enviMass v2.2

An environmental monitoring workflow for LC-HRMS data

Martin Loos, Matthias Ruff, Heinz Singer
2015

User manual

Contact:
Martin Loos, Eawag Dübendorf, P.O. Box 611, Switzerland
Telephone +41 (0)58 765 5118
Fax +41 (0)58 765 5311
martin.loos@eawag.ch

Contents

Abstract

enviMass version 2.0 is an automated workflow for the upload, preprocessing and analysis of measurements from high resolution mass spectrometry (HRMS) coupled to liquid chromatography (LC) in the environmental monitoring context. Specifically, enviMass 2.0 delineates temporal intensity trends from raw measurement data, providing a convenient user interface (UI) while building on the strength of the R environment and its associated packages. The workflow features:

- Data conversion & centroidization (utilizes ProteoWizard)
- Data import (.mzXML)
- EIC extraction
- Peak picking
- Mass recalibration
- Intensity normalization
- Quality control
- Profile extraction
- Trend detection
- Blind subtraction

While running from the UI, enviMass manages your project and all associated files in the background, providing interactive access to all project data – ranging from extracted ion chromatograms within selected samples to the full time profiles of temporal intensity changes.

Input data & system requirements

Input data: the workflow has been built for high resolution LC-ESI-HRMS data and tested on Thermo Orbitrap .raw measurements. enviMass requires centroided .mzXML formats, but allows to upload many other vendor formats if ProteoWizard is installed on the system (see Installation section). The enviMass peak picking algorithm fails for low resolution data or non-centroided input formats. Note that ProteoWizard may fail in centroiding certain formats.
Hardware requirements: as file sizes and data amounts to be processed are large, the user should assure that the system provides enough working memory. Thus, enviMass requires an absolute minimum of 4GB RAM but may often rely on more memory for certain formats and experiment types. A working memory of 16GB RAM is hence recommended. Hard disc ROM should be large enough to handle the size of series of LC-HRMS measurements.
OS requirements: if the R environment and a web browser other than Microsoft Internet Explorer (e.g., Google Chrome, Firefox) can be installed on our system, then your OS is fine. In other words, enviMass can be run on most Windows OS, MacOS and LINUX, both for 32bit and 64bit architectures. For uploading formats other than .mzML via ProteoWizard, the system requirements of the latter may apply, too. enviMass has so far been tested on Windows XP and Windows 7 only. As offline installation is restricted to

test users, your computer must have access to the internet. After installation, no further internet access is mandatory.

Installation

(1) Ensure you have administrator rights on your computer

(2) Install the latest R version from http://www.r-project.org/

(3) – offline R-package installation. Unzip the attached folders containing all necessary packages and their dependencies into the R library folder; you will often find this folder under the path *C:\Program Files\R\R-3.0.2\library.* Overwrite any existing R packages.

(3) – online R-package installation (not yet available): Install the R package enviMass: open your R browser and navigate to *R packages* and therein to *Install package from a local .zip file*. During installation, dependencies to other packages (e.g., enviPat, shiny, mgcv, mzR,…) are checked and more packages installed – ensure you are online!

(4) Install the latest release of any of the two web browsers Google Chrome or Firefox and make it your default browser.

(5) If you intend to read data other than .mzML or .mzXML formats, install ProteoWizard`s MSConvert from http://proteowizard.sourceforge.net/downloads.shtml ProteoWizard also requires your list separator to be , instead of ; . For this, navigate to *Control Panel -> Region and Language Settings -> Additional Settings -> List Seperator*. To make enviMass communicate with MSConvert, you must specify the path to a file called msconvert in the enviMass UI for each(!) of your projects. To do so, navigate to *Settings* and to *General* within an opened project. Type in the required path (e.g., *C:\Program Files\ProteoWizard\ProteoWizard 3.0.5140\msconvert* – with no file extension such as *\msconvert.exe*) and click *the Apply settings to project* button.

