

5

The coalescent with recombination

5.1 Introduction

Genetic recombination occurs in most organisms on earth, including eucaryotes, bacteria, and viruses. Recombination occurs by quite different mechanisms in these three different types of organisms. In eucaryotes recombination is usually associated with sexual reproduction. Thus, it is generally believed that sexual reproduction exists because it facilitates genetic recombination. No generally accepted theory is available to explain why genetic recombination is so common, but in its favour are arguments that genetic recombination ensures that favourable variants are brought together in the same sequence (or, equivalently, that sequences are created which do not harbour any deleterious variants), and that recombination can maintain more variation which may enable survival over an evolutionary time scale.

The simple coalescence process, including the various extensions in the previous chapter assumes no recombination and this body of theory can therefore strictly speaking only be used on non-recombining sequences. In animals, only Y-chromosomes and perhaps mitochondrion satisfies this constraint, and it is therefore pertinent for practical analysis that recombination can be build into the process. Fortunately, this is possible as shown by Hudson in 1983 very shortly after the coalescent process was first formulated. However, recombination adds much more complexity than any of the extensions discussed in Chapter 4, mainly because no single ‘coalescent tree’ can describe a sample of recombining sequences. A graph is required to describe the ancestral process which complicates the mathematical description considerably. However, the presence of recombination makes some estimators of evolutionary parameters more accurate than when applied on non-recombining sequences (see Chapter 6). Furthermore, recombination is the key force that enables us to perform linkage disequilibrium (LD) mapping in the search for genes causing common diseases (Chapter 7).

We will begin this chapter with a data example showing some of the hallmarks of recombination. We will discuss how the biology of recombination is captured by simple models of recombination and the alternative

form of genetic exchange called gene conversion. We will then formulate the coalescent process with recombination (and its gene conversion counterpart), and finally we will discuss several properties of this complex process that are important for understanding genetic variation in haplotypes and for our ability to detect the presence of recombination.

5.2 Data example with recombination

The Apolipoprotein E locus is an important human gene because variation in the gene is associated with increased risk of Alzheimer’s disease. Figure 5.1 shows the segregating sites in a large study by Fullerton et al. (2000), and the different haplotypes, including their frequencies in four different human populations. It can be seen that some haplotypes are much more common than others as expected under the basic coalescent model and that there are groups of similar haplotypes, which are quite different from

	73	308	471	545	560	624	832	1163	1522	1575	1998	2440	2907	3106	3673	3937	4036	4075	4951	5229	5361					
	C	C	A	C	A	T	G	G	G	C	G	G	T	T	C	C	C	C	A	G	T	J	C	N	R	Σ
																						1	0	1	2	4
						T										T						1	0	0	1	2
										T												0	0	0	1	1
																						0	0	1	4	5
																						0	0	0	1	1
																						0	0	0	1	1
																						0	0	0	1	1
																						1	5	2	8	16
																						0	1	0	0	1
																						8	19	11	7	45
																						0	0	1	0	1
																						0	0	2	0	2
																						1	0	0	0	1
																						3	5	0	0	8
																						1	0	0	0	1
																						2	3	0	0	5
																						8	3	3	1	15
																						0	0	0	2	2
																						15	6	11	11	43
																						1	0	0	0	1
																						1	1	4	2	8
																						0	0	1	0	1
																						2	0	0	0	2
																						1	0	0	1	2
																						0	0	1	1	2
																						1	0	0	0	1
																						1	0	0	0	1
																						0	0	0	2	2
																						0	0	1	0	1
																						0	5	9	1	15
																						0	0	0	1	1

Figure 5.1 The segregating sites in the study of the Apolipoprotein E gene by Fullerton et al. (2000). Shown are the thirty-one different haplotypes with twenty-one segregating sites compared to the corresponding nucleotides in a chimpanzee sequence. This enables us to see which variant is most likely to be ancestral. The frequencies of each of the thirty-one haplotypes in five human populations are also shown. J = Jackson (African-American), C = Campeche (Hispanic), N = North Karelia (Finnish), and R = Rochester (European-American).

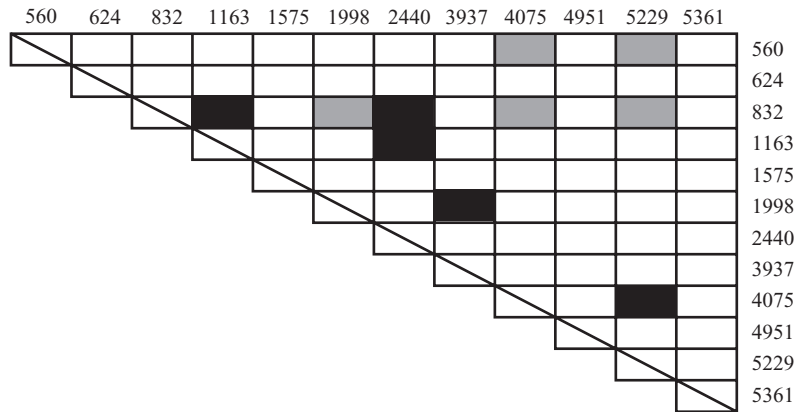


Figure 5.2 Matrix showing cases of significant LD between pairs of informative sites for the Rochester population. Grey hatching denotes significance at the 5% level, black at the 0.1% level. Informative sites are sites where both alleles are present in at least two sequences each.

each other. This can be caused by a deep split in the coalescent tree. However, there are major difficulties in fitting these sequences into a single gene tree as was done in the example of the Y-chromosome in Chapter 1. Many pairs of sites cannot be fitted on a single tree topology without assuming at least three mutations. The pair of sites are therefore said to be incompatible. As an example, for the Rochester population, 32% of all pairs of sites are incompatible. Genetic recombination creates incompatibilities since it allows different parts of the sequence to have topologically different trees. Another sign that recombination has occurred in this data set is the pattern of LD over the sequence. LD is a measure of non-random association of alleles at different sites and is therefore indirectly a measure of the correlation of genealogical trees for different segregating sites. Figure 5.2 shows which pairs of sites are in strong (statistically significant) LD. There is a weak tendency that highly significant LD is found for sites close to each other. Indeed, Figure 5.3 shows that LD is smaller the further apart the segregating sites are. Recombination leads to this pattern since sites far apart experience more recombination events between them than sites close together. Thereby, genealogical trees far apart becomes less correlated than trees close together.

The decrease in correlation with distance is the whole basis for LD mapping, since a site located close to a disease causing variant shares more history with the variant site than sites further away.

A mismatch distribution for the data set is shown in Figure 5.4. The distribution is unimodal which is consistent with exponential growth in the population (see Chapter 4). However, calculation of Tajima's D shows a value close to zero and not a negative value as expected under growth. Thus,

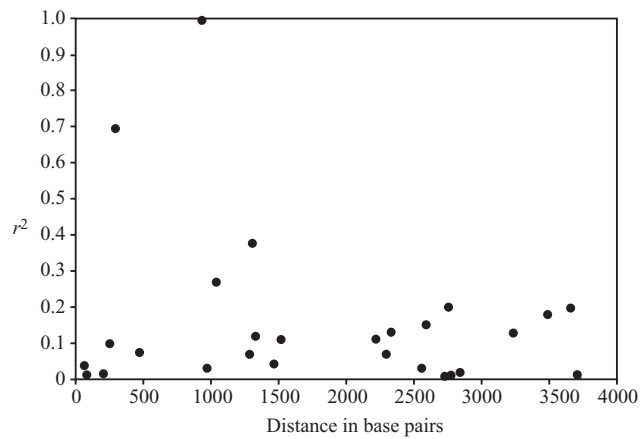


Figure 5.3 LD measured by the r^2 statistic (see Section 7.8 for a definition) using only sites with minor allele frequency 5%, in the Apolipoprotein E data set from the Rochester population.

