

Noise and Shadow: Statistical Methods for SAGE Data

Natalie Blades

Department of Biostatistics, Johns Hopkins University

Serial analysis of gene expression (SAGE) is a technique for obtaining information about gene expression. SAGE experiments provide insight into human disease by identifying disease-related genes and by suggesting possible therapeutic targets. Data from a SAGE experiment consist of long lists of gene identifiers and corresponding frequencies. A dominant proportion of these lists consist of genes which appear only a few times. Some of the low frequency tags represent low frequency mRNAs, but some are the result of sequencing errors. It is difficult to distinguish between these two cases.

We propose two methods for automatically adjusting for error in the observed counts: The first approach filters the data by exploiting the remarkable regularity in the frequency distributions of tags arising from these experiments. A statistical model is proposed to automatically discount low counts that cannot reliably be used for comparison of expression levels across conditions for a specific gene and to transform the cell counts to a scale that produces more reliable correlation and clustering of genome-wide expression profiles. Simulation studies indicate the proposed method presents a considerable advantage in identification of differentially expressed genes. The second approach considers the similarity in the sequences of frequently and infrequently occurring genes as a method for evaluating the integrity of an observation. A sequence error in just one base pair of a very common tag may result in the creation of a completely new, but similar, tag. Thus, low frequency tags which are very similar to other observed tags may just be shadows of those highly expressed genes. We attempt to reduce the effect of these shadows while not damping those tags which have no close neighbors, that is, those tags who differ by many base pairs from all other tags in the library.