Majority logic decoding: a discrete method for detecting differential expression in RNA-Seq data

Humberto Ortiz-Zuazaga[1], **Roberto Arce Corretjer**[1]
[1]Department of Computer Science, University of Puerto Rico Rio Piedras, San Juan, Puerto Rico

Abstract— *We present a novel method of analysis of RNA-Seq data based on majority-logic-decoding. We apply the analysis to a simulation of differential gene expression and compare to a typical statistical analysis with linear models. Our technique results in a markedly improved false positive rate.*

Keywords: next-gen sequencing, finite dynamical systems, differential expression

1. Introduction

Gene regulatory networks are a valuable tool in the analysis of microarray data, and in the description of biological systems. A well established current in microarray analysis is the reverse engineering problem: given a set of genes and a set of expression measurements under varying conditions, determine the nature of transcriptional regulation among the genes. A rich tradition of discrete Boolean approaches to this problem exists [1], [2], [3], [4], [5]. Recent research into finite fields as a richer and more efficient alternative to Boolean logic has proven fruitful [6], [7], [8]. We have developed a series of techniques for error-correction and clustering based on finite fields [9].

The same variety of tools is not available for RNA-Seq [10] analysis, although the two experimental techniques share goals and some analytical framework. Extending FDS to the analysis of RNA-Seq data will bring a new approach to the quickly growing corpus of RNA-Seq data. Our hypothesis is that FDS's discrete nature is suited to modeling digital expression measurements. To test this hypothesis, we apply our techniques to a simulated gene expression experiment.

2. Methods

2.1 Discretization of expression

1) Take base 2 logarithm of counts
2) Compute mean of the control samples counts of each transcript.
3) Subtract mean value of control samples from each treated sample count, to make counts zero centered
4) Compute standard deviation of treated samples
5) Divide each treated sample by the standard deviation
6) Pick a discretization threshold t
7) For each sample, if normalized counts $> 1t$, gene is upregulated, $< -1t$ gene is downregulated, otherwise no change
8) Compute majority logic decoding (mld) value over all samples for each gene

2.2 Majority logic decoding

Upregulated samples are encoded as '+', downregulated as '-', and unchanged as '0'. The discretization then yields a list of symbols for every sample of each gene. Majority logic decoding looks as the symbols for every sample and selects the symbol that appears in a majority of samples. This procedure has been adapted from a similar procedure described for microarray data in [9].

2.3 Verification

To validate our methods, we simulate gene expression counts and apply our techniques. We use *flux simulator* [11], a tool for generating simulated RNA-Seq data. We generate 20 random gene expression experiments using the *Drosphila melanogaster* genome release 70 from ENSEMBL [12], and the default flux-simulator parameters. These 20 runs are divided into 10 control samples and 10 treated samples. We randomly select 2000 transcripts from the 29,173 present in the simulated data. The 2000 are divided into 4 groups of 500 each, and we add 100 and 200 to the treated or the controls in each group, to simulate a spike-in experiment.

3. Results

Figures 1 shows the variance vs mean for the simulated data. The plot in Figure 2 shows the same plot for a similar sample of data from *Drosphila melanogaster* [13] from a public repository of RNA-Seq data [14].

We applied mld to the simulated spike-in data. Since the mld is sensitive to the choice of threshold value t, we sweep over a range of values and compute the false positive and false negative rate. We compare these rates to a linear model of differential expression using the `lmFit` [15] function in the `limma` [16] package from bioconductor [17]. Figure 3 shows the ROC curves for mld and the linear model. For $t = 1.3$, we correctly

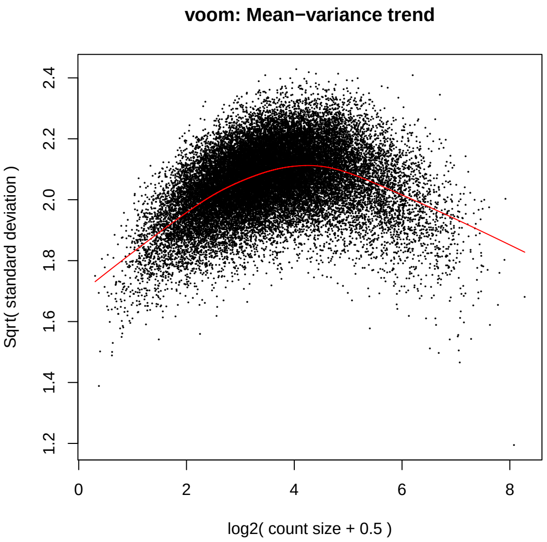

Fig. 1: Mean-variance relationship of simulated data.