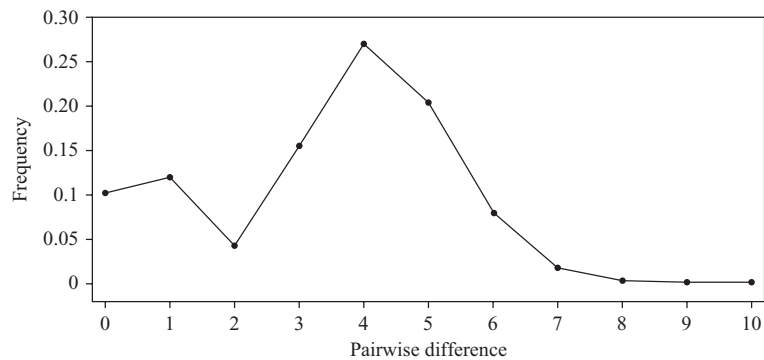


Figure 5.4 The mismatch distribution for the apolipoprotein E data set (see Section 2.5.3). If there is a strong rate of recombination the signature of a tree disappears and all pairs tend to have similar numbers of mismatches. So one should expect a slimmer distribution of mismatches around the same mean in the presence of recombination.

in this case it is more likely that recombination has caused the unimodal mismatch distribution because recombination shuffles variation between sequences, making them more equidistant.

5.3 Modelling recombination

5.3.1 Hudson's model of recombination

Hudson introduced recombination into the coalescent process and presented a simple model in 1983. Even though the model is very simple and does not

capture intricate details of the biology of recombination, it still forms the basis for most applications of coalescent theory to recombining sequences. It is very easy to simulate sequence data sets under the model. However, it is much more difficult to conduct inference under this model than under the coalescent models outlined in the previous chapters. This is because the structure needed to describe the relationship of a set of sequences is now a complicated graph, rather than a single tree. Even simple quantities concerning this graph are complicated or impossible to calculate analytically.

Hudson's model of recombination is illustrated in Figure 5.5. Choose a random point uniformly along the chromosome, then copy the genetic material from one parent chromosome to the left of this point and copy the genetic material from the other parent chromosome to the right of this point. The coalescent model has a time reversed perspective on this. Recombination splits the genetic material of a sampled sequence onto two different ancestors such that each position has exactly one ancestor. In this respect recombination events are the opposite of coalescent events that combine two sampled sequences into one ancestor. Hudson (1983a) formulated the coalescent process with recombination as competing exponentially distributed, and independent, waiting times for coalescent and recombination events. The parameter of the exponential distribution determining the coalescent intensity depends on the number of ancestors carrying ancestral material to the sample, whereas the parameter of the exponential distribution for the intensity of recombination depends on the recombination

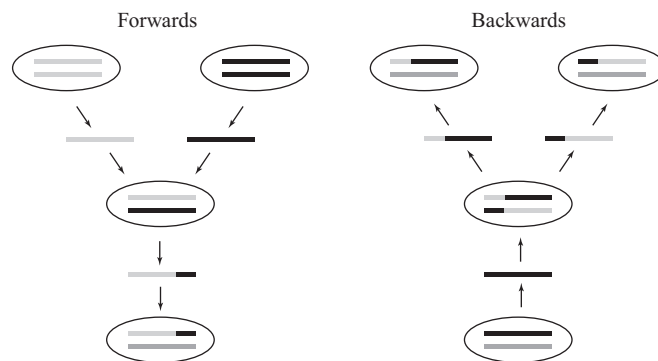


Figure 5.5 Hudson's recombination model on a continuous representation of a sequence. A sequence is made by recombination when an individual creates a haploid genome (sperm cell or egg). Looking forwards, two sequences are recombined into one recombinant sequence. Knowing the allelic states of the grandparent's chromosomes determines one of the child's two chromosomes; the other, the dark grey, originates from the second set of grandparents. Looking backwards, an individual chooses a chromosome from a parent. This chromosome is split onto two grandparental chromosomes. The child's dark grey chromosome is inherited through the other parent and the dark grey chromosomes in grandparents have unknown allelic states.

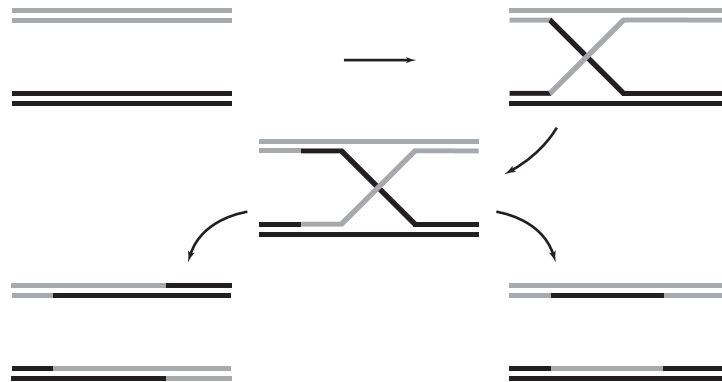


Figure 5.7 Recombination, gene conversion and the Holliday structure. In the first case a gene conversion with exchange of flanking region happens; in the second a pure gene conversion happens. The former is called a recombination event, the latter a gene conversion event. Only one of the two double-stranded DNAs is transmitted to the offspring.

double helices) of the Holliday structure, the creation of heteroduplex, the potential mismatch repair mechanism and the timing of the subsequent DNA replication, different daughter molecules will be created. What is of interest is the ancestry of individual nucleotides in the daughter molecules with respect to the parent nucleotides. Resolution of the Holliday structure after strand migration can occur either as a gene conversion event with crossing over (Figure 5.7 lower left) or as a gene conversion event without crossing over. Henceforth, we will reserve the phrase ‘recombination event’ for the former type of event, and the phrase ‘gene conversion event’ for the latter.

Holliday’s model is simplistic and more realistic models have been proposed (see Lewin 2003, Kauppi et al. 2004). However, Holliday’s model basically captures how recombination and gene conversion are believed to take place. Other models often lack the symmetry of the strand exchange shown in Figure 5.7.

5.3.2.3 *Relative rates of gene conversion versus recombination in eucaryotes*

The importance of gene conversion versus recombination in eucaryotes depends on the relative rates of single strand versus double strand breaks and the repair mechanisms. The relative frequency of gene conversion to recombination is dependent on how the Holliday junction is resolved. A random choice of cutting two of the four single strands that would create two homologous DNA sequences would lead to equal frequency of the two. In reality gene conversion can be much more frequent than recombination.

from the male set and one haploid from the female set. This process is illustrated in Figure 5.9. We consider a small region (sequence) within the genome. A recombination event takes place in the region with probability r , in which case a point is chosen uniformly along the paternal and maternal sequences and they recombine in that point. The existence of the gene conversion is ignored.

If this is viewed backwards in time, the effect of recombination will be that the ancestral material to a specific DNA sequences is found on two DNA sequences in the parent, which again came from two different grandparents, etc. In the generation before a sequence was created by recombination, there would have been one more sequence carrying ancestral material than after.

If we focus on a single point on the sequence, it will be inherited from one parent only, thus the Wright–Fisher model with recombination reduces to the Wright–Fisher model without recombination for each point on the sequence, but different points on the sequence are correlated instances of the Wright–Fisher process without recombination. The tree relating the sequences in a single position is called the *local tree* of that position. Thus, the genealogy of the whole sequence can be seen as a collection of local trees, one for each position.

