A Simulation Study of the Reliability of Recombination Detection Methods

Carsten Wiuf, Thomas Christensen,† and Jotun Hein‡*

*Department of Statistics, University of Oxford, United Kingdom; and †Department of Computer Science and ‡Institute of Biological Sciences, University of Aarhus, Denmark

There exist many methods to detect recombination or mosaic structure in a sample of DNA sequences. But how reliable are they? Four methods were investigated with respect to their power to detect recombination in simulated samples with different amounts of recombination and mutation. In addition, we investigated the impact of the shape of the underlying genealogy on their performances. We found that the methods detected far fewer recombinations than were theoretically possible and that methods based on the principle of incompatibility in general had more power than methods that did not make use of this principle explicitly. This seemed, in particular, to be the case for phylogenies generated under population expansion scenarios which result in long branches at the tips and small deep branches. In addition to the results obtained through simulations, a series of new theoretical results on recombination is presented.

Introduction

Most phylogeny reconstruction methods assume that there is a single underlying tree relating a set of homologous sequences. However, if the sequences have experienced recombination, there might be not just one tree, but a whole collection of trees, with each tree describing the history of a unique part of the alignment. Such sequences are said to have a mosaic structure (Maynard-Smith 1992); different parts of the alignment are likely to show different patterns of variation because they have different histories. If the sequences are analyzed using phylogenetic tools that do not take recombination into account, misleading or incorrect conclusions are likely to be drawn, and mosaic structures might wrongly be ascribed to evolutionary forces other than recombination.

Over the last two decades, many methods have been proposed to detect mosaic structures caused by recombination (see Crandall and Templeton [1999] for a review). The problem has been attacked at several distinct levels. At a first level, one could be interested in reporting whether an observed sample of sequences has experienced recombination in its history. This is a simple yes-or-no question—has recombination occurred? Several methods attack the problem at this level only (e.g., Sawyer 1989; Maynard Smith and Smith 1998). Confirming that recombination events have occurred in the sample's history, one can then go on to ask where the break points are located along the sequences. This problem is considerably harder and has also been addressed, e.g., Hein (1993) and Weiller (1998). Between these two levels of approach are methods that consist of manual inspection and division of the sequences into smaller regions (Stephens [1985], Maynard Smith [1992], and Jakobsen and Easteal [1996], among others), which are then tested for the presence of recombination. At a final level, one can attempt to reconstruct the entire history of the sample. In the parsimony approach taken by Hein (1993), the most parsimonious history is constructed assuming that the cost of a substitution compared with that of a recombination event is known. Recently, McGuire, Wright, and Prentice (2000) developed a similar method based on hidden Markov models in a Bayesian framework.

Each recombination event breaks the sequence alignment into two parts such that the left part of the alignment has a phylogenetic history potentially different from that of the right part (fig. 1). A recombination event can only be detected if the histories of the left and right parts are different. If the two recombining lineages coalesce before merging with any other lineage in the sample's history, no trace of the recombination event is left. On the other hand, if one (or both) of the recombining lineages merges with a nonrecombining lineage before the two recombining lineages coalesce, then the phylogenies on each side of the break point differ. As a consequence, the probabilities of a polymorphic site are not the same in the two phylogenies, and the recombination event is, in principle, detectable from sequence data.

This feature has been used in the methods by Sawyer (1989), Maynard Smith (1992), Grassly and Holmes (1997), and Weiller (1998) among others, but in very different ways. For example, the approach by Maynard Smith (1992) is based on the empirical distribution of polymorphic sites in the sample, whereas the method by Grassly and Holmes (1997) is based on statistical modeling of sequence evolution.

Other methods are within the framework of compatibility (Le Quesne 1969; Sneath, Sackin, and Ambler 1975). A site is compatible with a tree if the observed characters can be explained by $c - 1$ substitutional events, where c is the observed number of different characters in the given site. If more substitutional events are required, the site is said to be incompatible with the tree (fig. 2*A*). It is easy to check if there exists a tree such that two given sites are both compatible with the tree. If two given sites are not both compatible with the same tree, the incompatibility can be caused either by recurrent substitutions in one (or both) of the sites or by recombination (fig. 2*B*). The pair of sites is then said to be incompatible. The methods by Stephens (1985), Hein (1993), Fitch and Goodman (1991), Jakobsen and Easteal (1996), and Jakobsen, Wilson, and Easteal (1997),

Key words: DNA sequences, incompatibility, recombination, simulations.

Address for correspondence and reprints: Carsten Wiuf, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, United Kingdom. E-mail: wiuf@stats.ox.ac.uk.

Mol. Biol. Evol. 18(10):1929–1939. 2001

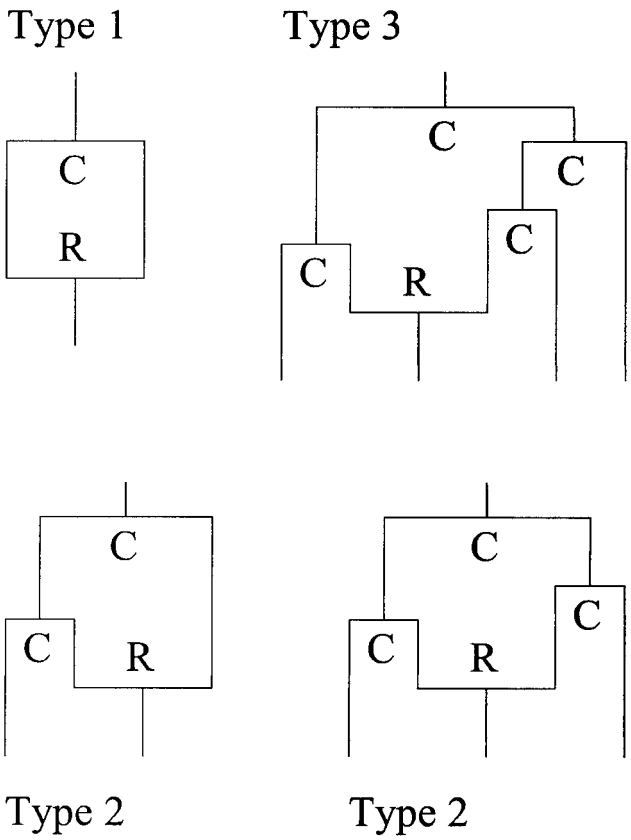

FIG. 1.—We distinguish three types (types 1, 2, and 3) of recombination events, which are shown in the figure (C = coalescence event; R = recombination event). Type 1: If the two recombining sequences coalesce before coalescing with any other lineage, the recombination event is undetectable; sites on both sides of the recombination break point share topology and branch lengths (left top corner). Type 2: One or two sequences merge with one of the two recombining sequences before the two recombining sequences merge (bottom two figures). The topology describing the sites to the left of the break point and the topology describing the sites to the right are identical. Branch lengths, however, differ. Type 3: If $i \geq 2$ sequences merge with the recombining sequences before they merge, then there are two different topologies. The genealogy in figure 2*B* depicts a type 3 recombination event which gives rise to two different topologies.

among others, are all based on the concept of compatibility. The approaches to the problem are, however, very different. For example, Hein (1993) and Fitch and Goodman (1991) develop parsimony procedures, whereas Jakobsen and Easteal (1996) develop a method based on comparisons of all adjacent columns in the sequence alignment.

