Accurate Computation of Likelihoods in the Coalescent with Recombination via Parsimony

Rune B. Lyngsø[1], Yun S. Song[2] and Jotun Hein[1]

[1] Department of Statistics, Oxford University, Oxford OX1 3TG, UK
[2] Computer Science Division and Department of Statistics, University of California, Berkeley, CA 94720, USA
`lyngsoe@stats.ox.ac.uk, yss@cs.berkeley.edu, hein@stats.ox.ac.uk`

Abstract. Understanding the variation of recombination rates across a given genome is crucial for disease gene mapping and for detecting signatures of selection, to name just a couple of applications. A widely-used method of estimating recombination rates is the maximum likelihood approach, and the problem of accurately computing likelihoods in the coalescent with recombination has received much attention in the past. A variety of sampling and approximation methods have been proposed, but no single method seems to perform consistently better than the rest, and there still is great value in developing better statistical methods for accurately computing likelihoods. So far, with the exception of some two-locus models, it has remained unknown how the true likelihood exactly behaves as a function of model parameters, or how close estimated likelihoods are to the true likelihood. In this paper, we develop a deterministic, parsimony-based method of accurately computing the likelihood for multi-locus input data of moderate size. We first find the set of all ancestral configurations (ACs) that occur in evolutionary histories with at most k crossover recombinations. Then, we compute the likelihood by summing over all evolutionary histories that can be constructed only using the ACs in that set. We allow for an arbitrary number of crossing over, coalescent and mutation events in a history, as long as the transitions stay within that restricted set of ACs. For given parameter values, by gradually increasing the bound k until the likelihood stabilizes, we can obtain an accurate estimate of the likelihood. At least for moderate crossover rates, the algorithm-based method described here opens up a new window of opportunities for testing and fine-tuning statistical methods for computing likelihoods.

1 Introduction

Estimating evolutionary parameters and making ancestral inference are an important part of molecular evolutionary genetics. Often, at the core of these studies is the problem of computing the probability of observing sample sequences for given parameter values. In the context of the coalescent model and its various extensions, closed-form formulas are generally not known for such likelihoods, and therefore several computationally intensive statistical methods have been

proposed for approximating them. Most of these statistical approaches fall into one of two categories, one based on Markov chain Monte Carlo methods—for examples, see [3, 25, 26, 44]—and the other on importance sampling methods, some notable examples being [5, 6, 9, 12–15, 41]. Both approaches involve sampling genealogies to estimate a sum over the genealogies consistent with the input data.

The problem of estimating recombination rates has received particular attention in the past and various methods have been proposed so far [9–12,23,26,28,30, 42]. (Henceforward, when we say recombination, we will mean crossover recombination.) Since computing the full likelihood in the coalescent with recombination is difficult, several approximation methods have been proposed. Hudson's [23] composite likelihood method is a popular approximation method which treats pairs of loci as being independent and takes a product of two-locus full likelihoods over all pairs of loci. (Different versions of the composite likelihood idea have also been suggested. E.g, see [10,11].) This method has been generalized to study the fine-scale crossover rate variation in the human genome [24, 32, 33].

On the algorithms side, much recent attention has focused on the problem of estimating the minimum number $R_{min}(D)$ of recombinations needed to derive a given set D of sequences, using some specified model of mutations. A commonly adopted model is the *infinite-sites model*, which implies that each site can mutate at most once in the entire evolutionary history of the sequences. Assuming that mutation model, it has been shown that computing $R_{min}(D)$ is NP-hard [4, 43], and previous algorithms that compute it exactly either work only on relatively small data sets [36, 38], or on problems with special structure [16–18]. Since there are no efficient algorithms to compute $R_{min}(D)$ for an arbitrary D, several papers have considered efficient computation of lower bounds on $R_{min}(D)$ [1, 2, 16–20, 22, 34, 37, 40], as well as practical upper bounds [40].

In a recent paper [29], we have made progress in making the exact computation of $R_{min}(D)$ more practical, significantly increasing the size of data sets that can be handled. Here, we extend some of the algorithmic ideas developed in that paper to address the aforementioned problem of computing likelihoods in the coalescent with recombination. To our knowledge, this is the first application of a parsimony-based algorithm to likelihood computations in the coalescent.

The main idea behind our approach goes as follows. Instead of attempting to sum over all genealogies, we sum only over a restricted subset of genealogies. To each genealogy, there corresponds a sequences of events, consisting of mutations, coalescences, and recombinations. When an event happens, going backwards in time, there is a change in ancestral configuration (AC) [39], defined as the set of all DNA sequences present at a particular point in time in the genealogy. Summing over all genealogies for D corresponds to summing of all paths of ACs consistent with D, i.e., with each path starting from the input data D and ending at an AC in which every site in the input data has found a common ancestor. In our work, we first find the set of all ACs that occur in evolutionary histories with at most k recombinations. Then, we compute the likelihood by summing over all evolutionary histories that can be constructed only using the ancestral

configurations in that set. We allow for an arbitrary number of recombination, coalescent and mutation events in the evolutionary history, as long as the transitions stay within that restricted set of ancestral configurations. By starting with $k = R_{min}(D)$ and incrementing the bound k gradually until the change in likelihood satisfies some stopping criteria, we can compute the likelihood accurately.

There exist well-defined recursions relating the probability of a given AC ψ to the probabilities of those ACs that can be reached from ψ using one event back in time [7,8,12–15,35]. Solving the system of recursion relations to evaluate the probability of $\psi = D$ effectively sums over all possible genealogies consistent with D. In our work, we systematically solve the system of recursion relations involving the probabilities of the ACs in the restricted set described above. Note that this effectively sums not only over genealogies with at most k recombinations, but over all genealogies that can be constructed using the ACs in the restricted set with an arbitrary number of recombination events.

