Poisson Processes and Poisson Distribution
CS9601 – Biological Sequence Analysis

Lucian Ilie

Department of Computer Science, University of Western Ontario
Poisson Process

- Assume a sequence of random events during some time interval, where randomness is defined by 1 and 2 below:
  1. Occurrences in disjoint intervals are independent
  2. There is a constant \( \lambda \) such that
     - Probability of one event in \((t, t + h)\) (\(h\) small) is \( \lambda h + o(h) \) (indep. of \(t\))
     - Probability of \( \geq 2 \) occurrences in \((t, t + h)\) is \( o(h) \) (v.small)

- \( N \) = number of events up to any time \( t \)
- \( P_j(t) \) = probability that \( N = j \) at time \( t \)

- We have:
  \[
  P_0(t + h) = P_0(t)(1 - \lambda h) + o(h)
  \]
  \[
  P_j(t + h) = P_{j-1}(t)(\lambda h) + P_j(t)(1 - \lambda h) + o(h)
  \]
Poisson Process

this gives:
\[
\frac{P_0(t+h)-P_0(t)}{h} = -\frac{P_0(t)(\lambda h)+o(h)}{h}
\]
\[
\frac{P_j(t+h)-P_j(t)}{h} = \frac{P_{j-1}(t)\lambda h-P_j(t)\lambda h+o(h)}{h}, j \geq 1
\]

letting \( h \to 0 \) we get
\[
\frac{d}{dt} P_0(t) = -\lambda P_0(t)
\]
\[
\frac{d}{dt} P_j(t) = \lambda P_{j-1}(t) - \lambda P_j(t), j \geq 1
\]

with initial conditions: \( P_0(0) = 1, P_j(0) = 0, j \geq 1 \)

this has the solution: \( P_0(t) = e^{-\lambda t}, P_j(t) = \frac{e^{-\lambda t}(\lambda t)^j}{j!}, j \geq 1 \)

we obtain that \( N \) has, at time \( t \), Poisson distribution with parameter \( \lambda t \)

\( \lambda \) is the average rate with which the events occur
Poisson and Binomial Distributions

- consider the time interval \((0, t)\) and divide into \(n = t/h\) pieces of length \(h\) each

\[
P_j(t) = \binom{n}{j} p^j (1 - p)^{n-j} = \frac{n!}{j!(n-j)!} \frac{(\lambda t)^j}{n^j} \left(1 - \frac{\lambda t}{n}\right)^{n-j}
\]

- the number of events follows a binomial distribution with probability \(p = \lambda h = \frac{\lambda t}{n}\)

\[
P_j(t) = \frac{(\lambda t)^j}{j!} \frac{n!}{(n-j)! n^j} \left(1 - \frac{\lambda t}{n}\right)^{n-j} \xrightarrow{n \to \infty} e^{-\lambda t} (\lambda t)^j \frac{1}{j!}
\]
Poisson Distributions

Poisson Distribution

\[ f(X(x)) \]

\( \lambda = 0.3 \)
\( \lambda = 1 \)
\( \lambda = 2 \)
\( \lambda = 4 \)
\( \lambda = 10 \)
Examples of Poisson Processes

- chromosomal crossovers
- spontaneous protein degradation
- shotgun sequencing
Chromosomal crossover

- **Mendel** studied the inheritance of **characters** in pea plants:
  - characters: seed shape (round\( (R) \), wrinkled\( (r) \)), seed color (yellow\( (Y) \), green\( (y) \)), height (tall\( (T) \), short\( (t) \)), etc.
  - kept track of characters in many generations
- each character determined by a **gene**
- **alleles** – versions of the same gene (dominant and recessive)
- dominant gene wins (genotype \( Ttrr \) gives tall and wrinkled phenotype)
- **1st law**: half of the gametes have one copy of a gene and half the other copy; individuals inherit one allele from each parent:
  - \( Tt \) gives half with \( T \) and half with \( t \)
  - \( rr \) gives all with \( r \)
  - \( Tr + tr \) gives \( Ttrr \)
Chromosomal crossover

