Genome Assembly: Lander-Waterman Statistics

Sorin Istrail

Center for Computational Molecular Biology
Department of Computer Science
Brown University, Providence
sorin@cs.brown.edu

May 2, 2011

Lander-Waterman Statistics

- The Lander-Waterman formulas[1] provide a statistical framework for estimating sequencing parameters needed to achieve particular levels of quality for large-scale sequencing projects.
- Suppose we have a long piece of DNA we want to sequence.
- We model the sequencing of fragments by breaking the DNA uniformly at random and then sequencing the DNA at the start of that break.
- **Assumption 1: DNA may be broken uniformly at random and all DNA bases are able to be sampled.**

[1] E. S. Lander and M. S. Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics, 2(3):231-239, April 1988.

Lander-Waterman Statistics Definitions

We define several parameters common to all sequencing projects for our calculations

- L = read length
- N = number fragments each of length L
- G = original sequence's length (genome length)
- $a = \frac{NL}{G}$ or the coverage

Lander-Waterman Statistics: The three fundamental questions

Q1. What is the mean portion of the genome covered by contigs?

Q2. What is the mean number of contigs?

Q3. What is the mean contig size?

Lander-Waterman Statistics Assumptions

We must make simplifications via assumptions because the model becomes too complex to analyze analytically. We define now our assumptions.

Assumption 1: DNA may be broken uniformly at random and all DNA region are able to be sampled.

Assumption 2: We assume $G >> L$ and therefore we can safely ignore end effects in our calculations (e.g. reads starting in [G-L+1,G]).

Assumption 3: No repeats on the target genome.

Assumption 4: No errors

Q1. What is the mean portion of the genome covered by contigs?

- The mean portion of the genome covered by one or more fragments is the probability that a point chosen at random is covered by at least one fragment.
- We can model with a Poisson model.
- Let Y = number of reads whose leftmost end point is located within the interval of length L. This follows a Poisson distribution with mean a.

Lander-Waterman Statistics

- The fragments are taken at random from the original full-length sequence so, if end effects are ignored, the left-hand ends of the fragments are independently distributed with a common *uniform distribution* over $[0, G]$.
- This implies that any such left-hand end falls in an internal $(x, x+h)$ with probability $\frac{h}{G}$ and that the number of fragments whose left-hand end falls in this interval has a binomial distribution with mean $\frac{NL}{G}$.
- If N is large and h is small, this distribution is approximately Poisson with mean $a = \frac{NL}{G}$.

Lander-Waterman Statistics

- The number Y of fragments whose left-end is located within an interval of length L to the left of a randomly chosen point, therefore, has a Poisson distribution with mean a, so that the possibility that at least one fragment arises in this interval is $1 - Prob(Y = 0) = 1 - e^{-a}$.

$$P(Y = 0) = e^{-a}\frac{a^0}{0!} = e^{-a}$$

$$P(Y = k) = e^{-a}\frac{a^k}{k!}$$

- $P(Y)$ together with other properties of homogeneous Poisson processes is enough to provide the answer to these basic questions.

Lander-Waterman Statistics

$$A1 : P(Y \geq 1) = 1 - e^{-a}$$

Genome covered by the reads means that every base on the genome is covered by at least one read.

Lander-Waterman Statistics

Example. Consider a Shotgun sequencing project. We have $a = 8$, $G = 10^4$, $L = 500$, $N = 1600$. Randomly pick the start position of the read (left most position). We have 1600 randomly picked points where reads start. Using $\frac{NL}{G}$ for the mean for the Poisson we can find the answer to Q1.

	2	4	6	8	10	12
A1: mean portion of the genome covered	0.86	0.98	0.997	0.9996	0.99995	0.999994

Q2. What is the mean number of contigs?

- Each contig has a *unique rightmost fragment* so that the formula np for the mean of the binomial distribution given below shows that the mean number of contigs is the number N of fragments multiplied by the probability that a fragment is the rightmost member of a contig.
- Mark all the rightmost fragments.
- Their number equals the number of contigs.
- The probability that no other fragment has its left-hand end point on the fragment in question is e^{-a}.
- $A2 = Ne^{-a} = Ne^{\frac{-NL}{G}}$ is the mean number of contigs.

Lander-Waterman Statistics

Example. Recall that $A2 = Ne^{-a}$. Let $G = 100000$ and $L = 500$. Then

	0.5	0.75	1	1.5	2	3	4	5	6	7
A2: mean portion of the genome covered	60.7	70.8	73.6	66.9	54.1	29.9	14.7	6.7	3.0	1.3

Q3: What is the average contig size?

$$A3 = \frac{e^a - 1}{\lambda} = L\frac{e^a - 1}{a}$$

where $\lambda = \frac{N}{G}$**,** $a = \frac{NL}{G}$**,** $\lambda L = a$**.**

Lander-Waterman Statistics

- By the Poisson approximation we create the distribution for reads on a genome.
- The distance between the starting point of one read to the starting point of the next overlapping read follows a geometric distribution which we approximate by an exponential distribution with $\lambda = \frac{N}{G}$. See Figure 1.

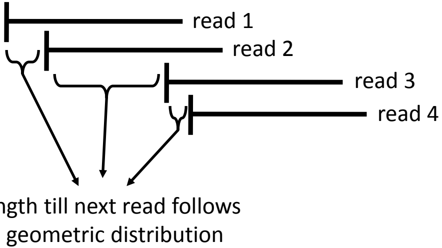


Figure: We approximate the geometric dist by an exponential dist.

Lander-Waterman Statistics

- The second fragment will overlap the first one if this distance is less than L in length of the first fragment. This occurs with prob $1 - e^{-a}$
- We then have a series of overlaps until the first failure. Treat overlaps as successes and non-overlaps as failures. The number of successive overlapping fragments before the first non-overlap has a geometric distribution whose mean is $e^a - 1$.