Getting started

To start the enviMass UI upon installation, open R and first type *library(enviMass)* and then *webMass()* into the console. Wait for the enviMass start page to open in your web browser. The R console keeps running meanwhile and must not be closed, working on your UI procedures and printing progress messages. However, as this is of no concern to the UI user other than to intercept error messages, R can be minimized.

On the start page, you have two options. You can either start a new project or open an existing one. In both cases, a project folder containing several subfolders and a logfile.emp is generated or accessed, respectively. For project consistency, these folders must not be manipulated other than via the enviMass UI. To start a new enviMass project, choose a project name, give a valid (existing) directory to contain your new project folder

(named after your project name) and press *Start*. Existing projects cannot be overwritten. To open an existing enviMass project, insert (or copypaste from, e.g., Windows Explorer) the path of your enviMass project folder (contains the logfile.emp) and press Open. For example, if your project is called *new_project* and its folder can be found under *C:\users\enviMass*, you must insert *C:\users\enviMass\new_project*. Hence, you proceed to the UI of an individual project. You can always navigate back to the enviMass start page by using the back button. To leave enviMass, hit exit – the browser UI disables and the R console is ready for new command inputs. Project results are stored and can be retrieved when a project is reopened. If you want to use ProteoWizard`s MSconvert, specifiy a proper path within the project settings (cp. installation section). At the time being, separate projects need to be established for positive and negative ionization mode series of measurements.

Within an individual enviMass project, you have access to several tabsets which convey different functionalities within a single project: uploading data sets and compounds, controlling parameter settings and workflow options and – after calculation – the different results on peak picking, data preprocessing, profiles and trend detection. In general, a project requires at least one measurement to be loaded to run any calculation. Calculations are initiated by pressing the *Calculate* button on the sidebar. The enviMass UI is not responsive during calculation and prints out the current state it is in on the sidebar table. Depending on the number of measurements loaded and the workflow steps chosen, enviMass calculations can range in anything between minutes to hours. enviMass always adjusts calculations depending on the current state of the project, avoiding redundant calculations while keeping track of changes in data upload, workflow modifications or parameter settings. Further information on the various enviMass functionalities can be found in the tabset sections below. Error messages and project states are printed in the textbox of the sidebar.

Within each project and all its associated tabsets, (a) samples, (b) compounds, (c) peaks and (d) profiles are all administered by unique numeric IDs. The latter are crucial to retrieve results in the UI and to store and access data in the background.

enviMass crash

A crash of enviMass results in the browser UI turning grey and unresponsive while relevant error messages may be printed in the R console for debugging. To return to the enviMass UI, navigate to the R console and press the *Esc* button on your keyboard. Then proceed by typing *webMass()* into your console.

tabsets

Files

The *Files* tabset allows you upload new LC-HRMS measurement files to an enviMass project. Note the Input data section to find out which input formats are eligible. The set of files can be continuously extended by new files, which are finally included into the

calculation outcomes (i.e., the *Results* tabset) after having re-run all calculations by pressing the *Calculate* button. The *Add LC-HRMS measurement* box specifies the metadata of each measurement, e.g., its ID, name, sampling date, sampling time, etc. Using the upload button, an LC-HRMS measurement (.mzML, .mzXML or any format supported by PW MSConvert, e.g. .raw) can be selected and loaded into your project. Note that, for upload *and* conversion from formats other than .mzML or .mzXML via PW MSConvert, the time enviMass needs to list the concerned measurement in the measurement table is somewhat longer than mirrored by the upload bar. To delete a measurement from the project set, use its ID given in the first column of the files table. Although you can choose the ionization mode of a measurement, this information is ignored at the moment – separate projects have to be set up for negative and positive ionization modes. If a measurement file is corrupted and cannot be loaded, it will not appear in the table of files. Instead, an error message is printed in the side bar.

