ROMANCE

Santiago Codesido Sánchez

Contents

1 Overview and purpose

This document describes the main functionalities of the ROMANCE (RObust Metabolomic Analysis with Normalized CE) software. Its purpose is to efficiently transform *capillary electrophoresis* (CE) data from the measured time scale to an experience-independent electrophoretic mobility scale. We first briefly review the theoretical basis of this transformation, and then document the software.

2 Theory basics

2.1 Migration in CE

In CE, an analyte traverses a capillary for a certain length L, after which it is detected (this is typically at the end of the capillary, but certain detection modes can be applied before). The analyte travels in a *background electrolyte* (BGE) that is moved along the capillary by an external electric field E. The speed induced in the BGE is effectively proportional to E, and we denote the proportionality constant by μ_{BGE}. The separation of analytes is a consequence of the analytes themselves responding differently to E, changing their relative speed w.r.t. the BGE. This is measured by the effective electrophoretic mobility, μ, that is an intrinsic property of each analyte for given BGE and environmental conditions (temperature, etc). The overall flow of the BGE can be increased i.e. by applying an external pressure, which contributes with a constant flow speed v_0. In all, the speed with which the analyte travels along the capillary is

$$v = (\mu + \mu_{BGE})E + v_0. \tag{1}$$

The main problem affecting CE is that μ_{BGE} can change greatly from run to run, even under as identical conditions as practically possible. The contribution of v_0 may be also hard to quantify directly. This does not hinder separation, since these contributions produce a constant shift in speed amongst the analytes.

The situation becomes more complicated if the electric field is not constant. In particular, something that can be found in experimental setups is a field ramp. The value of the electric field increases linearly from 0 until a maximum E_m over a time t_R, where we introduce t_H for convenience. This means that

$$E(t) = \begin{cases} \frac{t}{t_R} E_m & 0 \leq t \leq t_R, \\ E_m & t \geq t_R. \end{cases} \tag{2}$$

By integrating the speed over a time t_M, we find that the analyte travels a distance L,

$$L = \int_0^{t_M} v \, \mathrm{d}t = E_m(t_M - t_R/2)(\mu + \mu_{BGE}) + v_0 t_M. \tag{3}$$

Typically, we want to consider L as the length between the injection point of the capillarity and the detector, so that t_M is the *migration time*. One can solve for μ,

$$\mu = \frac{L - v_0 t_R/2}{E_m} \frac{1}{t_M - t_R/2} - \frac{v_0}{E_m} - \mu_{BGE}. \tag{4}$$

This can be used to translate the observed migration times, t_M, into the effective mobility μ. The trick to get rid of unwanted experimental parameters is to infer them by using markers.

2.2 Markers and mobility

Assuming we have a marker analyte A, with *observed* migration time t_A and *known* mobility μ_A, the following holds

$$\mu_A = \frac{L - v_0 t_R/2}{E_m} \frac{1}{t_A - t_R/2} - \frac{v_0}{E_m} - \mu_{BGE}. \tag{5}$$

Then, for the unknown analyte arriving at time t_M,

$$\mu = \mu_A + \frac{L - v_0 t_R/2}{E_m} \left(\frac{1}{t_M - t_R/2} - \frac{1}{t_M - t_R/2} \right). \tag{6}$$

This is enough if $t_R = 0$ (i.e. no ramp), or if v_0 and t_R are known. In a simplistic scenario, one could estimate v_0 by running a measurement under no electric field whatsoever, and seeing how long it takes for an injection to reach the detector. However, this would assume no interaction between the electric field and the hydrodynamic properties of the BGE. If $v_0 t_R$ can be neglected w.r.t. to the L, the length between injection and detection, one could get away with setting v_0.

A last approach, which gets rid of most external parameters, is to use a second marker,

$$\mu = \frac{(t_A - t_M)(t_B - t_R/2)\mu_B - (t_B - t_M)(t_A - t_R/2)\mu_A}{(t_A - t_B)(t_M - t_R/2)}. \tag{7}$$

2.3 Intensity normalization

Transforming the measure migration time into mobility removes the problem of BGE variability with regards to *identification*. It can still influence (semi-) *quantification*. We consider two main factors.

