Chapter 12 – Gene Genealogies

Noah A. Rosenberg

Program in Molecular and Computational Biology. University of Southern California, Los Angeles, California 90089-1113 USA. E-mail: noahr@usc.edu. Phone: 213-740-2416. Fax: 213-740-2437.

January 2, 2005

Introduction

Genetic variation at a locus among extant individuals can be viewed as the result of mutations on a scaffold of genetic relationships – a *gene genealogy*. Because patterns of genetic variation contain much information about phenomena such as hybridization, migration, species divergence, and changes in population size, an understanding of gene genealogies is helpful for the application of genetic variation to inference about evolutionary processes. As we will see, gene genealogies, which underlie numerous statistical methods for population genetic analysis, are useful in diverse areas of genetics and evolutionary biology, ranging from phylogenetics to genetic mapping.

The basic nature of the inheritance of genetic material is familiar: copies of corresponding stretches of the genome in different individuals are passed through a series of generations from some piece of DNA in a common ancestor of the individuals. The mutations that occur in transmission leave a pattern of similarities and differences in extant individuals that, albeit imperfectly, records the genealogical history in their DNA sequences. All the processes that affect this history – for example, the size of the population to which the individuals belong, which influences the length of time to the common ancestor – affect the outcome in the DNA sequences, the data available to us today. Thus, to learn about how the population has evolved, we need to know how evolutionary processes affect genealogies, and in turn, how genealogies affect genetic data.

In this chapter, I introduce gene genealogies, which describe relationships among copies of a locus in different individuals, through a discussion of their link to *pedigrees*, the structures that describe relationships among the individuals themselves. Two initial questions that might be asked about gene genealogies are:

(1) What schemes can be used to categorize gene genealogies, and what are the categories?

(2) What attributes do we expect gene genealogies to have in specific evolutionary scenarios?

After considering these issues – classification of genealogies and properties of random genealogies – I discuss a variety of examples that illustrate the use of gene genealogies for interpreting patterns of genetic variation.

Concepts

Pedigrees and Gene Genealogies

For haploid organisms, relationships of individuals and those of their genomes are equivalent: when a cell divides, the genomes of the offspring descend directly from the parental genome (but see Box 1). For diploids, however, the way in which genomes pass from parents to offspring is more complex. To understand the relationships between diploid genomes, rules that characterize the transmission process of genomes from parents to offspring – Mendel's laws of inheritance – can be used.

Consider an individual, and choose one of its parents. The law of segregation states that for any (autosomal) locus in the genome, (1) the individual has a copy of the locus from the chosen parent, and (2) with probability 1/2 this copy is inherited from the parent's maternal copy, and with probability 1/2 it is inherited from the parent's paternal copy. For two loci, the law of independent assortment states that whether the copy inherited at the first locus derives from the chosen parent's maternal or paternal copy does not depend on which grandparent produced the copy at the second locus. Genetic linkage between some pairs of loci produces exceptions to this rule; in these cases, however, modifications can be made to accommodate dependence between loci.

Suppose we are given a set of individuals *S*, whose biological relationships are represented by a pedigree (Figure 1i). Consider a locus randomly chosen from the genomes of the individuals. If we use the law of segregation to trace copies of the locus through the pedigree, starting with the set *S*, it is likely that we will eventually reach a single copy from which all copies in *S* descend (Figure 1iii).[a] All individuals in the figure are biologically ancestral to the individuals in *S* – that is, ancestors in terms of the pedigree. However, only a small fraction of the individuals in the pedigree, by being in lines of descent to *S* from the most recent common ancestor of the copies of the locus in *S*, are *genetically* ancestral at the locus. These genetic ancestors are the only individuals that affect the genotypic state at the locus for individuals in *S*. When we restrict our attention to these ancestors, we obtain the *gene genealogy* for the individuals at the locus.

Using the law of independent assortment, the grandparent from whom the copy from the chosen parent descends at one locus is independent of the one from whom the corresponding copy descends at a second locus. Applying this rule as we trace through a given pedigree, gene genealogies of two unlinked loci are independent. Because most diploid genomes have many independent loci, and thus, many independent gene genealogies, for any set of individuals, many paths are followed by at least one locus. Consequently, a pedigree of a set of individuals can be viewed as describing their "average" gene genealogy: proceeding through a pedigree, each path has the same probability. On average, all paths of a given length (that is, of a fixed number of generations) are taken by equal numbers of loci.

Examples considered by Wollenberg & Avise (1998), Derrida *et al.* (2000), and Rohde *et al.* (2004) make the relationship between pedigrees and gene genealogies apparent. The time until all humans share a common ancestor along the male or female line – that is, the time until the genetic ancestor for all human Y-chromosomes or mitochondrial genomes – has been estimated

[a] The exception in which a single copy is not necessarily reached is if life originated multiple times and the copies trace back to more than one of the original genomes.

at tens to hundreds of thousands of years. However, the most recent common ancestor (MRCA) in terms of the pedigree – the most recent individual to be part of the pedigree of all living humans – might have been surprisingly more recent, perhaps only 2,000-7,000 years ago (Rohde *et al.*, 2004). In other words, across all loci in the genome, the common ancestor for the gene genealogy whose MRCA is smallest may have lived in historical times.[b]

Terminology

This chapter uses the following definitions, which are generally standard, except where noted. The tips of gene genealogies represent *sampled lineages* (Figure 2). In general, each line that connects a descendant to an ancestor is a *lineage*. Nodes, which represent the joining of lineages in common ancestors as time proceeds backwards from the present, are *coalescences* or *coalescence events*. Lengths of time that separate coalescences from each other or from sampled lineages are *branch lengths*. A branch that separates two coalescences is *internal*; one that separates a sampled lineage from a coalescence is *external*. A coalescence at which two external branches join is a *cherry*. The *time to the most recent common ancestor* (T_{MRCA}) for a set of sampled lineages is the length of time from the present until the lineages first reach a common ancestor, their *most recent common ancestor* (MRCA). The T_{MRCA} for a genealogy is often called the *coalescence time*, although *coalescence times* can also refer to lengths of time between successive coalescences. The *root* node represents the MRCA for all sampled lineages in a genealogy; the two branches connected to the root are *basal*.

For a set of sampled lineages, a *locus* is a unit of DNA, ranging in size from a single base pair to a whole chromosome, in which no recombination has occurred in the genetic ancestors of the lineages since the time of their MRCA. In scenarios in which lineages derive from multiple populations, it often does not matter whether the populations are from the same species. Thus, except where otherwise specified, *species* is used to refer to the population of individuals who belong to a "species," and is sometimes interchangeable with *population*.

A *genealogy* or *gene genealogy* for n sampled lineages is a tree specified by the sequence of coalescences that reduce the n lineages to a MRCA, along with the *coalescence times* that separate these events. Two genealogies are *identical* if and only if they have the same sequence of coalescence events and the same coalescence times. A *subgenealogy* containing k of the n lineages includes the MRCA of these k lineages together with all parts of the genealogy that descend from this MRCA. Although it is possible to consider genealogies in which coalescences involve more than two lineages, it is assumed in this chapter that exactly two lineages join in each coalescence.

The major features of a genealogy can be captured in quantities that summarize its shape and size (Table 1). These quantities fall into three categories: (1) those that depend only on which lineages participate in coalescences, without regard to when coalescences occur; (2) those that

[b] Technically, there is no guarantee that any living person contains DNA descended from the pedigree MRCA studied by Rohde *et al.* (2004), as such segments of DNA may have disappeared over time through recombination. However, if the genome had infinitely many possible points at which recombination could occur, and if recombination only happened at each point at most once in evolutionary history, the pedigree MRCA would be the MRCA of the gene genealogy whose MRCA is smallest across all loci.

depend only on the coalescence times, without regard to which lineages participate in coalescences; (3) those that depend on both the lineages involved in coalescences and on the coalescence times.

Classification of Genealogies

We frequently have occasion to compare two or more genealogies. For example, to search for signatures of events with genome-wide effects, such as population splits, we can compare genealogies for different loci in the same set of individuals. To determine if a particular sample is suitably representative of a population, we can compare genealogies for the same locus in several samples.

We may be interested in whether or not two genealogies are identical; because identity of genealogies is rare, however, the equivalence or nonequivalence of attributes of the shapes of two genealogies – such as their *labeled topologies* – is more often of interest. Thus, it is useful to consider various ways in which shapes of genealogies can be classified; for convenience, each of several classification schemes is denoted here by a different letter.

Labeled Histories and Labeled Topologies. The *labeled history* of a genealogy is its sequence of coalescence events (Figure 3). Two genealogies of n lineages have the same *labeled history*, or are *H-equivalent*, if they have the same coalescences in the same temporal order. The number of possible labeled histories for genealogies of n lineages is $H_n=n!(n\text{-}1)!/2^{n-1}$ (Steel & McKenzie, 2001). Each genealogy of n lineages has one of H_n possible labeled histories, and each labeled history is the labeled history of some genealogy.

The genealogies in Figures 3i and 3ii have the same coalescence events, but in different sequences; therefore, they have different labeled histories. However, there is a sense in which these two genealogies are equivalent. The *labeled topology* of a genealogy is its unordered list of coalescence events.[c] Two genealogies of n lineages have the same *labeled topology*, or are *T-equivalent*, if they have the same coalescences, but not necessarily in the same order. The number of possible labeled topologies for genealogies of n lineages is $I_n=(2n\text{-}3)!/[2^{n-2}(n\text{-}2)!]$ (Felsenstein, 2004, table 3.1). Each genealogy of n lineages has one of I_n possible labeled topologies, and each labeled topology is the labeled topology of some genealogy.

Monophyly, Paraphyly, and Polyphyly. For genealogies whose sampled lineages derive from two species (or populations), *(A,B)*, we may be interested in how the lineages from the two species are interleaved in the genealogy. For each species, the sampled lineages from that species have a *monophyly status*: they are either *monophyletic* – that is, they comprise all the sampled descendants of their MRCA – or they are *not monophyletic*. Lack of monophyly requires that lineages of the other species be descendants of this MRCA. A genealogy of lineages from two species can be classified into one of four categories (Figure 4):

C1. *Monophyly of A and B*, or *reciprocal monophyly*. The lineages of each species are separately monophyletic.

[c] It is also possible to consider the *unlabeled* topology (Felsenstein, 2004, p. 29) and *unlabeled* history (Tajima, 1983, appendix 1) of a genealogy.

C2. *Paraphyly of B with respect to A* . The lineages of species *A* are monophyletic, and the lineages of species *B* are not monophyletic.

C3. *Paraphyly of A with respect to B.* The lineages of species *B* are monophyletic, and the lineages of species *A* are not monophyletic.

C4. *Polyphyly of A and B.* Neither the lineages of species *A* nor the lineages of species *B* are monophyletic.

Two genealogies of lineages from two species will be said to have the same *phyletic status* here if they classify into the same one of these four categories.

