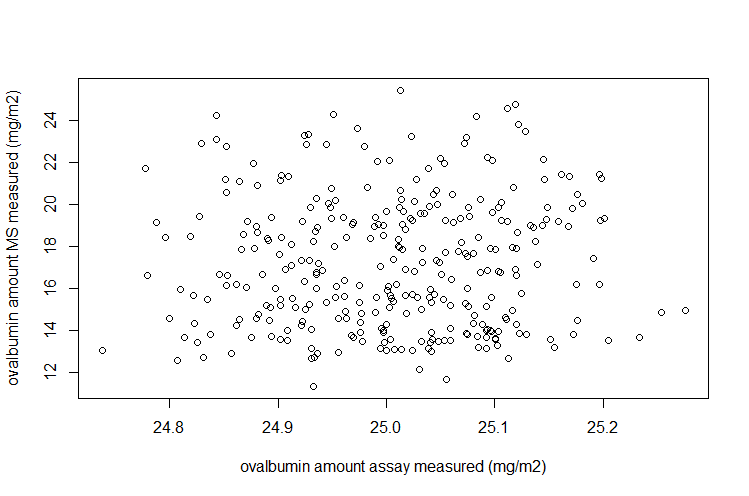
**Manual for DP14 Leaf Proteomics Project data processing pipeline (euc data)**

04/011/2016

Overview:

This document describes the pipeline for processing proteomics data for the *Eucalyptus* (inc. *Angophora* and *Corymbia*) species of the Leaf Proteomics Project. The pipeline involves the following sequential steps:

1. Calculate protein areas
   1. Load and stitch together SWATH data
   2. Filter SWATH data by false discovery rate (FDR): for a given peptide, there must be minimum of 3 samples with FDR < 0.01
   3. Removed proteins with reverse sequences from SWATH data (10 proteins, denoted by ‘RRRRR’ prefix on protein accession)
   4. Calculate protein areas using top2top3 method
      1. For a given peptide, peptide area = the sum of the top two most intense ions
      2. Then, for a given protein, protein area = the average of the three peptides with the greatest area
2. Calculate protein amounts using the ovalbumin peptide ‘GGLEP’ as a standard
   1. Find GGLEP top2 areas for each sample
   2. Find protein areas relative to GGLEP area
   3. Multiply by 5.64x10^-11 to get moles per cm2 (2.5e-6g / MW of ovalbumin)
   4. Multiply by protein molecular weight to get g/cm2
      1. Molecular weight is calculated using amino acid sequences using the R package *peptides*
   5. multiply by 10^07 to get mg/m2
   6. check MS derived values against protein assay values



1. Generate MAPMAN/Mercator functional annotations
   1. Constrain Egrandis\_Eglobulus\_chloroplast\_Myrtales\_At\_mt\_cRAP\_160405 protein sequence data to only include proteins reliably identified by MS.
   2. Convert to fastA format and upload to *http://www.plabipd.de/portal/mercator-sequence-annotation?p\_p\_id=Mercator\_WAR\_Mercatorportlet&p\_p\_lifecycle=0&p\_p\_state=maximized&p\_p\_mode=view*
   3. Download MAPMAN/Mercator annotation scheme
2. Associate proteins in SWATH data with functional annotations
   1. Protein amounts for functional groups of interest are created by summing the protein amounts of all proteins associated with that functional group
3. Generate outputs
   1. Protein functional group amount dataframes (both absolute quant. amounts and ‘relative’ amounts standardised by total protein amount in the sample) are merged with site-specific and phenotypic data on: average and recent (last year or similar) climatic conditions, modelled average aridity, soil and litter CNP, canopy openness, LMA and leaf water content, leaf CNP, leaf age, photosynthesis data (Amax).
   2. knitr reports can be generated which show scatterplots of relationships between predictors of interest and protein amounts for functional categories of interest

Structure:

The data processing code is split into a number of files, some of which source / depend on each other, and which must be run in the correct order to produce the desired output.

|  |  |  |
| --- | --- | --- |
| *Script file* | *Dependencies* | *Description* |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |