Journal Title Here, 2022, pp. 1–4

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Final Paper

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

[1]Computer Science Department, University of Maryland, Maryland, USA

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: Salmon, Tuna, Onefish, Twofish, Redfish, Bluefish

Introduction

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.

One 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, does not 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 gene. 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 [**?**] 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 [**?**]. 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 did not result in conclusive results.

```
salmon quant -i human_index -l IU -1 sample_1.fa.gz -2 sample_2.fa.gz -p 6 --gcBias --useEM \
    --validateMappings -o sample
salmon quant -i human_index -l IU -1 sample_1.fa.gz -2 sample_2.fa.gz -p 6 --gcBias --vbPrior \
    1e{from -5 to 1} --validateMappings -o sample
salmon quant -i human_index -l IU -1 sample_1.fa.gz -2 sample_2.fa.gz -p 6 --gcBias --useBoth \
    --vbPrior 1e{from -5 to 1} --validateMappings -o sample
```

Fig. 1. Commands used for the quantification of each sample using EM, VBEM, or both inference strategies, respectively.

We saw that using the average is very similar to EM when used with a higher prior and more similar to the VBEM output when used with a higher prior. In light of these results we have also included a section on future work.

Datasets

We used the same synthetic dataset from Love's 2018 method article on DTU analysis, which consists of 24 sets of unstranded inward pair-ended reads [**?**]. This data was generated using polyester 1.16.0 and alpine version 1.6.0. [**?**] We also utilized the Homo sapiens GRCh37 version 28 sequence as our human transcriptome reference, which was the same reference used for the generation of the simulated read. [**?**]

Methods

Combining the traditional EM algorithm with the VBEM algorithm

Since we are only interested in optimizing abundance estimation in the offline phase, we opted to directly change Salmon v1.10.3's source code. We introduce a new flag, `--useBoth`, which forces the program to run the offline phase twice: once with the traditional EM algorithm and once with the VBEM algorithm. The program then computes the average of the two different estimates for the $\boldsymbol{\alpha}$ vector before computing its final prediction for $\boldsymbol{\eta}$.

Running the new implementation on the polyester dataset

The commands used for the quantification of each sample using EM, VBEM, or both inference strategies, are provided in Fig 1.

Quantification was run on the 24 samples through a slurm batch script. For each run, 5GB of memory was allocated on the CBCB's Nexus partition. In total, every sample had 15 runs, each with different specifications for the inference method (1 EM, 7 VBEM, and 7 combo) and the prior value if applicable (10^{-5} to 10^{1}). All other parameters were kept the same among the 360 total runs:

1. We kept the selective alignment feature on for this experiment. This strategy is currently the default for the pipeline, and it allows Salmon to adopt a more sensitive approach during quasi-mapping.
2. In recent versions of Salmon, the library type can be automatically detected. Nevertheless, we elected to specify the library type for all the runs.
3. The `–gcBias` flag was passed to the pipeline as well, which allows Salmon to identify and remedy for GC biases. This extra step is essential for the analysis of our synthetic data since they were simulated from an already existing GC bias profile [**?**].

Results

We encountered different results depending on the prior selected. Overall, we can separate the outcome of combining both inference algorithms into two categories. The first of such categories includes runs with a prior between 10^{-5} and 10^{1}, inclusive. As shown in figures 1 and 2, both the MARD values and the Spearman correlation values are lower for estimates inferred from VBEM than for estimates inferred from both methods (Mann-Whitney U test, $P \leq 1.436 \cdot 10^{-7}$ for MARD, $P < 6.202 \cdot 10^{-14}$ for Spearman). However, MARD values tended to be similar between estimates inferred from EM and estimates inferred from both methods (Mann-Whitney U test, $P \geq 0.4803$), whereas the Spearman correlation values were slightly higher for EM (Mann-Whitney U test, $P \leq 0.03941$). Because the MARD metrics are negatively impacted by false discovery rates [**?**], we hypothesize that the results gathered from running both inference algorithms is affected by EM's tendency to collect more non-zero abundances than VBEM at low prior values.

This can also be concluded from the analysis of our second category, which includes runs with a prior of 1 or 10. Unlike with the first category, the MARD and the Spearman correlation values are strikingly lower for the estimates inferred from EM (Mann-Whitney U test, $P < 6.202 \cdot 10^{-14}$ for MARD, $P < 6.202 \cdot 10^{-14}$ for Spearman), whereas the difference in MARD between VBEM and the combination of both methods was insignificant for runs with a prior of 1 (Mann-Whitney U test, $P = 0.8944$). The difference between the Spearman correlation values and between the MARD values for runs with a prior of 10 was still significant (Mann-Whitney U test, $P = 5.675 \cdot 10^{-11}$ for MARD, $P < 6.202 \cdot 10^{-14}$ for Spearman). We predict that this would suggest that for high prior values, the negative impact from false discovery rates by the VBEM algorithm is offsetted by the more accurate abundance estimates from the EM algorithm.

