Simulated Regression Algorithm for Transcriptome Quantification

Adrian Caciula[1], Olga Glebova[1], Alexander Artyomenko[1], Serghei Mangul[2],
James Lindsay[3], Ion I. Măndoiu[3], and Alex Zelikovsky[1]

[1] Georgia State University, Atlanta GA, 30303, USA
[2] University of California, Los Angeles CA, 90095, USA
[3] University of Connecticut, Storrs CT, 06269, USA

RNA-Seq is a cost-efficient high-coverage powerful technology for transcriptome analysis. We propose a novel algorithm for transcriptome quantification from RNA-seq data (*SimReg*) which uses regression to find transcript frequencies for which the simulated read counts match the observed read counts.

SimReg first aligns the reads to existing transcript library and then counts the equivalent reads, i.e., reads aligned to the same set of transcripts. The bipartite graph with vertices corresponding to transcripts and read classes is split into connected components which can be treated independently. For each component we simulate high coverage reads and estimate $D_{\mathcal{R},\mathcal{T}} = \{d_{r,t}\}$, where $d_{r,t}$ is the portion of reads from transcript $t \in \mathcal{T}$ belonging to read class $r \in \mathcal{R}$.

Initial transcript frequencies are estimated by minimizing the squared deviation between observed read class frequency $O_{\mathcal{R}} = \{o_r\}$ and expected read class frequency $E_{\mathcal{R}} = D_{\mathcal{R},\mathcal{T}} \times F_{\mathcal{T}}$, where $F_{\mathcal{T}} = \{f_t\}$ are the portions of reads emitted by transcripts. The squared deviation is minimized by the following quadratic program: $\sum_{r\in\mathcal{R}} \left(\sum_{t\in\mathcal{T}} d_{r,t} f_t - o_r\right)^2 \rightarrow \min \mid \sum_{t\in\mathcal{T}} f_t = 1$ and $f_t \geq 0$.

Next *SimReg* repeatedly updates the frequency estimates by (1) simulating reads according to current estimates $F_{\mathcal{T}}$, (2) finding deviation between simulated and observed reads, $\Delta_{\mathcal{R}} = S_{\mathcal{R}} - O_{\mathcal{R}}$, (3) obtaining corrected read frequencies $C_{\mathcal{R}} = O_{\mathcal{R}} - \Delta_{\mathcal{R}}/2$, and (4) updating estimated transcript frequencies $F_{\mathcal{T}}$ based on corrected read class frequencies $C_{\mathcal{R}}$.

We tested *SimReg* on several test cases using simulated human RNA-Seq data. Experiments on synthetic RNA-seq datasets show that the proposed method improves transcriptome quantification accuracy compared to previous methods. The results show better correlation compared with currently best method *RSEM* [1].

References

1. Li, B., Dewey, C.: Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC bioinformatics **12**(1), 323 (2011)