g4db 0.48

Eric Largy

2020/04/29

Contents

1 Added features

```
# to mirror repository to team repo:
# D:\owmCloud\Projects\G4 database\g4dbr> git push --mirror https://github.com/g4db-team/g4dbr

#to do:
#- make NMR label as text for comma separated labels
```

1.1 v. 0.48

- Creation of a package to encapsulate the app, additinal functions and datasets
  - Creation of datasets in .Rda format to replace the .xlsx files containing the mass and isotope references
  - These can still be modified but require to convert the modified excel file into .Rda within the package using `devtools::use_data()`
  - Three functions: *g4db* to launch the app, *epsilon.calculator* to calculate extinction coefficients of oligonucleotides (used within g4db but can be used as a standalone function), and *nb_row_extract* that determines the number of rows in panelled figures to adjust their size within *g4db*.
  - Documentation of the functions
  - Documentation of the datasets
- Database load button replaced by file import button
  - Database no more hard-linked to code, now the user can specify whichever database they want
- Oligonucleotide selection now in two phases: bulk selection from *left sidebar* and refinment from *general information table*
  - All oligonucleotides now selected by default in *left sidebar* dropdown menu
  - *General information table* displays oligonucleotides selected in *left sidebar* (all by default)
  - Data displayed now filtered from the *general information table* (selectable rows)
  - Any oligonucleotide information field (including potential new ones) can now be used for data selection via the table with no UI/server code change
- *Buffer* selection now also possible by *electrolyte* and *cation* selection
  - Further selection by existing individual *electrolyte* + *cation* combination still possible
  - All values selected by default
  - Displayed data is intersection of *electrolyte*, *cation*, and *buffer* selections (e.g. if KCl is not selected in *cation* then TMAA + KCl data will not displayed even if it selected in *buffer*) _ When no *cation* is specified in the input data (e.g. sample in TMAA alone), a `none` value is automatically attributed to allow selection of cation-less data in the dropdown menu. Another consequence is that there a now empty *cation* row in the database
- *MS* can now be filtered by *tunes* and *replicates*
  - *Tune* names and *replicate* numbers filter now available in the *right sidebar* of the MS plot box
  - *Tune* names and *replicate* numbers are collected from the oligonucleotide- and electrolyte/cation/buffer-filtered MS data (i.e. only relevant *tunes* and *replicates* are proposed) _ Multiple selections possible _ Values default to the first available ones
- *MS* figure grid layout can now be customized:
  - All variable couples among *oligonucleotide*, *buffer*, *replicate* and *tune* can be plotted
  - Switch added to transpose the grid

  - Dynamic legend added, which maps the two non-selected variables to the colour of the spectra. E.g. if the grid is *oligonucleotide* x *buffer*, then colours will be attributed to *tune* x *replicate*. No two superimposed spectra can have the same colour and variations across all four variables are visible at once.
  - Label colours mirroring those of the spectra
- Dynamic *MS*, *CD* and *NMR* figure dimension: grid height expands with number of rows automatically
- *meltR* now manages data that cannot be baseline-subtracted
  - Export raw data button added in *left sidebar*
  - Absorbance normalised to [0,1] for plotting in the baseline-subtracted data figure (same y-axis scale)
- NMR peak labelling now accepts character strings

1.2 v. 0.37

- All data input tabs
  - The *buffer* field is now divided into two *buffer* and *cation* fields. Note that the field names are not used in the code, so it can be renamed as you wish at no extra cost (this is true of all variables ; only the relative position of the fields are used).
- UV-melting
  - Table switched to a wide format, with header rows for buffer, etc. for easier/faster data copy-pasting
  - No more field for specifying the *ramp* (cooling or heating), the automated assignment is used exclusively
- MS input
  - Added *replicate* and *tune* field for MS input
  - Added check for duplication of *replicate* and *tune* fields upon database edition
- MS label input
  - Table switched to a wider format (the long format became impractical after adding the new MS input field)
  - x = charge, y = labels
  - Exact same header as in MS input for quick copy-pasting, and easier to verify that all data to be imported has been labelled
- NMR label input
  - Table switched to a wider format, consistent with the changes of MS labels (different data formatting in different tabs is prone to confuse users)
  - x = peak number, y = chemical shift ; empty rows are still not a problem, making it easier to paste data for several oligos at once in certain cases
  - Same heaser as in NMR input, same remarks than for MS labels

2 Main features

2.1 General overview

DatAniRban is dedicated to the consolidated vizualisation of circular dichroism (CD), ^{1}H NMR, UV-melting and native ESI-MS data from selected G-quadruplex oligonucleotides whose structure has been deposited on

the PDB. To do so, it also contains tools to selectively import and automatically treat raw data. DatAniRban can therefore be used as a data treatment/vizualization software, regardless of the database.