Suppose now that sampled lineages derive from m species ($m \geq 2$). For each species, the lineages of that species are either monophyletic or not monophyletic. The ordered list of m monophyly statuses for the species is the *M-type* of the genealogy. Two genealogies of lineages from two or more species are *M-equivalent* if and only if they have the same M-type. Each genealogy of lineages from m species has one of 2^m possible M-types.

For each pair of species, the phyletic status of the lineages from the two species can potentially be either C1, C2, C3, or C4. The ordered list of $\binom{m}{2}$ phyletic statuses for the m species is the *P-type* of the genealogy. Two genealogies of lineages from two or more species are *P-equivalent* if and only if they have the same P-type. Note that for m=2, P-equivalence has the same meaning as M-equivalence. For m>2, however, each M-type is the M-type of some genealogy, but many of the $4^{\binom{m}{2}} = 2^{m(m-1)}$ possible P-types cannot be the P-type of any genealogy. For example, no genealogy for three species – *A*, *B*, and *C* – can have pairs (*A*,*B)* and (*A*,*C*) in category C2 while (*B*,*C*) is in C1.

Collapsed Genealogies. For $m \geq 2$, the phylogeny of m species – the genealogy of the species – has one of H_m possible labeled histories, and one of I_m labeled topologies. To ease comparison between gene genealogies and species phylogenies, it is convenient to classify genealogies of lineages from m species with the same classes as those used for the species phylogeny itself.

The collapsing algorithm in Rosenberg (2002) gives a procedure for mapping a genealogy of n lineages from m species ($n \geq m$) onto the set of H_m labeled histories or to the set of I_m labeled topologies. This algorithm maps a gene genealogy from many species onto a *collapsed genealogy* obtained by considering only the most recent interspecific coalescence for each species (Figure 5). Taking into account the order of these coalescences, the genealogy is mapped to its *collapsed labeled history* or *C-type*. Considering the coalescences but ignoring their order, the genealogy is mapped to its *collapsed labeled topology* or *D-type*. Two genealogies of lineages from two or more species are *C-equivalent* if and only if they have the same collapsed labeled histories, and *D-equivalent* if and only if they have the same collapsed labeled topologies. For m=3, because each labeled topology is consistent with only one labeled history, D-equivalence has the same meaning as C-equivalence. Each of the H_m labeled histories for m

lineages can be the collapsed labeled history for some genealogy of lineages from *m* species; similarly, each of the I_m labeled topologies for *m* lineages can be the collapsed labeled topology for some genealogy.

Random Genealogies

For a given collection of assumptions about the evolutionary process in a set of species – a model – it is of interest to know the probability distribution for a random genealogy, or the genealogy of a random sample of lineages. Such a model can be used to predict patterns of genetic variation for a randomly chosen locus under a specific set of conditions. Although we would like to make predictions under any model, much can be learned using a relatively simple model with one population.

The Coalescent Distribution

Consider a random sample of *n* lineages from a haploid population of constant size *N*, with $N>>n$. In each of a series of discrete generations, every lineage chooses a random parent from the previous generation. Under these assumptions, the same as those of the frequently-used *Wright-Fisher model* (Ewens, 2004), the probability distribution of the genealogy of *n* random lineages is closely approximated by the *coalescent distribution*, variously termed the *coalescent*, *n-coalescent*, *neutral* or *standard coalescent*, or *Kingman's coalescent* (Kingman, 1982; Hudson, 1983; Tajima, 1983; Nordborg, 2001).

Recall that a genealogy consists of two components: its sequence of coalescence events and its set of coalescence times. Under the coalescent, the coalescence times have exponential distributions, so that the time until *n* lineages reduce to *n*-1 has exponential distribution with mean $2/[n(n\text{-}1)]$ units of *N* generations. The sequence of coalescence events has a uniform distribution over the set of labeled histories: at any point in time, each pair of lineages has the same probability of being the next pair to experience a coalescence. This uniform distribution, the *Yule distribution* (Aldous, 2001), assigns probability $1/H_n$ to each labeled history. Note that under the coalescent, the probability distribution of the labeled topology of a random genealogy is not uniform: the probability that a random genealogy has labeled topology *t* equals $(2^{n-1}/n!)\prod_{i=3}^{n}(i-1)^{-d_i(t)}$, where $d_i(t)$ is the number of coalescences in the labeled topology from which exactly *i* sampled lineages descend (Brown, 1994; Steel & McKenzie, 2001). Table 1 lists additional properties of genealogies under the coalescent.

The utility of the coalescent derives from the fact that it describes the distribution of the genealogy of *n* lineages in diverse evolutionary models besides the Wright-Fisher model, such as scenarios with age structure, horizontal DNA transfer (Box 1), or separate sexes (Möhle, 2000; Nordborg & Krone, 2002). In each of these models, a parameter termed the *coalescence effective size*, or N_e, is required to transform the model into one for which the coalescent applies. In other words, for a given model, if it has a coalescence effective size, the probability distribution of a random genealogy under the model is obtained from the coalescent, substituting N_e for *N*. One useful case for which the coalescent distribution applies is that of diploidy: a diploid constant-sized population with $N/2$ males and $N/2$ females has coalescence effective size $2N$ (Nordborg, 2001).

Many models, however, including some that include time-varying population size, do not have coalescence effective sizes. That is, for every value of N, the distributions of genealogies under these models differ from the coalescent distribution for population size N. Despite the lack of a coalescence effective size, the labeled history of the genealogy under such models can still have the Yule distribution. For example, although changes in population size affect coalescence times, they do not alter the fact that all pairs of lineages are equally likely to coalesce.

Several strategies are available for determining the properties of models whose genealogies do not follow the coalescent distribution. It is sometimes possible to directly calculate or at least approximate the distributions of random genealogies. Alternatively, it may be possible to obtain the distributions from modified versions of the coalescent. However, the most general strategy for studying genealogies under complex models is simulation from sampled lineages back in time to their MRCA (Hudson, 1990). In fact, because backward simulations can often be performed rapidly, they are useful even when the coalescent distribution does apply. Their efficiency results from the fact that simulation from a small sample backwards in time to a MRCA requires that only a small number of random variables be generated. The forward approach, which entails simulation of whole populations for a long enough period of time to erase the effects of initial conditions, followed by extraction of genealogies of random sets of lineages, wastes considerable effort simulating lineages that are not ancestral to samples.

The coalescent distribution of genealogies is often taken as a "null" distribution, as it represents the behavior of a population under simple assumptions. To understand the impact of complex phenomena on genealogies, distributions of genealogies under various models can be compared to the coalescent qualitatively or quantitatively, using properties such as T_n or L_n from Table 1 (Donnelly, 1996; Uyenoyama, 1997). For example, it is often noted that genealogies from exponentially growing populations are more "star-like" than are those from constant-sized populations (Slatkin & Hudson, 1991). In quantitative terms, this observation reflects the fact that random genealogies under exponential growth have elevated values of ratios such as P_n/T_n and $L_n/(nT_n)$ (Rosenberg & Hirsh, 2003).

Population Structure

In models with subdivision of populations, by geography or by other variables, the coalescence sequence of a random genealogy does not follow the Yule distribution, as pairs of lineages from the same group are more likely to coalesce than are pairs from different groups. The distribution of the labeled history or labeled topology of a random genealogy may be of less interest, however, than such distributions as that of the M-type or the collapsed labeled topology. Under a given model, these distributions, only applicable for multiple populations (or species), can help in articulating the predictions that the model makes about the processes that it considers.

Two Populations. For two populations, the probability distribution of the phyletic status of a random genealogy is of interest. Consider the *island model*: two haploid populations of size N with a fraction m of the lineages in each population switching populations each generation. With samples of size 2 from each population, for small Nm, the probabilities of scenarios C1, C2, C3, and C4 (Figure 4) approximately equal 1-14Nm/3, 5Nm/3, 5Nm/3, and 4Nm/3, respectively

(Takahata & Slatkin, 1990). From these values, it is observed that as the migration rate decreases to zero, the probability of reciprocal monophyly increases to one.

The distribution of phyletic status can also be obtained (for any sample sizes) in the *two-population divergence model*, in which an ancestral population splits instantaneously into two descendant populations each of size N (Rosenberg, 2003), or (for small sample sizes) in a divergence model that allows descendant populations to be subdivided after divergence (Wakeley, 2000). In these cases, it is observed that at divergence, polyphyly is the most likely phyletic status, and as time progresses, reciprocal monophyly becomes most likely. In the two-population divergence model, reciprocal monophyly has probability 0.99 by $6N$ generations after divergence.

Although much is known about random genealogies under the island model (Takahata & Slatkin, 1990; Nath & Griffiths, 1993), the two-population divergence model (Takahata & Nei, 1985; Rosenberg, 2003), and other two-population models (Wakeley, 2000; Teshima & Tajima, 2002), the distributions of attributes of genealogies (Table 1) are more difficult to compute with two populations than with one. However, as in one-population models, backward simulation has proven useful for exploring these distributions in two-population scenarios (Hudson, 1990; Rosenberg & Feldman, 2002).

Three or More Populations. The probability distributions of C- or D-types for random genealogies, which are trivial for one or two populations, become interesting with three or more populations. Perhaps the most useful of these distributions is that of the collapsed labeled topology of a random genealogy. Suppose three populations descend from an ancestral population that split into two groups, one of which subsequently bifurcated again. Suppose also that the time between the bifurcations is t generations and that the population size between bifurcation events is constant at N haploid individuals. If one lineage is sampled from each population, the probability that the (collapsed) labeled topology of a random genealogy is the same as the labeled topology of the population phylogeny is $1-(2/3)e^{-t/N}$ (Pamilo & Nei, 1988). Each of the other two possible collapsed labeled topologies has probability $(1/3)e^{-t/N}$, so that as t increases to infinity, the probability of concordance of the labeled topologies of the gene genealogy and the phylogeny nears one. A similar calculation for arbitrary sample sizes shows that the probability of topological concordance increases more quickly with t if larger samples are used (Rosenberg, 2002).

As is true for the two-population case, probability distributions of complex aspects of genealogies in multi-population models remain elusive, except by simulation. However, some progress has been made in various scenarios (Pamilo & Nei, 1988; Wakeley, 1998; Wilkinson-Herbots, 1998).

Case Studies

Uses of Genealogies

The usefulness of gene genealogies arises from the fact that genetic variation can be viewed as the result of mutations occurring along the branches of genealogies (Figure 6). Thus, patterns of

genetic variation are affected by the attributes of the genealogies on which mutations have occurred. However, these genealogies are generally unknown. To address this issue, one of two main strategies can be adopted (Rosenberg & Nordborg, 2002; Hey & Machado, 2003): first, the genealogy can be estimated from the data, and the analysis based on the estimated genealogy. Alternatively, the coalescent and its extensions can be used to sample genealogies from a set of random genealogies consistent with the data, and the analysis averaged over these genealogies. The former approach has the limitation that basing the analysis on the estimated genealogy ignores uncertainty in the estimate. The latter approach, while statistically rigorous, can potentially require intensive computations, so that sometimes, it can only be applied approximately.

