

Integrating Gene Expression Studies Via Statistical Synthesis

Anna Elizabeth Campain¹ Jean Yee Hwa Yang¹
annac@maths.usyd.edu.au

¹ School of Mathematics and Statistics, Sydney Bioinformatics, Centre for Mathematical Biology, University of Sydney, F07 Sydney, NSW 2006, Australia

Abstract

Microarrays are one of the driving forces in biological research. With their explosion in popularity, an increase is also seen in the availability and size of public repositories. These data sets are by no means superannuate, but offer researchers a depth of information if used wisely. We proposed a method to integrate multiple data sets to aid in overcoming major meta-analysis hurdles such as lab-to-lab variation and platform dependent biases. Heuristically if a gene is truly differentially expressed (DE), it should be DE throughout a range of DE measures and experiments. Meta-Differential Expression via Distance Synthesis (mDEDS), is an extension of the DEDS method (Yang et al., 2004), which makes use of permutation tests on the multiple data sets and a plethora of expression measures synthesised through a distance equation to produce a candidate DE list.

Keywords:

meta-analysis, microarray, mDEDS

Motivation

As the demand to skilfully combine data sets grows, meta-analysis tools are being developed. Yet microarray experiments are not hassle free, systematic bias with the ability to mask true biological effects is caused mainly by two issues; the length of sample collection and hybridisation as well as varying array hybridisation protocols. To successfully integrate multiple data sets methods must transcend these biases.

Successful meta-analysis offers two fold advantages, first by increasing the sample size of an experiment the power of detecting an overall treatment effect is increased, secondly through meta-analysis one can assess the variability between studies.

Meta-Analysis

Meta-analysis can be defined as a synthesis, or at times review of results from data sets that are independent, but related. Some meta-analysis methods are simple variants of common classical statistical methods, others offer a sophisticated response to the issues faced specifically in the microarray environment. Appreciable effort and literature has been dedicated to these two issues.

The designed method mDEDS incorporates the strength of DEDS by culminating numerous DE measures. Comparisons of these results are made across data sets using permutation tests, and a distance measure, to assess DE significance. A consistent DE list is established despite the bias of data set or platform.

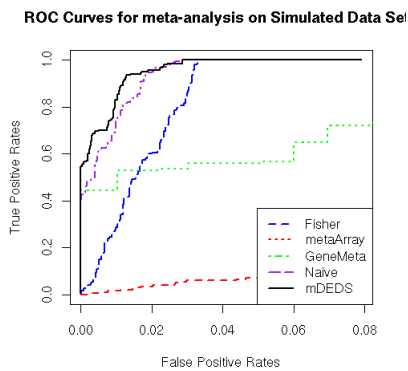


Figure 1: Five ROC curves, for differing meta-analysis methods.

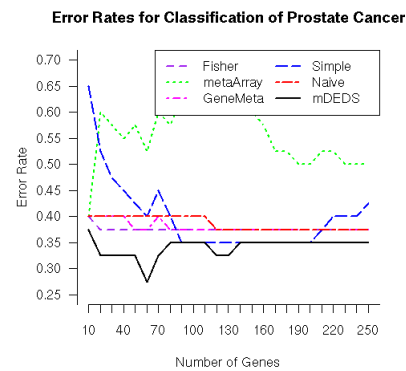


Figure 2: Error rates in the binary classification of three prostate cancer data sets.

Simulation and Case Study

We evaluated the performance of mDEDS using simulations and comparative analyses with prediction accuracy as a criterion. Currently available meta-methods; metaArray (Choi, H. et al., 2007), GeneMeta (Choi, J.K., et al. 2003), Fisher, naive (where a classifier was built purely off one platform) and simple (combined with no adjustment) as well as our proposed mDEDS were utilised within these two studies.

Two data sets were simulated based on the parametrisation of two different publicly available data sets. mDEDS was able to outperform other meta-analysis methods for a known DE list as illustrated with the ROC curves.

Meta-analysis was used to combined three publicly available prostate cancer data sets (Singh, et al. 2002, Varambally et al. 2005, Yu et al. 2004). Performance of meta-analysis methods was based on the construction of a binary classification system, between normal and cancerous prostates.

References

- [1] Choi, H., Shen, R., Chinnaiyan, A.M. and Ghosh, D. A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics* 8:364–384, 2007.
- [2] Choi, J.K., Yu, U., Kim, S. and Yoo, O.J. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 19(1):i84–i90, 2003.
- [3] Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 3(1):204–209, 2002.
- [4] Varambally, S., Yu, J., Laxman, B., Rhodes, D.R., et al. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell* 11(8):393–406, 2005.
- [5] Yang, Y.H., Xiao, Y., and Segal, M.R. Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics* 21(7):1084–1093, 2005.
- [6] Yu, Y.P., Landsittel, D., Jing, L., Nelson, J. et al. Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *Journal of Clinical Oncology* 22(14): 2790 – 2799, 2004.