Review on Hill and Oberg’s papers

A statistical model for iTRAQ data analysis by Hill

The author demonstrate the analysis of variance approach that assess the differential protein expression in a single model fit to the collection of reporter ion peak areas from all observed tadem mass spectra. Their model allows the analysis of data from multiple MudPIT runs.

They define the expected protein expression profile for a given condition as the collection of proteins and their corresponding amounts in a representative sample from population of interest. Furthermore, they define the complete proteome as the st of all proteins encoded in an organism’s genome, and the protein profile as the list of numbers giving the concentrations for each protein in that proteome. For the proteome containing I proteins, we express the concentration of the ith protein as Pi and the entire set of numbers giving the expression levels of each of those proteins as Pi. Let Ri.c represent the relative mount of protein i comparing condition c to the control condition. Rc denotes the condition or treatment type. Dki denote the observed amount og protein I relative to its expected amount for subject k.

The author proposed, since the iTRAQ measurements are med at the peptide level and therefore, the model must reflect the relationships between peptide and associated protein expression levels. First issue is the some tryptic peptides can be associated with more than one proteins, but these peptides are usually get eliminated during the database searching with the Protien-pilot software. However, since the Protein-Pilot software is processing each run separately, hence the issue of non-unique peptide may still occur when performing the analysis with several runs.

Post-translational modification and splice variants can affect individual peptide within a protein, hence the model that author proposed also captured the effect at the peptide level in addition to the condition specific protein levels effects discussed previously. The effect of the peptide is denoted by Fji and Gjic be the ratio of the expected amount of peptide j comparing to condition c to the control condition.

Moreover, the model is aimed to analyse the data from the multiple MudPIT runs. The loading, labelling and mixing difference across different iTRAQ experiments can be denoted by Vq,l where q denotes the tag and l denote the MudPIT runs. Additional term Bq is included for qth iTRAQ experiment as a constant proportion in the peptide profile to yield the expected reporter ion area profiles. All the other un-captured experimental error is denoted by Hikcqsl.

To complete the model, the author introduced a constant (or intercept) u.

Y is the response of log reporter ion peak area corresponding to protein I, peptide j condidiotn c, run q tag l and spectrum s. The model also contains explanatory or predictor variables. For this particular case, the term of interest is the ri,c which informs the differentially abundant proteins, the remaining terms can be thought of as normalisation terms.

The ratio estimation formula was also given in the Hill’s paper, which also included the condition (i.e. treatment) effects, protein by condition and peptide by condition. The author also included the equation for computing the 95% confidence interval. Note the ratios were in log scale, hence transformation is required. However, the author did not mention any methods on determining differentially abundant proteins are statistically significant, i.e. p-value.

Despite they used the ANOVA model to fit iTRAQ data, they did not present an ANOVA table which can help to illustrate how the data is decomposed and analysis is achieved.

Oberg’s paper is talking about the normalisation methods based on the same model that was described by Hill et al. One method that was proposed is called the iterative regression, this method is to iteratively estimate the run, tag, protein and peptide effects, and then adjust the raw abundance levels with these estimates by subtraction.

The author stated that sub-setting the data to protein-by-protein is not valid, because run effects should be global effects and should not be re-estimated for each protein. Moreover, it is not sensible that the run effect is high for some proteins but low for some proteins. However, we proposed that data should be fitted protein-by-protein, because every protein behaves differently, i.e. different variances for different proteins, hence it is invalid to test all of these proteins under the same variance, i.e. global variance.

In Oberg’s method did mention the testing the differentially abundant proteins for statistical significance. The author also mentioned briefly importance of multiple testing, which suggests the test is based on proteins-by-proteins despite the model fits every protein at the same time.