The main features are:

- Vizualisation of CD, UV-melting, ^{1}H NMR and ESI-MS data imported from a templated Excel file, or from the local database
- Collapsible and tabulated interface
- Automated data treatment
  - Conversion of CD to molar ellipticities
  - MS data normalization
  - ^{1}H NMR and MS peak labelling
  - Quick and user-friendly data filtering (oligonucleotide, buffer, x-axis range)
- Custom figures
  - Control over colors, size, and transparency of figures
  - Color palettes adapted to qualitative, sequential, and diverging data
  - Switch between overlaid and paneled figures for quick comparisons
  - Control over variables mapped in paneled figures
  - Automated colour mapping to non-paneled variables
  - Automated figure dimension change to accomodate multiple rows
- Robust database
  - Selective data importing (by technique/oligo/buffer/data range)
  - Duplicate detection/suppression
  - Password-protected edition
  - Traceability: automated deposition date and DOI link generation
  - Automated oligonucleotide and buffer list collection
  - Manages replication for MS and UV-melting data
  - Manages tuning for MS data
- Open
  - Easy-to-export data tables (practical for standalone data treatment)
  - Import template easy to read in other software

2.2 Use for data deposition

The raw data formatted in the template for DatAniRban can be deposited and viewed in several ways, which are open to other scientists without the need for proprietary software use. This approach is three-fold.

2.2.1 Template file in an Excel-like software

Once pasted into the input template, the data can be deposited as is. It can then be explored natively in Excel or any open-source equivalent. The data formatted is formatted in a non-ambiguous format, and should be properly labelled in the header cells. The template is also very amenable to pieces of software allowing header cell import/management, such as Origin, in which import script can be used.

2.2.2 Template file in DatAniRban

If deposited alongside DatAniRban, the raw data can also be visualized with this open-source software. The advantages over Excel/Origin for this particular application are numerous in terms of both ease and speed of use (data filtering, automated figures, etc.), and functionalities (peak labelling, normalisation/calculation, selective data export, etc.). See the *general overview* for more details.

2.2.3 Database file in DatAniRban

The use of DatAniRban also allows exporting selected datasets to a database were the data is consolidated and all calculation has already been performed. The database file can be deposited alongside DatAniRban, to enjoy all of its functionalities with faster performances, and control over submission authorship and dates.

---

3 Preliminary remarks on data formatting

There are two basic way to store data, that is in a wide or long format. In the *wide format*, different datasets are presented in different columns. In the table below, the data is shown with a single x column and one y column for each oligonucleotide that was analyzed.

	Wide format		
x	Oligo1	Oligo2	Oligo3
1	0.299070723587647	0.0168962816242129	0.672464023577049
2	0.878921636147425	0.769953901879489	0.118909144308418
3	1.69430859386921e-05	0.851018289336935	0.60637704632245
4	0.230462895939127	0.305208650650457	0.738826122134924
5	0.955561132868752	0.820542640751228	0.751897171139717
6	0.147223622538149	0.795151739614084	0.00561738875694573
7	0.555317368824035	0.475179521599784	0.969051418127492
8	0.374301896663383	0.598382666241378	0.589954387163743
9	0.19796669203788	0.00748656992800534	0.0969171042088419
10	0.39302771515213	0.535146594513208	0.28323801769875

It is easy to add data to such table, by simply pasting the new data set in a new column. It has two major drawbacks though:

1. If the x values are not shared, one need to add a new x column and the data will be mismatched (as in MS), leading to
2. It is globally harder to filter data by any given variables.

	Wide format mismatched				
x1	Oligo1	x2	Oligo2	x3	Oligo3
1	0.348155150422826	2	0.471839506411925	0	0.485694591188803
2	0.226233282359317	3	0.551377793308347	1	0.881106314714998
3	0.915780127746984	4	0.733992651337758	2	0.10820273286663
4	0.508837714325637	5	0.000484986696392298	3	0.677298142341897
5	0.170618709176779	6	0.947096308227628	4	0.362098056590185
6	0.777400587219745	7	0.162980386288837	5	0.132547377608716
7	0.779971096199006	8	0.950781109044328	6	0.900770844193175
8	0.0949956437107176	9	0.923481578705832	7	0.470615454483777
9	0.793374943546951	10	0.504868268733844	8	0.68203906645067
10	0.353502612793818	11	0.545238607330248	9	0.944358139531687

Conversely, in the *long format*, all data sets are stacked in the same columns but each variable is stored in its own column. In the example above, the data has three variables x, y, and the oligonucleotide name, leading to three columns.

