Decoding Coalescent Hidden Markov Models in Linear Time

Kelley Harris[1], Sara Sheehan[2], John A. Kamm[3], and Yun S. Song[2,3,4]

[1] Department of Mathematics, University of California, Berkeley
[2] Computer Science Division, University of California, Berkeley
[3] Department of Statistics, University of California, Berkeley
[4] Department of Integrative Biology, University of California, Berkeley
kharris@math.berkeley.edu, {ssheehan,yss}@eecs.berkeley.edu,
jkamm@stat.berkeley.edu

Abstract. In many areas of computational biology, hidden Markov models (HMMs) have been used to model local genomic features. In particular, coalescent HMMs have been used to infer ancient population sizes, migration rates, divergence times, and other parameters such as mutation and recombination rates. As more loci, sequences, and hidden states are added to the model, however, the runtime of coalescent HMMs can quickly become prohibitive. Here we present a new algorithm for reducing the runtime of coalescent HMMs from quadratic in the number of hidden time states to linear, without making any additional approximations. Our algorithm can be incorporated into various coalescent HMMs, including the popular method PSMC for inferring variable effective population sizes. Here we implement this algorithm to speed up our demographic inference method diCal, which is equivalent to PSMC when applied to a sample of two haplotypes. We demonstrate that the linear-time method can reconstruct a population size change history more accurately than the quadratic-time method, given similar computation resources. We also apply the method to data from the 1000 Genomes project, inferring a high-resolution history of size changes in the European population.

Keywords: Demographic inference, effective population size, coalescent with recombination, expectation-maximization, augmented hidden Markov model, human migration out of Africa.

1 Introduction

The hidden Markov model (HMM) is a natural and powerful device for learning functional and evolutionary attributes of DNA sequence data. Given an emitted sequence of base pairs or amino acids, the HMM is well-suited to locating hidden features of interest such as genes and promotor regions [2,5]. HMMs can also be used to infer hidden attributes of a collection of related DNA sequences. In this case, emitted states are a tuple of A's, C's, G's and T's, and the diversity of emitted states in a particular region can be used to infer the local evolutionary history of the sequences. When two sequences are identical throughout a long

R. Sharan (Ed.): RECOMB 2014, LNBI 8394, pp. 100–114, 2014.

genetic region, they most likely inherited that region identical by descent from a recent common ancestor. Conversely, high genetic divergence indicates that the sequences diverged from a very ancient common ancestor [1,15].

In recent years, coalescent HMMs such as the Pairwise Sequentially Markov Coalescent (PSMC) [15] have been used to infer the sequence of times to most recent common ancestor (TMRCAs) along a pair of homologous DNA sequences. Two other coalescent HMMs (CoalHMM [4,12,16] and diCal [24,25]) also tackle the problem of inferring genealogical information in samples of more than two haplotypes. These methods are all derived from the coalescent with recombination, a stochastic process that encapsulates the history of a collection of DNA sequences as an ancestral recombination graph (ARG) [13,29]. The hidden state associated with each genetic locus is a tree with time-weighted edges, and neighboring trees in the sequence are highly correlated with each other. Sequential changes in tree structure reflect the process of genetic recombination that slowly breaks up ancestral haplotypes over time.

The methods mentioned above all infer approximate ARGs for the purpose of demographic inference, either detecting historical changes in effective population size or estimating times of divergence and admixture between different populations or species. PSMC and CoalHMM have been used to infer ancestral population sizes in a variety of non-model organisms for which only a single genome is available [6,17,19,20,28,30], as well as for the Neanderthal and Denisovan archaic hominid genomes [18]. Despite this progress, the demographic inference problem is far from solved, even for extremely well-studied species like *Homo sapiens* and *Drosophila melanogaster* [7,9,15,23,27]. Estimates of the population divergence time between European and African humans range from 50 to 120 thousand years ago (kya), while estimates of the speciation time between polar bears and brown bears range from 50 kya to 4 million years ago [3,10,19]. One reason that different demographic methods often infer conflicting histories is that they make different trade-offs between the mathematical precision of the model and scalability to larger input datasets. This is even true within the class of coalescent HMMs, which are much more similar to each other than to methods that infer demography from summary statistics [8,11,21] or Markov chain Monte Carlo [7].

