Overview of RSEM

1 How to compute score(A)

1.1 Definitions

Let I be the number of possible isoforms of length less than a fixed constant L_0. Index the isoforms by i. Let $[I] = \{1, 2, \ldots, I\}$. Let len$(i)$ be the length of isoform i. Let base(i,k) be the base at position k of isoform i.

Let $A \subset [I]$ be an assembly. Let A' be another assembly.

Let $d(A,A') = A \Delta A'$, the symmetric difference.

Let $D = (R_1, \ldots, R_N)$ where R_n is a read. All reads are of length L. Let base(R_n,k) be the base at position k of read n.

Let $G = (G_1, \ldots, G_N)$ be the isoform indices; $G_n \in I$ is the index of the isoform that read n (putatively) comes from.

Let $O = (O_1, \ldots, O_N)$ be orientation indicators; $O_n \in \{0,1\}$ indicates the orientation of read n (putatively).

Let $S = (S_1, \ldots, S_N)$ be read start positions; read n (putatively) starts at base $S_n \in \mathbb{N}$ of its parent isoform.

Let $P(D,G,O,S|A) = \prod_{n=1}^{N} P(R_n|G_n,O_n,S_n)P(O_n|G_n)P(S_n|G_n)P(G_n|A)$. (We use the original RSEM model for now, because it is a bit simpler.)

Let $P(A|D) \propto P(D|A) = \sum_{(G,O,S)} P(D,G,O,S|A)$, where the sum is over all possible values.

Let score$(A') = \sum_A d(A',A)P(D|A) \propto \sum_A d(A',A)P(A|D)$ where the sum is over all possible assemblies. (Note that we define this backwards of what we intuitively want; that is ok because both forms are equally good for comparing scores.)

1.2 Distributional assumptions

Assume that all orientations are forward, so that $P(O_n = 0|G_n) = 1$. (Not crucial, but makes notation simpler.)

Assume that the start positions $S_n|G_n$ are uniform, so that $P(S_n|G_n) = 1/\text{len}(G_n)$. (Not crucial.)

Assume that the prior probabilities of each isoform are uniform within each assembly. Note that $P(G_n|A) = 0$ if $G_n \notin A$. Thus $P(G_n|A) = 1_A(G_n)1/|A|$. (The uniformity is not crucial.)

Assume that $P(R_n|G_n, O_n = 1, S_n) = \prod_{k=1}^{L} w_{k,\text{base}(R_n,k),\text{base}(G_n,S_n+k-1)}$, where w is some substitution matrix.

1.3 Fubini galore

Rewrite

$$\begin{aligned}
\text{score}(A') &= \sum_A d(A',A)P(D|A) \\
&= \sum_A d(A',A) \sum_{(G,O,S)} P(D,G,O,S|A) \\
&= \sum_A d(A',A) \sum_{(G,O,S)} P(R|G,O,S)P(O|G)P(S|G)P(G|A) \\
&= \sum_A d(A',A) \sum_{(G,O,S)} \prod_{n=1}^{N} P(R_n|G_n,O_n,S_n)P(O_n|G_n)P(S_n|G_n)P(G_n|A) \\
&= \sum_A |A'\Delta A| \sum_{(G,S)} \prod_{n=1}^{N} \Big(\prod_{k=1}^{L} w_{k,\text{base}(R_n,k),\text{base}(G_n,S_n+k-1)} \Big) \Big(1/\text{len}(G_n) \Big) \Big(1_A(G_n) 1/|A| \Big)
\end{aligned}$$

Note that in general

$$\begin{aligned}
\sum_G \prod_{n=1}^{N} x(G_n,n) &= \sum_{G_1} \sum_{G_2} \cdots \sum_{G_N} x(G_1,1)x(G_2,2)\cdots x(G_N,N) \\
&= \sum_{G_1} x(G_1,1) \sum_{G_2} x(G_2,2) \cdots \sum_{G_N} x(G_N,N) \\
&= \prod_{n=1}^{N} \sum_{G_n} x(G_n,n)
\end{aligned}$$

