Estimation of transcript length using contig length from assembly using single-end metatranscriptome reads of equal length

Tianyang Li
Department of Automation
Tsinghua University
tmy1018@gmail.com

TABLE I: Notation

Symbol	Meaning
L	The length of the transcript
k	The length of each read
N	The number of reads
$\tilde{L}$	The effective length of the transcript, defined as $L-k+1$
c	The length of the contig
$\tilde{c}$	The effective length of the contig, defined as $c-k$
$\hat{L}(\tilde{c})$	The estimated effective length of the transcript estimated using the effective contig length
x_+	$\max(0, x)$

I. The model

It is assumed here that the reads are independently and uniformly distributed along the transcript.

A. Notation

The notation used is shown in Table I.

B. Estimation of transcript length using contig length

It is assumed here that the contig used to estimate the transcript length is the only contig assembled using the N reads.

X_i denotes the starting position of the i^{th} read ($1 \leq i \leq N$). By assuming that the reads are independently and uniformly distributed along the transcript, X_i ($1 \leq i \leq N$) are i.i.d discrete uniform random variables with the parameter being $\tilde{L}$. However, by using de novo assembly to obtain the contig, X_i ($1 \leq i \leq N$) is unknow, and only $X_i - X_j$ ($1 \leq i, j \leq N$) are known.

Now consider the order statistics of $X_1, X_2, ..., X_N$

$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(N)}$$

It is evident that the joint distribution of $X_i - X_j$ ($1 \leq i, j \leq N$) does not depend on $\tilde{L}$ when $X_{(N)} - X_{(1)}$ is given. Thus, $X_{(N)} - X_{(1)}$ is a sufficient statistic for $\tilde{L}$.

When there is only one contig assembled using the N reads, we have

$$\tilde{c} = X_{(N)} - X_{(1)}$$

From previous works on order statistics [1], the distribution of $\tilde{c}$ given $\tilde{L}$ is

$$P_{\tilde{L}}(\tilde{c} = x) = \frac{\tilde{L} - c}{\tilde{L}^N}((x+1)^N - 2x^N + (x-1)_+^N) \quad (1)$$

for $x = 0, 1, ..., \tilde{L} - 1$.

It will be now shown that $\tilde{c}$ is complete. Indeed, for a function $g(\tilde{c})$ such that

$$E_{\tilde{L}} g(\tilde{c}) = \sum_{x=0}^{\tilde{L}-1} P_{\tilde{L}}(x) g(x) = 0$$

for $\tilde{L} = 1, 2, ...$, it is evident that $g(\tilde{c}) = 0$ for $\tilde{c} = 0, 1, ...$

By solving the recurrence equation

$$E_{\tilde{L}} \hat{L}(\tilde{c}) = \sum_{x=0}^{\tilde{L}-1} P_{\tilde{L}}(x) \hat{L}(x) = \tilde{L} \quad (2)$$

where $\tilde{L} = 1, 2, ...$, an unbiased estimator for $\tilde{L}$ using $\tilde{c}$ can be calculated.

Subsituting (1) into (2), we have

$$\sum_{x=0}^{\tilde{L}-1} (\tilde{L} - x)((x+1)^N - 2x^N + (x-1)_+^N) \hat{L}(x) = \tilde{L}^{N+1} \quad (3)$$

When $\tilde{L}$ is set to $\tilde{c}$ and $\tilde{c} + 1$ in (3) and one equation is subtracted from the other, we have

$$\sum_{x=0}^{\tilde{c}} ((x+1)^N - 2x^N + (x-1)_+^N) \hat{L}(x) = (\tilde{c}+1)^{N+1} - \tilde{c}^{N+1} \quad (4)$$

From (4), it is evident that

$$\hat{L}(\tilde{c}) = \frac{(\tilde{c}+1)^{N+1} - 2\tilde{c}^{N+1} + (\tilde{c}-1)_+^{N+1}}{(\tilde{c}+1)^N - 2\tilde{c}^N + (\tilde{c}-1)_+^N} \quad (5)$$

From the Lehmann-Scheffe theorem, we can know that $\hat{L}(\tilde{c})$ given in (5) is the minimum-variance unbiased estimator.

C. Probability of obtaining a single contig

The probability for obtaining a single contig with effective length $\tilde{c}$ from a transcript with effective length $\tilde{L}$ when there are N reads of lenth k are used to assemble the contig is calculated.

It is assumed here that overlaps between reads can be exactly detected, and there is no repeat sequences in the transcript.

