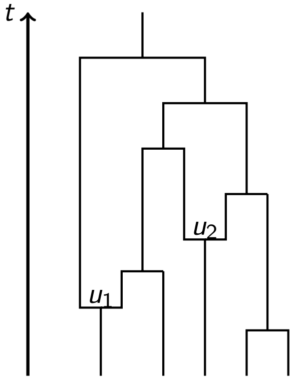
Recombination as a tree-valued process along the genome

Peter Pfaffelhuber
joint with Andrej Depperschmidt, Etienne Pardoux

Mainz, June 12, 2015

The Ancestral Recombination Graph (ARG)

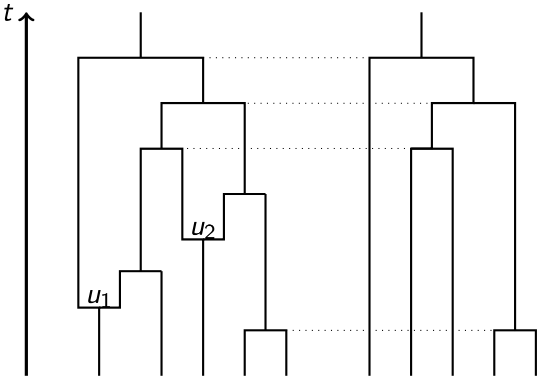
...along a genome $\mathbb{G} := [a, b], u_1 \leq u_2$



$$\mathcal{A} = \mathcal{A}^5$$

The Ancestral Recombination Graph (ARG)

...along a genome $\mathbb{G} := [0,1] \ni u_1, u_2$

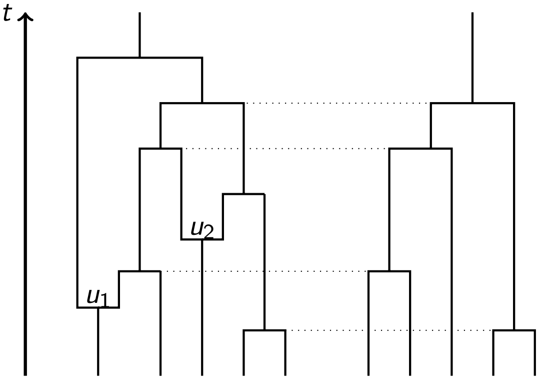


$\mathcal{A} = \mathcal{A}^5$

$\mathcal{T}_u(\mathcal{A}^5)$ $u < u_1, u_2$

The Ancestral Recombination Graph (ARG)

...along a genome $\mathbb{G} := [0,1] \ni u_1, u_2$

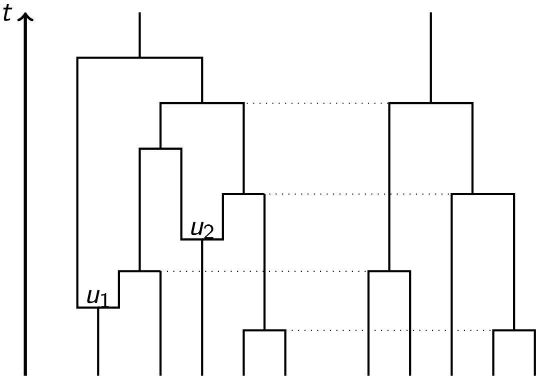


$\mathcal{A} = \mathcal{A}^5$

$\mathcal{T}_u(\mathcal{A}^5)$
$u_1 < u < u_2$

The Ancestral Recombination Graph (ARG)

...along a genome $\mathbb{G} := [0,1] \ni u_1, u_2$



$\mathcal{A} = \mathcal{A}^5$

$\mathcal{T}_u(\mathcal{A}^5)$
$u_1, u_2 < u$

Goal

- The ARG $\mathcal{A}$ from a population of size N gives rise to a tree-valued process $(\mathcal{T}_u^N)_{u\in\mathbb{G}}$.
- Goal 1: Check if

$$(\mathcal{T}_u^N)_{u\in\mathbb{G}} \xrightarrow[]{N\to\infty} \mathcal{T} = (\mathcal{T}_u)_{u\in\mathbb{G}}$$

  for some tree-valued process $\mathcal{T}$
- Goal 2: Study some properties of $\mathcal{T}$ (for $\mathbb{G} = (-\infty,\infty)$)

Convergence of N-ARGs

- Theorem 1: It holds that

$$\mathcal{T}^N \xRightarrow{N\to\infty} \mathcal{T}$$

on $\mathcal{D}_{\mathbb{M}}(\mathbb{G})$ for some process $\mathcal{T}$. The finite-dimensional distributions of $\mathcal{T}$ are given through the ARG.
The process $\mathcal{T}$ has almost surely finite variation with respect to

  - Gromov-Prohorov,
  - Gromov total variation and
  - Gromov-Hausdorff metrics.

