Documentation for Matlab code to estimate partitioning properties of non-polar chemicals based on GC×GC retention time data

Version 1.2

J. Samuel Arey and Deedar Nabi, EPFL, 6 August, 2015

Please cite the following article when publishing any results obtained by use of this software:

> Deedar Nabi, Jonas Gros, Petros Dimitriou-Christidis, and J. Samuel Arey, "Mapping environmental partitioning properties of nonpolar complex mixtures by use of GC×GC". *Environmental Science & Technology* 2014, vol 48, p 6814-6826.

The present algorithm is an updated and expanded implementation of the prior work:

> J. Samuel Arey, Robert K. Nelson, Li Xu, and Christopher M. Reddy, "Using comprehensive two-dimensional gas chromatography retention indices to estimate environmental partitioning properties for a complete set of diesel fuel hydrocarbons". *Analytical Chemistry* 2005, vol 77, p 7172-7182.

I. Getting started. What you need to plan before GC×GC analysis.

A. Decide on the GC×GC instrument program.

Ideally, run all of the samples with the same GC×GC instrument program. Use an instrument program that leads to good chromatography for your samples, and bear in mind the instrument program requirements discussed in section 4.6 of Nabi et al., *ES&T* 2014. Chief considerations include the following:

1. Ensure that you use the following stationary phases for the 1st and 2nd dimension columns: 100% methyl polysiloxane stationary phase for column 1 (Rxi-1MS or equivalent); and methyl 50% phenyl polysiloxane stationary phase for column 2 (BPX-50 or equivalent).
2. Avoid using a 1st dimension temperature ramp that exceeds 3 °C min^{-1}.
3. Bear in mind that partitioning property predictions are considered valid only for non-polar chemicals having boiling point ≤402 °C. This includes analytes that elute earlier than pentacosane (*n*-C_{25}) on the GC×GC chromatogram (see sections 4.5 and 4.6 of Nabi et al., *ES&T* 2014).
4. You will need to analyze a series of *n*-alkanes. The model is designed to estimate partitioning properties for the *n*-C_9 to *n*-C_{25} elution range, so ideally some or all of these *n*-alkanes should be included in the analysis. The supplied *n*-alkane members do not need to be a contiguous or regular set.
5. The 2nd dimension of the produced GC×GC chromatogram should have a "zero" retention time value when the GC×GC modulation occurs.

B. Choose a set of instrument calibration analytes.

To estimate partitioning properties with GC×GC, you will need to calibrate the model to the GC×GC instrument program. The calibration is represented by three model-fitted parameters named α_1, α_2, and α_3 (see section 4.3 of Nabi et al., *ES&T* 2014). Once you have fitted the three α parameters for a specific instrument program, the instrument program subsequently can be used to produce partitioning property predictions for as many samples as needed. However if you change any instrument program parameters that would lead to changes in analyte retention times (e.g., temperature, pressure, column length), then you must recalibrate the three α parameters.

To calibrate the three α parameters, you will need to know the GC×GC retention times of 15 or more identified non-polar analytes. These *instrument calibration analytes* can be compounds that are identified in the sample, or they can be separately run standards. What is important is to record the retention times of the instrument calibration analytes.

Guidelines for choosing the instrument calibration analytes are as follows. First, a minimum of 15 analytes is recommended, although more is better.

Second, the instrument calibration analytes must have known Abraham solvation parameters. Many (but not all) non-polar compounds have known Abraham parameters. Additionally, Abraham parameter datasets have undergone "revisions", and recent compilations may lead to the most reliable and consistent sets of values. Ideally the supplied Abraham parameters should come from experimental data and not computed estimates. Refer to Sections 3.1-3.2 and Tables S1 and S4 of Nabi et al., *ES&T* 2014, as well as references 22-38 in that work.

Third, the instrument calibration analytes should form a balanced chemical set and span the two-dimensional region of chromatogram for which you want make property predictions. Ideally, the instrument calibration analytes should be well-distributed throughout the chromatogram. Also, it is more important that the instrument calibration analyte set is balanced rather than large in number. The term "balance" is used to mean that different chemical types are represented proportionately in the set. It would be better if to choose a smaller calibration set that has a reasonably equivalent distribution among several different chemical types, rather than choosing a large set that is strongly biased toward only one or two chemical families.

Finally, the instrument calibration analytes can include some *n*-alkanes – in fact this is a good idea – as long as the instrument calibration set remains balanced.

II. Calibrating and applying the partitioning property estimation model.

Once the instrument calibration analytes have been analyzed using a designated GC×GC instrument program, then you are ready to calibrate and apply the partitioning property estimation model.

A. Organization of the model file directory. Where to find what.

The model code is organized as follows. From the base directory, three folders are present, called `users/`, `model_code/`, and `model_parameters/`.

```
~/.../users/
~/.../model_code/
~/.../model_parameters/
```

These three folder names should not be changed.

The user should only need to operate from within the folder called `users/`. Normally, nothing should be changed or adjusted in the `model_code/` and `model_parameters/` folders.

