Poisson Distribution in Genome Assembly
Poisson Example: Genome Assembly

- **Goal**: figure out the sequence of DNA nucleotides (ACTG) along the entire genome
- **Problem**: Sequencers generate random **short reads**

TABLE 9.1 Next-generation sequencing technologies compared to Sanger sequencing. Adapted from the companies’ websites, [http://en.wikipedia.org/wiki/DNA_sequencer](http://en.wikipedia.org/wiki/DNA_sequencer), and literature cited for each technology.

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read length (bp)</th>
<th>Reads per run</th>
<th>Time per run</th>
<th>Cost per megabase (US$)</th>
<th>Error (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche 454</td>
<td>700</td>
<td>1 million</td>
<td>1 day</td>
<td>10</td>
<td>0.1</td>
<td>99.90</td>
</tr>
<tr>
<td>Illumina</td>
<td>50–250</td>
<td>&lt;3 billion</td>
<td>1–10 days</td>
<td>–0.10</td>
<td>2</td>
<td>98</td>
</tr>
<tr>
<td>SOLiD</td>
<td>50</td>
<td>~1.4 billion</td>
<td>7–14 days</td>
<td>0.13</td>
<td>0.1</td>
<td>99.90</td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>200</td>
<td>&lt;5 million</td>
<td>2 hours</td>
<td>1</td>
<td>2</td>
<td>98</td>
</tr>
<tr>
<td>Pacific Biosciences</td>
<td>2900</td>
<td>&lt;75,000</td>
<td>&lt;2 hours</td>
<td>2</td>
<td>1</td>
<td>99</td>
</tr>
<tr>
<td>Sanger</td>
<td>400–900</td>
<td>N/A</td>
<td>&lt;3 hours</td>
<td>2400</td>
<td>0.1</td>
<td>99.90</td>
</tr>
</tbody>
</table>

- **Solution**: assemble genome from short reads using computers. **Whole Genome Shotgun Assembly**.
Short Reads assemble into Contigs
How many short reads do we need?

**Input**
- Low coverage:
  - A few pieces to assemble

**Output**
- many contigs, many gaps

**Input**
- High coverage:
  - many pieces to assemble

**Output**
- a few contigs, a few gaps
Where is the Poisson?

- **G** - genome length (in bp)
- **L** - short read average length
- **N** – number of short read sequenced
- **λ** – sequencing coverage redundancy = LN/G
- **x** - number of short reads covering a given site on the genome

\[
P(x) = \frac{\lambda^x e^{-\lambda}}{x!}
\]

Poisson as a limit of Binomial: For a given site on the genome for each short read Prob(site covered): p = L/G is very small. Number of attempts (short reads): N is very large. Their product (sequencing redundancy): λ = NL/G is O(1).
What fraction of genome is covered?

- **Coverage**: $\lambda = \frac{N L}{G}$,
  
  $X$ – r.v. equal to the number of times a given site is covered

  Poisson: $P(X=x) = \frac{\lambda^x \exp(-\lambda)}{x!}$

  $P(X=0) = \exp(-\lambda)$, $P(X>0) = 1 - \exp(-\lambda)$

- **Total length covered**: $G \times [1 - \exp(-\lambda)]$

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean proportion of genome covered</td>
<td>.864665</td>
<td>.981684</td>
<td>.997521</td>
<td>.999665</td>
<td>.999955</td>
<td>.999994</td>
</tr>
</tbody>
</table>

Table 5.1. The mean proportion of the genome covered for different values of $\lambda$
How many contigs?

- Probability that a given short read is the right end of a contig = no left ends of other reads fall within it.
- Left ends of each of $N-1 \approx N$ other short reads has Prob: $p = (L-1)/N \approx L/N$ to fall within given read. Probability that none do is $= \exp(-\lambda)$:

**Number of contigs:** $N_{\text{contigs}} = Ne^{-\lambda}$:

<table>
<thead>
<tr>
<th>$\lambda$</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>Mean number of contigs</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>

Table 5.2. The mean number of contigs for different levels of coverage, with $G = 100,000$ and $L = 500$. 
Average length of a contig?

