Juggling with offsets unlocks bulk RNA-seq tools for fast and scalable differential detection, usage and aberrant splicing analyses

A seminar by Lieven Clement

Department of Applied Mathematics, Computer Science and Statistics
Ghent University (Belgium)

Wednesday 6 Mar 2024 | 4.30 p.m.
Room Benvenuti
Department of Statistical Sciences

RNA-seq analysis has become indispensable in molecular biology and revolutionised nearly every aspect of our understanding of genomic function. The technology generates count data for every gene, which are used as a proxy for their corresponding expression levels in biological samples. The data analysis is typically done with well-established bulk RNA-seq tools that rely on negative binomial regression to prioritise genes for which the expression levels are associated with predictors such as disease status, treatment, age, etc.

However, RNA-seq also has become key for other applications which involve the analysis of proportions. Indeed, the biological process of alternative splicing gives rise to multiple expression products, i.e. transcripts, for a single gene and the dysregulation of this splicing process has been reported extensively as a cause for disease. This application involves assessing differences in transcript usage, i.e. the fraction of transcript counts on the total count of all transcripts of a particular gene.
Other medical applications involve the detection of outliers in transcript usage that are caused by patient specific genomic mutations for instance in the context of Mendelian diseases.
Finally, with the advent of single cell RNA-seq technologies the expression in single cells can be assessed, and for particular applications it is relevant to assess differences in detection, i.e. in the fraction of cells that express a particular transcript or gene.

For each of these applications bespoke tools have been developed.
However, we argue that these tools are suboptimal and computationally inefficient. Instead, we propose to build upon the popular negative binomial bulk RNA-seq tools and show how we can unlock them for differential detection, differential usage and aberrant splicing analysis by introducing offsets, which render the interpretation of the mean model parameters towards ratios.
This enables us to build a common workflow based on well-established software that only requires us to change the input counts and offsets to tailor it towards specific applications.
Finally, we also show how we can develop even more scalable algorithms for the unbiased estimation of the mean model parameters of count data by building upon quasi-likelihood and a smart choice of the mean variance relation.

Università degli Studi di Padova

Dipartimento di Scienze Statistiche