	Long format	
x	oligo.name	value
0	Oligo3	0.6724640
1	Oligo1	0.2990707
1	Oligo3	0.1189091
2	Oligo1	0.8789216
2	Oligo2	0.0168963
2	Oligo3	0.6063770
3	Oligo1	0.0000169
3	Oligo2	0.7699539
3	Oligo3	0.7388261
4	Oligo1	0.2304629
4	Oligo2	0.8510183
4	Oligo3	0.7518972
5	Oligo1	0.9555611
5	Oligo2	0.3052087
5	Oligo3	0.0056174
6	Oligo1	0.1472236
6	Oligo2	0.8205426
6	Oligo3	0.9690514
7	Oligo1	0.5553174
7	Oligo2	0.7951517
7	Oligo3	0.5899544
8	Oligo1	0.3743019
8	Oligo2	0.4751795
8	Oligo3	0.0969171
9	Oligo1	0.1979667
9	Oligo2	0.5983827
9	Oligo3	0.2832380
10	Oligo1	0.3930277
10	Oligo2	0.0074866
11	Oligo2	0.5351466

In this format, the number of columns is independent from the number of experiments. If a fourth variable was used, say a buffer name, a fourth column would have been added. Mismatched x values have no more impact, and the data can be readily sorted by any variable (Here by ascending `oligo.name` then ascending `x`). It is now much easier to filter the data by any given variable while conserving the data properly formatted in the table.

Below the data has been filtered by `oligo.name` (only *Oligo1* and *Oligo3* are selected) and `x` values between 2 and 5, and `value` > 0.5. It would have been easy to filter by oligonucleotide from a *wide format* table by not selecting the column, however filtering x and y values from tables with mismatched x scales must be performed column by column.

	Filtered long format	
x	oligo.name	value
3	Oligo3	0.7388261
4	Oligo3	0.7518972

The same goes for mapping variables to figures, in terms of what to plot on the x/y axes, which parameters control the data color, shape, size, etc., and creating paneled figures, (more details below), following the grammar of graphics that has been implemented in the ggplot2 package of R.

It is however very impractical to work with a *long format* table in Excel or the like, where it is necessary to stack each new data set and fill the variables such as oligo names, buffers, etc. manually. This is tedious and prone to errors particularly that it generates a lot more rows than in the *wide format* (each extra variable doubles the number of rows, assuming all the possible experiments across these variables have been performed).

To summarize, it is easier to prepare a wide table in Excel, and then work with a long table for data manipulation and visualization. Consequently, the *wide format* was selected for compiling data for importing into the database and an easy-to-fill template was created to do so. After the data is imported it is pivoted into the *long format* automatically.

---

4 Features and use of DatAnirban

4.1 General interface features

4.1.1 Organization

The interface is divided in 3 tabs that can be selected at the top of the screen:

- *database*, to visualize the content of the database,
- *ImportR*, to visualize new data and export all or part of it to the *database*,
- *meltR*, to visualize and treat UV-melting data, and export all or part of it to the *database* (via *ImportR*),

4.1.2 Sidebars

Each tab has a sidebar on the left-hand side, which contains a number of tools for data importing, exporting, filtering, and formatting. This *left sidebar* is collapsible to release some space for figures and tables on smaller screens.

The sidebar from the *database* and *ImportR* tabs also contain a color palette selection menu, and submenu for certain palettes having variations. The available palettes include:

- The well known Brewer palettes that include qualitative, diverging, and sequential palettes,
- Some discrete palettes from D3.js, a JavaScript library for producing interactive data visualizations,
- Several palettes inspired by the colors used by scientific journals/publishers (NPG, AAAS, NEJM, Lancet, JAMA, JCO, etc.)

Drop-down menus contains select all/deselect all buttons for quick data selection. The values from the *left sidebar* modifies the data for *all* the content of the tab. Each tab has a specific and independent *left sidebar*.

Given the amount of menus necessary for the *meltR* tab, a large portion of it is hosted by two collapsible and movable "hovering" panels.

4.1.3 Figures and tables

The figures and tables are hosted within collapsible and closable boxes, so that the user can select what data to display at any given time.

Each figure box from the *database* and *ImportR* tabs features a *right sidebar*. They contain filtering and data formatting filters that are applied *only* on the corresponding figure. These sidebars are collapsible as well, and hidden by default.