In this paper, we discuss a selection of methods and their performances. We are interested in evaluating the power of the methods based on randomly generated samples and sample histories. Rather than giving a full treatment of all available methods, we aimed at simplicity and chose four methods that covered the range of levels of ambition and different underlying frameworks.

Materials and Methods

Methods

Four different methods of detecting the presence/absence of recombination were chosen for evaluation:

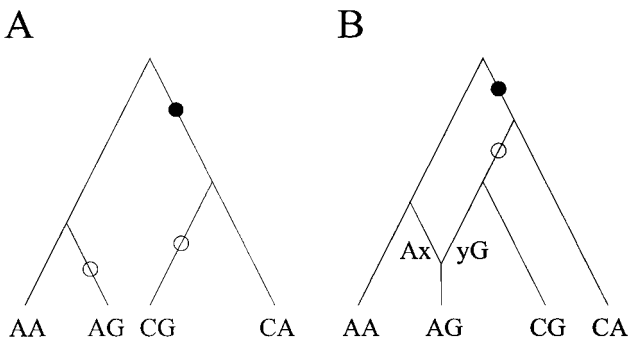

FIG. 2.—Compatibility/incompatibility. Four sequences are given, each consisting of 2 nt. The sample comprises four different haplotypes; in each site two different characters are present. *A* and *B*, Two possible genealogies explaining the history of the four sequences. The observed characters (A and C) in the first position can be explained by a single substitution (black dot) in *A* and hence are compatible with the tree. The characters in the second position cannot be explained by less than two substitutions (open circles) and are not compatible with the tree. In *B*, recombinations are allowed to take place and the observed pattern can be explained by two substitutions in total. The characters x and y in the two recombining sequences, Ax and yG, can be any of the four possible characters. Because no tree exists (without recombination) such that the four sequences can be explained by two substitution events, the two sites are incompatible.

those of Sawyer (1989), Maynard Smith (1992), Hein (1993), and Jakobsen and Easteal (1996). Of these methods, the first two are based on the distribution of polymorphic sites, and the latter two are based on the distribution of incompatibilities. All of these methods were developed to detect recombination in samples with high variation, (e.g., viral or bacterial samples, low recombination) but vary in their ambition levels as described in the introduction. For details and examples of their performances on real data, we refer readers to the original papers or the discussion of such methods in general by Crandall and Templeton (1999).

Sawyer (1989) developed the (inner) SSCF method to detect recombinations or gene conversions without reference to the history of the sample or where on the sequences the recombinations occurred. This method detects the presence of recombination by the appearance of long tracts of identities among pairs of sequences.

Maynard Smith (1992) proposed the Max Chi-Squared method, which searches for recombination break points by comparing the number of segregating sites on both sides of a putative recombination break point in a pair of sequences with the number of segregating sites in the rest of the sequences in the sample. The window size used in Max Chi-Squared was 300 (2 $\times$ 150). This gave the best power overall.

Jakobsen and Easteal's (1996) neighbor similarity score (NSS) method uses pairs of informative sites. It detects recombination by the tendency of neighboring positions to be more compatible than sites that are farther apart.

Hein (1993) developed RecPars, a parsimony algorithm to detect shifts in evolutionary history along the sequences. RecPars minimizes a combined cost of recombinations and substitutions necessary to explain a data set. The cost of a recombination (d) was set to 1.5,

Table 1
The Effects of Growth on Trees

		Tree Height				Branch Length			
n[a]	β[b]	2	3	$n-1$	n	2	3	$n-1$	n
5.....	0	62.5	20.8	10.4	6.2	48.0	24.0	16.0	12.0
	5,000	11.1	10.1	14.4	64.4	5.1	7.0	13.4	74.4
10.....	0	55.6	18.5	1.5	1.2	35.3	17.7	4.4	3.9
	5,000	8.6	5.9	12.0	46.7	2.2	2.3	14.0	60.2
15.....	0	53.6	17.9	0.6	0.5	30.8	15.4	2.4	2.2
	5,000	8.0	5.2	11.1	37.5	1.5	1.4	14.3	51.6
25.....	0	52.1	17.4	0.2	0.2	26.5	13.2	1.2	1.1
	5,000	7.6	4.8	9.9	27.0	0.9	0.9	14.3	40.4
50.....	0	51.0	17.0	0.0	0.0	22.3	11.2	0.5	0.5
	5,000	7.3	4.5	7.6	14.9	0.5	0.5	12.6	25.3

Note.—The columns under the Tree Height heading show the ratio (%) of the expectation of the time while there are k lineages, W_k, to the expected total tree height, $\Sigma_j W_j$, for $k = 2, 3, n-1$, and n. The columns under the Branch Length heading show the ratio (%) of the expectation of the branches while there are k lineages, kW_k, to the expected length of the entire tree, $\Sigma_j jW_j$, for $k = 2, 3, n-1$, and n.

[a] Sample size.

[b] Growth rate.

and that of a substitution (s) was set to 1 (based on initial simulations of sequences under different scenarios).

Evaluating Power

A permutation test to evaluate the power of the methods was used. Assume that $M(\rho)$ is a stochastic model of sequence evolution, and let ρ denote the recombination rate, where $\rho = 0$ implies no recombination (i.e., all sites have the same history), and $\rho = \infty$ implies that all sites have independent histories. The power, defined as the probability of rejecting the hypothesis H_0 of no recombination given the true model $M(\rho)$, $0 \leq \rho \leq \infty$, was assessed by comparing the output of a method under $M(\rho)$ with output from permutations of the alignment.

Unless $\rho = 0$ or $\rho = \infty$, the distribution of the permuted alignment will be different from the distribution of the original alignment. If there is no recombination ($\rho = 0$), all sites have the same history, and if all sites are unlinked ($\rho = \infty$), all sites have independent histories. In both cases, the distribution of the alignment is invariant under permutations. Thus, comparison of the output of a method under $M(\rho)$ with output from permutations of the alignment provides a test of H_0 jointly with the hypothesis that all sites are unlinked, H_∞, given the true model $M(\rho)$. The power is expected to increase from α%, the chosen significance level, for $\rho = 0$ until a certain point, and then to decrease to α% again for $\rho = \infty$.

In comparing the output of a method under $M(\rho)$ with output from permutations of the alignment, the essential assumption is an overall rate homogeneity across sites, that is, that no regions evolve faster than other regions. No specific model of sequence evolution (e.g., the Jukes-Cantor model) or a specific model of the phylogenetic process (including recombination) is assumed. This makes it attractive compared with the traditional method of comparing the output under $M(\rho)$ with output under $M(0)$, where a specification of $M(0)$ is required.