Although our deterministic approach can currently handle only small data sets—say, with about ten sequences and half as many sites—the work described here should prove useful for evaluating the performance of Monte-Carlo-based methods. Further, some pseudo-likelihood methods [23,32,33] are based on accurate likelihood calculations for few (typically 2) sites, and the method presented here significantly extends this capability.

2   Methods

We use D to denote a set of single nucleotide polymorphisms (SNPs) with two alleles at each site. We assume that the ancestral allele type is known. (This assumption is only made for ease of exposition. The approach presented here has a straightforward generalization to the case in which the ancestral allele type is unknown, albeit with steeper time and space complexity. Our software handles both cases.) In what follows, the ancestral allele is denoted by 0, while the mutant allele type is denoted by 1. For given mutation and recombination rates, our goal is to compute the probability of observing D under the coalescent with recombination and the infinite sites model of mutation.

2.1   Possible Events Back in Time

We assume that D contains m segregating sites with positions $s_1, \ldots, s_m$. We rescale the region to a unit interval between 0 and 1 so that $0 = s_1 < s_2 < \cdots < s_m = 1$. We use θ and ρ to denote, respectively, the population-scaled mutation and recombination rates for the unit interval. We assume that both recombination and mutation rates are constant over the interval. For given θ and ρ, the probability of observing D is obtained by integrating over the probabilities of all evolutionary histories that derive D. Tracing an evolutionary history backwards in time gives a path of ancestral configurations, reached from D through the following types of events back in time:

Mutation We assume the infinite sites model of mutation. So, for any particular site, if there is exactly one sequence carrying a 1 at that site, it may change to the ancestral type 0.

Recombination A sequence x breaks up into two new sequences with a breakpoint between sites i and $i+1$. One new sequence carries the prefix of x up to site i, followed by a suffix of length $m-i$ carrying non-ancestral material, denoted by $*$s. The other new sequence carries the suffix of x starting from site $i+1$, preceded by a prefix of length i carrying non-ancestral material, again denoted by $*$s. Recombination events where there is no ancestral material (0 or 1) either to the left or to the right of the breakpoint are ignored.

Coalescent Type 1 Two identical sequences find a common ancestor.

Coalescent Type 2 Two distinct sequences find a common ancestor if there is no site in which one sequence carries a 1 and the other a 0. Suppose that two sequences x and y are replaced by a single sequence z via coalescence. Then, z contains a 1 (respectively, 0) at site i if either x or y contains a 1 (respectively, 0) at site i. Otherwise, z has a "$*$" at site i.

See [12] for a more detailed description of the coalescent with recombination.

2.2   The Full Recursion

Griffiths and Marjoram [12] constructed a system of recursion relations satisfied by ancestral configurations, assuming a continuous model of recombination. Obtaining a closed-form solution to the recursions is out of reach, so they proposed using an importance sampling method to obtain numerical solutions. More efficient importance sampling approaches now exist for computing coalescent likelihoods by sampling genealogies [5, 6, 9, 41], but the recursions found by Griffiths and Marjoram still provide a transparent framework for computation. In what follows, we devise a deterministic algorithm for numerically solving the recursions accurately. To make the problem tractable, we assume a discrete model of recombination such that breakpoints occur only at the midpoints between consecutive segregating sites. Such a discretized model of recombination has been adopted by others in the past [9, 27].

To describe the recursions in more detail, we first need to define some notation. An ancestral configuration is a multiset of strings from $X = \{0, 1, *\}^m \setminus \{*^m\}$. With a chosen ordering on X, we use $\boldsymbol{n} \in \mathbb{Z}_{\geq 0}^{3^m-1}$ to uniquely specify an AC by listing the multiplicity of each element in X. A subscript (respectively, superscript) on $\boldsymbol{n}$ denotes decreasing (respectively, increasing) the multiplicity of string i by 1. For example, $\boldsymbol{n}_i$ denotes changing the component n_i to $n_i + 1$, while keeping n_j for $j \neq i$ unchanged. Then, the recursion relation satisfied by the probability $Q(\boldsymbol{n})$ of an AC $\boldsymbol{n}$ can be schematically written as

