Evaluation Project Documentation

Edited by Bo Li

September 20, 2011

1 Introduction

We use the graphical model showed in Figure 1. We denote the observed data as D. We denote $|D|$ as N, number of reads in the data. We assume there are M transcripts in the reference and they are numbered from 1 to M. In addition, our model has an extra "noise" transcript to account for reads coming from background noise, numbered as 0. θ is the probability distribution of a read is sequenced from a particular transcripts. We have $\theta_i = \tau_i l_i, i = 1...M$ and $|\theta| = M + 1$. For details about RSEM model, please see reference[1][2].

We denote an assembly (the reference set used in RSEM's model) as A.

This project's goal is to evaluate which assemble method performs better, given a fixed data set D. That is to say, we want to find a function f, such that given any two assemblers, for their assemblies A_1 and A_2 made from D, we have :

$$f(A_1) > f(A_2) \Leftrightarrow A_1 \text{ is better than } A_2$$

Currently, we have four candidates for f. They are likelihood score, BIC, model evidence by Monte Carlo approximation and model evidence by convex approximation. We want to show that the latter three performs better than the first one. Ideally, we also want to find that the latter two are better than BIC.

In the following four sections, I'll describe the four measures. In addition, I'll omit notation A int all following formulae. We just need to know for all formulae, "given A" is omitted.

2 Loglikelihood

First, pick up θ_{MLE} (MLE means maximum likelihood estimator) :