So

$$\text{score}(A') = \sum_A |A'\Delta A| \prod_{n=1}^{N} \sum_{G_n} \sum_{S_n} \Big(\prod_{k=1}^{L} w_{k,\text{base}(R_n,k),\text{base}(G_n,S_n+k-1)} \Big) \Big(1/\text{len}(G_n) \Big) \Big(1_A(G_n) 1/|A| \Big)$$

blah

Also note that

$$\sum_A |A'\Delta A| f(A) = \sum_{c=1}^{\infty} c \sum_{A:|A'\Delta A|=c} f(A)$$

blahblah

$$\begin{aligned}
\text{score}(A') &= \sum_A |A'\Delta A| \sum_{(G,S)} \prod_{n=1}^{N} \Big(\prod_{k=1}^{L} w_{k,\text{base}(R_n,k),\text{base}(G_n,S_n+k-1)} \Big) \Big(1/\text{len}(G_n) \Big) \Big(1_A(G_n) 1/|A| \Big) \\
&= \sum_A \Big(\sum_{i=1}^{I} 1_{A'}(i) + 1_A(i) - 2 \cdot 1_{A'\cap A}(i) \Big) \sum_{(G,S)} \prod_{n=1}^{N} \Big(\prod_{k=1}^{L} w_{k,\text{base}(R_n,k),\text{base}(G_n,S_n+k-1)} \Big) \Big(1/\text{len}(G_n) \Big) \Big(1_A(G_n) 1/|A| \Big) \\
&= \Big[\sum_{i=1}^{I} \sum_A (1_{A'}(i) - 2 \cdot 1_{A'\cap A}(i)) \sum_{(G,S)} \prod_{n=1}^{N} \Big(\prod_{k=1}^{L} w_{k,\text{base}(R_n,k),\text{base}(G_n,S_n+k-1)} \Big) \Big(1/\text{len}(G_n) \Big) \Big(1_A(G_n) 1/|A| \Big) \Big] \\
&+ \Big[\sum_{i=1}^{I} \sum_A 1_A(i) \sum_{(G,S)} \prod_{n=1}^{N} \Big(\prod_{k=1}^{L} w_{k,\text{base}(R_n,k),\text{base}(G_n,S_n+k-1)} \Big) \Big(1/\text{len}(G_n) \Big) \Big(1_A(G_n) 1/|A| \Big) \Big]
\end{aligned}$$

The first term can essentially be computed using Bo's code, so we focus on the second term for now.

blahblah

Note

$$\begin{aligned}
\text{score}(A') &= \sum_A d(A',A) \sum_{(G,S)} P(R|G,O,S)P(S|G)P(G|A) \\
&= \sum_A d(A',A) \sum_{(G,S)} \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A) \\
&= \sum_A d(A',A) \sum_{(G,S)} \prod_{n=1}^{N} \sum_{i=1}^{I} 1_{G_n=i} P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A) \\
&= \sum_A d(A',A) \prod_{n=1}^{N} \sum_{(G,S)} \sum_{i=1}^{I} 1_{G_n=i} P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A)
\end{aligned}$$

blahblah

Note

$$\begin{aligned}
\text{score}(A') &= \sum_A d(A',A) \sum_{(G,S)} P(R|G,O,S)P(S|G)P(G|A) \\
&= \sum_A d(A',A) \sum_{(G,S)} \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A) \\
&= \sum_A \sum_{(G,S)} d(A',A) \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A) \\
&= \sum_{(A,G,S)} \sum_{i=1}^{I} 1_{A'\Delta A}(i) \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A)
\end{aligned}$$

blah

What is $P(D|i \in A)$? it is

$$\begin{aligned}
P(D|i \in A) &= \sum_{A:i\in A} \sum_{(G,S)} \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A) \\
&= \sum_{(G,S)} [\prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)][\sum_{A:i\in A} \prod_{n=1}^{N} P(G_n|A)]
\end{aligned}$$