For the first factor, a correction is needed if the detection is sensitive to the concentration at the point of measurement/exit of the capillary. This is the case for UV detectors, and also for MS detectors whose ionization source is in a saturated regime, often called *concentration mode* [1]. In this scenario, analytes moving slower will have broader peaks, while the height remains only dependant on their spatial concentration. In particular, suppose the same analyte is observed in two runs, the first with half the speed of the second. The integrated area of the first peak will be twice as large as that of the second. Thankfully, the speed at the point of measurement can be easily related to observed migration time,

$$v_{\text{exit}} = \frac{L - t_R/2 v_0}{t_M - t_R/2}. \tag{8}$$

Changing the width of the peak would require effecting local shape changes to the electropherogram, which not only opens up a mathematical Pandora's box, but is probably a good way of ensuring one never observes the smaller peaks more susceptible to distortion. On the other hand, the total area can be corrected by changing the height of the peak, in particular, multiplying it by the exit speed, to compensate for the broadening of the slower ones.

[1] This is opposed to the unsaturated regime, or *mass mode*, where everything that comes out of the capillary is ionized.

Notice that the numerator of the exit speed is analyte-independent. Intensities are proportional to concentration (hopefully) up to an unknown response factor. Therefore, we can choose it arbitrarily as long as we wish to compare compounds within the same run. A sensible choice that maintains the order of magnitude of observed intensities is to correct intensities by

$$I(t) \mapsto \frac{\overline{t_A}}{t - t_R/2} \times I(t). \tag{9}$$

The "reference" time $\overline{t_A}$ can be taken to be the average of marker times of certain ensemble of runs, but its only purpose is to offer a judicious normalization.

To compare different runs, one would need to make sure that L is the same, and either $t_R = 0$ or v_0 and t_R also remain constant. If not, the full expression for the exit speed can be expressed in terms of the migration times of two markers, by using

$$L - t_R/v_0 = E_m \cdot (t_A - t_R/2) \cdot (t_B - t_R/2) \cdot \frac{\mu_A - \mu_B}{t_B - t_A}. \tag{10}$$

This can be then used as an inter-run normalization factor. In any case, normalization between runs typically includes many other factors (not least sample preparation) and is usually left to some form of post-processing such as normalization by a marker peak or PQN.

For the second factor, the detector may measure either a *snapshot* at the time of the scan (as UV detectors) or measure a number of *counts* between consecutive scans (as most MS detectors). In the latter case (counts) the transformation to mobility requires no particular correction: $I(t_i)$ represents the number of counts between two scans at times t_{i-1} and t_i, which should also be the number of counts between the corresponding[2] mobilities μ_i and μ_{i-1}.

But in the first case, the number of counts between times t_{i-1} and t_i should be interpreted as

$$I(t_i) \times (t_i - t_{i-1}) = I(t_i)\, \Delta t_i. \tag{11}$$

Because the transformation from times to mobilities is not linear, the distance between two time points is not proportional to the distance between the corresponding mobilities. This makes peaks heavily distorted w.r.t. their shape in the electropherogram. The intensities must be corrected by the ratio between distances in the time and mobility domain,

$$\frac{\Delta t_i}{\Delta \mu_i} \simeq \left.\frac{\mathrm{d}t}{\mathrm{d}\mu}\right|_{t=t_i}. \tag{12}$$

The approximation to the derivative holds when the distances between times are small relative to the times themselves, which should be the case if the scans are to have any use as an electropherogram. Then,

$$\frac{\mathrm{d}t}{\mathrm{d}\mu} = -\frac{E_m}{L - t_R/2v_0} \left(t_M - t_R/2\right)^2. \tag{13}$$

The minus sign just remembers that the transformation reverses the order, and again there is an irrelevant overall factor. The transformation that should be applied to correct for this effect is

$$I(t) \mapsto \left(\frac{t - t_R/2}{\overline{t_A}}\right)^2 \times I(t). \tag{14}$$

If both factors should be taken into account, as is the case with UV detectors, then the transformation should be their composition,

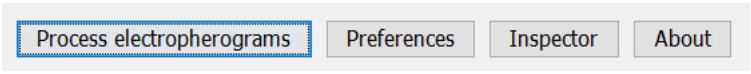
$$I(t) \mapsto \frac{t - t_R/2}{\overline{t_A}} \times I(t). \tag{15}$$

[2]One should take care to remember that the mobility transformation *reverses* the order of times (faster compounds arriver earlier).

