GCALIGNER 1.0

>> manual (20-06-13)

S. Dellicour[1] & T. Lecocq[2] (2013)

(1) Evolutionary Biology and Ecology, Free University of Brussels, av. FD Roosevelt 50, 1050 Brussels, Belgium.
(2) *Laboratoire de Zoologie*, University of Mons, Place du Parc 20, 7000 Mons, Belgium.

GCALIGNER and GCAKOVATS are available from *ebe.ulb.ac.be/ebe/Software.html*
Contact and bugs report:
Thomas.Lecocq@umons.ac.be
Simon.Dellicour@ulb.ac.be

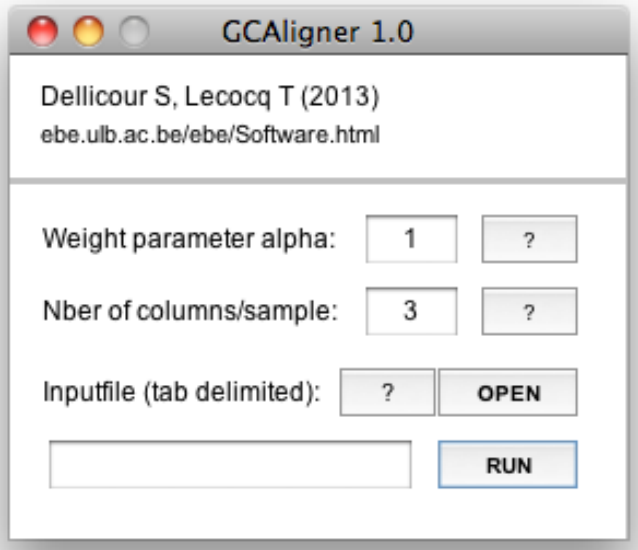

Figure 1: GCALIGNER interface.

1. GCALIGNER method

GCALIGNER aims to align GC data files originated from distinct injections/samples according to their retention times (or their Kovats retention indices, see 6.). It is designed to make the comparison of multiple samples easier. The alignment algorithm is based on the comparison between each retention time (RT), the following retention time in the same sample and its closest retention times in other samples.

The algorithm implemented in GCALIGNER builds a new matrix by analyzing each retention time row of the initial non-aligned dataset. The dataset is made up of the retention time linked to the interesting data (i.e. area integration of each peak). At each row (i), the same procedure is applied:

(1) detection of the smallest retention time $t_{S(i)}$ on this row.

(2) detection of the nearest (non-analyzed yet) retention time $t_{N(i)}$ from $t_{S(i)}$ on this row.

(3) $t_{N(i)}$ is judged "equivalent" (corresponding to the same compound as) 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 $\alpha\,(t_{N(i)} - t_{S(i)}) < (t_{S(i)} - t_{H(i-1)})$;

with α, a weight parameter allowing to systematically increase or decrease the rate of acceptance of a retention time as "equivalent" to $t_{S(i)}$;
$t_{S(i+1)}$, the retention time following $t_{S(i)}$ in row (i+1) on the same injection/sample ;
$t_{N(i+1)}$, the retention time following $t_{N(i)}$ in row (i+1) on the same injection/sample ;
$t_{H(i-1)}$, the highest retention time of the last aligned row in the final matrix.

(4) if $t_{N(i+1)} < t_{S(i+1)}$, $t_{S(i+1)}$ is replaced by $t_{N(i+1)}$ until the end of this row analysis.

Steps (2) to (4) are repeated until all the retention times of row (i) are analyzed. At each repetition of step (2), $t_{N(i)}$ is determined as the nearest retention time from row (i) which has not been analyzed yet in one of the preceding loop (2)-(4). When all the retention times of row (i) are analyzed, data corresponding to the set of "equivalent" retention times (i.e. area, relative area) are copied in the new final matrix. Data corresponding to non-"equivalent" retention times

are simply replaced by empty spaces in this final matrix and reported on the next row (i+1) for the following row analysis.