Simulation Algorithms and Theory

Two different setups were used to simulate sample histories. In the first setup, we used the coalescent process with recombination and exponential growth (Hudson 1983; Slatkin and Hudson 1991). The coalescent process emerges in a variety of contexts (Kingman 1982*b*), has been used in studies of viral data (e.g., Pybus, Holmes, and Harvey 1999; Rodrigo and Felsenstein 1999), and provides a natural foundation for simulation. However, the main objective of this simulation approach is to provide a stochastic tool to generate sample histories with shapes that depend on the choice of parameter values. There are two parameters: the recombination rate ρ and the growth rate β. In the population genetic context, $\rho = 2Nr$ and $\beta = Nb$, where N is the effective population size, r is the probability of a recombination per sequence per generation, b is the growth rate of the population per generation, and time is measured in units of N generations (Hudson 1983; Slatkin and Hudson 1991). In particular, if $b = 0$ ($\beta = 0$), the population is of constant size. Details of the simulation algorithm can be found in Hudson (1983) and Griffiths and Tavaré (1994).

Consider a single site. Let W_k, $k = n, n-1, \ldots, 2$, be the times between successive coalescent events; that is, W_k is the time while there are k lineages in the history of a particular site. The distribution of W_k depends on β: (1) If $\beta = 0$, W_k is exponentially distributed, $\text{Exp}(k(k-1)/2)$; and (2) if β is large (e.g., $\beta = 5{,}000$), $W_k/W_n \approx 0$, $k = 2, 3, \ldots, n-1$ (Griffiths and Tavaré 1998). This imposes fundamental differences on the shape of a tree (see also table 1):

1. If $\beta = 0$, the tree has long deep branches and short branches at the tips.
2. If β is large, the tree is starlike with long branches at the tips and short deep branches.

We let recombination happen uniformly along the sequences at rate $\rho/2$ per sequence. The number of re-

combination events, $R(n)$, in the history of a sample of size n has expectation

$$E_\beta[R(n)] = \frac{\rho}{2} E_\beta[L_n] \quad (1)$$

(the proof in Hudson [1983] is for $\beta = 0$, but it can easily be extended to a general β), where $L_n = \Sigma_{j=2}^{n} jW_j$ denotes the total branch length of the tree at a single site and E_β denotes expectation under growth with rate β. If $\beta = 0$, equation (1) reduces to

$$E_0[R(n)] = \rho \sum_{j=1}^{n-1} \frac{1}{j} \quad (2)$$

(Hudson and Kaplan 1985).

In the second setup, a different technique was applied to generate samples with one recombination event only. Let γ_k be the joint rate of coalescence and recombination events. The time from when there are k ancestral lineages until there are $k + 1$ lineages (recombination) or $k - 1$ lineages (coalescence) is thus exponentially distributed with parameter γ_k (fig. 3), and times between events are independent. The probability that the recombination occurs while there are k lineages is

$$q_k = \frac{k}{\gamma_k}\left(\sum_{j=2}^{n} \frac{j}{\gamma_j}\right)^{-1}, \quad (3)$$

which is defined in analogy with the coalescent with recombination (see the appendix, where this is discussed in more detail).

Mutations were added to a sample history using a Jukes-Cantor model with mutation rate $\theta/2$ (Jukes and Cantor 1969). In the population genetic context, $\theta = 2Nu$, where u is the probability of a mutation per sequence per generation. A nucleotide was assigned to the most recent common ancestor of each site by choosing randomly among the four different types. Whenever a substitution occurred, the substituted nucleotide had equal chances of being either one of the other three nucleotides. Thus, samples of sequences were generated under very simple conditions: one substitution rate along sequences and no constraints on this rate due to different kinds of substitutions, i.e., transversions versus transitions or synonymous versus nonsynonymous changes.

The first setup reflects the situation in which a real data set is under scrutiny. In the second setup, the phylogenetic signal from a single recombination event is explored.

Details of Setups

In all simulations, sequence length L was fixed at 1,000. Sample size n, mutation rate θ, recombination rate ρ, and growth rate β or the rates γ_k were varied. Between 1,500 and 2,800 samples of sequences and sequence histories were simulated for each choice of parameters, and the four methods were applied to all of the simulated samples. For each of these simulated samples, 200 permutations of the columns in the sequence alignment were performed, and the methods were run on the permuted data sets. For RecPars, 200 histories were simulated and 200 permutations performed due to a very time-consuming algorithm in RecPars. It was reported how many times the outcome from simulated samples deviated significantly at a 5% level from the outcome of permuted data sets.

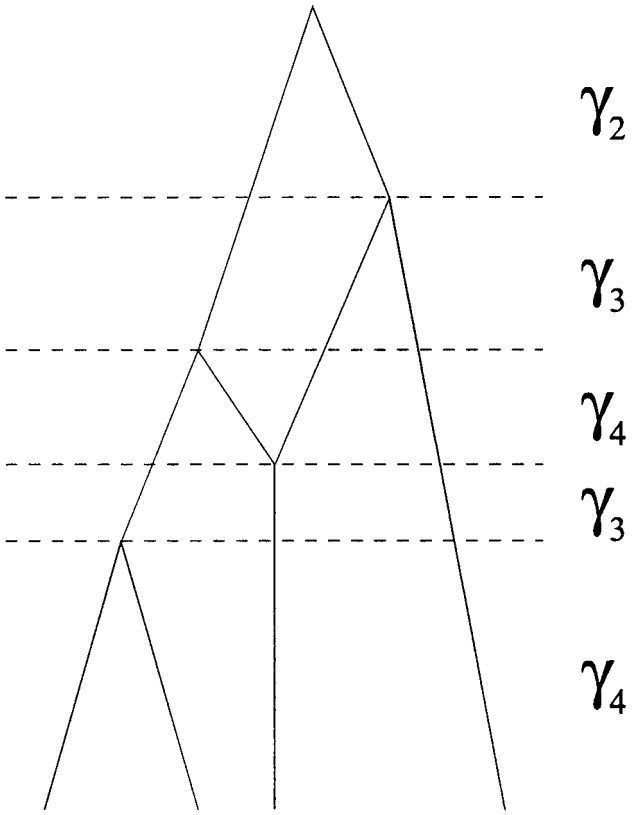

FIG. 3.—Example of a genealogy with one recombination. In this example, $\gamma_k = 1$ for all k, and the sample size is 4. The first event is a coalescence event, and the second is a recombination event. The time until all sites share a most recent common ancestor is (in this case) given by a sum of five exponential variables with rates 4, 3, 4, 3, and 2, respectively.

Setup 1: Histories with a Random Number of Recombination Events

In one set of simulations, n was fixed at 10, and θ was varied such that the expected sequence divergence p between two sequences was $p = 1\%$, 2.5%, 5%, 10%, and 20%. In a second set of simulations, θ was chosen such that $p = 10\%$ and n was varied over 5, 10, 15, 25, and 50. Thus, two sequences differed on average in p sites, but due to recombination, the distribution of polymorphic sites was not necessarily uniform along the sequences. Two choices of β were considered: $\beta = 0$ and $\beta = 5{,}000$. The relationship between p and θ for $\beta = 0$ is given by

$$\theta = \frac{3p}{3 - 4p} \quad (4)$$

(Kimura 1980; Kingman 1982*a*). No analogous expression of equation (4) is known for $\beta > 0$. For $\beta = 5{,}000$,

the relationship between p and θ was found by simulation.

