Estimation of alternative splicing isoform frequencies from RNA-Seq data

Marius Nicolae[1], Serghei Mangul[2], Ion Măndoiu[1], and Alex Zelikovsky[2]

[1] Computer Science & Engineering Department, University of Connecticut
371 Fairfield Way, Storrs, CT 06269
{man09004,ion}@engr.uconn.edu
[2] Computer Science Department, Georgia State University
University Plaza, Atlanta, Georgia 30303
{serghei,alexz}@cs.gsu.edu

1 Introduction

In this paper we focus on the IE problem, namely estimating isoform expression levels (interchangeably referred to as frequencies) from RNA-Seq reads, under the assumption that a complete list of candidate isoforms is available. While current transcript libraries are still incomplete, we expect their coverage to increase rapidly. Besides facilitating finer-resolution studies of isoform regulation, improvements in IE accuracy lead to direct improvements in GE estimates (estimating gene expression levels). Indeed, as shown in Section 3, genome-wide gene expression level estimates derived from isoform level estimates are much more accurate than those obtained directly from RNA-Seq data using isoform-oblivious GE methods such as the widely used counting of unique reads, the rescue method of [5], or the recent EM algorithm of [1]. Furthermore, improvements in IE accuracy are likely to also benefit methods for identification of novel transcripts based on iterative refinement frameworks similar to that proposed in [2].

Our main contribution is a novel EM algorithm for isoform frequency estimation from (any mixture of) single and paired RNA-Seq reads. A key feature of our algorithm, referred to as IsoEM, is that it exploits a largely ignored source of disambiguation information provided by the distribution of insert sizes, which is typically tightly controlled during library preparation as recommended by sequencing instrument manufacturers. The recently published [7] is the only other work we are aware of that incorporates insert size distribution in conjunction with paired read data. We show that modeling insert sizes is also highly beneficial in conjunction with single RNA-Seq reads.

Insert sizes contribute to increased estimation accuracy in two different ways. On one hand, insert sizes are an important source of disambiguation information. In IsoEM, insert lengths are combined with base quality scores, and, if available, read pairing and strand information to probabilistically allocate reads to isoforms during the expectation step of the EM algorithm. As in [4], the genomic locations of multireads are also resolved probabilistically in this step, further contributing to overall accuracy compared to methods that pre-select

a unique genomic location by ad-hoc filtering rules. On the other hand, insert size distribution is used to accurately adjust isoform lengths during frequency re-estimation in the M step of the EM algorithm; an equivalent adjustment was independently employed in the probabilistic model of [7].

We also present preliminary experimental results on synthetic datasets generated with various sequencing parameters and distribution assumptions. The results show that IsoEM algorithm significantly outperforms existing methods of isoform and gene expression level estimation from RNA-Seq data. Furthermore, we empirically evaluate the effect of sequencing parameters such as read length, read pairing, and strand information on estimation accuracy. We confirm the finding of [4] that, for a fixed total number of typed bases, longer reads are not necessarily better for estimation accuracy. In particular, for both single and paired read sequencing, 100bp reads are always dominated by 50bp reads. This suggests that there may be limited benefits from further increases in read length and that higher sequencing depth is more critical to expression estimation accuracy.

2 Methods

2.1 Read mapping

As with most RNA-Seq analyses, the first step of IsoEM is to map the reads. Our approach is to map the reads onto the library of known isoforms using any one of the many available aligners (we used Bowtie [3] with default parameters in our experiments). Although spliced alignment to the genome could also be employed (as was done in [7] – note that genome mapping is required for discovery of novel transcripts), preliminary experiments with TopHat [6] resulted in much fewer mapped reads and significantly increased mapping uncertainty, despite providing TopHat a complete set of annotated junctions and using paired reads of up to 100bp. Nevertheless, further increases in read length in conjunction with improvements in spliced alignment algorithms could make genome mapping more attractive in the future. To make our implementation compatible with both mapping approaches, coordinates of alignments onto known isoforms are converted to genome coordinates, and all subsequent operations are done in genome space.

2.2 Finding read-isoform compatibilities

The candidate set of isoforms for each read is obtained by putting together all genome coordinates for reads and isoforms, sorting them and using a line sweep technique to detect read-isoform compatibilities. During the line sweep, reads are grouped into equivalence classes defined by their isoform compatibility sets. This speeds up the E-steps of the EM algorithm by allowing the processing of an entire read class at once.

2.3 EM algorithm

The EM algorithm starts with the set of N known isoforms. For each isoform we denote by $l(j)$ its length and by $f(j)$ its (unknown) frequency. If we ignore library preparation and amplification biases, the probability that a read is sampled from isoform j is proportional with $(l(j) - \mu + 1)f(j)$ where μ is the mean fragment length from the sample preparation.

Thus, if the isoform of origin is known for each read, then the maximum likelihood estimator for $f(j)$ is given by $c(j)/(c(1) + \ldots + c(N))$, where $n(j)$ denotes the number of reads sampled from isoform j and $c(j) = n(j)/(l(j) - \mu + 1)$ is its length-normalized read coverage. Unfortunately, some reads match equally well multiple isoforms, so their isoform of origin cannot be established unambiguously. The EM algorithm (see Algorithm 1) overcomes this problem by simultaneously estimating the frequencies and imputing the missing read origin within an iterative framework.

