From multiple imputation to Bayesian framework in quantitative proteomics

Marie CHION
MAP5, UMR 8145, CNRS-Université Paris Cité, Paris France

Quantitative proteomics using liquid chromatography-mass spectrometry can identify and quantify several thousand proteins in a few hours of analysis. In particular, differential proteomics analysis compares the measured intensity of proteins between different conditions to determine those whose abundance varies "significantly". A particularity of these large datasets analysed is that they contain missing values between 5 and 15%.

One way of dealing with the problem of missing values is to impute them, i.e. to replace them with a value defined by the user or an algorithm. Thus, multiple imputation allows the imputation process to be iterated several times to obtain several complete data sets. These are then combined before applying conventional statistical tools. However, standard software for statistical analysis of proteomics data uses the average complete dataset and ignores the uncertainty induced by the random imputation process.

Therefore, we present a rigorous method of multiple imputation using Rubin's rules and a variant of the t-moderate test that accounts for the variability arising from both the original dataset and the multiple imputation process. Since the t-moderate test is based on a Bayesian hierarchical model, we also propose a fully Bayesian framework for differential proteomics analysis and discuss the place of multiple imputation in such a framework.