Iterative Linear Regression by Sector: Renormalization of cDNA Microarray Data and Cluster Analysis Weighted by Cross Homology.

David Finkelstein
Carnegie Institution of Washington
260 Panam Street Stanford, CA 94305
USA
650 325 1521 ext 341
finkel@genome.stanford.edu
CAMDA00 Dataset 1: Yeast

Finkelstein David B., Gollub Jeremy, Ewing Rob, Sterky Fredrik, Somerville Shauna, Cherry Michael J

Empirical evidence and observations validated by statistical tests have indicated that several distinct types of consistent measurement error can alter the interpretation of cDNA microarray data. Whenever possible models of error are derived during quality assessment and applied during data analysis. When measurement error is detectable and conforms to a defined model, corrections can be applied during renormalization. However, some measurement errors are detectable but less well defined. In such cases, parallel analyses are required to determine the significance of such effects. Furthermore, supporting biological evidence from a distinct method designed to detect the problem may be required. In the specific case of the Spellman data both well-defined problems and ambiguous problems were examined. First, the clearly detectable and definable measurement errors are corrected through renormalization. Reanalysis of the Spellman and Sherlock cell cycle data set begins with a new method of normalization that more accurately reduces the effects of outliers and spatial variation on the arrays. First, all intensity values are log transformed, then linear regression is performed separately on each sector. These sectors were produced by slotted printing pins. The Spellman data has four sectors and was printed with four distinct pins. Then these residuals are calculated for these four regression lines; one for each sector. Outliers (those residuals where |e| > 2 x std dev of e) are removed and the four regression functions are recalculated. If the difference between the value of r-squared of the new regression line is less than .001 of the old, then no further residuals are removed. Else, outliers are removed by the same test as above and the iterations continue. Once completely determined, the slope and intercept values are applied as correction factors to the log transformed channel 2 values. The result is that the function of log channel 1 and log channel 2 closely approximates y = x. Then these values are exponentiated, a new ratio is calculated and this ratio is put on the familiar log base2 scale. This renormalization alone has been demonstrated to substantially reduce the standard deviation of log2 ratios. Next, the ambiguous task of detecting the effect of cross-hybridization was examined. The yeast genome is fully sequenced, thus the sequences of PCR fragments were known. Therefore it is possible, with some error, to determine the likely number of transcripts that could cross-hybridize to a given PCR fragment. The correlation between the likelihood of cross-hybridization and the frequency of transcripts with cross-homology is difficult to assess without empirical evidence. It is important to note that modeling the molecular events during hybridization has proven difficult. Therefore, no analysis can be used to correct data. However, a technique can be applied as an informed post hoc method. In this way, such analysis may indicate where biological confirmation experiments are warranted, rather than supply a mathematical solution.

Keywords

normalization, linear regression, spatial bias, residual

Tools

web-based Perl Program available for data loaded on SMD. Algorithm fully detailed in pdf file.
Send requests for Perl code to jgollub@genome.stanford.edu

website

http://afgc.stanford.edu/afgc_html/site2Stat.htm