Estimating the distance between two sequences using paired-end reads

Shaun D Jackman, İnanç Birol
sjackman@bcgsc.ca
Canada's Michael Smith Genome Sciences Centre
Vancouver, British Columbia, V5Z 4E6, Canada

December 14, 2011

Abstract

Paired-end reads may be used to estimate the distance between two sequences. Comparing a statistic, such as the mean, of the sample population of fragment sizes to the global population of fragment sizes is a trivial but flawed estimator. The maximum likelihood estimator yields more accurate estimates.

Background

The goal is to estimate the size of the gap between two sequences.

Results

To estimate the distance between two sequences, we start by mapping paired-end reads to the two sequences, which are then ordered and oriented to agree with the orientation of the reads. We establish a coordinate system where $-l_1$ and -1 are the first and last base of the first sequence of length l_1, and 0 and $l_2 - 1$ are the first and last base of the second sequence of length l_2. An observed fragment size, x_i, is calculated for each pair by calculating the difference of the mapped position of the first sequenced base of each of the

two reads. This observed fragment size differs from the actual fragment size by the size of the gap between the two sequences, θ_0.

The final input is the distribution of fragment sizes of the library, which is derived empirically by mapping the reads to a reference sequence or sequences assembled *de novo* and determining the inferred fragment size distribution.

Estimator using the mean

At first glance, this task appears to be rather simple, and a simple solution presents itself readily. A reasonable estimate, $\hat{\theta}_{\text{mean}}$, of the size of the gap is the difference between the mean of the population, μ, and the mean of the sample, $\bar{x}$.

$$\hat{\theta}_{\text{mean}} = \mu - \bar{x}$$

Maximum likelihood estimator

The estimate of the distance between the two sequences can be improved by using the probability distribution in its entirety rather than a summary statistic. Let the probability of observing a fragment of size x selected at random from the population be $f_X(x)$, and the probability of observing a fragment of size x that spans a gap of size θ be $f_\theta(x)$. With a sample of n observed fragment sizes, $x_1, \ldots, x_n$, the likelihood that the two sequences are separated by a distance of θ bases is $\mathcal{L}(\theta \mid x_1, \ldots, x_n)$.

The most likely estimate of the size of the gap between the two sequences is the value $\hat{\theta}_{\text{MLE}}$ that maximizes the likelihood function, or conveniently, the log likeliehood function, since the log function is a monotonic transformation.

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \arg\max_\theta \mathcal{L}(\theta \mid x_1, \ldots, x_n) = \arg\max_\theta \prod_{i=1}^{n} f_\theta(x_i) \\ &= \arg\max_\theta \log \mathcal{L}(\theta \mid x_1, \ldots, x_n) = \arg\max_\theta \sum_{i=1}^{n} \log f_\theta(x_i)\end{aligned}$$

Distribution of fragment sizes that span the gap

The distribution of observed fragment sizes that span the gap is equal to the population distribution, $P(X = x)$, shifted by the size of the gap, θ. Since

we can only observe fragments that actually span the gap, we use Bayes' thereom to determine the conditional probability of observing a fragment of size x given that it spans the gap of size θ.

$$\begin{aligned} f_\theta(x) &= P(X = x + \theta \mid \text{fragment spans gap}) \\ &= \frac{P(\text{fragment spans gap} \mid X = x + \theta) P(X = x + \theta)}{P(\text{fragment spans gap})} \\ &\propto P(\text{fragment spans gap} \mid X = x + \theta) P(X = x + \theta) \end{aligned}$$

Assume the reads are sampled uniformly from the genome between the coordinates a and b, where $b - a$ is the size of the genome.

$$P(U = u) = \begin{cases} \frac{1}{b-a} & a \leq u < b \\ 0 & \text{otherwise} \end{cases}$$

The probability that a fragment of size X spans the gap is the probability that the fragment's left coordinate, U, falls to the left of the gap, and its right coordinate, $U + X$, falls to the right of gap.

$$\begin{aligned} \text{Let } w_\theta(x + \theta) &= P(\text{fragment spans gap} \mid X = x + \theta) \\ &= P(-l_1 \leq U < 0 \wedge \theta \leq U + X < l_2 + \theta \mid X = x + \theta) \\ &= P(-l_1 \leq U < 0 \wedge \theta \leq U + x + \theta < l_2 + \theta) \\ &= P(-l_1 \leq U < 0 \wedge 0 \leq U + x < l_2) \\ &= w(x) \end{aligned}$$