Compounds

Compound names of internal standards (IS) and targets must not contain tabs, umlaute, "," or ′ (the latter is sometimes used to mark structural positions in molecules) . Usage of any of these will cause enviMass to crash. Similarly, tags 1 to 3 must under no circumstances contain underscores ("_").

Include target and internal standard compounds into a project by using the input boxes in their respective tabsets. In each project directory, the folder *dataframes* contains .txt-files of these two classes of compounds (if present), namely *IS.txt* and *targets.txt*. You can use the *Import compound list* box to import these files from the *dataframes* folder of other projects.

You can also use spreadsheet-based software such as Excel or OpenOffice to define compound lists. To do so, copy *IS.txt* or *targets.txt* to another directory and open them in Excel as tab-delimited text-files and save them as such. Thereupon, use the *Import compound list* box to import these files. Do not open and save the *IS.txt* or *targets.txt* files in the *dataframes* folder directly; you must first copy them to another directory and import them from there. In addition, strictly adhere to the tab-separated formatting of the *IS.txt* and *target.txt* files as well as the individual formats of columns (especially date formatting). Best start from the standard files of *IS.txt* and *target.txt* that are generated as default when a new project is set up. Do not use umlaute in compound names. Do not confuse tab-delimited with space-delimited, e.g. when using a text editor for manipulating the files.

Data for ISs and targets can also be included into a new project by having an IS and/or target *dataframe* available in the R workspace before starting *webMass()* and a new project from within the UI. For exemplary dataframes of ISs and targets, have a look at the project defaults available in R using *data(IS)* and *data(targets)*.

Workflow options

This tabset lists the enviMass calculation steps, some of which can be dis – or enabled. To do so, use the radio buttons and press the *Apply settings to project button*. Workflow settings will be stored in the project and are still available when the project is being reopened. Note however that certain workflow steps rely on other down- or upstream

steps. For instance, you may disable the Profile extraction step while still selecting the Trend detection & blind subtraction step. Although the former is marked as disabled, it will be included in the workflow calculation since the latter step depends on it. Thus, enviMass manages the minimum workflow steps required based on the user selections in this *Workflow options* tabset.

Settings

Parameter settings for calculations and some general project settings can be managed in this tabset. Press the *Apply settings to project button* to make them permanent; thus, settings will still be available when the project is being reopened. Note that the consequences of changing parameter settings on the calculation speed depend on the stance of the concerned calculation in the overall workflow. For instance, changing the peak picking settings requires a time consuming recalculation of all peak lists from the raw files as all downstream workflow steps rely on these lists. In contrast, changing the time lags for the Trends only concerns a later step in the workflow, with little changes in overall calculation speed.

The user must ensure that all parameter settings are correct. For example, ensure that the time lags for trend detection do not exceed your time span of files. Not all settings can be checked a priori at the beginning of each enviMass calculation and may then lead to an abortion of a calculation at a later stage. In the best case, an error message is appears in the sidebar. In the worst case, enviMass crashes – check the crash section.

Results

After having run a calculation, the different outcomes are listed in various sub-tabsets in the *Results* tabset. Results excluded by settings in the *Workflow options* tabset will not be available. In enviMass, results are often itemized for individual peaks, files, compounds or profiles and must then be accessed by the relevant IDs. The latter relate for files and compounds to the IDs listed in the tables shown in the *Files* and *Compounds* tabsets, respectively.

Quality control

This tabset lists results on the quality control (QC) of individual samples (first two plots) and the intensity normalization (third plot). QC is based on the assumption that, although individual peaks may appear and disappear over the temporal sample sequence, the overall distribution of peak intensities changes little. Blank files are not checked in this quality control step and are thus not excluded from further processing. Topmost, the quantiles of the $\log_{10}$-transformed peak intensities within each sample are plotted. Based on these distributions, the second plot outlines samples that strongly deviate from the mean log-intensities of all other samples at quantiles. This deviation is derived both for the quantile with the maximum deviation (x-axis) and for the median over all quantiles (y-axis). If any of the two deviations exceed a threshold (dashed lines), the concerned sample will be regarded as an outlier and hence excluded from any further enviMass calculations. The numbers shown refer to the sample ID as specified in the Files tab. Coloring in the quantile plot refers to text coloring in the deviation plot.