The fact that genealogies underlie patterns of variation has been useful for developing interpretations of particular observations in genetic data. Allowing for mutations, the coalescent model has been used to make various predictions about the distribution of allele frequencies expected across sites in a set of DNA sequences (Tajima, 1989; Fu & Li, 1993). For example, the comparatively "star-like" nature of genealogies in populations undergoing expansions in size, compared to those from constant-sized populations, is reflected in an excess number of mutations along external branches. The D and D^* statistics of Fu & Li (1993), which are computed from DNA sequences sampled from a population, compare numbers of mutations along internal and external branches. Negative values of these statistics, reflecting an excess of external mutations, indicate that growth in size may have been important in the history of the population.

A need to use gene genealogies arises in many contexts in diverse organisms (Avise, 2000; Donnelly & Tavaré, 1997; Li & Fu, 1999; Knowles & Maddison, 2002; Slatkin & Veuille, 2002). Several examples are discussed below.

Molecular Phylogenetics

The inference of species genealogies (or phylogenies) from the distribution across species of a genetic character typically relies on the premise that if one lineage is sampled per species, then the genealogy for the character is identical to that of the species. If species are distantly related, this premise generally holds for the coalescence sequence of the gene genealogy, although the coalescence times of the gene genealogy are often considerably larger than those of the species genealogy (Figure 5). In this case, the problem of phylogenetic inference is to recover an underlying genealogy that has been obscured by the stochastic occurrence of mutations along its branches (Figure 6).

As we have seen, however, especially for closely related species, this basic premise may fail to hold. First, the lineages of one or more of the species may not be monophyletic, so that the choice of lineage affects the shape of the genealogy. Second, the gene genealogy often may have a different labeled topology from that of the species genealogy, so that the choice of locus affects the shape of the genealogy. When these scenarios have nontrivial probabilities, careful consideration of gene genealogies is important to phylogenetic inference. Generally, the solutions to the nonmonophyly and discordance problems involve use of many lineages per species and many independent genealogies, respectively.

A study by Wilson *et al*. (2003) addresses the problem of nonmonophyly of lineages for a set of 13 human populations. Assuming that the evolution of the populations followed a bifurcating tree, Wilson *et al*. aimed to estimate the genealogy of the populations. They genotyped 121 individuals for seven linked markers on the Y chromosome. They scanned the space of genealogies of 13 populations, for each population genealogy using the coalescent distribution to simulate gene genealogies of 121 lineages. Their numerical procedure, a Bayesian Markov chain Monte Carlo approach, guaranteed that the possible population genealogies and gene genealogies were visited during the scanning process with frequencies proportional to their likelihoods. Of the population genealogies visited by their population growth model, 91% included a monophyletic grouping of the 3 African populations. Such a grouping only has probability 1/132 for random labeled histories sampled from the Yule distribution. Thus, the analysis was quite confident in the monophyly of these populations.

Discordance between gene and species genealogies is considered in a study of a human, a gorilla, and a chimpanzee. Chen & Li (2001) used genetic data in a study of the classic "trichotomy problem," that of deciding which pair of species, among humans, gorillas, and chimpanzees, has the closest relationship. The divergence of the three species occurred during a short enough period of time that genealogies vary by locus. Unlike in the case of separate groups within the human population, however, the splits among these species occurred long enough ago that nonmonophyly is unlikely for genealogies representing only one of them; thus, attention can be restricted to one lineage per species. Of the gene genealogies estimated by Chen and Li – one for each of 53 non-coding regions – the majority (31/53) showed that the human and chimpanzee had the most similar DNA sequences, favoring a grouping of humans and chimpanzees. By computing a multinomial likelihood to measure the weight of the evidence, Chen and Li concluded that their data provided strong very strong support for the human-chimpanzee grouping.

Demographic History

Gene genealogies are frequently applied to the reconstruction of population histories from DNA sequences. The inference of population and species phylogenies is one example of this kind of application. A second is the quantitative estimation of parameters of population history, such as times of divergence or migration rates.

Morrell *et al*. (2003) sequenced nine loci in 25 individuals representing three populations of wild barley: two low-elevation groups from east and west of the Zagros mountains in southwest Asia, and one group from the mountainous region itself. They were interested in the amount of migration among the three populations. Using a procedure that searches the space of possible migration rates and gene genealogies, sampling regions of this space in proportion to their likelihoods of explaining the data, they estimated that ~1-2 migrants move from each population to each of the other two populations in every generation. Morrell *et al*. suggest that this observation could be a consequence of dispersal via seeds embedded in the fur of migratory animals, or of deliberate dispersal by ancient hunter-gatherer peoples.

Selected Genes and Speciation Genes

One of the aims of genome-wide studies is to identify loci that have been strongly affected by natural selection. Demographic phenomena, such as admixture and migration, affect individuals, and are reflected in patterns of genetic variability across whole genomes. Natural selection, however, is localized to particular regions of the genome. Thus, selected loci can potentially be identified through their deviations from genome-wide averages. One way in which such deviations can be identified is through anomalous properties of gene genealogies.

Using 10-20 individuals per species and a popular genealogical estimation method – the neighbor-joining algorithm – Machado & Hey (2003) inferred the genealogies for 16 regions in the genomes of three *Drosophila* species. Genealogies for regions on chromosomes X and 2 came closer to achieving monophyletic concordance – in which lineages from each species were monophyletic and the collapsed labeled topology matched the labeled topology of the species phylogeny – than did genealogies for regions on other chromosomes. Interestingly, laboratory studies have assigned to chromosomes X and 2 the highest densities of hybrid-sterility genes in the genome. Machado and Hey suggest a view in which genotypes on chromosomes X and 2 diverged earlier in speciation than did those of other chromosomes, as it was possible to produce hybrids with differing genotypes on other chromosomes long after hybrids with incompatible types on chromosomes X and 2 were no longer viable.

Experimental Design

Experimental studies of genetic variation require choices about sample sizes, numbers of markers, and statistical methods. Random genealogies can assist in deciding how to optimize studies to obtain maximal information about quantities of interest with minimal effort.

Pluzhnikov & Donnelly (1996) considered various ways of estimating the population mutation parameter θ, which measures the level of genetic diversity in a set of DNA sequences. Because longer branches in genealogies provide more opportunities for mutations to occur, the information that a data set contains about mutation parameters increases with the branch lengths of underlying genealogies. To improve the precision in an estimate of θ obtained from a set of DNA sequences, data can be added either by sampling new individuals for the same sequenced region or by increasing the length of the region. Because individual DNA sequences are correlated in that they result from the same genealogies, the addition of individuals provides new information about θ only if the new individuals represent parts of genealogies that have not yet been sampled. Lengthening the sequence provides additional loci at which recombination could have occurred. Because recombination causes neighboring loci to have different (though correlated) genealogies, additional sequence provides new information if recombination did indeed occur. Pluzhnikov and Donnelly used random genealogies to derive expressions for the variance of estimates of θ as a function of sample size and sequence length. They determined what allocation of resources to sample size and sequence length led to the smallest variance in the estimate of θ. For various values of θ and recombination rates, they found that samples of fairly small size (~3-10) were optimal, with most of the effort devoted to increasing the lengths of sequences from these individuals. Their optimal schemes can be used for future studies that aim to estimate θ.

A related use of gene genealogies for experimental design is in evaluating statistical methods. Ramos-Onsins & Rozas (2002) were interested in identifying tests useful for detecting population growth. Using extensions of the coalescent for population growth models, they simulated genealogies, on which they simulated mutations in order to obtain simulated data sets of DNA sequences. For each simulated data set, they applied 17 tests, observing that their own R_2 test and Fu's F_S test most frequently rejected the null hypothesis that the sequences were drawn from a constant population size model when indeed they were sampled from a growing population. Thus, investigators who wish to detect growth may be more successful if one of these two tests rather than one of the other 15 methods studied is used.

Genetics of Complex Traits

Many traits, including various human diseases, result from the interactions of multiple genetic factors. By searching for alleles that are found more frequently among individuals who have a trait than among those who do not, a genome can be narrowed to a small set of alleles that can be more directly tested for possible effects on the trait. These alleles must have originated as mutations in ancestors of the extant individuals who possess them. Thus, considering the genealogies on which these mutations occurred can help to make predictions about properties of trait loci; these predictions, in turn, can be used to design streamlined strategies to map the loci.

Using a random genealogy model, Pritchard (2001) studied the fraction of the individuals with a disease who possess the disease-susceptibility allele of highest frequency. In the model, mutations could occur from “normal” to “susceptibility” alleles and vice versa. Susceptibility alleles conferred elevated disease risks and selective disadvantages to their possessors. For various assumptions about mutation rates, selection coefficients, and human demographic history, random genealogies were simulated backwards to a MRCA, which was assumed to be a normal allele. For each mutation on the genealogy that changed a normal to a susceptibility allele, the number of descendants of that mutation in a sample was tabulated. The mutation rate from normal to susceptibility alleles was observed to be the most important determinant of the fraction of diseased individuals who possessed the most frequent allele. Except at very small values of this rate, only a small fraction of the diseased individuals descended from the highest-frequency mutation. Pritchard concluded that mapping strategies will be most effective if they account for the possibility that disease-susceptibility genes might have many low-frequency mutations, each of which is found in only a small proportion of diseased individuals.

Future directions

The use of gene genealogies has led to new ways of conceptualizing genetic variation. By viewing genetic variation as the result of mutations on branches of genealogies, it becomes possible to reason about the signatures of evolutionary phenomena in data by thinking about how these phenomena affect genealogies. The coalescent enables quantification of the resulting intuitions, and new insights about evolutionary processes continue to follow from the incorporation of new phenomena into genealogical models. Statistical approaches based on gene genealogies continue to find new applications, of which the examples above give only a short introduction.

By considering many possible random genealogies that *could* underlie the pattern of variation at a locus, and by treating independent loci as replicates of the evolutionary process, methods based on genealogies can enable estimation of population history parameters and measurement of the uncertainty in the estimates. Because many uses of gene genealogies cannot yet be incorporated in methods that both quantify uncertainty in estimates and evaluate relative support for alternative models (Knowles & Maddison, 2002), however, a major challenge is to develop methods applicable to the complex scenarios that are typically of interest. This endeavor requires computational improvements: while the simulation of random genealogies and data sets can usually be performed quickly, simulation of random genealogies from the conditional distribution of the genealogy given a specific data set is generally slow (Stephens, 2001). Use of approximate numerical techniques may lead to greater computational tractability (Hudson, 2001; Beaumont *et al.*, 2002). Such tools will be especially useful for forthcoming genome-wide data on genetic variation.

Computational infeasibility is a particular problem in regions with large amounts of recombination. Such regions produce a sequence of correlated genealogies, which can be simulated using an adaptation of the coalescent (Nordborg, 2001); however, most existing statistical tools apply only to individual regions with little or no recombination, or to unlinked collections of several such regions. Construction of computationally desirable models of genealogies that are not based on the coalescent may help to deal with this problem (Li & Stephens, 2003). Indeed, the development of models of gene genealogies and the statistical methods to which they give rise offers many new challenges for the genomic era.