In the first set of simulations, the rate of recombination was varied over $\rho = 2$, 4, 8, and 16 for $\beta = 0$ ($n = 10$). For $\beta = 5{,}000$, ρ was determined such that $E_\beta[R(n)] = E_0[R(n)]$, and the same amount of recombination was expected in samples simulated under $\beta = 0$ and $\beta = 5{,}000$. This gave $\rho = 800$, 1,600, 3,200, and 6,400, approximately, for $n = 10$. In the second set of simulations, for $\beta = 0$, $\rho = 4$, and for $\beta = 5{,}000$, $\rho = 2{,}200$ ($n = 5$), $\rho = 1{,}600$ ($n = 10$), $\rho = 1{,}300$ ($n = 15$), $\rho = 1{,}000$ ($n = 25$), and $\rho = 700$ ($n = 50$), approximately.

Setup 2: Histories with One Recombination Event Only

In the second setup, samples and histories were simulated according to equation (3) with one recombination event of type 2 or 3 only or with one recombination event of type 3 only. The sample size was 10 in all simulations.

Three different forms of the rates γ_k were chosen: (A) $\gamma_k = k(k-1)/2$, (B) $\gamma_k = 1$, and (C) $\gamma_k = n - k + 2$. (Note that $\gamma_k = n - k + 2$ is only meaningful because the number of recombinations is restricted to 1; thus, $k \le n + 1$ and $\gamma_k > 0$ for all k.) Form A gives phylogenies with short external branches and long internal branches, and form C gives phylogenies with long external branches and short internal branches. Form B is between forms A and C. The form of γ_k determines when the recombination event happens: for form A, most recombinations happen while there are few lineages, and for forms B and C, most recombinations happen while there are many lineages (see eq. 3). The recombination break point occurs between nucleotides 500 and 501. The parameter θ was adjusted such that the pairwise sequence divergence varied over $p = 1\%$, 2.5%, 5%, 10%, and 20%. This was accomplished by simulation.

Results

Setup 1: Histories with a Random Number of Recombination Events

The amount of detectable recombination (types 2 and 3, fig. 1) varies with ρ, β, and n. Events of type 1 and 2 occur in samples of arbitrary size, whereas events of type 3 occur only if $n \ge 4$. The numbers of recombination events of the three different types are denoted by $R_1(n)$, $R_2(n)$, and $R_3(n)$, respectively, and the expected numbers of these three types of recombination events for $\beta = 0$ are given by (see appendix)

$$E_0[R_1(n)] = \frac{2}{3}\left(1 - \frac{1}{n}\right)\rho,$$

$$E_0[R_2(n)] = \left\{C(n) - \frac{2}{3}\left(1 - \frac{1}{n}\right)\right\}\rho, \quad \text{and}$$

$$E_0[R_3(n)] = \left\{\sum_{j=1}^{n-1}\frac{1}{j} - C(n)\right\}\rho, \qquad (5)$$

where $C(n)$ is an increasing function in sample size n, tending toward $C(\infty) \approx 2.14$. For $n = 2$ and $n = 3$, we have $E_0[R_3(n)] = 0$, and $C(2) = 1$ and $C(3) = \frac{2}{3}$. Both the expectation of $R_1(n)$ and the expectation of $R_2(n)$ are bounded in n; in fact, $\frac{1}{3}\rho \le E_0[R_1(n)] \le \frac{2}{3}\rho$ and $\frac{2}{3}\rho \le E_0[R_2(n)] \le \{C(\infty) - \frac{2}{3}\}\rho \approx 1.47\rho$. If $\beta > 0$, no explicit expressions are known for the expectations of $R_1(n)$, $R_2(n)$, and $R_3(n)$, although the ratio $E_\beta[R_3(n)]/E_\beta[R(n)]$ increases with β, and the expectations of $R_1(n)$ and $R_2(n)$ are bounded in n (unpublished data). Most events in large samples are of type 3.

For sample size 10, an increase in power was observed with increasing recombination rate (figs. 4 and 5) with both $\beta = 0$ and $\beta = 5{,}000$. For $\beta = 0$, the expected number of events of types 2 and 3 goes from about 4.8 ($\rho = 2$) to about 35 ($\rho = 16$), with the expected total number of events going from 5.7 ($\rho = 2$) to about 45 ($\rho = 16$). The expected number of type 2 and 3 events is larger for $\beta = 5{,}000$ than for $\beta = 0$. For $\rho \ge 4$, the chance that there is at least one event of type 2 or 3 in a random sample is almost 1, essentially ruling out the possibility of reporting false positives, that is, reporting recombinations in sample histories without recombination. Generally, an overall increase in power was observed with increasing sequence divergence p; this was expected, because the number of polymorphic sites increases with p. An exception to this was SSCF (with $\beta = 0$), for which the power stayed roughly constant for fixed recombination rate. Note that for $\rho = 0$, recombinations were detected in about 5% of the simulations. This was expected, as the level of significance was set to 5%.

There were some interesting differences between the results for $\beta = 0$ and those for $\beta = 5{,}000$. First, except for RecPars, which was roughly insensitive to the value of β, all methods had more power for $\beta = 0$ than for $\beta = 5{,}000$ (standard errors were between 1 and 3.5 for RecPars and were less than 1.3 for the other methods; results not shown.) Second, SSCF had less power for $\beta = 0$ than did the other methods, except if $p = 1\%$, in which case SSCF actually had the most power. For $\beta = 5{,}000$, SSCF and Max-Chi Squared had low power (less than 20%) unless both ρ and p were high. Here, the incompatibility methods, RecPars and NSS, performed remarkably better. These differences can be explained by differences in the shapes of phylogenies (table 1) for the two β values. More mutations happen at the long branches near the tips of the phylogeny for $\beta = 5{,}000$ than for $\beta = 0$, since for $\beta = 0$, branches near the tips are short. As a consequence, for $\beta = 5{,}000$, the sequences tend to have similar patterns of mutations (in terms of number and how they are distributed along the sequence), and recombination becomes harder to detect using Max Chi-Squared or SSCF. Furthermore, a mutation (e.g., A→G) happening at a tip branch is not informative unless paralleled by a similar mutation (i.e., A→G) at another branch. Thus, the incompatibility-based methods, RecPars and NSS, are less affected by the change in tree shape than are the polymorphic-site-based methods, Max Chi-Squared and SSCF.