Mixing properties

- Theorem 2: Let $(\mathcal{T}_u)_{u\in(-\infty,\infty)}$ be as above and Φ, Ψ be polynomials. Then, there is $C = C_{\Phi,\Psi} > 0$ such that for all $u > 0$

$$\left|\mathbb{E}[\Psi(\mathcal{T}_0)\Phi(\mathcal{T}_u)] - \mathbb{E}[\Psi(\mathcal{T}_0)]\mathbb{E}[\Phi(\mathcal{T}_u)]\right| \le \frac{C}{u^2}.$$

- Surprise: From Jenkins et al, one would have guessed a lower order term

$$\mathbb{E}[\Psi(\mathcal{T}_0)\Phi(\mathcal{T}_u)] = \mathbb{E}[\Psi(\mathcal{T}_0)]\mathbb{E}[\Phi(\mathcal{T}_u)] + \mathcal{O}\left(\frac{1}{u}\right).$$

Formalizing genealogical trees

- **Leaves in genealogical trees** form a metric space

A tree is given by:

(X, r) complete and separable **metric** space

- $r(x_1, x_2)$ defines the genealogical distance of individuals x_1 and x_2

Formalizing genealogical trees

- **Leaves in genealogical trees** form a metric space

A tree is given by:

(X, r) complete and separable **metric** space, $\mu \in \mathcal{P}(X)$

- $r(x_1, x_2)$ defines the genealogical distance of individuals x_1 and x_2
- μ marks currently living individuals

Formalizing genealogical trees

- **Leaves in genealogical trees** form a metric space

State space of $\mathcal{T}$:

$$\begin{aligned}\mathbb{M} :=\{&\text{isometry class of } (X, r, \mu) : \\ &(X, r) \text{ complete and separable } \textbf{metric} \text{ space}, \ \mu \in \mathcal{P}(X)\}\end{aligned}$$

- $r(x_1, x_2)$ defines the genealogical distance of individuals x_1 and x_2
- μ marks currently living individuals

Gromov-Prohorov topology

- Polynpomials: Functions on $\mathbb{M}$ of the form

$$\Phi(X, r, \mu) := \int \phi(r(\underline{x}, \underline{x}))\mu^{\mathbb{N}}(d(\underline{x})$$

for $\underline{x} = (x_1, x_2, \ldots), \phi \in \mathcal{C}_b(\mathbb{R}^{\binom{\mathbb{N}}{2}})$ depending on finitely many coordinates

- The Gromov-Prohorov topology on $\mathbb{M}$ is given as the coarsest topology making all polynomials continuous

Example: Kingman measure tree

- Single locus: genealogical tree $\mathcal{T}^N$
- Theorem 4 in Greven, P, Winter (2009):
  There exists an $\mathbb{M}$-valued random variable $\mathcal{T}$ such that

$$\mathcal{T}^N \xRightarrow{N\to\infty} \mathcal{T}.$$

- Proof: Tightness by coming down from infinity; uniqueness since polynomials form a separating algebra of functions.

