StatsM254 Statistical Methods in Computational Biology Lecture 7 - 04/23/2014

Lecture 7

Lecturer: Jingyi Jessica Li *Scribe: Wanlu Liu*

1 Introduction

As the development of RNA-seq, it has become possible for genome wide transcriptome analysis. However, the identification of source and distribution of reads, modeling the transcript abundance and developin statistical method are still remaing challenges. In 2008, Mortazavi *et al.* [1] have developed reads per kilobase of the transcript per million mapped to the transcriptome (RPKM) to quantify the expression level of a transcript.

$$RPKM = \frac{total\ number\ of\ reads\ in\ a\ region}{\frac{length\ of\ the\ region\ in\ bp}{1000} \times \frac{total\ number\ of\ reads\ in\ the\ RNA-Seq\ data\ set}{10^6}}$$

RPKM simply counts the reads mapped to a specific isoformed normalized against the isoform length and the sequencing depth. Thus, it's hard to calculate the isoform specific RPKM because most short reads are mapped to a regions that is shared by more than one isoforms. In this paper [2], the authors developed a statistical model to compute the isoform-specific expression indexes and the uncertainties in the estimates.

The question is, for multiple isoforms, how to determine the isoform specific read counts? For example, in figure 1 for reads from exon 1 or exon 3, we cannot distinguish whehter it comes form isoform 1 or isoform 2. The challenge is how to distinguish those reads that are shared between isoforms. By modeling the reads that fall into multi-isform region with a poisson model for isoform expression, we can estimate the isoform abundance. The expression of each individual isoform is estimated by solving a convex optimization problem and statistical inferences about the parameters are obtained from the posterior distribution by importance sampling.

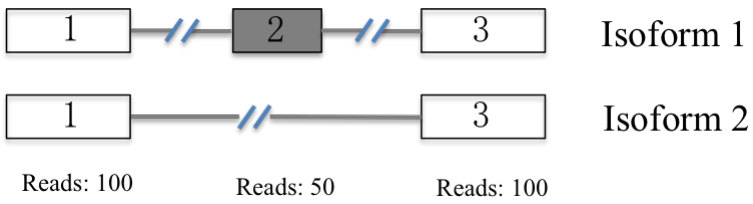

Figure1. Assume a gene have two isoforms and it's exon 1, exon2 and exon 3 have equal exon length. The reads mapped to exon 1 and exon 3 are 100 while the reads for exon 2 is 50.

2 Model

2.1 Notations

For isoform f:
l_f : length of isoform f
k_f : number of copies of transcripts of the isoform f
F : the collection of all isoforms,
Then, the total length of transcript is:

$$L = \sum_f k_f \cdot l_f$$

2.2 Assumption

The reads are sampled independently and uniformly from the collection of the transcripts with total length.
Given the assumption,
the probability of a read from isoform f is

$$\frac{k_f \cdot l_f}{L} \triangleq \theta_f \cdot l_f$$

Where θ_f is the expressed index of f, N is the total number of mapped reads in the gene, and X is the number of reads from isoform f.
Then,

$$X \sim Binominal(N, \theta_f l_f)$$

Based on approximation,

$$X \sim Poisson(\lambda = N\theta_f l_f)$$

For a gene with m exons and n isoform with expression indexes $(\theta_1, \ldots, \theta_n)$, suppose the exon length is $(l'_1, \ldots, l'_m)$, denote the C_{ij} be the indicator that

$$C_{ij} := I(\text{isoform } i \text{ contain exon } j)$$

Then,

Counts in exon $j \sim Poisson(\lambda = Nl'_j \sum_{i=1}^{n} \theta_i C_{ij})$,

Counts in junction of j and $k \sim Poisson(\lambda = Nl'_{jk} \sum_{i=1}^{n} C_{ij} C_{ik} \theta_i)$

We can get the likelihood based on the parameter estimation of $(\theta_1, \ldots, \theta_n)$.

References

[1] Mortazavi,A. et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat. Methods, 5, 621628.

[2] Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. Bioinformatics, 25(8), 1026-1032.