Table 2 shows the results for the second set of simulations with $p = 10\%$ and n varying. Consider $\beta = 0$:

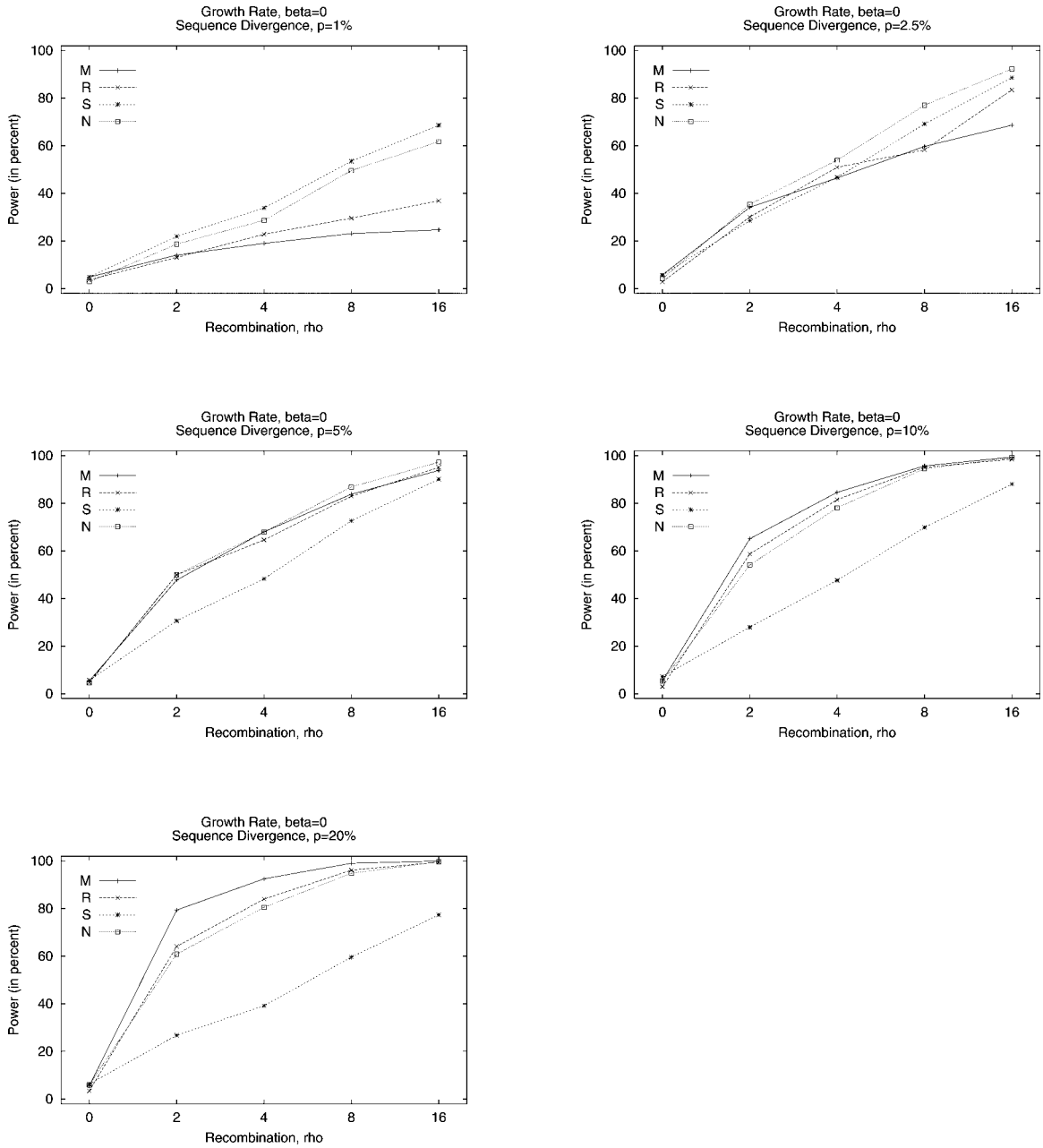

FIG. 4.—The power of the methods under setup 1 and $\beta = 0$. Power is plotted for various values of p and ρ and each of the four methods. The power increases with increasing ρ, which is expected as the number of recombination events increases with ρ. M = Max Chi-Squared; R = RecPars; S = SSCF; N = NNS. See the text for further explanation.

All four methods showed increasing power with increasing sample size. Next, consider $\beta = 5{,}000$. Here, SSCF and Max Chi-Squared showed drastically reduced power. The power of NSS was increased from $n = 5$ to $n = 10$, but decreased for $n > 10$ from 66% to 15%. In contrast, RecPars retained the power obtained under $\beta = 0$. The reduced power of SSCF and Max Chi-Squared can be explained by the fact that under population growth, even for $n = 50$, the branches of the tips compose a large part of the total branch length (table 1), and as consequence, most mutations happen near the tips. The fact that the power of NSS is increased from $n = 5$ to $n = 10$ is likely due to the higher percentage of type 3 events in the latter case. The decrease in power for larger sample sizes for this method was expected: The number of incompatible pairs of sites increases with sample size (for fixed ρ and p; results not shown). RecPars was not similarly affected, as it is not based on pairwise comparisons of sites. RecPars suggests a change in topology if a stretch of sites is consistently more economically explained by a new topology. Within such a stretch, there might be sites incompatible with the new topology (e.g., because of recombination or mutations); these sites will then be explained by mutations. This might account for the observed increase in power with n for this method.