Suggestions for further reading

Ewens, W. J. 2004. “Mathematical Population Genetics I. Theoretical Introduction.” Springer-Verlag, New York, 2nd edition.

Felsenstein, J. 2004. “Inferring phylogenies.” Sinauer, Sunderland, MA.

Hudson, R. R. 1990. Gene genealogies and the coalescent process, *Oxford Surv. Evol. Biol.* **7**, 1-44.

Maddison, W. P. 1997. Gene trees in species trees. *Syst. Biol.* **46**, 523-536.

Nordborg, M. 2001. Coalescent theory, in “Handbook of Statistical Genetics” (D. J. Balding, M. Bishop, and C. Cannings, eds), chapter 7, pp. 179-212, Wiley, Chichester, UK.

Rosenberg, N. A. and Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Rev. Genet.* **3**, 380-390.

The well-known reviews of Hudson (1990) and Maddison (1997) cover gene genealogies and the coalescent, and the relationship of gene genealogies to species phylogenies, respectively. A rich and thorough survey by Nordborg (2001), supplemented by our somewhat less mathematical addendum (Rosenberg & Nordborg, 2002), provides a more recent treatment. Material on gene genealogies is expertly embedded in the context of theoretical population genetics by Ewens (2004) and in the context of phylogenetics by Felsenstein (2004).

Acknowledgments

I thank Steve Finkel, Peter Morrell, Mark Tanaka, John Wakeley, Jeff Wall, and Jason Wolf for extensive comments on a draft of this chapter.

Box 1. Horizontal Inheritance

Individuals of some organisms can inherit DNA from individuals other than their parents. This is particularly true for certain haploids, who can replace DNA that they "vertically" inherit from parents with DNA "horizontally" inherited from other individuals of the same species, individuals of other species, or the surrounding environment (Bushman, 2002). Such organisms have two types of coalescence, vertical and horizontal.

Because of horizontal inheritance, genealogies in many haploid species might not follow the pattern of bifurcation of genomes expected for haploids. With horizontal transfer, haploid genealogies contain many of the complexities seen in gene genealogies of diploids. Just as recombination enables different parts of the genomes of diploids to have distinct genealogies, horizontal DNA transfer leads to differing genealogies for different parts of a haploid genome. Analogously, as migration in diploids can lead different multi-population genealogies to have different collapsed labeled topologies, horizontal inheritance among individuals from different species can cause such discordances in haploid genealogies.

Recall that in diploids, discordance of collapsed labeled topologies does not require migration among populations. Similarly, in haploids, such discordance can arise even if no horizontal transfers occur between individuals of different species. In other words, discordance of collapsed labeled topologies for genealogies for several regions of a genome can result from horizontal transfer *between* species or *within* species. At the same time, however, horizontal transfers between or within species need not lead to discordance.

In bacterial studies, it is of interest to identify which genes have and have not been transferred across species, and for those that have been transferred, to identify the donor species (Eisen, 2000; Koonin, 2003). Because any shape for a haploid genealogy can be produced by many different combinations of horizontal transfers within and between species, it is important to quantitatively evaluate the relative support for different scenarios. Such an endeavor might be advanced by connecting horizontal transfer models to the coalescent.

A Horizontal Transfer Model

Consider a random sample of n individuals from a haploid population of constant size N in a closed environment, with $N>>n$. Suppose that the individuals have independently and identically distributed lifespans that follow exponential distributions with mean 1 generation. When an individual dies, another individual randomly chosen from the population duplicates to replace it. These are the basic assumptions of the *Moran model*, a frequently-used neutral model in population genetics (Ewens, 2004).

Looking backwards in time from the sample of n individuals, the waiting time until one of the individuals arose from its parent is exponentially distributed with mean $1/n$ generations. The probability that this origin is a (vertical) coalescence is the probability that the parent is ancestral to the other n-1 sampled individuals, or $(n\text{-}1)/(N\text{-}1)$. Using basic properties of exponential random variables, the time until a vertical coalescence is exponentially distributed with mean $(N\text{-}1)/[n(n\text{-}1)]$ generations. Genealogies in this model follow the coalescent distribution with coalescence effective size $(N\text{-}1)/2$.

Now suppose that for each individual, the waiting time until its DNA at a locus of interest is replaced by DNA horizontally transferred from another individual in the population is exponentially distributed with mean $1/\lambda$ generations. Such transfers could potentially occur by *conjugation*, *transduction*, or *transformation*, procedures in which DNA is transferred between cells via plasmids, viruses, or the extracellular environment, respectively (Bushman, 2002).

Assuming that horizontal transfers in different individuals are independent, the waiting time (backwards in time) until one of the lineages experiences a horizontal transfer event (as the recipient of DNA) is exponentially distributed with mean $1/(n\lambda)$ generations. If the individual that donates DNA during this transfer is an ancestor to one of the other n-1 sampled lineages, an event that has probability $(n\text{-}1)/(N\text{-}1)$, horizontal coalescence occurs. If this donor is not an ancestor to the n-1 lineages, no coalescence takes place. As before, using the properties of exponential random variables, the time until a horizontal coalescence is exponentially distributed with mean $(N\text{-}1)/[\lambda n(n\text{-}1)]$ generations.

Considering the vertical and horizontal processes simultaneously, the time until a coalescence of either type has exponential distribution with mean $(N\text{-}1)/[(1+\lambda)n(n\text{-}1)]$ generations. This distribution has the same form as in models that only include vertical coalescence. In other words, the waiting times in this model follow the coalescent distribution with coalescence effective size $(N\text{-}1)/[2(1+\lambda)]$.

Implications of the Model

In comparison with a model that includes vertical coalescence only, the horizontal transfer model has shorter waiting times until coalescence, so that lineages find a MRCA more rapidly. This is sensible, as horizontal inheritance enables genes to diffuse rapidly through a population. The amount by which horizontal transfer speeds up coalescence depends on λ, which measures the mean number of horizontal transfers experienced by a random individual at the locus of interest during a lifetime of average length. If λ is very small – that is, if most cells die before experiencing any transfers, the presence of horizontal transfer has little effect on genealogies, and most coalescences are vertical.

The horizontal transfer model has a coalescence effective size, so that the coalescent distribution applies to its genealogies. Thus, in the same way used for models without horizontal transfer, it can potentially be generalized to allow multiple genes, populations, or species. This could enable methods originally designed for such problems as the estimation of migration rates (Beerli & Felsenstein, 2001; Nielsen & Wakeley, 2001) to be applied to estimation of horizontal transfer rates within and among species, and to probabilistic determination of the sources of observed apparent transfers.

References

Aldous, D. J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, *Stat. Sci.* **16**, 23-34.

Avise, J. C. 2000. "Phylogeography: The History and Formation of Species", Harvard University Press, Cambridge, MA.

Beaumont, M. A., Zhang, W., and Balding, D. J. 2002. Approximate Bayesian computation in population genetics, *Genetics* **162**, 2025-2035.

Beerli, P. and Felsenstein, J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach, *Proc. Natl. Acad. Sci. USA* **98**, 4563-4568.

Brown, J. K. M. 1994. Probabilities of evolutionary trees, *Syst. Biol.* **43**, 78-91.

Bushman, F. 2002. "Lateral DNA Transfer", Cold Spring Harbor Press, Cold Spring Harbor, New York.

Chen, F.-C. and Li, W.-H. 2001. Genomic divergences between humans and other Hominoids and the effective population size of the common ancestor of humans and chimpanzees, *Am. J. Hum. Genet.* **68**, 444-456.

Derrida, B., Manrubia, S. C., and Zanette, D. H. 2000. On the genealogy of a population of biparental individuals, *J. theor. Biol.* **203**, 303-315.

Donnelly, P. 1996. Interpreting genetic variability: the effects of shared evolutionary history, *in* "Variation in the Human Genome", pp. 25-40, Wiley, Chichester, UK.

Donnelly, P. and Tavaré, S., eds. 1997. "Progress in Population Genetics and Human Evolution", Springer, New York.

Durrett, R. 2002. "Probability Models for DNA Sequence Evolution", Springer-Verlag, New York.

Eisen, J. A. 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis, *Curr. Op. Genet. Devel.* **10**, 606-611.

Ewens, W. J. 2004. "Mathematical Population Genetics I. Theoretical Introduction", Springer-Verlag, New York, 2nd edition.

Felsenstein, J. 2004. "Inferring Phylogenies", Sinauer, Sunderland, MA.

Fu, Y.-X. and Li, W.-H. 1993. Statistical tests of neutrality of mutations, *Genetics* **133**, 693-709.

Hey, J. and Machado, C. A. 2003. The study of structured populations – new hope for a difficult and divided science, *Nature Rev. Genet.* **4**, 535-543.

Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination, *Theor. Pop. Biol.* **23**, 183-201.

Hudson, R. R. 1990. Gene genealogies and the coalescent process, *Oxford Surv. Evol. Biol.* **7**, 1-44.

Hudson, R. R. 2001. Two-locus sampling distributions and their application, *Genetics* **159**, 1805-1817.

Kingman, J. F. C. 1982. On the genealogy of large populations, *J. Appl. Prob.* **19A**, 27-43.

Knowles, L. L. and Maddison, W. P. 2002. Statistical phylogeography, *Mol. Ecol.* **11**, 2623-2635.

Koonin, E. V. 2003. Horizontal gene transfer: the path to maturity, *Mol. Microbiol.* **50**, 725-727.

Li, N. and Stephens, M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data, *Genetics* **165**, 2213-2233.

Li, W.-H. and Fu, Y.-X. 1999. Coalescent theory and its applications in population genetics, *in* "Statistics in Genetics" (M. E. Halloran and S. Geisser, eds), pp. 45-79, Springer-Verlag, New York.

Machado, C. A. and Hey, J. 2003. The causes of phylogenetic conflict in a classic *Drosophila* species group, *Proc. R. Soc. Lond. B* **270**, 1193-1202.

McKenzie, A. and Steel, M. 2000. Distributions of cherries for two models of trees, *Math. Biosci.* **164**, 81-92.

Möhle, M. 2000. Ancestral processes in population genetics – the coalescent, *J. theor. Biol.* **204**, 629-638.

Morrell, P. L., Lundy, K. E., and Clegg, M. T. 2003. Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite migration, *Proc. Natl. Acad. Sci. USA* **100**, 10812-10817.

Nath, H. B. and Griffiths, R. C. 1993. The coalescent in two colonies with symmetric migration, *J. Math. Biol.* **31**, 841-852.

Nielsen, R. and Wakeley, J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach, *Genetics* **158**, 885-896.

Nordborg, M. 2001. Coalescent theory, *in* "Handbook of Statistical Genetics" (D. J. Balding, M. Bishop, and C. Cannings, eds), chapter 7, pp. 179-212, Wiley, Chichester, UK.