Statistical Evaluation

The statistical evaluation of the different inference algorithms was performed with R. Because only the original simulated read counts were made available to, we conducted our comparisons on the estimated number of reads given by Salmon. We chose two statistical metrics for our analysis: The mean absolute relative difference, or MARD, was already introduced in the original Salmon publication. It is the average of the absolute relative difference ARD for each transcript i, where x_i is the original simulated read count, and y_i is the read count reported

Table 1. Spearman correlation table for first four samples and their subsamples.

	Sample 1_{01}	Sample 1_{02}	Sample 2_{01}	Sample 2_{02}	Sample 3_{01}	Sample 3_{02}	Sample 4_{01}	Sample 4_{02}
EM	0.932	0.932	0.933	0.932	0.937	0.936	0.934	0.933
both - 10^{-5}	0.933	0.933	0.934	0.933	0.939	0.938	0.935	0.934
VBEM - 10^{-5}	0.952	0.951	0.953	0.951	0.957	0.956	0.954	0.952
both - 10^{-4}	0.933	0.933	0.934	0.933	0.939	0.938	0.935	0.934
VBEM - 10^{-4}	0.952	0.951	0.953	0.951	0.957	0.956	0.954	0.952
both - 10^{-3}	0.933	0.933	0.934	0.933	0.939	0.938	0.935	0.934
VBEM - 10^{-3}	0.952	0.951	0.953	0.951	0.957	0.956	0.954	0.952
both - 10^{-2}	0.933	0.933	0.934	0.933	0.939	0.938	0.935	0.934
VBEM - 10^{-2}	0.952	0.951	0.953	0.952	0.957	0.956	0.954	0.952
both - 10^{-1}	0.933	0.933	0.934	0.933	0.939	0.937	0.935	0.934
VBEM - 10^{-1}	0.951	0.950	0.953	0.951	0.957	0.955	0.954	0.951
both - 10^{0}	0.755	0.754	0.755	0.754	0.755	0.755	0.755	0.754
VBEM - 10^{0}	0.751	0.751	0.751	0.751	0.752	0.752	0.751	0.751
both - 10^{1}	0.736	0.735	0.736	0.736	0.738	0.737	0.736	0.736
VBEM - 10^{1}	0.715	0.715	0.715	0.716	0.717	0.717	0.715	0.715

by Salmon.

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

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

The Spearman correlation, also described in the original publication, calculates the relationship between the rank for our results and the rank for the simulated data. It was computed through the core.test function where we indicated a two sided alternative hypothesis.

Spearman Correlation

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

Mann–Whitney U Test

MATH Additionally, we were also required to measure the differences between the three inference algorithms. We used the Mann-Whitney test (also known as the Wilcoxon test) to calculate the p value for such analysis, and computations were also performed in R with the wilcox.test function

Discussion

We were unable to definitively say whether our method improved on Salmon. The output of using the average of the EM and VBEM results produced similar results to just EM or VBEM by themselves. A lower prior, 10^{-5} to 10^{-1}, resulted in similar results to EM while priors greater than 10^{0} resulted in similar results to VBEM. At no point does it appear that using the average of both methods significantly improved the output, but it also did not significantly harm the results. This is likely attributed to our method of combining EM and VBEM. Even though we did not find conclusive results, there are next steps to further this work.

Future Work

In our results we saw that using the average of both VBEM and EM is very similar to EM when used with a lower prior and more similar to the VBEM output when used with a higher prior. We have hypothesized this is because we were taking the average even when one of the results was a zero. In future work, we would like to prioritize zero abundances over non-zero abundances. Instead of taking the average between value and zero, we want to record the value as zero. If neither result is a zero, we wish to take the average of those values. We believe this could improve our estimates.

In the current timeframe, we were unable to expand on our projec, but in future work, would like to expand upon ways to best combine the results of the EM and VBEM algorithms. Currently, our code is only taking the mean of the two results. However, we would also like to experiment with the weighted average or the Bayesian model average of these two algorithms also taking into account various prior sizes. We hypothesize that this would further improve Salmon's results [**?**]. Additionally, we also would like to test our approach on real life data. Although we had picked out a dataset consisting of sequencing reads from NK cells, T cells, and tumor collected from matched soft tissue sarcoma and peripheral blood [**?**], we were pressed for time and resources, and therefore we had to abandon this portion of our project.

We would also like to verify the time it takes to run our combined approach. It would be beneficial to run multiple trials to gain an idea of the time it takes to run Salmon using our solution. We would like to verify that running this approach will not greatly impact the runtime of Salmon. The runtime is expected to increase, since we are running both the EM and VBEM algorithms. However, we would like to verify that it is not detrimental to the use of Salmon.

In future work, it will be informative 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 so, is the change in uncertainty significantly. Salmon already supports this analysis through the use of a flag.

Supplementary

Fig. 2. MARD table for first four samples and their subsamples.

Author contributions statement

Must include all authors, identified by initials, for example: S.R. and D.A. conceived the experiment(s), S.R. conducted the experiment(s), S.R. and D.A. analysed the results. S.R. and D.A. wrote and reviewed the manuscript.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. This work is supported in part by funds from the National Science Foundation (NSF: # 1636933 and # 1920920).