perm.pub/dsi: Fs4CjTor07Ssbb69Qcggp5yiBA0/0.3
Additional formats and editions available online.

Edition 0.3

Obsolete

Author date: 2023-03-06

Citation:
E. Castedo Ellerman (2023) "Gametic lineage space" perm.pub https://perm.pub/dsi: Fs4CjTor07Ssbb69Qcggp5yiBA0/0.3

Gametic lineage space

E. Castedo Ellerman (castedo@castedo.com)

Abstract

STAGE: WORKING DRAFT

DOCUMENT TYPE: Definition Document

OBJECTIVES

- Formal mathematical definition to be used in stochastic process model under development.
- Solicit feedback, in particular, on choice of wording for definitions.
- Precise mathematical definitions for technical discussions relating to ancestral recombination graphs

Gametic genealogy

A *gametic genealogy* is a convenient mathematical formalism of the genealogy of a population from the perspective of gametes. Mathematically, it is a quadruple

$$(\mathsf{Gam}, \mathsf{Mate}, \mathsf{Par}, \mathsf{Fert})$$

with components

- Gam, the set of underlying gametes,
- Mate, the set of zygotes formed by the fusion of egg gametes and sperm gametes,
- Par, a mapping from child gametes to parent zygotes, and
- Fert, a mapping from zygotes to fertilization time.

For convenience, given a *gametic genealogy*,

- Gam_0 denotes the set of egg gametes,
- Gam_1 denotes the set of sperm gametes, and
- Mate_* denotes the mapping from gametes to the zygotes they formed during fertilization.

Formally, a *gametic genealogy* must satisfy the following conditions.

1. $\mathsf{Gam}_0 \cup \mathsf{Gam}_1 = \mathsf{Gam}$ and $\mathsf{Gam}_0 \cap \mathsf{Gam}_1 = \emptyset$.
2. $\mathsf{Mate} \subset \mathsf{Gam}_0 \times \mathsf{Gam}_1$ and forms a one-to-one mapping between Gam_0 and Gam_1.

3. Par is a function $C \mapsto \mathsf{Mate}$, where C is a subset of Gam representing child gametes.
4. Fert is a function $\mathsf{Mate} \mapsto \mathbb{R}$ such that for all child gametes $g \in \operatorname{dom} \mathsf{Par}$,

$$\mathsf{Fert}(\mathsf{Mate}_*(g)) > \mathsf{Fert}(\mathsf{Par}(g)) \ .$$

$\operatorname{dom} \mathsf{Par}$ denotes the domain of Par, that is, the set of child gametes.

Gametic lineage space

A *gametic lineage space* is a mathematical formalism representing the lines of transmission of genetic information via gametes of a population over time. It is a triplet

$$(\mathsf{Loc}, G, \mathsf{Lin})$$

where

- Loc is the set of all genomic locations,
- G is a gametic genealogy $(\mathsf{Gam}, \mathsf{Mate}, \mathsf{Par}, \mathsf{Fert})$, and
- Lin is a function $\mathsf{Loc} \times \mathsf{Gam} \mapsto 2^{\mathsf{Gam}}$ mapping a genomic position in a gamete to the set of gametes that transmitted genetic information to that position in that gamete.

For every location $\ell \in \mathsf{Loc}$ and gamete $g \in \mathsf{Gam}$, $\mathsf{Lin}(\ell, g)$ is the lineage ending at gamete g via locus ℓ and it must satisfy the condition

$$\mathsf{Lin}(\ell, g) = \{g\} \cup \mathsf{Lin}(\ell, \mathsf{Par}(g)_i) \text{ for either } i = 0 \text{ or } i = 1$$

when $g \in \operatorname{dom} \mathsf{Par}$, otherwise $\mathsf{Lin}(\ell, g) = \{g\}$.

$\mathsf{Par}(g)_0$ and $\mathsf{Par}(g)_1$ are the maternal and paternal gametes, respectively, that fertilized the parent of g.

Stochastic gametic lineage space

A *stochastic gametic lineage space* is a gametic lineage space extended to model a random gametic lineage.

From a stochastic gametic linage space, a time-indexed family of probability distributions $\{P_t\}_{t \in I}$ is induced.

