Chapter 13

Nonredundant Representation of Ancestral Recombinations Graphs

Laxmi Parida

Abstract

The network structure that captures the common evolutionary history of a diploid population has been termed an ancestral recombinations graph. When the structure is a tree the number of internal nodes is usually $\mathcal{O}(K)$ where K is the number of samples. However, when the structure is not a tree, this number has been observed to be very large. We explore the possible redundancies in this structure. This has implications both in simulations and in reconstructability studies.

Key words: Ancestral recombinations graph, ARG, Redundancies, Minimal descriptor, Coalescent, Wright–Fisher, Population simulators, Nonredundant

1. Introduction

In keeping with the theme of the book, we study in this chapter the common evolutionary history of a diploid population. This common history is a phylogeny with the extant members at the terminal or leaf nodes. The internal nodes of the topology are some common ancestors while the edges can be viewed as conduits for the flow of genetic material. The direction on the edges represents the direction of flow. A directed edge from node v_1 to node v_2 is to be interpreted as v_1 being an ascendant of v_2 or v_2 is a descendant of v_1. The topology has no cycles since, no matter what the underlying model, a member is not an ancestor of itself. Thus, the topology is always a directed acyclic graph (DAG). Under uni-parental (unilinear) transmission each member at a generation derives all its genetic material from only one parent whereas under a biparental model a member derives the material from two parents. Then does this simple difference in inheritance in the two models have an effect

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods, Volume 2*,
Methods in Molecular Biology, vol. 856, DOI 10.1007/978-1-61779-585-5_13,

on the overall topology of the common evolutionary history? Under uniparental model a unit has only one ancestor (ascendant) in an earlier generation while under biparental model a unit can have multiple ancestors. But in both models, a unit can have multiple descendants at a future generation. Thus, the DAG for only the uniparental model is guaranteed to be a tree.

One of the primary genetic events shaping an autosomal chromosome is *recombination* which is a process that occurs during meiosis that results in the offsprings having different combinations of homologous genes, or chromosomal segments, of the two parents. The topology incorporating this has been called the ancestral recombinations graphs (ARGs) and is an annotated network structure that captures the common evolutionary history of the extant haplotypes. This subject is also discussed in the chapter on "Ancestral Population Genomics" in this book. The random mathematical object, ARG, was introduced in the context of modeling population evolution in the field of population genetics (1, 2). Thus, the ARG is not only used for modeling population evolution (3), but is also the object of interest in the reconstruction of the evolution history from the haplotypes of extant samples (4, 5). For the latter, the ARG is viewed as a phylogeny of the extant samples. The reader must keep this general view of ARG in mind for the chapter.

In summary, the topology of the evolutionary history of a diploid population is a rather complicated network that represents the flow of the genetic material down to the extant units. See Fig. 1 for a visualization of the ARG that simulates the history of 210 samples or extant units (see the figure caption for details). The complexity of this combinatorial structure begs the following question: *Is it possible to identify a substructure that really matters to the extant units?* The problem addressed in this chapter is the extent of topological redundancies, if any, in such structures. This understanding of redundancy is useful both for reconstruction as well as simulation studies. While in the former it is possible to obtain an algorithm-independent bound on the recoverability of common history, in the latter it has the potential for producing simpler simulation systems. In any case the issue of redundancy of a model is never an irrelevant mathematical question to ask.

2. Background

The ideal population or Wright–Fisher model assumes some properties of the evolving population such as constant population size and nonoverlapping generations. While these conditions appear nonrealistic at first glance, the assumptions are reasonable for the

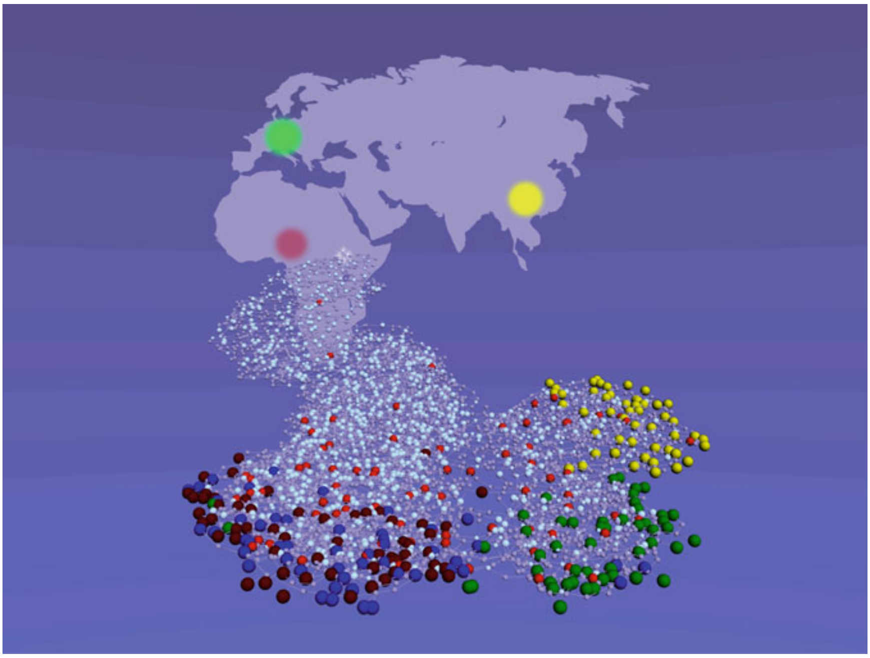

Fig. 1. The terminal (leaf) nodes are as follows: the 60 *brown nodes* represent African samples, the 50 *blue nodes* African-American samples, the 50 *yellow nodes* Asian samples and the 50 *green nodes* European samples. The internal *cyan* and *red nodes* are recombination nodes and *gray nodes* are coalescent nodes. The simulation was generated with COSI (2) and the visualization using Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/). The *red* recombination nodes are the ones reconstructed by the method in (1).

purposes of the study of the genetic variations at the population level. In fact, models with varying population size and/or overlapping generations can be reparameterized for an equivalent Wright–Fisher model (see texts such as ref. 3, 6). Yet another property of the evolving Wright–Fisher population is panmixia. Panmictic means that there is no substructuring of the population due to mating restrictions caused by mate selection, geography, or any other such factors. Thus the model assumes equal sex ratio and equal fecundity. Figure 2a shows the complete pedigree history of four ($K = 4$) samples with a population size of eight males and eight females ($N = 8$). The network structure is a random graph written as $G_{\mathrm{PG}}(K, N)$. An ARG, which tracks some fixed locus on all the K samples, is a subgraph of this complete pedigree history and an instance is shown in Fig. 2b. To mimic the genetic diversity patterns seen in worldwide human populations, it is important to also weave in other influencing factors such as different migration, (site) selection, and expansion models.