Gromov-Prohorov metric

- Recall the Prohorov metric of two probability measures μ, ν

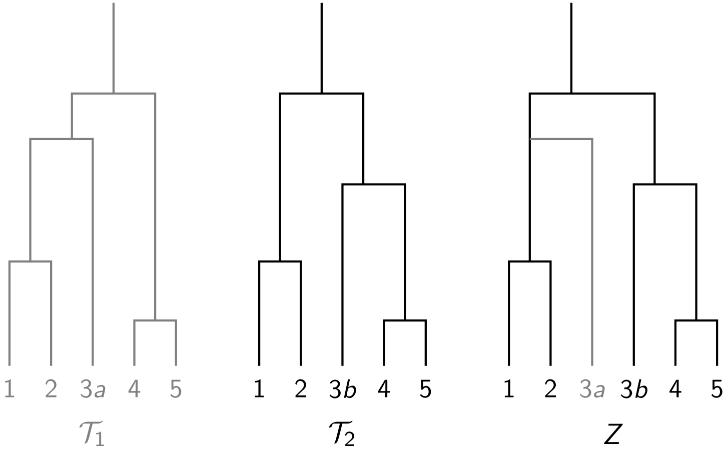
$$d_P(\mu,\nu) := \inf\{\varepsilon > 0 : \mu(A) \leq \nu(A^\varepsilon) + \varepsilon, A \text{ closed } \}$$

- Let (X_i, r_i, μ_i) be mm-spaces, $\varphi_1 : X_1 \to Z$ for $i = 1, 2$ be isometric embeddings into a common complete and separable metric space (Z, d).
- The **Gromov-Prohorov metric** is defined by

$$d_{\mathrm{GP}}((X_1, r_1, \mu_1), (X_2, r_2, \mu_2)) := \inf_{\varphi_1, \varphi_2, Z} d_{\mathrm{P}}((\varphi_1)_* \mu_1, (\varphi_2)_* \mu_2).$$

- Theorem (Gromov; Greven, P, Winter, 2009): The Gromov-Prohorov metric is complete and metrizes the Gromov-Prohorov topology.

Example: $d_{\mathrm{GP}}(\mathcal{T}_1, \mathcal{T}_2) \leq 1/5$



Total variation distance

- If Z is countable, the total variation distance of probability measures μ, ν on Z is given by

$$d_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2} \sum_{z \in Z} |\mu(z) - \nu(z)|. \tag{1}$$

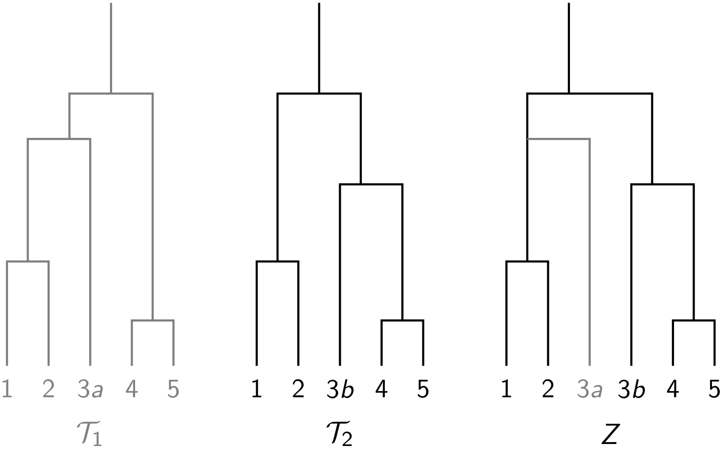
- Recall that

$$d_{\mathrm{P}}(\mu, \nu) \le d_{\mathrm{TV}}(\mu, \nu).$$

- The *Gromov total variation distance* is defined by

$$d_{\mathrm{GTV}}((X_1, r_1, \mu_1), (X_2, r_2, \mu_2)) := \inf_{\varphi_1, \varphi_2, Z} d_{\mathrm{TV}}((\varphi_1)_* \mu_1, (\varphi_2)_* \mu_2).$$

Example: $d_{\mathrm{GTV}}(\mathcal{T}_1, \mathcal{T}_2) \leq 1/5$



Proof of Theorem 1: Main steps

- Fdd-convergence: similar to convergence of Kingman measure tree.
- Tightness: Find $C > 0$ such that

$$\limsup_{N\to\infty} \mathbb{E}[d_{\mathrm{GTV}}(\mathcal{T}_{-h}^N, \mathcal{T}_0^N) \cdot d_{\mathrm{GTV}}(\mathcal{T}_0^N, \mathcal{T}_h^N)] \leq Ch^2.$$

Auxiliary distance

- Within the ARG, define

$$d^{u,v}(\mathcal{T}_u^N, \mathcal{T}_v^N) := \frac{\#\left\{\begin{array}{l} i \text{ leaf in } \mathcal{T}_u^N, \text{ hit by a splitting event} \\ \text{marked with } U \in [u,v] \text{ before reach-} \\ \text{ing the root of } \mathcal{T}_u^N \end{array}\right\}}{N}$$

