing the peak area of each compound by the total peak area of all compounds in the sample. No correction factor was used to calculate the RA of individual compounds.

Alignment of each dataset by GCALIGNER was performed from raw datasets (unaligned datasets; see Supporting Information Table S1) (Weight parameter $\alpha = 0.25$; Number of columns/injection = 3; columns are RT, peak area, and RA). In order to check aligned datasets (match between homologous RT), we determined the CLGS composition by GC–MS on a Finigan GCQ with a DB-5ms nonpolar capillary column (5% phenyl (methyl) polysiloxane stationary phase; 30 m × 0.25 mm × 0.25 μm) and an ion trap in electron impact mode "full scan (300–600)." The chromatographic conditions were identical to those of GC–FID. Compounds were identified in Xcalibur™ 2.0. by using their mass spectra

compared to those at the National Institute of Standards and Technology library (NIST, USA) using NIST MS Search 2.0. We quantified the error rate of alignments as:

$$\text{Error rate} = 100\left[\frac{N\,tr_{(mis)}}{N\,tr_{(alig)}N_s}\right] \quad (2)$$

in which $N\,tr_{(mis)}$ is the number of misaligned RTs, $N\,tr_{(alig)}$ is the number of RTs to align (number of lines in the corrected aligned dataset) and N_S is the number of samples to align.

3 Results and discussion

3.1 Test case results

We detected the following error rates: 1.24% in the *B. bimaculatus* dataset (2496 RT aligned), 2.67% in the *B. ephippiatus* dataset (1800 RT aligned) and 0.30% in the *B. flavifrons* dataset (671 RT aligned) (see Supporting Information; GCALIGNER alignments: Supporting Information Table S2; corrected alignments: Supporting Information Table S3). All errors were found to be related to abundant compounds with a large variability of RA that led to large gaps between homologous RTs (Fig. 1).

In the final datasets, we only kept compounds for which RAs were recorded to be higher than 0.1% for at least one sample [22]. We detected 41 chemical compounds in the CLGS of *B. bimaculatus* (main compounds: hexadecenol acetate and geranylgeranyl acetate), 64 in *B. ephippiatus* (main compounds: hexadecenol, hexadecanol), and 58 in *B. flavifrons* (main compound: geranylgeranial) huj (Supporting Information Table S4).

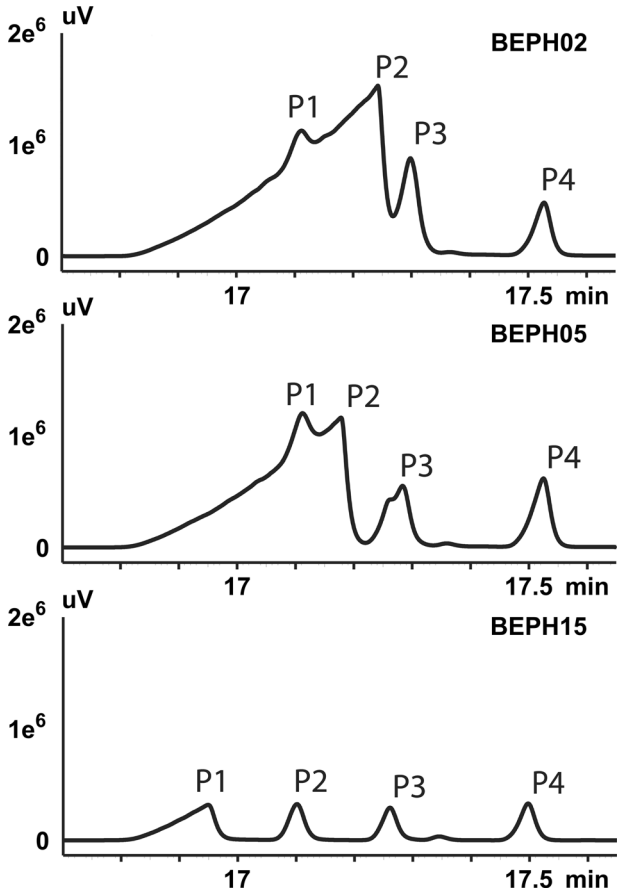

Figure 1. Algorithm limitations. Chromatograms of three *Bombus ephippiatus* (samples = BEPH02, BEPH05, BEPH15) between 16.7 and 17.7 min. Large gaps between homologue RTs (ex: P1 from BEPH02 and BEPH05 versus P1 from BEPH15) due to the large variability of abundance can lead to misalignments.

3.2 Software limitations

The main limitation of GCALIGNER resides in the fact that the last RT of each sample can not be aligned due to the alignment algorithm being based on the comparison between each RT and the RT in the following row.

Other limitations leading to misalignments are mainly the consequence of comparisons between samples of different composition: (i) comparison between compounds with large gaps between homologous RTs (e.g. homologous compounds with very different abundances); (ii) comparison between samples analyzed under different chromatographic conditions. The method based on RT needs to maintain exact identical chromatographic conditions. This last limitation can be largely resolved by calculating *I* with GCKOVATS.

