Chromosome Painting.
Limit Theorems for the partitioning process.

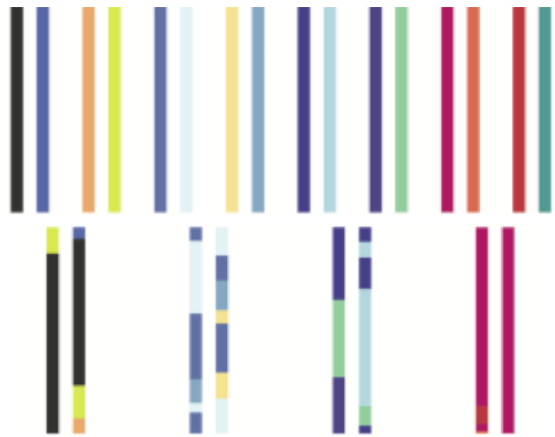
Emmanuel Schertzer.
Joint work with A. Lambert and V. Miro Pina.

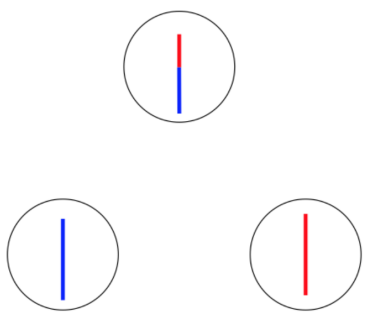
August 10, 2017

COLLÈGE
DE FRANCE
1530

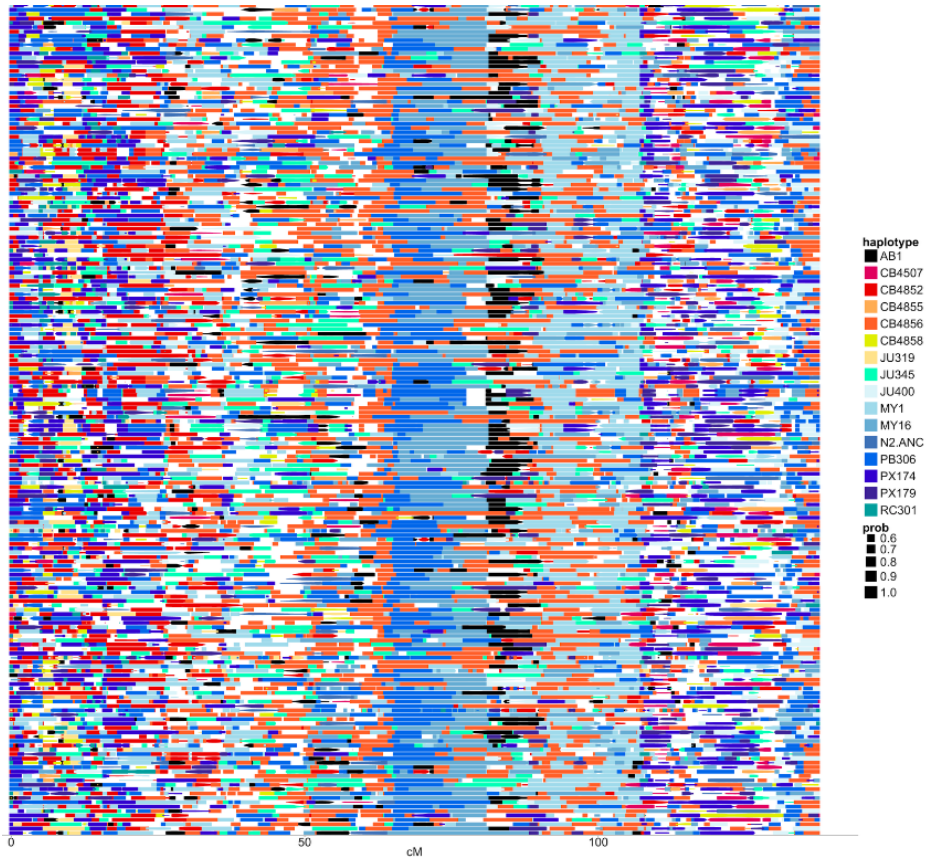
Chromosome painting: Experimental populations of Caenorhabditis elegans (Teotonio et al ('12))

- Start with 16 individuals.
- Build a population of size $\sim 10^4$ by random intercross
- Let it evolve during during 140 generations at controlled population size.
- Genotype 180 sequences.





Chromosome 3



- Segment = maximal connected set of of points sharing the same color.
- Cluster = maximal set of points sharing the same color.