As discussed earlier, if the locus under study is always transmitted from a single parent, then the topology of the evolutionary history is a tree (i.e., no closed paths in the directed graph). The mitochondrial genome and nonrecombining Y chromosome satisfy this property. The former is always transmitted from the mother

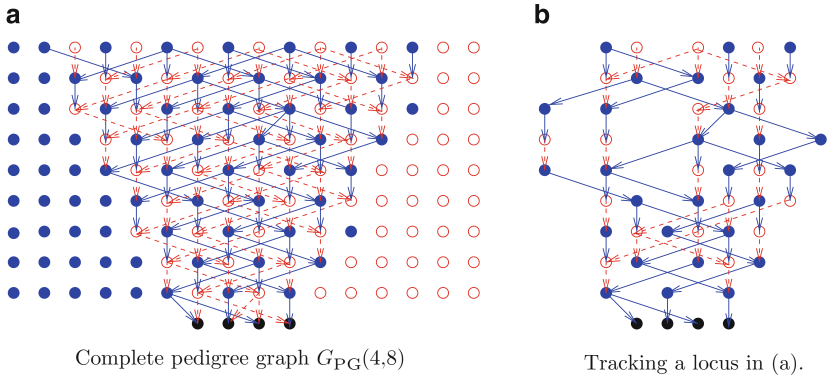

Fig. 2. (**a**) The first ten generations of the relevant part of the complete pedigree graph (G_{PG} (K, N) with $K = 4$ and $N = 8$). The solid (*blue*) dots represent one gender, say males and the hollow (*red*) dots represent the other gender (females). Each row is a generation with the direction on edges indicating the flow of the genetic material and the four extant units are at the bottom row, i.e., row 0. Under the Wright–Fisher population model, there are equal number of males and females in each row and the two distinct parents, one male and one female from the immediately preceding generation are randomly chosen. (**b**) Tracking a locus gives a subgraph of (**a**).

and the latter from the father. However, if the locus is on the autosome or even the X chromosome then the genetic material may be transmitted from two parents. This implies that the topology of the evolutionary history is no longer a tree, but a network (i.e., it may have closed paths in the directed graph). Thus, due to the occurrence of genetic exchange event, such as recombination, the common evolutionary history can no longer be captured by a tree. The network that captures both the genetic exchange event (such as recombinations) and events that do not exchange genetic material between parents (such as mutations) is the ARG. For simplicity of exposition we call the class of latter events as *nonexchange* events.

Notice that this important distinction in the topological characteristics arises simply from the basic locus-inheritance model, that is uniparental or biparental. The rest of the model characteristics define the depth (or age) distribution of the nodes. Thus, it is important to note the subtlety that an ARG is a random object and there are many (infinite) *instances* of the ARG. Usually, when we say that a topological property holds for the ARG, we mean that the property that holds for every instance of the ARG, i.e., the property holds with probability 1. Note that some may hold for a subset of instances (such as unboundedness).

Focusing on the topology of the ARG and its effect on the samples provides us with insights to identify vertices that "do not matter." Modeling these as missing nodes in the ARG leads to a core that preserves the essential characteristics. The random object ARG is defined by at least two parameters: K, the number of extant samples and $2N$, the population size at a generation. A Grand Most Recent Common Ancestor (GMRCA) plays an important role in restricting the zone of interest in the common evolutionary structure. A GMRCA is defined as a unit whose genetic material is ancestral to all the genetic materials in all the extant samples (6). Thus, while the relevant common evolutionary history of some $K > 1$ units is potentially unbounded, it is reasonable to bound this structure of interest with this single GMRCA. Thus *when a GMRCA exists, it is unique* and we say the ARG is *bounded*. When an ARG has no GMRCA, we call it *unbounded*.

The least common ancestor (LCA) of a set of vertices V in a graph is defined as a common ancestor of V with no other common ancestor of V on any path from the LCA to any vertex of V. A combinatorial treatment, based on random graphs, of the ARG is presented in (7). The directed graph representation is acyclic, a root is analogous to a GMRCA, and the leaf nodes to the extant samples. Though tantalizingly similar GMRCA and LCA do not define the same entity in an ARG. The edges (or nodes) of the ARG must be annotated with the genetic material it transmits. The absence of any annotation leads to the *ancestor without ancestry* paradox: It is possible for an individual with finite amount of genetic material to have an infinite number of unrelated (i.e., no genetic flow between any pair) ancestors. This paradox is averted by annotating the ARG (7).

3. A Combinatorial Definition of ARG

The random object ARG is usually parameterized by three essential parameters: K the number of extant samples, $2N$ the population size, and recombination rate r (see texts such as ref. 3 for a detailed description). The following theorem is paraphrased from (7):

Theorem 1. *Every ARG G on $K > 1$ extant samples is the topological union of some $M \geq 1$ trees (or forests).*

The alternative definition of an ARG suggested by this theorem is illustrated in Fig. 3. Here an ARG, defined on four (K) extant samples, is decomposed into three (M) trees. Note that M is the number of nonmixing or completely linked segments in the extant samples. In both the models, all the samples are of same length say l and additionally the length of each of the M segments is specified as $l_1, l_2, \ldots, l_M$ with $\sum_{i=1}^{M} l_i = l$, in the latter.

We describe the graph G (ARG) here. Although the figures do not show the direction of the edges to avoid clutter, the direction is toward the more recent generation (or the leaves). In other words, the leaf (extant) nodes have no outgoing edges and the root node has no incoming edges. The edges of the ARG are annotated with genetic events and these labels are displayed in the illustrations. See Fig. 4a for an example. An edge in G is defined to have multiple *strands*. In the illustrations, the multiple strands are shown as distinct colors, each color corresponding to one of the component

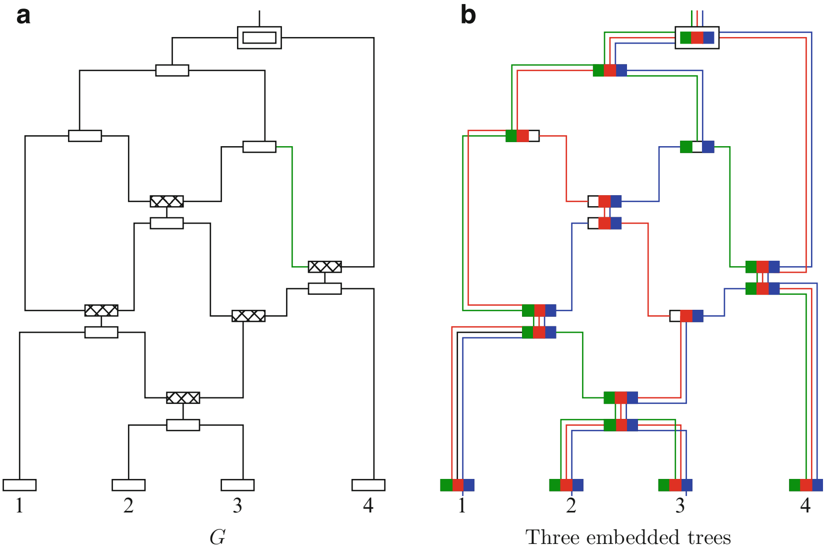

Fig. 3. Here $K = 4$ and the extant samples are numbered 1, 2, 3, and 4. The hatched nodes are the genetic exchange nodes. (**a**) The topology of an ARG, where the GMRCA is marked by an additional rectangle (*on top*). (**b**) A possible embedding of (**a**) by three trees (shown in *green, red*, and *blue*, respectively).

trees $1 \leq i \leq M$. Between any pair of vertices v_1 and v_2, no two strands can be of the same color. Thus, the number of multiple strands, corresponding to an edge, between a pair of vertices can be no more than M. An i-path from node v_1 to node v_2 is a path where all the edges in the path are on the component tree i.

The annotations on the edges play a critical role since it is these annotations that ultimately shape the units on the leaf nodes. In the chapter, samples refer to extant samples. The two kinds of genetic events represented in the graph are genetic (1) nonexchange and (2) exchange events. While the former is modeled by the genetic exchange nodes, the latter is modeled by labels on the edges. To keep this discussion simple, let the nonexchange genetic event correspond to single nucleotide polymorphisms (SNPs). The set of labels of edge v_1v_2 is written as $lbl(v_1v_2)$. Then $x_i \in lbl(v_1v_2)$ is a label on strand i of edge v_1v_2. For example in Fig. 4a, the labels on the green tree are the SNPs a, b, c, d. Also, the exact position of the SNP on the genome does not matter. However, in the ARG, a particular ordering of the M trees is assumed and hence the SNPs of each of the M trees respect this order (this is reflected in the sample definitions below where green is the leftmost segment and blue the rightmost). Each strand of an edge is labeled by a set of genetic events (SNPs), possibly empty. A node with multiple ascendants (parents) is called a *genetic-exchange* node. A node with multiple descendants (children) is a *coalescent* node. Note that a node can be both a coalescent as well as a genetic-exchange node. In the figure a genetic-exchange node is hatched.

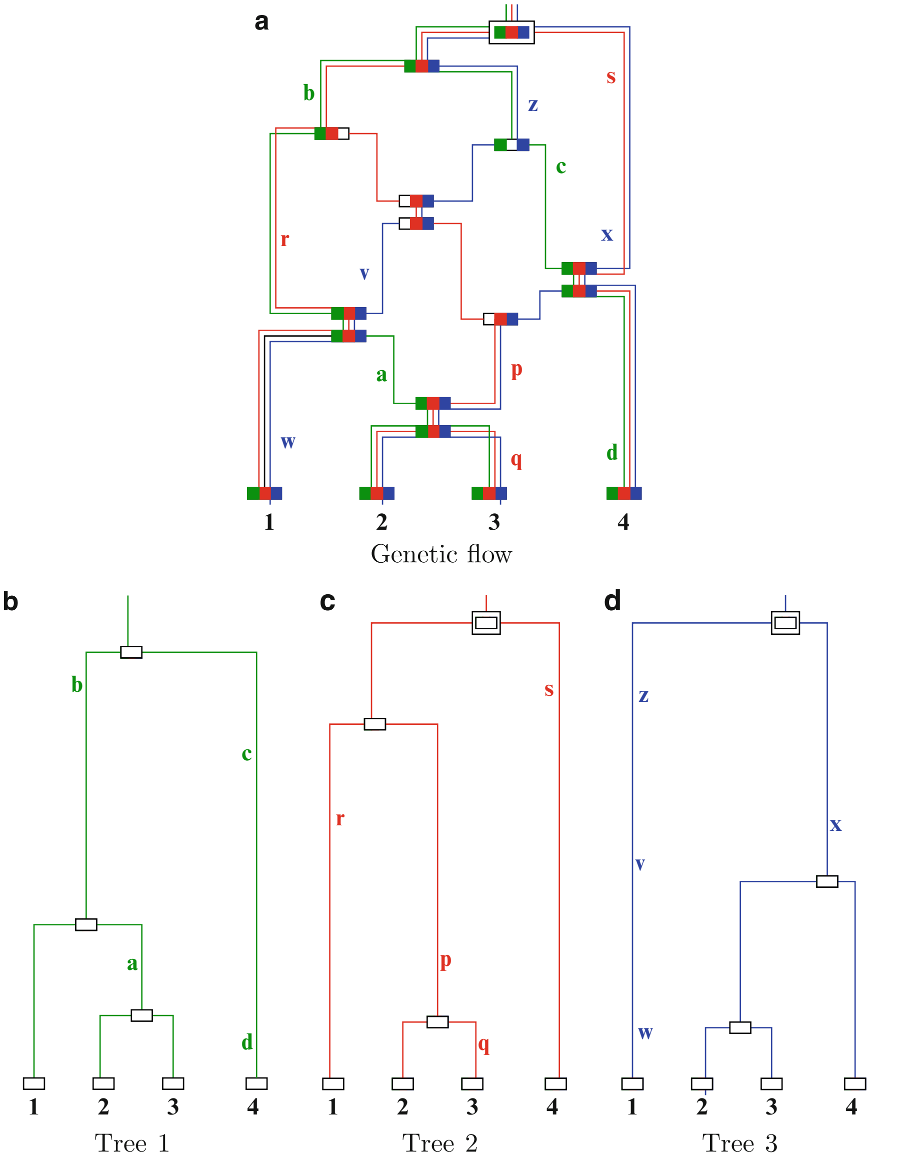

Fig. 4. (**a**) Genetic event labels on the edges. At each node the nonmixing segment corresponding to the embedded tree is shown in the same color as that of the tree. The three embedded trees are shown separately in (**b**), (**c**), and (**d**).

Next, we define the samples represented by the graph instance G of the ARG. This is denoted as $S(G)$ which is a set of K sequences which is also the number of leaf nodes in G. Each sequence is obtained simply by "flowing" the genetic event labels of tree i, $1 \leq i \leq M$, along paths of color i all the way down to the leaf (samples) units. In other words, for each extant unit u on G, let the corresponding sequence be $s(u)$ ($\in S(G)$). Each label is associated with a chromosomal position and its exact location on the sequence really does not matter in this framework. However, we

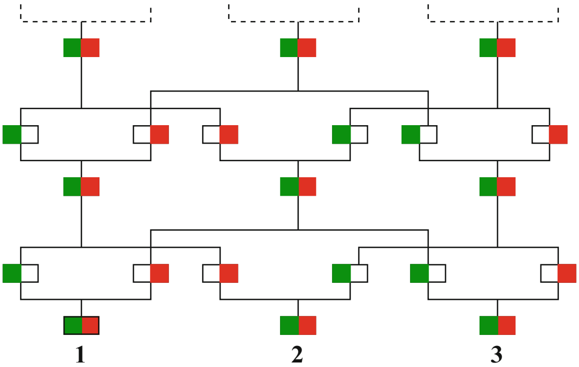

Fig. 5. Example of an unbounded ARG. Here $K = 3$ corresponding to the samples numbered 1, 2, and 3 and $M = 2$, for the two segments colored *red* and *green*. The pattern of vertices and edges can be repeated along the *dashed edges* to give an unbounded structure.

use the value of the label to define the sequence $s(u)$. Let $\Pi(s(u))$ denote the elements of $s(u)$. Then $\Pi(s(u)) = \bigcup_{i=1}^{M} \{x_i | x_i \in lbl(v_1 v_2)$ and there exists an i - path from v_2 to $u.\}$. Although the exact location does not matter, the labels of a strand (tree or color) i are adjacent on the chromosome sequence $s(u)$. Let s_1, s_2, s_3, and s_4 be the sequences corresponding to the extant units marked $1, 2, 3$, and 4, respectively in Fig. 4a. Assigning colors and a relative ordering to the strand labels, the aligned four samples are:

$$S(G) = \begin{array}{c} (s_1) \\ (s_2) \\ (s_3) \\ (s_4) \end{array} \left\{ \begin{array}{cccccccccccc} - & b & - & - & - & - & r & - & v & w & - & z, \\ a & b & - & - & p & - & - & - & - & - & x & -, \\ a & b & - & - & p & q & - & - & - & - & x & -, \\ - & - & c & d & - & - & - & s & - & - & x & - \end{array} \right\}. \tag{1}$$

The "—" here is to be interpreted as the ancestral allele.

To summarize,

1. An ARG G must satisfy the following
   - (a) (topology) Every node v in G must have multiple children or multiple parents (since chains are not informative).
   - (b) (annotations) The nonexchange genetic event label (say, SNP) corresponding to a position on the samples must transmit down to at least one extant sample.
2. Further, a nontrivial G must encode at least $M - 1$ genetic exchange events.

It is quite possible to have unbounded ARGs, i.e., ARGs with no GMRCA. Figure 5 shows such an example. See the "Exercise" for other families of unbounded structures on the Wright–Fisher population.

4. Redundancies in an ARG

How do we identify redundancies in the topology of an ARG? Studying the effect of the topology on the samples provides us with insights to identify vertices that "do not matter." Modeling these as missing nodes in the ARG leads to a core that preserves the essential characteristics.

To maintain biological relevance, a "missing" node is modeled by the following vertex removal operation. Note that in an ARG, each node has an implicit depth associated with it that reflects its age (in generations). An alternative view is that the edge length denotes the age. Note that in the following the age of the nodes does not change and the new edges get the edge length from the ages of the nodes they connect. Given G and a node v in G, $G \backslash \{v\}$ is obtained in the following steps. This is not the only possible definition of vertex removal, but it is a simple and natural one and is used in this chapter

1. For each child $v_{c,i}$ of v, that is in the embedded tree $1 \leq i \leq M$
   (a) (adding new edges) This child is connected by a new edge to $v_{p,i}$, a parent of v in i.
   (b) (annotating the new edges) The new edges between $v_{p,i}$ and $v_{c,i}$ are annotated as follows: for each strand i, the label of the new edge is the union of the labels on the i-path from $v_{p,i}$ to $v_{c,i}$. Next if a label x_i appears on multiple new outgoing edges of $v_{p,i}$, then it is removed from all but one of the outgoing edges. *(This is to avoid introducing parallel mutations, i.e., the same label appearing multiple times on the embedded tree i.)*
2. The node v with all the edges incident on it are removed from G.

4.1. Samples-Preserving Transformation

Two distinct ARGs G and G' are *samples preserving* if and only if $S(G) = S(G')$. When two instances are samples preserving, all the allele statistics, including allele frequencies, LD decay, and so on are identical in the two.

A node v of G is called *nonresolvable* if $S(G) = S(G \backslash \{v\})$. The intuition is that if removing the node v has no effect on the samples, then no algorithm can detect the node using only the samples. Node v is called *resolvable* if $S(G) \neq S(G \backslash \{v\})$. Again, the intuition is that some algorithm may be able to detect the node in this case.

4.2. Structure-Preserving Transformation

Next we identify the vertices in G that determine the topology (as well as the branch lengths) in the M embedded trees. Given G and G', if each of the M embedded trees in G and G' are identical in topology as well as branch lengths (in generations), then G' preserves the structure of G and vice versa.

Note that the embedded trees (also called marginal trees) are very important in an ARG and critical in defining the ARG: Not just the topology but also the branch lengths, which represent the time (in generations) to the next coalescent event. Then is it possible to characterize a node that can lead to structure-preserving transformation? A coalescent vertex in G is *t-coalescent* if and only if it is also a coalescent node in at least one of the M embedded trees. In fact the following is proved in (8).

Theorem 2. *If $G' \leftarrow G \backslash U$ and no t-coalescent vertex of G is in U, then G' is structure-preserving.*

In other words, if a set of coalescent nodes that are not t-coalescent are removed from G to obtain G', then G and G' are structure preserving. With this useful property, we are ready to zero-in on a core preserving structure.

4.3. Minimal Descriptor

We begin with the following theorem (8) that relates t-coalescent with resolvability.

Theorem 3. *A resolvable coalescent node v is also t-coalescent in G.*

The theorem shows that the vertices that ensure the invariance of the branch lengths of each embedded tree are also resolvable, leading to the following definitions.

1. An ARG G is a *minimal descriptor* if and only if every coalescent vertex, except the GMRCA, is t-coalescent.
2. An ARG G_{md} is a minimal descriptor of G if and only if (a) G_{md} is a minimal descriptor, (b) G_{md} preserves the structure of G, and (c) G and G_{md} are samples preserving, i.e., $S(G) = S(G_{\mathrm{md}})$ holds.

Given G, let U be the set of all coalescent vertices in G, other than the GMRCA, that is not t-coalescent. Let $G' \leftarrow G \backslash U$. By the definition of a minimal descriptor and the following statement, G' is a minimal descriptor.

If v_1 is a t-coalescent vertex in G and v_2 is not, then v_1 continues to be a t-coalescent vertex in $G \backslash \{v_2\}$. Further if V_1 is a set of t-coalescent vertices in G, and none of the vertices in V_2 is, then each $v \in V_1$ continues to be t-coalescent in $G \backslash V_2$.

The following gives a constructive description of a minimal descriptor. *Let G' be a minimal descriptor of G. Then G' is biologically and evolutionarily relevant as*

1. *(Structure preserving) the embedded (marginal) trees of G and G' are identical.*
2. *(Samples preserving) the allele statistics (including allele frequencies, LD decay) in the samples in both G and G' are identical.*

5. Properties of Minimal Descriptor

Although, a minimal descriptor of an ARG is not unique (see Subheading 8), it nevertheless has very interesting properties. Figure 6 shows an example of a minimal descriptor of an ARG.

1. *Boundedness*. It is quite surprising that even an unbounded ARG G always has a bounded minimal descriptor. It takes some mathematical ingenuity to prove this and the interested reader is directed to (8) for details. We just illustrate this through an example here in Fig. 7a.
2. *Overlap of genetic segments*. This is a local property of a node that can be potentially used in designing sampling algorithms. Let v be a coalescent node, except the GMRCA, in a minimal descriptor ARG with descendants $u_1, u_2, \ldots, u_l$, for some $l > 1$. Then for each descendant u_i of v there exists another descendant u_j of v overlapping with u_i, $1 \leq i \neq j \leq l$. Figure 7b shows an example. Note that it is adequate that the overlap is only pairwise.
3. *Small size*. The number of vertices in a minimal descriptor ARG is not just guaranteed to be finite (by 1 above) but is also quite small. Let n_c be the number of coalescent events, n_e be the number of genetic exchange events, and n_v be the number of

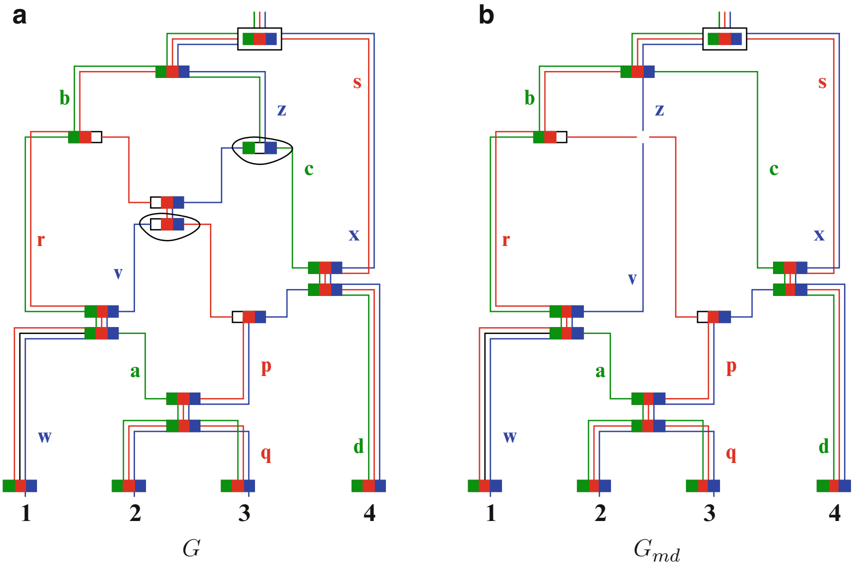

Fig. 6. Overall picture: (**a**) A generic ARG and all its genetic flow, thus defining the samples $S(G)$. The two marked nodes are not t-coalescent. (**b**) A minimal descriptor, G_{md} as it preserves the structure of G. Although the graphs are clearly topologically very different, yet they define exactly the same samples, i.e., $S(G) = S(G_{md})$ and G_{md} preserves the structure of G.