In analogy with the Wright–Fisher model without recombination, we may ignore the existence of individuals and describe the process as acting only on individual sequences as illustrated in Figure 5.10. When the number of sampled sequences is small relative to the population size, and inbreeding is limited the quantitative effects of ignoring which sequences are present in the same individual are negligible. The only difference to the genuine haploid model is that each sequence must choose two parents and that a sequence is created after letting the parent sequences recombine (if they do). If the recombination rate is low, there is a high probability of no recombination and the new sequence will be a direct copy of one of the two parent sequences.

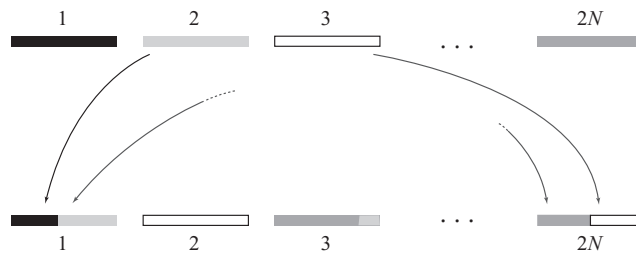


Figure 5.10 The haploid Wright–Fisher model with recombination. Each haploid individual chooses two parents that make the individual’s chromosome by recombination. Real biological individuals can be made by packing pairs of haploid individuals together. Describing the process as a purely chromosome focused process has negligible consequence for the actual probabilities involved. Recombination happens with probability r .



Figure 5.11 In a large population a sequence created by recombination is always made up of two sequences that have no other descendants in the sample. Thus, the left part of the figure is possible, whereas the right part is impossible because one ancestral sequence is involved in both a recombination event and a coalescence event at the same time.

5.5 Algorithms

Recombination and coalescent are competing processes that determine the graph structure of the genealogy of the sample, with recombination causing splitting and coalescent causing merging of sequences when looked at backwards in time. Analogous to the definition of the scaled mutation rate θ , the scaled recombination rate is defined as $\rho = 4Nr$. It is convenient to represent a sequence by a continuous interval (infinite sites model) of length $\rho/2 = 2Nr$, i.e. the expected number of recombination events in the population in one generation. ρ is called the scaled recombination rate or the population recombination rate. There are two alternative algorithms to construct the coalescent with recombination process, Hudson's back-in-time algorithm and Wiuf and Hein's (1999b) spatial algorithm. They illustrate different aspects of the process. Hudson's algorithm is simplest and most useful in the majority of applications.

5.5.1 The ancestral recombination graph

Assume the history of n sampled sequences is being described going backwards in time and that the first event encountered is a recombination event. How will that affect the number of ancestors to the sample and the distribution of genetic material ancestral to the sample? Before the recombination event (i.e. closer to the present), there were n sequences each carrying the ancestral material to the n sequences in the sample. After the recombination event (further back in time), one of the sequences had two ancestor sequences—one carrying ancestral material to the left of the recombination break point and one carrying ancestral material to the right of the recombination point.

If time is measured discretely in generation, the time until a recombination event occurs is geometrically distributed with parameter $r = \rho/4N$. Tracing a sequence back in time, the probability that it was created by recombination j generations back is

$$P(T_R = j) = r(1 - r)^{j-1}, \quad (5.1)$$

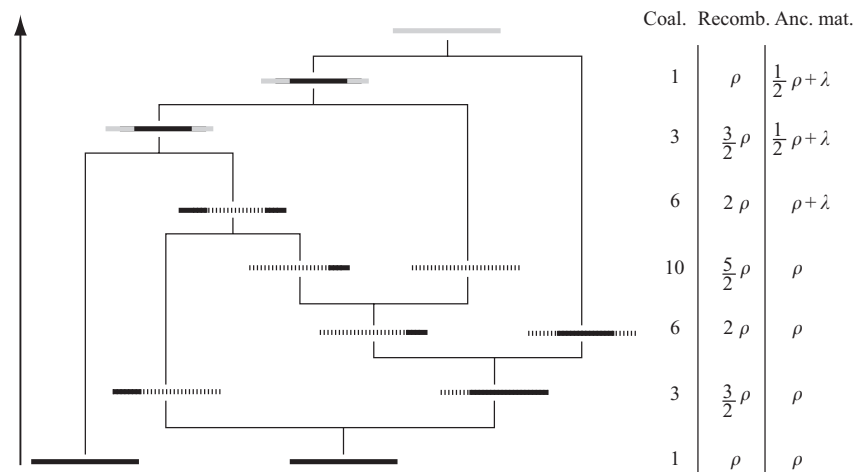


Figure 5.12 Black lines represent the sample sequences or ancestral sequence material to these. Dotted lines represent non-ancestral material. Light grey lines indicate that a MRCA has been found. The non-ancestral piece formed after the first coalescence event between two non-consecutive pieces of ancestral material is trapped material. Also shown is the rate of coalescence and recombination, and the amount of material spanned by ancestral material. λ is the length of the black bar in the sequence with dashed ends.

either a coalescent event (with rate $k(k-1)/2 = 1$) or a recombination event (with rate $k\rho/2 = \rho$) could occur. In this example the first event is a recombination event. After the event there are three sequences with ancestral material to the two sampled sequences. The next two events are also recombination events. In one of the two events a sequence is created with no material ancestral to the sample. The rate of a coalescence is now 10, while the rate of recombination is 2.5ρ .

The fourth event is the first coalescent event that also traps a piece of non-ancestral material between two pieces of ancestral material. As long as the flanking regions are linked, their genealogical histories are identical, so if one segment coalesces into another sequence, so does the other. After three more coalescence events all the ancestral material from the two sampled sequences have found common ancestry, in fact have found a GMRCAs. There are two MRCA: one is also the GMRCAs which is the MRCA of the middle island of ancestral material, the other is the MRCA of the two flanking islands of ancestral material. This MRCA is created at the second coalescent event. When two pieces of ancestral material are bridged together they share fate as long as they are not cut by recombination again. The material between the two pieces is called *trapped material*.

5.5.1.2 Discrete versus continuous sequences

Real sequences have a discrete number of base pairs rather than an infinite number of sites. The infinite sites model described in the previous section can be converted to a discrete model by dividing the continuous interval

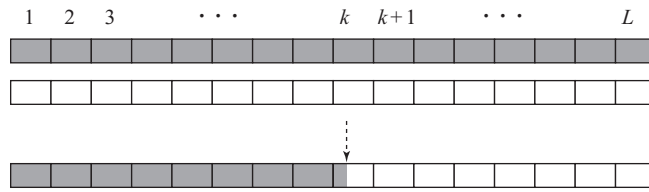


Figure 5.13 Hudson's recombination model in a continuous representation of a sequence with the discrete structure superimposed. The probability of a recombination within a specified dinucleotide is small—about 10^{-8} in humans per generation. In the continuous representation a sequence 1000 bp long in a population with effective population size of 10,000, will then be $\rho/2 = 2 \cdot 10,000 \cdot 1000 \cdot 10^{-8} = 0.2$ long. A recombination event within the k th nucleotide counts as a break between the k th and the $(k + 1)$ th nucleotide in the discrete model.

into equally sized fragments corresponding to nucleotides. The difference between the models is illustrated in Figure 5.13. Thereby there is a constant scaled rate of recombination ρ/L per nucleotide where L is the number of nucleotides. For small sequences of length L , r is approximately Lr_0 , where r_0 is the probability of a recombination per generation per nucleotide. We will continue to use the infinite sites model if not stated otherwise.

