, , pp. 1–2

doi:

Final Project

Nathalie Bonin,[1] Annie Dai[1] and Emma Shroyer[1]

[1]Computer Science Department, University of Maryland, Maryland, USA
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Salmon provides a fast and accurate way to estimate transcript abundance. This method uses a dual-phase inference algorithm and takes into account fragment biases for transcript verification. However, there are some areas where Salmon could improve. **?** The inference algorithms it uses, EM and VBEM, are only point estimates and cannot be checked if there is no ground truth available. Further, the choice of prior when using the VBEM algorithm can affect the accuracy of the results. In this work, we test how averaging the inference model predictions affect the accuracy of the inference estimation during Salmon's offline phase.

Key words: keyword1, Keyword2, Keyword3, Keyword4

Background

A key characteristic of many newly-formed eukaryotic RNA sequences is the inclusion of non-coding and coding regions, which are often referred to as introns and exons. Prior to translation, transcript maturation must occur via splicing, a process that revolves around the displacement of introns from the mRNA sequence. However, depending on environmental conditions and sequence variations, some exons are often removed from the mRNA as well, eventually resulting in the creation of a different protein isoform. Unfortunately, the many factors related to alternative splicing are still poorly understood for most eukaryotic proteins, hence the reasoning behind the continuation of transcriptome projects.

An essential facet to transcriptomics is the sequencing techniques used to generate transcript fragments for data analysis. Prior approaches to transcriptome quantification involved the use of microarrays, which has important limitations such as relatively high risks of cross-hybridization, low range of detection, and complicated normalization methods for differential expression analysis **?**. RNA-Seq, on the other hand, doesn't face such limitations and retains microarrays' advantage over classical sequencing such as high-throughput capabilities and the use of inexpensive technology. It has been revolutionary in furthering the field of transcriptomics, specifically with regards to eukaryotic cell transcriptomes. In particular, recent RNA-seq experiments yielded meaningful results in quantifying expression difference across different tissues **?**, distinguishing host-pathogen interactions **?**, and identifying the effect of structural variants on splicing regulators **?**.

Equally important to transcriptomics is abundance quantification. Estimations can be done through an alignment to a reference transcriptome sequence; nevertheless, a constant struggle with quantifying eukaryotic samples is to map read fragments back to the original transcript as opposed to the original. Although the detection of protein family abundances can be useful on its own, oftentimes researchers are more so interested in transcript abundances. Since isoforms of the same protein will retain common exons, identifying the source of each read is a tedious process that requires complex statistical methods. In the past, Salmon was developed to address this particular issue. Given a known transcript and a set of sequenced fragments, it quantifies the relative abundance of each transcript in the sample using a dual-phase inference procedure. Salmon improves accuracy compared to other models, because it takes into account sample-specific biases to improve accuracy. When these biases are not accounted for, calculations like the false discovery rate cannot be controlled for. Using multiple inference steps, Salmon improves its abundance estimates using either the VBEM or EM inference algorithms. However, both methods have drawbacks. Both algorithms return point estimates of abundances, and we are uncertain about the returned estimates without ground truth. Further, the choice of Bayesian priors used in the VBEM algorithm also has considerations. A small prior leads to sparser results than EM. However, a larger prior may result in more estimated non-zero abundance than EM. Prior simulated tests [Salmon documentation] show that VBEM with a small prior can lead to more accurate estimates. However, there has been much research into improving the results of these algorithms such as model averaging [citation]. Given the trade offs of the two algorithms, we sought to know if the averaged prediction for EM and VBEM inference is better than one algorithm or another. Further, if this ensemble estimate is better, could it be improved by varying the Bayesian priors. We have found that the average of these outputs [tie in results]

EM & VBEM

The Expectation Maximization (EM) algorithm optimizes the likelihood of the parameters given the data. This algorithm returns a point estimate of the abundances. This algorithm is the default of Salmon.

The Variational Bayesian Expectation Maximization (VBEM) algorithm accounts for the sparsity of the data. It takes a Bayesian nucleotide prior that controls for the sparsity of the data. The default prior for Salmon was 1×10^{-2}.

Each transcriptome $\mathcal{T}$ is composed of transcripts t. Each transcript is a nucleotide sequence which can be described through its length l, effective length $\tilde{l}$, and its count c—which is the number of times that t occurs in a given sample. There are M transcripts in a given transcriptome.

The probability that a given fragment originates from a transcript t depends on the length of that transcript relative to all other transcripts in the transcriptome. We define this nucleotide fraction η for a given transcript t as

$$\eta_i = \frac{c_i \cdot \tilde{l}_i}{\sum_{\boldsymbol{j}=1}^{M} c_j \cdot \tilde{l}_j}$$

We obtain a transcript fraction τ for a given transcript t by normalizing its nucleotide fraction against the effective length of all transcripts.

$$\eta_i = \frac{c_i \cdot \tilde{l}_i}{\sum_{\boldsymbol{j}=1}^{M} c_j \cdot \tilde{l}_j}$$

Let's say that the true nucleotide fraction for a transcript t is $\boldsymbol{\eta}$. We can describe the probability of observing a set of sequenced fragments $\mathcal{F}$ as a

$$\Pr\left(\mathcal{F}|\boldsymbol{\eta}, Z, \mathcal{T}\right) = \prod_{j=1}^{N} \Pr\left(f_j|\boldsymbol{\eta}, Z, \mathcal{T}\right) = \prod_{j=1}^{N}\sum_{i=1}^{M} \Pr\left(t_i|\boldsymbol{\eta} \cdot \Pr\left(f_j|t_i, z_{ij} = 1\right)\right)$$

where N is the number of fragments in $\mathcal{F}$, Z is a relationship matrix where $z_{ij} = 1$ when fragment f_j is derived from t_i.