$$
Z(\boldsymbol{n}, \theta, \rho)\, Q(\boldsymbol{n}) = \underbrace{\sum_{} c(\boldsymbol{n}, i)\, Q(\boldsymbol{n}_i)}_{\substack{\text{coalescent type 1} \\ x_i \text{ with } x_i}} + \underbrace{\sum_{} c(\boldsymbol{n}, i, j, k)\, Q(\boldsymbol{n}_{ij}^k)}_{\substack{\text{coalescent type 2} \\ x_i \text{ with } x_j \to x_k}}
$$
$$
+ \theta \underbrace{\sum_{} c(\boldsymbol{n}, k)\, Q(\boldsymbol{n}_i^k)}_{\substack{\text{mutation} \\ x_i \to x_k}} + \rho \underbrace{\sum_{} c(\boldsymbol{n}, s_1, \ldots, s_m, , i, j, k)\, Q(\boldsymbol{n}_k^{ij})}_{\substack{\text{recombination} \\ x_k \to x_i \text{ and } x_j}},
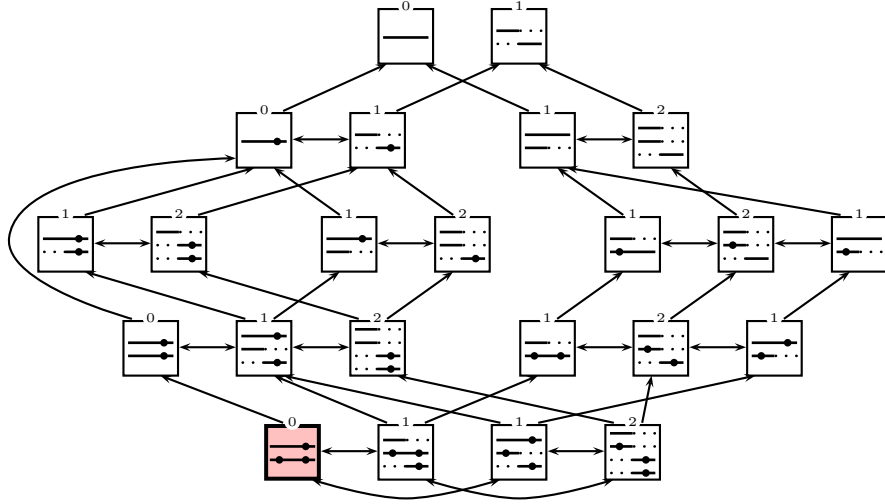\tag{1}
$$

Fig. 1. Graph of all ACs for $D = \{01, 11\}$. Each node (box) corresponds to an AC, with "—" denoting an ancestral segment, "$\cdot\cdot$" a non-ancestral segment, and "$\bullet$" a site carrying the mutant allele. The highlighted node at the bottom left corresponds to D. A directed edge joins an AC x to an AC y if there is an event (coalescence, mutation, or recombination) that transforms x to y. Each node is labeled with the minimum value of k for which that AC is in $\mathcal{C}_k(D)$, i.e. $\mathcal{C}_k(D)$ consists of all the nodes labeled k or less. ACs connected by horizontal bi-directional arrows form a strongly connected component of the graph. The probability of D for $\theta = 2$ and $\rho = 1$ is 0.125 when only ACs from $\mathcal{C}_0(D)$ are used, 0.193 when ACs from $\mathcal{C}_1(D)$ are used, and 0.202 when the full equation system is used. If cyclic structures of the recursions are eliminated by requiring that coalescing sequences have at least one site where they both carry ancestral material, as proposed in [31], the above probabilities reduce to 0.125, 0.172, and 0.174, respectively. This suggests that our parsimony-based approximation method of restricting the set of ACs is more accurate than forbidding certain classes of events.

where $c(\cdot)$ denote combinatorial coefficients that depend on the factors specified in the argument; $Z(\boldsymbol{n}, \theta, \rho)$ is a normalization constant; $x_i, x_j, x_k \in X$; and summations are performed over the events described in the previous subsection. Shown in Fig. 1 is an example of "unwrapping" the above recursion for an input data set D containing two length-2 sequences 01 and 11. In total there are 23 ACs for D, shown as rectangular boxes in Fig. 1, explained further below.

2.3   Restricting the Recursion

The discretized recombination model described above considerably reduces the number of possible ACs, from infinite to finite. Since (1) describes a system of linear equations, we could in principle find the probability of D by constructing and solving this equation system. It would correspond to summing over all genealogies that can derive D. However, although finite, the number of possible

ACs for a given data set grows extremely fast with the size of the data set [39], and exact computation remains infeasible for practical purposes. As mentioned in Sect. 1, the main idea behind our work is to sum over a restricted subset of genealogies, rather than over all genealogies. We achieve this by solving a restricted system of recursions. First, we find the set of all ACs each occurring in at least one possible evolutionary history for D with at most k recombinations, but with arbitrary coalescent and mutation events. Then, solving the system of recursions restricted to that set of ACs corresponds to computing the likelihood by summing over all evolutionary histories that can be constructed only using the ACs in that set. Note that this is more general than summing over the genealogies with at most k recombinations. As long as transitions remain within the restricted set of ACs, our method allows for an arbitrary number of recombination events in a genealogy.

The method can be used either with a fixed value of k determined from $R_{min}(D)$, or increasing k until a stopping criteria is met. The simplest stopping criteria is to continue until the change in likelihood becomes less than some specified small number ϵ. From our experiment, we suggest using a stopping criteria based on diminishing returns, stopping when the change in likelihood begins to decrease.

Formally, we define the *k-neighborhood* $\mathcal{N}_k(\boldsymbol{n})$ of an AC $\boldsymbol{n}$ to be the set of all ACs reachable from $\boldsymbol{n}$ with no more than k recombinations. The *inverse k-neighborhood* of $\boldsymbol{n}$ is defined as $\mathcal{N}_k^{-1}(\boldsymbol{n}) = \{\boldsymbol{n'} \mid \boldsymbol{n} \in \mathcal{N}_k(\boldsymbol{n'})\}$. Finally, the *k-configurations* for D is defined as $\mathcal{C}_k(D) := \bigcup_{i=0}^{k} \left[\mathcal{N}_i(D) \cap \mathcal{N}_{k-i}^{-1}(\mathcal{A}) \right]$, where $\mathcal{A}$ denotes the set of ACs in which every site has found a common ancestor and $\mathcal{N}_i^{-1}(\mathcal{A}) := \bigcup_{\boldsymbol{a} \in \mathcal{A}} \mathcal{N}_i^{-1}(\boldsymbol{a})$. Note that $\mathcal{C}_k(D)$ is the set of ACs that can occur in histories with at most k recombinations. Fig. 2 illustrates these concepts.

