Normalization in Scaffold Q+

Intensity-Based Normalization

Pre-processing

The raw intensity data is acquired from the spectra and purity-corrected as appropriate. The data is transformed by applying a logarithm (base 2). Within each MS-Sample a "missing value" is assigned as the larger between (a) the minimum positive logged intensity acquired and (b) the value whose z-score is -4 for the distribution of all logged values in the MS-Sample. These missing values are applied to all intensities which either had raw value zero, or fall below the "Minimum Dynamic Range." If the option is selected, spectra with missing values in the reference channel are removed.

Removal of spectra from quantitation: A spectrum can be removed for any of the following reasons

1. In Label-Free we only pick one spectrum to represent the precursor.
2. It has no quantitative data.
3. It is non-exclusive.
4. If a Spectrum Quality Filter has been applied. Currently we have one called Reference Value Required which removes spectra after normalization if their reference value is marked as a missing value.

Iterative Median Polish

A version of Tukey's median polish is applied iteratively to normalized the data. In what follows, all computations are on the logged data, and "average" denotes either median or mean depending on the mode the user has selected. (Median is the default mode.) The normalization has four steps:

1. *Inter-Sample Normalization*: A normalization factor consisting of the global average minus the within-MS-Sample average is added to each data point.
2. *Intra-Sample Normalization*: A normalization factor consisting of the within-MS-Sample average minus the within-channel average is added to each data point.

3. *Peptide/Spectrum Normalization*: For each protein, the averages across the values in each spectrum are brought into alignment by adding a per-spectrum normalization factor (average of these averages minus the particular spectrum's average).
4. *Intensity Weighting*: A weight is assigned to each spectrum based on a t-statistic derived from percent deviations from channel averages.

Steps 1-4 are repeated three times.

Post-processing

A standard deviation estimate is derived for each spectrum, based on smoothed within-protein deviations of spectral values from averages binned by total intensity. The weight of each spectrum is divided by the total number of spectra matched to its peptide, providing a form of intermediate peptide-level averaging when subsequently computing protein-level values.

What are the details of the normalization calculations?

For inter-sample normalization:

You have a bunch of raw intensity values… stored as multiplexes (eg, 4 channels values per spectrum for iTRAQ-4) and these spectra are in MS samples. Note that we are in log space, so they're really log2(int) values. Every one of these log2(int) values is adjusted by adding **(average value over all data points in any MS sample in any channel)** -[minus] **(average over all data points in the same MS sample in any channel in that MS sample)**.

For intra-sample normalization:

Every one of these log2(int) values is adjusted by adding **(average value over all data points in the same MS sample in any channel)** -[minus] **(average over all data points in the same MS sample in the same channel)**

*and average means weighted mean/median.

And since logb(x) = C * log2(x) where C = 1/log2(b)

…all log2(int) would just change by a constant