5.5.1.3 Improvements of the basic algorithm

The ARG just described is simple but in many cases unnecessarily time consuming to simulate. In particular when the recombination rate is relatively high. There are two reasons for this:

1. Any given point on the sequence may have reached a MRCA long before the GMRCAs.
2. The recombination break point may create ancestors that do not carry material ancestral to the sample.

It is straightforward to make the algorithm more efficient by adjusting for either of these factors. The first factor can be taken care of if a record of the number of ancestors of each position is kept. As soon as a position has found a MRCA it is no longer counted as ancestral material, but as non-ancestral. This reduces the time complexity of the algorithm considerably.

The second factor can be dealt with in several ways. The simplest way is to keep the recombination rate at $k\rho/2$. However, once a recombination occurs, it is recorded whether both recombining sequences carry material ancestral to the sample or not. If only one does, the other can be discarded and the number of ancestral sequences is unaffected by the recombination event. As a consequence, the recombination event does not affect the genealogy at all. An even more efficient way is to make sure that only

recombination events that change the distribution of ancestral material are feasible. To do so one keeps track of the ancestral material present on all sequences. A recombination event only affects the genealogy if it occurs in the region spanned by the left and right endpoints of ancestral material on the recombined sequence. This region might include segments of non-ancestral trapped material as well (see Figure 5.12). The intensity, A , of a recombination event is $0 < A \leq k\rho/2$, where A is the sum of material spanned by left and right endpoints of ancestral material.

For recombination rates $\rho > 10$ much efficiency is gained by taking both factors into account, but the efficiency is not greatly increased by keeping track of A as compared to just keeping track of whether all ancestors carry ancestral material or not. An ARG trimmed according to the two factors (and keeping track of A) is called Hudson's algorithm or Hudson's graph. It is not possible to reduce the algorithm further.

5.5.2 Sampling ARGs: Not back in time, but along sequences

The coalescent process has been described as a process starting with n leaves and then coalescing pairs of sequences until a MRCA sequence has been reached. There is an alternative algorithm that moves along the sequences and modifies the genealogy as recombination break points are encountered (Wiuf and Hein 1999b). The basic idea is the following: (1) simulate the genealogy for the first position in the sequences; (2) find the first break point as one moves towards the right end of the sequences; (3) choose the sequence that undergoes recombination and modify the genealogy accordingly. Intuitively it is clear that there is a formulation of the coalescent with recombination fulfilling (1)–(3) because we could run through all local trees starting at the left endpoint moving rightwards. However, it is less obvious how the algorithm probabilistically should be formulated. The details are given in Wiuf and Hein (1999b); here we will only elaborate on the intuitive formulation of the algorithm.

It works by building up a graph stepwise, a graph that is embedded in the ARG and that contains Hudson's graph. The graph for a given position t in the sequences is called the *local graph* for t and it has the local tree for t embedded. The local graph of position zero is always the local tree but for all other positions the local graph might differ from the local tree. The local graph for t is described relative to the starting point of the sequences. Thus if the starting point is moved (e.g. ignoring a part of the sequences) the local graph for t is given relative to the new starting point.

The algorithm is illustrated in Figure 5.14 with three sequences. A genealogy is sampled for three sequences that describes the relationship of the sequences in the left-most point, zero. This is the local graph in position zero which is an ordinary coalescent tree. Now we would like to move along the sequences scanning for the first break point, that is, the first point

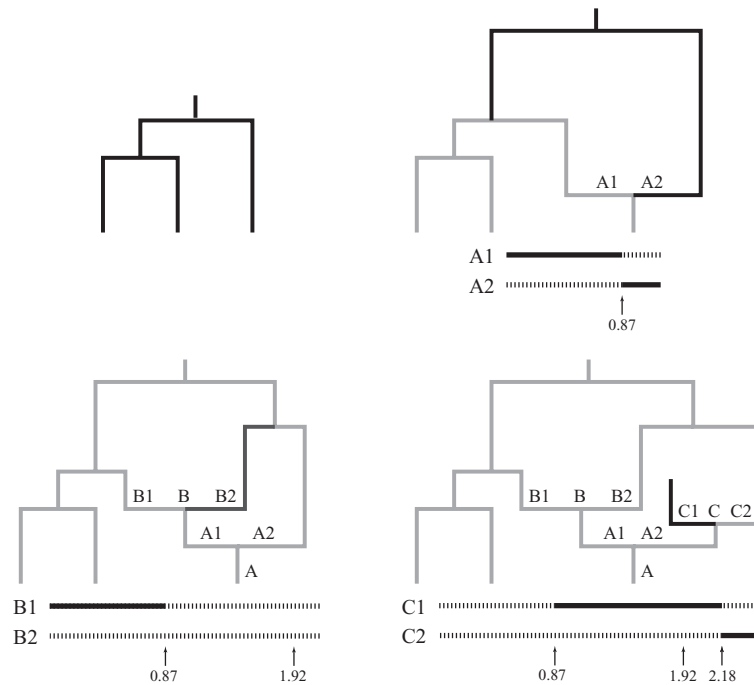


Figure 5.14 The spatial algorithm. The steps are explained in detail in the text. Events that do not contribute to the genealogy of the sequences might be included in the graph created by the spatial algorithm. A graph with three events A, B, and C, is shown creating sequences A1, A2, B1, B2, C1, and C2, respectively. The second recombination event creates an ‘empty’ sequence: If C2 created after the third event merges with B2 then event B becomes part of the genealogy, whereas if C2 merges with anything else B is not part of the genealogy.

described by a different genealogy than the one describing position zero. Where this point is, depends on the total length of the local graph (tree) in position zero a large graph spans many more generations than a small graph (tree) and the first break point would be closer to 0 in a large graph than in a small graph (Figure 5.15).

In the example, assume the total branch length is 1.8. A variable is taken from an exponential distribution with 1.8 as intensity parameter—here the outcome of the variable is 0.87. Now choose a uniformly random point on the first local tree and postulate a recombination at that point. All positions from 0 to 0.87 ($0 \leq t \leq 0.87$) share the same local tree or local graph; from position 0.87 ($0.87 < t$) the graphs are different. The newly created sequence coalesces with the local graph for position 0.87. In the example the sequence coalesces with the root sequence and the local graph in position at 0.87 is not a tree but a graph. Assume the total length of all branches in the local graph for 0.87 is 3.3. The steps are now repeated. An exponential variable with intensity 3.3 is drawn—here 1.05—and adding it to 0.87 to find the location of the next break point. Now choose a recombination event

which is true at least for large n and/or ρ . Now the numbers in equation (5.6) and (5.7) can be quite different, for example, $E(R_n^*)$ increases at an exponential rate with ρ , whereas $E(R_n)$ is linear in ρ . It is thus natural to expect that Hudson's algorithm is advantageous computationally. This is indeed the case: Let E be the total number of events in the graph (irrespective of what algorithm is used), then

$$E = 2R + n - 1, \quad (5.8)$$

where R denotes the number of recombination events. This number depends on the chosen algorithm (and is a stochastic variable in itself). Each recombination event creates a new sequence and an extra coalescent event is required to complete the genealogy, thereby adding two events to E . The number $R_n + R_n^T$, where R_n^T is the number of recombination events in trapped material, is always a lower bound to R because all events in trapped and ancestral material modify the genealogy. Thus

$$2R_n + 2R_n^T + n - 1 \leq E \leq 2R_n^* + n - 1. \quad (5.9)$$

The ARG has $R = R_n^*$, whereas Hudson's algorithm has $R = R_n + R_n^T$. The expectation of R_n^T has not been found explicitly, but can be bounded upwards by

$$E(R_n^T) \leq \rho(\rho + 1) \left(\sum_{i=1}^{n-1} \frac{1}{i} \right)^2. \quad (5.10)$$

The bound is crude as can be seen by comparison with $E(R_n^*)$ which depends on $\sum_{i=1}^{n-1} 1/i$ only. For increasing n and fixed ρ , $E(R_n)$, $E(R_n + R_n^T)$, and $E(R_n^*)$ increase at similar rates, whereas for fixed n and increasing ρ , $E(R_n^*)$ increases at a much faster rate than $E(R_n + R_n^T)$. Thus, there is a huge gain in time spent on computation for large ρ in choosing Hudson's algorithm instead of the ARG. The spatial algorithm, in contrast, produces more events than given by the lower bound in equation (5.9). Simulation results show that the spatial algorithm becomes computationally heavy for large ρ , though it is not known whether E increases like e^ρ .

