Monte-Carlo Regression Algorithm for Isoform Frequency Estimation from RNA-Seq Data

Adrian Caciula, Alex Zelikovsky
Department of Computer Science
Georgia State University
Atlanta, Georgia 30303
Email: {acaciula, alexz}@cs.gsu.edu

Serghei Mangul
Department of Computer Science
University of California
Los Angeles, CA 90095
Email: serghei@cs.ucla.edu

James Lindsay and Ion Mandoiu
Deptment of Computer Science &
Engineering, University of Connecticut
Storrs, CT 06269
Email: {james.lindsay, ion}@engr.uconn.edu

***Abstract*—We propose a Monte-Carlo Regression based method for isoform frequency estimation from RNA-Seq reads.**

I. Introduction

Reducing isoform frequency estimation error rate is critical for detecting similar transcripts or unraveling gene functions and transcription regulation mechanisms, especially in those cases when one isoform is a subset of another. Figure 1 shows a gene with sub-transcripts from human genome (hg19).

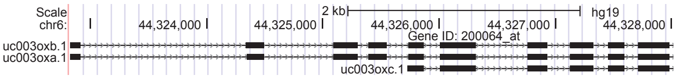

Fig. 1. Screenshot from Genome browser [1]

The most accurate existing tools exploit the expectation-maximization method for the maximum likelihood approach (see e.g., IsoEM [2]), but such methods tend to skew the estimated frequency toward super-transcripts. In this paper we propose to apply a more accurate regression-based estimation.

II. MCReg Algorithm

A. Observed Read Distribution

The first step of $MCReg$ is to map the paired-end reads onto the library of known isoforms using an ungapped aligner (e.g., Bowtie [3]). We assume that the fragment length distribution is normal $\mathcal{N}(\mu, \sigma^2)$ with the mean fragment length $\mu \in R$ and the standard deviation σ estimated from the read alignments.

We partition all reads into set of classes $\tilde{R}$, where each class $\tilde{r} \in \tilde{R}$ consists of reads that can be emitted by the same subset of transcripts. The observed frequency of $\tilde{r}$ is the sum of frequencies of all reads belonging to $\tilde{r}$.

B. MC-Based Estimation of Expected Read Distribution

Let R' be the set of all possible reads and let $\tilde{R}'$ be the partition of R' into read classes. For each transcript $t \in T$ and $\tilde{r}' \in \tilde{R}'$, we estimate $d_{t,\tilde{r}'} = Pr(\tilde{R}' = \tilde{r}' | T = t)$ using Monte Carlo method – we simulate reads from t ($|R'|$ is proportional to the adjusted length of the transcript t, $l_t = |t| - \mu + 1$) and find the portion of them belonging to $\tilde{r}'$. Let f'_t be the portion of reads emitted by t, then the expected frequency of the class $\tilde{r}'$ is estimated as follows:

$$e_{\tilde{r}'} = \sum_{t \in T} f'_t d_{t,\tilde{r}'} \qquad (1)$$

C. Regression-Based Estimation of Isoform Frequencies

Regression-based estimation of f'_t's minimizes squared deviation between observed and expected read frequencies

$$\text{minimize:} \quad \sum (e_{\tilde{r}'} - o_{\tilde{r}})^2 \qquad (2)$$

Substituting (1) in (2) we obtain the following program

$$\begin{aligned} &\text{minimize:} && \sum (\sum_{t \in T} f'_t d_{t,\tilde{r}'} - o_{\tilde{r}})^2 \\ &\text{subject to:} && \sum_{t \in T} f'_t = 1 \text{ and } f'_t \geq 0 \ , \forall t = 1...|T| \end{aligned} \qquad (3)$$

The least-square formulation (3) can be solved with any constrained quadratic programming solver. Finally, the isoform frequencies f_t's can be obtained from f'_t's using adjusted transcript lengths l_t's

$$f'_t = f_t l_t / \sum_{k \in T} f_k l_k \Rightarrow f_t = (f'_t / l_t) / \sum_{k \in T} f'_k / l_k \qquad (4)$$

III. Results

We validated $MCReg$ on $chr1$ from $hg19$ which contains a total of 5509 transcripts (from 1990 genes). We have simulated $10M$ paired-end reads of length $100bp$ with the mean fragment length $\mu = 500$. Frequency estimation accuracy was assessed using the coefficient of determination r^2. For $IsoEM$ $r^2 = 0.92$, while for $MCReg$ $r^2 = 0.97$. The results shows better correlation compared with $IsoEM$ especially because of those cases of sub-transcripts where $IsoEM$ skewed the estimated frequency toward super-transcripts.

Acknowledgments

This work has been partially supported by two Collaborative Research Grant from Life Technologies, awards IIS-0916401 and IIS-0916948 from NSF, and Agriculture and Food Research Initiative Competitive Grant no. 201167016-30331 from the USDA National Institute of Food and Agriculture.

References

[1] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, Haussler, and D., "The Human Genome Browser at UCSC," *Genome Research*, vol. 12, no. 6, pp. 996–1006, Jun. 2002.

[2] 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.

[3] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, p. R25, 2009.

978-1-4673-1321-6/12/$31.00 ©2013 IEEE