Segment size

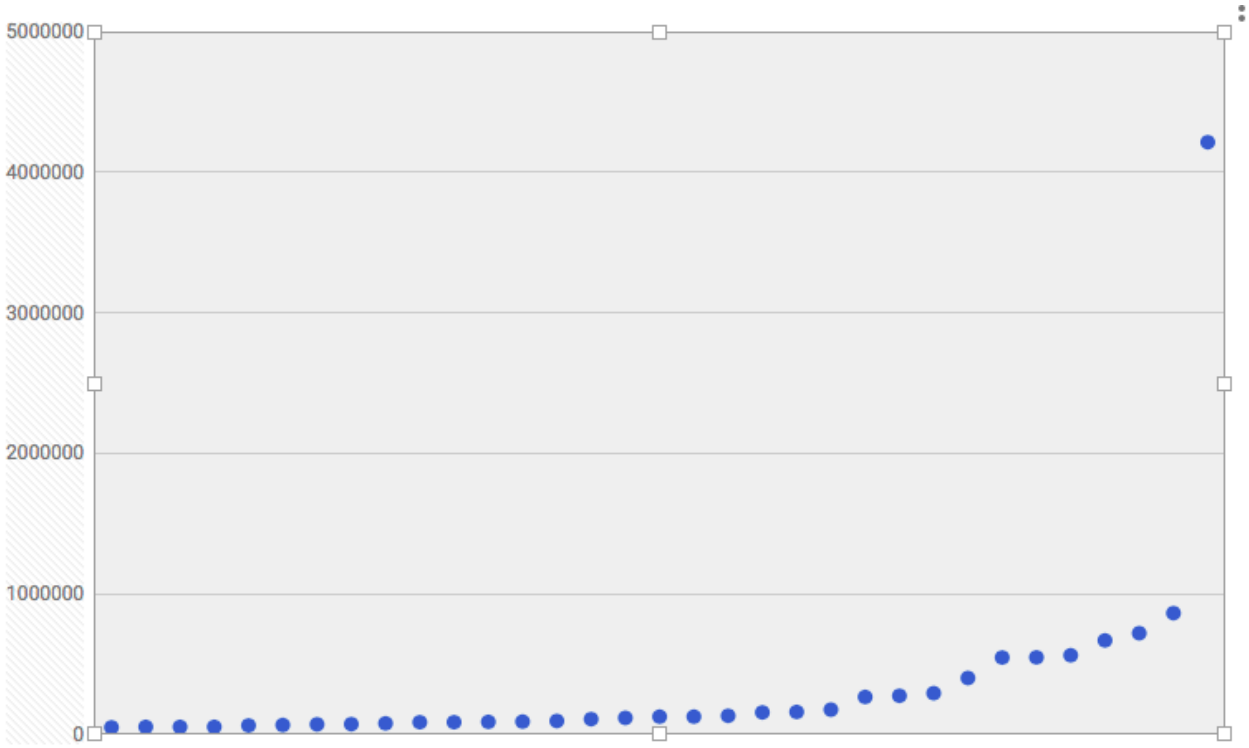


Figure: Segment size in increasing order

- Question : How to explain the outlier ?
- Intrinsic randomness of the population or result of natural selection ?
- Need of reliable statistical tests.

Chromosome painting

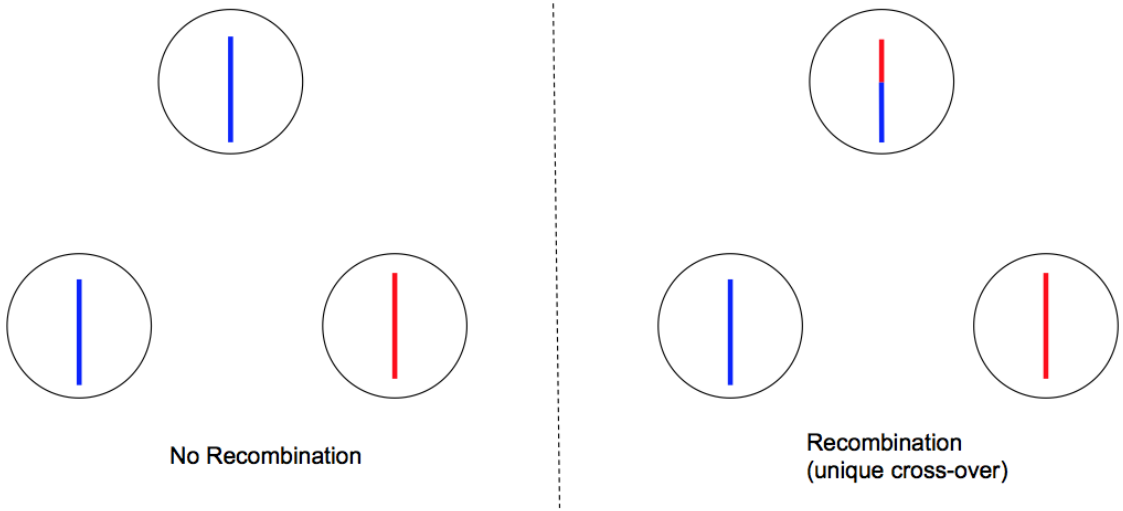
- Segment = maximal connected set of of points sharing the same color.
- Cluster = maximal set of points sharing the same color.