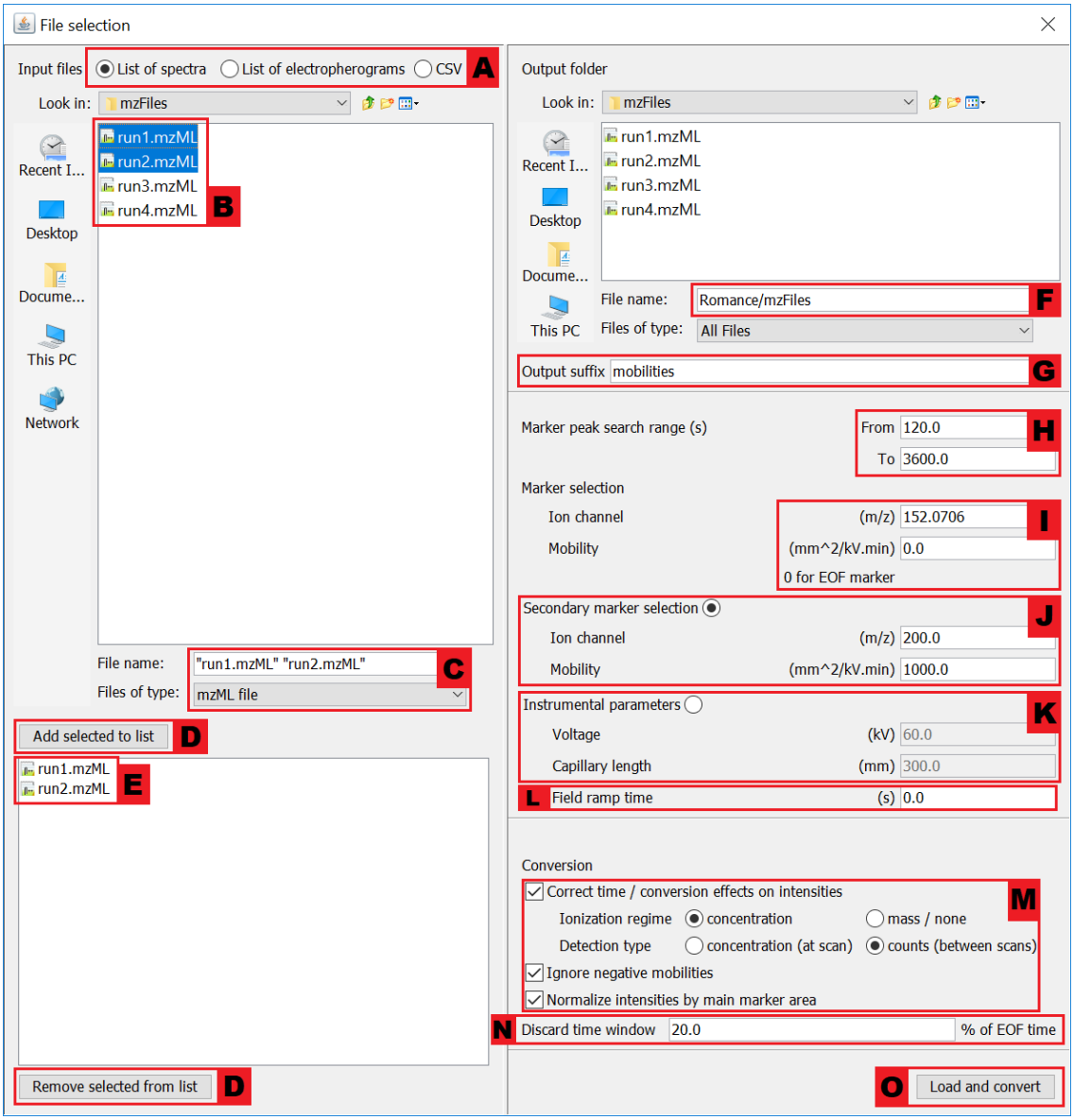
3 User interface

3.1 Electropherogram processing

The main functionality of ROMANCE can be accessed directly with the "Process electropherograms" button in the main bar,



