Differential Expression in Microarray Data: Comparison of various approaches

Course Project for STAT675 (Rice University) or GS010103 (GSBS)

Objective: One of the primary aims of microarray experiments is detecting differentially expressed genes under multiple conditions. This project aims at comparing various approaches: Frequentist, Empirical Bayes and Bayesian to identify genes/features that may be differentially expressed in the two (or more) groups. There are two main parts:

(a) Identification of differentially expressed genes

(b) A evaluation of the stability of the list say via False Discover Rates (FDR), some measure of significance values, posterior probabilities etc.

You can use the following steps as guidelines for the project.

1. Pick a microarray dataset(s) of your choice. There are a lot of publicly available datasets. Some examples include the Breast Cancer dataset of Hedenfalk et al (2001), Colon Cancer dataset of Alon et al (1999) and the Leukemia dataset of Golub et al(1999). These are now textbook datasets and have been analyzed a lot in literature. If you have your own data thats fine too, but it should fall under the *large p small n* paradigm.

2. Choose your favorite frequentist method(s). Suggested software include include SAM, SNOMAD etc. A good reference is Parmigiani et al (2003) which provides an outline of the methods along with links to the softwares. Also see Dr. Hu's project for this course.

3. Another popular method is Empirical Bayes method of Newton et al (2004). See `http://www.stat.wisc.edu/~newton/research/arrays.html` for details about the paper and software.

4. The Bayesian methods you could consider are BAM and BRIDGE. The relevant websites for links to papers and softwares are `http://www.bamarray.com` and

   `http://www.stat.ubc.ca/~raph/Software/BiocRPackages/index.html`.

5. For Aim (b) of the project use the following. Suppose $(x_{i1}, x_{i2}, \ldots, x_{in})$ are n replicate measurements of differential expression levels for gene i, for $i = 1, \ldots, p$.

   (1) Suppose x_{ij} are i.i.d $N(\mu_i, 1)$ and we want to test $H_{0i} : \mu_i = 0$ against $H_{1i} : \mu_i \neq 0$ for each gene i. Perform a frequentist test for each gene and obtain the corresponding p-values. (Already done in homework)

   (2) For each value of FDR = 0.01, 0.02, 0.05, 0.2, 0.5, obtain a list of significant genes based on the p-values and the multiple comparison procedure in BH(1995).

   (3) Assuming μ_i are independent $N(0, 100)$ (100 is the variance), obtain the posterior distribution of μ_i and compute the posterior probability of $|\mu_i| > 1$. (Already done in homework)

   (4) Apply the Bayesian FDR procedure described in Newton et al. (2004) based on the posterior probabilties obtained in (3). For each value of FDR = 0.01, 0.02, 0.05, 0.2, 0.5, obtain a list of significant genes. Are these lists similar to the ones in (2)?

   (5) According to the Gamma/Gamma/Bernoulli model in Newton (2001), how would you apply a Normal/Normal/Bernoulli model here? What will be the probabilities used to test $H_{0i} : \mu_i = 0$ against $H_{1i} : \mu_i \neq 0$? (Note, the posterior probability of $\mu_i = 0$ equals 0) (Already done in homework)

   (6) Implement your Normal/Normal/Bernoulli model and apply the Bayesian FDR procedure to obtain lists of significant genes.

6. You should critically evaluate all the methods. Discuss the merits and demerits of the procedures. For bonus points you can also conduct a simulation study to evaluate the performance of the methods under various scenarios.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad Patterns of Gene Expression revealed by Clustering Analysis of Tumor and Normal Colon Tissues probed by Oligonucleotide Arrays. *Proc. Natl. Acad. Sci. USA*, 96, 6745-6750.

Golub T. R., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek., M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caliguiri, M., Bloomfield, C. and Lender, E. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286, 531-537.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A. and Trent J. (2001). Gene Expression Profiles in Hereditary Breast Cancer. *New England Journal of Medicine*, 344, 539-548.