Nordborg, M. and Krone, S. M. 2002. Separation of time scales and convergence to the coalescent in structured populations, *in* "Modern Developments in Theoretical Population Genetics" (M. Slatkin and M. Veuille, eds), chapter 12, pp. 194-232, Oxford University Press, Oxford.

Pamilo, P. and Nei, M. 1988. Relationships between gene trees and species trees, *Mol. Biol. Evol.* **5**, 568-583.

Pluzhnikov, A. and Donnelly, P. 1996. Optimal sequencing strategies for surveying molecular genetic diversity, *Genetics* **144**, 1247-1262.

Pritchard, J. K. 2001. Are rare variants responsible for susceptibility fo complex diseases? *Am. J. Hum. Genet.* **69**, 124-137.

Ramos-Onsins, S. E. and Rozas, J. 2002. Statistical properties of new neutrality tests against population growth, *Mol. Biol. Evol.* **19**, 2092-2100.

Rohde, D. L. T., Olson, S., and Chang, J. T. 2004. Modelling the recent common ancestry of all living humans, *Nature* **431**, 562-566.

Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees, *Theor. Pop. Biol.* **61**, 225-247.

Rosenberg, N. A. 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model, *Evolution* **57**, 1465-1477.

Rosenberg, N. A. and Feldman, M. W. 2002. The relationship between coalescence times and population divergence times, *in* "Modern Developments in Theoretical Population Genetics" (M. Slatkin and M. Veuille, eds), chapter 9, pp. 130-164, Oxford University Press, Oxford.

Rosenberg, N. A. and Hirsh, A. E. 2003. On the use of star-shaped genealogies in inference of coalescence times, *Genetics* **164**, 1677-1682.

Rosenberg, N. A. and Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms, *Nature Rev. Genet.* **3**, 380-390.

Saunders, I. W., Tavaré, S., and Watterson, G. A. 1984. On the genealogy of nested subsamples from a haploid population, *Adv. Appl. Prob.* **16**, 471-491.

Slatkin, M. and Hudson, R. R. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations, *Genetics* **129**, 555-562.

Slatkin, M. and Veuille, M., eds. 2002. "Modern Developments in Theoretical Population Genetics", Oxford University Press, Oxford.

Slowinski, J. B. and Guyer, C. 1989. Testing the stochasticity of patterns of organismal diversity: an improved null model, *Am. Nat.* **134**, 907-921.

Steel, M. and McKenzie, A. 2001. Properties of phylogenetic trees generated by Yule-type speciation models, *Math. Biosci.* **170**, 91-112.

Stephens, M. 2001. Inference under the coalescent, *in* "Handbook of Statistical Genetics" (D. J. Balding, M. Bishop, and C. Cannings, eds), chapter 8, pp. 213-238, Wiley, Chichester, UK.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations, *Genetics* **105**, 437-460.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, *Genetics* **123**, 585-595.

Takahata, N. and Nei, M. 1985. Gene genealogy and variance of interpopulational nucleotide differences, *Genetics* **110**, 325-344.

Takahata, N. and Slatkin, M. 1990. Genealogy of neutral genes in two partially isolated populations, *Theor. Pop. Biol.* **38**, 331-350.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. 1997. Inferring coalescence times from DNA sequence data, *Genetics* **145**, 505-518.

Teshima, K. M. and Tajima, F. 2002. The effect of migration during the divergence, *Theor. Pop. Biol.* **62**, 81-95.

Uyenoyama, M. K. 1997. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants, *Genetics* **147**, 1389-1400.

Wakeley, J. 1998. Segregating sites in Wright's island model, *Theor. Pop. Biol.* **53**, 166-174.

Wakeley, J. 2000. The effects of subdivision on the genetic divergence of populations and species, *Evolution* **54**, 1092-1101.

Wilkinson-Herbots, H. M. 1998. Genealogy and subpopulation differentiation under various models of population structure, *J. Math. Biol.* **37**, 535-585.

Wilson, I. J., Weale, M. E., and Balding, D. J. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities, *J. Roy. Statist. Soc. Ser. A* **166**, 155-187.

Wollenberg, K. and Avise, J. C. 1998. Sampling properties of genealogical pathways underlying population pedigrees, *Evolution* **52**, 957-966.

Figure Legends

Figure 1. Pedigrees and gene genealogies. (i) The pedigree for a set of six individuals from the current generation. Empty squares and circles represent males and females, respectively. (ii) Application of the law of segregation to a randomly chosen locus, conditional on the pedigree. This diagram shows the transmission paths of a particular locus through the pedigree. Shaded squares and circles respectively represent paternal and maternal copies of a locus. (iii) Genealogy of the copies of the locus present in the most recent generation of individuals, showing only the transmissions that contribute to the current generation. (iv) Abstracted genealogy obtained by rearranging the order of the copies. Time proceeds downwards (in this figure and in subsequent figures).

Figure 2. A genealogy for seven sampled lineages. Cherries are marked with open circles. Each lineage is separated from the root by exactly three branches, except for lineage 5, which is separated by only two branches. The two subgenealogies that coalesce at the root have four and three lineages (lineages 1, 2, 3, and 4 in the subgenealogy on the left, and lineages 5, 6, and 7 in the one on the right). W_n is the length of time during which the sampled lineages have exactly n ancestors. T_{MRCA} for the lineages is represented by the height of the genealogy, $T_n = \sum_{n=2}^{7} W_n$. The total length of all branches in the genealogy is $L_n = \sum_{n=2}^{7} nW_n$. This length is the sum of E_n, the sum of the lengths of the external branches (marked E), and I_n, the sum of the lengths of the internal branches (marked I). Basal branches are marked B.

Figure 3. Labeled histories and labeled topologies for example genealogies. The sequences of coalescence events are: (i) (1,3), (2,5), ((2,5),4), ((1,3),((2,5),4)); (ii) (2,5), (1,3), ((2,5),4), ((1,3),((2,5),4)); (iii) (2,5), ((2,5),4), (1,3), ((1,3),((2,5),4)); (iv) (1,4), (2,3), ((2,3),5), ((1,4),((2,3),5)). The genealogies in (i), (ii), and (iii) have the same coalescence events and therefore have the same labeled topology, ((1,3),((2,5),4)). Because the order of the coalescences differs for (i), (ii), and (iii), however, these genealogies have different labeled histories. Although its coalescence times equal those of (i), the genealogy in (iv) has different coalescences from those of (i), (ii), and (iii); thus, it differs from the other genealogies both in labeled history and in labeled topology.

Figure 4. The four phyletic statuses possible for a genealogy of lineages sampled from two species. Thick lines represent the divergence of an ancestral species into two descendant species, *A* and *B*. Thin lines represent the genealogy of the sampled lineages from the two species. C1–monophyly of *A* and *B*. C2–paraphyly of *B* with respect to *A*. C3–paraphyly of *A* with respect to *B*. C4–polyphyly of *A* and *B*.

Figure 5. Classification of a genealogy of lineages from four species. Shaded circles indicate interspecific coalescence events used in determining the collapsed genealogy. Note that coalescences of lineages from two species occur prior to species divergence. The genealogy can be classified as follows. M-type: *A* – monophyletic, *B* – not monophyletic, *C* – not monophyletic, *D* – not monophyletic. P-type: (*A*,*B*)–C2, (*A*,*C*)–C2, (*A*,*D*)–C1, (*B*,*C*)–C4, (*B*,*D*)–C3, (*C*,*D*)–C4. C-type: the collapsed labeled history of the genealogy is the sequence (*C*,*D*), (*A*,*B*), ((*A*,*B*),(*C*,*D*)). D-type: the collapsed labeled topology of the genealogy is ((*A*,*B*),(*C*,*D*)). The collapsed labeled topology of the genealogy and the labeled topology of the species phylogeny match, but the collapsed labeled history of the genealogy and the labeled history of the species phylogeny differ.

Figure 6. Mutations on genealogies. Suppose the genealogy in Figure 2 describes a locus with three nucleotides, and that the MRCA has genotype AGC. Mutations on the genealogy lead to the genotypes shown for the sampled lineages. Because mutations can be viewed as random events that occur along the branches of a genealogy, random samples of genotypes can be obtained by placing random mutations on simulated random genealogies. Mutations may obscure the underlying relationships of the lineages: although lineages 1 and 2 are closely related, they differ in genotype; at the same time, 1 and 5 are distantly related, but are identical in genotype.

Figure A (inside Box 1). Genealogies for two loci. Each locus, one with a darkly drawn and the other with a lightly drawn genealogy, is sampled in a set of four individuals. Vertical and horizontal coalescences are marked with shaded and empty squares, respectively. Because horizontal transfer of DNA from one individual to another usually involves pieces of DNA that are small relative to the genome size, it is unlikely that two loci, unless separated by a short distance, would experience horizontal coalescence in the same individual. Because of horizontal inheritance,

the two genealogies have different labeled topologies: ((1,2),(3,4)) and (((1,2),3),4) for the dark and light genealogies, respectively.

Figure B (inside Box 1). Genealogies with discordant collapsed labeled topologies caused by horizontal transfer among individuals of (i) different species, and (ii) the same species; genealogies with concordant collapsed labeled topologies despite horizontal transfer among individuals of (iii) different species, and (iv) the same species. The thick lines represent the species phylogeny; shaded circles depict interspecific coalescences used in determining collapsed genealogies. A “species” is interpreted to be a group of individuals, each of which has the property that most of its genome coalesces with other individuals in the group, horizontally or vertically, more recently than it does with individuals not in the group. *Speciodendric* loci are those loci whose collapsed labeled topologies match that of the species phylogeny.

Tables

Table 1. Attributes of a genealogy with n lineages ($n \geq 2$).