- **Length of a genome covered:**
  \[ G_{\text{covered}} = G \cdot P(X>0) = G \cdot (1 - \exp(-\lambda)) \]

- **Number of contigs** \( N_{\text{contigs}} = N \cdot e^{-\lambda} \)

- **Average length of a contig** =
  \[ <L> = \frac{\sum_i L_i}{N_{\text{contigs}}} = \frac{G_{\text{covered}}}{N_{\text{contigs}}} = \frac{G \cdot (1 - \exp(-\lambda))}{N \cdot e^{-\lambda}} = \frac{L \cdot (1 - \exp(-\lambda))}{\lambda \cdot e^{-\lambda}} \]

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean contig size</td>
<td>1,600</td>
<td>6,700</td>
<td>33,500</td>
<td>186,000</td>
<td>1,100,000</td>
</tr>
</tbody>
</table>

Table 5.3. The mean contig size for different values of \( \lambda \) for the case \( L = 500 \).
Estimate

- Human genome is $3 \times 10^9$ bp long
- Chromosome 1 is about $G = 0.25 \times 10^9$ bp
- Illumina generates short reads $L = 100$ bp long
- What number of reads $N$ are needed to completely assemble the 1st chromosome?
- The formula to use is: $1 = N_{contigs} = Ne^{-\lambda} = Ne^{-NL/G}$
- Answer: $N = 4.4 \times 10^7$ short (100bp) reads
  Test: $4.4e7 * \exp(-4.4e7 * 100 / 0.25e9) = 0.9997$
- What coverage redundancy $\lambda$ will it be?
  Answer: $\lambda = NL/G = 17.6$ coverage redundancy
How much would it cost to assemble human genome now?

- Human Genome Project: $2.7 billion in 1991 dollars.
- Now a de novo full assembly of the whole human genome would now cost $3 \times 10^9 \times 17.6/10^6 \times 0.1$/MB = $5300
- 2nd genome (and after) would be even cheaper as we would already have a reference genome to which we can map short reads. (Puzzle: picture on the box)
- But, this is a naïve estimate. In reality there are complications. See next slides:
What spoils these estimates?

There were 100s of matches while one expects \(<< 1\) match:

\[
2 \cdot 3 \times 10^9 \cdot 4^{-18} = 0.08 \ll 1
\]

DNA repeats make assembly difficult
Repeats are like sky puzzle pieces
How many repeats are in eukaryotic genomes?

Data for mouse genome obtained in 1961 (sic!) using DNA denaturation and renaturation curves.

**FIGURE 8.6** The complexity of genomic DNA can be estimated by denaturing then renaturing DNA. This figure (redrawn from Britten and Kohne, 1968) depicts the relative quantity of mouse genomic DNA (y axis) versus the logarithm of the frequency with which the DNA is repeated. The data are derived from a $C_0 t_{1/2}$ curve, which describes the percent of genomic DNA that reassociates at particular times and DNA concentrations. A large $C_0 t_{1/2}$ value implies a slower reassociation reaction. Three classes are apparent. The fast component accounts for 10% of mouse genomic DNA (arrow A), and represents highly repetitive satellite DNA. An intermediate component accounts for about 20% of mouse genomic DNA and contains repeats having from 1000 to 100,000 copies. The slowly reassociating component, comprising 70% of the mouse genome, corresponds to unique, single-copy DNA. Britten and Kohne (1968) obtained similar profiles from other eukaryotes, although distinct differences were evident between species. Used with permission.

© 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.
Companion Website: www.wiley.com/go/pevsnerbioinformatics
Distinct components in complex genomes

- **Highly repeated DNA**
  - $R$ (repetition frequency) $\geq 100,000$
  - Almost no information, low complexity

- **Moderately repeated DNA**
  - $10 < R < 10,000$
  - Little information, moderate complexity

- **“Single copy” DNA**
  - $R = 1$ or $2$
  - Much information, high complexity

Slide by Ross Hardison, Penn State U.
Almost all transposable elements in mammals fall into one of four classes.