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \log P(D|\theta)$$

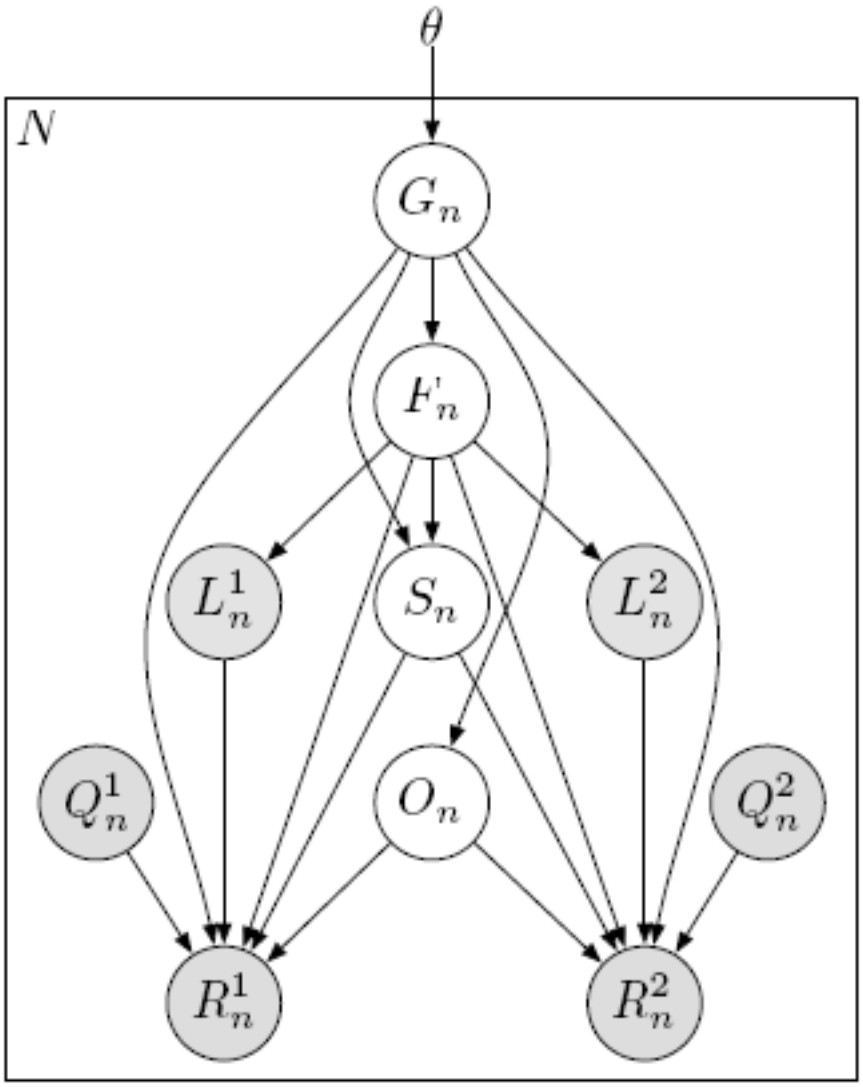

Figure 1: RSEM's graphical model

Then Loglikelihood score is defined as

$$\log P(D|\theta_{MLE})$$

3 Bayesian information criterion

We are interested in $P(D)$, model evidence.

$$P(D) = \int P(D|\theta)p(\theta)d\theta$$

Under certain condition [3], by the Laplace approximation, we have

$$\log P(D) \simeq \log P(D|\theta_{MAP}) + \log P(\theta_{MAP}) + \frac{M}{2}\log(2\pi) - \frac{1}{2}\log|H|$$

θ_{MAP} is the Maximum a posteriori estimator. It is the model of the posterior distribution $P(\theta|D)$. H is the Hessian matrix of second derivatives of the negative log posterior at θ_{MAP}.

If we further assume the Gaussian prior, then in the asymptotic case, we have

$$\log P(D) \simeq \log P(D|\theta_{MAP}) - \frac{1}{2}M\log N$$

The above formula is what we used in this project for BIC. Because we assume θ follows $Dir(1)$, the MAP estimator is the same as MLE estimator.

However, because the "certain condition" is not satisfied here, we do not prefer this measure.

For details, please read P213-P217 of Pattern Recognition and Machine Learning(PRML).

4 Model evidence by Monte Carlo approximation

Our goal is to compute the *model evidence*, $P(D)$. Using Bayes rule, we can express the model evidence as

$$P(D) = \frac{P(D|\theta')P(\theta')}{P(\theta'|D)} \tag{1}$$

Here, θ' can be any particular value of the parameters. For example, we might choose $\theta' = \theta_{PME}$ for numerical issues. PME means posterior mean estimator.The numerator of this fraction is easily computed, as it is simply the product of the likelihood and the prior. The challenge is to compute the denominator, $P(\theta'|D)$. One way to compute this value is via sampling of the latent variables,

Z, from their posterior distribution:

$$\begin{aligned} P(\theta'|D) &= \sum_z P(\theta', z|D) && (2)\\ &= \sum_z P(\theta'|z, D)P(z|D) && (3)\\ &= \sum_z P(\theta'|z)P(z|D) && (4)\\ &\approx \frac{1}{N_s}\sum_{i=1}^{N_s} P(\theta'|z^{(i)}) && (5) \end{aligned}$$

where $z^{(1)}, \ldots, z^{(N_s)}$ are samples from $P(z|D)$, possibly via Gibbs sampling.

After we get $P(D)$, the f is defined as $\log P(D)$.

5 Model evidence by convex approximation

This is another way to approximate $P(D)$ and our goal is to calculate $\log P(D)$ here, too.

Refresh: There are N reads and M transcripts. So $|\theta| = M + 1$ (including the noise transcript).

The data likelihood $\log P(D|\theta)$ can be decomposed as follows:

$$\begin{aligned} \log P(D|\theta) &= \sum_Z q(Z) \log \frac{P(D, Z|\theta)}{P(Z|D, \theta)} \frac{q(Z)}{q(Z)}\\ &= \sum_Z q(Z) \log \frac{P(D, Z|\theta)}{q(Z)} + \sum_Z q(Z) \log \frac{q(Z)}{P(Z|D, \theta)}\\ &= F(q, \theta) + KL(q(Z)||P(Z|D, \theta))\\ F(q, \theta) &= \sum_Z q(Z) \log \frac{P(D, Z|\theta)}{q(Z)} \end{aligned}$$

For any given θ^*, let $q(Z) = P(Z|D, \theta^*)$, we have

$$\log P(D|\theta) \geq F(P(Z|D, \theta^*), \theta)$$

In addition, when $\theta = \theta^*$, $logP(D|\theta) = F(P(Z|D, \theta^*), \theta)$ for that $KL(q(Z)||P(Z|D, \theta)) = 0$.