- 2\textsuperscript{nd} law: different genes are inherited independently
- Example:
  - female genotype: $TtRr$ (eggs: $TR$, $Tr$, $tR$, $tr$ 25% each)
  - male genotype: $TtRR$ (sperm: $TR$, $tR$ 50% each)
  - Punnett square

<table>
<thead>
<tr>
<th></th>
<th>$TR(1/2)$</th>
<th>$tR(1/2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TR(1/4)$</td>
<td>$TTRR(1/8)$</td>
<td>$TtRR(1/8)$</td>
</tr>
<tr>
<td>$Tr(1/4)$</td>
<td>$TTRr(1/8)$</td>
<td>$TtRr(1/8)$</td>
</tr>
<tr>
<td>$tR(1/4)$</td>
<td>$TtRR(1/8)$</td>
<td>$ttRR(1/8)$</td>
</tr>
<tr>
<td>$tr(1/4)$</td>
<td>$TtRr(1/8)$</td>
<td>$ttRr(1/8)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>genotype</th>
<th>prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TTRR$</td>
<td>$1/8$</td>
</tr>
<tr>
<td>$TtRR$</td>
<td>$2/8 = 1/4$</td>
</tr>
<tr>
<td>$TTRr$</td>
<td>$1/8$</td>
</tr>
<tr>
<td>$TtRr$</td>
<td>$2/8 = 1/4$</td>
</tr>
<tr>
<td>$ttRR$</td>
<td>$1/8$</td>
</tr>
<tr>
<td>$ttRr$</td>
<td>$1/8$</td>
</tr>
</tbody>
</table>

- exceptions from these laws: genes come in chromosomes (independent inheritance applies to genes on different chromosomes), some genes have more than two alleles, etc.
Morgan discovered that genes do not combine in predicted proportions because of chromosomes.

- Genes that are closer have a higher chance to stay together.
- E.g., the outcomes $GT$, $Gt$, $gT$, $gt$ appear in proportions $40\%$, $10\%$, $10\%$, $40\%$, resp.
- Recombination rate is $20\%$.
Chromosomal crossover

\[ R_{AB} = \text{recombination between } A \text{ and } B \]

that is, odd number of crossover between A and B

\[
\begin{align*}
\text{Prob}(R_{AC}) &= \text{Prob}(R_{AB}) \text{Prob}(R_{BC}) + \text{Prob}(R_{AB}) \text{Prob}(R_{BC}) \\
&= \text{Prob}(R_{AB})(1 - \text{Prob}(R_{BC})) + (1 - \text{Prob}(R_{AB})) \text{Prob}(R_{BC})
\end{align*}
\]

- Haldane’s formula: \( \text{Prob}(R_{AB}) = \frac{1}{2}(1 - e^{-2 \text{dist}(A,B)}) \)
- distance measured in Morgans (\(M\))
- if distance = 0.01\(M\), the recombination rate is
  \[
  \frac{1}{2}(1 - e^{-0.02}) = 0.009900663 \approx 1\%
  \]
- if recombination rate is 1\% = 0.01, then distance is
  \[
  -\frac{1}{2} \ln(1 - 2 \cdot 0.01) = 0.01010135 \approx 0.01M
  \]
Chromosomal crossover

- on the right: Mouse chromosome 5 linkage map (Mouse Genome Database)
- distance between Xmv34 and LMBR1 is $15.80 - 14.00 = 1.80 \text{cM} = 0.018M$
- recombination rate is $\frac{1}{2}(1 - e^{-2 \cdot 0.018}) = 0.01767985 = 1.767985\%$
Probability distribution of crossover

- events = crossovers
- rate = 1 crossover per M (\( \lambda \))
- occurrence of crossovers in disjoint intervals is independent
- probability of two crossovers very close (\( h \)) to each other is very very small
- the theory of Poisson processes says that the distribution of the number of crossovers between two given genes \( A \) and \( B \) at distance \( d \) has Poisson distribution with parameter \( \lambda d \)
Determining the Poisson parameter from data

- assume the number of crossovers between two genes from 100 gametes was, experimentally:

<table>
<thead>
<tr>
<th># of crossovers</th>
<th># gametes</th>
<th>total # of crossovers</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>64</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>≥4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>total</td>
<td>100</td>
<td>44</td>
</tr>
</tbody>
</table>

- average number of crossover per gamete = $44/100 = 0.44$
- Poisson parameter $\lambda d = 0.44$; so they are $0.44M$ apart
- Poisson distribution: $\frac{e^{-0.44}(0.44)^k}{k!}$
### Determining the Poisson parameter from data

The Poisson parameter can be determined from data using the formula:

\[
\text{Poisson prob.} = \frac{e^{-0.44} (0.44)^k}{k!}
\]

<table>
<thead>
<tr>
<th>k = # of crossovers</th>
<th>observed prob.</th>
<th>Poisson prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.64</td>
<td>0.6440364</td>
</tr>
<tr>
<td>1</td>
<td>0.29</td>
<td>0.283376</td>
</tr>
<tr>
<td>2</td>
<td>0.06</td>
<td>0.06234273</td>
</tr>
<tr>
<td>3</td>
<td>0.01</td>
<td>0.0091436</td>
</tr>
<tr>
<td>≥4</td>
<td>0</td>
<td>0.00110127</td>
</tr>
</tbody>
</table>
Pure-Birth Process

- initial value of \( N \) is \( k \)
- the probability that \( N \) increases from \( j \) at \( t \) to \( j + 1 \) in the interval \((t, t + h)\) is \( \lambda_j h + o(h) \)
- as for Poisson we obtain:

\[
\frac{d}{dt} P_k(t) = -\lambda_k P_k(t)
\]
\[
\frac{d}{dt} P_j(t) = \lambda_{j-1} P_{j-1}(t) - \lambda_j P_j(t), j = k + 1, k + 2, \ldots
\]
Yule Process

- pure-birth process with $\lambda_j = j\lambda$
- the growth of a population is proportional with current size
- the probability distribution is

$$P_j(t) = \binom{j - 1}{j - k}(e^{-\lambda t})^k (1 - e^{-\lambda t})^{j-k}, j \geq k$$
Polymerase Chain Reaction

- used to amplify short pieces of DNA
- start with a primer of length \( k \)
- that is extended by sequential addition of single base pairs
- pure-birth process with \( \lambda_j = m - j \)
- we have \( \frac{d}{dt} P_j(t) = (m - j + 1)P_{j-1}(t) - (m - j)P_j(t), j \geq k + 1 \)
- the probability distribution is

\[
P_j(t) = \binom{m - k}{j - k}(1 - e^{-t})^{j-k}(e^{-t})^{m-j}, j \geq k
\]

\[
P_j(t) = \binom{n}{i}(1 - e^{-t})^i(e^{-t})^{n-i}, i \geq 0
\]
the lengths of the time intervals between events in a Poisson process
or
until the first event

(memoryless property says there is no difference)

Examples:

the distance between crossovers (or to the first crossover)
the time from one protein degradation to the next (or the time until
the first degradation)

\( X = \) the time until the first event

\[
\text{Prob}(X > x) = P_0(x) = e^{-\lambda x} \quad \text{(that is, no event until time } x) \\
F_X(x) = \text{Prob}(X \leq x) = 1 - e^{-\lambda x} \\
f_X(x) = \frac{d}{dx} F_X(x) = \lambda e^{-\lambda x}
\]
Shotgun sequencing

- DNA sequence – needed in order to do any analysis
- no technology can read a whole genome sequence
  - human genome – 3 billion nucleotides
- **Sanger** technology – can read 500 - 1000 consecutive nucleotides from one end of a DNA sequence
  - Fredrick Sanger – Nobel Prize, Chemistry, 1980 (and 1958)
- locations of fragments in the genome are not known
- **shotgun sequencing**
  - use overlaps between fragments to assemble them
Shotgun sequencing – the method

- start with many copies of a genome
- genome length $\approx 3$ billion
- fragment them and sequence reads
- read length $\approx 500$
- find overlapping reads