- Then,

$$d_{\mathrm{GTV}}(\mathcal{T}_u^N, \mathcal{T}_v^N) \leq d^{u,v}(\mathcal{T}_u^N, \mathcal{T}_v^N)$$

and

$$d^{0,-h}(\mathcal{T}_0^N, \mathcal{T}_{-h}^N), d^{0,h}(\mathcal{T}_0^N, \mathcal{T}_h^N)$$

conditionally independent given $\mathcal{T}_0^N$.

Main step

- Lemma:

$$\mathbb{E}[d^{0,h}(\mathcal{T}_0^N, \mathcal{T}_h^N)|\mathcal{T}_0^N] \leq h \sum_{k=2}^{N} S_k.$$

- Corollary: There is $C > 0$ such that

$$\begin{aligned}
&\mathbb{E}[d_{\mathrm{GTV}}(\mathcal{T}_{-h}^N, \mathcal{T}_0^N) \cdot d_{\mathrm{GTV}}(\mathcal{T}_0^N, \mathcal{T}_h^N)] \\
&\qquad \leq \mathbb{E}\big[\mathbb{E}[d^{0,-h}(\mathcal{T}_0^N, \mathcal{T}_{-h}^N)|\mathcal{T}_0^N] \cdot \mathbb{E}[d^{0,h}(\mathcal{T}_0^N, \mathcal{T}_h^N)|\mathcal{T}_0^N]\big] \\
&\qquad \leq Ch^2
\end{aligned}$$

Bounding the Gromov-Hausdorff distance

- Recall the Hausdorff metric of two sets $A, B \subseteq Z$

$$d_H(A, B) := \inf\{\varepsilon > 0 : A \subseteq B^\varepsilon, B \subseteq A^\varepsilon\}$$

- Let (X_i, r_i, μ_i) be mm-spaces, $\varphi_1 : X_1 \to Z$ for $i = 1, 2$ be isometric embeddings into a common complete and separable metric space (Z, d).
- The **Gromov-Hausdorff metric** is defined by

$$d_{\mathrm{GH}}((X_1, r_1, \mu_1), (X_2, r_2, \mu_2)) := \inf_{\varphi_1, \varphi_2, Z} d_{\mathrm{P}}(\varphi_1(X_1), \varphi_2(X_2)).$$

- Bound the time when a recombinant line coalesces back into the tree leads to a $C > 0$ such that

$$\mathbb{E}[d_{\mathrm{GH}}(\mathcal{T}_0^N, \mathcal{T}_h^N)] \leq Ch.$$

This implies finite variation in Gromov-Hausdorff sense.

Mixing properties

- Theorem 2: For $n \in \mathbb{N}$ let Φ, Ψ be polynomials and $(\mathcal{T}_u)_{u \in (-\infty,\infty)}$ be as above. Then, there is $C = C_{\Psi,\Phi} > 0$ such that for all $u > 0$

$$\left|\mathbb{E}[\Psi(\mathcal{T}_0)\Phi(\mathcal{T}_u)] - \mathbb{E}[\Psi(\mathcal{T}_0)]\mathbb{E}[\Phi(\mathcal{T}_u)]\right| \leq \frac{C}{u^2}.$$

Proof of Theorem 2: Main idea

- Let Φ, Ψ be polynomials of degree 2, i.e. only depend on a single genealogical distance. Then, for an ARG $\mathcal{A}^4$, and distances R_0, R_u,

$$\mathbb{E}[\Psi(\mathcal{T}_0)\Phi(\mathcal{T}_u)] = \mathbb{E}\big[\phi(R_0(1,2))\psi(R_u(3,4))\big].$$

- $R_0(1,2), R_u(3,4)$ are independent unless $R_0(1,2) = R_u(3,4)$. But

$$\mathbb{P}(R_0(1,2) = R_u(3,4)) = \mathcal{O}\Big(\frac{1}{u^2}\Big)$$

(whereas

$$\mathbb{P}(R_0(1,2) = R_u(1,2)) = \mathcal{O}\Big(\frac{1}{u}\Big).)$$