1 Outline

- Lander-Waterman Statistics: Statistical theory that guides the planning of a genome assembly project.

2 Lander-Waterman Statistics

2.1 The analysis of one DNA sequence

Definitions

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

We have a long piece of DNA we want to sequence. The model for breaking the DNA into pieces: break the DNA uniformly at random and then sequencing the DNA at the start of that break. This assumes that all pieces of DNA are able to be sampled.

Because \( G >> L \) we can ignore end effects. 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 \((x, x+L)\) with probability \( \frac{L}{G} \) 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 \( L \) is small, this distribution is approximately Poisson with mean \( \frac{NL}{G} \). 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 \( 1 - Prob(Y = 0) = 1 - e^{-a} \).

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

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

This calculation together with other properties of homogeneous Poisson processes is enough to provide the answer to these basic 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?
We must make simplifications via assumptions because the model becomes too complex to analyze analytically. For instance, protein folding is one of the grand challenges in computer science. Modeling the problem with the exact physics is intractable. Computational modeling of protein folding defines some objective function and a set of assumptions. Hopefully the solution to the computational model is close to reality. Most computational models try to minimize free energy in some sense. In the HP model, you simplify proteins as a sequence of hydrophobic and hydrophilic amino acids and try to maximize hydrophobic contacts. When we talk about models we first must make the assumptions explicit and then draw the conclusions.

Assumption 1: No repeats on the target genome. Assumption 2: \( L << G \) Assumption 3: no errors

2.2 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 when \( N \) is large and \( h \) is small. The binomial distribution doesn’t model very well when \( N \) is very large and \( p \) is very small. Let \( Y \) = number of reads whose leftmost end point is located within the interval of length \( L \). This follows a Poisson distribution with mean \( a \).

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

\[
P(Y = 0) = e^{-a}
\]

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

The probability of Q1 = the probability that at least one fragment has its left end in the interval \( L \) immediately to the left of this point. Genome covered by the reads means that every base on the genome is covered by at least one read.

2.3 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. This is the probability that no other fragment has its left-hand end point on the fragment in question. This is equal to \( e^{-a} \). So \( Ne^{-a} = Ne^{a-NL/G} \) is the mean number of contigs.

2.4 Outline of Q3

1. 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.

2. 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 \( 1 - e^{-a} \)

3. 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 \( e^a - 1 \).
4. 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.

5. The mean of a sum of a random number of random variables show that the mean total of these distances is $\frac{e^a - 1}{\lambda} - L$

6. Adding $L$ from the last fragment we obtain the mean contig size: $A_3 = L e^a - 1$

Example 1. For a particular BAC we have $G = 125$kb, $L = 480$ (Sangar sequencing). We want $a = 10$. How many reads do we need? The number of reads $N = A_1 G = \frac{10 \times 125000}{480} = 2604$ reads. When we sequence we either sequence the sense or antisense strand. The first thing people do when sequencing is reverse compliment all strands and expect assembly algorithm to assemble both strands. So we conclude that we need $2 \times 2604$ reads.

Example 2. Shotgun sequencing. 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. We now postulate a Poisson distribution for the read placement. $p = \frac{N}{G} = 0.016$. Thus we have “success” with $p = 0.016$ and “failure” with $1 - p = 0.98$. Using $N L / G$ for the mean for the Poisson we can find the answer to Q1.

<table>
<thead>
<tr>
<th>A1: mean portion of the genome covered</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.86</td>
<td>0.98</td>
<td>0.997</td>
<td>0.9996</td>
<td>0.99995</td>
<td>0.999994</td>
</tr>
</tbody>
</table>

Example 3. Recall that $A_2 = Ne^{-a}$. Let $G = 100000$ and $L = 500$. Then

<table>
<thead>
<tr>
<th>A2: mean portion of the genome covered</th>
<th>0.5</th>
<th>0.75</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60.7</td>
<td>70.8</td>
<td>73.6</td>
<td>66.9</td>
<td>54.1</td>
<td>29.9</td>
<td>14.7</td>
<td>6.7</td>
<td>3.0</td>
<td>1.3</td>
</tr>
</tbody>
</table>
A3 will be analyzed thoroughly in a future class.