Therefore, assume a dirichlet prior of $\alpha_i = 1$, we have $P(D) \geq \int_\theta p(\theta) e^{F(P(Z|D,\theta^*),\theta)} d\theta$ for any θ^*. We use $\theta^* = \theta_{MLE}$.

Because

$$F(P(Z|D, \theta^*), \theta) = \sum_{i=0}^{M} c_i^* \log \theta_i + \sum_Z P(Z|D, \theta^*) \log \frac{P(D|Z)}{P(Z|D, \theta^*)}$$

We have

$$\begin{aligned} P(D) &\geq \int_\theta p(\theta) e^{F(P(Z|D,\theta^*),\theta)} d\theta \\ &= e^{\sum_Z P(Z|D,\theta^*) \log \frac{P(D|Z)}{P(Z|D,\theta^*)}} \int_\theta p(\theta) \prod_{i=0}^{M} \theta_i^{c_i^*} d\theta \\ &= e^{\sum_Z P(Z|D,\theta^*) \log \frac{P(D|Z)}{P(Z|D,\theta^*)}} \frac{\Gamma(M+1)\prod_{i=0}^{M}\Gamma(c_i^*+1)}{\Gamma(M+1+N)} \end{aligned}$$

So

$$\begin{aligned} \log P(D) &\geq \log\Gamma(M+1) + \sum_{i=0}^{M} \log\Gamma(c_i^*+1) - \log\Gamma(M+1+N) + \sum_Z P(Z|D,\theta^*) \log \frac{P(D|Z)}{P(Z|D,\theta^*)} \\ &= \log\Gamma(M+1) + \sum_{i=0}^{M} \log\Gamma(c_i^*+1) - \log\Gamma(M+1+N) + \sum_{n=1}^{N} \sum_{z_{ni}\in\pi_n^x} P(z_{ni}|r_n,\theta^*) \log \frac{P(r_n|z_{ni})}{P(z_{ni}|r_n,\theta^*)} \end{aligned}$$

To have a better understand of this part, I'd suggest to read P450-P455 of PRML.

6 Comparison of Approx score and BIC

Let $\theta^* = \theta_{MLE}$.

Because $\theta_{MAP} = \theta_{MLE}$ if $\theta \sim Dir(1)$, we have

$$\begin{aligned} BIC &= \log P(D|\theta_{MAP}) - \frac{1}{2} M \log N \\ &= \log P(D|\theta_{MLE}) - \frac{1}{2} M \log N \\ &= \log P(D|\theta^*) - \frac{1}{2} M \log N \end{aligned}$$

For *Approx*, we have

$$Approx = \log\Gamma(M+1) + \sum_{i=0}^{M} \log\Gamma(c_i^*+1) - \log\Gamma(M+1+N) + \sum_Z P(Z|D,\theta^*) \log \frac{P(D|Z)}{P(Z|D,\theta^*)}$$

From section 5, we have $logP(D|\theta^*) = F(P(Z|D,\theta^*),\theta^*)$, where

$$F(P(Z|D,\theta^*),\theta^*) = \sum_{i=0}^{M} c_i^* \log\theta_i^* + \sum_Z P(Z|D,\theta^*) \log \frac{P(D|Z)}{P(Z|D,\theta^*)}$$

So

$$\sum_Z P(Z|D,\theta^*)\log\frac{P(D|Z)}{P(Z|D,\theta^*)} = logP(D|\theta^*) - \sum_{i=0}^{M} c_i^*\log\theta_i^*$$