5.6 The effect of a single recombination event

Assume that only one recombination event has happened in the history of a sample and that the break point is in p . To the left of p there will be one local tree and to the right there will be another local tree. Two neighbour trees cannot be any two trees: one must be obtainable from the other by

Type 3: The tree topologies change (see Figure 5.19). This can be defined negatively by being the remaining cases, but it might as well be defined by the number of sequences that merge with the recombining sequences before they coalesce with each other. In order for the topology to change two or more sequences must merge with the recombining sequences, before the two recombining sequences merge again.

Figure 5.20 illustrates the three types of events on a single tree. The probabilities of different categories of events can be calculated and are tabulated in Table 5.1. One surprising feature is the very high frequency of invisible recombinations for even a high number of sequences. This implies that methods trying to detect recombinations by detecting change in tree topologies will miss the majority of recombinations. The probability of an invisible recombination event will go to zero as the number of sequences increases, but very slowly.

These probabilities can be calculated using simple combinatorial arguments. The probability for an invisible recombination is the easiest. Assume the recombining sequences are created while there are $k \leq n$

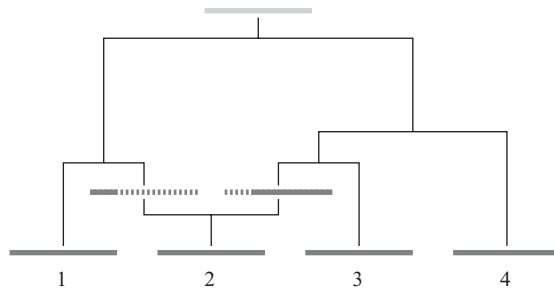


Figure 5.19 Genealogical history with one recombination resulting in different tree topologies on each side of the recombination spot. Dark grey: ancestral material, light grey: the MRCA, dotted lines: non-ancestral.

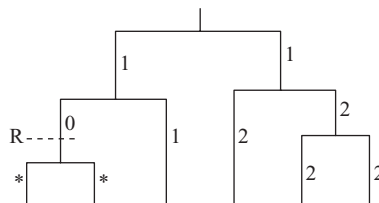


Figure 5.20 The genealogical consequences of recombination and recombination. If the recombination happens at the edge labelled 0, recombination to this edge will not change tree or branch lengths, recombination to the branches labelled 1 will give the same topology, but change some branch lengths. Recombination to the remaining branches will change the topology. It is impossible to recombine to the branches labelled ‘*’, because they are more recent than the recombination event.

A similar recursion can be written for the probability, p_{k2} , of a branch length change. The solution is

$$p_{k2} = \frac{8}{9k} + \frac{8}{(k+1)k^2(k-1)} \left(\frac{1}{3} + 4 \sum_{j=1}^{k-2} \frac{1}{j} \right). \quad (5.15)$$

The probability, p_{k3} , of a topology shift follows from $p_{k3} = 1 - p_{k1} - p_{k2}$. Closed expressions for the probabilities, P_n^2 and P_n^3 , of type 2 and 3 events, respectively, can be found similar to P_n^1 . However, they do not reduce to simple expressions.

5.7 The number of recombination events

In this section we list some moments of variables that count different types of recombination events.

As stated in equation (5.6) the expected number of recombination events in ancestral material is

$$E(R_n) = \rho \sum_{i=1}^{n-1} \frac{1}{i}. \quad (5.16)$$

The expectation is linear in ρ such that histories of sequences of double length have double the number of recombination events on average. The variance of R_n can be expressed in terms of the correlation $f_n(x)$ between the total branch lengths of two local trees separated by $x/2$ recombination units (the whole sequence is $\rho/2$ units)

$$\text{Var}(R_n) = \rho \sum_{i=1}^{n-1} \frac{1}{i} + 2 \int_0^\rho (\rho - x) f_n(x) dx_n \quad (5.17)$$

(Hudson 1983a). The last term is for large ρ of order $\log(\rho)/\rho$, thus disappearing with increasing ρ . For $n = 2$, $f_n(x)$ is known:

$$F_2(x) = \frac{18 + x}{18 + 13x + x^2}.$$

The events that count in R_n can be divided into three subtypes (Section 5.6) according to whether

Type 1: The topologies and branch lengths are the same for local trees close to the break point (Figure 5.17).

Type 2: The unrooted topologies are the same, but branch lengths differ (Figure 5.18).

Type 3: The unrooted topologies differ (Figure 5.19).

$Q(\mathbf{A}, \mathbf{M}, \mathbf{n})$ is the probability or density of the $(\mathbf{A}, \mathbf{M}, \mathbf{n})$ configuration conditional on the positions of the mutations and the beginnings and ends of ancestral segments. The first sum is over all possible coalescence events of identical sequences, the second sum is over all coalescence of sequences with the same configuration of mutations in common ancestral material, the third sum is over all mutations of multiplicity one (singletons) and the last combined sum and integral is over all possible ways to generate a sequence by recombination. This recursion is initialised by $Q(\mathbf{A}, \mathbf{M}, \mathbf{n}) = 1$, where $\mathbf{n} = (1)$, \mathbf{A} corresponds to one interval of ancestral material covering the complete sequence and \mathbf{M} corresponds to no mutations. If the scaled recombination rate ρ is zero, the second and fourth terms can be ignored and the recursion reduces to the recursion in (2.27). This initialisation corresponds to waiting until the GMRCAs, which is computationally slow. It can be initialised, when an ancestral sequence only has segments that are the ancestors to all the sequences in the sample.

This recursion is illustrated in Figure 5.21. The ancestral states of the mutations are assumed to be known. The present configuration ‘a’ has four mutations at four positions (infinite site assumption). Alleles with identical configurations of mutations and ancestral material can coalesce as shown in the ‘b’ configuration, where the third and fourth sequences from the present configuration have found a common ancestor. Sequences can also coalesce if they are identical on the ancestral segments as shown in the ‘c’ configuration, where the first and second sequences in the present configuration coalesce. The ‘d’ configuration illustrates the event of removing a mutation (the rightmost). The ‘e’ configuration shows a recombination event that splits the fifth allele in the present configuration into two new sequences. The recombination could have been anywhere in material spanned by ancestral material (also in trapped material) and all these possibilities must be integrated out.

The recursions for probability of data generated by histories with recombination are harder because they cannot be used to find the probability of the data even with infinitely much computing power as the number of ancestral states multiplies ad infinitum. Even if the number of ancestral configurations had been bounded, the necessity of integration in the last term makes the computations very hard.