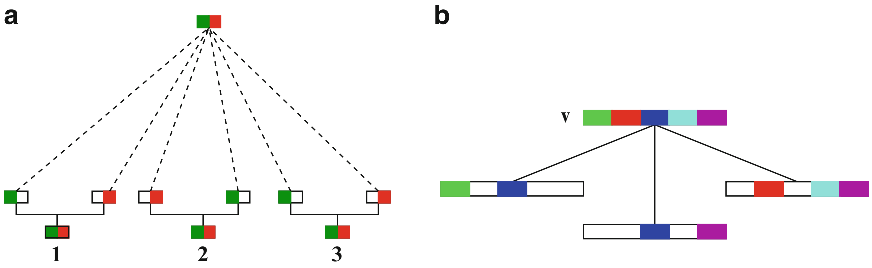

Fig. 7. (**a**) Bounded G_{md} of unbounded G of Fig. 5. (**b**) Pairwise overlap of genetic segments in the children of node v.

vertices, excluding the leaf nodes, in a nontrivial minimal descriptor ARG. Then

$$\begin{aligned} 1 \le n_c &\le M(K-1)+1, \\ 0 \le n_e &\le K(M-1)+M(K-1), \\ n_v &= \mathcal{O}(MK). \end{aligned}$$

This property is surprising, since most current simulators produce an extremely large number of internal nodes. It appears that most of them have no effect either on the marginal tree structures or on the samples. We end this discussion with this interesting observation.

6. Population Simulators

A modelless approach to simulations is to take an existing population sample S and perturbs it to obtain S' that has similar properties as S. However, here we discuss systems that explicitly model the population evolution evolving under the Wright–Fisher model (9). It is important to point out that literature abounds with population simulation systems and the list of simulators mentioned here is by no means complete. However, the attempt here is to classify them based on the underlying approaches. The simulation systems are aligned along two approaches: forward and backward. In the former the simulation of the events proceeds forward in time, that is from past to present. While this is a natural direction to proceed a trickier approach is to simulate backward in time that is from present to past. In principle, this is more economical in space and time. In both approaches an implicit phylogeny structure is constructed. We call the reduced version of this as the ARG in Fig. 8. An internal node in an ARG is either a coalescent node or a genetic exchange node but not neither. A mathematically interesting approach is to simulate the time to the next coalescent, or recombination, event without explicit simulation of every generation.

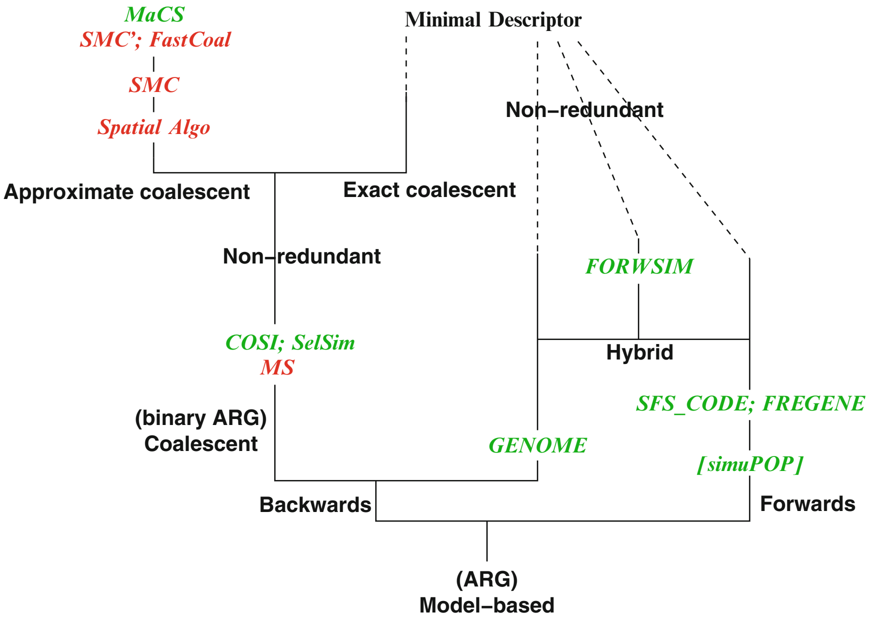

Fig. 8. A classification of the model-based (hence an associated ARG) population evolution systems based on their underlying architectures. The software systems are shown either in *red* or *green*. The systems in *green* additionally incorporate selection and/or demographics to produce genetic diversity patterns that somewhat reflect the current populations. *Bottom to top*: Backward and forward are the two basic schemes with hybrid as a combination of the two. Coalescent is a mathematically interesting backward scheme whose ARG topology characterizes it as a binary ARG. A set of simulators are listed here as approximate coalescent which are attempts at removing redundancies in the underlying binary ARG. The minimal descriptor, by its definition, is a nonredundant representation of the ARGs resulting from all the schemes (and additionally it is an exact coalescent model, hence the bifurcation in the coalescent "lineage" above).

The coalescent model captures this in the backward model. Figure 8 gives a classification of a few simulators along these lines.

The primary output for the simulators is the K sample (genetic) sequences, given the population size N along with other parameters. The primary genetic exchange event captured in the simulators is recombinations, although some simulators also incorporate gene exchange. Realistic worldwide human population requires the modeling of at least two more classes of parameters: (1) selection-related and (2) migration-related parameters. Due to the inherent complexity of the variations in the human population, the simulators generally handle population at the level of continents, that is, African, Asian, and European. Most of the programs do not make the ARG available. The authors of cosi made the internal ARG accessible to us (which has been visualized in Fig. 1).

6.1. Forward Simulators

Forward simulation is conceptually the simpler of the two approaches. An advantage of this approach is its easy adaptability to diverse evolutionary forces. simuPOP (10) is an individual-based forward simulation environment. The system also allows for interactive evolution of populations. For ease of use, many basic population genetics models are available through their "cookbooks." This is a suitable system for experimentations since the user can engineer complex evolutionary scenarios in the environment.

Next we discuss a few simulators that directly provide the population samples based on a set of input parameters. SFS_CODE (11) is a forward simulator that additionally handles effects of migration, demographics, and selection. The migration model is the general island model with complex demographic histories. FREGENE (12) additionally incorporates selection, recombination (crossovers and gene conversion), population size and structure, and migration.

6.2. Backward Simulators

In the software *GENOME* (13), the authors simulate the coalescent and recombination events at every generation proceeding backward in time. The standard coalescent model, however, simulates the time to the next event. However, GENOME models an evolutionary history, more general than the standard coalescent model. In the random graphs framework in (7), the *genetic exchange model* or *mixed subgraph* represented this more general model. In this chapter, to avoid confusion in terminologies, such a general model is simply called the generic ARG or just ARG. On the other hand, the standard coalescent model is called the *binary* ARG, for reasons discussed below.