Lander-Waterman Statistics

- If n frags form a contig, the total fragment lengths of the contig is the sum of random distances + L.
- The mean of these random distances is $\frac{1}{\lambda} - \frac{L}{e^{a}-1}$. See Figure 2.

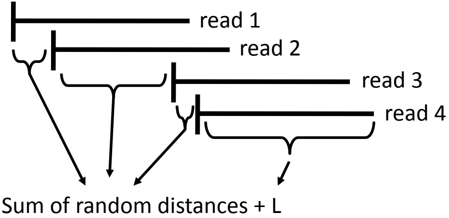


Figure: The contig length is given by the sum of random distances + L.

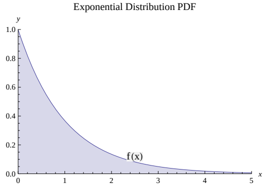
- The mean of a sum of a random number of random variables show that the mean total of these distances is $\frac{e^{a}-1}{\lambda} - L$
- Adding L from the last fragment we obtain the mean contig size: $A3 = L\frac{e^{a}-1}{a}$

Lander-Waterman Statistics

- Reads are place uniformly at random along the genome that is the left end point is picked and you read out L bases.
- The mean contig is found by considering the left-hand ends of a succession of fragments starting with the initial left-hand fragment of a given contig.
- The Poisson distribution can approximate the Binomial very well when the number of trials is very large and the probability p is very small.
- Using the Poisson approximation for the read distribution, the distance from the left-hand end of this first fragment and the left-hand end of the next fragment has a *geometric* distribution.
- This distribution is closely approximated by the exponential distribution with parameter $\lambda = \frac{N}{G}$. For example: $N = 1600$, $G = 10^4$, then $p = \frac{N}{G} = \frac{1600}{10^4}$. We make this approximation here.

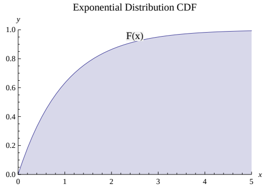
The Exponential Distribution

A continuous random variable having this distribution has range $[0, +\infty]$ and density function $f_x(x) = \lambda e^{-\lambda x}, x \geq 0$.



Lander-Waterman Statistics

A single parameter $\lambda > 0$ characterizes the distribution. The mean is $\frac{1}{\lambda}$ and the variance is $\frac{1}{\lambda^2}$. It has cumulative distribution function $F_x(x) = \lambda - e^{=\lambda x}, x \geq 0$



The Relation of the Exponential to the Geometric Distribution

Taking $\lambda = 1$ and plugging the CDF into the mean of the geometric distribution: $\frac{p}{1-p} = \frac{1-e^{-a}}{1-(1-e^{-a})} = \frac{1-e^{-a}}{1-(1-e^{-a})} = \frac{1-e^{-a}}{e^{-a}} = \frac{1}{e^{-a}} - 1 = e^{a} - 1$

Lander-Waterman Statistics

- Assume the fragments are of constant size L.
- The second fragment will overlap the first one if this distance is less than the length L of the first fragment.
- This occurs with probability

$$\int_0^L \lambda e^{-\lambda x} dx = 1 - e^{-a}$$

Lander-Waterman Statistics

- A further overlap occurs if the next fragment to the right of the second fragment overlaps the second fragment.
- The contig is built this way until such an overlap fails to occur.
- We think of an overlap as "success" and a non-overlap as a "failure."
- The number of successive overlapping fragments, before the first non-overlap has a geometric distribution whose mean is $\frac{p}{1-p} = e^{a} - 1$ where $p = 1 - e^{-a}$.
- If n fragments from a contig, total length of the contig is the length L of the final fragment together with the sum of the $n-1$ random distances between the left-hand end of any given fragment and the left-hand of the next fragment in the contig to the right.

Lander-Waterman Statistics

The conditional distribution of an exponential random variable X given that $0 \leq X \leq L$ is

$$\frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda L}}, \qquad 0 \leq x \leq L$$

The mean of this distribution is

$$\int_0^L x \cdot \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda L}} dx = \frac{1}{\lambda} - \frac{L}{e^{\lambda L - 1}}$$

Lander-Waterman Statistics

The mean of these random distances is $\frac{1}{\lambda} - \frac{L}{e^{a}-1}$ where $\lambda = \frac{N}{G}$ and $a = \frac{NL}{G}$

The Exponential Distribution vs. Random n for sum of n Random variables The X_i are independently and identically distributed random variables with mean μ. Let $S_n = X_1 + ... + X_n$ with mean of $S_n = n\mu$. However this is assuming n is fixed. We need n to also be a random variable which we will denote by $\mathcal{N}$. Assume $\mathcal{N}$ is independent from $X_1, ..., X_n$. The sum is now denoted S. The mean of S is $E(S) = E(\mathcal{N})E(X)$

Lander-Waterman Statistics

- So for us the mean of the sum of a random number of random variables is according to above $E(S)$ formula.
- The mean of the random distances is $\frac{1}{\lambda} - \frac{L}{e^a - 1}$ where $\lambda = \frac{N}{G}$
- The mean of the sum of a random variables of random mean is $\frac{e^a - 1}{\lambda} - L$.
- Upon adding length L for the last contig we get

$$A3 = \frac{e^a - 1}{\lambda} = L\frac{e^a - 1}{a}$$

where $\lambda = \frac{N}{G}$, $a = \frac{NL}{G}$, $\lambda L = a$.

Lander-Waterman Statistics

Example with $L = 500$.

a	2	4	6	8	10
mean contig size	1,600	6,700	33,500	186,000	1,100,000

Table: An example of the mean contig size calculation for a fixed cover and read length.