Our proposed method of computing the probability of the input data set D is to set $Q(\boldsymbol{n}) = 0$ if $\boldsymbol{n} \notin \mathcal{C}_k(D)$ and apply the recursion in (1) if $\boldsymbol{n} \in \mathcal{C}_k(D)$. For a data set with n sequences and m segregating sites, $\mathcal{C}_{2n(m-1)}(D)$ will be equal to the set of all ACs for D, since any AC can be reached from D using at most $n(m-1)$ recombinations and an AC in $\mathcal{A}$ can then be reached using at most $n(m-1)$ additional recombinations. Therefore, for sufficiently large k, our method becomes equivalent to solving the full equation system.

2.4   Algorithmic and Implementation Details

The k-neighborhoods $\mathcal{N}_k(D)$ can be computed incrementally by increasing k one by one. However, in our work the entire k-neighborhood $\mathcal{N}_k(D)$ is not needed; only the k-configurations $\mathcal{C}_k(D) = \bigcup_{i=0}^{k} \left[\mathcal{N}_i(D) \cap \mathcal{N}_{k-i}^{-1}(\mathcal{A}) \right]$ are needed. First, note that we can determine whether $\boldsymbol{n} \in \mathcal{N}_{k-i}^{-1}(\mathcal{A})$ by checking whether the minimum number of recombinations needed to derive $\boldsymbol{n}$ is at most $k-i$. To compute that minimum number, we can employ the algorithm described in [29]. Second, $\boldsymbol{n} \in \mathcal{N}_i(D) \cap \mathcal{N}_{k-i}^{-1}(\mathcal{A})$ only if there exists an AC $\boldsymbol{n'} \in \mathcal{N}_{i-1}(D) \cap \mathcal{N}_{k-i+1}^{-1}(\mathcal{A})$ such that $\boldsymbol{n} \in \mathcal{N}_1(\boldsymbol{n'})$. Using these ideas, we can find $\mathcal{C}_k(D)$ without having to explore the entire set $\mathcal{N}_k(D)$, which can be significantly larger than
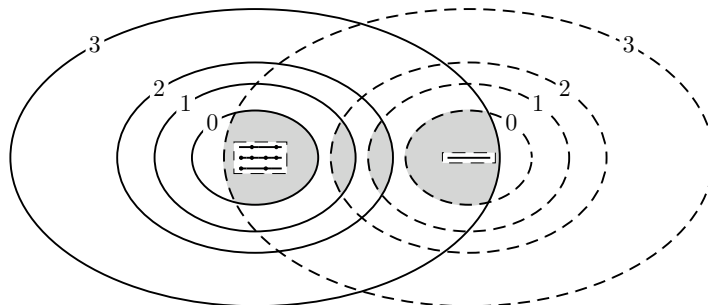
Fig. 2. Illustration of k-neighborhoods $\mathcal{N}_k(D)$ and k-configurations $\mathcal{C}_k(D)$. Neighborhoods around the input data D are shown as solid ellipses and inverse neighborhoods $\mathcal{N}_i^{-1}(\boldsymbol{a})$ around a particular most-recent-common-ancestor AC $\boldsymbol{a} \in \mathcal{A}$ are shown with dashed ellipses. The minimum number of recombinations required for the data set is 3, and the regions corresponding to the set $\mathcal{C}_3(D)$ of 3-configurations are shaded in gray.

$\mathcal{C}_k(D)$. Pictorially, in Fig. 2 we enumerate only the ACs in the shaded areas and their one-event neighbors, rather than the full k-neighborhoods of D. In this way, we can achieve a large reduction in both time and space requirement.

A dependency graph corresponding to the systems of recursions in (1) is a graph with one node for each AC and a directed edge from $\boldsymbol{n}$ to $\boldsymbol{n}'$ if $\boldsymbol{n}'$ appears on the right hand side of the recursion (1) for $\boldsymbol{n}$. Once $\mathcal{C}_k(D)$ has been found, we determine the strongly connected components of the dependency graph and the directed acyclic graph connecting them. Then, subsystems of recursions corresponding to the strongly connected components are solved in reverse topological order. This reduces the time complexity from $O(|\mathcal{C}_k(D)|^3)$ to $O(|\mathcal{C}_k(D)| \times M^2)$, where M is the size of the largest strongly connected component.

In our implementation, a coarse grained *a priori* separation and sorting of connected components is obtained by sorting the ACs in $\mathcal{C}_k(D)$ by their total number of 0s and 1s. Going backwards in time in any evolutionary history, the total number of 0s and 1s will be non-increasing. This means that if the total number of 0s and 1s in $\boldsymbol{n}$ is larger than that in $\boldsymbol{n}'$, then $Q(\boldsymbol{n}')$ does not depend on the value of $Q(\boldsymbol{n})$, thus allowing the recursions to be solved slice by slice in the order of increasing total number of 0s and 1s.

We have implemented our algorithm in C, using the UMFPack library. Our software is called `cob`, available at `http://www.stats.ox.ac.uk/~lyngsoe/section26/` under the Lesser Gnu Public License.