All tables are sort-able and filterable to assist in exploring rich data sets, and find specific data points rapidly. Filtering the tables do *not* alter the figures. Each column can be selectively hidden, and some of the less interesting ones are hidden by The data is presented in *long format*, which makes it easier to filter through, and to map variables into figures, because each variable is contained in its own column.

All tables can be exported as .csv, .xlsx, or in the clipboard. Note however that this data is in a long format, that is not necessarily easy to work with with a piece of software like Excel, but is much more powerful to map different variables into figures.

4.2 Database

4.2.1 Input data

The *Database* tab allows to consult the data contained into a database file, by selecting the oligonucleotide(s) of interest, and where necessary, specific buffers. This tab is read only (database modifications must be performed in the *Import R* tab), but allows exporting all or part of the data.

The database data is contained in an Excel (.XL) file, in the *long format*. Although it is not formatted to be easily consulted nor modified in Excel, it is very much possible.

The general info (*info*) and data of each experimental method (*CD*, *NMR*, *UV-melting*, *ESI-MS*) are stored in their own tab within the Excel file, allowing to selectively read and write the database (see below).

The database is loaded by clicking on the *load database* button from the sidebar. By default, no data will be displayed to avoid long waiting times, particularly if the database is large. To start visualizing data, the *oligonucleotide(s)* of interest must be selected from the drop-down menu just below. The list of *oligonucleotides* is collected automatically from the info *panel*. It is therefore important to maintain this info accurately when important new experimental data.

By default, all available *buffers* are selected ; they are automatically collected from the *CD* and *UV-melting* data. The buffers from MS and ^{1}H NMR data are not collected ; their data is filtered immediately upon importing to save on memory use. It can be implemented if necessary, but it is unlikely that their buffers mismatch with those from *CD* and *UV-melting* (otherwise the database wouldn't be very consistent). It is always possible to trick the database by adding a fake data point with the desired buffer, if the problem were to arise punctually.

Once at least one oligonucleotide has been selected, the data will be displayed.

4.2.2 Visualization

4.2.2.1 General information. This groups all the information data by the user. The DOI is automatically made into an hyperlink for quick access to the relevant publication presenting the high-resolution structure. The deposition date is added automatically when the data is imported through *ImportR*.

From the sequence are computed the number of each and all nucleotides, the atomic composition, and the average and mono-isotopic masses. The two latter are calculated based on what was published in Anal. Chem., although it was streamlined as there is no need for isotopic distribution calculation, nor non-natural isotopic abundances.

4.2.2.2 Circular Dichroism. The circular dichroism data is presented as a scatter plot colored by buffer and shaped by oligonucleotide (both changeable). The data can be shown in mdeg or molar ellipticities (default), and filtered by wavelength. Molar ellipticities are calculated automatically from the supplied mdeg data.

The spectra can be all overlaid (default), overlaid by oligonucleotide or buffer, or not overlaid at all, to ease the comparison of data sets.

The selected data is also shown in a table under the plot. Further filtering of the table does not alter the plot, nor does filtering from the the plot's *right sidebar*.

4.2.2.3 NMR. The ^{1}H NMR spectra are presented as a line plot colored by oligonucleotide (changeable). The chemical shift is presented in descending order, as is tradition, and can be filtered. The same stacking options than for CD are available, but the default is unstacked for clarity. No data with different buffers was submitted, so it is unclear whether buffer (un)stacking is necessary (but it is implemented nonetheless). The line size can be changed.

A data table is available, with the same remarks as previously.

4.2.2.4 UV-melting. The UV-melting data is presented in two scatter plots. On the left-hand side, the fitted raw data is shown, where the right-hand side plot presents the baseline-subtracted data. The former shows how the latter was obtained, and in particular the fit that was used to determine the baselines and the thermodynamic parameters. The latter is more appropriate for visual determination of Tm, comparison across samples, and determination of the amount of folded species at any given temperatures.

The data is colored by `id`, which is unique for any given oligonucleotide-buffer-ramp-replicate combination. A paired color palette is particularly well-suited for this type of data visualization. Besides the colors, the temperature range, line size and transparency and point size can be changed.

A data table is available, with the same remarks as previously.

4.2.2.5 Native ESI-MS. The MS data is not plotted when one or several oligonucleotides are selected until the button *plot MS* has been clicked on. That is because the amount of data to plot can be quite large, leading to slow plotting speeds. To avoid long refresh times of the software every time a new oligo/buffer is selected/deselected, plotting (and re plotting) only occurs when desired.

The spectra are shown unstacked, in an oligonucleotide/buffer grid. Labels for species defined when importing can be shown or not. The colors, m/z range and line size can be changed.

A table will be added.