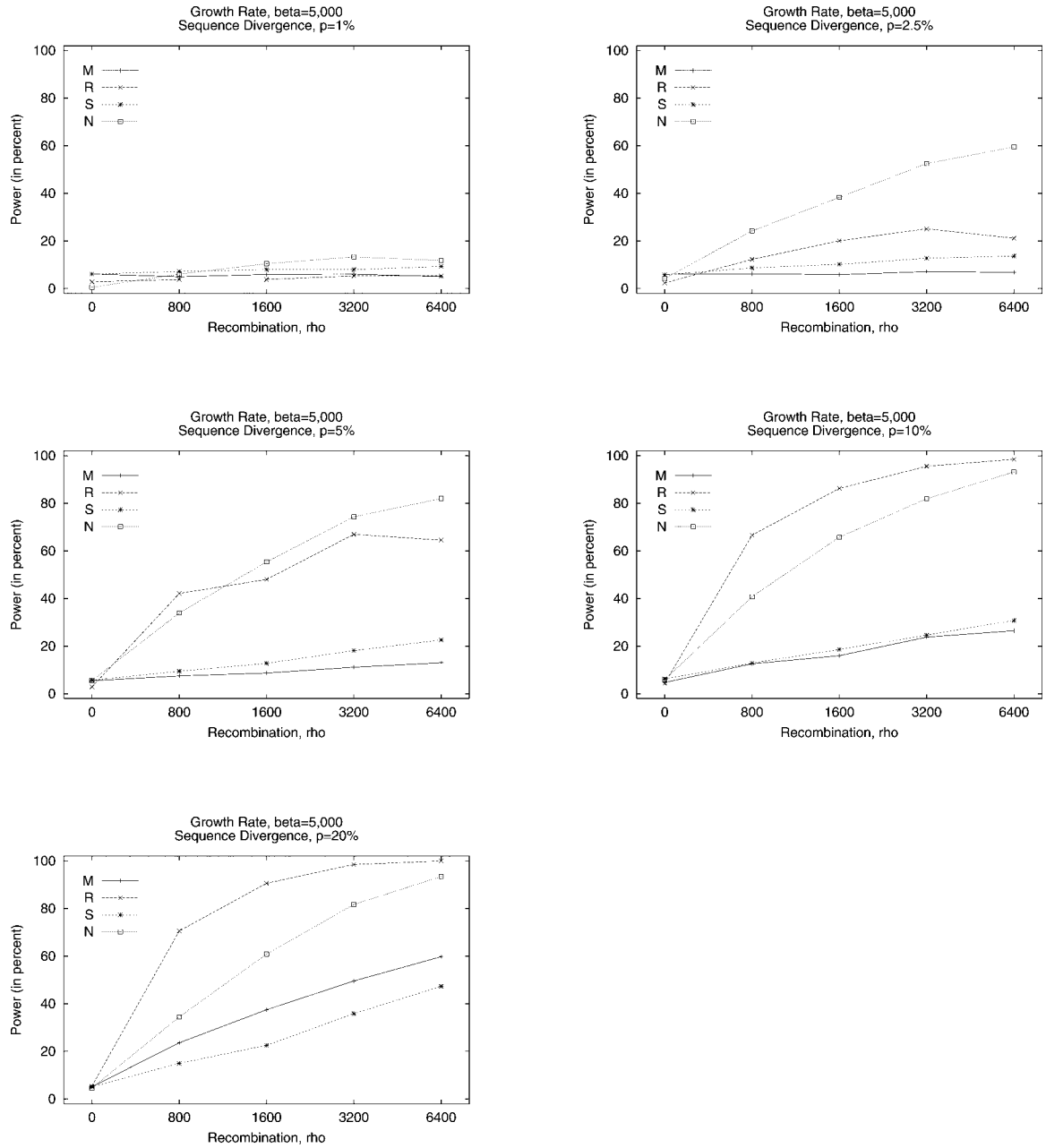

FIG. 5.—The power of the methods under setup 1 and $\beta = 5{,}000$. Power is plotted for various values of p and ρ and each of the four methods. The power increases with increasing ρ, which is expected as the number of recombination events increases with ρ. Furthermore, the power is generally lower than the power obtained with the same parameter values under $\beta = 0$. See the text for further explanation.

Table 2
Power for Different Sample Sizes

	$\beta^a = 0$				$\beta = 5{,}000$			
n^b	S	M	N	R	S	M	N	R
5	41	68	46	38	15	16	49	56
10	48	85	78	84	19	16	66	81
15	55	90	86	92	18	17	57	93
25	68	94	93	92	17	17	35	94
50	80	96	94	—	16	21	15	—

NOTE.—The powers of the four methods are shown as percentages. S = SSCF; M = Max Chi-Squared; N = NSS; R = RecPars.

[a] Growth rate.

[b] Sample size.

Setup 2: Histories with One Recombination Only

The probabilities p_{ik}, $i = 1, 2, 3$, that the recombination event is of type $i = 1, 2, 3$, given that it occurs while there are k lineages, are

$$p_{k1} = \frac{2}{3k}, \tag{6}$$

$$p_{k2} = \frac{8}{9k} + \frac{8}{(k+1)k^2(k-1)}\left(\frac{1}{3} + 4\sum_{j=1}^{k-2}\frac{1}{j}\right), \quad \text{and} \tag{7}$$

$$p_{k3} = 1 - p_{k1} - p_{k2} \tag{8}$$

(see *appendix*). If $k = 2$ or 3, $p_{k3} = 0$; otherwise, $p_{k3} > 0$. Furthermore, p_{k1} is decreasing for $k \geq 2$, p_{k2} is decreasing from $k = 3$ onward, and p_{k3} is increasing for $k \geq 2$. The overall chance that a recombination event is of type $i = 1, 2, 3$, given that it occurs, is

$$p_i = \sum_{k=2}^{n} p_{ki} q_k, \tag{9}$$

which, for specific choices of γ_k, $k = 2, \ldots, n$, can be calculated. Hudson and Kaplan (1985) found p_3 in the coalescent model with recombination.

Conditioned on exactly one event of type 2 or 3, the ratio of type 3 events to type 2 events changes with the form of γ_k: If $\gamma_k = k(k-1)/2$, $r = p_3/(p_2 + p_3) = 38\%$; if $\gamma_k = 1$, $r = 75\%$; and if $\gamma = 10 - k + 2$ ($n = 10$), $r = 82\%$. This should be reflected in NSS and RecPars, and to lesser extent in Max Chi-Squared and SSCF, because these methods are not based on the distribution of incompatible sites in the sample. For example, in the first case, NSS and RecPars should recover less than 38% of the recombination events unless some events of type 2 also are detected. In contrast, Max Chi-Squared and SSCF could potentially recover 100% because they are designed to detect either of the two types. Conditional on type 3 events, only the ratio is, of course, 100% for all methods.

For simplicity, let us denote by T3 the case in which the conditioning is on type 3 events only and let us denote by T23 the case in which the conditioning is on type 2 or 3 events. In case T23, the power was highest under (A) $\gamma_k = k(k-1)/2$ and lowest under (C) $\gamma_k = 10 - k + 2$, except for RecPars, where form B had the most power (fig. 6; standard errors ranged from 1 to 3 for RecPars and were less than 1.2 for the other methods; results not shown). Two factors are in play. First, more recombination events happen under C than under A while there are many lineages. This implies that the chance of a type 3 event is higher under C than under A, because a type 3 event requires at least four sequences. As a consequence, the chance of detecting a recombination should increase. (This argument also accounts for the observed higher power under T3 than under T23.) Second, under C, the phylogenies tend to be more starlike, and less information about topology is available from the sequences. The latter effect seemed to be of the most importance. It is interesting that the power of RecPars and NSS did not go up as the upper bound for detecting a recombination went up from 38% under A, through 75% under B, to 82% under C. This must be ascribed, again, to the fact that recombination is harder to detect in starlike trees.

Discussion

In this paper, we discussed the power of four different methods of detecting recombination. Generally, we find that all of them are based on statistics that capture features characteristic of samples that have experienced recombination in their history. However, we also find that all of the investigated methods detect far less recombination than is theoretically possible. In general, methods based on the principle of incompatibility seem to have more power than methods that do not make use of this principle explicitly. This seems, in particular, to be the case if the phylogeny has long branches at the tips and small deep branches. Figure 6 suggests that recombination in genealogies that are similar to genealogies sampled from Kingman's (1982*a*) coalescent is much easier to detect than recombination in starlike genealogies.

Both RecPars and Max Chi-Squared would perform better if the recombination cost or the window size were set individually for each simulated sample. The values chosen here are those that overall gave the highest power. In general, it would be difficult, if not impossible, to assign a value to the recombination cost in RecPars from information in data. The value chosen here (1.5 times the cost of a mutation) seems to work fine despite the wide range of shapes of genealogies and the amount of recombination, and it might be taken as a sensible starting value in analysis of real data. In contrast, the window size in Max Chi-Squared can easily be varied, and one can choose the size that gives the most significant output. However, this will not necessarily guarantee the highest power.

The mutation scheme applied in the simulations is very simple and not realistic for most real data. Rates of mutation tend to vary and to be higher in some regions of the sequences than in other regions. This will inevitably give more variable sequence patterns and affect the power negatively.