ACGTA

ACGTA

merge them into contigs

AACCATG...
Sequencing the human genome

- **BAC sequencing**
  - Bacterial Artificial Chromosomes
  - public effort
  - start with many copies of the genome
  - break them into pieces of length 150,000 – 350,000
  - shotgun sequence each BAC
  - assemble BACs using overlaps

- **Whole genome shotgun**
  - Celera Genomics (private company)
  - shotgun sequence the whole genome
  - assembly is much more difficult
  - requires higher coverage
Shotgun sequencing – basic parameters

- $G =$ genome length (in nucleotides)
- $L =$ read length (assume 500 nucleotides)
- $N =$ number of reads sequenced
- $NL =$ number of nucleotides in all sequenced reads
- $a = \frac{NL}{G} =$ coverage (average number of times each nucleotide in the whole genome is sequenced)
- because $G \gg L$, end effects are ignored below
Shotgun sequencing – modeling

- reads taken at random from the sequence
- left-hand ends of reads are iid uniformly distributed over $[0..G]$ (end effect ignored)
- prob. that a fixed read starts at a given position is $1/G$
- prob. to have a left-hand end at any given position is $N/G$
- consider an interval $I = (x, x + h)$; the number $Y$ of left-hand ends in $I$ has binomial distribution with $p = N/G$ (mean $= Nh/G$)
- this is approximately Poisson with parameter $Nh/G$
- if $I$ has length $L$, we have Poisson with $a$: $f_Y(y) = \frac{e^{-a}a^y}{y!}$
- probability of at least one read having its left-hand end in $I$ is $1 - \text{Prob}(Y = 0) = 1 - e^{-a}$
Mean proportion of the genome covered by contigs

- A nucleotide at position $x$ is in a gap if no read starts within the interval $[x - L + 1..x]$
- The probability for this to happen is $e^{-a}$
- Mean amount of gaps is $Ge^{-a}$
- Mean amount of genome covered is $G(1 - e^{-a})$
- Mean proportion of genome covered is $1 - e^{-a}$

To have 99% genome covered we need $a = -\ln(0.01) \approx 4.6$

- 1% of human genome is 30 million nucleotides
- For $a \geq 8$ the mean proportion covered does not increase

<table>
<thead>
<tr>
<th>$a$</th>
<th>Covered genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.8646647</td>
</tr>
<tr>
<td>4</td>
<td>0.9816844</td>
</tr>
<tr>
<td>6</td>
<td>0.9975212</td>
</tr>
<tr>
<td>8</td>
<td>0.9996645</td>
</tr>
<tr>
<td>10</td>
<td>0.9999546</td>
</tr>
<tr>
<td>12</td>
<td>0.9999939</td>
</tr>
</tbody>
</table>
Mean number of contigs

- each contig has a rightmost read
- a read is rightmost if no read starts within it; probability $= e^{-a}$
- if the rightmost contigs are "successes," we have a binomial distribution with $N$ and $e^{-a}$
- the mean number of contigs is $Ne^{-a} = Ne^{-NL/G}$
  - highest for $a = 1$
  - few reads means few (small) contigs and many reads means few (large) contigs

<table>
<thead>
<tr>
<th>$a$</th>
<th>mean number of contigs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1819592</td>
</tr>
<tr>
<td>0.75</td>
<td>2125649</td>
</tr>
<tr>
<td>1</td>
<td>2207277</td>
</tr>
<tr>
<td>1.5</td>
<td>2008171</td>
</tr>
<tr>
<td>2</td>
<td>1624023</td>
</tr>
<tr>
<td>3</td>
<td>896167</td>
</tr>
<tr>
<td>4</td>
<td>439575</td>
</tr>
<tr>
<td>5</td>
<td>202138</td>
</tr>
<tr>
<td>6</td>
<td>89235</td>
</tr>
<tr>
<td>7</td>
<td>38299</td>
</tr>
</tbody>
</table>
Conditional exponential distribution