4.3 ImportR

The ImportR tab's purpose is to selectively import raw data into the database. As a corollary of this primary function, it gives allows automated data treatment and Visualization of CD, ^{1}H NMR, UV-melting and MS data.

4.3.1 Input data

The raw data must be supplied in a template .xlsx file, in a wide format (except for UV-melting so far) that is with one column per dimension with extra information being filled into a header.

To open said file in *ImportR*, click on *Browse...* at the top of the *left sidebar* and select the file.

The file is divided into seven tabs designed to contain raw data (UV, CD, NMR, MS), general oligonucleotide information (info), or peak labeling data (NMR and MS labels).

4.3.1.1 info. Five fields must be filled, i.e.:

- `oligo` that is the name of the oligonucleotide, preferably a PDB code,
- `sequence`, in the 5' to 3' direction, without spaces or dashes. If present, only the extinction coefficient will be affected, although this can be corrected.
- `submitted_by` is the initials of whoever submits the data, for traceability purpose,
- `DOI` is the DOI of the paper linked to the PDB deposition, it is converted into a link automatically in *ImportR.*
- `Topology` contains a description of the structure that can be as long or short as necessary; it is displayed as is in the *database.*

All the other fields that cab be seen in the corresponding table in *ImportR* and *database* are calculated by the software.

	A	B	C	D	E
1	oligo	sequence	submitted_by	DOI	Topology
2	2M4P	TTGTGGTGGGTGGGTGGGT	EL	10.1038/nature755	parallel
3	2LEE	TAGGGCGGGAGGGAGGGAA	EL	10.1021/ja208483v	parallel

Figure 1: Info template

4.3.1.2 CD. The CD data must be pasted in two columns, below the header, with the wavelenght in the first column and the ellipticity in mdeg in the second column. The oligonucleotide and buffer names, the cuvette path length, and the oligonucleotide concentration must be given in the header rows. It is important to keep buffer names consistent with the database content.

For every new data set (new oligonucleotide/buffer), the next two column must be used and so forth. Even if the wavelength axis is the same, it must be specified again; this allows dealing with mismatched axes.

	A	B	C	D	E	F	G	H
1	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*
2	oligonucleotide	2M4P	oligonucleotide	2M4P	oligonucleotide	2M4P	oligonucleotide	2LEE
3	buffer	TMAA	buffer	TMAA+KCl	buffer	Kp+KCl	buffer	TMAA
4	pathlength (cm)	0.4	pathlength (cm)	0.4	pathlength (cm)	0.4	pathlength (cm)	0.4
5	oligo concentration	10	oligo concentration	10	oligo concentration	10	oligo concentration	10
6	wavelength (nm)	CD (mdeg)	wavelength (nm)	CD (mdeg)	wavelength (nm)	CD (mdeg)	wavelength (nm)	CD (mdeg)
7	350	0.04499791	220	1.044384219	350	-0.337267	350	0.089797941
8	349.8	0.056435374	221.3	0.490227894	349.8	-0.343757	349.8	0.088003829
9	349.6	0.062846686	222.6	-0.06577531	349.6	-0.341989	349.6	0.085259893
10	349.4	0.069086502	223.9	-0.6226228	349.4	-0.338243	349.4	0.088768965
11	349.2	0.073901581	225.2	-1.17738596	349.2	-0.34166	349.2	0.091209485
12	349	0.080471197	226.5	-1.72105464	349	-0.351237	349	0.099388524
13	348.8	0.090312429	227.8	-2.33325576	348.8	-0.354535	348.8	0.111485588

Figure 2: CD template. Four spectra have been added. Note that one of the x-axis is mismatched

4.3.1.3 NMR. The ^{1}H NMR template follows the same principle: two columns per oligonucleotide/buffer for the chemical shift and intensity, and two header rows for the oligonucleotide and buffer names.

It is preferred to not import data that will not be visualized in the *database*. The user can either paste selectively the data into the template, or the full spectrum and use the filtering tools in *ImportR*. For online use, it is advised to paste selectively the data to avoid long uploading times.

4.3.1.4 NMR labels. This tab is used to submit ^{1}H NMR peak labelling information. The first column must be filled with peak numbers, in any order. For each oligonucleotide, a column must be added with the oligonucleotide name as title and the chemical shift of the corresponding peak below. All cells do not have to be filled, only those for which a given oligonucleotide has a peak with this number.