FORWSIM (14) simulates the Wright–Fisher population of constant size under natural selection at multiple sites, moving forward in time. However, the authors describe this as a forward–-backward simulator, since they simulate only those chromosomes in the next generation that can potentially contribute to the future population. This handling of multiple generation in a single step is possible only by some backward insight. Hence in Fig. 8, this is classified as a hybrid scheme. Additionally, it also models self-fertilization, making it a possible candidate for plant populations.

The Standard Coalescent. Coalescent theory provides a continuous-time approximation for the history of a relative small sample of extant units from a large population. Under this framework, the genealogy of a sample of DNA sequences is modeled backward in time and mutations (neutral) are superposed on the structure to generate sequence polymorphism data. Hudson introduced *MS* the seminal implementation to sample sequences from a population evolving under the Wright–Fisher model. COSI (15) is an implementation of simulation with the addition of human population demographics to the coalescent model. In fact, the same parameters were used in the forward simulator FREGENE discussed above. SelSim (16) is yet another simulator based on the coalescent framework that incorporates natural selection. It is important to point out a subtlety here. Usually under the coalescent model, the coalescence is between exactly two lineages and multiple genetic events do not occur in the same generation in the common evolutionary history. These simplifications help in defining the model as an ordered sequence of events as well as in estimating the time from one event to the next. Thus in these simulators, every node has no more than two descendants and no more than two ascendants, hence is called the binary ARG.

Approximate Standard Coalescent. While the above methods generate events backward in time, an orthogonal approach, introduced in (17), samples the events along the sequence. This is called the *Spatial Algorithm* (SA) and one of its characteristic effects is that the density of recombination breakpoints increases as one moves along the sequence. Another (perhaps related) characteristic of SA is that the process is not Markovian. The *Sequentially Markov Coalescent* (18) introduces modifications to the process to make the structure Markovian. Based on this model, in *FastCoal* (19), the authors use an additional heuristic of retaining only a subset of local trees while moving along the sequence. MaCS (20) is an implementation including human population demographics. It turns out that all the models discussed here, including the Markovian structure, only approximate the standard coalescent model. While each model is defined algorithmically as a sequence of precise steps, yet the reason for this lack of exactness is not clear enough to provide algorithmic modifications to close or reduce the gap with the standard model. These simulators that address redundancies are labeled "approximate coalescent" in Fig. 8.

6.2.1. Minimal Descriptor

The minimal descriptor is a compact version of the ARG which is both samples preserving and structure preserving. It is a nonredundant structure that can be extracted from any ARG, no matter its underlying model. The model could be based on forward or backward simulations or even backward coalescent. Notice that any probability measure, such as the above, immediately induces (by push forward) a measure on the space of minimal descriptors. Thus when the ARG is binary coalescent, it models the underlying standard coalescent exactly. Figure 8 illustrates this generality of the minimal descriptor.

Assume that the "true" probability space of the ARGs is the one implicated by the Wright–Fisher model. In fact, the standard coalescence also does not exactly capture the Wright–Fisher for high enough recombination rate (see ref. 21). To address the issue of the true probability space, Parida (7) defines a natural measurable space over the combinatorial pedigree history structures and presents a sampling algorithm based on it.

Any method that directly samples the space of minimal descriptors, such as in a statistical sampling setting say, needs to (implicitly) incorporate an underlying probability space. For instance, incorporation of the standard coalescent primarily manifests itself as the problem of estimation of branch lengths in the structures.

7. Conclusion

Population evolution models are important to understand the differences and similarities in individual genomes, particularly due to the explosion of data in this area. While these faithfully model the genetic dynamics of the evolving population, their structure is usually very large involving tens of thousands of internal nodes for say a few hundred samples with a thousand SNPs each. The complexity of this combinatorial structure raises the question of redundancies in this structure. This chapter addressed this precise question and gave mathematical description of such a substructure. This is important not only for simulations and reconstruction purposes, but also opens the door for a comprehensive understanding of genetic dynamics that ultimately shape the chromosomes.

8. Exercises

1. Construct an instance of $G_{\mathrm{PG}}(4, 3)$ with no LCAs.

   What is the probability of an instance of $G_{\mathrm{PG}}(4, 3)$ having no LCAs?

   (*Hint*: see ref. 7 for the definition of a natural probability measure).

2. (a) What is the difference in topology of a pedigree history graph and ARG?

   (*Hint*: How many parents must a diploid have?)

   (b) When tracing a haploid, at most how many parents can the extant unit have? Why? Does this hold for a unit at every generation? (Hint: Fig. 9a.)

3. Is it possible to assign labels to the nodes of the ARGs in Fig. 9b, c, why?

4. Argue that the number of resolvable nodes decreases with depth of the nodes.

5. Argue that an ARG may have multiple minimal descriptors. (*Hint*: Fig. 10.)

Acknowledgments

I would like to thank Marc Pybus for generating the visualization of the ARG produced by COSI to show the world populations (Fig. 1). I am grateful to the anonymous referees whose comments have substantially improved the exposition.