Note

$$\begin{aligned}
\sum_{A:i\in A} \prod_{n=1}^{N} P(G_n|A) &= \sum_{A:i\in A} \prod_{n=1}^{N} 1_A(G_n) 1/|A| \\
&= \sum_{c=1}^{\infty} \sum_{A:i\in A,|A|=c} \prod_{n=1}^{N} 1_A(G_n) 1/c \\
&= \sum_{c=1}^{\infty} c^{-N} \sum_{A:i\in A,|A|=c} \prod_{n=1}^{N} 1_A(G_n) \\
&= \sum_{c=1}^{\infty} c^{-N} \binom{L_0-1}{c-1}
\end{aligned}$$

The last line follows because there are $\binom{L_0}{c}$ possible assemblies with c isoforms in each, so there are $\binom{L_0-1}{c-1}$ possible assemblies that have both isoform i and $c-1$ other isoforms.

So

$$\text{score}(A') \propto \sum_A d(A',A)P(A|D) \qquad = \sum_A d(A',A)\sum_{i=1}^{I} P(i \in A|D) = \sum_A \sum_{i=1}^{I} 1_{i \in A' \Delta A} P(D|A)$$

blahblah

Say $d(A',A) = |A' \Delta A|^N$. Then

$$\begin{aligned}
\text{score}(A') &= \sum_A d(A',A) \sum_{(G,S)} P(R|G,O,S)P(S|G)P(G|A) \\
&= \sum_A \Big(\sum_{i=1}^{I} 1_{A' \Delta A}(i) \Big)^N \Big(\sum_{(G,S)} \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A) \Big) \\
&= \sum_A \Big(\prod_{n=1}^{N} \Big(\sum_{G_n=1}^{I} 1_{A' \Delta A}(G_n) \Big)^N \Big)^{1/N} \Big(\sum_{(G,S)} \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A) \Big) \\
&= \sum_A \Big(\prod_{n=1}^{N} \sum_{G_n=1}^{I} 1_{A' \Delta A}(G_n) \Big) \Big(\sum_{(G,S)} \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A) \Big) \\
&= \sum_A \Big(\sum_G \prod_{n=1}^{N} 1_{A' \Delta A}(G_n) \Big) \Big(\sum_{(G,S)} \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A) \Big) \\
&= \sum_A \sum_G \Big(\prod_{n=1}^{N} 1_{A' \Delta A}(G_n) \Big) \Big(\sum_S \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A) \Big) \\
&= \sum_A \sum_G \sum_S \prod_{n=1}^{N} 1_{A' \Delta A}(G_n) P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A) \\
&= \sum_G \sum_S \sum_A \prod_{n=1}^{N} 1_{A' \Delta A}(G_n) P(R_n|G_n,S_n)P(S_n|G_n)P(G_n|A) \\
&= \sum_G \Big(\sum_A \prod_{n=1}^{N} 1_{A' \Delta A}(G_n) P(G_n|A) \Big) \Big(\sum_S \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n) \Big) \\
&= \sum_G \Big(\sum_A \prod_{n=1}^{N} 1_{A' \Delta A}(G_n) 1_A(G_n) 1/|A| \Big) \Big(\sum_S \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n) \Big) \\
&= \sum_G \Big(\sum_A \prod_{n=1}^{N} 1_{A \setminus A'}(G_n) 1/|A| \Big) \Big(\sum_S \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n) \Big)
\end{aligned}$$