- for a small interval \((x, x + h)\) included in \((a, b)\) we have

\[
\text{Prob}(x < X < x + h | a < X < b) = \frac{\text{Prob}(x < X < x + h)}{\text{Prob}(a < X < b)}
\]

- \(h \to 0\) gives \(f_{X|a<X<b}(x) = \frac{f_X(x)}{\int_a^b f_X(u)du}\)

- the conditional distribution of an exponential variable \(X\) given that \(0 < X < L\) is

\[
\frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda L}}, 0 \leq x < L
\]

- and has the mean

\[
\int_0^L x \cdot \frac{\lambda e^{-\lambda x}}{1 - e^{-\lambda L}} = \frac{1}{\lambda} - \frac{L}{e^{\lambda L} - 1}
\]
Sum of randomly many random variables

- \( S = S_n = X_1 + X_2 + \cdots + X_n, \) \( X_i \) indep.
- if pdf of \( X_i \) is \( q_i(t) \), then pgf of \( S \) is \( \prod_{i=1}^{n} q_i(t) \) (induction)
- if \( X_i \) are iid, then pgf of \( S \) is \((q(t))^n\)
- assume Prob\((N = n) = P_n\), then

\[
\text{Prob}(S = s) = \sum_n P_n \text{Prob}(S = s \mid N = n) \\
= \text{coef. of } t^s \text{ in } \sum_n P_n(q(t))^n \\
= \text{coef. of } t^s \text{ in } p(q(t))
\]

- therefore the pgf of \( S \) is \( p(q(t)) \)
- using \( \mu_Y = \left( \frac{d}{dt} p_Y(t) \right)_{t=1} \), we obtain

\[
E(S) = E(N)E(X)
\]
Mean contig size

- left-hand ends – Poisson($N/G$)
- the distance between left-hand ends of reads has exponential distribution
- $f_X(x) = \frac{N}{G} e^{-\frac{N}{G} x}$
- a contig is a succession of overlapping reads – we need exponential distribution conditioned by $X < L$
- if a contig has $n$ reads, then its total length is $L$ plus the sum of the $n-1$ random distances between left-hand ends of reads
- the mean of these distances is $\frac{1}{\lambda} - \frac{L}{e^a - 1}$
- the mean of the sum of a random number of random variables is $E(S) = E(N)E(X)$, that is $(e^a - 1)\left(\frac{1}{\lambda} - \frac{L}{e^a - 1}\right) = \frac{e^a-1}{\lambda} - L$
- the mean contig size is $\frac{e^a-1}{\lambda} = L \frac{e^a-1}{a}$

<table>
<thead>
<tr>
<th>$a$</th>
<th>mean contig size</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1,600</td>
</tr>
<tr>
<td>4</td>
<td>6,700</td>
</tr>
<tr>
<td>6</td>
<td>33,500</td>
</tr>
<tr>
<td>8</td>
<td>186,000</td>
</tr>
<tr>
<td>10</td>
<td>1,100,000</td>
</tr>
</tbody>
</table>
Human whole genome shotgun sequencing

\[ G = 3,000,000,000, \quad L = 500 \]