- What is the size of a typical segment ?
- What is the length, diameter of a typical cluster ?
- How many segments, clusters on a given interval ?
- etc.

An Haploïd Wright Fisher Model with Recombination

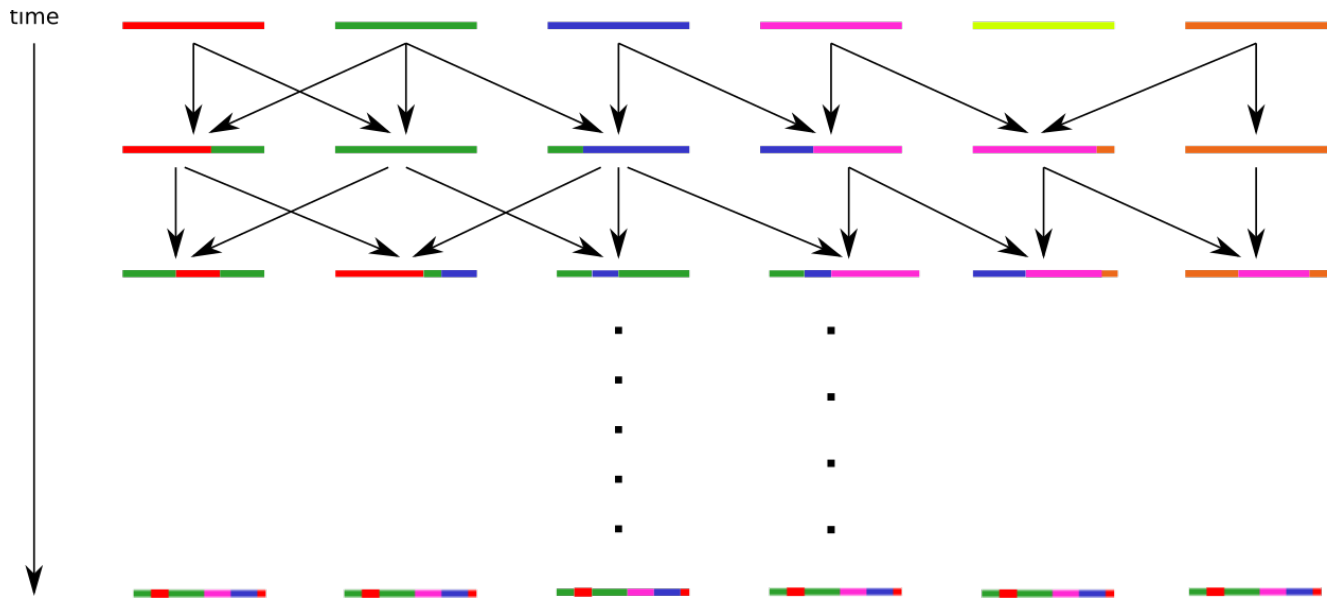
- Population of constant size N.
- Haploïd population: Each individual carries one chromosome of size R. $(R < N)$
- Discrete time dynamics:

time 0 Each chromosome is uniformly colored with a distinct color.
time 1 Each individual chooses 2 parents from the previous generation:
proba $1 - \frac{R}{N}$ copies one parent chromosome.
proba $\frac{R}{N}$ (Recombination event): a cross-over occurs.



An Haploïd Wright Fisher Model with Recombination

- At time 1, the population consist of N individuals, whose unique chromosome is either uniformly colored, or is partioned into two segments of distinct colors.



- After k steps, each chromosome is a mosaïc of colors, each colors corresponding to the genetic material of an ancestral individual.

- No mutation
- By genetic drift, the system a.s. reaches fixation after a finite (random) time, i.e., every individual in the system carries the same genetic material, and the system stops evolving.