which opens up the following dialog asking for the processing parameters.



The UI elements are:

- **A**: Type of input files. Output files are written in the same format.
  - List of spectra: an mzML file with a list of scans with a time tag, containing a full mass spectrum each (usually from untargeted, TOF MS).

  - List of electropherograms: an mzML file with a list of mass channels, containing a full electropherogram each (usually from targeted, QQQ MS).
  - CSV: a comma-delimited file. Electropherograms are written in the vertical direction, with column $2k-1$ containing the times of run $k = 1, 2, \ldots$; and column $2k$ containing the measures intensities/counts. The header of column $2k-1$ is taken as the (mass or otherwise) channel. Useful for UV data.

- **B**: File explorer. Used to navigate the filesystem and select the input files.
- **C**: File type selector (all, mzML or CSV).
- **D**: Add and remove files from **B** to **E**.
- **E**: List of input files. Contains all the files that will be parsed when clicking **O**.
- **F**: Output folder. Where the converted files will be stored.
- **G**: Suffix for output files. The name of each output file is the name of the corresponding input, plus this suffix, and its extension. Critical to avoid overriding input files when output is sent to the same directory.
- **H**: Marker search range. To determine the position of each marker, ROMANCE has to perform peak integration. This places bounds on the time interval. It is mainly useful if more than one peak is expected in the channel of the marker.
- **I**: Main marker parameters. The marker channel is logically where the marker peak will be integrated. It must be an m/z channel for mass spectrometry, *not* the neutral mass of the analyte, since at this point of the processing of the data nothing is known about adducts and the like. If the file was read as a CSV, this can be any of the headers identifying the channels (i.e. a wavelength). The mobility is the one associated to the main marker. If the electroosmotic flow (EOF) is taken as the marker, the mobility should be set to zero. Otherwise, it should be determined independently. The units are suggested by ROMANCE, but any scale can be used - the output mobilograms will reflect it accordingly.
- **J** and **K**: Other parameters required for processing. Two options:
  - **J**: Use another marker. As seen in the theory section (7), the rest of the instrumental parameters can be determined from a secondary marker. The same criteria as for the main marker apply to these fields.
  - **K**: Provide explicitly the instrumental parameters, as in (6).