The bottom plot compares the raw (i.e., un-normalized) $\log_{10}$-transformed peak intensity distributions across all samples and blanks, using boxplots. Peak intensities are then

normalized and standardized to a median intensity across all samples (red dashed line). Numbering of the boxplots refers to the sequence of files as listed in the Files tab.

EIC & Peaks

This tab visualizes the raw data (centroided measurements) that makes up a peak (red) and the EIC (grey) within which it is picked. The EIC constituting a certain peak in a certain sample (or blank) can be accessed via the IDs of both a sample and a peak therein. Sample IDs refer to the IDs provided in the Files tab; strictly speaking, they can represent both sample or blank files. The two topmost panels depict chromatogram intensities over the full elution time range and the elution time window containing intensities above zero, respectively. The bottom panel depicts the mass deviation of the EIC, scaled to the elution time window of the middle panel.

Processing

The Processing tabset shows the m/z recalibration outcomes over the full mass range for each sample or blank, as selectable by their IDs provided in the Files tab (top panel, dots: mass deviation of internal standards and/or target compounds, red line: modelled recalibration function used for m/z recalibration). In addition, a second plot shows a histogram of the intensities of all the centroided files in a sample (white) versus those contained in the picked peaks only (red) for the selected sample. Typically, the latter capture the upper range of intensities above the mode of the former, while less peaks are picked for low-intensity files (Figure 1). Finally, the bottom plot depicts the m/z and RT location of all peaks picked in the selected sample.

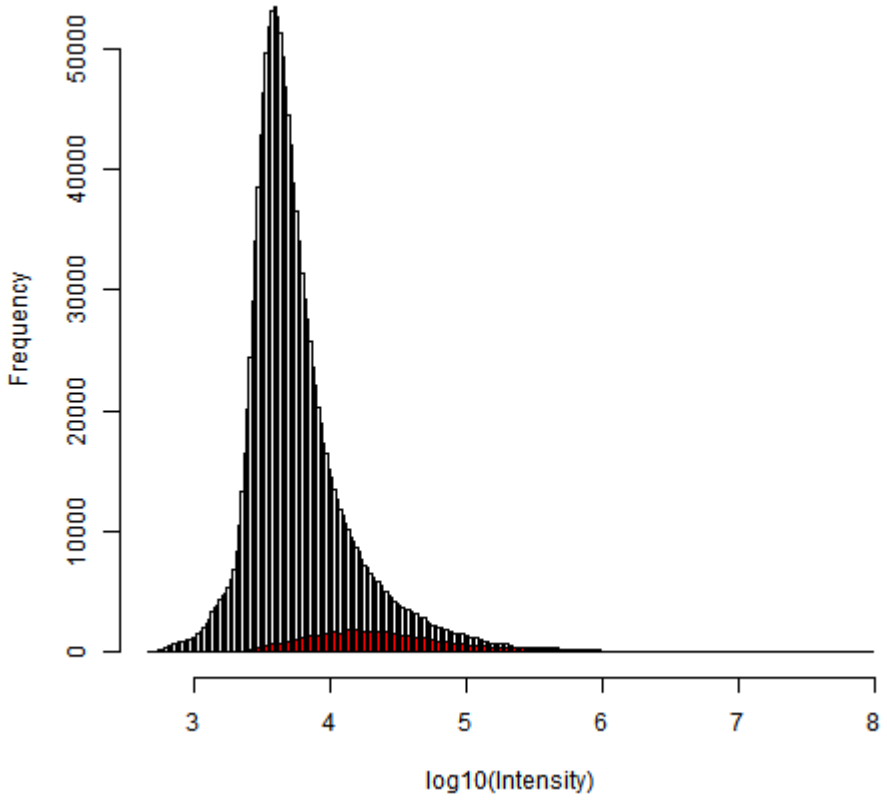

Figure 1: Intensity distribution of measurements in a sample. White: all measurements. Red: measurements related to a peak.