This shows that $w_\theta(x + \theta)$, the probability that a fragment of size $x + \theta$ spans the gap, is independent of the size of the gap, θ, and depends only on the observed size of the fragment, x.

$$
\begin{aligned}
w(x) &= P(-l_1 \leq U < 0 \wedge 0 \leq U + x < l_2) \\
&= P(-l_1 \leq U < 0 \wedge -x \leq U < l_2 - x) \\
&= P(\max(-l_1, -x) \leq U < \min(0, l_2 - x)) \\
&= \sum_{i=\max(-l_1,-x)}^{\min(0,l_2-x)-1} P(U = i) \\
&= \sum_{i=\max(-l_1,-x)}^{\min(0,l_2-x)-1} \frac{1}{b-a} \\
&= \max(0, \min(0, l_2 - x) - \max(-l_1, -x)) \frac{1}{b-a} \\
&\propto \max(0, \min(0, l_2 - x) - \max(-l_1, -x)) \\
&= \max(0, \min(0, l_2 - x) + \min(l_1, x)) \\
&= \max(0, \min(x, l_1, l_2, l_1 + l_2 - x))
\end{aligned}
$$

Without loss of generality, assume $l_1 \leq l_2$.

$$
w(x) \propto \begin{cases} x & 0 \leq x < l_1 \\ l_1 & l_1 \leq x < l_2 \\ l_1 + l_2 - x & l_2 \leq x < l_1 + l_2 \\ 0 & \text{otherwise} \end{cases}
$$

$$
f_\theta(x) \propto f_X(x + \theta) w(x)
$$

$$
f_\theta(x) = \frac{f_X(x + \theta) w(x)}{\sum_{j=1}^{\infty} f_X(j + \theta) w(j)}
$$

Solving the maximum likelihood estimator

We now substitute the distribution of observed fragment sizes, $f_\theta(x)$, into the formula for the maximum likelihood estimator.

$$\begin{aligned}\mathcal{L}(\theta \mid x_1, \ldots, x_n) &= \prod_{i=1}^{n} f_\theta(x_i) \\ &= \prod_{i=1}^{n} \frac{f_X(x_i+\theta)w(i)}{\sum_{j=1}^{\infty} f_X(j+\theta)w(j)} \\ &= \frac{\prod_{i=1}^{n} f_X(x_i+\theta)w(i)}{\left(\sum_{j=1}^{\infty} f_X(j+\theta)w(j)\right)^n}\end{aligned}$$

$$\log \mathcal{L}(\theta \mid x_1, \ldots, x_n) = \sum_{i=1}^{n} \log f_X(x_i+\theta) + \sum_{i=1}^{n} \log w(i) - n \log \sum_{j=1}^{\infty} f_X(j+\theta)w(j)$$

$$\begin{aligned}\hat{\theta}_{\mathrm{MLE}} &= \arg\max_\theta \log \mathcal{L}(\theta \mid x_1, \ldots, x_n) \\ &= \arg\max_\theta \left[\sum_{i=1}^{n} \log f_X(x_i+\theta) + \sum_{i=1}^{n} \log w(i) - n \log \sum_{j=1}^{\infty} f_X(j+\theta)w(j) \right] \\ &= \arg\max_\theta \left[\sum_{i=1}^{n} \log f_X(x_i+\theta) - n \log \sum_{j=1}^{\infty} f_X(j+\theta)w(j) \right]\end{aligned}$$

Finding the value of θ that maximizes the likelihood function is an optimization problem. When the range of possible values of θ is small, that is when the fragment size of the sequencing library is small, it is reasonable to calculate exhaustively every value of $\mathcal{L}(\theta)$ to find the maximum.

Conclusion

This distance estimation algorithm is implemented by the *ABySS* assembly software in the utility *DistanceEst*, which requires as its input the distribution of fragment sizes of the sequencing library and a *SAM*-formatted file of paired-end reads that map to different sequences.

Acknowledgements

Jared Simpson implemented a maximum likelihood estimator for estimating distances between sequences in the first release of the software *ABySS*.

References

ABySS: A parallel assembler for short read sequence data. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. Genome Research, 2009-June.