Deterministic Regression Algorithm for Transcriptome Frequency Estimation

Adrian Caciula*, Olga Glebova*, Alexander Artyomenko*, Serghei Mangul†, James Lindsay‡,
Ion I. Mândoiu‡ and Alex Zelikovsky*
*Georgia State University, Atlanta, GA 30303
Emails: {acaciula, glebova, aartyomenko, alexz}@cs.gsu.edu
†University of California, Los Angeles CA, 90095
Email: serghei@cs.ucla.edu
‡ University of Connecticut, Storrs CT, 06269
Emails: {james.lindsay, ion}@engr.uconn.edu

Abstract—We present a deterministic version of our novel Monte-Carlo Regression based method MCReg [1] for transcriptome quantification from RNA-Seq reads. Experiments on simulated and real datasets demonstrate better transcriptome frequency estimation accuracy compared to that of the existing tools which tend to skew the estimated frequency toward super-transcripts.

I. METHOD

We propose a novel improvement for transcriptome quantification from RNA-seq data (MCReg2) which uses a deterministic method to compute expected read frequency, and regression to find transcript frequencies for which the simulated read counts match the observed read counts.

MCReg2 pipeline consists of the following steps. First, minimum fragment length and read length is estimated from reads uniquely mapped to transcript annotation. Then, based on estimated lengths, simulated reads for each position in each transcript are generated. The matching bipartite graph \( M = (T \cup R, E) \) is constructed where nodes correspond to reads and transcripts, and edges correspond to valid alignments of all observed reads to all transcripts. Next step is to map observed reads to read classes. This step is especially challenging if the following scenario is encountered: there is an empty simulated read class. The easiest way to resolve such case is to discard all corresponding observed reads, but such solution obviously worsens the transcriptome quantification results. The proposed solution consists of using a special option of Bowtie read alignment tool [2], namely "--best", and start checking whether reported "best" reads fit into any simulated classes. "Best" according to the Bowtie option is in terms of number of mismatches and in terms of the quality values at the mismatched position(s).

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 \( D_{R,T} = \{d_{r,t}\} \), where \( d_{r,t} \) is the portion of reads from transcript \( t \in T \) belonging to read class \( r \in R \).

By minimizing the squared deviation between observed read class frequency \( O_R = \{o_r\} \) and expected read class frequency \( E_R = D_{R,T} \times F_T \), initial transcript frequencies are estimated, where \( F_T = \{f_t\} \) are the portions of reads emitted by transcripts. The following quadratic program:

\[
\sum_{r \in R} (\sum_{t \in T} d_{r,t}f_t - o_r)^2 \rightarrow \min |\sum_{t \in T} f_t = 1 \text{ and } f_t \geq 0 \text{ minimizes the squared deviation.}
\]

II. RESULTS

We tested MCReg2 on both simulated and real human RNA-Seq datasets. Experiments on chromosome 1 synthetic RNA-seq dataset show that the proposed method improves transcriptome quantification accuracy compared to other similar methods. The results show better correlation (0.995) for MCReg2 compared with currently best methods MCReg (0.97), RSEM [3] (0.924), and IsoEM [4] (0.92).

For the real dataset a subset of human transcripts was used, where transcripts were quantified independently by NanoString assay [5] (a total of 109 genes were targeted by 141 distinct probes). MCReg2 reports a correlation of 0.8 showing a better performance than RSEM which reports only 0.75 correlation. However, for this particular dataset, the IsoEM performance is of overall best (0.85).

REFERENCES