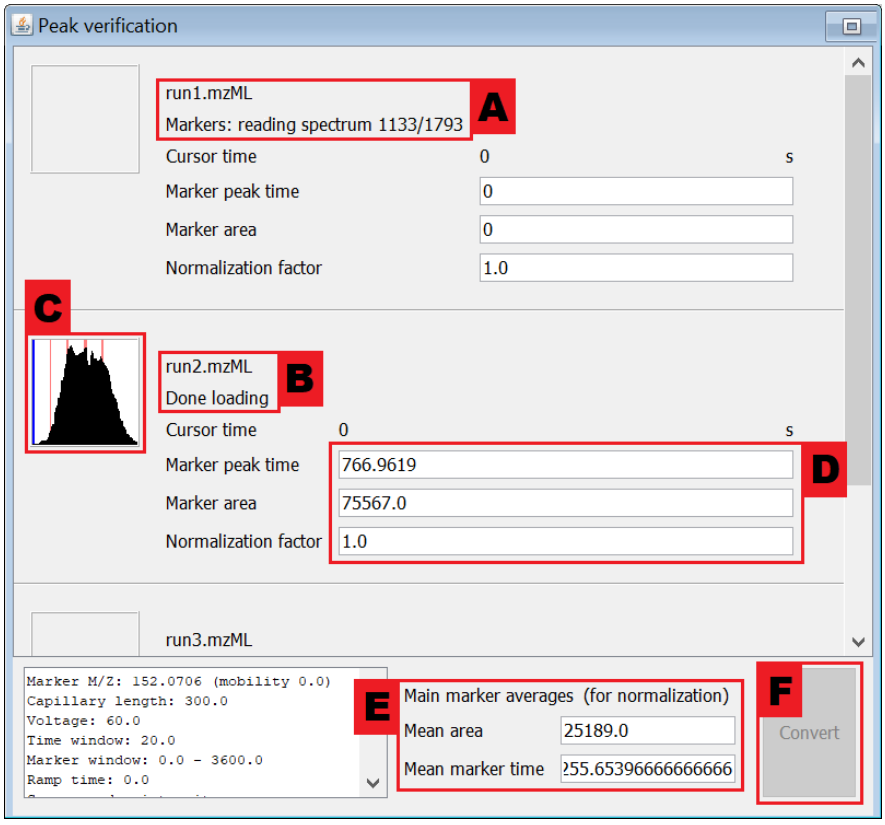
  Again, in either case, ROMANCE has no way of knowing the actual units provided. However, this step requires extra care. The units *must* be consistent with those chosen for the main marker, or the transformation will simply be wrong.
- **L**: Field ramp time. This instrumental parameter must always be provided explicitly. Units must be consistent with those present in the input files.
- **M**: Conversion options.
  - Intensity correction. "Ionization regime" corresponds to correction (9), and "Detection type" to (14). In case of doubt, MS with μL/min flows in the CE typically corresponds to [*concentration, counts*], MS with nL/min flows in the CE to [*mass, counts*], and UV detection to [*concentration, concentration*].
  - Ignore negative mobilities. Normally, separations are run in separate negative and positive modes to efficiently extract analytes with opposite responses to the electric field. In a single run, analytes that move against the BGE take longer to arrive at detection, or may directly not arrive at all. If this option is selected, only analytes with positive mobilities (w.r.t. to the response of the BGE) are written in the output. Otherwise, *two output files* are produced per input file, one for negative and one for positive mobilities, written in absolute value. This is to ensure that mobilograms lie always on the positive real axis, to avoid frictions with later peak-picking software platforms.

  - Normalize intensities by main marker area. If the main marker is i.e. some compound that has been added a posteriori to samples, in known amount, it can be used to perform inter-run normalization of the intensities. Each run is multiplied by the quotient between the average of the marker areas between all runs, and its particular main marker area.

- **N**: Discard time window. If scans are started as soon as the BGE starts moving, the very first are surely noise. Mathematically, zero times correspond to infinite mobilities. Such singularity also induces the heaviest peak shape deformations close to it. For these reasons, it is *necessary* to cut the start of the electropherogram.

- **O**: When everything is set, click to run.

Once all the input parameters are set, ROMANCE opens a new internal window showing the progress of the importing process:



Each input file has its own *verification panel*, that includes a preview of the peak and the automatically detected parameters.

- **A**: File name and status (reading). ROMANCE loads files in parallel, limited by the number of cores of the machine (for XML indexing and parsing) and the I/O bandwidth.