Exact inference of the posterior distribution of ARGs given data is a very challenging problem, the major reason being that the space of hidden states is infinite, parameterized by continuous coalescence times. In practice, when a coalescent HMM is implemented, time needs to be discretized and confined to a finite range of values. It is a difficult problem to choose an optimal time discretization that balances the information content of a dataset, the complexity of the analysis, and the desire to infer particular periods of history at high resolution. Recent demographic history is often of particular interest, but large sample sizes are needed to distinguish between the population sizes at time points that are very close together or very close to the present.

In a coalescent HMM under a given demographic model, optimal demographic parameters can be inferred using an expectation-maximization (EM) algorithm. The speed of this EM algorithm is a function of at least three variables: the length

L of the genomic region being analyzed, the number n of sampled haplotypes, and the number d of states for discretized time. In most cases, the complexity is linear in L, but the complexity in n can be enormous because the number of distinct n-leaved tree topologies grows super-exponentially with n. PSMC and CoalHMM avoid this problem by restricting n to be very small, analyzing no more than four haplotypes at a time. diCal admits larger values of n by using a *trunk genealogy* approximation (see [22,24,25] for details) which is derived from the diffusion process dual to the coalescent process, sacrificing information about the exact structure of local genealogies in order to analyze large samples which are informative about the recent past.

To date, all published coalescent HMMs have had quadratic complexity in d. This presents a significant limitation given that small values of d lead to biased parameter estimates [16] and limit the power of the method to resolve complex demographic histories. PSMC is typically run with a discretization of size $d = 64$, but diCal and CoalHMM analyses of larger datasets are restricted to coarser discretizations by the cost of increasing the sample size. In this paper, we exploit the natural symmetries of the coalescent process to derive an alternate EM algorithm with linear complexity in d. The speedup requires no approximations to the usual forward-backward probabilities; we perform an exact computation of the likelihood in $O(d)$ time rather than $O(d^2)$ time using an augmented HMM. We implement the algorithms presented in this paper to speed up our published method diCal, which is equivalent to PSMC when the sample size is two, yielding results of the same quality as earlier work in a fraction of the runtime. We have included the speedup in the most recent version of our program diCal; source code can be downloaded at `http://sourceforge.net/projects/dical/`.

2 Linear-Time Computation of the Forward and Backward Probabilities

We consider a coalescent HMM $\mathcal{M}$ with hidden states $S_1, \ldots, S_L$ and observations $x = x_1, \ldots, x_L$. For PSMC, S_ℓ is the discretized time interval in which two homologous chromosomes coalesce at locus ℓ, while x_ℓ is an indicator for heterozygosity. The method diCal is based on the conditional sampling distribution (CSD) which describes the probability of observing a newly sampled haplotype x given a collection $\mathcal{H}$ of n already observed haplotypes. In diCal, the hidden state at locus ℓ is $S_\ell = (H_\ell, T_\ell)$, where $H_\ell \in \mathcal{H}$ denotes the haplotype in the "trunk genealogy" (see [22]) with which x coalesces at locus ℓ and $T_\ell \in \{1, \ldots, d\}$ denotes the discretized time interval of coalescence; the observation $x_\ell \in \mathcal{A}$ is the allele of haplotype x at locus ℓ. For $n = |\mathcal{H}| = 1$, diCal is equivalent to PSMC. In what follows, we present our algorithm in the context of diCal, but we note that the same underlying idea can be applied to other coalescent HMMs.

2.1 A Linear-Time Forward Algorithm

We use $f(x_{1:\ell}, (h, j))$ to denote the joint forward probability of observing the partial emitted sequence $x_{1:\ell} := x_1, \ldots, x_\ell$ and the hidden state $S_\ell = (h, j)$ at