5.9 The number of segregating sites

With and without recombination the expected numbers of S_n and η_i are the same, that is,

$$E(S_n) = \theta \sum_{i=1}^{n-1} \frac{1}{i}, \quad (5.23)$$

and

$$E(\eta_i) = \frac{\theta}{i}, \quad (5.24)$$

for $i = 1, \dots, n - 1$. However, the variances changes, for example, the variance of S_n becomes

$$\text{Var}(S_n) = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \frac{2\theta^2}{\rho^2} \int_0^\rho (\rho - x) f_n(x) dx, \quad (5.25)$$

where $F_n(x)$ is defined as in equation (5.17).

The variance attains the largest value for $\rho = 0$ and decreases towards $\theta \sum_{i=1}^{n-1} 1/i$ with increasing ρ . The reason behind this is that with increasing ρ , S_n sums mutations over many (almost) independent trees, thus reducing the variance.

5.10 The coalescent with gene conversion

Wiuf and Hein (2000) and Wiuf (2000a) included simple models of gene conversion into Hudson's model of recombination. It is assumed that the time to a gene conversion event is exponentially distributed (in the continuous time approximation) with a parameter $\gamma = 4Ng$, where g is the probability that a gene conversion tract initiates within a sequence in one generation.

The tract length in nucleotides is drawn from a specified distribution: Empirical evidence points to a geometric distribution with parameter $q > 0$. In the infinite sites approximation the tract length becomes exponentially distributed with intensity $Q = qL$ such that $1/Q$ is the mean length of the gene conversion tract. If a gene conversion event occurs, the first break point is chosen uniformly on the sequence, and the second break point is chosen a distance away from the first as determined by a random number from the tract length distribution. The upper part of Figure 5.22 shows how a gene conversion event distributes the ancestral material on two different ancestors. Note that a gene conversion event may only have one break point within the sequence if the tract extends beyond the end of the sequence, or if a tract initiates outside the sequence but ends within. Therefore some events will be indistinguishable from recombination events.

The time to the next event for a set of k sequences is exponentially distributed with parameter

$$\frac{k(k-1)}{2} + \frac{\rho}{2}k + \frac{\gamma Q^*}{2}k, \quad (5.26)$$

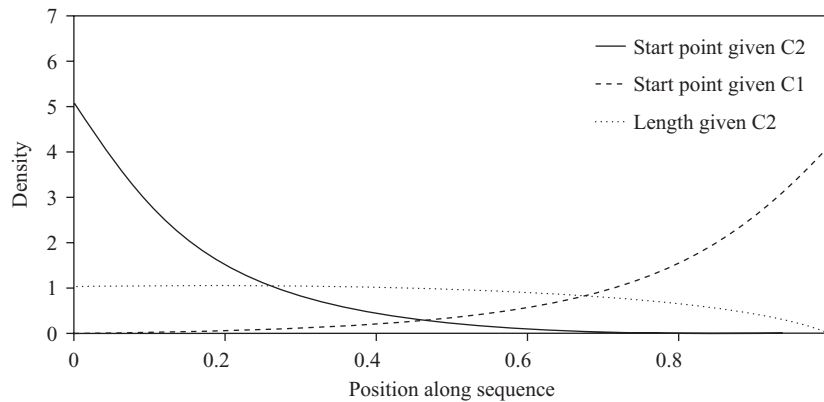


Figure 5.23 The probabilities of the gene conversion tract covering an interval. C2 is the event that the gene conversion tract is contained in the interval and C1 the event that it is not. The almost flat curve is the density of the length of the tract given C2. The strongly descending curve is the density of the start position given C2. The strongly ascending curve is the density of the start position given C1.

recombination—taking a subtree and moving it. The end of the gene conversion will look like the reverse of the start of the gene conversion—taking the subtree and moving it back again. Obviously this creates a change over short distances, but not over long distances.

5.11 Gene trees with recombination—from incompatibilities to minimal ARGs

Genealogical histories without recombinations are much easier to represent than genealogies with recombinations. This section discussed some purely combinatorial and representational questions that are encountered, when analysing data subject to recombination. A proper statistical treatment of these issues belongs to the future.

First, the phylogenetic consequences of recombination are discussed, followed by some considerations and examples of the difficulty in reconstructing recombinations. Then three central statistics of the data are discussed: The minimal number of trees compatible with the sample, which also is a lower bound to the number of recombination events in the sample's history, an improved lower bound to the number of recombination events, and lastly the minimal ARG describing the data.

Reconstructing—in a classical phylogenetic sense—a full history of a set of sequences is close to impossible. The complexities introduced by recombination enters at several levels. Within the infinite site model recombination

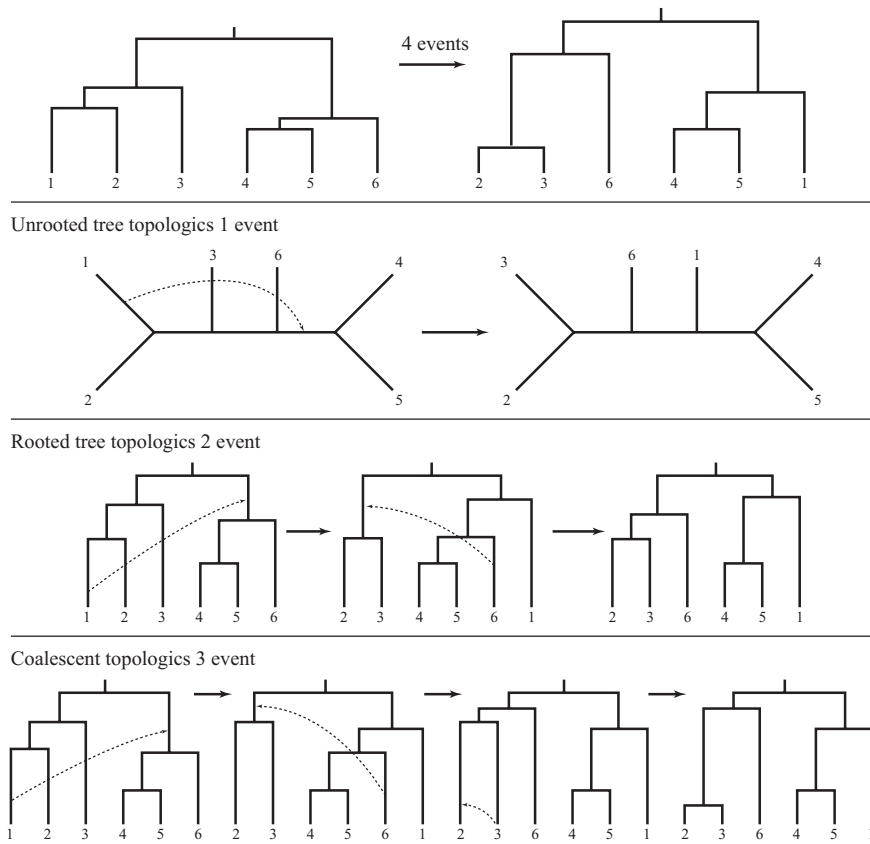


Figure 5.24 Various distance measures on trees and topologies. The unrooted topologies are one subtree transfer away, the rooted two, the age-ordered three, and the trees with full specification of times four transfers away.

