Poster: Empirical Comparison of Protocols for Sequencing-Based Gene and Isoform Expression Profiling

Marius Nicolae and Ion I. Măndoiu
Department of Computer Science & Engineering
University of Connecticut, 371 Fairfield Rd., Unit 2155, Storrs, CT, 06269-2155, USA
Tel: +1 860 4863784; Fax: +1 815 3018557; E-mail: {man09004,ion}@engr.uconn.edu

I. Introduction

Massively parallel sequencing is quickly replacing microarrays as the technology of choice for performing gene expression profiling. Two main transcriptome sequencing protocols have been proposed in the literature. The most commonly used one, referred to as RNA-Seq, generates short (single or paired) sequencing tags from the ends of randomly generated cDNA fragments. An alternative protocol, referred to as Digital Gene Expression (DGE) [1], or high-throughput sequencing based Serial Analysis of Gene Expression (SAGE-Seq) [2], generates single cDNA tags using an assay including as main steps transcript capture and cDNA synthesis using oligo(dT) beads, cDNA cleavage with an anchoring enzyme, and release of cDNA tags using a tagging enzyme whose recognition site is ligated upstream of the recognition site of the anchoring enzyme. In this we poster present the results of a simulation study comparing the accuracy achieved by the two protocols based on state-of-the-art inference algorithms using the expectation-maximization (EM) framework.

II. Methods and Results

Two EM algorithms were used to infer gene and isoform expression levels: IsoEM [3] for RNA-Seq and a newly designed EM algorithm for DGE. RNA-Seq and DGE data was generated for 20 different human tissues based on gene expression levels defined in the GNFAtlas2 table. For DGE we modeled several cutting enzymes from the Restriction Enzyme Database, both assuming complete digestion (cutting probability $p = 1$) and partial digestion with $p = .5$.

Table I shows the Median Percent Error (MPE) for isoform and gene expression levels inferred from simulated DGE and RNA-Seq data sets with 30M tags of length 21 averaged over the 20 tissues analyzed. We included DGE results for three anchoring enzymes: DpnII [1] with recognition site GATC, NlaIII [2] with recognition site CATG, and CviJI with degenerate recognition site RGCY (R=G or A, Y=C or T). The last column shows the percentage of isoforms cut by each enzyme.

Table I
MPE for isoform and gene expression levels inferred from DGE and RNA-Seq datasets with 30M tags of length 21

Protocol		Isoform MPE Avg.	Isoform MPE S.D.	Gene MPE Avg.	Gene MPE S.D.	Isof. Cut
DGE	GATC $p = 1$	15.4	0.71	5.7	1.14	94%
	GATC $p = .5$	15.1	0.90	4.4	0.98	
	CATG $p = 1$	10.4	0.91	3.3	0.57	96%
	CATG $p = .5$	12.0	0.71	2.8	0.36	
	RGCY $p = 1$	9.3	0.86	2.6	0.26	98%
	RGCY $p = .5$	10.4	0.84	2.3	0.27	
RNA-Seq		12.0	0.82	5.4	0.68	N/A

III. Conclusion

For isoform expression inference, RNA-Seq and DGE give comparable results whereas for gene expression DGE has better accuracy when enough isoforms are cut by the restriction enzyme used. DGE accuracy improves with the percentage of isoforms cut and is better for partial digestion compared to complete digestion. Using enzymes with degenerate recognition sites, such as CviJI, yields better accuracy than using enzymes from published studies [1], [2].

Acknowledgment

Work supported in part by NSF awards IIS-0546457 and IIS-0916948

References

[1] Y. Asmann, E. Klee, E. A. Thompson, E. Perez, S. Middha, A. Oberg, T. Therneau, D. Smith, G. Poland, E. Wieben, and J.-P. Kocher, "3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer," *BMC Genomics*, vol. 10, no. 1, p. 531, 2009.

[2] Z. J. Wu, C. A. Meyer, S. Choudhury, M. Shipitsin, R. Maruyama, M. Bessarabova, T. Nikolskaya, S. Sukumar, A. Schwartzman, J. S. Liu, K. Polyak, and X. S. Liu, "Gene expression profiling of human breast tissue samples using SAGE-Seq," *Genome Research*, vol. 20, no. 12, pp. 1730–1739, 2010.

[3] M. Nicolae, S. Mangul, I. I. Mandoiu, and A. Zelikovsky, "Estimation of alternative splicing isoform frequencies from RNA-Seq data," in *Proc. WABI*, ser. Lecture Notes in Computer Science, V. Moulton and M. Singh, Eds., vol. 6293. Springer, 2010, pp. 202–214.

978-1-61284-852-5/11/$26.00 ©2011 IEEE