

# Meta-analysis based on control of False Discovery Rate

Saamyadipta Pyne,<sup>1</sup> Steve Skiena,<sup>2</sup> Bruce Futcher<sup>3</sup>

**Keywords:** Meta-analysis, False Discovery Rate, FDR, P-values, Microarrays

## 1 Introduction

High-throughput technology like microarrays can be used to simultaneously examine thousands of features, such as all the known genes or regulatory regions of an organism, and obtain a relative measure of their expression in a particular experiment. Two common concerns of microarray bioinformatics are to combine with high power the significance values for each feature across a fixed number of independent analogous experiments, and to control the proportion of false positives among those features declared significant.

A variety of existing methods address either issue; in general however, these issues are addressed independently and sequentially. We present a novel algorithm to simultaneously address the two requirements in an inter-dependent way. The input to our method is a  $N \times L$  matrix  $\mathbf{M}$  such that the  $(i, j)^{th}$  entry of  $\mathbf{M}$  is the significance measure, given by a p-value, for  $i^{th}$  feature in  $j^{th}$  experiment. Typically  $N$  is several thousands, while  $L$  is a small constant. We intend to compute a combined p-value for every row of  $\mathbf{M}$ .

While attempting to combine significance from every experiment for a particular feature, it is difficult to guard against false negatives. Even a single sufficiently poor entry, possibly spurious, could skew the combined statistic of an otherwise truly significant feature enough to prevent *any* test of the joint null hypothesis from rejecting it, thereby forcing the test to lose power. To fix this, the ordinary Fisher product was extended in [9] with the Truncated Product Method (TPM) whereby only those p-values which “clear” (i.e., are less than or equal to) some pre-specified cutoff value  $\tau$  contribute to the combined product.

The cutoff ( $\tau$ ) obviously helps to increase the power of the test by letting the worst p-values of a feature to be ignored. It also guards against the case when a combination of semi-significant p-values might spuriously suggest high significance. Given its usefulness, it seems natural to extend the choice of  $\tau$  such that it is neither arbitrary nor necessarily the same for all the experiments. A meaningful choice of  $\tau$  could be guided by the important aim of controlling the False Discovery Rate (FDR) in multiple testing for the significance of the given large set of features due to a particular experiment.

Several techniques to obtain meaningful p-value cutoffs for genome-wide lists of microarray results have been suggested [8]. For a chosen FDR level  $\alpha$ , we thus obtained a p-value cutoff  $\tau_{j, \alpha'}$  for each experiment  $j$ , where  $\alpha'$  is such a value that would bound the FDR after combination to  $\alpha$ . We generalize TPM whereby only those p-values which “clear” their respective cutoffs form the present product for which we also compute the probability distribution and hence the combined p-value.

Yet another parameter  $K$  allows us to impose a *consensus* requirement that the product for every feature  $g$  be formed of at least  $K$  p-values (of  $g$ ) which clear their respective cutoffs.

---

<sup>1</sup>Department of Computer Science, Stony Brook University, NY 11794, Stony Brook, USA. E-mail: [spyne@cs.sunysb.edu](mailto:spyne@cs.sunysb.edu)

<sup>2</sup>Department of Computer Science, Stony Brook University, NY 11794, Stony Brook, USA. E-mail: [skiena@cs.sunysb.edu](mailto:skiena@cs.sunysb.edu)

<sup>3</sup>Department of Molecular Genetics and Microbiology, Stony Brook University, Stony Brook, NY 11794, USA. E-mail: [bfutcher@ms.cc.sunysb.edu](mailto:bfutcher@ms.cc.sunysb.edu)

Among all possible choices of  $m$  p-values ( $K \leq m \leq L$ ) with the above property, suitably chosen values of  $\alpha'$  yield the smallest (i.e., the most significant) product of  $g$  for the maximum number of its p-values  $p_j$  clearing  $\tau_{j,\alpha'}$  while the overall FDR level  $\alpha$  is preserved.

## 2 Experimental Results

We applied our meta-analysis method to three well-known genome-wide transcription factor (TF) local binding data sets due to [3], [7] and [1]. We indexed the intergenic regions in the datasets by their transcriptional target “gene”-s (we use this sweeping term for mere convenience) to obtain a combined list of 6401 genes over 3 experiments (**L** [3], **S** [7] and **H** [1]) containing p-values for the local binding of different TF proteins. Meta-analysis results for the protein Swi4 are presented below, where the combined result is validated with the help of percentile ranks of shortlisted 207 genes for Swi4 (and the TF MBF) due to [2].

Meta-analysis of the **L**, **S**, and **H** data sets with  $\alpha = 0.06$  and  $K = 2$  yields a screened subset **LSH**<sub>0.06,2</sub> of 106 genes common to all the sets, which is declared as significant. Of these 106 genes, 67 belong also to the set **I** due to [2], with a rank correlation  $|\rho| = 0.71$ . The high significance of the genes in **LSH**<sub>0.06,2</sub> is strongly supported by the data in **I** with the median of the **I**-specified percentile ranks of the **LSH**<sub>0.06,2</sub> genes being as high as 99.04 which corresponds to the topmost 20% of the set **I**.

Without the consensus and corroboration available to our meta-analysis, however, at the same level (0.06) of FDR, the correlations of the genes common to **I** and each of the individual data sets **L**<sub>0.06</sub>, **S**<sub>0.06</sub>, and **H**<sub>0.06</sub> drop to as low as 0.63. On the other hand, in the absence of the cutoffs to control FDR, the set of top 67 genes due to ordinary Fisher product combination of **L**, **S** and **H** also gives a similar lower correlation of 0.65. For  $K = 1$ , the correlation drops to 0.64, which proves the effectiveness of the consensus parameter.

Detailed meta-analysis is also performed with the much bigger number of recent experiments on genome-wide cell-cycle expression in fission yeast due to [6], [5], and [4].

## References

- [1] Harbison, C.T., *et al.* 2004. Transcriptional Regulatory Code of a Eukaryotic Genome. *Nature* 431: 99-104.
- [2] Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M., Brown, P.O. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409:533-8.
- [3] Lee, T.I., *et al.* 2002. Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*. *Science* 298: 799-804.
- [4] Oliva, A., Rosebrock, A., Ferrezuelo, F., Pyne, S., Chen, H., Skiena, S., Futcher, B., Leatherwood, J. The Cell Cycle Regulated Genes of *Schizosaccharomyces pombe*. In press at *Public Library of Science* (PLOS Biology), pending modifications.
- [5] Peng, X. *et al.* 2005. Identification of Cell Cycle-regulated Genes in Fission Yeast. *Molecular Biology of the Cell* 16:1026-1042.
- [6] Rustici, G. *et al.* 2004. Periodic gene expression program of the fission yeast cell cycle. *Nature Genetics* 36:809-817.
- [7] Simon, I., *et al.* 2001. Serial Regulation of Transcriptional Regulators in the Yeast Cell Cycle. *Cell* 106:697-708.
- [8] Storey, J.D., Tibshirani, R. 2003. Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences USA* 100: 9440-9445.
- [9] Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H., Weir, B.S. 2002. Truncated product method for combining p-values. *Genetic Epidemiology* 22: 170-185.