Let $\alpha(G,A') = \sum_A \prod_{n=1}^{N} (1_{A \setminus A'}(G_n)/|A|)$. Note

$$\begin{aligned}
\alpha(G,A') &= \sum_A \prod_{n=1}^{N} (1_{A \setminus A'}(G_n)/|A|) \\
&= \sum_A |A|^{-N} \prod_{n=1}^{N} 1_{A \setminus A'}(G_n) \\
&= \sum_{c=1}^{\infty} c^{-N} \alpha_c(G,A')
\end{aligned}$$

where $\alpha_c(G,A') = \sum_{A:|A|=c} \prod_{n=1}^{N} 1_{A\setminus A'}(G_n)$ is the number of assemblies of cardinality c that contain all of $\{G_n : n = 1,\ldots,N\}$. I think we can probably find a closed form for $\alpha(G,A')$ but I haven't done so yet. Continuing,

$$\begin{aligned}\text{score}(A') &= \sum_G \alpha(G)\Big(\sum_S \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)\Big) \\ &= \sum_G \sum_S \alpha(G) \prod_{n=1}^{N} P(R_n|G_n,S_n)P(S_n|G_n)\end{aligned}$$

Note that each read's set of variables is independent, so we can do:

1. For $n = 1,\ldots,N$, sample a bunch of G_n^t, S_n^t, $t = 1,\ldots,T$, using Gibbs sampling or some other MCMC method, and throw away the S_n^t.
2. Repeatedly sample G by sampling G_n uniformly from $\{G_n^t\}_{t=1}^T$, and sum $\alpha(G)$ for these.

Note that

1. If $|A| = c$, then $P(G_n|A) = 1_A(G_n)1/c$.
2. $A'\Delta A = (A' \cap A^c) \cup (A'^c \cap A)$. Note that

$$|\{A : G_n \in A' \cap A^c\}|$$

What is $\alpha(G_n) := \sum_A \prod_{n=1}^{N}(1_{A\setminus A'}(G_n)/|A|)$? It is:

$$\begin{aligned}\alpha(G_n) &= \sum_A \prod_{n=1}^{N}(1_{A\setminus A'}(G_n)/|A|) \\ &= \sum_A |A|^{-N} \prod_{n=1}^{N} 1_{A\setminus A'}(G_n) \\ &= \sum_{c=1}^{\infty} c^{-N}\alpha_c(G_n)\end{aligned}$$

where $\alpha_c(G_n) = \sum_{A:|A|=c} \prod_{n=1}^{N} 1_{A\setminus A'}(G_n)$ is the number of assemblies of cardinality c that contain all of $\{G_n : n = 1,\ldots,N\}$. Let $|G|$ be the number of unique isoforms in G, i.e., $|G| = |\{G_n : n = 1,\ldots,N\}|$. There are $I_0 - |A'|$.

2 RSEM

Fixed quantities:

1. N - number of fragments, indexed by n.
2. M - number of isoforms (excluding the noise isoform), indexed by i; $i = 0$ is the noise isoform.
3. ℓ_i - length of the ith isoform.
4. $k \in \{1,2\}$ - indexes left or right read.

Parameters:

1. $\boldsymbol{\theta} = (\theta_0,\ldots,\theta_M) \in \Delta_{M+1} \subset \mathbb{R}^{M+1}$ - prior probabilities of a fragment being derived from each transcript; i.e., $P(G_n = i|\boldsymbol{\theta}) = \theta_i$.

Random variables:

1. $G_n \in [0,M]$ - the index of the isoform that fragment n is derived from; $P(G_n = i|\boldsymbol{\theta}) = \theta_i$.