Symbol	Description of attribute	Expected value over random genealogies generated by the coalescent distribution *	Reference
Attributes that depend only on the coalescence sequence of the genealogy			
C_n	the number of cherries ($n \geq 3$)	$n/3$	McKenzie & Steel (2000)
$X_{n,1}$	the number of branches that separate a randomly chosen lineage from the root	$2\sum_{i=2}^{n} 1/i$	Steel & McKenzie (2001)
$X_{n,2}$	the number of branches that separate the MRCA of two randomly chosen lineages from the root	$\frac{2}{n-1}\left(n-1-2\sum_{i=2}^{n} 1/i\right)$ **	Steel & McKenzie (2001)
Y_n	the number of branches that separate two randomly chosen lineages from each other	$\frac{4n+4}{n-1}\left(\sum_{i=2}^{n} 1/i\right)-4$	Steel & McKenzie (2001)
$\{l, n-l\}$	numbers of lineages in the two subgenealogies that coalesce at the root	***	Tajima (1983), Slowinski & Guyer (1989)
Attributes that depend only on the coalescence times of the genealogy			
$W_{n,k}$	the length of time it takes for n lineages to coalesce to k lineages ($1 \leq k \leq n$) ****	$2(n-k)/(nk)$	Tajima (1983)
T_n	the time to the most recent common ancestor	$2(n-1)/n$	Tajima (1983), Tavaré *et al*. (1997)
L_n	the total length of time in all branches	$2\sum_{i=1}^{n-1} 1/i$	Hudson (1990), Tavaré *et al*. (1997)
Attributes that depend on both the coalescence sequence and the coalescence times of the genealogy			
E_n	the total length of time in external branches	2	Fu & Li (1993), Durrett (2002)
I_n	the total length of time in internal branches	$2\sum_{i=2}^{n-1} 1/i$	Fu & Li (1993), Durrett (2002)
P_n	the average coalescence time for two randomly chosen lineages	1	Tajima (1983), Durrett (2002)
B_n	the average length of time in a basal branch	$2\left(1/n+\sum_{i=2}^{n-1} 1/i^2\right)$	Uyenoyama (1997)

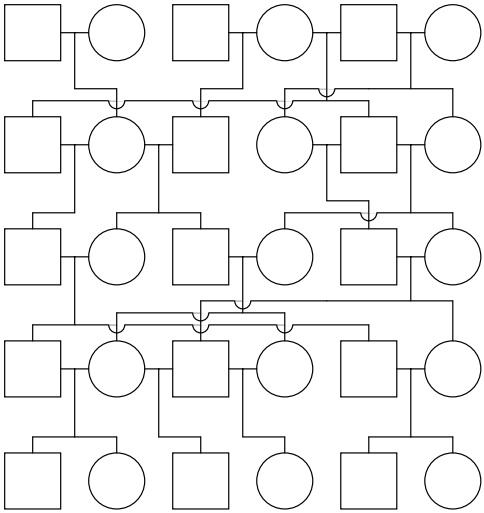
* For genealogies generated by the coalescent distribution, coalescence sequences follow the Yule distribution. Therefore, the expected values for attributes that depend only on the coalescence sequence do not utilize the exponential distribution of coalescence times under the coalescent. For attributes that depend on the coalescence times, times are measured in units of N generations, where N is the population size.

** For each k, with $1 \leq k \leq n$, we can consider $X_{n,k}$, the number of branches that separate the MRCA of k randomly chosen lineages from the root. Under the coalescent, $P[X_{n,k}=0]=(k-1)(n+1)/[(k+1)(n-1)]$ (Saunders *et al*., 1984; Steel & McKenzie, 2001).

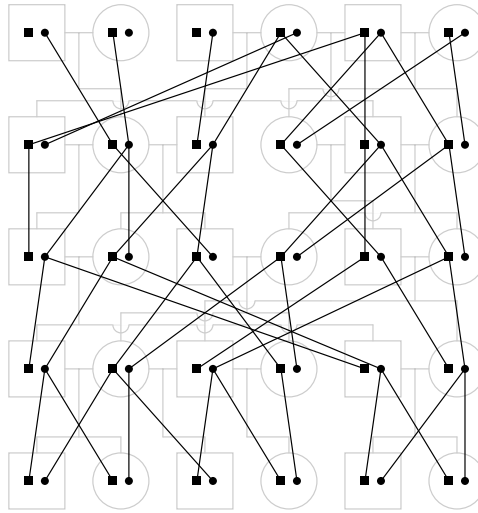
*** For each l, with $1 \leq l \leq \left\lfloor \frac{n}{2} \right\rfloor$, $\{l, n-l\}$ has probability $2/(n-1)$, with the exception that if n is even, $\{n/2, n/2\}$ has probability $1/(n-1)$.