This iterative process is performed until all the data of the initial matrix have been copied in the final aligned matrix. While the algorithm is partially based on the comparison between retention times of one specific row (i) and retention times of the following row (i+1), data of the last row are simply copied at the end of the final matrix without any alignment. In the same logic, for the analysis of the first line, $t_{H(i-1)}$ is simply equal to 0.

Users are invited to begin with the default value of α = 1. Users can then decrease (i.e. α = 0.5, 0.25, 0.125) or increase (i.e. α = 2, 3, 4) this parameter in order to check if it can give better preliminary alignment results on their own initial dataset. The choice of the α value will depend on the average difference between compounds' retention times.

2. GCALIGNER input files

GCALIGNER needs **input files in a « .txt » tab delimited** format. Here is an example:

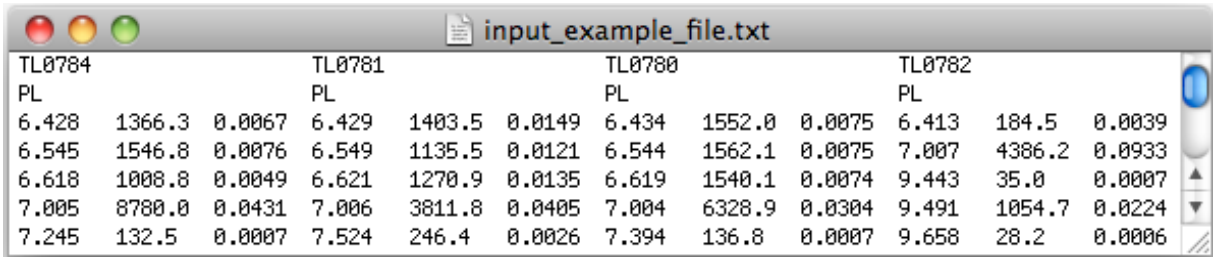

Figure 2: example of an input file.

- The **first column of each sample/injection must always be the RT** column. The program does not read the other columns, which will only be copied in the output file (in this example, the two others columns correspond to the integration areas and the relative integration areas). The number of additional columns to the RT column is not limited but has to be specified before running the program (cfr. (3) Running GCALIGNER).

- **Empty data are not allowed**.

- The two first lines are not read. They can contain any useful information for users but **input files need to present two lines** before the different columns per sample.

3. Running GCALIGNER

A double click on the executable file will start the program. Users then need to specify the location of the input file. Finally, users have to precise two others parameters (figure 1):

- the weight parameter alpha (α) allowing to systematically increase or decrease the acceptance rate of a RT "equivalence" (corresponding to the same compound signal) to another RT. Users are invited to begin with the default value of α = 1 and then to decrease (i.e. α = 0.5, 0.25, 0.125) or increase (i.e. α = 2, 3, 4) this parameter in order to check if it can give better alignment results on their own initial dataset.

- the number of columns per injection/sample (including the compulsory first columns with the retention times).

4. GCALIGNER output files

Output files are in an Excel format (« .xls »). Because the algorithm is based on the comparison between each RT and its following RT, **the last line of each injection is not aligned.** These last lines are simply added at the end of the output file. Here is an output file example:

	A	B	C	D	E	F	G	H	I	J	K	L
1	TL0784			TL0781			TL0780			TL0782		
2	PL			PL			PL			PL		
3	6.428	1366.3	0.0067	6.429	1403.5	0.0149	6.434	1552.0	0.0075	6.413	184.5	0.0039
4	6.545	1546.8	0.0076	6.549	1135.5	0.0121	6.544	1562.1	0.0075			
5	6.618	1008.8	0.0049	6.621	1270.9	0.0135	6.619	1540.1	0.0074			
6	7.005	8780.0	0.0431	7.006	3811.8	0.0405	7.004	6328.9	0.0304	7.007	4386.2	0.0933
7	7.245	132.5	0.0007									
8	7.393	61.5	0.0003				7.394	136.8	0.0007			
9	7.516	615.4	0.0030	7.524	246.4	0.0026	7.515	653.3	0.0031			
10	7.665	169.7	0.0008				7.670	295.8	0.0014			
11	7.954	250.4	0.0012				7.951	167.2	0.0008			
12							7.996	124.6	0.0006			
13	8.083	2165.1	0.0106	8.084	400.9	0.0043	8.083	1772.0	0.0085			

Figure 3: example of an output file.