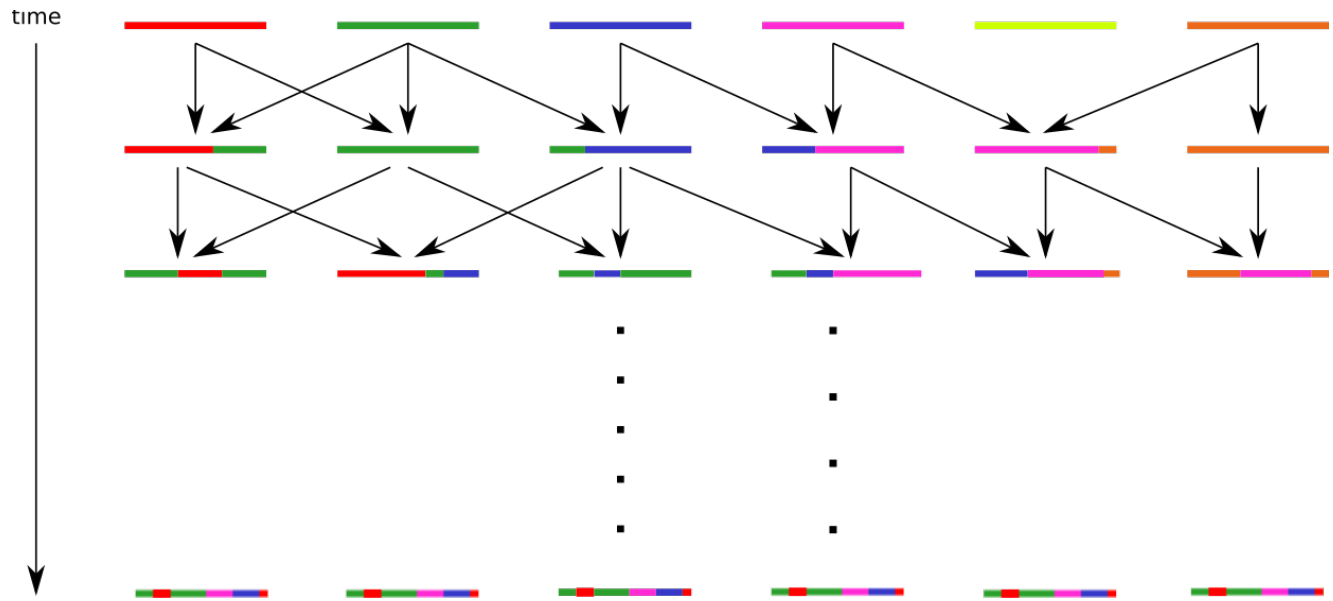


Figure: 6 segments. 4 clusters

- (N, R)-Partitioning process Π_N^R: partition of colors of the system at equilibrium (for a population of size N with chromosomes of size R.)

k-point partition

- $\mathcal{P}_k := \{ \text{ partitions of } \{1, \cdots, k\} \}$.
- Let $x_1 < \cdots < x_k \in [0, R]$. Define

$$\tau^{x_1, \cdots, x_k} \circ \Pi_N^R \in \mathcal{P}_k$$

the partition of $\{1, \cdots, k\}$ induced by $(x_1, \cdots, x_k)$ on $\{1, \cdots, k\}$, e.g.