4.3.1.5 MS. The MS template is identical to the NMR template, albeit that column one is *m/z*.

It is strongly advided to only paste the necessary data to vizualise as it can get quite heavy otherwise, which is an issue for uploading and plotting times. A function was designed to quickly remove data points, by selecting a user-defined m/z range, and removing all data points whose intensity is lower than that of the average of a user-supplied baseline range. In the figure below, the MS spectra of 2M4P in TMAA was plotted from the full raw data (300-3150 *m/z*, focused on 1150-1650 *m/z*), that is 7.2 MB.

	A	B	C	D
1	*x*	*y*	*x*	*y*
2	**oligonucleotide**	2M4P	**oligonucleotide**	2LEE
3	**buffer**	TMAA+KCl	**buffer**	TMAA+KCl
4	**Chemical shift**	**Intensity**	**Chemical shift**	**Intensity**
5	12.999957	-550	12.99979	485
6	12.999655	-637	12.999488	341
7	12.999352	-774	12.999185	178
8	12.999049	-936	12.998882	8
9	12.998747	-1090	12.99858	-151
10	12.998444	-1205	12.998277	-288
11	12.998141	-1271	12.997974	-396
12	12.997838	-1298	12.997671	-476
13	12.997536	-1311	12.997369	-538

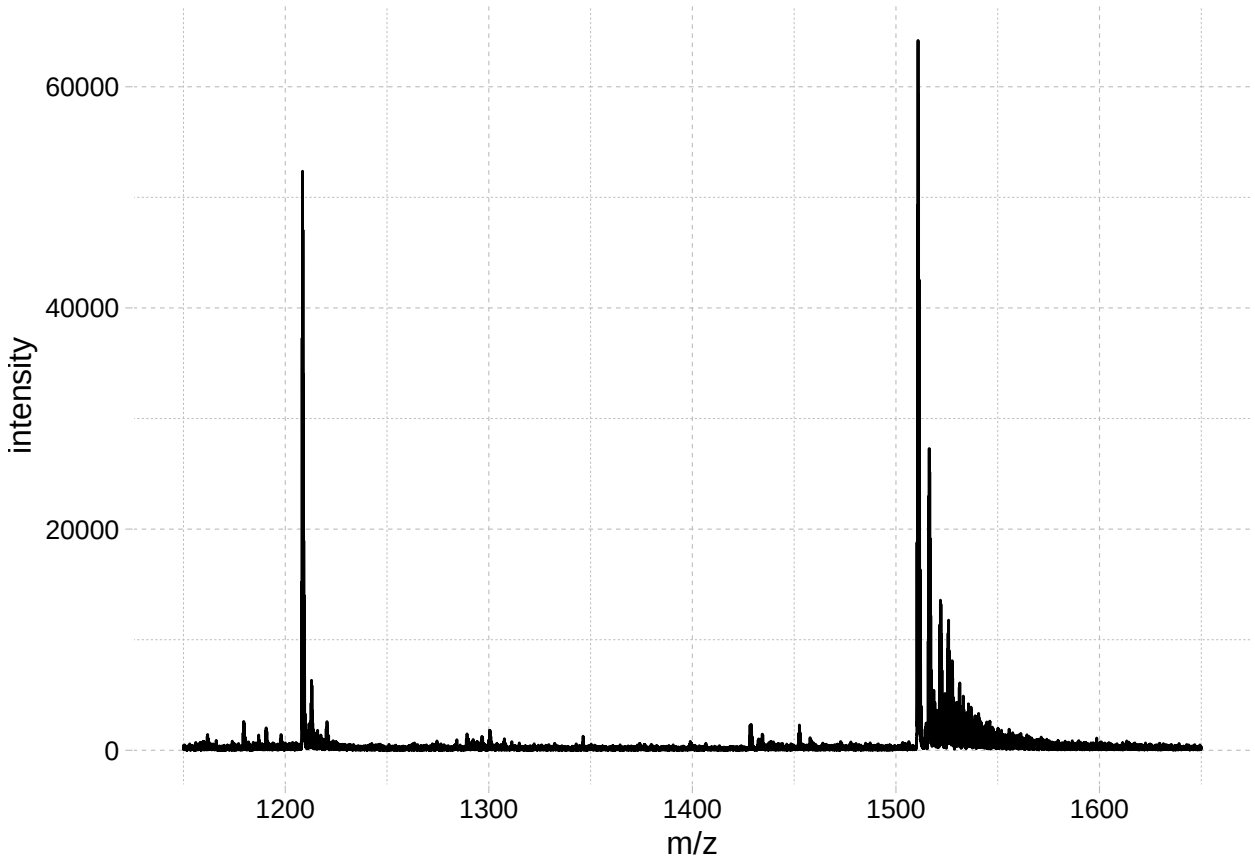
Figure 3: NMR template

	A	B	C
1	**peaks**	**2M4P**	**2LEE**
2	3	11.85	11.633
3	4		11.328
4	5	11.37	11.127
5	6	11.277	
6	7		11.731
7	8	11.778	11.439
8	9	11.419	11.354
9	10	11.287	
10	11		11.91
11	12	11.744	11.294
12	13	11.286	11.155
13	14	11.088	
14	15		11.234
15	16	11.545	11.234
16	17	11.595	10.768
17	18	11.207	

