Title: 'R for Proteomics' training proposal

==========

Primary contact:

Jan Sklenar (jan.sklenar@tsl.ac.uk)

Summary:

Proteomics is increasingly used in many research projects. While the throughput of mass spectrometers used in proteomics has increased in recent years, data processing work-flows are still a recognised bottleneck. Proteomics users struggle with large datasets. Slow algorithms and proprietary and free software often require manual intervention during data processing. That has a negative effect on reproducibility and throughput. Truly configurable tools to suit the changing requirements are rare.

Recently there has been a substantial development of R package 'R for Proteomics' (RfP) in Cambridge by L. Gatto et al. We believe that RfP is a powerful independent open source data pipeline that allows the development of customized work-flows. At the same time, it provides high quality visualization available in R, a facility mostly missing from other software packages. For this reason, it could complement our data processing and, if appropriate, become an alternative to our currently used software.

We would like to introduce the package, train ourselves and integrate it into our toolbox to serve the proteomics community in Norwich. The desired outcome should be increase in the reproducibility of data analyses and our ability to provide clearer results of protein identifications to the users in many projects we collaborate on.

We have an agreement with RfP developer to come on site and provide hands on training.

Team:

Jan Sklenar

(Jan.Sklenar@sainsbury-laboratory.ac.uk)

The Sainsbury Laboratory, Norwich

Proteomics and mass spectrometry support specialist

Familiar with many proteomics software and the requirements of various project that can be accomplished using given mass spectrometry platform. Constantly on search for the tools to improve our technology and data processing. I provide training in proteomics to the users to understand the details of proteomics experiments and the corresponding data types. Beginner in R, I actively use R scripts to visualize proteomics results.

Laurent Gatto

(lg390@cam.ac.uk)

Computational Proteomics Unit, University of Cambridge

Laurent is a computational biologist that, since 2010, specialises in mass spectrometry-based proteomics data analysis. He is the main developer of the 'R for Proteomics' suite of software that includes over

10 R/Bioconductor packages for the analysis and comprehension of proteomics data; these include packages to access raw MS and identification data, data to efficiently and conveniently manipulate and process quantitative data, software dedicated to the utilisation of machine learning algorithms such as clustering and classification to quantitative proteomics data, as well as infrastructure for data visualisation and quality control.

Marielle Vigouroux

(marielle.vigouroux@jic.ac.uk)

John Innes Centre, Norwich

Marielle is a computational biologist, who provides bioinformatics support to the NRP proteomics community and administrates the Mascot server.

Mainly uses R for analysing genomics data.

Govind Chandra

(govind.chandra@jic.ac.uk)

John Innes Centre, Norwich

Computational biologist

Govind a computational biologist and an R-expert who teaches an introductory R course for NRP staff.

The proteomics community (Proteomics and mass spectrometry support specialists)

in Norwich consists of the following people:

Frank Menke <Frank.Menke@sainsbury-laboratory.ac.uk>

The Sainsbury Laboratory, Norwich

Gerhard Saalbach (Gerhard.Saalbach@jic.ac.uk)

John Innes Centre, Norwich

Paul Derbyshire (Paul.Derbyshire@sainsbury-laboratory.ac.uk)

The Sainsbury Laboratory, Norwich

Proposal:

There are two groups of support specialists in the Norwich Research Park, R-oriented bioinformaticians and proteomics experts that would benefit from closer interactions. We would like to organize a workshop aiming to bridge the gap between these groups and provide new tools to both groups. We would like to cover the basics of proteomics and then use RfP package to improve quality and performance of proteomics data analyses. We would like to arrange for a series of training to be delivered by L. Gatto, both in Norwich and Cambridge. We propose the training to take place six times within six months to provide enough time for learning, self-study and problem solving. For convenience and exchange of ideas with the proteomics computational unit in Cambridge, the training should take place both in Norwich and Cambridge.

We believe that the proposed interactions with experts in computational proteomics will bring inspiration and fresh ideas to our mission. To cultivate this atmosphere, we are going to engage all the proteomics experts and users in Norwich regardless of their ability to produce R- code. The results of the proposed training, the code generated, will become freely available to the wider proteomics community.

First, we are going to provide proteomics training to computational scientists involved to improve their understanding of proteomics, types of experiments and the related data structures.

Secondly, RfP training will cover the basics of raw data assessment, peptide matching, spectral counting, reporter ion quantitation and visualization of the results.

Third, we are going to asses our current data analysis work-flows in terms of efficiency and reproducibility. We will use RfP to complement our workflows and, where applicable, we will prepare and test an alternative solution. That should result in the development of scripts to automate repeated tasks e.g. comparison of several group of samples differentially changing in label free or label based experiment and visualize the results (See the outcomes section for more examples). The datasets that we can use for the development are are available in our raw data archive.

While our current data processing pipeline consists of mostly commercial software that is reasonably fast for a typical project, it is not practical for large dataset analysis. Since RfP is scalable to high performance computing we will explore the possibility of using a computer cluster (local or cloud computing based) to accelerate the analyses of large datasets.

All established procedures and practices will be documented on the github platform for future reference. All the training material that will be compiled or developed for these courses will also be released for the wider community to make use of.

Benefits and outcomes:

We believe that RfP could bring bioinformatics and proteomics in Norwich closer together.

The main focus of the proposed training will be better automation, increased reproducibility of data analysis and, better visualization of the results. This can be used in many projects involving proteomics and mass spectrometry including, but not exclusively, Open Plant.

Furthermore, this project will allow us to directly collaborate and exchange knowledge with the computational proteomics unit in Cambridge. We expect to re-evaluate our current approaches, reinforce or develop alternatives in following areas:

1) Our main protemics workflow: Mascot Daemon - Mascot - Scaffold

2) Integrating visualization in our current work-flows in label based or label free experiments

3) Quality control protocol to benchmark and troubleshoot LC-MS/MS systems

4) Execute large dataset processing on a computer cluster

5) We expect more ideas will rise from the proposed in-depth training

Sponsor for the research and cost centre:

Govind Chandra (govind.chandra@jic.ac.uk)

John Innes Centre, Cost centre: CSTR1-B09-S

I confirm that I have the full support of the sponsor listed above and that they can be added to the OpenPlant Fund mailing list to receive project updates (to which they can unsubscribe at any time).

Budget:

Cloud computing costs and workflows development	£1500
Notebook	£600
Travelling and accommodation expenses	£800
Catering	£200