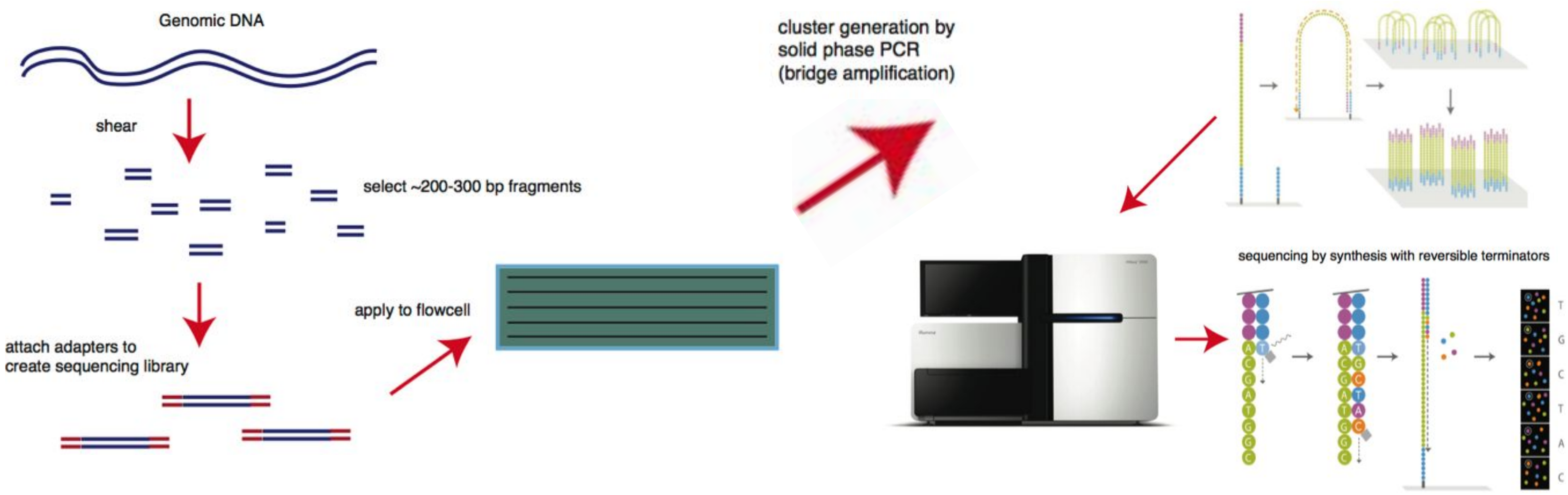
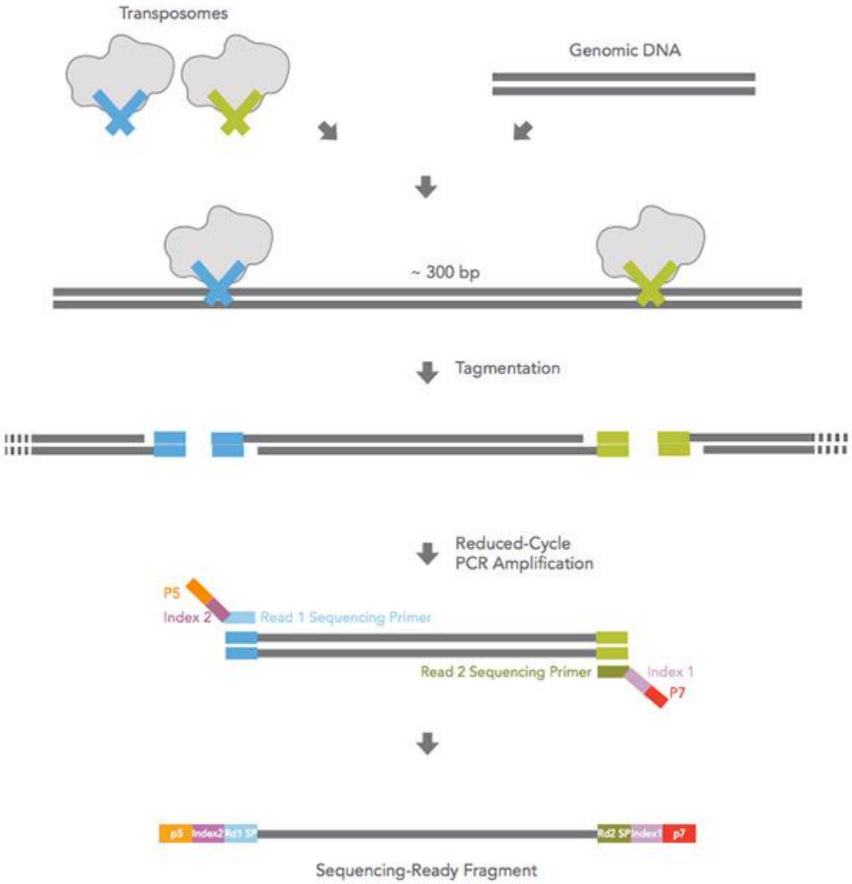
Illumina-sequencing emulator

