Protein ratio calculation and significance testing

Peptide/protein ratio calculation

Relative protein quantitation and assessment of statistical significance of the observed expression changes are derived from log-ratios of peptides unambiguously associated with the protein. For calculating peptide log-ratio, the average of log-transformed intensities in all samples belonging to the same experimental group is computed. Value obtained from controls is then subtracted from value from experimental samples. Protein/cluster ratio reported by Diffprot is calculated as a median of peptide ratios.

Test statistics

The statistical significance of protein ratios is estimated using a resampling procedure. The test statistics computes median ratio for a protein/cluster, giving a penalty for inconsistency of peptide intensities inside experimental group, adjusted in order to account for the fact that quantification is less precise for low intensity peaks. The test statistics is defined as:

$$s = median\left(\frac{i(p,s_1) - i(p,s_2)}{d(p)} \; s_1 \in E, s_2 \in C, p \in P\right)$$
$$- median\left(\frac{\left(\left(i(p,s_1) - i(p,s_2)\right)\right)}{d(p)} \; s_{1,} s_2 \in C, s_1 \neq s_{2,} \; p \in P\right)$$

where *i(p, s)* denotes logarithm of peptide *p* intensity in sample *s, E* –experimental group, *C* – control group, P – all peptides from tested protein, d(p) – median intensity difference for peptides with similar intensity levels. The estimation of the d(p) as a function of peptide intensity is precomputed according to the following procedure: first all peptides are divided into 100 quantiles according to their median intensity, and next median of absolute intensity difference inside one sample group is then calculated for peptides within each quantile and smoothed using LOWESS.

In order to estimate the null distribution, test statistic values are calculated for a large number of peptide sets, with sizes equal to the size of the tested protein, chosen at random from the whole dataset. These sets may be considered as "random proteins". As the null distribution depends on number of peptides in a protein and amount of missing data, it has to be calculated separately for each protein. Given the test statistics value for the current protein r_0 and a set of values r_i for N

permutations the p-value is defined as:

$$p = \frac{\sum_{i=1}^{N} I(r_i \geq r_o)}{N}$$

where *I(•)* is the indicator function equal 1 if the condition in the parenthesis true, and 0 otherwise. The resulting p-values, are adjusted for multiple hypothesis testing using a procedure that controls the false discovery rate (FDR).

Proteins present in only one experimental group (i.e. expressed uniquely in controls or experimental samples) are also taken into consideration during the computation. They are not assigned with protein ratio, just an “up” or “down” flag, depending on the analytical group they are absent in.