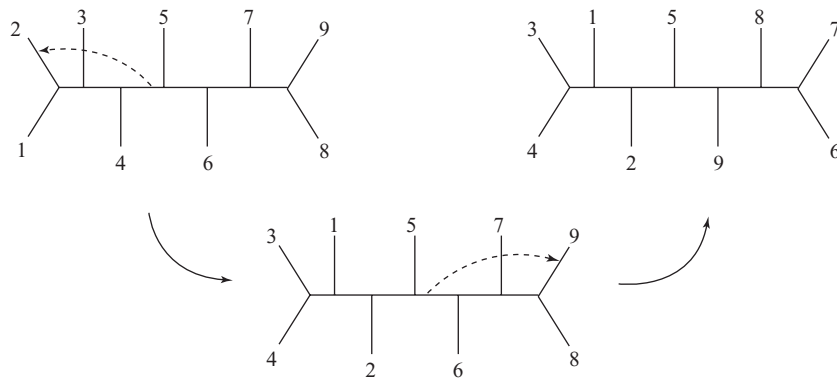


Figure 5.25 An example of where the two distance measures differ. The two topologies are two subtree transfers away. The second violates the time constraint imposed by the coalescent process. Example by Thomas Christensen.

If the genealogies are converted into two unrooted topologies, then the first can be converted into the second by one subtree transfer—by moving the little tree only containing leaf 1 over to sit outside (4, 5) (see second highest panel of the figure). It is possible to assign ages to all internal nodes and to the roots, and then the resulting genealogies would be a subtree transfer from each other. But these two genealogies would not be the same as the two original genealogies.

If the information concerning the root is retained, that is, giving rooted tree topologies, then these will be two subtree transfers apart (see second-lowest panel of the figure). The single operation used on the unrooted topologies would have six sitting on the wrong side of the root. This can be rectified by an extra transfer.

If the internal ordering and the root is kept, that is, coalescent topologies are considered, then three subtree transfers are needed (see lowest panel of the figure). The internal nodes might now be labelled (7–10), so the lower numbered labels are the more recent. If the two transfers are used from the rooted topologies above, then internal nodes 9 and 10 have not been changed. The resulting coalescent topology has the right unrooted topology, but the age ordering of 9 and 10 are wrong relative to the target coalescent topology and a subtree transfer has to be postulated to move the ancestor of 2 and 3 further down.

If the information in the complete genealogies is kept, and for instance the dates of the roots are different, then a subtree transfer would be necessary to move the root to the correct time (top panel).

In these different cases there is in general no reason to believe that the best path (a series of transfers) of the simpler problem is always a part of the path of the more complicated problem.

In the coalescent with recombination the number, ζ_i , of events that involve a transfer of a subtree of size i has mean

$$E(\zeta_i) = \rho \frac{1}{i}, \quad (5.29)$$

for $i = 1, \dots, n - 1$. Thus it is more likely that small trees are moved rather than large trees.

When dealing with sequence data the situation is different, since local trees cannot be observed. If a number of SNP polymorphisms is observed each SNP polymorphism corresponds to a partition of the sequences into two subsets. All that can be said about the genealogy underlying the polymorphism is that the topology belongs to a certain class of topologies, namely the class of topologies that all have the given partition. This implies that even less can be said about the coalescent topology that also includes an age-ordering of all coalescent events in the history of the SNP. If two topologies, T_1 and T_2 , or two classes of topologies, each corresponding to

a SNP, are compared, one has to take into account that each subtree transfer applied to change T_1 into T_2 imposes constraints on the age-ordering. Otherwise there would not be a coalescent topology (or a realisation of the coalescent process) that is in agreement with the series of subtree transfers. Figure 5.25 provides an example of two topologies where the number of subtree transfers differ according to whether age-ordering constraints are taken into account or not. A subtree can only be moved in its root and thus two subtree transfers could impose conflicting constraints. In Figure 5.25 two trees with nine leaves represented as unrooted tree topologies are two subtree transfers away from each other. But if one tries to pick coalescent topologies that would correspond to these events, it is impossible because the two subtree transfers give conflicting information about the subtrees that must exist in the coalescent topology. In most cases, and for small number of sequences, the two ways of counting the distance between topologies agree.

If an infinite sites model is assumed, the sequences cannot be represented by a gene tree if all four combinations of 0 and 1 are present in two columns. At least one recombination event is required to explain these two columns. But what is the least number, R_M (M is for minimum), of events required to explain the whole sample? What is the least number, T_M , of gene trees required to explain the sample? Are the two numbers related? In general, $R_M \geq T_M - 1$, because there must at least be one recombination event between any two trees (if there were zero events between two trees they could be collapsed into a single tree). Further, each site is compatible with a gene tree so at most S_n gene trees are required to explain the whole sample, that is, $T_M \leq S_n$. The number of recombination events is also bounded upwards: $R_M \leq (n - 1)(S_n - 1)$. If $n - 1$ of the sequences are each split into S_n fragments (each fragment is a single nucleotide) using $(n - 1)(S_n - 1)$ recombination events, a history that is compatible with the infinite sites assumption can easily be constructed. A simple example is shown in Figure 5.26. This bound is in general a crude overestimate.

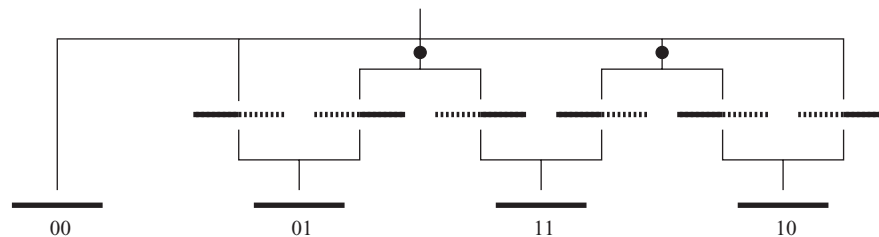


Figure 5.26 There is always a history with at most $(n - 1)(S_n - 1)$ recombination events that explains the sample. In the example $n - 1 = 3$ of the sequences are each broken into $S_n = 3$ pieces that subsequently can coalesce to the unbroken sequence.

two regions A–D and D–F. In the first site B was left out, in the second D. Note that $T_M = 4$, because A and C are incompatible, so are C and D, and E and F. It is always true that $H_M \geq T_M$.

5.11.3.1 How to find H_M

The algorithm provided here is from Myers and Griffiths (2003). Define H_{ij} by

$$H_{ij} = \max\{b_{ik} + H_{kj} \mid k = i + 1, \dots, j - 1\} \quad (5.30)$$

and

$$H_M = H_{1S_n},$$

with boundary conditions $H_{ii} = 0$, $H_{i,i+1} = b_{i,i+1}$, and $b_{ij} = 1$ if sites i and j are incompatible, and 0 otherwise. H_{ij} is the bound obtained for the regions spanned by the site i and j . For large data sets the algorithm can be quite slow because many partitions have to be tried out to find the optimal combinations of sites. This algorithm creates larger and larger intervals with a maximal combination of bounds on small segments and smaller intervals with maximal bounds on recombination events.

5.11.4 Minimal ARGs

One of the most famous data sets in the history of molecular population genetics was published by Martin Kreitman in 1983. It contained 11 sequences each about 3200 bp long of 11 alleles of ADH genes from *Drosophila melanogaster*. It had forty-three segregating sites. Twenty-eight of these are informative and could contain information in parsimony sense about recombination events. Determining the number of recombination events in this data set has been the motivation for at least four methods: A paper by Hudson and Kaplan from 1985 that produced 5 as a lower bound on the necessary recombinations events in the history of these data. Using Myers and Griffiths' H_M gave a lower bound of 6. Song and Hein (2003) produced a lower bound of 7 and subsequently Song and Hein (2004a) proved that this bound can be realised in an ARG and that no higher lower bound is possible. This minimal ARG is shown in Figure 5.30.

The method obtaining this minimal ARG is very slow indeed and cannot be used on data sets of more than 9–10 sequences and is thus not practical for large data sets at present.