We want to obtain $\boldsymbol{\alpha}$, which is the estimated number of reads from each transcript. We describe the maximum likelihood estimates as:

$$\mathcal{L}\left(\boldsymbol{\alpha}|\mathcal{F}, \boldsymbol{Z}, \mathcal{T}\right) = \prod_{j=1}^{N}\sum_{i=1}^{M} \hat{\eta}_i \Pr\left(f_j|t_i\right)$$

Written in terms of equivalence classes $\boldsymbol{C}$

$$\mathcal{L}\left(\boldsymbol{\alpha}|\mathcal{F}, \boldsymbol{Z}, \mathcal{T}\right) = \prod_{\mathcal{C}^j \in \boldsymbol{\mathcal{C}}} \left(\sum_{t_i \in \boldsymbol{t}^j} \hat{\eta}_i w_t^j\right)^{d^j}$$

The abundances $\hat{\boldsymbol{\eta}}$ are computed directly from $\boldsymbol{\alpha}$

$$\hat{\eta}_i = \frac{\alpha_i}{\sum_j \alpha_j}$$

Then we apply an update function

$$\alpha_i^{u+1} = \sum_{\mathcal{C}^j \in \boldsymbol{\mathcal{C}}} d^j \left(\frac{\alpha_i^u w_i^j}{\sum_{t_k \in \boldsymbol{t}} j \alpha_k^u w_k^j}\right)$$

Until the maximum relative difference in $\boldsymbol{\alpha}$ is

$$\Delta\left(\alpha^u, \alpha^{u+1}\right) = \max \frac{\left|a_i^u - \alpha_i^{u+1}\right|}{\alpha_i^{u+1}} < 1 \times 10^{-2}$$

for all $\alpha_i^{u+1} > 1 \times 10^{-8}$, at which point we derive the estimated nucleotide fraction

$$\hat{\eta}_i = \frac{\alpha_i'}{\sum_j \alpha_j'}$$

Variational Bayes Optimization

Optionally, the we can apply variational bayeseian optimization where the update function is

$$\alpha_i^{u+1} = \sum_{\mathcal{C}^j \in \boldsymbol{\mathcal{C}}} d^j \left(\frac{e^{\gamma_i^u} e_i^j}{\sum_{t_k \in \boldsymbol{t}} j e^{\gamma_k^u} w_k^j}\right)$$

where

$$\gamma_i^u = \Psi\left(\sum_{k=1}^{M} \alpha_k^0 + \alpha^u + _k\right)$$

where Ψ is the digamma function and the expected value of nucleotide fractions can be expressed as

$$\mathbb{E}\left(\eta_i\right) = \frac{\alpha_i^0 + \alpha_i'}{\sum_j \alpha_j^0 + a_j'} = \frac{a_i^0 + \alpha_i'}{\hat{\alpha}^0 + N}$$

$$\hat{\alpha}^0 = \sum_{i=1}^{M} \alpha_i^0$$

Datasets

We used synthetic datasets used in Salmon. We generated our synthetic data using polyester 1.16.0 and alpine version 1.6.0. [citation] There was ground truth available for this dataset.

We additionally used real human datasets. We used a dataset consisting of NK cells, T cells, and tumor from matched soft tissue sarcoma and peripheral blood.

We also used the human transcriptome reference: Homo sapiens GRCh38 cDNA sequence. These were gene models built from alignments of the human proteome and alignments of human cDNAs.

Methods

Experiment

We began by editing Salmon's source code. We wrote a bash script that downloaded our datasets. We also had a script that ran our samples using EM, VBEM, and a combination of the VBEM and EM algorithms. First we ran unaltered Salmon using EM and VBEM. For each pass that used VBEM, we specified a prior 10-5 through 101. We then ran the same datasets using the average of both the EM and VBEM algorithms. We ran the EM and VBEM algorithms on the same data, producing an array. FOr each index in this array, the data was averaged and inserted into a new array that was used the final calculations. We again used priors of 10-5 through 101 when considering the VBEM algorithm. We ran this on [Nathalie's laptop] Finally, we compared the results of these passes using the ground truth of the synthetic data. We calculate the MARD of our results. We also calculated the Spearman Correlation and a two-sided Mann-Whitney U test.

Results

Initially, we ran EM and VBEM with varying priors by themselves without any alterations to Salmon. We found the following results:

Evaluation

$$\mathrm{ARD}_i = \begin{cases} 0 & \text{if} \quad x_i = y_i = 0 \\ \dfrac{|x_i - y_i|}{x_i - y_i} & \text{otherwise} \end{cases}$$

$$\mathrm{MARD} = \frac{1}{M} \sum_{i=1}^{M} \mathrm{ARD}_i$$

Spearman Correlation

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Mann–Whitney U Test

Discussion

We were unable to expand on our project with the time given but in future work, we would like to expand upon ways to best combine the results of the EM and VBEM algorithms. Currently we are only taking the mean of the two results. However, we would also like to experiment with the weighted average of these two algorithms also taking into account various prior sizes. We would like to see if this also improves the estimates of Salmon.

We would also like to verify the time it takes to run our combined approach. We would like to verify that running this approach will not greatly impact the runtime of Salmon. While we expect the runtime to increase; however, we would like to verify that it is not detrimental to the use of Salmon.

In future work, it will be beneficial to get a measure of uncertainty when running our trials. We were unable to run this during our initial experiments. However, we would like to know if there is a level of uncertainty with our results when we use the average of the EM and VBEM algorithms. We would like to know if this uncertainty increases or decreases, and if it does so significantly.