Profiles

Profiles represent the intensity variation of a set of peaks over time. Ideally, these peaks are all caused by the same isotopologue of the same adduct of the same chemical compound that can at least be once detected in the sequence of samples. For each sample time point not more than one peak exists in a profile. However, as required for the blank subtraction, one peak from the sample and several peaks from blank files my coincide at one time point. Often less blank than sample files are available, e.g. if samples and blanks are acquired every day versus only once per week, respectively. In this case, blank peak intensities are interpolated for the missing sample time points to provide an estimate for the background intensities.

Each profile is analyzed for the occurrence of unusual intensity trends to detect spills or other temporal patterns of concern. On the one hand, this requires the extraction of periods with intensities increasing in time. On the other hand, an extracted trend is compared to all other trends observable in a profile to delineate unprecedented trends. Third, trends must be corrected for blank intensities.

Trend detection is accomplished by calculating the mean intensity at each observed time point in a profile over a time-lag window of n previous peak intensities. If this mean intensity is higher than the observed intensity, it is set to the observed intensity (grey line in Figure 2). Consequently, this mean intensity is either lower than the observed one for an increasing trend or equal to it for a decreasing trend from some point of time onwards. A trend period is subsequently extracted for any consecutive time points for which the mean lag intensity is smaller than the observed one. Within such a consecutive trend period, the trend intensity and apex time point is set to the maximum peak intensity and its date, respectively. However, fluctuating intensities may lead to the detection of a multitude of low-intensity trends in each profile, some of which may not even be distinguishable from the blank intensity. Therefore, a filtering step is included: a trend is only retained if it exceeds (a) the mean trend intensity by x times the standard trend intensity deviation of all other unfiltered trends and (b) the (interpolated) blank intensity by y times. Overall, two trend intensities have to be distinguished. The global trend intensity is that of the maximum trend in a profile. The current trend intensity is the intensity of the latest peak that falls into a trend. The latter may coincide with the global trend intensity or may be absent if no trend is detected for the latest time point. In other words, not all profiles have a trend intensity at the most recent time point. Note that time lags of different length n must be utilized to cover trends occurring at different temporal scales.

Finally, if a current trend intensity is very high as compared to the full distribution of global trend intensities over all profiles it deserves closer attention on whether a spill alert may be issued (red dots in Figure 3). The set of global trend intensities evolves over time as new files are added; current trend intensities of concern may turn into global trend intensities.

There are three sub-tabs available in the Profiles tab. The first, Summary, lists all available profiles and their intensity distributions. Several options are available to filter this list to extract profiles that, e.g., fall into certain ranges of m/z or RT, range above

blind intensities for at least one trend, do not appear in the blind files at all or show a current trend intensity. The latter is the default filter after an enviMass calculation: the profile list only contains profiles of current concern, namely those with a current trend intensity, ranked in order of decreasing intensity. Profiles can also be sorted by the maximum intensity a profile has for all the peaks it contains (maximum intensity selected in the Sort profile list by: selection). While the second tab (Newest trends) compares global vs. current intensities as exemplified in Figure 3, the last tab (Single Profile) gives access to individual profiles, either based on the selection and ordering done in the above mentioned Summary tab (cp. Entry # in profile list) or by the ID of a certain profile.