Goal:

obtain a fast method of generating Illumina-like result of sequencing of the given genome



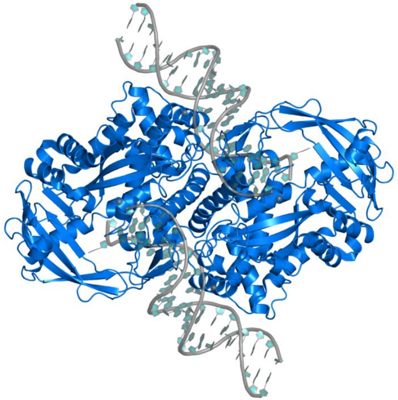
Transposon insertion fragmentation



First step conjecture:

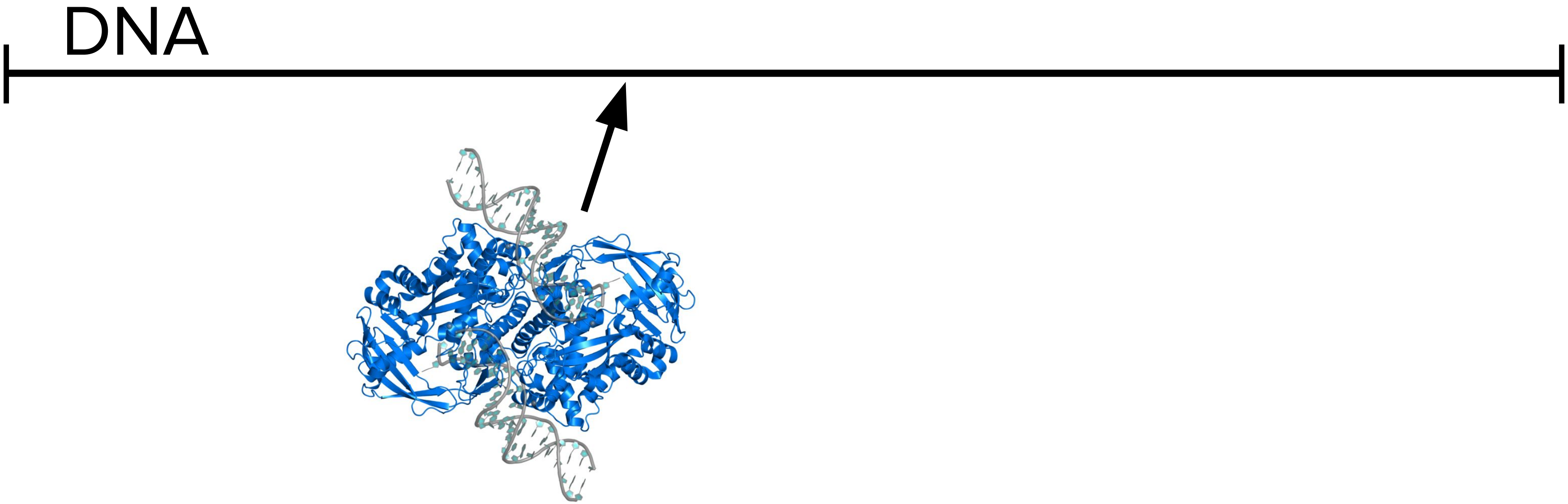
- uniformity of probability distribution of transposon landing site