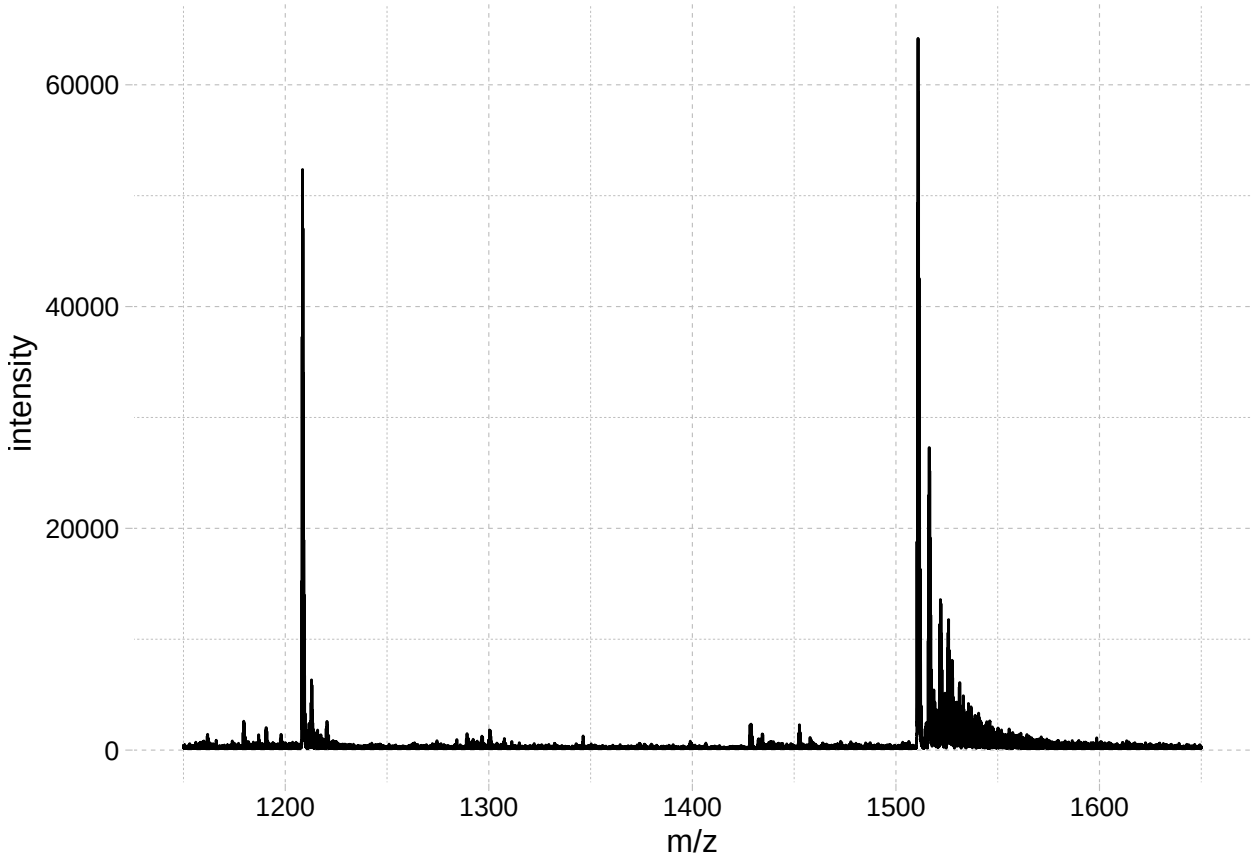
Figure 4: NMR labels template. Note that both oligonucleotides have completely different labelling