  - If $x_1 \sim \cdots \sim x_k$ under Π_R^N then

$$\tau^{x_1, \cdots, x_k} \circ \Pi_N^R = \{1, \cdots, k\}.$$

  - If $x_i \sim x_j$ for $i \neq j$ then

$$\tau^{x_1, \cdots, x_k} \circ \Pi_N^R = \{1\} \cdots \{k\}.$$

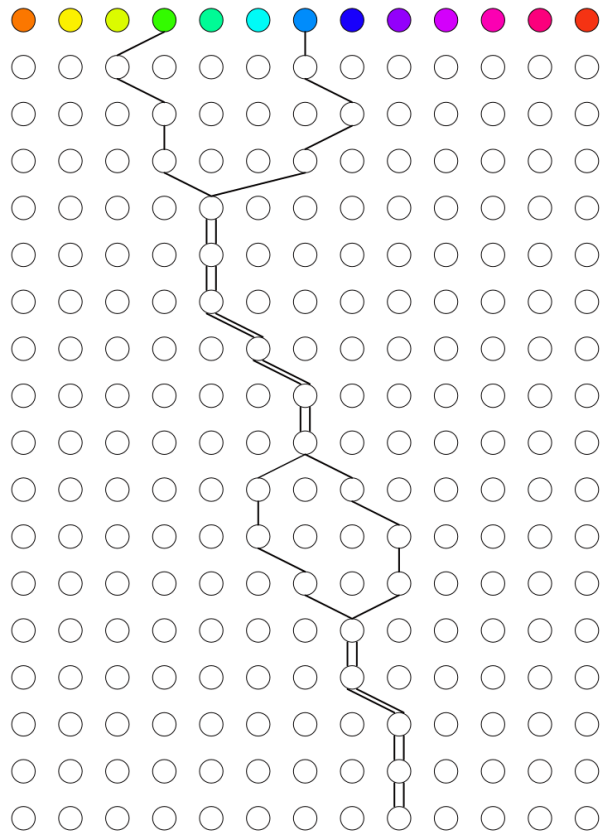
- Question 1 : What can we say about the law $\mathcal{L}\left(\tau^{x_1, \cdots, x_k} \circ \Pi_N^R\right)$ of the k-point partition ?
- Question 2 : Asymptotics when $N \to \infty$

2-point partition. The Ancestral Recombination Graph

- Consider two loci $x_1 < x_2 \in [0, R]$ on the same chromosome in the extant population.
- Follow their ascendances as time goes backward.
- At each generation, the common line of ascent $\{1, 2\}$ splits into $\{1\}\{2\}$ with probability

$$\frac{1}{N} |x_2 - x_1|$$

- At each generation, the singleton lines $\{1\}$ and $\{2\}$ coalesce with probability $1/N$.
- x_1, x_2 carry the same color iff their lines coincide at $-\infty$



2 point partition – 2 point motion

- $(X_N^{x_1,x_2}(t);\ t \geq 0)$ valued in $\mathcal{P}_2$.
- Coalescence $(\{1\}\{2\} \to \{1,2\})$ with probability 1/N.
- Fragmentation $(\{1,2\} \to \{1\}\{2\})$ with probability $\frac{1}{N}|x_2 - x_1|$.
- Let $\mu_N^{x_1,x_2}$ be the invariant measure of the 2-point motion.

Proposition

- Answer 1 : $\mathcal{L}(\tau^{x_1,x_2} \circ \Pi_N^R) \ = \ \mu_N^{x_1,x_2}$
- Answer 2 :

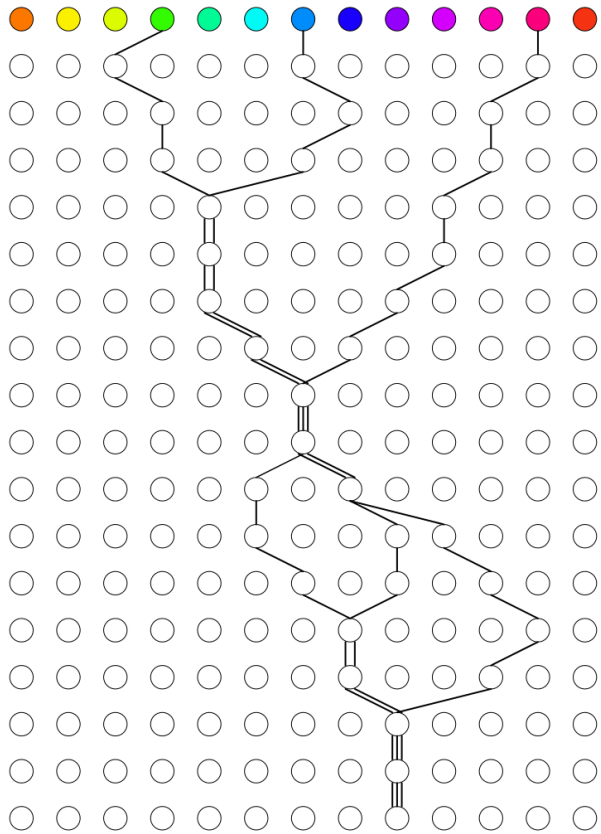
$$\left(X_N^{x_1,x_2}(tN);\ t \geq 0\right) \Longrightarrow (X_\infty^{x_1,x_2}(t);\ t \geq 0)$$

where the limiting process is the continuous time coagulation-fragmentation process with