5. GCALIGNER limitations and authors recommendations

Alignment based on RT clearly shows limits in cases of large gaps between two samples corresponding RT. This can be the case for example if GC analyses are performed on different columns or on the same column but at different degree of column usury. A way to solve this problem is to use Kovats retention indices (Kovats 1958) instead of retention times (see the next point: "Kovats indices and GCKOVATS 1.0").

The aim of GCALIGNER is to perform a first preliminary alignment on large GC dataset and accelerate analysis of multiple GC datasets. We strongly recommend to users to check all the alignment results obtained with this method. The authors decline any responsibility in case of errors/misalignments.

6. Kovats indices and GCKOVATS 1.0

Alignment based on RT with large RT gaps between samples for identical compounds (e.g. samples analyzed at different degrees of columns usury or in different columns) has no meaning. This problem can be avoided by converting all the RT into Kovats retention indices (Kovats 1958). Kovats retention indices (KRI) are normalised system-independent measures (see also http://en.wikipedia.org/wiki/Kovats_retention_index). The KRI of a given compound is computed with the following formula based on the two n-alkanes whose RT surrounding the RT of this compound:

$$KRI = 100\left(n_s + (n_l - n_s)\frac{RT_{sample} - RT_s}{RT_l - RT_s}\right)$$

with:

- n_s: the number of carbon atoms in the smaller n-alkane.
- n_l: the number of carbon atoms in the larger n-alkane.

Here we propose GCKOVATS 1.0, a program that converts RT into KRI. This conversion requires prior calibration points that can be obtained by GC analyses of a standard solution of several alkanes (e.g. from C6 to C26). This calibration will be used as reference to convert RT of samples to KRI. A new calibration should be performed after any changes in methods or parameter (typically each day, depending the user and the expected precision).

GCKOVATS needs the same input file as GCALIGNER and an additional tab-delimited text file displaying the RT of several or all the n-alkanes from the calibration solution (first column) and the number of carbon atoms of each alkane (second column). Figure 4 gives an example of this additional input file. Please note that the two first lines are not read. They can contain any useful information for users but, as for GCALIGNER, this input file needs to present two lines before the n-alkanes RT and number of carbon atoms. **It is also important to underline that the range of samples RT have to be included into the range of n-alkanes RT**.

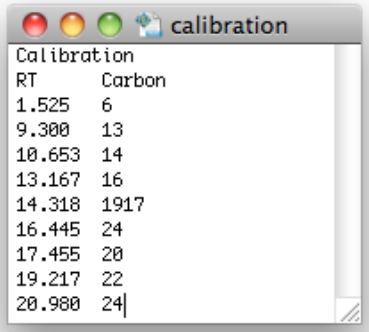

Figure 4: second input file for GCKOVATS.

7. Softwares availability

GCALIGNER and GCKOVATS are both available free from the Evolutionary Biology & Ecology (ULB) lab website *ebe.ulb.ac.be/ebe/Software.html* where this user manual and examples of data files are also available. GCALIGNER and GCAKOVATS are java executable softwares running on any operating system.

8. Acknowledgements

We are grateful to Audrey Coppée, Loïc Dohet, Thibaut Demeulemeester, Laurent Grumiau, Louis Hautier, Patrick Lhomme, Nicolas Meurisse and Thomas Vantorre for help in testing the software and for all their useful comments and advices. 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).

9. How to cite GCALIGNER and GCKOVATS

Dellicour S, lecocq T (2013). GCALIGNER 1.0: An alignment program to compute a multiple sample comparison data matrix from large eco-chemical datasets obtained by GC. *Journal of Separation Science* (*in press*).

10. References

Kovats E (1958). Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone". *Helv. Chim. Acta* **41 (7)**: 1915–32.