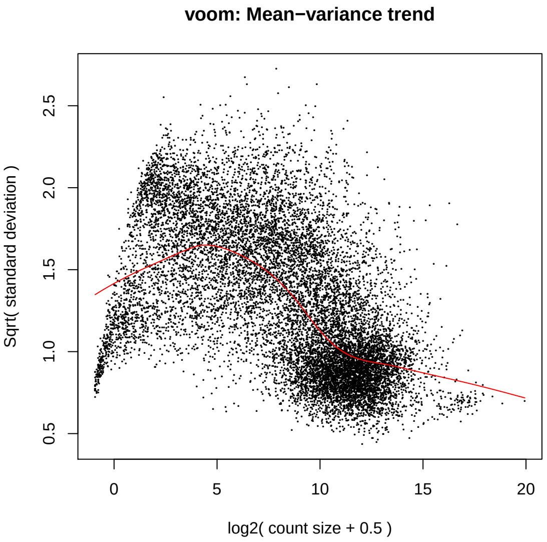

Fig. 2: Mean-variance relationship of real data.

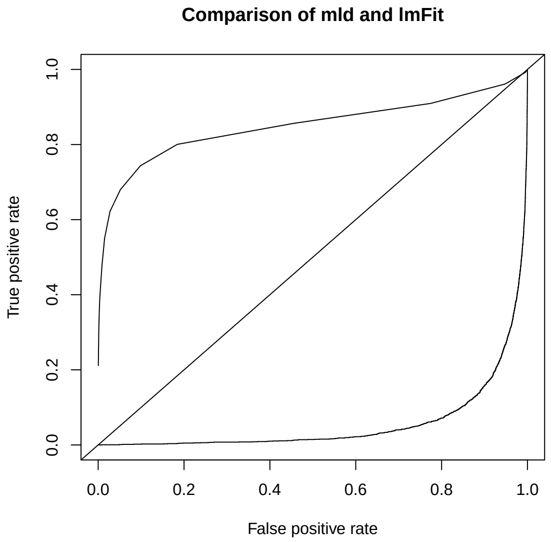

Fig. 3: ROC curve for mld (upper curve) and a linear model fit (lower curve) on the simulated spike-in experiment.

identified 1436 out of 2000 positives, while predicting 2779 false positives from the 27,166 negatives.

The linear model fit can be used to predict the probability of differential expression for each transcript. Figure 4 is a plot of the log odds of differential expression versus the log ratio of expresison for the 2000 spike-in transcripts.

4. Discussion

Our method of simulating differential expression of RNA-Seq data seems to produce data very similar to real RNA-Seq data, although Figure 1 shows reduced variance and mean expression compared to Figure 2.

We have described a modification of majority logic decoding to handle discrete gene expression data such as produced by RNA-Seq experiments. We have tested the method by comparing with linear models such as those produced by `lmFit`. On simulated spike-in data, our method results in a markedly improved sensitivity. Figure 4 shows that linear modelling of the spike-in transcripts predicts log ratios of expression different from zero for almost all spike-ins, but the large majority of spike-in transcripts show no statistical support for differential expression. It is unlikely that refinements of the linear model would be able to distinguish the spike-in genes from the negatives. The mld technique, conceptually simpler, demonstrates better specificity, while keeping the false positive rate low.

5. Future studies

The spike-in experiment we simulated is a simple gene expression experiment. We want to extend our simulation technique to allow simulation of biologically relevant differential expression. In particular, we would like to simulate a gene regulatory network response and then use our mld technique to recover changes in expression at different stages. This more complex simulation will allow us to test other FDS techniques and adapt them to RNA-Seq data. The next step after that would be to apply these techniques to real RNA-Seq data.

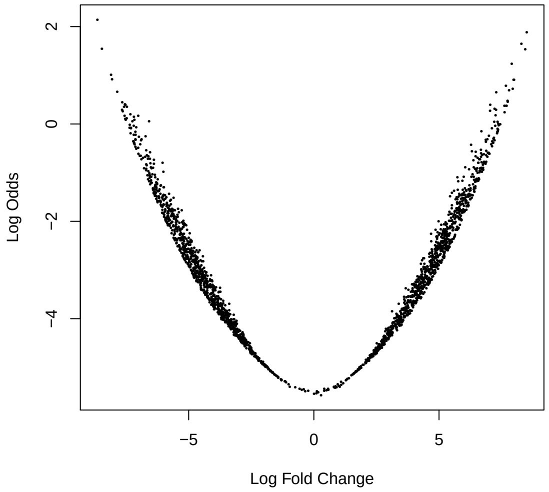

Fig. 4: Volcanoplot of spike-in genes.

