Poisson (4)

N fragments of length L $\quad G$ = genome length

coverage $= \frac{NL}{G} = c$

Hypothèse de départ : simplification

- left-hand ends of fragment are independently distributed with a uniform distribution in $[0\ G] \Rightarrow [1\ G]$ plutôt

Any such left-hand end falls in an interval $[x, x+h]$ with probability $\frac{h}{G}$

The number of left-hand ends that fall in this interval have a binomial distribution with probability of success $\frac{h}{G}$

The mean number of fragment left-hand ends that fall in $[x, x+h]$ is $\frac{Nh}{G}$

- we assume this distribution is approximatively Poisson with $\lambda = \frac{Nh}{G}$
- The number Y of fragments whose left-hand end is located within an interval L to the left of a randomly chosen point has a Poisson distribution with mean $\lambda = \frac{NL}{G} = c$

The probability that at least 1 fragment arises in this interval is

$$\mathbb{P}(Y \geq 1) = 1 - \mathbb{P}(Y=0) = 1 - \frac{\lambda^0}{0!}e^{-\lambda} = 1 - e^{-c}$$

(1) Mean proportion of the genome covered by contigs

This is the probability that a point chosen at random is covered by at least 1 fragment.
This is the probability that at least 1 fragment starts in the interval of length L to the left of this point. From above this is $\mathbb{P}(Y \geq 1) = 1 - e^{-c}$

(2) Mean number of contigs

- Each contig has a unique right-most fragment.
- Mean number of contigs is the number of fragments N multiplied by the probability that a fragment is the right-most member of a contig.
- The latter probability is the probability that no other fragment starts in the fragment in question. This is $\mathbb{P}(Y=0) = e^{-c}$
- Thus mean number of contigs $= Ne^{-c}$

(3) Mean contig size

- one considers the left-hand end of a succession of fragments starting with the initial left-hand fragment on a given contig
- Under Poisson approximation the distance between the left-hand ends of these fragment has a geometric distribution

See Annex 3

- Geometric distribution is closely approximated by the exponential distribution with parameter $\lambda = \frac{N}{G}$.

See Annex 2

- The 2^{nd} fragment will overlap with the 1^{st} one if the distance between their left-hand ends is $\leq L$. This occurs with probability $\int_0^L \lambda e^{-\lambda x} dx = 1 - e^{-\lambda L} = 1 - e^{-c}$
- A further overlap occurs if the next fragment to the right of the 2nd fragment overlaps that second fragment.

Poisson (2)

(3) Mean contig size (cont.)

- We define an overlap as a failure and a non-overlap as a success
- probability of a success from above $p = e^{-c}$
- the number k of failures before a success has a geometric distribution with probability mass function $(1-p)^k p$. The mean number of failures is given by $\frac{1-p}{p} = \frac{1-e^{-c}}{e^{-c}} = e^{c}-1$
- The mean number of failures $\equiv$ the mean number of overlapping fragments per contig.
- If n fragments form a contig the total length of the contig is the length L of the final fragment with the sum of $n-1$ random distances between the left-hand end of any given fragment an the left-hand end of the next overlapping fragment to the right

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

$$\frac{f_X(x)}{\int_0^L f_X(u)\,du} = \frac{\lambda e^{-\lambda x}}{1-e^{-\lambda L}} \qquad 0 \le x < L \quad (2.51)$$

the mean of this distribution is

$$\mu = \int_0^L x \frac{\lambda e^{-\lambda x}}{1-e^{-\lambda L}}\,dx = \frac{1}{\lambda} - \frac{L}{e^{\lambda L}-1} \qquad (2.52)$$

The mean of a sum of random variables $X_1, X_2 \ldots X_n$ when n is itself a random variable (N) is

$$E(S) = \mathbb{E}(N)\,\mathbb{E}(X) \quad (2.81)$$

Eq 2.52 shows that the mean of this random distances (between the left-hand end of any fragment and the left-hand end of the next overlapping fragment to the right is

$$\mathbb{E}(X) = \frac{1}{\lambda} - \frac{L}{e^{c}-1}$$

The mean number of fragments in a contig is $\mathbb{E}(N) = e^{c}-1$ [see above]

From Eq 2.81 (the mean of a sum of a random number of random variables) we get that the mean total of this distance is

$$(e^{c}-1)\left[\frac{1}{\lambda} - \frac{L}{e^{c}-1}\right] = \frac{e^{c}-1}{\lambda} - L$$

Adding the length L of the final fragment the mean contig size is $\frac{e^{c}-1}{\lambda} = \boxed{L\frac{e^{c}-1}{c}}$