Ref: OT-37

Title:

Statistical framework for the analysis of high-throughput quantitative proteomics data.

Summary:

Researchers from CNIC have developed and validated a general statistical framework, the WSPP model, for the analysis of quantitative proteomics results, that very accurately describes the technical variability of data for a representative set that includes the most common Stable Isotope Labeling (SIL) methods. In addition, the model allows a systematic comparison and integration of data from different experiments.

The WSPP model separately considers the variances produced during (i) protein extraction and manipulation, (ii) peptide generation from their corresponding proteins and labeling, and (iii) generation of quantitative information from the mass spectra. Each of these steps are analyzed independently, providing for the first time the required flexibility to account for the different nature of error sources associated to the different methods. The general validity of the model was demonstrated by confronting 48 experimental distributions against 18 different null hypotheses.

The WSPP model provides the first integrated standard of general validity for the analysis of quantitative proteomics data obtained by stable isotope labeling.

Innovative aspects:

Mass spectrometry (MS)-based proteomics allow the identification and relative quantification of thousands of proteins in a single study. However, existing models are highly specific to each SIL method and mass spectrometer, making them unsuitable for examining data from different laboratories, judging experimental quality on the basis of unified criteria, handling, comparing, and integrating multiple measurements, or interpreting the complete set of experimental results from different SIL approaches as a whole. Moreover, most models and statistical significance tests are based on normality assumptions that have not been tested despite the fact that heterogeneity of variance has been documented in all SIL methods. These techniques are based on peptide-centric measurements, and the lack of general models leads to the subjective choice of a method for combining multiple peptide readings to estimate protein ratios. This problem is further aggravated by the undersampling that characterizes SIL-based MS analysis: the number of peptides that quantify a protein is variable and cannot be controlled between experiments, a nontrivial fact that complicates mathematical modeling.

The WSPP model is the first generally applicable statistical framework for the analysis of data generated with SIL-MS technologies, provides a detailed description of the behavior of technical variance, and by analyzing it independently at the spectrum, peptide, and protein levels, the model is able to capture separately the specific error sources of each SIL and MS method, demonstrating that error distributions are accurately modeled in all cases at the three levels.

No models have been formulated previously for quantitative proteomics data that decompose technical variance into two or more components. Besides, in WSPP model, the averages are calculated following error propagation theory so that the statistical weight with which each value contributes to the average is exactly the inverse of its local variance, and the variances of each one of the averaged values are known with accuracy. In the WSPP model, all of the elements, even single hits, are assigned a local variance, and this is done on the basis of only four parameters that are estimated from the analysis of the whole collection of data.

This statistical model also provides for the first time a framework to compare and integrate results obtained using different quantitative approaches in different kinds of mass spectrometers.

Competitive advantages:

The model generates the integration at each level taking into account the separate variances according to error propagation theory so that the specific variance of each value at any level is accurately estimated. The model efficiently resolves the under-sampling problem, providing a framework to analyze the data at each level using unique normal distributions. In addition, the statistical framework also allows comparing and integrating results obtained using different SIL techniques so that full control over variance is maintained in the integrated data, opening the possibility of making further integrations at upper levels.

This statistical framework efficiently resolves the problems of variance heterogeneity, data integration, and under-sampling and provides a statistically sound method for testing the quality of quantitative experiments and detecting experimental deviations.

This statistical algorithm is the first one that has been demonstrated to be of general validity for a wide range of different SIL and MS approaches. It also resolves the problem of undersampling and at the same time provides a framework to integrate data and a robust algorithm to estimate variances, which are of general applicability.

Key words: Quantitative proteomics, stable isotope labeling, statistical analysis

Contact:
Technology Transfer Office – TTO
CNIC
proyectos_otri@cnic.es