DNA



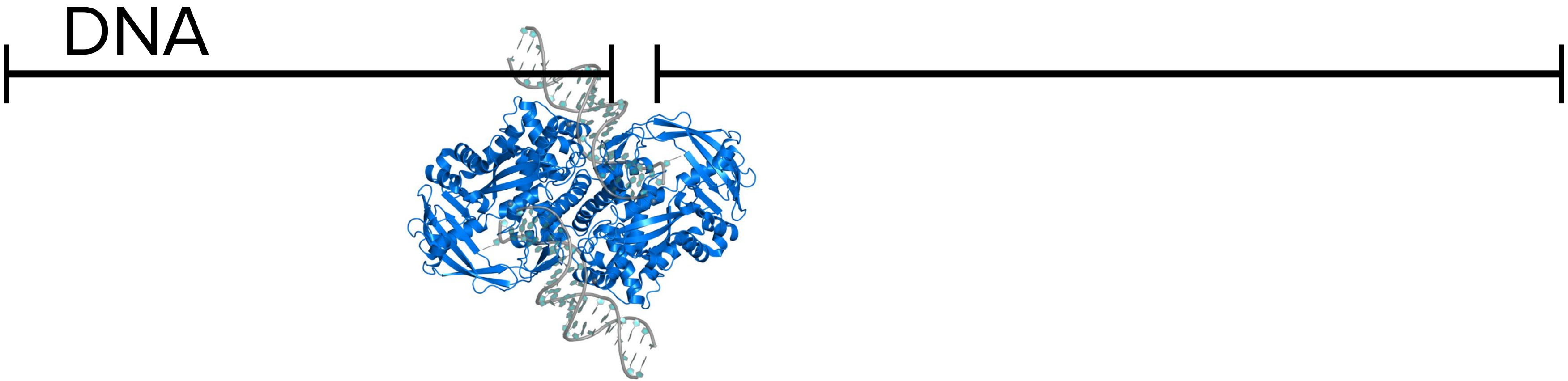
First step conjecture:

- uniformity of probability distribution of transposon landing site



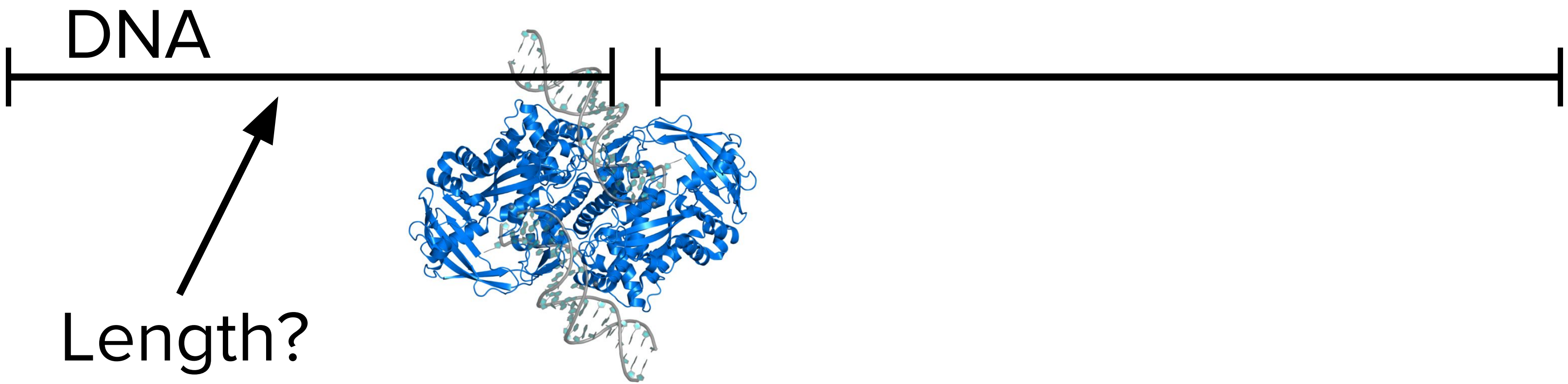
First step conjecture:

- uniformity of probability distribution of transposon landing site



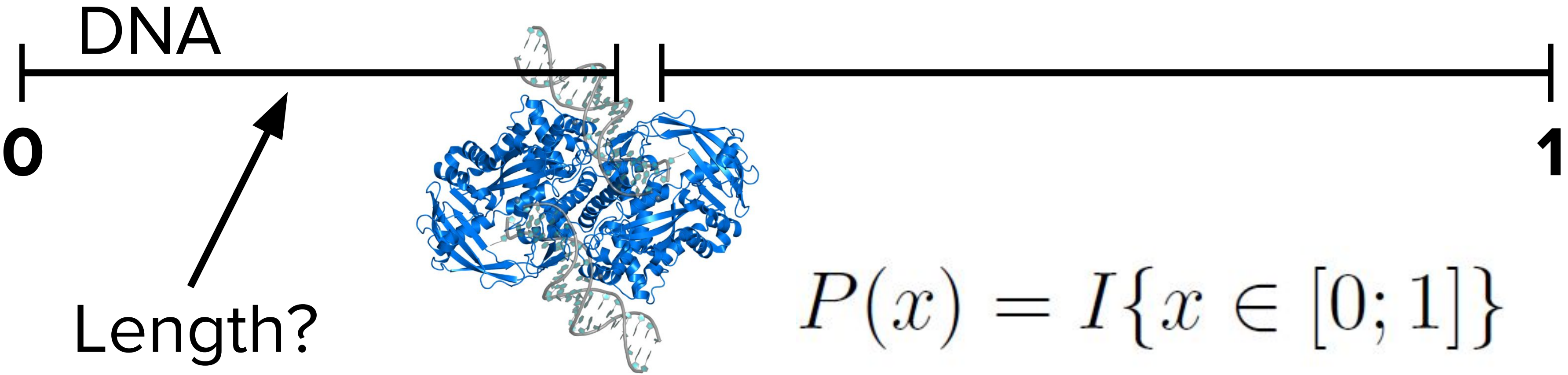
First step conjecture:

- uniformity of probability distribution of transposon landing site



First step conjecture:

- uniformity of probability distribution of transposon landing site



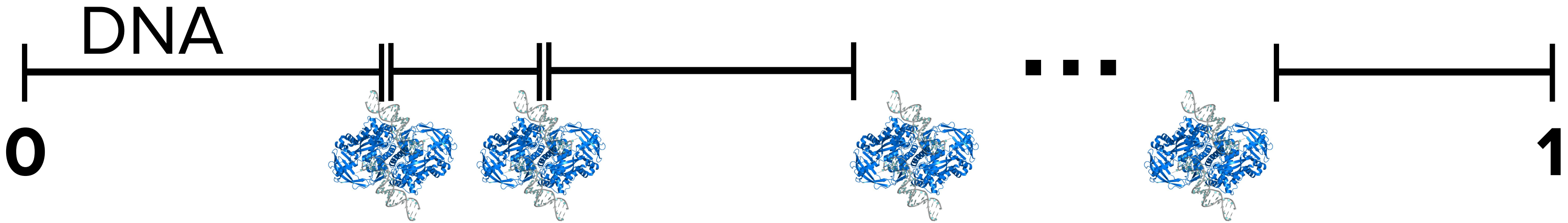
$$P(x) = I\{x \in [0; 1]\}$$

$$P\{\text{Length} : \text{Length} > x\} = 1 - P\{\text{Length} : \text{Length} \leqslant x\} =$$

$$= 1 - \int_{-\infty}^{x} P(\tau)\mathrm{d}\tau = \boxed{1 - x,\ x \in [0, 1]}$$

First step conjecture:

- uniformity of probability distribution of transposon landing site



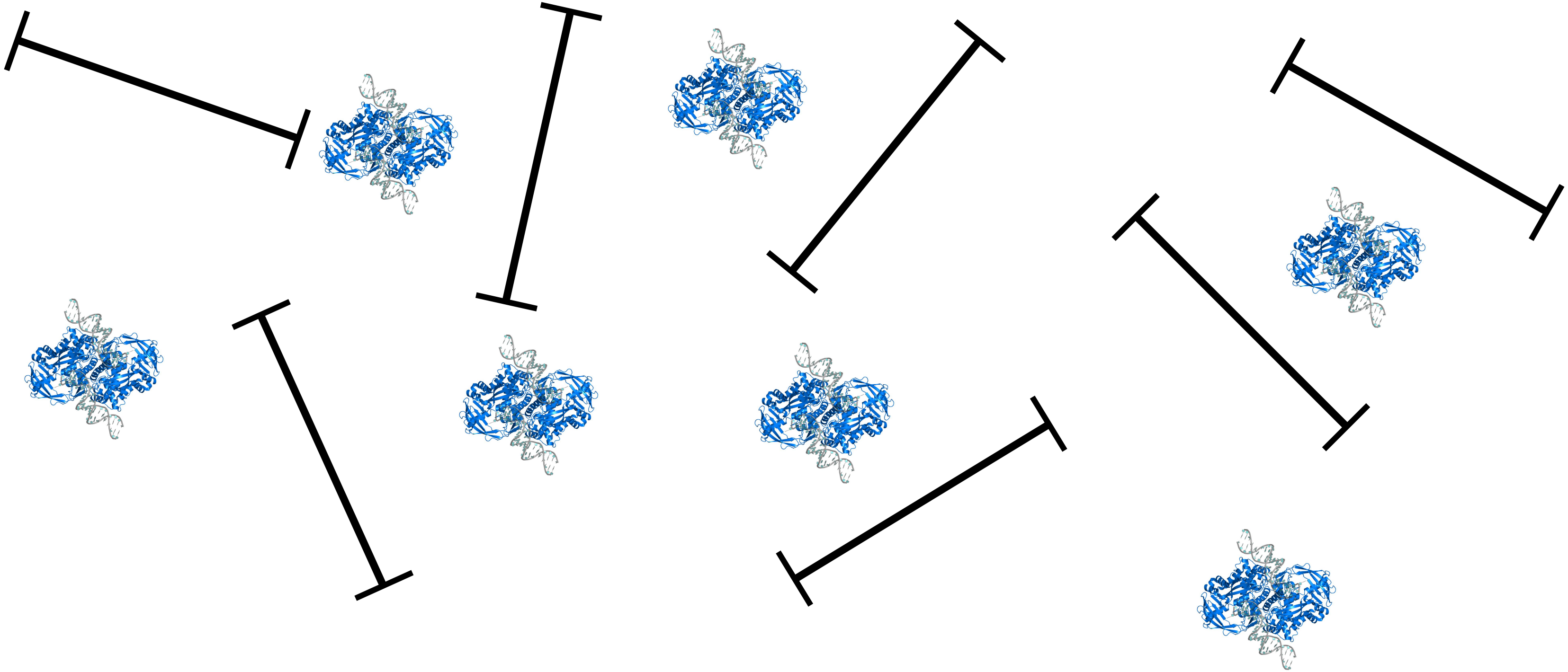
What if N transposons?

$$P_N\{\text{Length} : \text{Length} > x\} = (1-x)^N$$

$$P_N(x) = \mathrm{d}F_N(x)/\mathrm{d}x = -\mathrm{d}P_N\{\text{Length} : \text{Length} > x\}/\mathrm{d}x = \boxed{N * (1-x)^{N-1}}$$

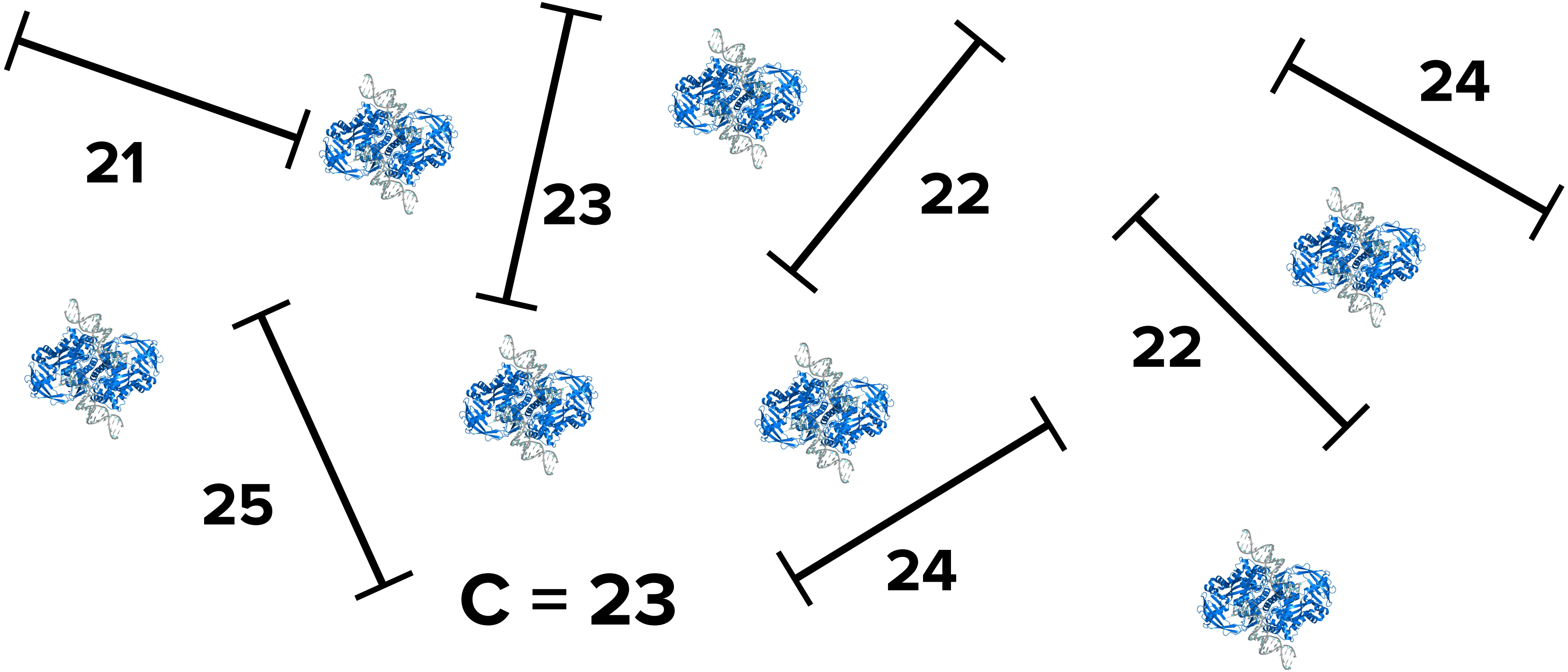
Second step conjectures:

- Concentration of transposons: C [number per DNA molecule]
- Number of cleavages per DNA is distributed according to Poisson



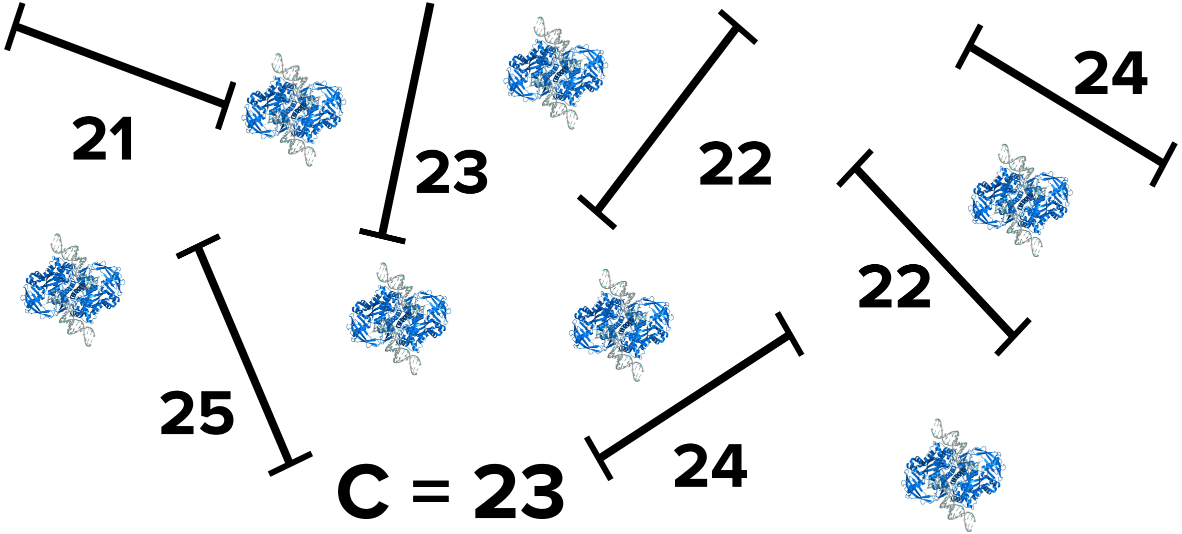
Second step conjectures:

- Concentration of transposons: C [number per DNA molecule]
- Number of cleavages per DNA is distributed according to Poisson



<u>Second step conjectures:</u>

- Concentration of transposons: C [number per DNA molecule]
- Number of cleavages per DNA is distributed according to Poisson



Number of cleavages:

Poisson(C)

$$P\{x = N\} = \frac{\exp(-\mathrm{C}) * \mathrm{C}^N}{N!}$$

Second step conjectures:

- Concentration of transposons: C [number per DNA molecule]
- Number of cleavages per DNA is distributed according to Poisson

$$P(x) = \sum_{N=0}^{+\infty} P(x|N)P(N) =$$

$$= \sum_{N=0}^{+\infty} N * (1-x)^{N-1} * \frac{\exp(-\mathrm{C}) * \mathrm{C}^N}{N!} =$$

$$\boxed{= \exp(-C) + C * \exp(-C * x)}$$

Number of cleavages:

Poisson(C)

$$P\{x = N\} = \frac{\exp(-\mathrm{C}) * \mathrm{C}^N}{N!}$$