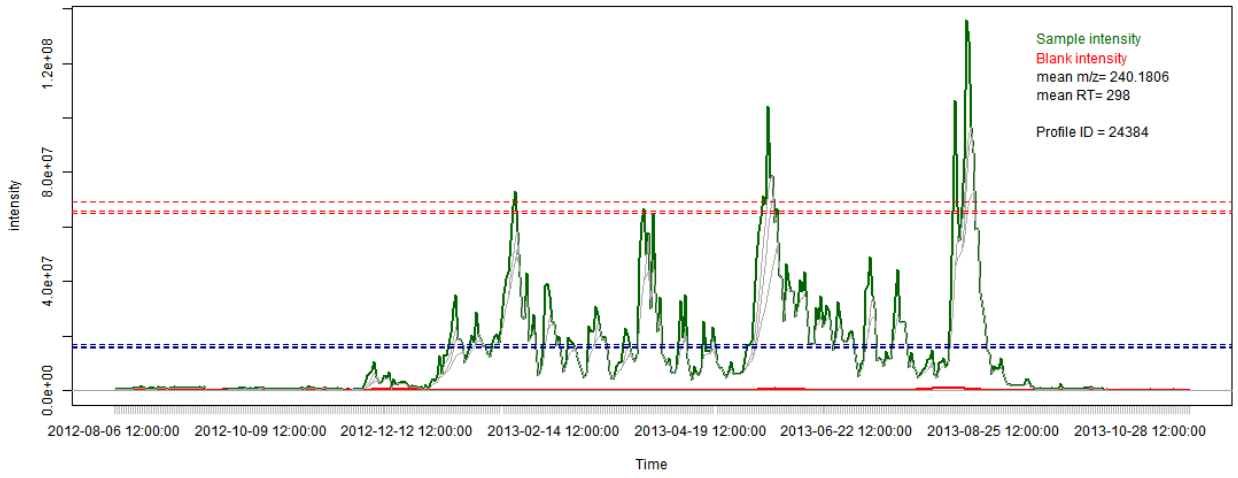

Figure 2: Illustration of the trend detection scheme in a single profile. Intensity profile of sample peaks (green solid line) and blank peaks (red solid line), the time-lagged intensity mean for trend extraction (grey solid line), the mean trend intensity (blue dashed line) and the trend intensity threshold (red dashed line). Note that, as lags are calculated for different time lags, different values exist for the mean trend intensities and the trend thresholds.