**** The special case $W_{n,n-1}$ is often abbreviated to W_n.

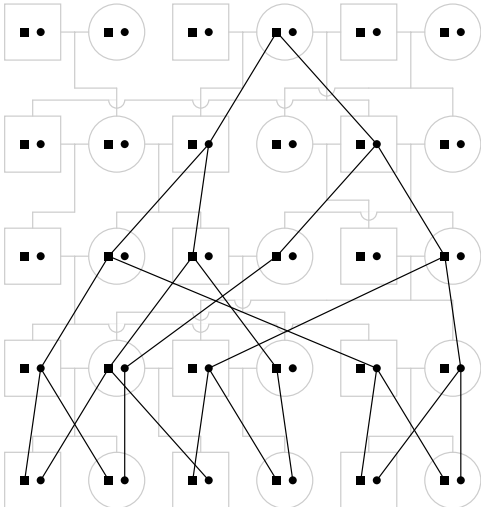
i



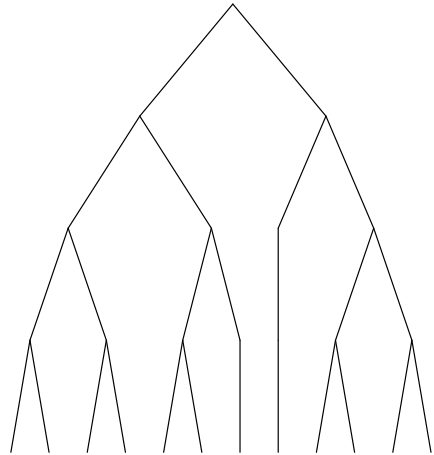
ii

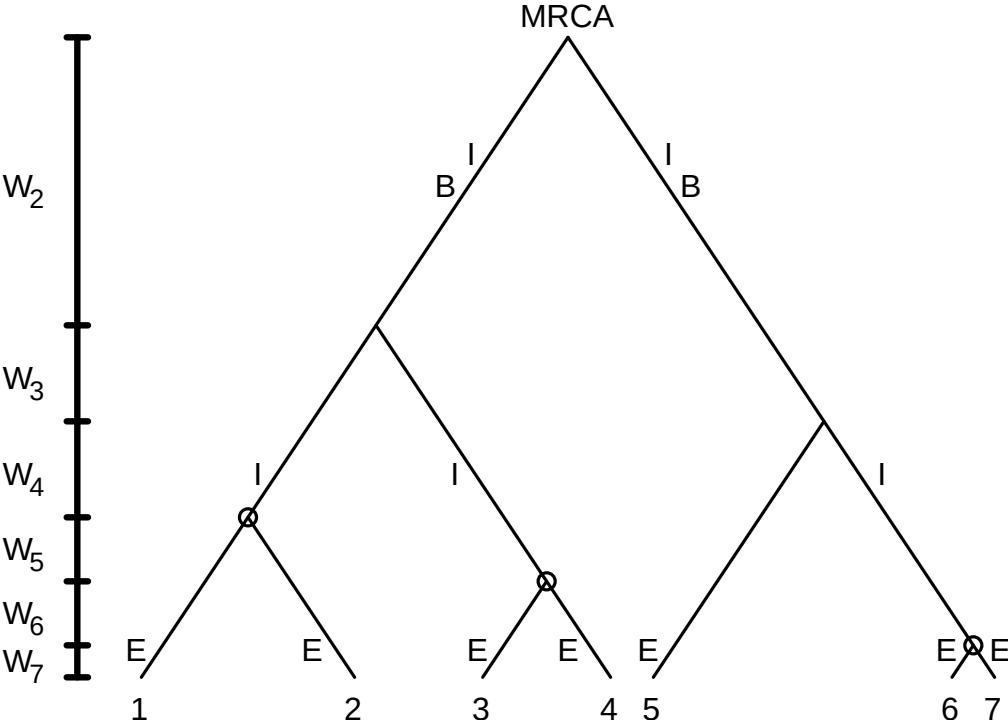


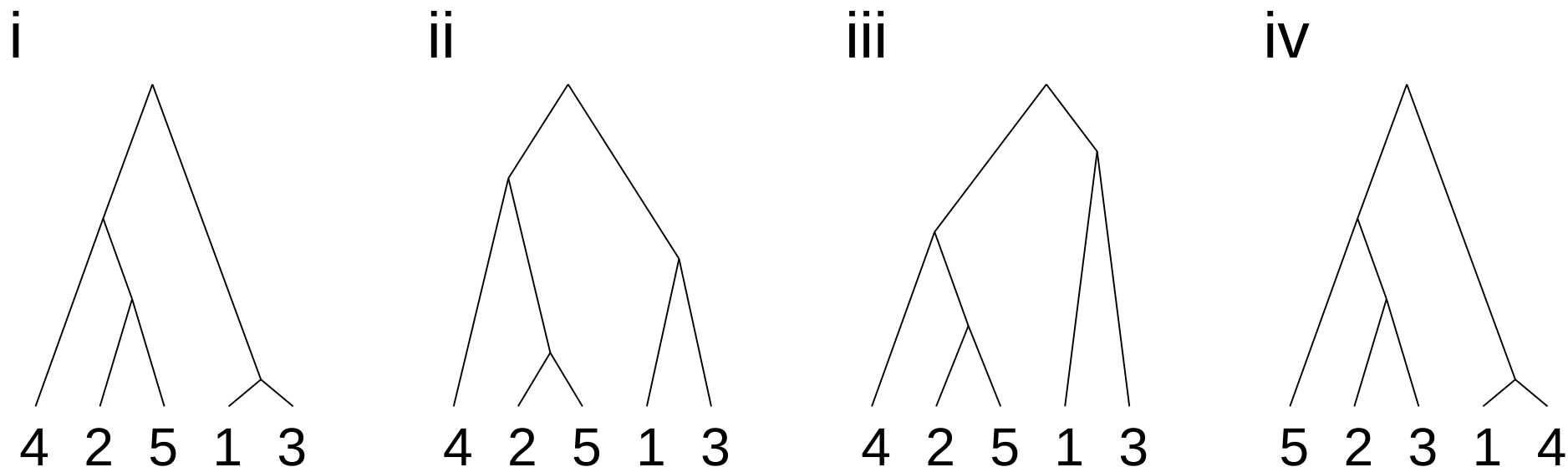
iii

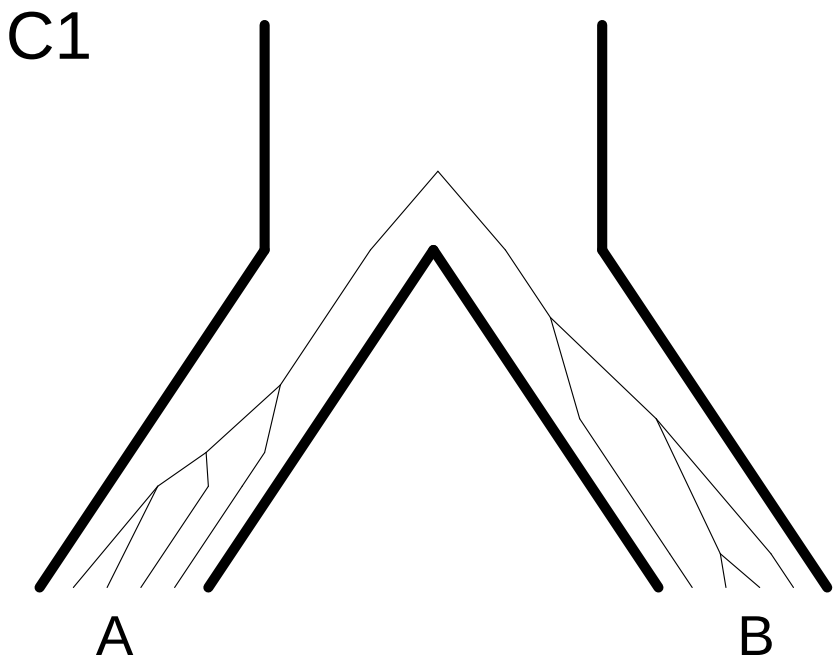


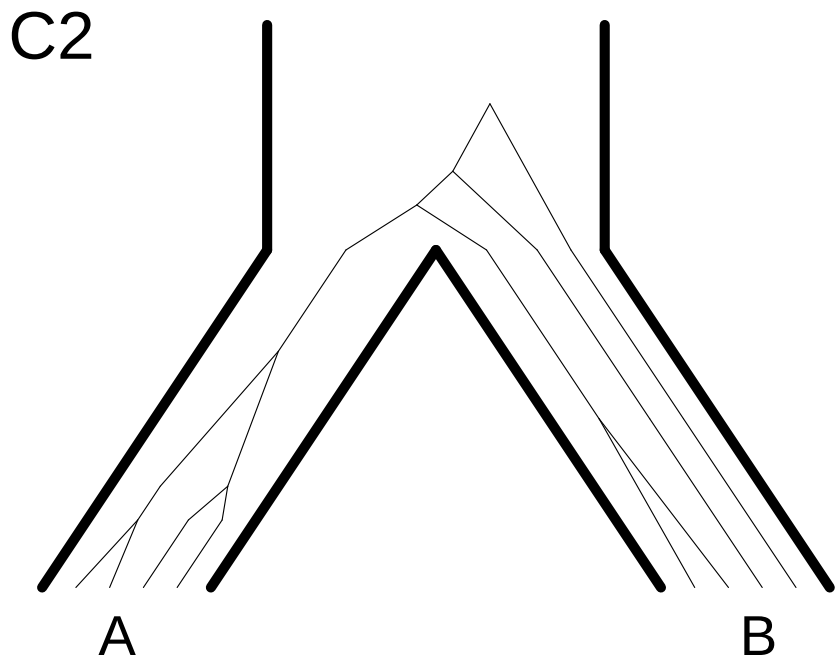
iv

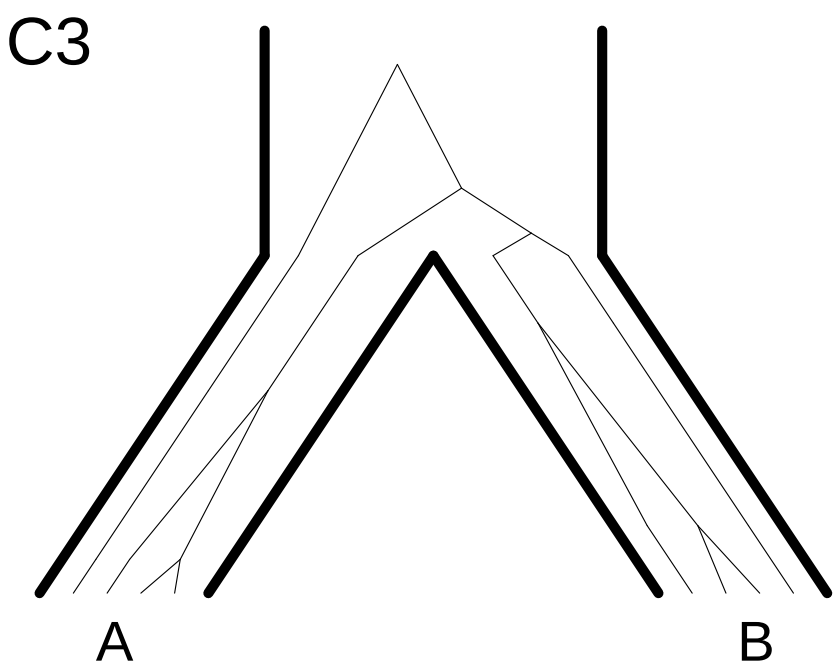


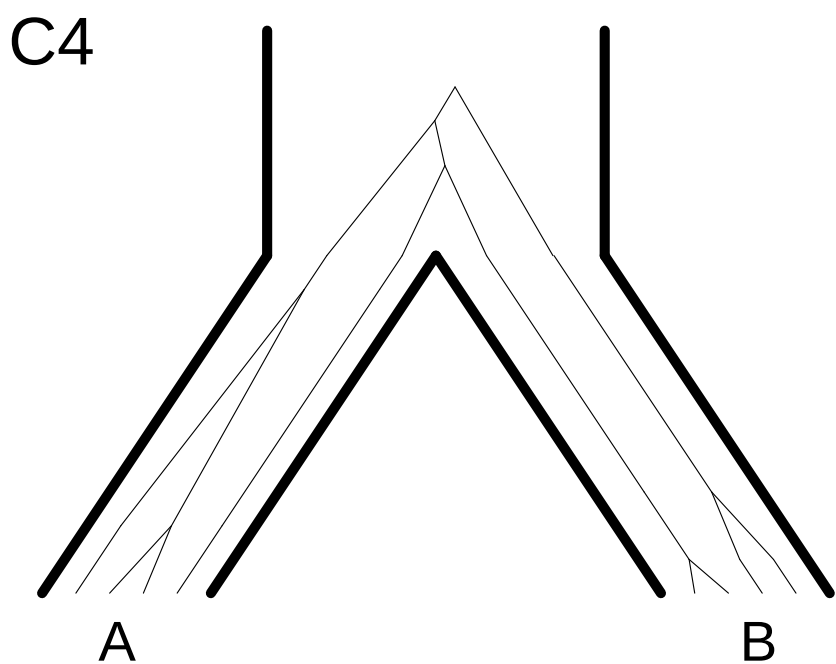

MRCA
I
B
I