Within the folder called `users/`, the organization of folders and files is user-defined. The user can define directory paths with the following two model variables:

`input_path`. This variable indicates the directory path location of the input files. The input path variable is set in the file called `main.m`, and it assumes that `main.m` is located in the directory `~/.../users/`. The `input_path` variable also assumes that the indicated directory exists. Example:

```
input_path = 'users/Columbia_input/';
```

`output_path`. This variable indicates the directory path location of the output files. Example:

```
output_path = 'users/Columbia_output/';
```

B. Prepare the model input files.

The model requires three input files. The input files must be placed in the directory designated by `input_path` before you can run the model. The names and contents of the input files are explained below.

The Matlab code will assume that a given unique GC×GC instrument program is represented by a capital letter ranging from A to Z. Whenever you run the model, you will designate the GC×GC instrument program based on this assigned capital letter.

Based on the instrument program name (A, B, C, D, …), you must ensure that the model input files are named as shown below. The contents of the input files may be generated by hand, or you may copy/paste data into the input files directly from an Excel spreadsheet. Example input files names are given below for a program called 'B'.

`retention_times_alkanes_progB.dat`
This file contains three columns of data describing the *n*-alkane series used in the GC×GC program. These are: the carbon number (an integer), retention time 1 (units of minutes), and retention time 2 (units of seconds).

The *n*-alkane series is an important model input: the model is designed to calculate partitioning properties only for solutes that fall within the 1st dimension retention time span of the *n*-alkane series. So no property predictions will be made for solutes that fall outside of this elution range. The input *n*-alkane members do not need to be a contiguous, incremental set of carbon numbers: the algorithm will use whatever members are provided.

`retention_times_calibration_progB.dat`
This file contains 8 columns of data describing the instrument calibration analytes used in the GC×GC program. These are: retention time 1 (minutes), retention time 2 (seconds), and the 6 Abraham parameters of the instrument calibration analytes. The Abraham parameters should be given in the sequence: $A\ B\ S\ E\ V\ L$.

`retention_times_test_progB.dat`
This file contains retention time 1 (minutes) and retention time 2 (seconds) of "test" analytes for which you want to make partitioning property predictions, separately from the calibration analytes. If you do not have any test analytes, then you must nonetheless supply this file name – just leave the file empty.

C. Adjust the parameter settings of the model.

Adjust the parameter settings that appear in the first 30-40 lines of `main.m`. This file can be read and modified from within Matlab or using a generic text editor. This is the only Matlab file that you need to adjust, for normal use of the code.

User-adjusted parameters in `main.m` are as follows: `program_flag`, `modulation_period`, `acquisition_rate`, `group_flag`, `mapped_properties`, `plot_flag`, `output_path`, `input_path`, and `prompt_output`.

Most of the parameters are self-explanatory. However some additional explanation is given below.

The `program_flag` indicates the instrument program, given by a capital letter (A, B, C, D, …), to be decided by you. Example:

```
program_flag = 'A';
```

The `group_flag` parameter is used to adjust the training set that is used to fit the eq 5 coefficients (see section 4.2 of Nabi et al., *ES&T* 2014). The `group_flag` parameter is set to a value of 0 by default, meaning that the entire training set will be used to set the eq 5 coefficients (Table S4 of the research article). If you are making predictions for hydrocarbons only, we have found that a hydrocarbons-only training set gives better regression statistics compared to a fit of the entire nonpolar compound set. With the `group_flag` parameter set to a value of 1, the model will use the hydrocarbons-only training set for the fitting of eq 5 coefficients. Training set regression statistics for all 11 partitioning properties are given as output in the file called `eq5_training_set_fit_statistics.dat`, discussed further below.

The `mapped_properties` parameter is used to decide which properties will be mapped as a contour plot onto the GC×GC chromatogram space. The user enters a vector of integers here with values ranging from 1 to 11. The user can enter as many values as desired. Example:

```
mapped_properties = [1 7];
```

The above assignment would mean that the user requests a contour plot of two properties: the pure liquid vapor pressure and liquid aqueous solubility. The integers assigned to the properties follow the sequence of the properties listed in the file `~/.../model_parameters/properties_list.txt`. The integer assignments are:

1 pure liquid vapor pressure
2 enthalpy of vaporization
3 hexadecane-air partition coefficient
4 dry octanol-air partition coefficient
5 organic carbon-air partition coefficient
6 air-water partition coefficient
7 pure liquid aqueous solubility
8 wet octanol-water partition coefficient
9 organic carbon-water partition coefficient
10 dissolved organic carbon-water partition coefficient
11 bioconcentration factor

D. Run the model.

The Matlab code is straightforward to use.

To start, you may want to place the Matlab console on the right side of your screen. Plots will appear on the left side of the screen.

Make sure that the Matlab working directory points to `~/.../users`. Then at the Matlab prompt, type:

```
>> main
```

The model may require several seconds to run.