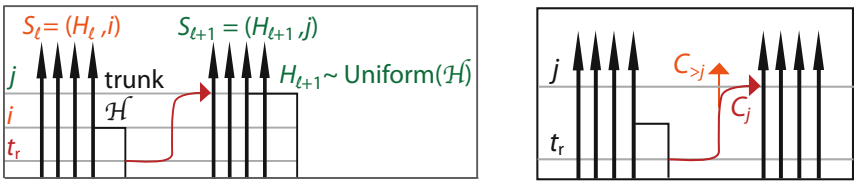


Fig. 1. (a). Here, we illustrate a transition from hidden state $S_\ell = (h_n, i)$ to hidden state $S_{\ell+1} = (h_k, j)$ that proceeds via recombination at time t_r. The probability of this transition does not depend on the identity of the haplotype h_k.(b). As a recombined lineage floats through time interval j, it can either coalesce with the trunk (event C_j) or keep floating (event $C_{>j}$) and eventually coalesce with the trunk in a more ancient time interval.

locus ℓ. The probability of transitioning from state (h', k) at locus ℓ to state (h, j) at locus $\ell + 1$ is denoted by $\phi(h, j \mid h', k)$, the stationary probability of state (h, i) is denoted $\zeta(h, i)$, and the emission probability of the observed allele $x_\ell = a$ given coalescence at $T_\ell = j$ onto haplotype h with allele $h_\ell = b$ at locus ℓ is denoted by $\xi(a \mid b, j)$. When ℓ is obvious from the context, we sometimes use $\xi(a \mid s) := \xi(a \mid h_\ell, j)$ for $s = (h, j)$. Explicit expressions for $\zeta(h, i)$, $\phi(h, j \mid h', k)$, and $\xi(a \mid b, j)$ in the context of our program diCal are given in [24].

The forward probabilities are computed using the recursion

$$f(x_{1:\ell+1}, (h, j)) = \xi(x_{\ell+1} | h_{\ell+1}, j) \cdot \sum_{k=1}^{d} \sum_{h' \in \mathcal{H}} f(x_{1:\ell}, (h', k)) \cdot \phi(h, j | h', k), \quad (1)$$

which contains nd terms. Since there are also nd possibilities for $S_{\ell+1} = (h, j)$, it should naively take $O(n^2 d^2 L)$ time to compute the entire forward dynamic programming (DP) table $\{f(x_{1:\ell}, S_\ell)\}_{\ell=1}^{L}$. The key to achieving a speed-up is to factor (1) in a way that reflects the structure of the coalescent, exploiting the fact that many transitions between different hidden states have identical probabilities.

After a sampled lineage recombines at time t_r between loci ℓ and $\ell + 1$, it will "float" backward in time from the recombination breakpoint until eventually coalescing with a trunk lineage chosen uniformly at random (Figure 1a). This implies that $\phi(h, j|h', k) = \phi(h, j|h'', k)$ whenever $h' \neq h$ and $h'' \neq h$, and exploiting this symmetry allows the forward table to be computed in $O(nd^2 L)$ time. This speed-up was already implemented in the algorithm described in Paul *et al.* [22].

Another symmetry of the transition matrix, not exploited previously, can be found by decomposing the transition from locus ℓ to locus $\ell + 1$ as a sequence of component events. In particular, let R_i be the event that a recombination occurs during time interval i, and let $\overline{R}$ be the event that no recombination occurs between ℓ and $\ell + 1$. Then we have that

$$\phi((h,j) \mid (h',k)) = \frac{1}{n} \sum_{i=1}^{\min(j,k)} \left(\mathbb{P}(R_i, T_{\ell+1} = j \mid T_\ell = k) + \mathbb{1}_{\{(h,j)=(h',k')\}} \mathbb{P}(\overline{R} \mid T_\ell = k)\right), \tag{2}$$

where $\mathbb{1}_E = 1$ if the event E is true or 0 otherwise. The factor $1/n$ corresponds to the probability that the sampled lineage coalesces with haplotype $h \in \mathcal{H}$ in the trunk genealogy.

