NORMALIZATION FOR PROTEOMICS DATA

Terry M. Therneau, Ann L. Oberg, Douglas W. Mahoney, Jeanette E. Eckel-Passow, Christopher J. Malone, H. Robert Bergen, III

Mayo Clinic

Shotgun proteomics attempts to identify all of the proteins in a biological sample, along with their amounts. However, if the resultant data is cast into a matrix (in the manner of microarray data) with samples in one dimension and proteins or peptides in the other, 80-95% of the cells will be missing. This appears to be due, at least in part, to the geometric nature of the proteome: if there are n proteins at some concentration x, there may be 2-3 n at concentration x/2. This means that no matter what the resolution of the hardware, over 1/2 of the compounds that we can detect will be at limit of sensitivity and hence random in their appearance. This abundance of missing values is fatal to simplistic methods of normalization.

It has been previously shown that normalization for microarrays can be cast as a regression or ANOVA model. Typically, the data from microarray experiments are balanced with regards to the normalizing factors. This balance lends nicely to using marginal means to estimate the regression parameters, significantly reducing the computational challenge of high dimensional data. Proteomic data on the other hand is highly unbalanced with regards to normalization factors due to the missing data. We have found that the regression approach can nevertheless provide accurate estimates of normalization parameters. Classical methods for iterative solution of linear models, such as the Gauss-Seidel algorithm, are used to deal with the large dimension of the data.

An increasingly common data analysis method in proteomics is spectral counting, where the actual intensities for each observed peptide are ignored and only a simple count of the number of times the peptide is detected is retained. For this type of data, it is natural to impute a zero count to every missing value. However, this creates a data matrix with far more zeros than the standard Poisson assumption. In this situation, we find that by using a zero-inflated Poisson model to deal with the surfeit of zeros, we are able to adequately model the normalization factors and provide better standard errors for treatment contrasts than more simple summaries for the counts.