III. Interpreting and using the model output.

A. Output appearing in the Matlab console.

The output appearing in the Matlab console depends on the parameter setting you have chosen for the `prompt_output` parameter. The default parameter setting is `'normal'`, which will lead the model to produce the following output information in the Matlab console:

```
Fitted alpha_1 and alpha_2 values (eq 6) are:

    0.2360   -0.1614

Bootstrap uncertainty estimates of alpha_1 and alpha_2 are:
    0.0070    0.1033

The resulting r^2 and RMSE values of eq 6 fitted u_1 values are:
    0.9900    0.1176
```

The above results indicate that the model has determined values of $\alpha_1 = 0.236 \pm 0.007$ and $\alpha_2 = -0.16 \pm 0.10$ using the instrument calibration analyte set, based on a regression fit of eq 6 (see section 4.3 of Nabi et al., *ES&T* 2014).

Eq 6 produces GC×GC-estimated values of the parameter log L_1, which is equivalent to u_1, for each analyte. According to the output shown above, the regression fit has produced a correlation coefficient of $r^2 = 0.990$ and root-mean-squared-error of RMSE = 0.12, between the GC×GC-estimated u_1 values and the reference u_1 values, for the set of calibration analytes. The reference values are obtained using the Abraham solvation model for the stationary phase of the GC×GC 1st dimension column.

The user should inspect the above output results carefully to ensure that the regression fits of α_1 and α_2 are robust. In particular:

1. Good fit statistics for eq 6, ideally $r^2 \geq 0.98$ and RMSE ≤ 0.15.
2. Tolerable uncertainties for the α values, ideally an α_1 uncertainty ≤ 0.01 and an α_2 uncertainty ≤ 0.2.

Subsequently the model will output the following results.

```
Now fitting alpha_3 with a nonlinear optimization of eq 7.

Local minimum found.

Optimization completed because the size of the gradient is less than
the default value of the function tolerance.

<stopping criteria details>

Conducting a bootstrap uncertainty analysis of alpha_3. This may take a minute.
The fitted alpha_3 value is:
    0.8026

The bootstrap uncertainty estimate of alpha_3 is:
    0.1001

The r^2 and RMSE values of eq 7 fitted u_2 values are:
    0.9004    0.0843
```

This means that a value of $\alpha_3 = 0.80 \pm 0.10$ has been assigned, based on a non-linear fit of eq 7. Fit statistics for eq 7 are shown, finding a $r^2 = 0.90$ and RMSE = 0.084 for u_2 values of the calibration set. As with the first two α values, the user should inspect the statistics of the eq 7 fit in order to ensure that α_3 is well-determined. You will want to see:

1. Good fit statistics for eq 7, ideally $r^2 \geq 0.85$ and RMSE ≤ 0.1.
2. An α_3 uncertainty ≤ 0.2. (*Value is tentative. JSA will update.*)

Finally, α_3 has the physical interpretation of (approximately) representing the second dimension hold-up time. This equivalence is not exact, due to the presence of inactive column sections in the instrument. Nonetheless, the α_3 value should have a physically reasonable value. For example, α_3 should not have a negative (<0) or imaginary value. Typically expected values of α_3 would be between 0 and 1 s. If the user observes an unreasonable value for α_3, this is a sign that something else is likely wrong.

B. Names and contents of the output files.

The code will create several output files in the directory `output_path`. The naming convention for output files is:

"output_file_progB.dat"

for output data that you have generated from instrument program 'B'. The model will overwrite existing files, if they have the same names as the target output names of the model.

The output files are:

`predicted_properties_test_progB.dat`

This file contains the GC×GC-predicted partitioning properties for the analyte test set. Each row corresponds to a chemical for which retention time data was provided in the input file `retention_times_test_progB.dat`. Each column corresponds to a partitioning property, according to the sequence shown in section II.C above. This file can be imported directly into Excel or copy/pasted into Excel.

`ASM_predicted_properties_calib_progB.dat`

This file contains the partitioning property predictions given by the Abraham solvation model for the instrument calibration analytes. Each row corresponds to a chemical for which information was provided in the input file `retention_times_calibration_progB.dat`. Each column corresponds to a partitioning property, according to the sequence shown in section II.C above. This file can be imported directly into Excel or copy/pasted into Excel.

`eq5_training_set_fit_statistics.dat`

This file contains the eq 5 fitted λ coefficient values, the uncertainties assigned to each λ coefficient, and eq 5 regression fit statistics for the training set, for each partitioning property. Each row corresponds to a partitioning property, according to the sequence shown in section II.C above. The columns follow the format:

λ_1 λ_2 λ_3 σ_{λ_1} σ_{λ_2} σ_{λ_3} RMSE r^2

where $\lambda_j \pm \sigma_{\lambda_j}$ refers to the 95% confidence interval of λ_j.

Contacts

For any questions, comments, or bug reports, please contact:
J. Samuel Arey: arey@alum.mit.edu
Deedar Nabi: deedarnabi@gmail.com

Acknowledgements

Special thanks to Jonas Gros (EPFL) and Bob Swarthout (WHOI) for agreeing to beta-test the code during the development of this documentation.