Workshop : Novel Transcript Reconstruction from Paired-End RNA-Seq Reads Using Fragment Length Distribution

Serghei Mangul*, Adrian Caciula*, Ion Mandoiu† and Alex Zelikovsky*
*Department of Computer Science, Georgia State University, Atlanta, GA 30303
Email: {serghei,acaciula,alexz}@cs.gsu.edu
†Department of Computer Science & Engineering, University of Connecticut, Storrs, CT 06269
Email : ion@engr.uconn.edu

Recent advances in DNA sequencing have made it possible to sequence the whole transcriptome by massively parallel sequencing, commonly referred as RNA-Seq. RNA-Seq is quickly becoming the technology of choice for transcriptome research and analyses. RNA-Seq allows to reduce the sequencing cost and significantly increase data throughput, but it is computationally challenging to use such RNA-Seq data for reconstructing of full length transcripts and accurately estimate their abundances across all cell types. The common computational problems include: gene and isoform expression level estimation [1], find transcriptome quantification, transcriptome discovery and reconstruction. To solve these problems it is required to have scalable computational tools.

A number of recent works have addressed the problem of transcriptome reconstruction from RNA-Seq reads. These methods fall into three categories: "genome-guided", "genome-independent" and "annotation-guided". Genome-guided methods first map all reads to the reference genome and then use spliced reads to reconstruct transcriptome. Such methods, also called "ab initio", have been proposed in [2]. Rather than mapping reads to the reference genome first, genome-independent methods such as Trinity or transAbyss directly assemble reads into the transcripts. A commonly used approach for such methods is de Brujin graph utilizing "k-mers" rather than reads for the graph construction. Annotation-guided methods such as RABT Assembly[3] and DRUT [4] explicitly use existing annotations for the transcriptome reconstruction.

In this work, we propose a novel statistical "genome-guided" method called "**T**ransciptome **R**econstruction using **I**nteger **P**rograming" (TRIP) that incorporates fragment length distribution into novel transcript reconstruction from paired-end RNA-Seq reads. To reconstruct novel transcipts, we create a splice graph based an exact annotation of exon boundaries and RNA-Seq reads. The exact annotation of exons can be obtained from annotation databases (e.g., Ensembl) or can be inferred from aligned RNA-Seq reads. A splice graph is a directed acyclic graph (DAG), whose vertices represent exons and edges represent splicing events. We enumerate all maximal paths in the splice graph using a depth-first-search (DFS) algorithm. These paths correspond to putative transcripts and are the input for the TRIP algorithm.

To solve the transcriptome reconstruction problem we must select a set of putative transcripts with the highest support from the RNA-Seq reads. We formulate this problem as an integer program. The objective to select the smallest set of putative transcripts that yields a good statistical fit between the fragment length distribution empirically determined during library preparation and fragment lengths implied by mapping read pairs to selected transcripts.

Preliminary experimental results on synthetic datasets generated with various sequencing parameters and distribution assumptions show that TRIP has increased transcriptome reconstruction accuracy compared to previous methods that ignore fragment length distribution information.

Acknowledgments. This work has been partially supported by NSF award IIS-0916401,NSF award IIS-0916948, Agriculture and Food Research Initiative Competitive Grant no. 201167016-30331 from the USDA National Institute of Food and Agriculture and Second Century Initiative Bioinformatics University Doctoral Fellowship.

REFERENCES

[1] M. Nicolae, S. Mangul, I. Mandoiu, and A. Zelikovsky, "Estimation of alternative splicing isoform frequencies from rna-seq data," *Algorithms for Molecular Biology*, vol. 6:9, 2011. [Online]. Available: http://www.almob.org/content/6/1/9

[2] C. Trapnell, B. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. van Baren, S. Salzberg, B. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." *Nature biotechnology*, vol. 28, no. 5, pp. 511–515, 2010. [Online]. Available: http://dx.doi.org/10.1038/nbt.1621

[3] A. Roberts, H. Pimentel, C. Trapnell, and L. Pachter, "Identification of novel transcripts in annotated genomes using rna-seq," *Bioinformatics*, 2011.

[4] S. Mangul, A. Caciula, I. Mandoiu, and A. Zelikovsky, "Rna-seq based discovery and reconstruction of unannotated transcripts in partially annotated genomes," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, nov. 2011, pp. 118 –123.

978-1-4673-1321-6/12/$31.00