Missing Thinking in Shotgun Proteomics

One of Murphy’s Laws of Proteomics is that your favorite protein is never in the results list. A corollary is that the protein you most want to quantify will have the most missing data. Life is a bitch. You need to put John Prine’s debut album (John Prine, 1971, Atlantic) on the turntable and get some perspective.

Missing data in shotgun proteomics is a like trying to cross a well-used cow pasture. You really want to pay attention to where you step. The thinking about this problem does not start with how to impute values for the missing data. It needs to start with what are the measurement characteristics of the proteins or peptides with missing data.

This topic is about quantitative shotgun proteomics data. We do not need to think about all of the quantitative methods no one uses. What we do at this moment in time are mostly three types of methods: label-free MS1-feature-based quantification, isobaric labeling, and (the new kid on the block) label-free DIA. I only have detailed knowledge of TMT experiments (and spectral counting). I have some passing knowledge of the other two methods. We need some big picture understanding of each of these methods to understand what missing data means.

In a basic label-free quantitative study, we need high-quality chromatography and high-quality ion current measurements. What do I mean by high quality? We need liquid chromatography where the same peptides elute at very similar (and stable) retention times run-to-run. We need gradient conditions and sample loads to make sure that we have sufficient separation and appropriate peak shapes and widths to be able to quantify our proteins of interest (actually, it is the peptides from our proteins of interest). In many workflows, we will want to match the retention times across the runs. We can find a set of features common to all runs, and match those, or we can add RT standards and match those. We typically want to do this so that we can use identifications acquired from all runs in a comprehensive way (often called match between runs). Spectral libraries are also used, and retention time matching is important in addition to m/z matching.

High quality ion current measurements mean that the mass spectrometer instrument is operating in a narrow acceptable performance range. You can think of instrument performance as a sawtooth where we start at some best place (instrument clean and calibrated). Chromatography elements are also a factor in getting to this best place. Once you start at the best place, it is all downhill from there. As the instrument is used, measured intensities will decrease over time. Typically, some standards are interleaved with samples. We know what intensities to expect from the standard and can see when those have decreased to a point where we need to clean and recalibrate the instrument. Lather, rinse, repeat.

Needing to keep the instrument in an acceptable performance range, instruments having ranges of linear response, instruments having a lower limit of detection, and missing data are all intimately related.

* The poor runs determine the overall
* Zero replacement versus imputation
* Linear models can ignore NAs instead of imputing something
* Orbitraps hide the noise level – makes it hard to estimate