- Coalescence at rate 1
- Fragmentation at rate $|x_2 - x_1|$

3-point motion. ARG with three sites

- Consider three loci $\{x_1, x_2, x_2\}$ with $x_1 < x_2 < x_2$.
- At each generation, the three lines of ascent split
  - $\{1, 2, 3\} \to \{1, 2\}\{3\}$ with proba $\frac{1}{N}|x_2 - x_3|$.
  - $\{1, 2, 3\} \to \{1\}\{2, 3\}$ with proba $\frac{1}{N}|x_1 - x_2|$.
- Not exchangeable: rate of fragmentation depends on relative position of the sites.
- At each generation, each pair of lines coalesce with probability $1/N$.
- x_1, x_2, x_3 carry different colors iff their lines of ascent don't coincide at $-\infty$



$$\mathcal{L}(\tau^{x_1,x_2,x_3} \circ \Pi_N^R) \;=\; \mu_N^{x_1,x_2,x_3}.$$

k point partition – k point motion

- Let $x_1 < \cdots < x_k$ in $[0, R]$.
- There is a natural k-point motion $(X_N^{x_1,x_2,\cdots,x_k}(t); t \in \mathbb{N})$ valued in $\mathcal{P}_k$ describing the ancestry of loci t units of time in the past.
- Answer 1 : $\mathcal{L}(\tau^{x_1,x_2,\cdots,x_k} \circ \Pi_N^R) = \mu_N^{x_1,x_2,\cdots,x_k}$
- Answer 2 : Convergence to a continuous time process on $\mathcal{P}_k$ with rates:

coalescence groups of lineages coalesce at rate 1.

fragmentation group of lineages $\{\sigma(0) < \cdots < \sigma(j) < \sigma(j+1) < \cdots < \sigma(K)\}$ splits into two parts :

$$\{\sigma(0) < \cdots < \sigma(j)\} \text{ and } \{\sigma(j+1) < \cdots < \sigma(K)\}$$

at rate $z_{\sigma(j+1)} - z_{\sigma(j)}$.

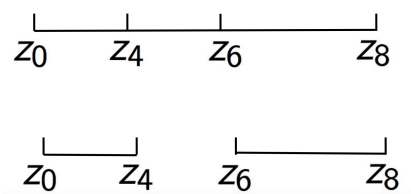


Figure: Fragmentation between z_4 and z_6 at rate $|z_4 - z_6|$

The partitoning process

Theorem

There exists a unique random variable Π_∞ valued in the set of locally finite partition of $\mathbb{R}$ such that for every $x_1 < \cdots < x_k$,

$$\mathcal{L}\left(\tau^{x_1,\cdots,x_k} \circ \Pi_\infty\right) \;=\; \mu_\infty^{x_1,\cdots,x_k}$$

(i.e., k-point partition described in terms of the invariant measure of the limiting k-point motion)

- $\left(\mu_\infty^S\right)_{S:S\subset\mathbb{R},|S|<\infty}$ is consistent (i.e., $\mu_\infty^{x_1,\cdots,x_n}$ is identical in law to the random partition induced by $\mu_\infty^{x_1,\cdots,x_n,x_{n+1}}$ on the first n coordinates).
- Let D be a dense set of $\mathbb{R}$. Define the skeleton $\Pi_\infty(D)$ such that

$$\forall x_1,\cdots x_n \in D,\; \tau^{x_1,\cdots,x_n} \circ \Pi_\infty(D) \;=\; \mu_\infty^{x_1,\cdots,x_n}$$

- Show that $\Pi_\infty(D)$ is locally finite.
- Take right limits to define Π_∞.

Large Population, Long Chromosome

Proposition

For every $R > 0$, as $N \to \infty$

$$\Pi_N^R \implies \Pi_\infty^R \quad \text{(convergence of } N\text{-point partitions)}$$

where Π_∞^R is the restriction of Π_∞ to $[0, R]$.

Question: What can we say about Π_∞^R on an interval of large size ? (For humans $R \approx 5 \times 10^4$)

Cluster covering the origin

Define

$$\mathcal{L}_R \;=\; \frac{1}{\log(R)} \int_0^R 1_{0\sim x} dx$$

the length of the cluster covering the origin.

Theorem (Lambert, Miro Pina, S.)

$$\lim_{R\to\infty} \mathcal{L}_R \;=\; \mathcal{E}(1) \;\text{ in law.}$$

Number of segments and clusters

Theorem (Lambert, Miro Pina, S.)

Let S_R be the number of segments in the interval $[0, R]$. Then

$$\lim_{R\to\infty} \frac{1}{R} S_R = 1 \text{ a.s.}$$

Typical size of a cluster on $[0, R]$ is of the order $\log(R)$. Thus, the number of clusters M_R should be of the order $R/\log(R)$.