3   Results

When facing a hard computational problem, one usually needs to choose the right balance between accuracy and speed. In this section we explore these two aspects of our method. We will assess the quality of the approximation proposed in the previous section, by characterizing the behavior of the likelihood itself and
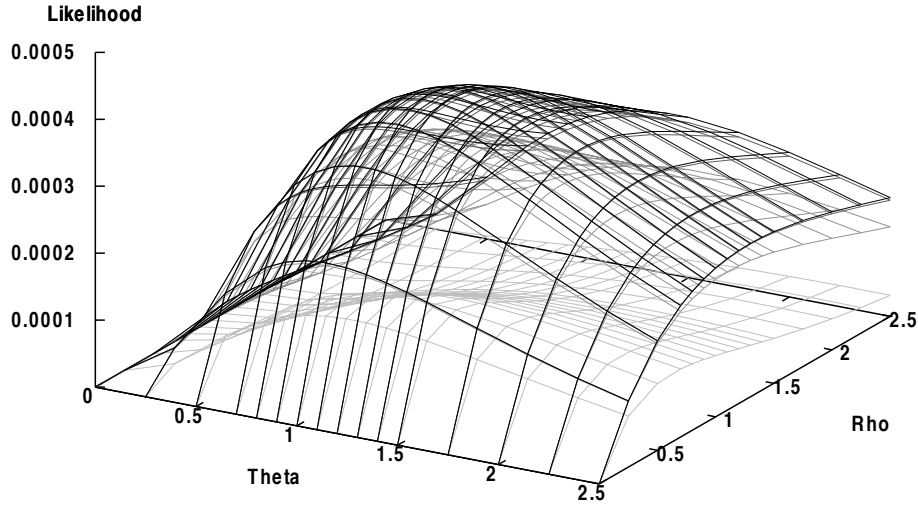
Fig. 3. Likelihood surfaces for $D = \{010, 010, 101, 101, 110\}$ computed using the ACs in $\mathcal{C}_1(D)$ (light gray), $\mathcal{C}_2(D)$ (medium gray), $\mathcal{C}_3(D)$ (dark gray) and $\mathcal{C}_{11}(D)$ (black). Both θ and ρ range from 0.0 to 2.5.

also by studying the accuracy of the maximum likelihood estimates (MLEs) of the population-scaled mutation and recombination rates θ and ρ, respectively.

3.1   Comprehensive Analysis of Small Data Set

We first study a small data set $D = \{010, 010, 101, 101, 110\}$ with segregating sites at positions 0, 0.75 and 1. The minimum number $R_{min}(D)$ of recombinations for this data set is 1, and the size of $\mathcal{C}_1(D)$ is 74. It turns out that all possible ACs can occur in evolutionary histories with at most 11 recombinations. The size of $\mathcal{C}_{11}(D)$ is 400,820. This is sufficiently small that the full system of recursions can be solved in a reasonable time, allowing us to track the accuracy and resource requirement for the approximation based on $\mathcal{C}_k(D)$, as k is varied from 1 to 11.

For a grid of θ and ρ values between 0.0 and 2.5, Fig. 3 shows four likelihood surfaces computed using four different levels of approximation: based on $\mathcal{C}_1(D)$, $\mathcal{C}_2(D)$, $\mathcal{C}_3(D)$, and the full equation system (i.e., $\mathcal{C}_{11}(D)$). The likelihood surfaces for the remaining $\mathcal{C}_k(D)$ have been left out as they are sandwiched between the $\mathcal{C}_3(D)$-surface and the $\mathcal{C}_{11}(D)$-surface, and these are already very similar. Numerical values of the likelihood obtained from the $\mathcal{C}_3(D)$-based equation system and that from the $\mathcal{C}_{11}(D)$-based system differ by little, at least for the range of θ and ρ shown in Fig. 3. Consequently, MLEs θ^*, ρ^* for the two cases are very similar, with $\theta^* = 1.1426$, $\rho^* = 0.9631$ for $\mathcal{C}_3(D)$ and $\theta^* = 1.1426$, $\rho^* = 0.9753$ for $\mathcal{C}_{11}(D)$. Even the $\mathcal{C}_2(D)$-surface is not far off from the $\mathcal{C}_{11}(D)$-surface, with MLEs $\theta^* = 1.1523$ and $\rho^* = 0.8240$, which is beginning to show a trend of underestimating ρ. This trend becomes even more pronounced with the
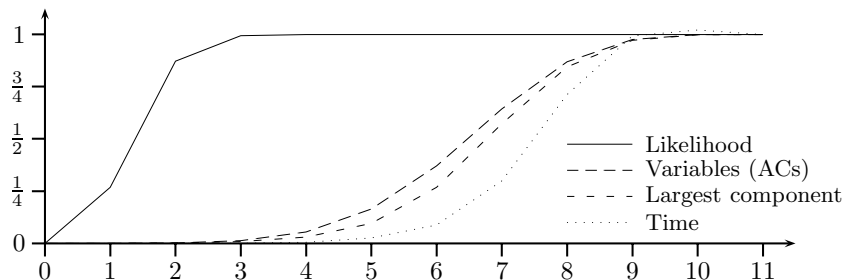
Fig. 4. The likelihood of $D = \{010, 010, 101, 101, 110\}$, the number of ACs (or variables), the size of the largest strongly connected component, and the running time, each plotted against the value of k in $\mathcal{C}_k(D)$, as a fraction of the corresponding values for the full recursion system. The full recursion system produced a likelihood of 4.05×10^{-4} and had $400, 820$ ACs in total. The largest strongly connected component contained $15, 998$ ACs and the computation took 404.2 seconds.

$\mathcal{C}_1(D)$-based equation system, with estimates $\theta^* = 1.2427$ and $\rho^* = 0.4370$. This $\mathcal{C}_1(D)$-based method also significantly underestimates the likelihood for most values of θ and ρ. As our heuristic is based on ignoring ACs that only contribute to the likelihood through evolutionary histories with many recombinations, for fixed θ, not surprisingly the difference between the approximated and the true likelihoods increases with increasing ρ. For fixed ρ and varying θ, the difference between the approximated and the true likelihoods tend to correlate more with the magnitude of the likelihood than with the value of θ.