The limitations of GCKOVATS are in fact those of the Kovats retention indices. This method cannot compensate for all processes that impact the gas chromatographic separation and reduce or modify column selectivity (e.g. initial acquisition delays or entirely different temperature programs). Moreover, different compounds might coincidentally coelute at the same RT, which somewhat limits the scope of GC and gives rise to the need for more enhanced analytical methods like GC–MS.

In conclusion, the aim of GCALIGNER is to perform a first preliminary alignment on large datasets. We thus strongly advise users to check all the alignment results obtained with this method.

3.3 Software availability

GCALIGNER and GCKOVATS are freely available from the website ebe.ulb.ac.be/ebe/Software.html where a user manual and examples of data files can also be found. GCALIGNER is a program written in java that runs on any operating system. The input file and the single output file containing the aligned dataset are tab delimited text files. See the software manual for further details.

We are grateful to B. Frérot, I. Valterová, N. Elvinger, L. Dohet, L. Grumiau, L. Hautier, P. Lhomme, N. Meurisse, N. Brasero, P. Rasmont, and T. Vantorre for help in testing the

software or for all their useful comments and advice. Special thanks are also due to K. Urbanová for her help in compound determination. This research project was funded by the Belgian Fonds National pour la Recherche Scientifique (FNRS) and the Fonds pour la Recherche dans l'Industrie et l'Agriculture (FRIA).

The authors have declared no conflict of interest.

4 References

[1] Meinwald, J., Eisner, T., *Proc. Natl. Acad. Sci. USA* 2008, *105*, 4539–4540.

[2] Hartman, T., *Proc. Natl. Acad. Sci. USA* 2008, *105*, 4541–4546.

[3] Youngsteadts, E., Nojima, S., Häberlein, C., Schulz, C., Schal, C., *Proc. Natl. Acad. Sci. USA* 2008, *105*, 4571–4575.

[4] Martin, S. J., Helanterä, H., Drijfhout, F. P., *Biol. J. Linn. Soc.* 2008, *95*, 131–140.

[5] Lecocq, T., Vereecken, N. J., Michez, D., Dellicour, S., Lhomme, P., Valterová, I., Rasplus, J.-Y., Rasmont, P., *PLoS One* 2013, *8*, e65642.

[6] Medeiros, P. M., Simoneit, B. R. T., *J. Sep. Sci.* 2007, *30*, 1516–1536.

[7] Vereecken, N. J., Mant, J., Schiestl, F. P., *Behav. Ecol. Sociobiol.* 2007, *61*, 811–821.

[8] Lhomme, P., Ayasse, M., Valterová, I., Lecocq, T., Rasmont, P., *PloS One* 2012, *7*, e43053.

[9] Bertsch, A., Schweer, H., Titze, A., Tanaka, H., *Insect Soc.* 2005, *52*, 45–54.

[10] Lecocq, T., Lhomme, P., Michez, D., Dellicour, S., Valterová, I., Rasmont, P., *Syst. Entomol.* 2011, *36*, 453–469.

[11] Bloemberg, T. G., Gerretzen, J., Lunshof, A., Wehrens, R., Buydens, L. M. C., *Anal. Chim. Acta* 2013, *781*, 14–32.

[12] Chae, M., Shmookler Reis, R. J., Thaden, J. J., *BMC Bioinformatics* 2008, *9*(Suppl 9), S15.

[13] Hoffmann, N., Stoye, J., *Bioinformatics* 2009, *25*, 2080–2081.

[14] Koh, Y., Pasikanti, K. K., Yap, C. W., Chan, E. C. Y., *J. Chromatogr. A* 2010, *1217*, 8308–8316.

[15] Li, Z., Wang, J.-J., Huang, J., Zhang, Z.-M., Lu, H.-M., Zheng, Y.-B., Zhan, D.-J., Liang, Y.-Z., *J. Sep. Sci.* 2013, *36*, 1677–1684.

[16] Zhang, Z.-M., Liang, Y.-Z., Lu, H.-M., Tan, B.-B., Xu, X.-N., Ferro, M., *J. Chromatogr. A* 2012, *1223*, 93–106.

[17] Mayfiel, H. T., Bertsch, W., *J. Comput. Appl. Lab.* 1983, *1*, 130–136.

[18] Lavine, B. K., Mayfield, H., Kromann, P. R., Faruque, A., *Anal. Chem.* 1995, *67*, 3846–3852.

[19] Kovats, E., *Helv. Chim. Acta* 1958, *41*, 1915–1932.

[20] IUPAC, *Compendium of Chemical Terminology*, Blackwell Scientific Publications, Oxford 1997.

[21] De Meulemeester, T., Gerbaux, P., Boulvin, M., Coppée, A., Rasmont, P., *Insect Soc.* 2011, *58*, 227–236.

[22] Terzo, M., Valterová, I., Rasmont, P., *Chem. Biodivers.* 2007, *4*, 1466–1471.

