Short Abstract: An Integer Programming Approach to Novel Transcript Reconstruction from Paired-End RNA-Seq Reads

Serghei Mangul[1], Adrian Caciula[1], Nicholas Mancuso[1], Olga Glebova[1], Ion Mandoiu[2], and Alex Zelikovsky[1]

[1] Department of Computer Science, Georgia State University, Atlanta, GA 30303
Email: {serghei, acaciula, nmancuso, oglebova, alexz}@cs.gsu.edu
[2] 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 high-throughput RNA sequencing (RNA-seq)[1]. RNA-Seq is becoming a technology of choice for transcriptome analyses [2] which allows to reduce the sequencing cost and significantly increase data throughput, but it is computationally challenging to use such data for reconstructing full-length transcripts and accurately estimating their abundances across all cell types.

The common applications of RNA-seq are *gene expression level estimation (GE)*, *transcript expression level estimation (IE)* [3] and *novel transcript reconstruction (TR)*. A variety of new methods and tools have been recently developed to tackle these problems.

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 transcripts, we create a splice graph based on exact annotation of exon boundaries and 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 model. The objective is 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.

The following parameters are used in the proposed mathematical model:

Symbol	Description
N:	Total number of reads ;
p :	Paired-end read ;
j :	Index of paired-end read p, $1 \leq j \leq N$;
i :	Index of standard deviation, $0 \leq i \leq 4$;
t :	Candidate transcript ;
$K(k)$:	Number (index) of transcripts t, $1 \leq k \leq K$;
$T_i(p_j)$:	Set of candidate transcripts on which paired-read p_j can be mapped with a fragment length between $i-1$ and i standard deviation, $0 \leq i \leq 4$;
$T_4(p_j)$:	set of candidates transcripts within more than 3 standard deviation ;
$y(t_k)$:	1 if candidate transcript t_k is selected, and 0 otherwise;
$x_i(p_j)$:	1 iff the read p_j is mapped between $i-1$ and i standard deviation, and 0 otherwise;

Objective function of this model is to minimize the total number of possible candidate transcripts, as shown in equation (1).

$$(1)\ minimize \sum_{t \in T} y(t)$$

$$(2)\ \sum_{t \in T_i(p)} y(t) \geq x_i(p), \forall p, i = \overline{1,4}$$

$$(3)\ N_{reads} * (n(s_i) - \epsilon) \leq \textstyle\sum_p x_i(p) \leq N * (n(s_i) + \epsilon)$$

$$(4)\ \textstyle\sum_i x_i(p) = 1$$

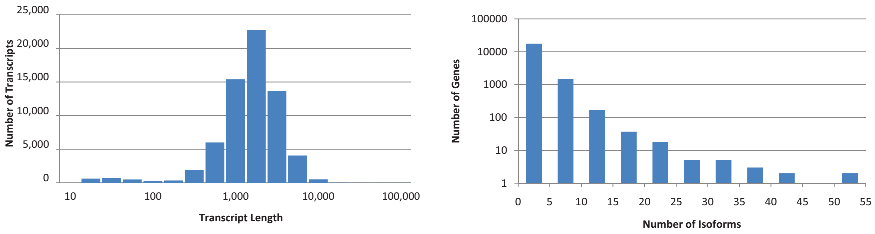
Equation (2) implies that for each paired-end read p_j with non-empty set $T_i(p_j)$, at least one transcript is selected (this first constraint allows to select multiple transcripts for the same read). Note that all $y(t) = 1$ because we only consider the transcripts in non-empty set $T_i(p_j)$ of that particular paired-read p_j, for which t had already been selected (otherwise $T_i(p)$ would be empty). All $x_i(p_j) = 1$ because if p_j is not mapped within standard deviation i then $T_i(p_j)$ is empty set, i.e., p_j will not be chosen for this loop since we only consider ”p with non-empty $T_i(p)$”. In the worst case read p_j is mapped for sure with standard deviation 4 (i.e., $x_4(p) = 1$ which ensures that at least one transcript is selected for read p_j, even if it is with a high standard deviation.

Equation (3) implies that the sum of all paired-end reads mapped within standard deviation i equals total number of paired-end reads expected within standard deviation i ($\pm\epsilon$). If a fragment length is approximately normal then about 68% of the fragments are within one standard deviation of the mean (mathematically, $\mu \pm s$, where μ is the arithmetic mean), about 95% are within two standard deviations ($\mu \pm 2s$), and about 99.7% lie within three standard deviations ($\mu \pm 3s$). This is known as the 68-95-99.7 rule, or the empirical rule. Let s_1, s_2 and s_3 be expected portion within one, two, and three standard deviations from the mean. (From statistics we know that $s_1 = 0.68$, $s_2 = 0.95$ and $s_3 = 0.99$).

The number of paired-end reads that have been mapped within a standard deviation i should be equal, more or less ϵ, with the expected value (ϵ varies between 0.01 ad 0.05, because we can have errors and map the same read to different transcripts with same standard deviation 1, so we want to limit to only one.

Paired-end reads p are short and may be mapped to different transcripts, therefore we have equation (4) which ensures that each paired-end read p is mapped in only one category of standard deviation (i.e. standard deviation sets are mutually exclusive).

For our simulation we have used Human genome UCSC annotations, GNFAtlas2 gene expression levels with uniform/geometric expression of gene transcripts. The fragment lengths follow a normal distribution with a mean length of 500 and a standard definition of 50.



(a) (b)

Fig. 1. Distribution of transcript lengths (a) and gene cluster sizes (b) in the UCSC dataset

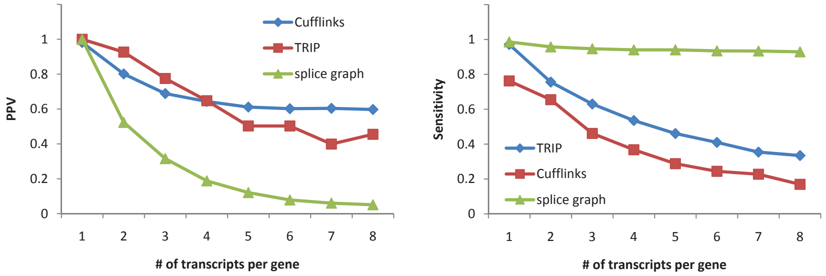
Preliminary experimental results on synthetic datasets generated with various sequencing parameters and distribution assumptions show that TRIP has increased transcriptome reconstruction accuracy for genes with less than 4 transcripts compared to previous methods that ignore fragment length distribution information.

Following [4], we use sensitivity and Positive Predictive Value (PPV) to evaluate the performance of different methods. Sensitivity is defined as portion of the annotated transcript sequences being captured by candidate transcript sequences as follows:

$$Sens = \frac{TP}{TP + FN}$$

PPV is defined portion of annotated transcript sequences among candidate sequences as follows:

$$PPV = \frac{TP}{TP + FP}$$



(a)

(b)

Fig. 2. Flowchart for TRIP: (a) Positive Predictive Value (PPV) and (b) Sensitivity

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. A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nature methods*, 2008. [Online]. Available: http://dx.doi.org/10.1038/nmeth.1226
2. Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics." *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, 2009. [Online]. Available: http://dx.doi.org/10.1038/nrg2484
3. 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
4. I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. Mandoiu, P. Balfe, and A. Zelikovsky, "Inferring viral quasispecies spectra from 454 pyrosequencing reads," *BMC Bioinformatics*, vol. 12, no. Suppl 6, p. S1, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/S6/S1