B
W2
W3
W4
W5
W6
W7
I
I
I
E
E
E
E
E
E
E
1
2
3
4
5
6
7

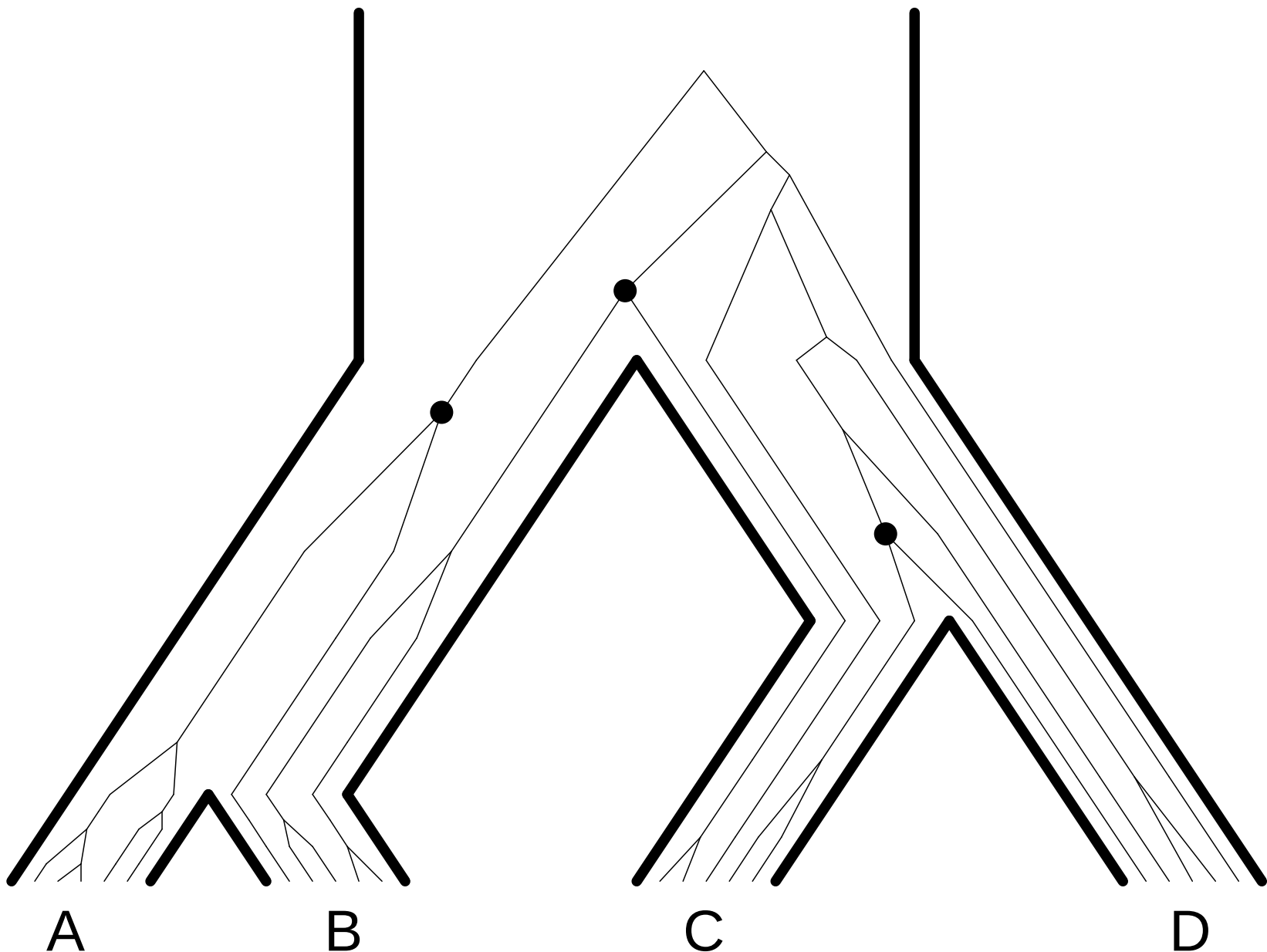

i
ii
iii
iv
4 2 5 1 3
4 2 5 1 3
4 2 5 1 3
5 2 3 1 4

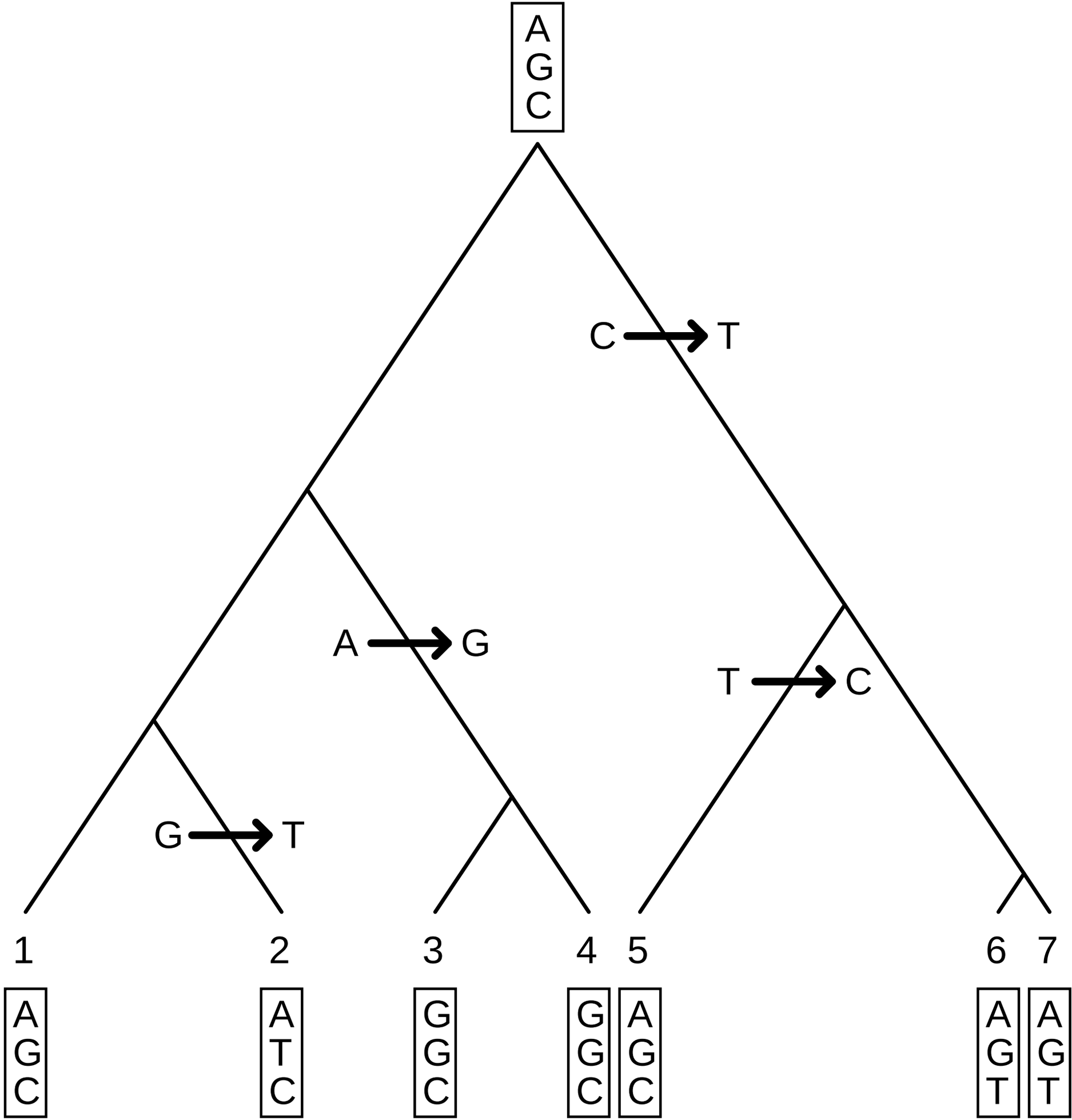

C1
A
B

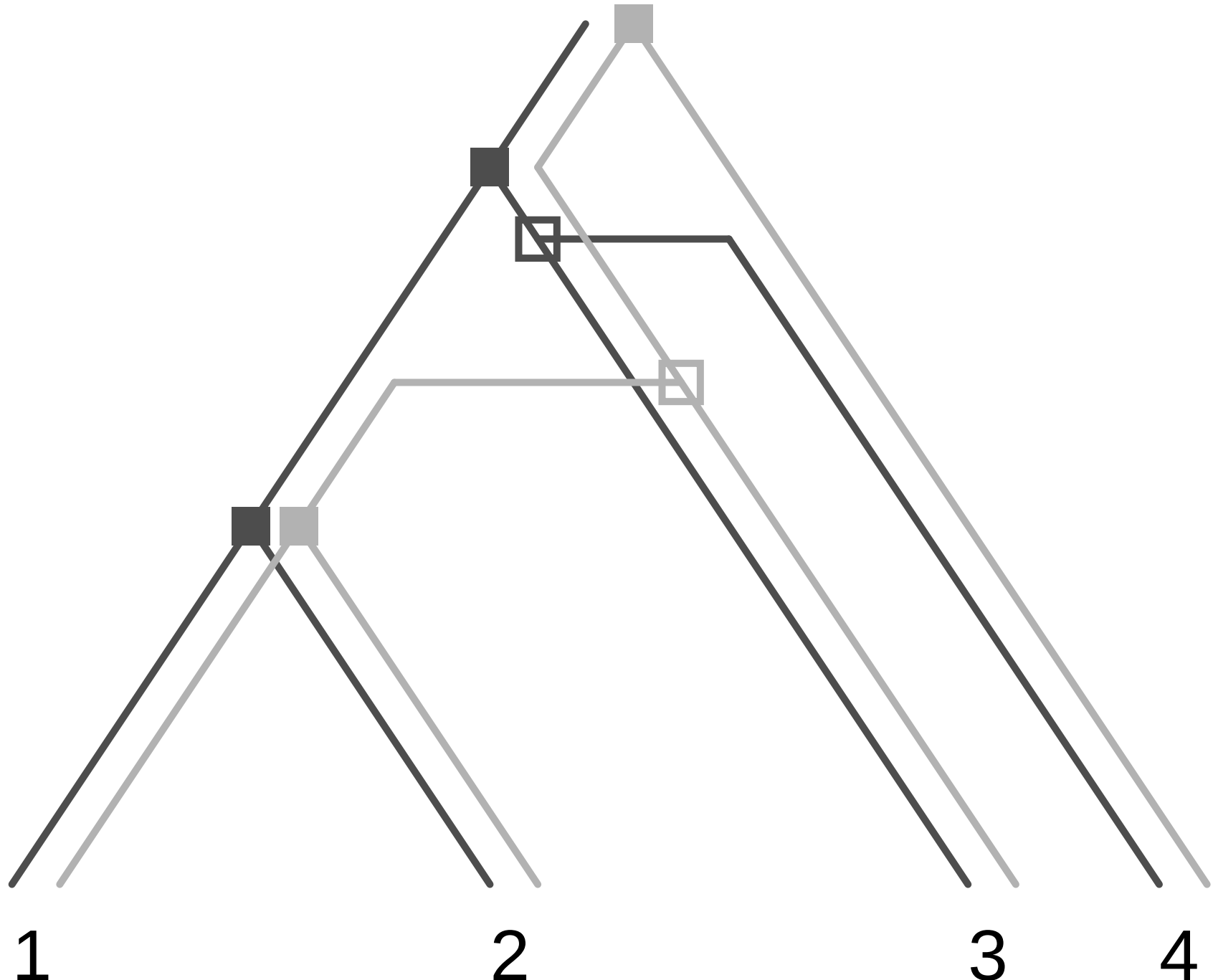

C2
A
B

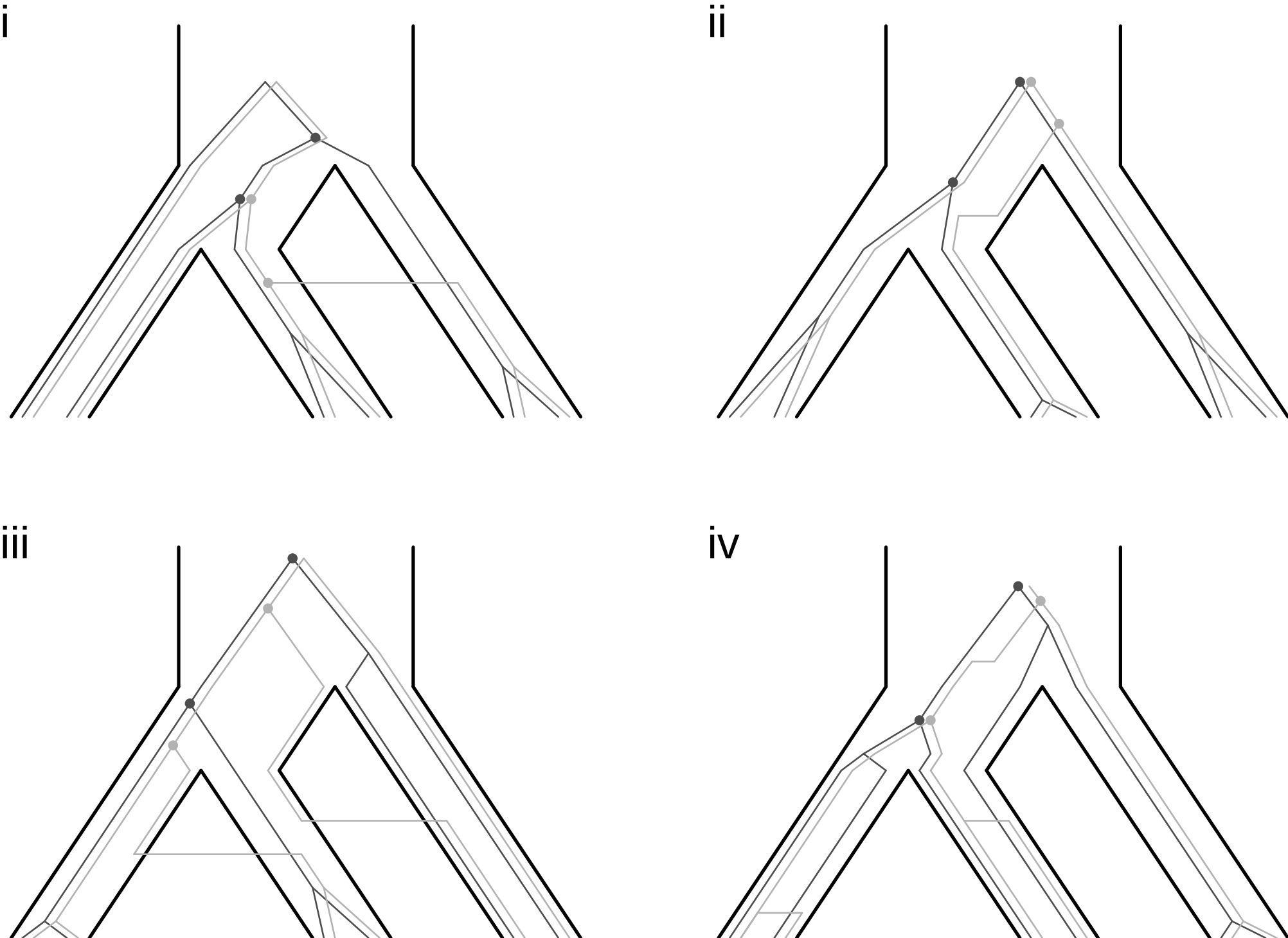

C3
A
B


C4
A
B


A
B
C
D


A
G
C
C → T
A → G
T → C
G → T
1
2
3
4
5
6
7
A
G
C
A
T
C
G
G
C
G
G
C
A
G
C
A
G
T
A
G
T


1
2
3
4


i
ii
iii
iv