<table>
<thead>
<tr>
<th>coverage</th>
<th># of reads (millions)</th>
<th>% genome covered</th>
<th>mean no. of contigs</th>
<th>mean contig size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>3</td>
<td>0.393469340</td>
<td>1819592.0</td>
<td>648.7</td>
</tr>
<tr>
<td>1.0</td>
<td>6</td>
<td>0.632120559</td>
<td>2207276.6</td>
<td>859.1</td>
</tr>
<tr>
<td>1.5</td>
<td>9</td>
<td>0.776869840</td>
<td>2008171.4</td>
<td>1160.6</td>
</tr>
<tr>
<td>2.0</td>
<td>12</td>
<td>0.864664717</td>
<td>1624023.4</td>
<td>1597.3</td>
</tr>
<tr>
<td>3.0</td>
<td>18</td>
<td>0.950212932</td>
<td>896167.2</td>
<td>3180.9</td>
</tr>
<tr>
<td>4.0</td>
<td>24</td>
<td>0.981684361</td>
<td>439575.3</td>
<td>6699.8</td>
</tr>
<tr>
<td>5.0</td>
<td>30</td>
<td>0.993262053</td>
<td>202138.4</td>
<td>14741.3</td>
</tr>
<tr>
<td>6.0</td>
<td>36</td>
<td>0.997521248</td>
<td>89235.1</td>
<td>33535.7</td>
</tr>
<tr>
<td>7.0</td>
<td>42</td>
<td>0.999088118</td>
<td>38299.0</td>
<td>78259.5</td>
</tr>
<tr>
<td>8.0</td>
<td>48</td>
<td>0.999664537</td>
<td>16102.2</td>
<td>186247.4</td>
</tr>
<tr>
<td>9.0</td>
<td>54</td>
<td>0.999876590</td>
<td>6664.1</td>
<td>450115.8</td>
</tr>
<tr>
<td>10.0</td>
<td>60</td>
<td>0.999954600</td>
<td>2724.0</td>
<td>1101273.3</td>
</tr>
<tr>
<td>11.0</td>
<td>66</td>
<td>0.999983298</td>
<td>1102.3</td>
<td>2721506.4</td>
</tr>
<tr>
<td>12.0</td>
<td>72</td>
<td>0.999993856</td>
<td>442.4</td>
<td>6781408.0</td>
</tr>
<tr>
<td>13.0</td>
<td>78</td>
<td>0.999997740</td>
<td>176.3</td>
<td>17015861.2</td>
</tr>
<tr>
<td>14.0</td>
<td>84</td>
<td>0.999999168</td>
<td>69.8</td>
<td>42950117.3</td>
</tr>
<tr>
<td>15.0</td>
<td>90</td>
<td>0.999999694</td>
<td>27.5</td>
<td>108967212.4</td>
</tr>
<tr>
<td>16.0</td>
<td>96</td>
<td>0.999999887</td>
<td>10.8</td>
<td>277690922.5</td>
</tr>
<tr>
<td>17.0</td>
<td>102</td>
<td>0.999999959</td>
<td>4.2</td>
<td>710439757.5</td>
</tr>
<tr>
<td>18.0</td>
<td>108</td>
<td>0.999999985</td>
<td>1.6</td>
<td>1823888003.8</td>
</tr>
<tr>
<td>19.0</td>
<td>114</td>
<td>0.999999994</td>
<td>0.6</td>
<td>4696902630.6</td>
</tr>
<tr>
<td>20.0</td>
<td>120</td>
<td>0.999999998</td>
<td>0.2</td>
<td>12129129860.2</td>
</tr>
</tbody>
</table>
## Human BAC sequencing