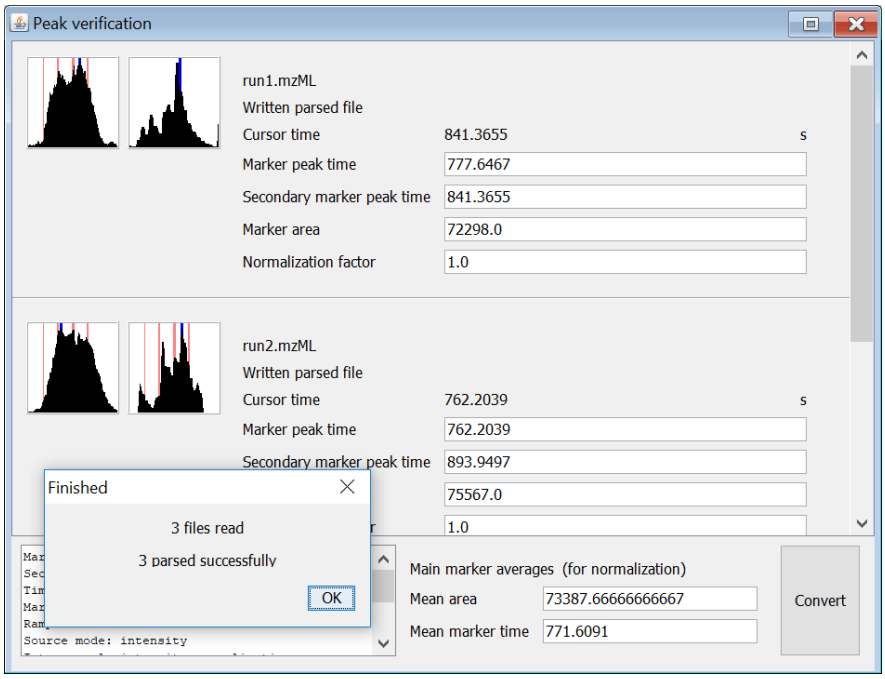
- **B**: File name and status (loading). The status will also include any I/O, parsing or peak integration errors. If any file fails to be properly loaded, it will be discarded for the current conversion batch.

- **C**: Peak preview thumbnail. Shows the integration of the peak in the selected channel. If two markers are employed, the preview is divided in two, with the left corresponding to the main marker, and the right to the secondary. A double right click in the preview will set the marker peak time (in **D**) to the cursor's (blue line) location. One can also zoom in/out by dragging the cursor, and zoom completely out by left-clicking. The red lines show the center and $\pm\sigma$ of the automatically detected peak.

The peak position and area are determined by starting at a the highest point in the histogram, and updating its mean and standard deviation as the scanning region is increased. The algorithm stops when a sufficient region around the peak has been scanned, by using Chebyshev's rule. This provides an extremely robust peak integration mechanism for *isolated peaks*, but it is not suited for very large amounts of, or overlapping, peaks. Markers should normally be a clear presence in their channel, but if ideal conditions are not met, the user should be very careful to left-click on the thumbnail and explore the general aspect of the electropherogram.