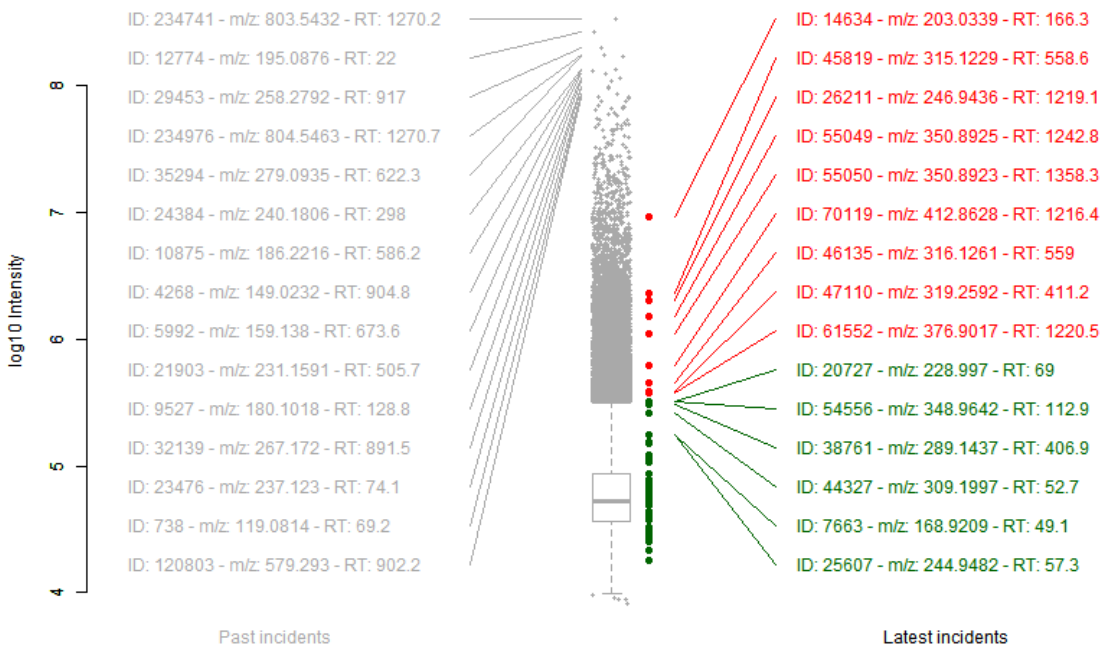

Figure 3: Boxplot of global trend intensities (grey, left side) and current intensities (colored dots, right side),listing the IDs, mean m/z and mean retention time (RT) of the underlying profiles. The red dots signify profiles with current trend intensities in the outlier range of the global incidents; green dots symbolize those below.

License

GNU General Public License version 3 (GPLv3).

Data formats

When using enviMass, make sure to have consistent data formatting for any inputs to the workflow, namely:

- Chemical Formulas: chemical formulas contain only letters, numbers and square brackets. Letters are all followed by a number of atom counts; if missing, the number is automatically set to 1. Element symbols must consist of an upper case letter, possibly followed by lower case letters. To refer to individual isotopes (e.g., from isotope labelling of a molecule, e.g., N5 vs. [15]N2N3), square brackets may precede the capital letter. Anything else that may sometimes be part of a standard chemical formula (e.g., (+), (C4H3)2, dashes,...) is not permissible.

- Retention time (RT), absolute or tolerances, are all given in seconds. This is due to standard .mz(X)ML format set to seconds. Note that this may deviate from other softwares, which often use minutes as default.

- The temporal sequence of measurements and all related parameters (e.g., lags for trend detection) is set in days. If you want to deal on temporal ranges of hours instead, divide by 24. For example, to include lags of 1h, 2h and 5h use lags = 0.042, 0.083, 0.21. Despite set at day scales, timing precision is set to the second level. Dates have to be given as YYYY-MM-DD and time as HH:MM::SS.

- Mass m: relative molecular mass and thus dimensionless, sometimes referred to using atomic mass or Dalton as unit. Mass tolerances can be given in ppm or absolute values (NOT mmu), with ppm= (mass differences / mass)*1E6.

- Charge z: integer.

- Compound names of internal standards (IS) and targets MUST NOT contain tabs, umlaute, “," or ′. Usage of any of these will lead enviMass to fail.

enviMass Internal

Structure

enviMass is a LC-MS workflow tool running from within the R statistical environment [1], loadable as a package and in turn depending on a number of packages from R CRAN and Bioconductor [2]. All dependencies are automatically checked & loaded during installation. The user interface (UI) is shown in a browser (client), with R scripts on the server side, run as a local host (IP-address 127.0.0.1) as provided by the R-package shiny [3]. The UI is based on shiny`s reactivity model and has a few HTML and JS contents (e.g., divs, busy-messages) that are extensions to the shiny`s standard widgets. Detailed tutorials on how to use shiny can be found on its website.
The uncompiled enviMass package is structured as following: (a) the R folder contains R functions to be run during server sessions, (b) the src folder contains C/C++ code to be used within some of the R functions for speeding up calculations, (c) the inst\webmass folder contains the shiny UI and server files, (d) inst on contains logo.gif and a manual.pdf which is copied during installation into shiny`s www folder to make it available from within a browser. The UI is – as required by shiny – run by the server.r and ui.r files, which in turn source a large number of other server_x.r- and ui_x.r-files. The sourcing can be nested, e.g., server_calculation.r sources further files for each step of the workflow calculation, such as server_calculation_profiling.r. The files contain a number of reactive (and sometimes conditional) dependencies between server and client which cannot be covered in short (these dependencies are modelled as shiny`s reactive values, reactive expressions and observers). However, the files can be categorized as:

- **server.r**: main shiny server file; mainly sources other files.
- **server_calculation.r** & its sourced **server_calculation_x.r**: observer for the calculation button; does most workflow calculations and manages their dependencies.
- **server_obs_Add.r**: manages (loading, deleting, sorting) compounds and files.
- **server_obs_res_meas.r**: manages result inquiries.
- **server_startup.r**: to establish a new project or to open an existing one.
- **server_variables_in.r** and **server_variables_out.r**: manages in- and outgoing project parameters.