Let $p_1(n)$ denote the probability that n positions out of the $\tilde{L}$ positions in the transcript are the starting positions of the reads. Using the inclusion-exclusion principle, we have

$$\begin{aligned} p_1(n) &= \binom{\tilde{L}}{n}(\frac{n}{\tilde{L}})^N(1-\binom{n}{1}(\frac{n-1}{n})^N+\binom{n}{2}(\frac{n-2}{n})^N-...) \\ &= \binom{\tilde{L}}{n}\frac{\sum_{i=0}^{n}(-1)^i\binom{n}{i}(n-i)^N}{\tilde{L}^N} \\ &= \binom{\tilde{L}}{n}\frac{n!S(N,n)}{\tilde{L}^N} \end{aligned} \tag{6}$$

where $S(n,k)$ denotes Stirling numbers of the second kind.

Let $p_2(n,\tilde{c})$ denote the probability that a single contig with effective length $\tilde{c}$ can be obtained when there are n distinct starting positions for the reads. Let $Y_1 \le Y_2 \le ... \le Y_n$ denote the n distinct starting positions of the reads, and $Y_0 = 1$ and $Y_{n+1} = \tilde{L}$. We can see that $p_2(n,\tilde{c})$ is equal to the probability that $Y_n - Y_1 = \tilde{c}$ and $Y_{i+1} - Y_i < k$ for $i = 1,2,...,n-1$. Now, let $z_i = Y_{i+1} - Y_i$ for $i = 0,1,...,n$. Then, in order to calculate $p_2(n,\tilde{c})$, the number of integer solutions satisfying

$$z_0 + z_1 + ... + z_n = \tilde{L} - 1 \tag{7}$$

where $z_0 + z_n = \tilde{L} - 1 - \tilde{c}; z_0, z_n \ge 0$ and $0 < z_i < k; i = 1,2,...,n-1$.

The number of integer solutions when we only require $z_0, z_n \ge 0$ and $z_i > 0; i = 1,2,...,n-1$ is

$$\binom{\tilde{L}}{n} \tag{8}$$

The number of integer solutions for $z_0 + z_n = \tilde{L} - 1 - \tilde{c}; z_0, z_n \ge 0$ is

$$\binom{\tilde{L}-\tilde{c}}{1} = \tilde{L} - \tilde{c} \tag{9}$$

Now assume $n \ge 2$.

The number of integer solutions for $z_1 + z_2 + ... + z_{n-1} = \tilde{c}$ when we only require $z_i > 0$ for $i = 1,2,...,n-1$ is

$$\binom{\tilde{c}-1}{n-2} \tag{10}$$

The number of integer solutions for $z_1 + z_2 + ... + z_{n-1} = \tilde{c}$ when $z_i > 0$ for $i = 1,2,...,n-1$ and there are m known terms among $z_1, z_2, ..., z_{n-1}$ larger than or equal k is

$$\binom{\tilde{c}-m(k-1)-1}{n-2} \tag{11}$$

where it is assumed that $\tilde{c} \ge m(k-1)$.

Using the inclusion-exclusion principle, the number integer solutions for $z_1 + z_2 + ... + z_{n-1} = \tilde{c}$ when $0 < z_i < k$ is

$$\sum_{m=0}^{\lfloor\frac{\tilde{c}-1}{k-1}\rfloor}(-1)^m\binom{n-1}{m}\binom{\tilde{c}-1-m(k-1)}{n-2} \tag{12}$$

Using (8), (9), and (12), we can get

$$p_2(n,\tilde{c}) = \tag{13}$$

$$\begin{cases} 1 & \text{if } n = 1, \tilde{c} = k-1 \\ 0 & \text{if } n = 1, \tilde{c} \ne k-1 \\ \frac{\tilde{L}-\tilde{c}}{\binom{\tilde{L}}{n}}\sum_{m=0}^{\lfloor\frac{\tilde{c}-1}{k-1}\rfloor}(-1)^m\binom{n-1}{m}\binom{\tilde{c}-1-m(k-1)}{n-2} & \text{if } n \ne 1 \end{cases} \tag{14}$$

Thus, the probability for obtaining a single contig with effective length $\tilde{c}$ from a transcript with effective length $\tilde{L}$, which is denoted as $Q_{\tilde{L}}(\tilde{c})$, is

$$Q_{\tilde{L}}(\tilde{c}) = \sum_{n=1}^{\tilde{L}} p_1(n)p_2(n,\tilde{c}) \tag{15}$$

References

[1] H. David and H. Nagaraja, *Order statistics*, ser. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. John Wiley, 2003.