### $G = 300,000$, $L = 500$

<table>
<thead>
<tr>
<th>coverage</th>
<th># of reads</th>
<th>% genome covered</th>
<th>mean no. of contigs</th>
<th>mean contig size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>300</td>
<td>0.393469340</td>
<td>182.0</td>
<td>648.7</td>
</tr>
<tr>
<td>1.0</td>
<td>600</td>
<td>0.632120559</td>
<td>220.7</td>
<td>859.1</td>
</tr>
<tr>
<td>1.5</td>
<td>900</td>
<td>0.776869840</td>
<td>200.8</td>
<td>1160.6</td>
</tr>
<tr>
<td>2.0</td>
<td>1200</td>
<td>0.864664717</td>
<td>162.4</td>
<td>1597.3</td>
</tr>
<tr>
<td>2.5</td>
<td>1500</td>
<td>0.917915001</td>
<td>123.1</td>
<td>2236.5</td>
</tr>
<tr>
<td>3.0</td>
<td>1800</td>
<td>0.950212932</td>
<td>89.6</td>
<td>3180.9</td>
</tr>
<tr>
<td>3.5</td>
<td>2100</td>
<td>0.969802617</td>
<td>63.4</td>
<td>4587.9</td>
</tr>
<tr>
<td>4.0</td>
<td>2400</td>
<td>0.981684361</td>
<td>44.0</td>
<td>6699.8</td>
</tr>
<tr>
<td>4.5</td>
<td>2700</td>
<td>0.988891003</td>
<td>30.0</td>
<td>9890.8</td>
</tr>
<tr>
<td>5.0</td>
<td>3000</td>
<td>0.993262053</td>
<td>20.2</td>
<td>14741.3</td>
</tr>
<tr>
<td>5.5</td>
<td>3300</td>
<td>0.995913229</td>
<td>13.5</td>
<td>22153.8</td>
</tr>
<tr>
<td>6.0</td>
<td>3600</td>
<td>0.997521248</td>
<td>8.9</td>
<td>33535.7</td>
</tr>
<tr>
<td>6.5</td>
<td>3900</td>
<td>0.998496561</td>
<td>5.9</td>
<td>51087.8</td>
</tr>
<tr>
<td>7.0</td>
<td>4200</td>
<td>0.999088118</td>
<td>3.8</td>
<td>78259.5</td>
</tr>
<tr>
<td>7.5</td>
<td>4500</td>
<td>0.999446916</td>
<td>2.5</td>
<td>120469.5</td>
</tr>
<tr>
<td>8.0</td>
<td>4800</td>
<td>0.999664537</td>
<td>1.6</td>
<td>186247.4</td>
</tr>
<tr>
<td>8.5</td>
<td>5100</td>
<td>0.999796532</td>
<td>1.0</td>
<td>289045.2</td>
</tr>
<tr>
<td>9.0</td>
<td>5400</td>
<td>0.999876590</td>
<td>0.7</td>
<td>450115.8</td>
</tr>
<tr>
<td>9.5</td>
<td>5700</td>
<td>0.999925148</td>
<td>0.4</td>
<td>703090.9</td>
</tr>
<tr>
<td>10.0</td>
<td>6000</td>
<td>0.999954600</td>
<td>0.3</td>
<td>1101273.3</td>
</tr>
</tbody>
</table>
Variable read length

- Assume read lengths is $L$ (between 400 and 600) distributed with density $f_L(\ell)$.

- If $P$ is a point in genome, the probability that $P$ is covered by a read having its left-hand end in the interval $(P - x, P - x + h)$ is the probability of having a read starting in that interval, $\frac{N}{G} h$, times the probability that the length of that read exceeds $x$, $\int_x^{600} f_L(\ell) d\ell$.

- The mean number of reads covering $P$ is
  \[
  \frac{N}{G} \int_0^{600} (1 - F_L(x)) dx = \frac{N}{G} E(L)
  \]

- The mean proportion of the genome covered becomes $1 - e^{-\frac{N}{G} E(L)}$. 
Next Generation Sequencing

- different technology
- much cheaper than Sanger
- much shorter reads

Comparison

<table>
<thead>
<tr>
<th>Technology</th>
<th>454 FLX/Roche</th>
<th>Solexa/Illumina</th>
<th>Solid/ABI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read Length</td>
<td>200-300</td>
<td>25-35</td>
<td>35</td>
</tr>
<tr>
<td>Sequence Yield</td>
<td>100Mb</td>
<td>1Gb</td>
<td>3Gb total</td>
</tr>
<tr>
<td>Accuracy</td>
<td>¿6 homopolymer probl.</td>
<td>near 100%</td>
<td>near 100%</td>
</tr>
<tr>
<td>Time/run</td>
<td>7.5 hrs</td>
<td>3-5 days</td>
<td>8 days</td>
</tr>
<tr>
<td>Run/week</td>
<td>5</td>
<td>1 or 2</td>
<td>1</td>
</tr>
<tr>
<td>Cost per run</td>
<td>$8,200</td>
<td>$3,976</td>
<td>$3,650</td>
</tr>
<tr>
<td>Cost per Mb</td>
<td>$49.80/Mb</td>
<td>$3.98/Mb</td>
<td>$2.43Mb</td>
</tr>
</tbody>
</table>

1 University of California at Riverside: http://genomics.ucr.edu/about/reports/SequencerComparison1207.Table.pdf
References

- Wikipedia
- G. Tesler, Statistical Methods in Bioinformatics, UCSD course.