- **LINEs**
  - Autonomous
  - ORF1
  - ORF2 (pol)
  - AAA
  - Length: 6–8 kb
  - Copy number: 850,000
  - Fraction of genome: 21%

- **SINEs**
  - Non-autonomous
  - A, B
  - AAA
  - Length: 100–300 bp
  - Copy number: 1,500,000
  - Fraction of genome: 13%

- **Retrovirus-like elements**
  - Autonomous
  - gag
  - pol
  - env
  - Length: 6–11 kb
  - Copy number: 450,000
  - Fraction of genome: 8%
  - Non-autonomous
  - (gag)
  - Length: 1.5–3 kb
  - Copy number: 450,000
  - Fraction of genome: 8%

- **DNA transposon fossils**
  - Autonomous
  - Transposase
  - Length: 2–3 kb
  - Copy number: 300,000
  - Fraction of genome: 3%
  - Non-autonomous
  - Length: 80–3,000 bp

*Slide by Ross Hardison, Penn State U.*
De Bruin Graph

Nodes: k-mers
Edges: connects k-mers within **sequenced** k+1-mer on a short read
Simplified De Bruin Graph
join all k-mers on a path without branching

Assembly = find Eulerian walk visiting each edge once

Slide by Sorin Istrail, Brown U.
Edge-disjoint loops are a problem: multiple solutions

A graph can have multiple Eulerian walks, only one of which corresponds to the original superstring.

Right: graph for ZABCDABEFABY, \( k=2 \)

Alternative Eulerian walks:

- \( ZA \to AB \to BE \to EF \to FA \to AB \to BC \to CD \to DA \to AB \to BY \)
- \( ZA \to AB \to BC \to CD \to DA \to AB \to BE \to EF \to FA \to AB \to BY \)

These correspond to two edge-disjoint directed cycles joined by node \( AB \)

\( AB \) is a repeat: \( ZABCDABEFABY \)

Adapted from a slide by Ben Langmead, Johns Hopkins U.
How to assemble genome with repeats?

- **Answer:** longer reads
- **But:**
  - cheap sequencing = short reads

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche 454</td>
<td>700</td>
</tr>
<tr>
<td>Illumina</td>
<td>50–250</td>
</tr>
<tr>
<td>SOLiD</td>
<td>50</td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>200</td>
</tr>
<tr>
<td>Pacific Biosciences</td>
<td>2900</td>
</tr>
<tr>
<td>Sanger</td>
<td>400–900</td>
</tr>
</tbody>
</table>

Example of disjoint loops
A gallery of useful discrete probability distributions
Geometric Distribution

- A series of Bernoulli trials with probability of success \(= p\). continued until the first success. \(X\) is the number of trials.
- Compare to: Binomial distribution has:
  - Fixed number of trials = \(n\).
  - Random number of successes = \(x\).
- Geometric distribution has reversed roles:
  - Random number of trials, \(x\)
  - Fixed number of successes, in this case 1.
  - Success always comes in the end: so no combinatorial factor \(C^n_x\)
  - \(P(X=x) = p(1-p)^{x-1}\) where:
    \(x-1 = 0, 1, 2, \ldots\), the number of failures until the 1\(^{st}\) success.
- NOTE OF CAUTION: Matlab, Mathematica, and many other sources use \(x\) to denote the number of failures until the first success. We stick with Montgomery-Runger notation
Geometric Mean & Variance

• If \( X \) is a geometric random variable (according to Montgomery-Bulmer) with parameter \( p \),

\[
\mu = E(X) = \frac{1}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(1-p)}{p^2}
\]  

(3-10)

• For small \( p \) the standard deviation \( \sim = \text{mean} \)

• Very different from Poisson, where it is variance = mean and standard deviation = mean\(^{1/2} \)
Geometric distribution in biology

- Each of our cells has mitochondria with 16.5kb of mtDNA inherited only from our mother
- Human mtDNA has 37 genes encoding 13 proteins, 22+2 tRNA & rRNA
- Mitochondria appeared 1.5-2 billion years ago as a symbiosis between an alpha-proteobacterium (1000s of genes) and an archaeaon (of UIUC’s Carl R. Woese fame)
- Since then most mitochondrial genes were transferred to the nucleus
- Plants also have plastids with genomes related to cyanobacteria
Time to the last common (maternal) ancestor follows geometric distribution

- **Constant population** of N women
- **Random number** of (female) offsprings. Average is 1 (but can be 0 or 2)
- **Randomly pick two women.** Question: how many generations T since their last maternal ancestor?
- T is a random variable. What is its PMF: P(T=t)?
  - Answer: P(T=t) follows a geometric distribution
- Do these two women have the same mother? Yes: “success” in finding their last common ancestor (p=1/N). P(T=1)=1/N.
- No? “failure” (1-p=1-1/N). Go to their mothers and repeat the same question.
  - P(T=t)=(1-1/N)^t-1(1/N) \approx (1/N) \exp(-T/N)
  - T can be inferred from the density of differences on mtDNA =2\mu T