As k in $\mathcal{C}_k(D)$ varied from 1 to 11, the change in some key features of the computation is plotted in Fig. 4 where the likelihood was computed at the MLEs $\theta^* = 1.1426$, $\rho^* = 0.9753$ from the full equation system. All plots exhibit an S-curve behavior. A very encouraging feature of these plots is that the turning point for the likelihood plot occurs much earlier than the turning points for the time, the largest strongly connected component, and the equation system size plots. In fact, the likelihood all but coincides with the exact value before any of the other features shows any increasing tendency. Identifying the turning point of an assumed S-curve behavior of the likelihood is the basis of our diminishing-returns stopping criteria described before.

3.2   Average Behavior on Simulated Data Sets

The data set analyzed above is just one example. To study the average behavior, we used Hudson's [21] program `ms` to generate simulated data under the coalescent with recombination. We generated 100 data sets for a given number of sequences and a given number of sites. We considered two to six sequences with either two or three sites. Hudson's program actually uses a finite-sites model of recombination, requiring the user to specify the number of sites. In our study, all simulations were carried out with $10, 000$ sites in the recombination model. We set $\rho = 5$ and used `-s` option to fix the number of segregating sites. For each data

Table 1. Average behavior of maximum likelihood estimation based on the near-minimal history restriction. For a given n and m, 100 data sets with n sequences and m sites were simulated using Hudson's [21] program `ms`. LR denotes the estimated likelihood relative to the true likelihood (denoted LH) computed using the full equation system, while $|\Delta\theta^*|$ and $|\Delta\rho^*|$ denote average absolute deviation from the true MLEs (θ^* and ρ^*) obtained using the full equation system. Running times are given in seconds. The columns under "Diminishing Returns" are for incrementing the number of recombinations until differences of likelihoods between increments no longer increases. The column labeled "k" lists the average value of k for which the final solution was obtained using $\mathcal{C}_{r+k}(D)$-configurations, where $r := R_{min}(D)$.

	$\mathcal{C}_r(D)$				$\mathcal{C}_{r+1}(D)$				$\mathcal{C}_{r+2}(D)$			
$n \times m$	LR	$|\Delta\theta^*|$	$|\Delta\rho^*|$	Time	LR	$|\Delta\theta^*|$	$|\Delta\rho^*|$	Time	LR	$|\Delta\theta^*|$	$|\Delta\rho^*|$	Time
2×2	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.0000	0.0000	0.02
3×2	0.99	0.01	0.06	0.00	1.00	0.01	0.03	0.10	1.00	0.0000	0.0003	0.12
4×2	1.00	0.01	0.09	0.00	1.00	0.01	0.05	0.16	1.00	0.0000	0.0014	0.19
5×2	0.98	0.02	0.12	0.02	1.00	0.01	0.06	0.21	1.00	0.0002	0.0018	0.36
6×2	0.98	0.02	0.17	0.09	0.99	0.01	0.08	0.33	1.00	0.0002	0.0034	0.56
2×3	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.10	1.00	0.0000	0.0000	0.10
3×3	0.99	0.02	0.05	0.00	1.00	0.01	0.03	0.24	1.00	0.0007	0.0031	0.31
4×3	0.98	0.03	0.10	0.05	1.00	0.01	0.05	0.45	1.00	0.0016	0.0057	0.88
5×3	0.99	0.01	0.07	0.09	1.00	0.01	0.04	0.59	1.00	0.0000	0.0036	0.90

	Diminishing Returns					Full Equation System			
$n \times m$	LR	$|\Delta\theta^*|$	$|\Delta\rho^*|$	Time	k	LH	θ^*	ρ^*	Time
2×2	1.00	0.0000	0.0000	0.002	1.00	0.15	5.08	0.00	0.10
3×2	1.00	0.0000	0.0000	0.15	1.27	0.06	3.88	0.07	0.19
4×2	1.00	0.0000	0.0000	0.28	1.29	0.05	4.16	0.09	0.50
5×2	1.00	0.0002	0.0002	0.70	1.76	0.03	3.41	0.14	1.12
6×2	1.00	0.0000	0.0014	0.94	1.57	0.02	3.45	0.21	2.34
2×3	1.00	0.0000	0.0000	0.10	1.00	0.16	5.86	0.00	1.20
3×3	1.00	0.0000	0.0000	0.62	1.16	0.04	4.55	0.05	18.7
4×3	1.00	0.0000	0.0004	2.11	1.28	0.02	3.91	0.13	326
5×3	1.00	0.0000	0.0000	2.28	1.18	0.01	3.10	0.07	11918

set, we determined the MLE of θ and ρ by iterating eight times the likelihood computation on a five-by-five grid of θ and ρ, refining around the (θ, ρ) pair that yielded the highest likelihood. For each simulated data set, four different computations were done: using $\mathcal{C}_k(D)$-configurations, for $k = r, r+1, r+2$, where $r := R_{min}(D)$, or using the diminishing returns stopping criteria. Simulated data sets were sufficiently small that it was possible to solve the full equation system, thus allowing approximations to be compared with the true value.