Therefore

$$Approx = \log P(D|\theta^*) - [\log\Gamma(M+1+N) - \log\Gamma(M+1) - \sum_{i=0}^{M}\log\Gamma(c_i^*+1) + \sum_{i=0}^{M} c_i^*\log\theta_i^*]$$

According to Stirling's approximation,

$$\log n! \sim \frac{1}{2}\log(2\pi n) + n\log n - n$$

For the penalty term of Approx, we have

$$\begin{aligned}
& \log\Gamma(M+1+N) - \log\Gamma(M+1) - \sum_{i=0}^{M}\log\Gamma(c_i^*+1) + \sum_{i=0}^{M} c_i^*\log\theta_i^* \\
\sim\ & [\frac{1}{2}\log 2\pi + (M+N+\frac{1}{2})\log(M+N) - (M+N) - (\frac{1}{2}\log 2\pi + (M+\frac{1}{2})\log M - M)] \\
& -[\sum_{i=0}^{M}(\frac{1}{2}\log 2\pi + (c_i^*+\frac{1}{2})\log c_i^* - c_i^*) - \sum_{i=0}^{M} c_i^*\log\frac{c_i^*}{N}] \\
=\ & [(M+N+\frac{1}{2})\log(M+N) - (M+\frac{1}{2})\log M - N] \\
& -[\frac{M+1}{2}\log 2\pi + \frac{1}{2}\sum_{i=0}^{M}\log c_i^* - N + N\log N] \\
\geq\ & [(M+N+\frac{1}{2})\log(M+N) - (M+\frac{1}{2})\log M - N] - [\frac{M+1}{2}\log 2\pi + \frac{M+1}{2}\log\frac{N}{M+1} - N + N\log N] \\
=\ & [(M+\frac{1}{2})\log(M\frac{M+N}{M}) - (M+\frac{1}{2})\log M] + [N\log(N\frac{M+N}{N}) - N\log N] \\
& -\frac{M+1}{2}\log\frac{N}{M+1} - \frac{M+1}{2}\log 2\pi \\
=\ & (M+\frac{1}{2})\log\frac{M+N}{M} - \frac{M+1}{2}\log\frac{N}{M+1} + N\log\frac{M+N}{N} - \frac{M+1}{2}\log 2\pi \\
>\ & \frac{M}{2}\log\frac{M+N}{M} + N\log(1+\frac{M}{N}) - \frac{M+1}{2}\log 2\pi \\
=\ & \frac{M}{2}\log(N\frac{M+N}{NM}) + N\log(1+\frac{M}{N}) - \frac{M+1}{2}\log 2\pi \\
=\ & \frac{M}{2}\log N + [\frac{M}{2}\log\frac{\frac{M}{N}+1}{M} + N\log(1+\frac{M}{N}) - \frac{M+1}{2}\log 2\pi]
\end{aligned}$$

The $\geq$ is by applying Jensen's Inequality to $\sum_{i=0}^{M}\log c_i^*$ and the equality is reached by taking $c_i^* = \frac{N}{M+1}$.

The $>$ is due to the fact that $\frac{M+N}{M} > \frac{N}{M+1}$.

If $N \gg M$, $\frac{M}{2}\log N$ dominates the penalty term of Approx, and therefore Approx will behave like BIC.

However, if $N \sim cM$, where c is a constant, we have

$$\begin{aligned} penalty &> \frac{M}{2}\log cM + [\frac{M}{2}\log\frac{\frac{M}{cM}+1}{M} + cM\log(1+\frac{M}{cM}) - \frac{M+1}{2}\log 2\pi] \\ &= \frac{M}{2}\log M - \frac{M}{2}\log M + [\frac{M}{2}\log c + \frac{M}{2}\log(\frac{1}{c}+1) + cM\log(1+\frac{1}{c}) - \frac{M+1}{2}\log 2\pi] \\ &= O(M) \end{aligned}$$

That is to say, the penalty term is in the order of $O(M)$ instead of $O(M\log N)$.

