SASeq: A Selective and Adaptive Shrinkage Approach to Detect and Quantify Condition-Specific Transcripts using RNA-Seq

Tin Nguyen, Nan Deng and Dongxiao Zhu[1],

Keywords: RNA-Seq, condition-specific isoform expression, tissue-specific isoform expression, quadratic programming, shrinkage, constrained least square, exonic coverage.

1 Abstract

Detection and quantification of condition-specific transcripts using RNA-Seq is a central task to transcriptomics research. Initial efforts on mathematical or statistical modeling of read counts or per-base exonic expression signal have been successful but may face an increasing risk of model misspecification and overfitting. This is because the number of reference transcripts in the database is much larger than that of the condition-specific transcripts expressed under a single biological condition, and the difference is getting larger with the accelerated augmentation of transcripts database. To improve the prediction accuracy, a subset of candidate condition-specific transcripts may be selected at the same time of estimating their relative abundances. The standard shrinkage approaches, such as Lasso, select condition-specific transcripts by shrinking all the transcript abundances to zero. However, the blind shrinkage does not necessarily lead to the set of condition-specific transcripts. Informed shrinkage approaches, motivated by exonic coverage signal, are thus desirable.

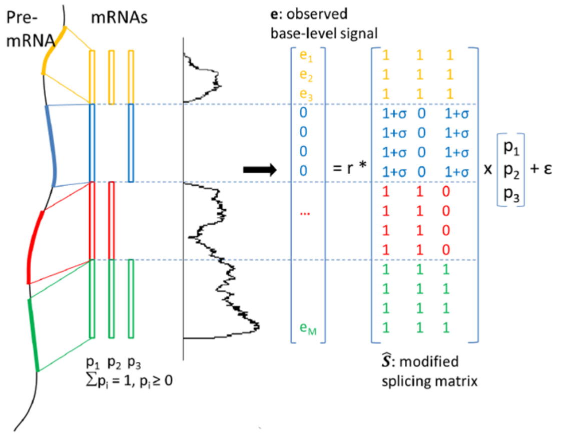

Figure 1: Select and quantify condition-specific transcripts using the observed per-base exonic expression signal (single sample case). r represents the gene expression abundance parameter, $\mathbf{p}$ represents abundance proportions of the three isoform transcripts, and $\mathbf{e}$ represents the observed per-base exonic expression signal. In this example, the gene has three reference transcripts and four annotated exons. The second transcript skips the second exon whereas the third transcript skips the third exon. A tuning parameter σ is introduced to penalize the selected (blue) regions of the selected transcripts (the first and the third) having no exonic expression signal. The shrinkage level is adaptively adjusted according to the exonic expression level over the optimization iterations.

We propose a new mathematical model of the observed exonic expression signal and the underlying transcript structure, we introduce a tuning parameter to penalize the selected regions in the selected transcripts that were not supported by the observed exonic expression signal, and we develop a constrained least square algorithm to adaptively adjust the shrinkage level based on the exonic expression signal (Figure 1). We implement and integrate the new method into our existing GUI system, SAMMate `http://asammate.sourceforge.net/`, to detect and quantify condition-specific transcripts. Our tool takes a variety of RNA-Seq data formats, such as fasta, fastq, SAM or BAM, as input and output transcript abundance through a few mouse clicks. Using simulation studies, our methods compare favorably with selected competing methods in terms of both time complexity and accuracy. We also demonstrate the potential applications by analyzing a real-world RNA-Seq data set.

The ever-increasing reference transcript numbers in the database as well as their error rates increase the risk of model misspecification and overfitting. Therefore, model selection via shrinkage opens a promising avenue to future research in this area. The key novelty of our approach is to shrink down the transcript abundance proportions that were not supported by the observed per-base exonic expression signal and adaptively adjust the shrinkage level accordingly. Our approach is an informed shrinkage approach, which has important differences from the Lasso type of blind shrinkage approach. The former permits a more accurate selection and shrinkage of the transcripts and their regions. In addition to working with transcripts database, SASeq is flexible enough to detect and quantify active transcripts from transcriptome assembly outputs in gtf format, where the model misspecification is also an outstanding issue due to the excessive assembly bias and errors.

[1] Department of Computer Science, Wayne State University, Detroit, MI 48202, USA. E-mail: `dzhu@wayne.edu`