Theorem (Lambert, Miro Pina, S.)

Let $\epsilon > 0$ and let $M_{R,\epsilon}$ be the number of clusters in the interval $[0, R]$ whose length is greater than $\epsilon \log(R)$. Then

$$\lim_{\epsilon\to 0} \lim_{R\to\infty} \frac{\ln(R)}{R} M_{R,\epsilon} = 1 \text{ in law.}$$

Number of Clusters Continued

Conjecture (Wiuf and Hein 97)

There exists a constant c such that $\frac{\ln(R)}{R} M_R \to c$ (in law, a.s. ?), with $c \approx 1.38 > 1$

For humans chromosome 1: $R \approx 5 \times 10^4$, and thus, the number of ancestors for chromosome 1 is approximatively $M_R \approx 6400$.

Idea of the proofs.

Proof for the Cluster Size at the Origin

- We aim at proving that

$$\lim_{R\to\infty} \quad \mathcal{L}_R \;=\; \mathcal{E}(1) \;\text{ in law.}$$

  where $\mathcal{L}_R$ is the length of the cluster at 0 on $[0, R]$.
- Main Idea: Method of moments.
- Using Carleman's condition, it is enough to show that

$$\lim_{R\to\infty} \mathbb{E}\left(\mathcal{L}_R^n\right) \;=\; n!$$

Proof for the Cluster Size at the Origin

$$
\begin{aligned}
\mathbb{E}(\mathcal{L}_R^n) &= \frac{1}{\log(R)^n}\mathbb{E}\left(\left(\int_0^R \mathbb{1}_{0\sim z} dz\right)^n \right) \\
&= \frac{1}{\log(R)^n}\mathbb{E}\left(\int_{[0,R]^n} \mathbb{1}_{0 \sim z_1 \cdots \sim z_n} dV \right) \\
&= \frac{1}{\log(R)^n} \int_{[0,R]^n} \mathbb{P}(0 \sim z_1 \cdots \sim z_n) dV(\mathbf{z}) \\
&= \frac{R^n}{\log(R)^n} \times \\
&\qquad \frac{1}{R^n} \int_{[0,R]^n} \mu_\infty^{\mathbf{z}}(\{1, \cdots, n+1\}) dV(\mathbf{z}) \to n!
\end{aligned}
$$

where $\mu_\infty^{\mathbf{z}}$ is the invariant distribution for the $N+1$ motion corresponding to $\mathbf{z} = (z_0 = 0, z_1, \cdots, z_n)$.

Estimating unlikely configuration in the n-point motion

Need to estimate

$$\frac{1}{R^n}\int_{[0,R]^n} \mu_\infty^{\mathbf{z}}(\{1,\cdots,n+1\})dV(\mathbf{z}) = E\left(\mu_\infty^{0,z_1,\cdots,z_n}(\{1,\cdots,n+1\})\right)$$

where the z_i's are chosen uniformly at random on $[0,R]$.

- Typical configuration : $\min_{i\neq j}|z_i - z_j| = O(R)$.
- Coagulation at rate 1. Fragmentation at rate $O(R)$ is much more frequent. And thus

$$\mu_\infty^{\mathbf{z}}\left(\{1\}\cdots\{n\}\right) \;=\; 1 - o(1)$$

(branching approximation)

- Need to estimate higher order terms.

Order of a partition

Definition

Let $\pi \in \mathcal{P}_n$. π is of order k iff it can be obtained from $\{1\}\cdots\{n\}$ by k successive coalescence events.

- $\{i,j\} +$ singletons is of order 1
- $\{i,j,k\} +$ singletons is of order 2. Three scenarios:

$$\begin{aligned}
\{i\}\{j\}\{k\}\cdots &\to \{i,j\}\{k\}\cdots \to \{i,j,k\}\cdots \\
\{i\}\{j\}\{k\}\cdots &\to \{i,k\}\{j\}\cdots \to \{i,j,k\}\cdots \\
\{i\}\{j\}\{k\}\cdots &\to \{k,j\}\{i\}\cdots \to \{i,j,k\}\cdots
\end{aligned}$$

- $\{i,j\},\{k,l\} +$ singletons is of order 2.
- $\cdots$
- $\{1,2,\cdots,n\}$ is of order $n-1$.