- **ui.r**: main shiny interface file; mainly sources other files, contains header.
- **ui_busy.r**: JS for being-busy notifications
- **ui_mainPanel.r**: shown within project
- **ui_mainPanel_startup.r**: startup panel
- **ui_sidebar.r**

Each enviMass project consists of a set of folders with raw data, results, exports, graphs, …, and a logfile.emp (= R list object). The latter stores the project settings, parameters, calculation state and workflow options; it is initialized, loaded first and updated for the project path whenever a new enviMass project is established, opened or calculated. Data acquired during calculation are stored as R objects in one of the named project folders, foremost the result folder. Upon loading a logfile.emp, these data and the workflow/parameter settings stored in logfile.emp are loaded into the R workspace and

the UI, respectively, using server_startup.r and server_variables_in.r. In turn, observers in server_variables_out.r write changing project settings into logfile.emp and save the latter. Similarly, logfile.emp entries for the calculation state of a project are updated and saved within many of the server_calculation_x.r files.

Extending

When adding tools to the existing workflow, make sure to include and check the (minimum) following entries in the package *inst* folder *ui_x.r* and *server_x.r* files, as well as underlying R function files in the package *R* folder:

(1) Add parameters and workflow steps to *ui_mainPanel* to have them in the interface

(2) Add defaults for parameters and workflow *logfile.emp* settings to *newproject(),* which sets up any new project

(3) Add defaults for parameters and workflow *logfile.emp* settings to *checkproject()*, which checks any existing project, e.g., ahead of new calculations

(4) Add parameters / workflow steps to *server_variables_out* to write from the UI to *logfile.emp*; this must also include dependencies between different workflow steps.

(5) Add parameters / workflow steps to *server_variables_in* to write from *logfile.emp* to the UI

(6) Add to *server_Add* if necessary – mainly for data upload during UI runs.

(7) Add R function files to the R folder.

(8) Upate their helpfiles in the man folder.

(9) Calculation process: embed tool R functionality in the *server_calculation* file and its various sourced subfiles *server_calculation_x*, at the appropriate *at* level. During calculation, entries must be updated at *logfile$workflow_steps, logfile$summary* and *logfile$to_be_redone* - both for tool execution and tool skip. Make entries to table *summa* as to monitor progress during calculations. Define global and local variables simultaneously.

(10) For package consistency: add dependencies and R function calls in the *DESCRIPTION* file of the enviMass package.

(11) Include and save current calculation results: .png to project pics folder and *output$,* tables/lists/… to project results folder and output$. In specific cases, set dummies in *newproject()*.

(11) Load older results via *server_startup*; check if files exist – either directly or via the *logfile.em*p .

(12) Certain calculations are not managed by *logfile$workflow_steps, logfile$summary* and *logfile$to_be_redone* alone. For example, isotope pattern calculations are run per compound; tags are used in the concerned tables to trigger calculation in combination with the *logfile$* entries.

Older projects may become corrupt after enviMass modifications. Workaround: define a new project via the UI, close it and copypaste its *logfile.emp* to the older project folder, replacing the older *logfile.emp* file.

References

[1] R version 3.0.2, 2013. The R foundation for statistical computing, Vienna, Austria. http://www.R-project.org.

[2] Bioconductor: Open software development for computational biology and bioinformatics R. Gentleman, V. J. Carey, D. M. Bates, B.Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, and others 2004, Genome Biology, Vol. 5, R80

[3] RStudio and Inc. (2013). shiny: Web Application Framework for R. R package version 0.8.0. http://CRAN.R-project.org/package=shiny