2. $F_n \in \mathbb{N}$ - the length of fragment n. Follows "global fragment length distribution" λ_F, truncated to the transcript length: $P(F_n = x|G_n = i) \propto \lambda_F(x)1_{[1,\ell_i]}(x)$.
3. $O_n \in \{0,1\}$ - the orientation of fragment n; $O_n = 1$ means that fragment n is reverse complemented with respect to its parent isoform. Either $P(O_n = 0|G_n \neq 0) = 1$ (strand-specific protocol) or $= 1/2$ (not strand-specific).
4. $S_n \in \mathbb{N}$ - start position of fragment n within its parent isoform; there are several choices for its distribution:
   (a) Uniform RSPD: $S_n|G_n = i \sim \text{Uniform}([1,\ell_i])$ if there are poly(A) tails, or $\text{Uniform}([1,\ell_i - F_n + 1])$ if not.
5. $L_n^k \in \mathbb{N}$ - the length of the fragment n's k read; follows "global read length distribution" λ_R, truncated to the fragment length: $P(L_n^k = y|F_n = x) \propto \lambda_R(y)1_{[1,x]}(y)$.
6. Q_n^k - the quality score of fragment n's k read.
7. R_n^k - fragment n's k read. Either: $P(R_{n,j}^k = r|Q_{n,j}^k = q, C_{n,j}^k = c) = \varepsilon(r,q,c)$. Or: ...
8. $C_{n,j}^k$ - the true character at position j of fragment n's k read.

3 Error

We define a transcriptome to be $(\mathbf{I},\mathbf{X})$ where

1. $\mathbf{I} = (\mathbf{I}_1,\ldots,\mathbf{I}_M)$ are the isoforms present. $\mathbf{I}_i = (I_{i,1},\ldots,I_{i,\ell_i})$. $I_{i,j}$ is the character at position j of isoform i.
2. $\mathbf{X} = (X_1,\ldots,X_M)$ are the expression levels of each isoform, measured by number of copies present in the transcriptome.

We define an assembly to be $(\mathbf{C},\mathbf{X})$ where ... same as transcriptome, substituting $\mathbf{C}$ for $\mathbf{I}$.

OR:

We define a transcriptome to be

The statistical model used by RSEM can be represented by the directed graphical model shown in Figure 4. Compared to our original statistical model [7], this model has been extended in four ways. First, PE reads are now modeled, using a pair of observed random variables, R1 and R2. For the case of SE reads, R2 is treated as a latent random variable. Second, the length of the fragment from which a read or pair of reads is derived is now modeled and is represented by the latent random variable F. The distribution of F is specified using a global fragment length distribution ÎżF, which is truncated and normalized given that a fragment is derived from a specific transcript of finite length. That is, where âĎŞi is the length of transcript i. The use of a fragment length distribution for RNA-Seq quantification was first introduced by [6] for paired-end data and later described by [4] for single-end data.

thumbnailFigure 4. The directed graphical model used by RSEM. The model consists of N sets of random variables, one per sequenced RNA-Seq fragment. For fragment n, its parent transcript, length, start position, and orientation are represented by the latent variables Gn, Fn, Sn and On respectively. For PE data, the observed variables (shaded circles), are the read lengths (and), quality scores (and), and sequences (and). For SE data, , , and are unobserved. The primary parameters of the model are given by the vector Îÿ, which represents the prior probabilities of a fragment being derived from each transcript.

A third extension allows the lengths of reads to vary (such as for 454 data). The length of a read is represented by the observed random variable L (or L1 and L2 for PE reads). Similar to the fragment length model, the distribution of L is specified using a global read length distribution ÎżR, which is truncated and normalized given a specific fragment length. In symbols, . Lastly, the quality scores for a read are now used to model the probability of that read's sequence. The quality score string for a read is represented by the random variable Q. For the purposes of quantification, we do not specify a distribution for the Q random variables, as they are observed and not dependent on any of the other random variables (i.e., we are only interested in the conditional likelihood of the reads given their quality scores). Rather than rely on the theoretical probabilities of errors implied by the quality scores, we use an empirical error function, Îţ. Given that read position i has quality score qi and is derived from the reference character c, the conditional probability

of the read character ri is P(ri|qi, c) = Îţ(ri, qi, c). If quality scores are not available or reliable, then our position and reference character-dependent error model [7] may be used.