The sequence length can be varied. In setup 1, the effect of varying the sequence length L is equivalent to the effect of varying the recombination rate ρ; all pairs (L, ρ) of constant ratio $C = \rho/L$ produce similar outcomes. In setup 2, there is only one recombination event, and an increased sequence length increases the chance of detecting the recombination event: more information is available. Max Chi-Squared is unaffected, as a sliding-window approach is adopted. In real sequences there will be an upper limit to the size of a segment without recombination. For example, most reported recombinant segments in HIV are less than 1,000 nt long (see, e.g., http://hiv-web.lanl.gov), the sequence length chosen in our studies.

In general, it is difficult to set up guidelines concerning which method to choose in an analysis of real sequences. If long branches are expected at the tips, compatibility methods might be preferred. An indication

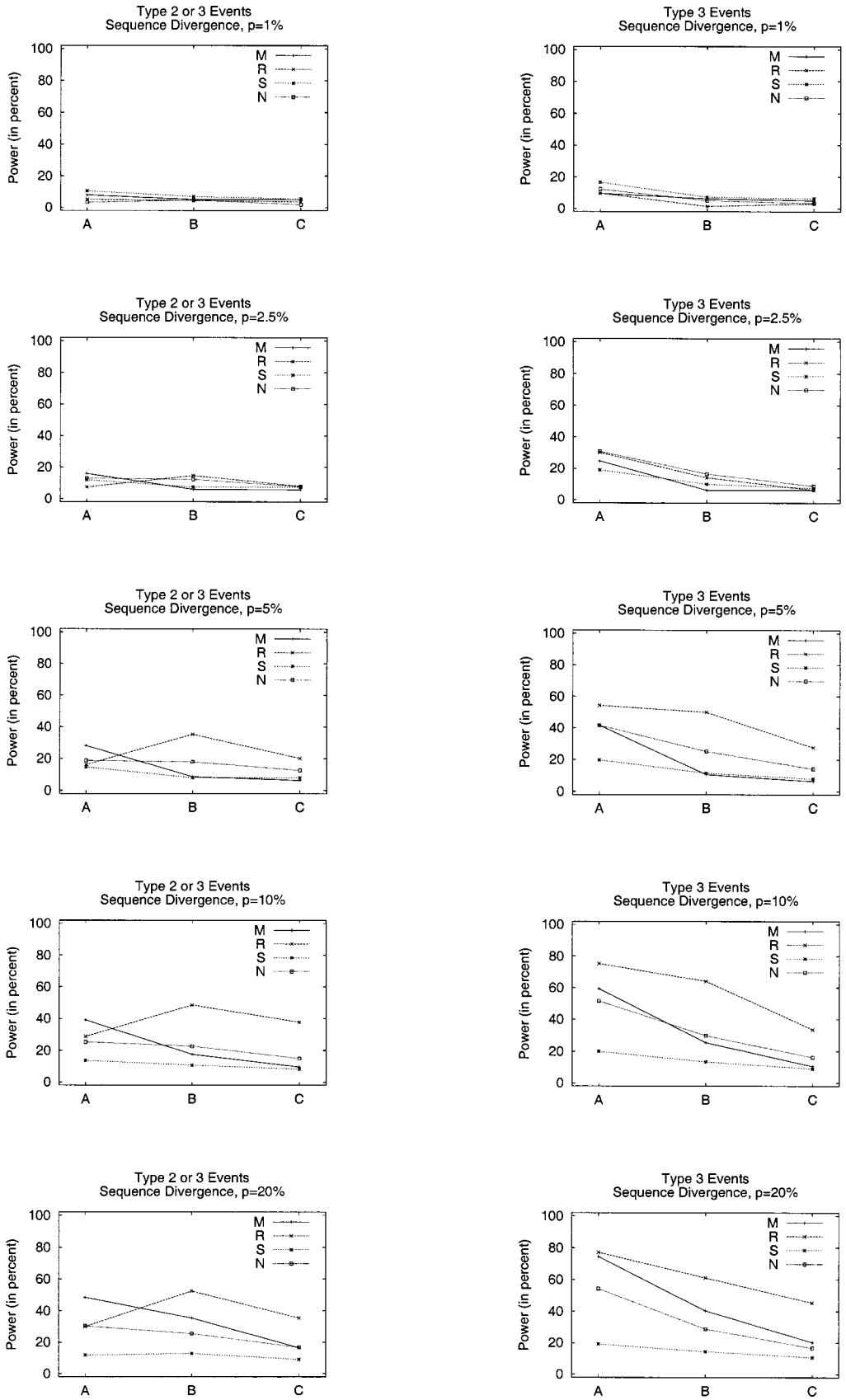

FIG. 6.—The powers of the methods under setup 2. Power is plotted for various choices of γ_k and p and each of the four methods. The curves correspond to the following values of α_k: (A) $\alpha_k = k(k-1)/2$, (B) $\gamma_k = 1$, and (C) $\alpha_k = n - k + 2$. Conditional on type 2 or 3 events, the predicted upper limits for the power of RecPars and NNS is 38%, 75%, and 82%, respectively, for A, B, and C. See the text for further explanation.

of long branches could be a large number of singletons in the sample. It is advisable to use a set of methods based on different principles.

Acknowledgments

M. Schierup, A.-M. Krabbe Pedersen, K. Strimmer, and D. Posada are thanked for reading and commenting on the manuscript. S. Sawyer is thanked for suggesting the use of variable coalescent rates. C.W. was supported by grant BBSRC 43/MMI09788 and by the Carlsberg Foundation, Denmark. The project was partly supported by Basic Research in Computer Science (BRICS), a center under the Danish National Research Foundation. Part of the research for this article was carried out while J.H. and C.W. visited the Isaac Newton Institute, University of Cambridge.

APPENDIX

Consider a model with coalescence rates α_k, $k = 2, 3, \ldots, n$, and recombination rate $\rho/2$ per sequence. All waiting times are independent of each other. The ordinary coalescent with recombination has $\alpha_k = k(k-1)/2$ (Hudson 1983), although in general α_k cannot be derived from a population model or a model of species evolution. Given k lineages, the time until an event is exponential with parameter $\gamma_k = \alpha_k + k\rho/2$. This event is a coalescence with probability α_k/γ_k and is otherwise a recombination. The probability of exactly one recombination event before all sites share an ancestor is

$$\frac{\rho}{2}\sum_{k=2}^{n}\frac{k\alpha_{k+1}}{\gamma_k\gamma_{k+1}}\prod_{j=2}^{n}\frac{\alpha_j}{\gamma_j} \tag{10}$$

(using arguments similar to Hudson and Kaplan's [1985]), and the probability that the recombination event happens while there are k lineages is

$$\frac{k\alpha_{k+1}}{\gamma_k\gamma_{k+1}}\Bigg/\sum_{j=2}^{n}\frac{j\alpha_{j+1}}{\gamma_j\gamma_{j+1}}, \tag{11}$$

$k = 2, \ldots, n$. Let p_{ki}, $i = 1, 2, 3$, be the probability that a recombination event is of type i given it happens while there are k lineages. Immediately after the event, there will be $k+1$ lineages. The probabilities p_{k1}, $k = 2, \ldots, n$, fulfill the recursion

$$p_{k1} = \frac{2}{(k+1)k} + \frac{(k-1)(k-2)}{(k+1)k}p_{(k-1)1}, \tag{12}$$

with boundary condition $p_{21} = 1/3$. The recursion has solution $p_{k1} = 2/(3k)$, which is equation (6). Concerning p_{k2}, the following recursion is found:

$$p_{k2} = \frac{4(k-1)}{(k+1)k}q_{k-1} + \frac{(k-1)(k-2)}{(k+1)k}p_{(k-1)2}, \tag{13}$$

with boundary conditions $q_2 = 1$ and $p_{22} = 2/3$. The q_k values fulfill the same recursion as p_{k1}. Recursion (13) can be solved and gives equation (7). The form of p_{k3} follows from $1 - p_{k1} - p_{k2}$.