	A	B	C	D	E	F	G	H
1	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*
2	**oligonucleotide**	2M4P	**oligonucleotide**	2M4P	**oligonucleotide**	2LEE	**oligonucleotide**	2LEE
3	**buffer**	TMAA	**buffer**	TMAA+KCl	**buffer**	TMAA	**buffer**	TMAA+KCl
4	**m/z**	**Intensity**	**m/z**	**Intensity**	**m/z**	**Intensity**	**m/z**	**Intensity**
5	1150.0299	125	1150.0012	194	1150.0006	178	1150.0592	877
6	1150.0357	129	1150.0071	160	1150.0064	241	1150.065	932
7	1150.0416	120	1150.013	156	1150.0123	316	1150.0709	1092
8	1150.0474	172	1150.0188	148	1150.0181	330	1150.0767	1297
9	1150.0533	164	1150.0657	150	1150.024	367	1150.0826	1656
10	1150.0592	128	1150.0716	173	1150.0299	407	1150.0884	1920
11	1150.0943	126	1150.0774	169	1150.0357	400	1150.0943	1993
12	1150.1002	195	1150.0833	152	1150.0416	390	1150.1002	1983
13	1150.106	214	1150.0891	163	1150.0474	419	1150.106	1827
14	1150.1119	180	1150.095	192	1150.0533	441	1150.1119	1628
15	1150.1177	169	1150.1008	206	1150.0592	412	1150.1177	1611
16	1150.1236	203	1150.1067	189	1150.065	342	1150.1236	1664
17	1150.1295	211	1150.1126	149	1150.0709	313	1150.1295	1774
18	1150.1353	178	1150.1184	151	1150.0767	236	1150.1353	1780
19	1150.1412	128	1150.1243	159	1150.0826	182	1150.1412	1844

Figure 5: MS template



After applying the filtering function, only 589 KB of data remains with no visible loss in terms of visualization.



4.3.1.6 MS labels. This tab is aimed at providing the database with the species to label in the MS spectrum. It differs from the NMR label tab, where one must supply the chemical shift of each label. Here, the user must simply supply the name of the species that must be labelled, that is *M* for the unaducted oligonucleotide, MK for a single potassium adduct, MK2 for a two-potassium adduct species, and so forth.

In order to label the different charge states and buffers independently, the first two columns must contain their respective values. All following columns must be titled with the oligonucleotide name and the species name below. Not all cells must be filled, only those for which a species must be labelled for a given oligonucleotide/buffer/charge.

4.3.1.7 UV-melting. The UV-melting tab is neither completely in a wide or long format. It will be made into a wide format and described here in the coming days.

The temperature can be supplied in Celcius or Kelvin, *meltR* will convert it automatically to Kelvin where necessary.

4.3.2 Data vizualisation and selection

The interface for vizualisation and all filters and options are roughly the same as in the *Database* tab. The main difference is the absence of UV-melting plots, as they have the dedicated *meltR* tab.

The behavior for data plotting upon importing data is also the same, with the MS plot being subject to an additional button click.

	A	B	C	D
1	charge	buffer	2M4P	2LEE
2	4	TMAA	M	M
3	4	TMAA	MK	MK
4	4	TMAA+KC	MK	
5	4	TMAA+KC	MK2	MK2
6	4	TMAA+KC	MK3	MK3
7	4	TMAA+KC	MK4	MK4
8	4	TMAA+KC	MK5	MK5
9	5	TMAA	M	M
10	5	TMAA	MK	MK
11	5	TMAA+KC	MK	
12	5	TMAA+KC	MK2	MK2
13	5	TMAA+KC	MK3	MK3
14	5	TMAA+KC	MK4	MK4
15	5	TMAA+KC	MK5	MK5

Figure 6: MS labels template. Note the difference in labelling between oligonucleotides and buffer. Both charge states have been labelled the same however

Importantly, the oligonucleotide and buffer selections, as well as the figure range filters, *do* condition what will and will not be exported to the database.

4.3.3 Database edition

To add new data to the database, the data must be opened in *ImportR*, then selected and filtered (oligonucleotide/buffer/x-axis range) as desired. If only a partial import is desired (e.g. only CD data), it is not necessary to deal with the other techniques. Although all the data will be displayed (expect MS by default), only the selected ones will be exported (as explained below). If UV-melting data must be imported, then it must be first treated in *meltR*.

To write the selected data to the database, the database must first be loaded from within the *database* tab, then the techniques to export must be switched on in the *left sidebar*. All techniques for which the toggle is off will *not* be written into the database, regardless of them being displayed in *ImportR*.

A password must be supplied to avoid accidental database edition, and finally a click on *Write to db* will edit the database. For each technique, as well as the *info*, the software will look for and remove duplicated data (by technique, oligonucleotide, buffer, and x-axis value). It remains therefore possible to submit additional buffers or extended data ranges to already existing oligonucleotides in the database. For instance, one can submit a 800-2500 m/z MS spectrum to an oligonucleotide for which a 1000-2000 m/z already exists in the database; the software will append the 800-1000 and 2000-2500 m/z data to it, without duplicating the points from the original database data.

4.4 MeltR

Still to come...