6. Acknowledgments

The authors received partial support from the PR-INBRE grant (P20GM103475) from the National Institute for General Medical Sciences of the National Institutes of Health.

References

[1] T. Akutsu, S. Kuahara, O. Maruyama, and S. Miyano, "Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions," in *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, H. Karloff, Ed. ACM Press, 1998.

[2] T. E. Ideker, V. Thorsson, and R. M. Karp, "Discovery of regulatory interactions through perturbation: Inference and experimental design," in *Pacific Symposium on Biocomputing*, 2000, pp. 302–313.

[3] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *J. Theor. Biol.*, vol. 22, pp. 437–467, 1969.

[4] ——, *The Origins of Order*. New York, Oxford: Oxford University Press, 1993.

[5] S. Liang, S. Fuhrman, and R. Somogyi, "REVEAL, a general reverse engineering algorithm for inference of genetic network architectures." *Pac Symp Biocomput*, pp. 18–29, 1998.

[6] R. Laubenbacher and B. Stigler, "Dynamic networks," *Adv. in Al. Math.*, vol. 26, pp. 237–251, 2001.

[7] O. Moreno, D. Bollman, and M. A. Aviñó-Diaz, "Finite dynamical systems, linear automata and finite fields," *2002 WSEAS Int. Conf. on System Science Alied Mathematics & Computer Science and Power Engineering Systems*, pp. 1481–1483, 2002, also to appear in the International Journal of Computer Research.

[8] M. A. Aviñó-Díaz, E. Green, and O. Moreno, "Applications of finite fields to dynamical systems and reverse engineering problems," *Proceedings of the 19th ACM Symposium on Applied Computing - SAC*, 2004.

[9] H. Ortiz-Zuazaga, S. Peña de Ortiz, and O. Moreno de Ayala, "Error correction and clustering gene expression data using majority logic decoding," in *Proceedings of The 2007 International Conference on Bioinformatics and Computational Biology (BIOCOMP'07)*, Las Vegas, Nevada, USA, June 2007.

[10] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The transcriptional landscape of the yeast genome defined by rna sequencing," *Science*, vol. 320, no. 5881, pp. 1344–9, Jun 2008.

[11] T. Griebel, B. Zacher, P. Ribeca, E. Raineri, V. Lacroix, R. Guigó, and M. Sammeth, "Modelling and simulating generic RNA-Seq experiments with the flux simulator," *Nucleic Acids Research*, vol. 40, no. 20, pp. 10 073–10 083, Nov. 2012. [Online]. Available: http://nar.oxfordjournals.org/content/40/20/10073

[12] P. Flicek, I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. García-Girón, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A. K. Kähäri, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. P. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, and S. M. J. Searle, "Ensembl 2013," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D48–55, Jan 2013.

[13] B. R. Graveley, A. N. Brooks, J. W. Carlson, M. O. Duff, J. M. Landolin, L. Yang, C. G. Artieri, M. J. van Baren, N. Boley, B. W. Booth, J. B. Brown, L. Cherbas, C. A. Davis, A. Dobin, R. Li, W. Lin, J. H. Malone, N. R. Mattiuzzo, D. Miller, D. Sturgill, B. B. Tuch, C. Zaleski, D. Zhang, M. Blanchette, S. Dudoit, B. Eads, R. E. Green, A. Hammonds, L. Jiang, P. Kapranov, L. Langton, N. Perrimon, J. E. Sandler, K. H. Wan, A. Willingham, Y. Zhang, Y. Zou, J. Andrews, P. J. Bickel, S. E. Brenner, M. R. Brent, P. Cherbas, T. R. Gingeras, R. A. Hoskins, T. C. Kaufman, B. Oliver, and S. E. Celniker, "The developmental transcriptome of drosophila melanogaster," *Nature*, Dec. 2010.

[14] A. C. Frazee, B. Langmead, and J. T. Leek, "ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets," *BMC Bioinformatics*, vol. 12, p. 449, 2011, PMID: 22087737. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22087737

[15] G. Smyth, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, "Limma: linear models for microarray data." in *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, 2005, pp. 397–420.

[16] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Stat Appl Genet Mol Biol*, vol. 3, p. Article3, 2004.

[17] R. C. Gentleman, V. J. Carey, D. M. Bates, *et al.*, "Bioconductor: Open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, p. R80, 2004. [Online]. Available: http://genomebiology.com/2004/5/10/R80