However, please note that these results are for lower bounds of Approx's penalty term. We still need some the same results for upper bounds in order to claim the results. But at least, these results give us some intuition which match the plots we generated.

7 Overlap Length

Focus on single end reads only. Also assume that reads are strand-specific and there is no sequencing error.

Assume read length is r, overlap size is o and transcript/contig i has length l_i.

We have $\theta_i = \frac{c_i}{N}$, $\tau_i \propto \kappa_i = \frac{\theta_i}{l_i - r + 1}$ and BIC is equal to:

$$BIC = \sum_{i=0}^{M} c_i \log\frac{\theta_i}{l_i - r + 1} - \frac{1}{2}M\log N$$

Because when N is big, *Approx* and possibly *Gibbs* scores approach BIC, we only focus on BIC score here.

Suppose we want to combine contigs a and b. After combination, we have a new score:

$$BIC^{new} = \sum_{i=0, i\neq a,b}^{M} c_i \log\frac{\theta_i}{l_i - r + 1} + (c_a + c_b)\log\frac{\theta_a + \theta_b}{l_a + l_b - o - r + 1} - \frac{1}{2}(M-1)\log N$$

Let $\Delta = BIC^{new} - BIC$, we have

$$\begin{aligned}
\Delta &= (c_a + c_b)\log\frac{\theta_a+\theta_b}{l_a+l_b-o-r+1} - c_a\log\frac{\theta_a}{l_a-r+1} - c_b\log\frac{\theta_b}{l_b-r+1} + \frac{1}{2}\log N \\
&= (c_a + c_b)\log\frac{\kappa_a(l_a-r+1)+\kappa_b(l_b-r+1)}{l_a+l_b-o-r+1} - c_a\log\kappa_a - c_b\log\kappa_b + \frac{1}{2}\log N \\
&= c_a\log\frac{(l_a-r+1)+\frac{\kappa_b}{\kappa_a}(l_b-r+1)}{l_a+l_b-o-r+1} + c_b\log\frac{\frac{\kappa_a}{\kappa_b}(l_a-r+1)+(l_b-r+1)}{l_a+l_b-o-r+1} + \frac{1}{2}\log N
\end{aligned}$$

If $\tau_a = \tau_b$, we have $\kappa_a = \kappa_b$ and

$$\Delta = (c_a + c_b)\log\frac{l_a+l_b-2r+2}{l_a+l_b-o-r+1} + \frac{1}{2}\log N$$

If $o = r-1$, we have $\Delta = \frac{1}{2}\log N > 0$. This means that if two contigs have same expression level and overlap $r-1$, merge them will always increase the score.

When $0 \le o < r-1$, we need

$$c_a + c_b \le \frac{\frac{1}{2}\log N}{\log\frac{l_a+l_b-o-r+1}{l_a+l_b-2r+2}}$$

Because $\frac{1}{2}\log N$ will be a small number (if $N = 10^8$, $\frac{1}{2}\log N \approx 9$), it requires the counts of a and b be very small in order to reward the merge. For example, if $N = 10^8$ and you have two single reads with $o = 0$, it is always beneficial to merge them together.

In fact, contig length l_a and l_b also play roles. The longer the l_a and l_b, the bigger $c_a + c_b$ is allowed. Especially, when $l_a + l_b \to \infty$, overlap size does not affect and we should always merge the two together.

But the effect of l is much less than the effect of c due to l appears in the log term. However, we might also say that if both contigs are low expressors, we might want to merge them together to increase BIC score.

8 References

[1] Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., Dewey, C. N. (2010). **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics*, 26(4), 493-500.

[2] Li, B. and Dewey, C. N. **RSEM: accurate quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics*, 12:323. (Highly accessed)

[3] Schwarz, G. (1978). **Estimating the dimension of a model.** *Annals of Statistics* **6**, 461-464.

[4] Bishop, C. M. (2006). **Pattern recognition and machine learning**. *Springer*