Babraham Bioinformatics

Proteomics Exercises

Version 2025-03

Licence

Exercise 1: Obtaining Raw Data

In this exercise we're going to look for the data for a particular study which will be the data we're going to use for the subsequent exercises.

We're going to find both the data and the reference sequences we'd need to be able to analyse this study.

Finding Data in Proteome Central

Go to https://proteomecentral.proteomexchange.org/ in your browser.

We want to find a yeast (Saccharomyces cerevisiae) study performing label free quantitation.

From the full list of studies you can use the links on the left to limit your view to only yeast studies

How many yeast datasets are present in the repository?

Use the search box to search for eif2a – this should now restrict your results to a single study (PXD043985) which is the dataset we are going to use.

Click on the accession code to go through to the details page for the study. Check that you can find the following information:

- When was this data submitted?
- What sort of mass spec generated the data?
- Does this data include measurements of post-translational modifications?
- Which underlying database does this data come from?

You will see that the data in proteome central doesn't give any details of the samples deposited, nor the files which are available. For those you'll need to move through to the actual database in which the data resides.

Collecting data from PRIDE

Use the link on proteome central to move through to the PRIDE entry for this data (https://www.ebi.ac.uk/pride/archive/projects/PXD043985). Read through the details of the submission to see what samples you expect to see.

Have a look and see how the data was originally processed and analysed. Which program was used for the searches?

Look at the list of files available to download. Can you identify the original raw data, and the quantified results? How many samples are there, and can you match these to those described in the text at the top? In this study there is an "experimentalDesignTemplate.txt" file, which is not a required file. Have a look at the contents of this file to see if that helps in interpreting which file is which.

If you wanted more details about the study then can you find the paper which describes this study.

Finding Uniprot Reference Proteomes

If we wanted to re-process this data then we will need a reference proteome against which to search. In the project description can you see an exact description of which sequences were originally used to process the data?

We are going to find a suitable reference sequence set in Uniprot so that we can reprocess the data.

Go to https://www.uniprot.org/proteomes/ in your browser.

See if you can find the Saccharomyces cerevisiae proteome.

- How many proteins are in this proteome?
- How does this compare to the number of genes?

You will see that you can download the full proteome from

https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000002311/UP000002311_559292.fasta.gz

Exercise 2: Quality Control Reports

Running PSM level QC in R

Work your way through the `quality_control.Rmd` file in R to see how the various QC plots are constructed and ensure that you understand what each one is showing and how to interpret it.

[If you have time] PTXQC Reports

If you visit:

https://www.bioinformatics.babraham.ac.uk/training/proteomics/Proteomics_Course_Data/QC%20Reports/

You will see links to a number of PTXQC reports which you can examine.

- **ptxqc_demo_report.html** is the demo file which comes with the program
- **ptxqc_yeast_report.html** is from PXD043985 and compares a full proteome to an affinity pulldown with eIF2a
- **ptxqc_worm_report.html** is from PXD025694 and compares WT worms to cey3 KO worms
- **ptxqc_mouse_report.html** is from PXD003155 and compares WT and ATXN2 knockdown animals

For the mouse sample we also have ptxqc_mouse_relaxed_report.html which is a reprocessed version of the same file following identification of a problem in the original processing.

You can have a look at all of these datasets to see if there are any issues you might want to talk to whoever created the data about, or take into account during your analysis.

Demo Report

This is an example report which the authors of the program distribute. You can look through the different plots which are provided and read the information associated with each of them. Some of the plots show properties which are very specific to the way the mass spec was run to create the data, but others show the properties of the samples themselves, or the output from the search results.

Yeast Report

There are a few things to note on this report. You will see that the biggest flags on the QC are due to the low total number of peptides and proteins detected. The expected value here was probably set on a more complex proteome such as human or mouse. The lower number of data points in this data is going to largely be driven by the fact that this is a yeast, with a much less complex proteome.

You will also see in some of the plots that there are two distinct sets of samples. One has much lower complexity, signal intensity and higher contamination that the other. Given the design of the experiment can you see why this could occur?

Worm Report

The worm data again shows a somewhat lower amount of data than the QC program would like, but this will also be because of the lower proteome complexity in worm.

The samples here are much more homogeneous than the yeast, but you should see that one sample is behaving somewhat differently to the others. This would probably be something you'd want to note in your later analysis in case any changes you saw were being driven by the oddly behaving sample.

Mouse Report

In the mouse report we see problems with the alignment of the LC timings in the different runs. Look at the alignment plots and see if the green line appears to run through the centre of the flat portion of the run. If the difference in retention times is too large then maxquant may not have found the correct offset for the run.

In the relaxed report we reran the analysis but allowed 30mins difference in retention time, rather than the default 10. Have a look at the equivalent plots in the relaxed report and see if allowing the additional flexibility helped. Do you think there might still be problems with these samples?

Exercise 3: Visualising and Exploring

Introduction

In this exercise you will work on the protein level summarisation of the data. You will repeat some of the QC from the PSM level data, but only some of the metrics are readily available to you at the protein level.

You will look at the quantitation of the data and the possible normalisation.

You will look at data imputation options and compare the number of missing values between samples.

You will calculate condition level summaries of the data and plot out scatterplots to compare these to get an idea of what is actually changing in the data.

Exercise

Work your way through the `protein_exploration.Rmd` file to go through all of the steps described above. You can plot additional graphs for some of the comparisons not directly shown as well as the ones which are included already.

The aim of this section is to ensure you are happy with the quantitation of your data, and that you understand the relationships between your samples. You should also have a good impression of what is changing between your samples.

Exercise 4: Statistical Analysis

Introduction

In this final part of the analysis of this data you will look at different statistical options for comparing the 4 conditions in this study. We'll look at using ANOVA, T-Tests and LIMMA to find proteins which are differentially enriched between different conditions.

We will also look at different ways of visulising the results of the statistical tests.

Exercise

Work your way through the `statistical_analysis.Rmd` file to run statistics on the data exported from the visualising and exploring section.

Look at the various tables and plots to ensure you understand what the test is checking for, and how trustworthy you think the results are.