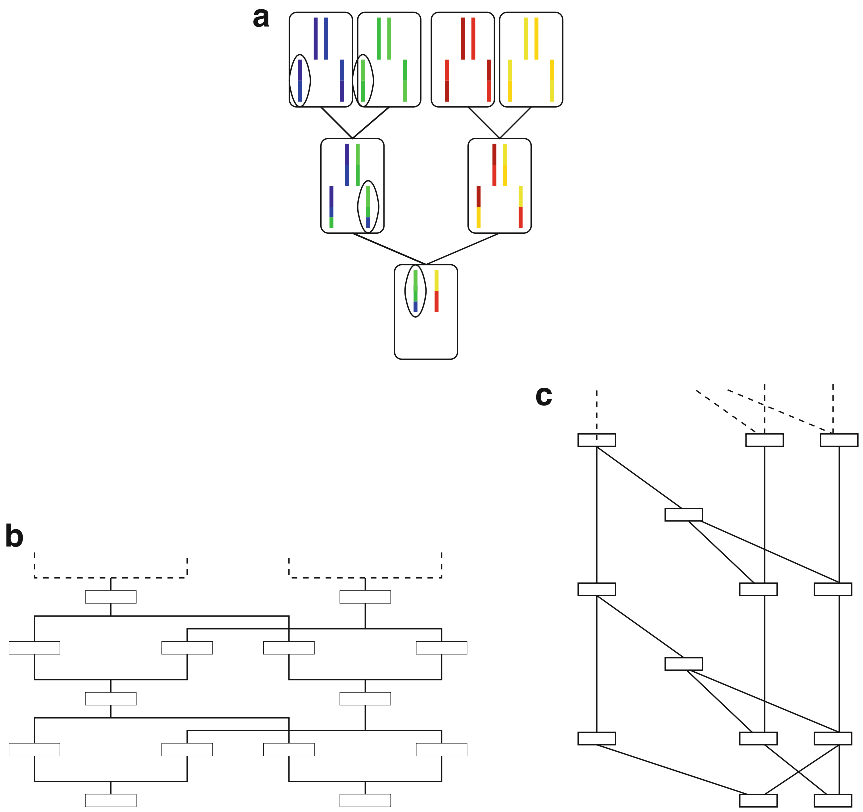

Fig. 9. (**a**) Tracking haploids in diploids. (**b**) and (**c**) The pattern of connectivity is repeated in both to produce infinite graphs.

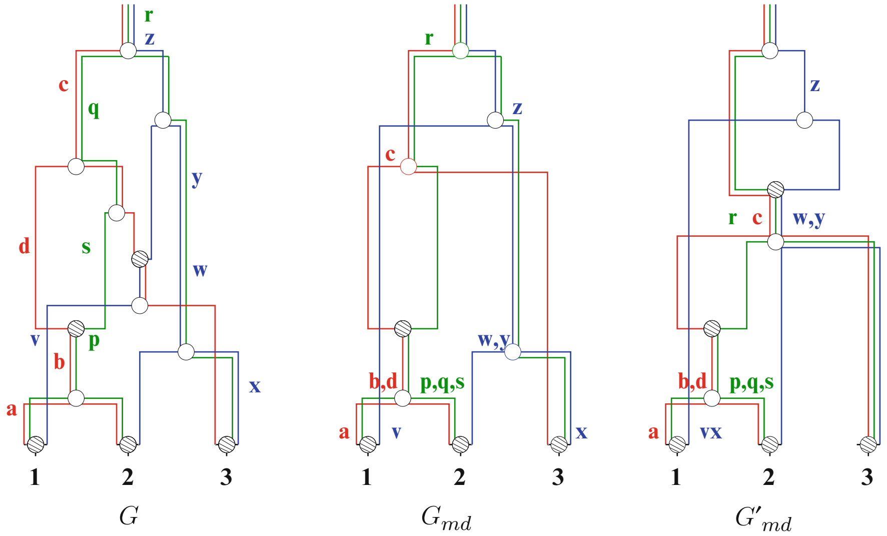

Fig. 10. G_{md} and G'_{md} are minimal descriptors of G.

References

1. R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, April 1983.
2. R. C. Griffiths and P. Marjoram. An ancestral recombinations graph. *Progress in Population Genetics and Human Evolution (P Donnelly and S Tavare Eds) IMA vols in Mathematics and its Applications*, 87:257–270, 1997.
3. Jotun Hein, Mikkel H. Schierup, and Carsten Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford Press, 2005.
4. Laxmi Parida, Marta Melé, Francesc Calafell, Jaume Bertranpetit, and Genographic Consortium. Estimating the ancestral recombinations graph (ARG) as compatible networks of SNP patterns. *Journal of Computational Biology*, 15 (9):1–22, 2008.
5. Marta Mele, Asif Javed, marc Pybus,, Francesc Calafell, Laxmi Parida, Jaume Bertranpetit, and Genographic Consortium.
6. M.A. Jobling, M. Hurles, and C. Tyler-Smith. *Human Evolutionary Genetics: Origins, Peoples and Disease.* Mathematical and Computaional Biology Series. Garland Publishing, 2004.
7. Laxmi Parida. Ancestral Recombinations Graph: A Reconstructability Perspective using Random-Graphs Framework. *to appear in Journal of Computational Biology*, 2010.
8. Laxmi Parida, Pier Palamara, and Asif Javed. A minimal descriptor of an ancestral recombinations graph. *BMC Bioinformatics*, 12(Suppl 1): S6, 2011. http://www.biomedcentral.com/1471-2105/12/S1/S6.
9. R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338, Feb 2002.
10. Bo Peng* and Marek Kimmel. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 21:3686–3687, 2005.
11. RD. Hernandez. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, 24:2786–2787, 2008.
12. Marc Chadeau-Hyam, Clive J Hoggart, Paul F O'Reilly, John C Whittaker, Maria De Iorio, and David J Balding. Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics*, 9, 2008. DOI = doi:10.1186/1471-2105-9-364.
13. Liming Liang, Sebastian Zllner, and Goncalo R. Abecasis. Genome: a rapid coalescent-based whole genome simulator. *Bioinformatics*, 23 (12):15651567, 2007.
14. Badri Padhukasahasram and Paul Marjoram and Jeffrey D. Wall and Carlos D. Bustamante and Magnus Nordborg. xploring Population Genetic Models With Recombination Using Efficient Forward-Time Simulations. *Genetics*, 178(4):24172427, 2008.
15. S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, 15:1576–1583, Nov 2005.
16. Spencer CC and Coop G. SelSim: a program to simulate population genetic data with natural selection and recombination. *Bioinformatics*, 12:20:3673–5, 2004.
17. Carsten Wiuf and Jotun Hein. Recombination as a point process along sequences. *Theoretical Population Biology*, 55:248–259, 1999.
18. Gilean McVean and Niall Cardin. Approximating the coalescent with recombination. *Phil. Trans. R. Soc. B*, 360:1387–1393, Sep 2005.
19. P. Marjoram and J. D. Wall. Fast coalescent simulation. *BMC Genetics*, 7(16), Jan 2006.
20. G. K. Chen, P. Marjoram, and J. D. Wall. Fast and flexible simulation of DNA sequence data. *Genome Res.*, 19:136–142, Jan 2009.
21. Joanna L. Davies, Frantiek Simank, Rune Lyngs, Thomas Mailund, and Jotun Hein. On recombination-induced multiple and simultaneous coalescent events. *Genetics*, 177:2151–2160, December 2007.