If $\rho \approx 0$, equation (11) reduces to

$$\frac{k}{\gamma_k}\Bigg/\sum_{j=2}^{n}\frac{j}{\gamma_j} = \frac{k}{\alpha_k}\Bigg/\sum_{j=2}^{n}\frac{j}{\alpha_j}, \tag{14}$$

and times between events (coalescence and recombination) are exponential with $\gamma_k = \alpha_k$. We adopt this as our simulation scheme in setup 2. Equation (14) is identical to equation (3).

Under the coalescent with recombination and $\beta = 0$, the expectations of $R_i(n)$, $i = 1, 2, 3$, are given by (using arguments similar to Hudson and Kaplan's [1985])

$$E[R_i(n)] = \rho\sum_{k=2}^{n}\frac{1}{k-1}p_{ki}, \tag{15}$$

which, by insertion of p_{k1} and p_{k2}, gives the first two equations in (5). The expectation of $R_3(n)$ can be found from the relation $R(n) = R_1(n) - R_2(n) - R_3(n)$ or in Hudson and Kaplan (1985). Hudson and Kaplan found that

$$E_0[R_3(n)] = 16\sum_{k=4}^{n}\frac{1}{(k+1)k^2(k-1)^2} \times\left\{\sum_{i=2}^{k-2}\frac{1}{i}\left\{\sum_{j=2}^{i}j^2(j+1)\right\}\right\}\rho. \tag{16}$$

Using $\Sigma_{l=1}^{m} l = m(m+1)/2$, $\Sigma_{l=1}^{m} l^2 = m(m+1)(2m+1)/6$, and $\Sigma_{l=1}^{m} l^3 = m^2(m+1)^2/4$, by decomposition of the fraction

$$\frac{1}{(k+1)k^2(k-1)^2} = \frac{1}{2(k-1)^2} + \frac{1}{k^2} + \frac{1}{k} + \frac{1}{4(k+1)} - \frac{5}{4(k-1)},$$

it is found that

$$\frac{E_0[R_3(n)]}{\rho} = \sum_{i=1}^{n-1}\frac{1}{i} - 36\sum_{i=1}^{n-1}\frac{1}{i^2} - 48\sum_{i=1}^{n-1}\frac{1}{i^2}\sum_{j=1}^{i-1}\frac{1}{j} + 114 + \frac{7}{9} + o(n) = \sum_{i=1}^{n-1}\frac{1}{i} - C(n).$$

Here, the term of order $o(n)$ is

$$-\frac{8(6n^2+9n+4)}{n^2(n+1)}\sum_{i=1}^{n-1}\frac{1}{i} - \frac{2(371n^2+506n+156)}{9n^2(n+1)},$$

and $C(n)$ is short for the sum of all terms except the first. The calculations were checked by computer simulations. For n increasing, $C(n)$ tends to approximately 2.14.

LITERATURE CITED

CRANDALL, K. A., and A. R. TEMPLETON. 1999. Statistical methods for detecting recombination. Pp. 153–176 *in* K. A. CRANDALL, ed. The evolution of HIV. Johns Hopkins University Press, Baltimore, Md.

FITCH, D. H. A., and M. GOODMAN. 1991. Phylogenetic scanning: a computer-assisted algorithm for mapping gene conversion and other recombinational events. Comput. Appl. Biosci. **7**:207–215.

GRASSLY, N. C., and E. C. HOLMES. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol. Biol. Evol. **14**:239–247.

GRIFFITHS, R. C., and S. TAVARÉ. 1994. Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B Biol. Sci. **344**:403–410.

———. 1998. The age of a mutant in a general coalescent tree. Stochastic Models **14**:273–295.

HEIN, J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. **36**:396–405.

HUDSON, R. R. 1983. Properties of the neutral allele model with intergenic recombination. Theor. Popul. Biol. **23**:183–201.

HUDSON, R. R., and N. KAPLAN. 1985. Statistical properties of the number of recombination events in the history of DNA sequences. Genetics **111**:147–164.

JAKOBSEN, I., and S. EASTEAL. 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. CABIOS **12**:291–295.

JAKOBSEN, I., S. R. WILSON, and S. EASTEAL. 1997. The partition matrix: exploring variable phylogenetic signals along nucleotide sequence alignments. Mol. Biol. Evol. **14**:474–484.

JUKES, T. H., and C. CANTOR. 1969. Evolution of protein molecules. Pp 21–123 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KINGMAN, J. F. C. 1982*a*. The coalescent. Stochastic Processes Appl. **13**:235–248.

———. 1982*b*. Exchangeability and the evolution of large populations. Pp. 97–112 *in* G. KOCH and F. SPIZZI-CHINO, eds. Exchangeability in probability and statistics. Amsterdam, North-Holland.

LE QUESNE, W. J. 1969. A method of selection of characters in numerical taxonomy. Syst. Zool. **18**:202–205.

MCGUIRE, G., F. WRIGHT, and M. J. PRENTICE. 2000. A Bayesian model for detecting past recombination events in DNA multiple alignments. J. Comp. Biol. **7**:159–170.

MAYNARD SMITH, J. 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. **34**:126–129.

MAYNARD SMITH, J., and N. H. SMITH. 1998. Detecting recombination from gene trees. Mol. Biol. Evol. **15**:590–599.

PYBUS, O. G., E. C. HOLMES, and P. H. HARVEY. 1999. The mid-depth method and HIV-1: a practical approach for testing hypotheses of viral epidemic history. Mol. Biol. Evol. **16**:953–959.

RODRIGO, A. G., and J. FELSENSTEIN. 1999. Coalescent approaches to HIV population genetics. Pp 233–272 *in* K. A. CRANDALL, ed. The evolution of HIV. Johns Hopkins University Press, Baltimore, Md.

SAWYER, S. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6**:526–536.

SLATKIN, M., and R. R. HUDSON. 1991. Pairwise comparisons of mitochondria DNA sequences in stable and exponentially growing populations. Genetics **129**:555–562.

SNEATH, P. H. A., M. J. SACKIN, and R. P. AMBLER. 1975. Detecting evolutionary incompatibilities from protein sequences. Syst. Zool. **24**:311–332.

STEPHENS, J. C. 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2**:539–556.

WEILLER, G. F. 1998. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol. Biol. Evol. **15**:326–335.

KEITH CRANDALL, reviewing editor

Accepted June 15, 2001