White Paper: RNA-Seq Methods

Introduction

RNA-seq is quickly becoming an invaluable next-generation tool for the study of RNA content in a sample, offering an unprecedented view of the transcriptome. The Partek RNA-Seq workflow allows you to calculate transcript-level expression of a gene's isoforms and determine alternative splicing of the gene. This white paper explains the methods used for estimating expression levels of transcripts and quantifying alternative splicing of a gene.

Isoform Expression Quantification

Given a set of known isoforms and a set of reads that map to the genome, Partek® Genomics Suite™ estimates the most likely relative expression levels of each isoform using an expectation/maximization algorithm similar to the one given in Xing, Y et al. Partek's algorithm differs from Xing, Y et al. in that it (1) quantifies isoform expression levels across the whole genome at the same time rather than each gene separately, and (2) normalizes by transcript length to account for the transcript fragmentation step in RNA-seq.

Partek starts from a set K of known isoforms of lengths $L_1 \ldots L_{|K|}$, and a set *M* of aligned reads that map to the genome (possibly at multiple locations). An alignment is *compatible* with an isoform if the following conditions are met:

- The alignment is contained entirely within the exon of the isoform OR is aligned to one of the isoform's junctions.
- For strand-specific assays, the strand of the alignment agrees with the strand of the isoform.

Let A be a $|K| \times |M|$ indicator matrix such that $a_{k,m} = 1$ if the mth alignment is compatible with isoform k, and zero otherwise. Let B be a $|K| \times |M|$ matrix such that $b_{k,m}$ is the probability alignment *m* comes from isoform *k*. Partek estimates $\theta = (p_1 \ldots p_{|K|})$, where p_k is the relative proportion of isoform k such that $\sum_k p_k = 1$. θ is estimated using an E/M algorithm consisting of the following steps:

E-step: $b_{k,m} = (a_{k,m} p_k^{(t)}) / (\sum_k a_{k,m} p_k^{(t)})$.

M-step: Let $n_k = \sum_m b_{k,m}$.

$$\begin{cases} p_k^{(0)} = 1/|K|, \\ p_k^{(t+1)} = \dfrac{n_k / L_k}{\sum_k n_k / L_k} \end{cases}$$

These two steps are repeated until θ converges or a maximum number of iterations is reached.

Transcript Level Analysis

In the previous section, the number of read counts for each isoform of a gene was estimated. These results will be used to estimate transcript-level differential expression and transcript-level alternative splicing. Each of these will be discussed in turn.

Transcript-level Differential Expression

From *S* samples, Partek determines if a transcript is differentially expressed across those samples (e.g. transcript expression between normal tissue and tumor tissue). For a given transcript, let n_s be the number of reads in sample *s* that are assigned to that transcript. n_s is modeled as being drawn from a binomial distribution with probability p_s and total number of mapped reads N_s. The null hypothesis assumes that the transcript is expressed the same across samples, $\forall_{i,j \in S} p_i = p_j$. In this case, p_s is estimated as $\hat{p}_{Null} = \sum_s n_s \Big/ \sum_s N_s$. The alternative hypothesis assumes that the transcript expression is the same across samples and estimates each p_s as $\hat{p}_{Alt} = n_s / N_s$. The log likelihood ratio (LLR) can be written as

$$LLR = \sum_s n_s \log(\hat{p}_{Alt} / \hat{p}_{Null})$$

Since the number of mapped reads in a sample N_s is large, we can use the approximation that $2 * LLR$ follows a χ^2 distribution with $(S-1)$ degrees of freedom. Partek then calculates a p-value based on the χ^2 distribution and reports that in a column of the RNA-seq results spreadsheet.

Alternative-Splicing Quantification

To determine if the relative expression levels of a gene's isoforms are changed across different samples (e.g. normal versus tumor tissues), set up a contingency table *C* with *S* rows (number of samples) and *K* columns (number of isoforms) for each gene. $c_{s,k}$ is the number of reads that fall within isoform *k* of sample *s,* which was estimated using the E/M algorithm previously discussed. Partek then calculates a χ^2 statistic on the contingency table *C* based on the log likelihood ratio and

computes the associated p-value, which is reported in the RNA-seq results spreadsheet.

References

Xing, Y., Yu, T., Wu, Y.N., Roy, M., Kim, J., & Lee, C. An expectation-maximization algorithm for probalisitic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res* 2006, 34:3150-3160.