The method scans the sequences and at each column keeps track of the cost of the minimal histories ending in all possible coalescent topologies in that column. Let $d(T_1, T_2)$ be the smallest number of recombination events needed to convert T_1 into T_2 , and $s(T, i)$ be the number of substitutions

and

$$R_M = \min_T \{W(T, S_n)\}. \quad (5.32)$$

$W(T, i)$ is the minimum number of recombination events required to explain the first i sites if the tree in site i is T .

The minimisation has to be over all T_1 s at the $(i - 1)$ th column and the recursion above will give a value for all possible T_2 at column i . The recursion is initialised by $W(T, 1) = s(T, 1)$. The initialisation states that the cheapest history for the first column given a given coalescent topology, is just the cost of that column using that coalescent topology. The recursion states that the cheapest history for the first i columns given that the relationship of the sequences in column i is T_2 must be the optimal combination of the history up to the $(i - 1)$ th column and the cost of adding the i th column. This cost has both a substitution cost from explaining column i using T_2 and the recombination cost of transforming T_1 into T_2 .

This minimisation algorithm has some resemblance to the spatial coalescent–recombination algorithm, but there is a crucial difference. The cost of changing to a new coalescent topology only depends on the coalescent topology in the previous column, while the corresponding probability would depend on the complete ARG for all the columns before the present column.

Applying this recursion will give a minimal set of coalescent topologies that can be combined into a single ARG. Figure 5.30 shows the above algorithm applied to the Kreitman data set, when no double mutations are allowed in any positions, which accelerates the algorithm. Recursion (5.31) has $T_M - 1$ as outcome if $d(T_1, T_2)$ is defined as $d(T_1, T_2) = 1$ if T_1 and T_2 have different topologies and 0 otherwise.

5.11.5 Topologies, recombination, and compatibility

Genealogies with recombinations pose challenging problems. The combinatorics of trees has been much studied, while similar efforts have not yet been undertaken for cases involving recombination.

It is obvious from the discussion of minimal ARGs, compatibility and the probability of tree topology changing recombination events, that detecting individual recombination events by inspection of the sample will only reveal a small fraction. Figure 5.31 show two sets of simulations for eight sequences with $\theta = 15$ and $\theta = 40$. This illustration shows the inherent difficulty in reconstructing recombination events. The dashed and solid lines are the expected number of recombination events and topology changing recombination events as a function of ρ . The lower set of *, o, and ■ are the events recovered by the minimal ARG method, H_M and T_M , respectively for simulations using $\theta = 15$. The upper set are from simulations using

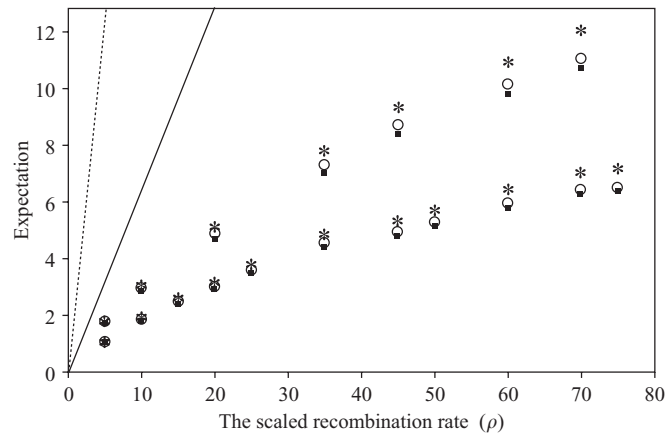


Figure 5.31 The dashed and full lines show the expected number, $E(R_n)$, of recombination events and the expected number, $E(R_n^3)$, of topology changing recombination events, respectively, for $n = 8$. The *, o, and ■ symbols show the performance of the minimal ARG method (R_M), the haplotype bound (H_M) and the Hudson–Kaplan bound (T_M) respectively, for $\theta = 15$ (lower curves) and $\theta = 40$ (upper curves). The discrepancy between reconstructed and expected number of recombination events is very large indeed. The larger the mutation rate, the larger the fraction of reconstructed recombination events. As θ goes to infinity one would expect that all topology changing recombination events would be detected. (Adapted from Song and Hein 2004a.)

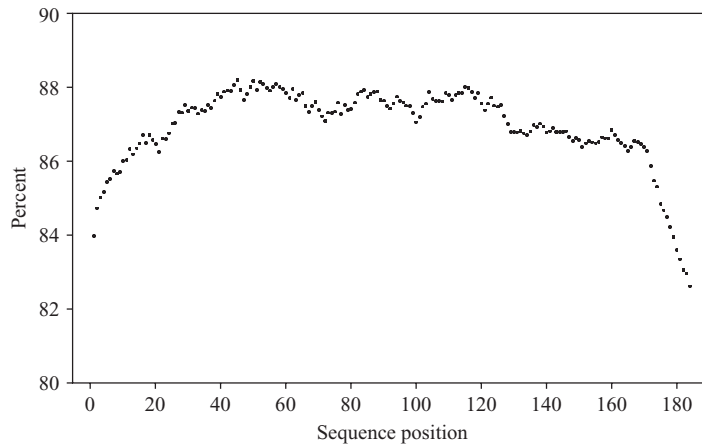


Figure 5.32 This figure is based on a sample of sequences of length $\rho/2 = 5$, (corresponding to 180 nucleotides) and 2000 simulations. In addition to the arch-shaped form due to better phylogenetic information in the middle of a sequence relative to the flanking regions, there is an asymmetry that could be caused by the dynamical algorithms choice to place necessary recombination events as rightward as possible. (Adapted from Song and Hein 2004a.)

$\theta = 40$. It is immediately seen that the recovered recombination events is only a small fraction of the total number of recombination events and also of the number of visible recombination events. It is also intuitive that a higher number of mutations will allow the detection of more recombination events.

The conclusions from investigations trying to reconstruct recombination events are thus negative, which is of course unfortunate. However, if the goal is to find the genealogical relationship between the sequences, the possibilities are better. First, a reasonable measure of similarity of genealogical histories needs to be defined. Comparing ARGs is an undeveloped topic relative to comparing classical phylogenies. If the local trees of a reconstructed ARG are compared with the local trees of the true ARG, comparing ARGs reduces to comparing trees. In Figure 5.32 one such comparison is shown for seven sequences with $\rho = 10$ and $\theta = 75$. The similarity of two trees is measured as the percentage of identical bipartitions they induce. Unrooted trees with seven leaves will have four such bipartitions corresponding to four internal edges. Again, if θ grew arbitrary large, this score would converge to 1.0. But for this simulation the average score along the sequence is around 0.85, which means that the local tree on average would have all or three bipartitions correct out of four possible. The shape of this score along the sequence is interesting as well and is readily interpretable: The local trees in the middle have more mutations to guide it towards the right tree, while at the borders of the sequence, there will only be mutations to one side.

In the approaches discussed, compatibility is a very simple and useful concept, that unfortunately catches too little of the underlying tree structure, while the method used to find the minimal ARG above seems very brute force and exhausting in testing all trees at informative positions that are compatible with that position.

Recommended reading

- Griffiths, R. C. and Marjoram, P. (1997), An ancestral recombination graph, in P. Donnelly and S. Tavaré, eds., *Progress in Population Genetics and Human Evolution*, Springer Verlag, pp. 257–270.
- Hudson, R. R. (1983a), ‘Properties of a neutral allele model with intragenic recombination’, *Theor. Popul. Biol.* **23**(2), 183–201.
- Hudson, R. R. and Kaplan, N. L. (1985), ‘Statistical properties of the number of recombination events in the history of a sample of DNA sequences’, *Genetics* **111**(1), 147–164.
- Lewin, B. (2003), *Genes VIII*, Oxford University Press.