**Project 4 of BIO316 – Metabolomics Data Analysis**

**Instructions:**

1. This assignment comprises 25% of the total mark for the BIO316 module (5 credits).
2. If you have any questions regarding the techniques, please refer to the practical slides. Everything required in this assignment has been fully covered with detailed procedures.
3. A screen shot of the key relevant section of the result is enough. Do NOT copy the entire result to your report. Write a maximum of 200 words of texts to interpret your results for each question.
4. Clearly label each question (and sub-question) in your report. Submit your assignment report in Word format directly to ICE before the deadline; late submissions will be penalized according to university policy.

**Problem Scenario**

The example used is compound concentration data obtained by targeted (i.e. quantitative) metabolic profiling of 1H NMR spectra of urine samples collected from 57 cancer patients. There are two groups of patients – **Cachexia (Y)** refers to the group with significant skeletal muscle loss; **Cachexia (N)** refers to the group with no obvious skeletal muscle loss. Cachexia is defined as the loss of weight, muscle atrophy, fatigue, weakness and significant loss of appetite in someone who is not actively trying to lose weight. Cachexia is often seen in end-stage cancer, and in that context is called "cancer cachexia". The exact mechanism behind cachexia is poorly understood, but there is probably a role for inflammatory cytokines, such as tumor necrosis factor-alpha (TNF-α) - which is also nicknamed cachexin, Interferon gamma (IFNγ), and Interleukin 6 (IL-6). The goal here is, by using methods provided in MetaboAnalyst, to identify metabolites that are significantly different between these two groups of cancer patients (cachexic vs. non-cachexic). These metabolites could serve as potential early-stage biomarkers for detecting cachexia and for exploring its underlying metabolic basis. The data file is ready on ICE.

**Task 1:** Data processing and normalization (15 points).

1. Check the data format and upload the data to MataboAnalyst. Check the data integrity (this step will be launched automatically) and briefly describe your findings. if missing values are detected, replace the zero values with a small positive value (half of the minimum positive number detected in the data), otherwise, skip this step and proceed to the normalization step (5 points).
2. Normalize to a reference sample (refer to the lab notes for the general rule of reference sample selection) and choose “Log normalization” for data transformation. Compare the distribution of normalized and original concentrations (5 points). Briefly describe the outcome and the purpose of data normalization (5 points).

**Task 2:** Identification of significantly different metabolites (50 points).

1. Use a Volcano Plot to compare the size of the fold change and statistical significance level of metabolites between the two sample groups. Use a fold change threshold 1.5 and adjusted P-value threshold 0.1. Save your results and briefly interpret your findings (10 points).
2. Perform Partial-Least Squares Discriminant Analysis (PLS-DA) as a more sophisticated way of detecting more candidates. Save your results and interpret your findings (10 points). Compare the difference between the Volcano Plot method and PLS-DA (10 points).
3. Try another approach, Significance Analysis of Microarray (SAM), to select significant features that distinguish between cachexic and non-cachexic patients (10 points). Briefly describe how SAM works (10 points).

**Task 3:** Consensus results and functional analysis (15 points).

1. Find metabolite candidates consistently identified by all the three approaches and interpret the results (10 points).
2. For the consensus results, identify their biological function(s) and possible roles in pathways using MetaboAnalyst Pathway Analysis (<https://www.metaboanalyst.ca/faces/upload/PathUploadView.xhtml> (10 points).