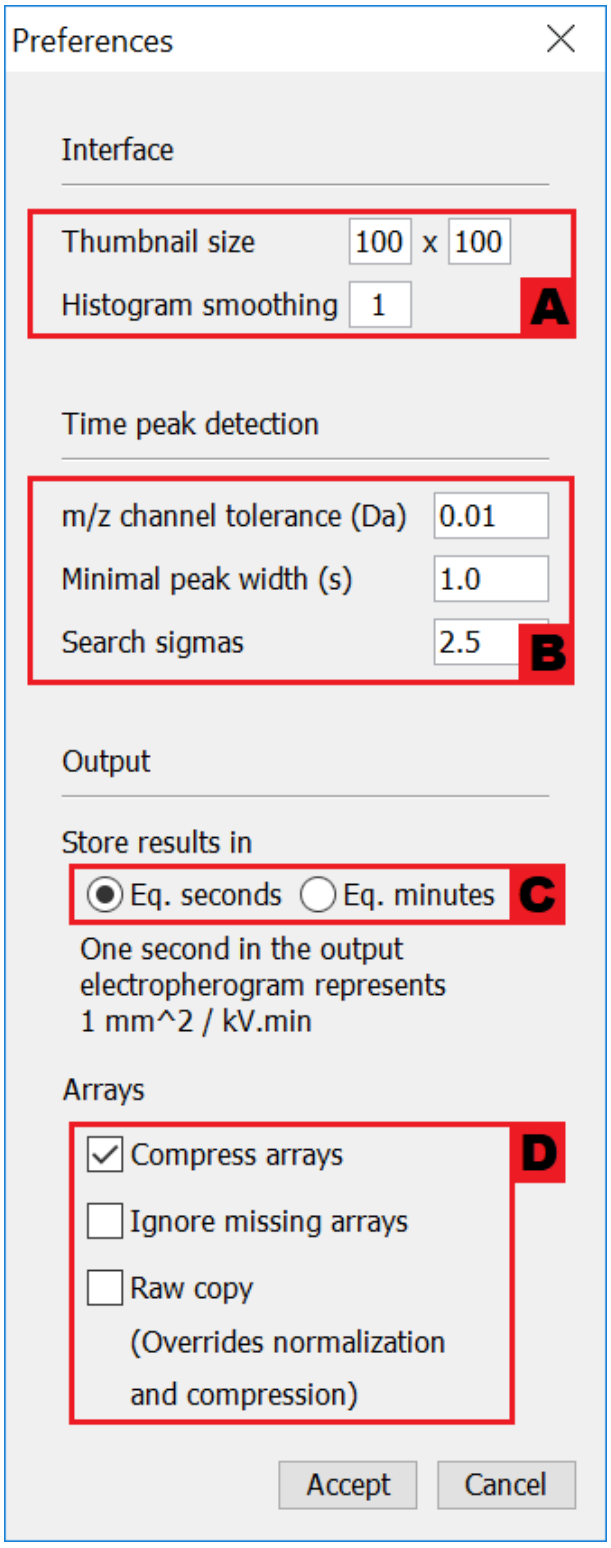
- **D**: The computed peak time for the marker in the given file. If inter-run normalization was selected, one can also specify the peak area and an overall normalization factor. The latter is useful for i.e. diluted samples.
- **E**: Inter-run values. The computed means that are used for correction and normalization (see 9). Can be changed as needed.
- **F**: Convert. Will only become available once all input files have been read/failed reading. Once the conversion is launched, changing the parameters in the preview panels will no longer have any effect.

When all output files are written ROMANCE will bring up a notification indicating the number of successful conversions. The following screenshot shows this, together with the two-marker preview windows, and the use of the cursor to select marker times differing from those suggested by the ROMANCE peak integrator.



3.2 Preferences

The "Preferences" button opens a dialog for the configuration of general ROMANCE options.



- **A**: Graphic options. Thumbnail size changes the display size of the peak thumbnail in the preview panel before processing. Histogram smoothing is the number of pixels of smoothing applied to the histograms.
- **B**: The tolerance is used to determine which channel to select as the marker electropherogram in electropherogram and CSV files, and to pick spectrum points in spectra files. The other two parameters apply to the peak integrator. "Minimal peak width" is used as the minimum scanning bandwidth for peak integration, although the actual integrated peak width can be smaller. "Search sigmas" is the number of sigmas at which scanning stops following Chebyshev's rule for peak integration.
- **C**: Since most peak-picking software expect time, and not mobility, to be the dependent coordinate of electropherograms, ROMANCE exports the results in "equivalent seconds", corresponding to the indicated mobility units. Depending on the downstream peak-picking software, it may be more convenient to store results in the scale of minutes.
- **D**: Options for mzML arrays. "Compress arrays" (zlib) reduces greatly the already rather bloated size of mzML files. However, downstream software may not be able to handle this too well, so arrays can be stored in their raw form. "Ignore missing arrays" skips any scan (for list of spectra) or channel (for list of electropherograms) that does not contain any array. Normally this should indicate some sort of malformation in the data, and the default behaviour is to raise an exception and ignore the whole affected file. "Raw copy" is only used for files with lists of spectra. Instead of generating a new mzML file, this options outputs the input file *as-is*, with only the time tags in the corresponding XML leaves modified to contain the mobility instead. Consequently, no correction of the intensity array is possible. This can be useful is some downstream software is expecting some very specific mzML metadata present in the input file and that ROMANCE ignores.

3.3 Inspector

ROMANCE provides an inspector for mzML files containing lists of spectra. Each individual spectrum can be loaded and shown as a histogram, together with the internal XML code containing all the mzML metadata. The electropherogram for a given channel can also be generated in the corresponding tab.