Algorithm 1 EM algorithm

```
assign random values to all f(i)
while not converged do
  initialize all n(j) to 0
  for each read r do
    sum = Σ_{j∈compatible(r)} w_{r,j} f(j)
    for each isoform j compatible with r do
      n(j)+ = w_{r,j} f(j)/sum
    end for
  end for
  s = Σ_j n(j)/(l(j) − μ + 1)
  for each isoform j do
    f(j) = n(j)/(l(j) − μ + 1)/s
  end for
end while
```

Some of the reads can match multiple positions in the genome. We will call these positions alignments. Each alignment a can in turn be compatible with multiple isoforms that overlap at that position of the genome. For paired end reads, an alignment consists of the two positions where the two reads in the pair align with the genome. For each alignment we define three random variables: Q_a, F_a and O_a. $Q_a = P(a)$ represents the quality of the alignment against the genome computed from the sequencing quality scores of the associated read and from comparison with the genome sequence at the given position(s). $F_a = P(a|i)$ represents the probability of the fragment length needed to produce alignment a from isoform i. For paired end reads, the length of the fragment can be inferred from the positions of the two reads in isoform i. For single reads, we can only estimate a maximum fragment length: if the alignment is on the same strand as the isoform, we use the distance from the start of the alignment to the end

of the isoform, otherwise we use the distance from the end of the alignment to the start of the isoform. $O_a = P(a|i, o)$ is the probability of alignment a coming from isoform i if we know whether the technology used for sequencing always samples from the coding strand of the isoform (for pairs, whether the first read comes from the coding strand). Putting it all together, the "weight" of mapping read r on isoform i is obtained by summing the product of the three random variables over all the alignments of read r in isoform i: $w_{i,r} = \sum_a Q_a F_a O_a$.

3 Experimental Results

3.1 Simulation setup

We tested IsoEM on simulated human RNA-Seq data. The human genome sequence (hg18, NCBI build 36) was downloaded from UCSC together with the coordinates of the isoforms in the KnownGenes table. Genes were defined as clusters of known isoforms defined by the GNFAtlas2 table. The dataset contains a total of 66803 isoforms pertaining to 19372 genes.

We compared IsoEM to several existing algorithms for solving the IE and GE problems. For IE we included in the comparison the analogues of the Uniq and Rescue methods used for GE [5], an improved version of Uniq (UniqLN) that estimates isoform frequencies from unique read counts by normalizing with the adjusted isoform length that excludes ambiguous positions, and the RSEM algorithm of [4]. For the GE problem, the comparison included the Uniq and Rescue methods, our implementation of the EM algorithm described in [1] (GeneEM), and estimates obtained by summing isoform expression levels inferred by RSEM and IsoEM.

Frequency estimation accuracy was assessed using the *error fraction (EF)* and *median percent error (MPE)* measures used in [4] along with the coefficient of determination, r^2. Accuracy was computed against true frequencies, not against estimates derived from true counts, as in [4].

3.2 Comparison between methods

For 30M reads of length 25 generated under a geometric isoform distribution assumption, Figure 1 shows the error fraction at different thresholds for isoform and gene expression levels. This plot makes more apparent the relative performance of compared algorithms, as well as the significant difference in accuracy between IE and GE. The variety of methods included in the comparison allows us to discern the relative contribution of various algorithmic ideas to estimation accuracy. The importance of appropriate length normalization is demonstrated by the IE accuracy gain over Uniq of UniqLN – clearly larger than that achieved by ambiguous read reallocation as implemented in the IE version of Rescue. Proper length normalization is also the basis for the large accuracy gain of IsoEM and RSEM over isoform oblivious GE methods. The importance of modeling insert sizes even for single read data is underscored by the significant IE and GE accuracy gains of IsoEM over RSEM.

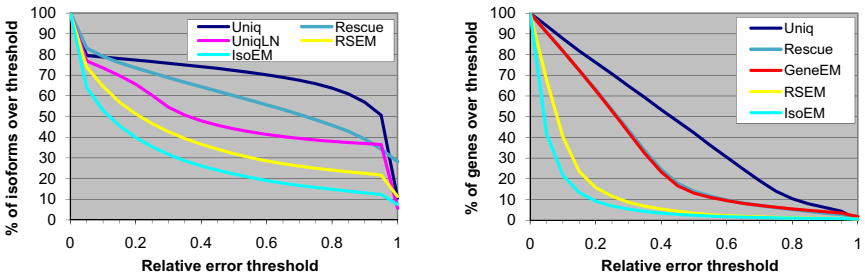

Fig. 1. Error fraction at different thresholds for isoform (left panel) and gene (right panel) expression levels inferred from 30M reads of length 25 simulated assuming geometric isoform expression.

Acknowledgments

This work was supported in part by NSF awards IIS-0546457, IIS-0916401, and IIS-0916948.

References

1. N. Zaitlen B. Paşaniuc and E. Halperin. Accurate estimation of expression levels of homologous genes in rna-seq experiments. In *Proc. RECOMB*, pages 397–409, 2010.
2. Wei Li Jianxing Feng and Tao Jiang. Inference of isoforms from short sequence reads. In *Proc. RECOMB*, 2010.
3. Ben Langmead, Cole Trapnell, Mihai Pop, and Steven Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.
4. Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, February 2010.
5. Ali Mortazavi, Brian A A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 2008.
6. Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, May 2009.
7. Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, May 2010.