Results are summarized in Table 1. Both the likelihood itself and MLE of θ are quite accurate even for the computation based on $\mathcal{C}_r(D)$-configurations, while MLE of ρ becomes quite accurate when the equation system is expanded to $\mathcal{C}_{r+2}(D)$-configurations. Applying the diminishing returns stopping criteria does
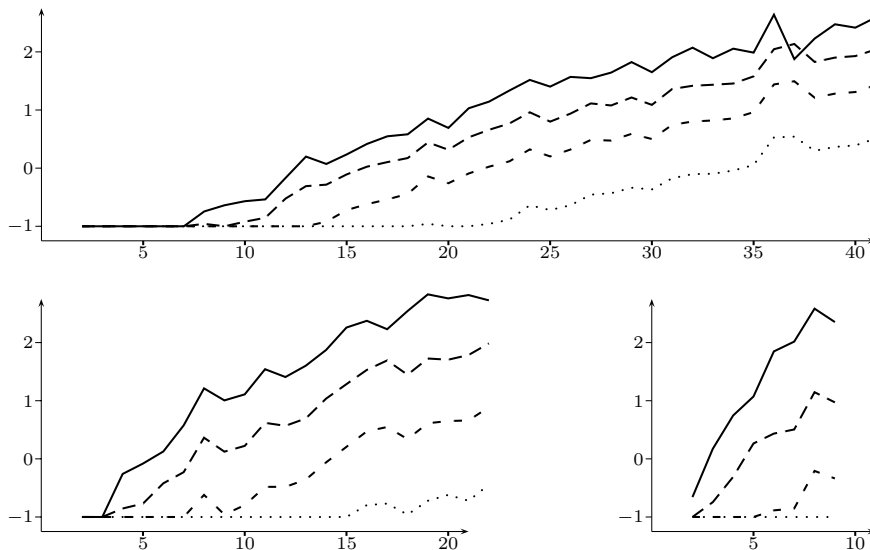
Fig. 5. Logarithm of running times in seconds as function of number of sequences for data sets with two segregating sites (top), three segregating sites (bottom left) and four segregating sites (bottom right). For each data set the likelihood at $\theta = 2$ and $\rho = 5$ was computed based on $\mathcal{C}_r(D)$ (shown in " $\cdots\cdots$ "), on $\mathcal{C}_{r+1}(D)$ (" - - - "), on $\mathcal{C}_{r+2}(D)$ (" – – "), and using the diminishing returns stopping criteria ("——"). Average running times of less than 0.1 second were truncated to 0.1 second.

slightly better than using $\mathcal{C}_{r+2}(D)$-configurations, both in terms of accuracy and time. All in all, our method is quite accurate, while being substantially faster than using the full equation system.

Being able to compare results to the true values severely limits the data set sizes that can be investigated. Even for data sets with just five sequences and three segregating sites, we experienced an average running time of more than three hours to solve the full equation system at 200 (θ, ρ) grid points to obtain the MLEs. (In contrast, our method required only a few seconds on average to obtain very accurate estimates. See Table 1.) To investigate how large of a data set our method can handle, we simulated data sets with more number of sequences, while keeping the number of segregating sites to two, three or four. We again used Hudson's [21] program ms with $\rho = 5$. For each data set, we computed the likelihood for $\theta = 2$ and $\rho = 5$. Average computation times averaged over ten simulated data sets are plotted in Fig. 5.

4   Discussion

In this paper, we have developed a novel parsimony-based, deterministic approach for accurately computing the likelihood under the coalescent with recombination. Given enough computation time and memory, our method can, in

principle, compute the exact likelihood, by finding all ancestral configurations for a given data set and then solving the full system of recursions. However, the size of the input data for which this can actually be done is severely limited. For a data set with only five sequences and three segregating sites, it currently takes several hours to obtain accurate MLEs of θ and ρ by computing the exact likelihood. Perhaps this is not so surprising, given that the total number of ACs grows very rapidly with the number of sequences and more so with the number of sites [39].

Our approximation method is based on restricting the probability recursions to certain ACs, namely those that occur in evolutionary histories with a near-minimal number of recombinations. The restricted system of recursions can be solved several orders of magnitude faster than the full recursion system, with no noticeable loss of accuracy. It dramatically increases the size of data sets for which one can compute the likelihood by solving the recursion system. For example, our approximation method takes only a few minutes to compute the probability of a data set with twenty sequences and three sites, while, in the same amount of time, one can only compute the probability of a data set with five sequences and three sites using the full equation system. However, even with the techniques introduced here, our method is limited to moderate-sized data sets. Despite the enormous reduction in time requirement of our method compared to the exact computation, the complexity of the problem grows so astronomically fast with data size that the speedup is dwarfed in comparison. For further details on this matter, we again refer to our previous work [39], where the growth of $\mathcal{C}_{R_{min}}(D)$ as a function of data size was also investigated. We believe that new insights—e.g., regarding symmetries in the recursion structure, allowing ACs to be lumped together—are required for making this kind of algorithm-based approach applicable to large data sets.

Even so, the work presented here should be useful to the researchers in statistical genetics. For moderate-sized data sets, our method can be used to develop benchmarks with very well-characterized likelihoods. Such studies can be valuable for evaluating the performance of existing and new sampling-based approaches, and for fine-tuning them. Further, as some pseudo-likelihood methods [23, 32, 33] use likelihood calculations for few (typically 2) sites, the method developed here should be useful for improving such methods.

References

1. V. Bafna and V. Bansal. The number of recombination events in a sample history: conflict graph and lower bounds. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:78–90, 2004.