> *TO DO: Need to rework space formalism to clarify over what is the σ-algebra: issues #42.*

For convenience we define the set of all lineages that contain a gamete $g \in \mathsf{Gam}$ as

$$B(g) := \{\mathsf{Lin}(\ell, s) : g \in \mathsf{Lin}(\ell, s), \ell \in \mathsf{Loc}, s \in \mathsf{Gam}\} \ .$$

Formally, a *stochastic gametic lineage space* is a quintuple

$$(G, I, \{S_i\}_{i \in I}, \mathcal{F}, \mu)$$

where

- G is a gametic lineage space $(\mathsf{Loc}, (\mathsf{Gam}, \mathsf{Mate}, \mathsf{Par}, \mathsf{Fert}), \mathsf{Lin})$

- I is an index set of points in time with $\mathrm{rng}\,\mathsf{Fert} \subset I$,
- $\{S_t\}_{t\in I}$ is a time-indexed collection of sets of living zygotes,
- $\mathcal{F}$ is a σ-algebra ($sigma$-field) over $\mathrm{rng}\,\mathsf{Lin}$, and
- μ is a measure on $\mathcal{F}$

which satisfy the following conditions

- $B(g) \in \mathcal{F}$ for all $g \in \mathsf{Gam}$, and
- $\mu(S_t)$ is defined and finite for all $t \in I$.

Every gametic lineage space induces a time-indexed family $\{P_t\}_{t\in I}$ of probabilities spaces measurable on σ-algebra F. This defines the probability of lineages which end in a zygote alive at time t.

> *TO DO: Need to clarify relationship between F and* Loc *and* Mate *for when they are uncountable.*

An embedded Ancestral Recombination Graph

An ancestral recombination graph [1] [2] [3] of a sampled population is embedded in a gametic lineage space. We formally show the exact embedding using the gARG formalism [4].

We start by defining the *genetic legacy* of a gamete $g \in \mathsf{Gam}$ for sample population $S \subseteq \mathsf{Gam}$ to be

$$\mathsf{Leg}(g, S) := \{(\ell, d) \in \mathsf{Loc} \times S : g \in \mathsf{Lin}(\ell, d)\} .$$

This genetic legacy is the genetic material that survives in the sample population S originally copied from ancestral gamete g (with or without mutations).

Genetic legacy for a sample population S induces the following equivalence relationship over pairs of gametes g_1 and g_2 in Gam:

$$g_1 \simeq_S g_2 \;:=\; \mathsf{Leg}(g_1, S) = \mathsf{Leg}(g_2, S) .$$

We denote the resulting equivalence class containing $g \in \mathsf{Gam}$ as

$$[g]_S \;:=\; \{g' : \mathsf{Leg}(g', S) = \mathsf{Leg}(g, S)\} .$$

In this equivalence relationship, gametes are considered equivalent if they have the same genetic legacy for the sample population S.

A convenient choice for an embedded gARG [4] is to set the gARG nodes (vertices) to be the equivalence classes:

$$\mathsf{Nodes}(S) := \{[g]_S : g \in \mathsf{Gam}\} .$$

The (unannotated) graph edges of the gARG are chosen as child-parent node pairs $(C, P) \in \mathsf{Nodes}(S)^2$ where

$$\mathsf{Par}(g)_i \in P \text{ for some } g \in C \text{ and some } i \in \{0, 1\} .$$

In the gARG, annotations are added for each graph edge (pair of child and parent nodes). This annotation is the set of locations through which genetic information has been copied from

parent to child. In the following interpretation, the only locations of interest are those for which genetic information has been transmitted into the sample population S. With this interpretation, the annotation for edge (C, P) is

$$\{\ell \in \mathsf{Loc} : C \cup P \subseteq \mathsf{Lin}(\ell, g) \text{ for some } g \in S\} .$$

Acknowledgements

Thanks to Daria Shipilina and Nick Barton for sharing their preprint [5] and discussing the conjecture in edition 0.1 of this document relating to their preprint.

Changes from edition 0.1

- add section about embedded ARG
- removed conjecture relating to [5]

References

1. Griffiths RC, Marjoram P. An Ancestral Recombination Graph. In: Friedman A, Miller W, Donnelly P, Tavaré S, editors. Progress in Population Genetics and Human Evolution. New York, NY: Springer New York; 1997. pp. 257–270. doi:10.1007/978-1-4757-2609-1_16
2. Hein J, Schierup MH, Wiuf C. Gene genealogies, variation and evolution: A primer in coalescent theory. Oxford ; New York: Oxford University Press; 2005.
3. Wakeley J. Coalescent theory: An introduction. Greenwood Village, Colo: Roberts & Co. Publishers; 2009.
4. Wong Y, Ignatieva A, Koskela J, Gorjanc G, Wohns AW, Kelleher J. A general and efficient representation of ancestral recombination graphs. https://archive.softwareheritage.org/swh:1:rev:7df4f1995028cc676a6c1b231e8d7a024666b5fc; 2022.
5. Shipilina D, Stankowski S, Pal A, Chan YF, Barton N. On the origin and structure of haplotype blocks. Preprints; 2022 Feb. doi:10.22541/au.164425910.09070763/v1