OpenPlant R for Proteomics report of August 2018

Marielle Vigouroux, Jan Sklenar, Govind Chandra

Project Title: *R for Proteomics* training

Report Title: Report of 31 July 2018

Summary

Report and outcomes

As proposed, training sessions were held on the following dates and venues and attended by (with some absences) Jan Sklenar, Frank Menke, Marielle Vigouroux, Gerhard Saalbach, Lisa Breckels, Laurent Gatto and Govind Chandra.

Date	Place
9 January	TSL Norwich
15 January	JIC Norwich
2 February	JIC Norwich
21 March	Department of Mathematics, Cambridge
20 April	Department of Mathematics, Cambridge
8 May	Department of Mathematics, Cambridge
29 June	Department of Mathematics, Cambridge

All sessions were 4 to 6 hours long with flexible short breaks for coffee and lunch. The first one was a presentation by Laurent Gatto followed by a detailed discussion in which a plan of action was agreed upon. We explained in detail to Laurent how proteomics data is generated in JIC and TSL, how we process it, what questions we ask of the data and, what more or different we would like to do with the data we generate. We agreed upon time to be spent on methods of labelled and label-free quantification data.

In the second session Laurent gave an introduction to R, RStudio and Bioconductor and introduced us to the packages of R which are relevant for proteomics data analysis. This included using them to demonstrate the analysis of test / toy data and map out the internal data structures

used by these packages. This level of understanding is crucial for confident and flexible use of these packages.

In the remaining sessions we mostly worked on real data provided by Jan Sklenar for spectral counting type of analysis and, that provided by Gerhard Saalbach for labelling based quantification assays. Real data posed several challenges such as working with large files as they are in reality and also that the annotation provided with the data is sometimes wrong. Laurent showed us how to deal with big data sizes as well as how to clean up and reorganise data for efficient analysis in R.

We also learned about some free alternatives to proprietary programs (for example, MSGF+ for peptide spectral matching) which can be used for preliminary / exploratory data analysis. This is particularly useful because it removes the limitation of being able to do the analysis only on those computers on which the proprietary software is installed. Doing so and by comparison with Mascot search engine (our workhorse) we dicovered an incompatibility of outputs and requested rectification, which has already been done. Thus we contributed to MSnbase development.

An unforeseen benefit of this project has been seeing a real R professional (Laurent) working with R. Not only did we learned the use of R functions in ways we would never have thought of but we also learned about R functions we did not even know existed. And some of us have significant experience with R!!

- R markdown files for recording steps in data analysis as it happens.
- tibbles (data frames with additional features and capabilities).
- shiny for application development.
- Some very useful plotting packges especially one which provides an alternative to Venn diagrams.

Although not linearly, the following were covered in depth over these training sessions.

- MSnbase
- Structure of MSnExp objects which hold raw.
- Structure of MSnset objects which hold quantification data.
- Raw data file formats such as mzML and mzXML.
- Ways of bringing together quantification data and identification data.
- Statistical analysis facilities built into MSnID and also the use of other packages such as msmsTests for analysis of differential expression.
- A handful of graphical packages which can be used for the visualisation of data and results.

Right from the beginning, all the code written during these training sessions was archived on Github so it could be accessed from anywhere. After each session, more comments were added to the code and then the code got uploaded to a directory of its own in the Github repository which all of us can access.

The primary Github repository for this project is located at
https://github.com/lgatto/r-for-proteomics-tsl

Marielle and Jan have also started working on forks of Laurent's Github repository.

- Jan's fork: https://github.com/Sjan1/r-for-proteomics-tsl
- Marielle's fork: https://github.com/mariellev/r-for-proteomics-tsl

These (especially Jan's fork) are where we will continue to add code and comments as we continue to advance our skills in using the R packages introduced to us by Laurent.

We made some peculiar demands of proteomics data analysis and management. In this process we discovered some niggles in the *R for Proteomics* packages which Laurent mostly fixed between sessions. Our data analysis demands also suggested to him some alternative and new approaches which he is planning to implement in the packages in the future. We initiated development of better annotation of fragment ions in MS/MS spectra than had been available in R. The weakness discovered was immediately added to the hot working list to be fixed by the developers of MSnbase as soon as possible. We pointed out search engine output incompatibility. So, in some small ways, this project has also led to the improvement of the packages.

We can read raw data and search results into R objects. One of the most important outcome was drafting a standard sample meta information format to be read in and connected with raw and search data. This mechanism had to be discussed several times, to allow both automation and flexibility required with various experimental design. Inspired by tidyverse.org, we used a long format of data to describe experimental design, linking in one table meta-information with raw files and search files. Then we use this table to drive all the other subsequent data processing. This design allow sample annotation to be available anytime during the data analysis and provide the cornerstone piece of information for experimental protocol and a project outcome, while still very close to data. We create documentation (automatically, on the fly) that describes the experimental design, calculate number of replicates, and check integrity of the file names before we start the main processing.

We are very close to plotting visual pictures generated from inputs: raw data (mzML), search data (mzid) and sample meta data (table). To complete this task, we need to keep meeting for roughly another couple of months, which we currently do on weekly bases.

On an individual level different members of the team benefited differently:

- Frank and Gerhard realized what is available and will be able to ask to get their data processed with our new tools.
- Jan became more independent in R and is able to use all we produced.
- Govind and Marielle learned new libraries and developed better understanding proteomics experiments and corresponding data formats.
- Laurent and Lisa benefited from several suggestions for improvements to their MSnbase package.

We all agreed to stay connected in future and meet occasionally, as we grow in our use of R for Proteomics data analysis. This is one of the greatest achievements of this project. We were able to bring two communities together and started working on the solutions that are generally available, yet still require a lot of communication, effort and learning to bring into everyday use. With every session we are becoming more efficient as we understand each other.

Changes to team

None

Expenditure

Item	Amount
Laptop	992.00
Book: Dynamic Documents with R	45.00
Travel and Subsistence (between Cambridge and Norwich)	678.00
Total	1716.00

Would you like to claim the £1,000 follow-on fund?

Yes, we would like to claim the £1,000 follow-on fund to be used as described below.

Follow on Plans

The training sessions have now ended. Jan, Marielle, Gerhard and Govind get together on a weekly basis to repeat at a slower pace what was done during the training sessions and also to explore new possibilities. Some things were only demonstrated to us during the training sessions due to constraints of time. We are now applying these to actual data. We still run into issues which we are unable to resolve ourselves and have to ask Laurent for help and guidance which he still provides very willingly.

At this time, we expect our meetings (which we have with a laptop connected to a projector or a large display) to continue indefinitely (with varying frequency) because we are all learning useful things in the process of writing code together. None of this would have been possible without encouragement and support from the OpenPlant Fund.

Laurent has now moved to Brussels to start his own group. In a few months time (sometime in November?) we would like to invite him to visit Norwich for a couple of days. During this visit we will show him what we are doing and also ask for help with problems that we are unable to solve ourselves. We will also take this opportunity to seek his comments about how we could improve our workflows both in quality and efficiency. The funds remaining and requested will be used to cover travelling expenses for Laurent between Brussels and Norwich and also his stay in Norwich for 2 or 3 nights.

Programming for scientific data analysis is a fast moving field and rather difficult to keep up with without a lot of reading. We also propose to buy some books related to this project which will be shared amongst the team (and more widely with anyone on site).

Any remaining funds will be returned.