2. V. Bafna and V. Bansal. Improved recombination lower bounds for haplotype data. In *Proc. 9th Annual Intl. Conf. on Research in Computational Molecular Biology (RECOMB)*, pages 569–584, 2005.
3. M. Beaumont. Detecting population expansion and decline using microsatellites. *Genetics*, 153:2013–2029, 1999.
4. M. Bordewich and C. Semple. Computing the minimum number of hybridization events for a consistent evolutionary history. discrete. *Applied Mathematics*, 155:914–928, 2007.
5. M. De Iorio and R. C. Griffiths. Importance sampling on coalescent histories. I. *Adv. Appl. Prob.*, 36:417–433, 2004.
6. M. De Iorio and R. C. Griffiths. Importance sampling on coalescent histories. II: Subdivided population models. *Adv. Appl. Prob.*, 36:434–454, 2004.
7. S. N. Ethier and R. C. Griffiths. The infinitely-many-sites model as a measure valued diffusion. *Ann. Probab.*, 15:515–545, 1987.
8. S. N. Ethier and R. C. Griffiths. On the two-locus sampling distribution. *J. Math. Biol.*, 29:131–159, 1990.
9. P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159:1299–1318, 2001.
10. P. Fearnhead and P. Donnelly. Approximate likelihood methods for estimating local recombination rates. *J. R. Statist. Soc. B*, 64:657–680, 2002.
11. Paul Fearnhead and Nick G. C. Smith. A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. *Am. J. Hum. Genet.*, 77:781–794, 2005.
12. R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, 3:479–502, 1996.
13. R. C. Griffiths and S. Tavaré. Ancestral inference in population genetics. *Stat. Sci.*, 9:307–319, 1994.
14. R. C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. *Proc. R. Soc. London B.*, 344:403–410, 1994.
15. R. C. Griffiths and S. Tavaré. Simulating probability distributions in the coalescent. *Theor. Popul. Biol.*, 46:131–159, 1994.
16. D. Gusfield. Optimal, efficient reconstruction of Root-Unknown phylogenetic networks with constrained recombination. *J. Comput. Sys. Sci.*, 70:381–398, 2005.
17. D. Gusfield, S. Eddhu, and C. Langley. The fine structure of galls in phylogenetic networks. *INFORMS J. on Computing, special issue on Computational Biology*, 16:459–469, 2004.
18. D. Gusfield, S. Eddhu, and C. Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *J. Bioinf. Comput. Biol.*, 2:173–213, 2004.
19. J. Hein. Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.*, 98:185–200, 1990.
20. J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, 36:396–405, 1993.
21. R. R. Hudson. Generating Samples under the Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338, 2002.
22. R. Hudson and N. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147–164, 1985.
23. R. R. Hudson. Two-locus sampling distributions and their application. *Genetics*, 159:1805–1817, 2001.
24. International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, 2005.

25. M. K. Kuhner, J. Yamato, and J. Felsenstein. Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. *Genetics*, 140:1421–1430, 1995.

26. M. K. Kuhner, J. Yamato, and J. Felsenstein. Maximum likelihood estimation of recombination rates from population data. *Genetics*, 156:1393–1401, 2000.

27. F. Larribe, S. Lessard, and N.J. Schork. Gene Mapping via the Ancestral Recombination Graph *Theor. Popul. Biol.*, 62:2150–229, 2002.

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

29. R. B. Lyngsø, Y. S. Song, and J. Hein. Minimum recombination histories by branch and bound. In Rita Casadio and Gene Myers, editors, *Proceedings of the 5th Workshop on Algorithms in Bioinformatics (WABI)*, number 3692 in Lecture Notes in Bioinformatics, pages 239–250. Springer, 2005.

30. G. McVean, P. Awadalla, and P. Fearnhead. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, 160:1231–1241, 2002.

31. G. McVean and N. Cardin. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 360:1387–93, 2005.

32. G.A.T. McVean, S. Myers, S. Hunt, P. Deloukas, D.R. Bentley, and P. Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304:581–584, 2004.

33. S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310:321–324, 2005.

34. S. R. Myers and R. C. Griffiths. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163:375–394, 2003.

35. K. L. Simonsen and G. A. Churchill. A Markov chain model of coalescence with recombination. *Theor. Popul. Biol.*, 52:43–59, 1997.

36. Y. S. Song and J. Hein. Parsimonious reconstruction of sequence evolution and haplotype blocks: Finding the minimum number of recombination events. In *Proc. of Workshop on Algorithms in Bioinformatics 2003*, pages 287–302, Berlin, Germany, 2003. Springer-Verlag LNCS.

37. Y. S. Song and J. Hein. On the minimum number of recombination events in the evolutionary history of DNA sequences. *J. Math. Biol.*, 48:160–186, 2004.

38. Y. S. Song and J. Hein. Constructing minimal ancestral recombination graphs. *J. Comput. Biol.*, 12:147–169, 2005.

39. Y. S. Song, R. B. Lyngsø, and J. Hein. Counting all possible ancestral configurations of sample sequences in population genetics. *IEEE Transactions on Computational Biology and Bioinformatics*, 3(3):239–251, 2006.

40. Y. S. Song, Y. Wu, and D. Gusfield. Efficient computation of close lower and upper bounds on the minimum number of needed recombinations in the evolution of biological sequences. *Proc. of ISMB 2005, Bioinformatics*, 21:i413–i422, 2005.

41. M. Stephens and P. Donnelly. Inference in molecular population genetics. *J.R. Stat. Soc. Ser. B*, 62:605–655, 2000.

42. J. D. Wall. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.*, 17:156–63, 2000.

43. L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. *J. Comput. Biol.*, 8:69–78, 2001.

44. I. J. Wilson and D. J. Balding. Genealogical inference from microsatellite data. *Genetics*, 150:499–510, 1998.