- Order of partition is a measure of the order of magnitude of its likelihood under $\mu_\infty^{\mathbf{z}}$.
- Lemma There exists c such that for every $\mathbf{z} = (z_1, \cdots, z_n)$, and every π such that $\text{Order}(\pi) = k$

$$\mu_\infty^{\mathbf{z}}(\pi) \leq \frac{c}{\min_{i \neq j} |z_i - z_j|^k} = O(\frac{1}{R^k}).$$

- Idea of the proof : Define $Y^{\mathbf{z}} = \text{Order}(X^{\mathbf{z}})$. The process is a non-markovian birth death process valued in $\{0, \cdots, n\}$ (coagulation (resp., fragmentation) induces positive (resp., negative) jumps).
- For every k, one can construct a true birth-death process Z such that

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T \mathbb{P}(Z(t) = k)\, dt \geq \lim_{T \to \infty} \frac{1}{T} \int_0^T \mathbb{P}(Y^{\mathbf{z}}(t) = k)\, dt = \mu_\infty^{\mathbf{z}}(\text{Order}(\pi) = k)$$

- Idea of the coupling: accelerate the excursions away from k.

- Need to compute $\mu^{\mathbf{z}}(\pi)$ for $\min_{i \neq j} |z_i - z_j| = O(R)$.
- Let $M^{\mathbf{z}}$ the transition matrix of the n point motion associated to $\{z_1, \cdots, z_n\}$:

$$\mu^{\mathbf{z}} M^{\mathbf{z}} \; = \; 0, \quad \mu^{\mathbf{z}}(\mathcal{P}_n) \; = 1.$$

- For every π of order $k \in \{1, \cdots, n-1\}$

$$M^{\mathbf{z}}(\pi, \pi)\mu^{\mathbf{z}}(\pi) \; = \underbrace{\sum_{\tilde{\pi} \; : \; \text{Order}(\pi)=k-1} \mu^{\mathbf{z}}(\tilde{\pi}) M^{\mathbf{z}}(\tilde{\pi}, \pi)}_{\textit{coalescence}} + \underbrace{\sum_{\tilde{\pi} \; : \; \text{Order}(\pi)=k+1} \mu^{\mathbf{z}}(\tilde{\pi}) M^{\mathbf{z}}(\tilde{\pi}, \pi)}_{\textit{fragmentation}}$$

- Neglecting the fragmentation part of the equation

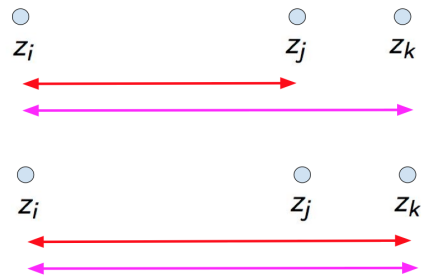
$$M^{\mathbf{z}}(\pi, \pi)\mu^{\mathbf{z}}(\pi) \; \approx \underbrace{\sum_{\tilde{\pi} \; : \; \text{Order}(\pi)=k-1} \mu^{\mathbf{z}}(\tilde{\pi}) M^{\mathbf{z}}(\tilde{\pi}, \pi)}_{\textit{coalescence}}$$

which provides a recurrence relation on the order.

Energy of a coalescence scenario

- $\{1,2,3\}$ is of order 2. Three scenarios:

$$
\begin{aligned}
s1: &\quad \{1\}\{2\}\{3\}\cdots \rightarrow \{1,2\}\{3\} \cdots \rightarrow \{1,2,3\} \\
s2: &\quad \{1\}\{2\}\{3\}\cdots \rightarrow \{1,3\}\{2\} \cdots \rightarrow \{1,2,3\} \\
s3: &\quad \{1\}\{2\}\{3\}\cdots \rightarrow \{2,3\}\{1\} \cdots \rightarrow \{1,2,3\}
\end{aligned}
$$



We define the energy of a coalescence scenario as the inverse of the product of the successive cover lengths at each step of the scenario.

$$
\begin{aligned}
\mathcal{E}(s1,\mathbf{z}) &= \frac{1}{z_2 - z_1} \times \frac{1}{z_3 - z_1} \\
\mathcal{E}(s2,\mathbf{z}) &= \frac{1}{z_2 - z_1} \times \frac{1}{z_2 - z_1}
\end{aligned}
$$

- Theorem : For every $z_1 < \cdots < z_n$:

$$\lim_{R \to \infty} R^k \mu_\infty^{R\mathbf{z}}(\{1, \cdots, n\}) \;=\; \sum_s \mathcal{E}^{R\mathbf{z}}(s)$$

where the sum is taken over every possible coalescence scenario from $\{1\}, \cdots \{n\}$ to $\{1, \cdots, n\}$.