If a recombination occurs in time interval i, the sampled lineage will start to "float" freely back in time until it either coalesces in i or floats into the next time interval $i+1$ (Figure 1b). Specifically, we let $C_{>i}$ denote the event where the sampled lineage recombines at or before i and floats into $i+1$, and C_i denote the event where the recombined lineage coalesces back in interval i. Noting that $\mathbb{P}(R_i, C_i \mid T_\ell = i')$ and $\mathbb{P}(R_i, C_{>i} \mid T_\ell = i')$ are independent of i' whenever $i' > i$, and that coalescence happens as a Markov process backwards in time, we obtain

$$\begin{aligned}\mathbb{P}(R_i, T_{\ell+1} = j \mid T_\ell = k) = {} & \mathbb{1}_{i=j=k} \cdot \mathbb{P}(R_i, C_i \mid T_\ell = i) \\ & + \mathbb{1}_{i=j<k} \cdot \mathbb{P}(R_i, C_i \mid T_\ell > i) \\ & + \mathbb{1}_{i=k<j} \cdot \mathbb{P}(R_i, C_{>i} \mid T_\ell = i) \cdot \prod_{k=i}^{j-1} \mathbb{P}(C_{>k+1} \mid C_{>k}) \\ & + \mathbb{1}_{i<\min(j,k)} \cdot \mathbb{P}(R_i, C_{>i} \mid T_\ell > i) \cdot \prod_{k=i}^{j-1} \mathbb{P}(C_{>k+1} \mid C_{>k}).\end{aligned} \tag{3}$$

Explicit formulas (specific to the method diCal) for the above terms are provided in the supporting information available at `http://www.eecs.berkeley.edu/~yss/publications.html`.

By combining (2) with (3) and then collecting terms in (1), we can remove the sum over $T_\ell = k$ when computing $f(x_{1:\ell+1}, S_{\ell+1})$. In particular, we define additional forward probabilities

$$f\left(x_{1:\ell}, T_\ell = k\right) := \mathbb{P}(x_{1:\ell}, T_\ell = k) = \sum_{h' \in \mathcal{H}} f(x_{1:\ell}, S_\ell = (h',k)), \tag{4}$$

$$f\left(x_{1:\ell}, T_\ell > k\right) := \mathbb{P}(x_{1:\ell}, T_\ell > k) = \sum_{k'=k+1}^{d} \sum_{h' \in \mathcal{H}} f(x_{1:\ell}, S_\ell = (h',k')), \tag{5}$$

$$\begin{aligned} f\left(x_{1:\ell}, R_{\le j}, C_{>j}\right) &:= \sum_{i=1}^{j} \mathbb{P}(x_{1:\ell}, R_i, C_{>i}, \ldots, C_{>j}) \\ &= \sum_{i=1}^{j} \left\{ \left[\prod_{i'=i+1}^{j} \mathbb{P}(C_{>i'} \mid C_{>i'-1}) \right] \right. \\ &\left. \times \left[f\left(x_{1:\ell}, T_\ell = i\right) \mathbb{P}(R_i, C_{>i} \mid T_\ell = i) + f\left(x_{1:\ell}, T_\ell > i\right) \mathbb{P}(R_i, C_{>i} \mid T_\ell > i) \right] \right\}. \end{aligned} \tag{6}$$

Then, (1) can be written as

$$f(x_{1:\ell+1},(h,j)) = \xi(x_{\ell+1} \mid h_{\ell+1}, j) \cdot \Bigg[\frac{1}{n} f\left(x_{1:\ell}, R_{\leq j-1}, C_{>j-1}\right) \mathbb{P}(C_j | C_{>j-1}) + \frac{1}{n} f\left(x_{1:\ell}, T_\ell > j\right) \mathbb{P}(R_j, C_j \mid T_\ell > j) + \frac{1}{n} f\left(x_{1:\ell}, T_\ell = j\right) \mathbb{P}(R_j, C_j \mid T_\ell = j) + f(x_{1:\ell},(h,j)) \mathbb{P}(\overline{R} \mid T_\ell = j)\Bigg]. \tag{7}$$

This can be seen by noting that the first three terms in the sum correspond to the terms for $i < j$, $i = j < k$, and $i = j = k$, respectively when putting together (1) and (2). Alternatively, (7) follows from directly considering the probabilistic interpretation of the terms $f(x_{1:\ell}, *)$ as given by (4), (5), and (6).

The required values of $f\left(x_{1:\ell}, R_{\leq i}, C_{>i}\right)$ and $f\left(x_{1:\ell}, T_\ell > i\right)$ can be computed recursively using

$$f\left(x_{1:\ell}, T_\ell > i\right) = f\left(x_{1:\ell}, T_\ell > i+1\right) + f\left(x_{1:\ell}, T_\ell = i+1\right), \tag{8}$$

$$f\left(x_{1:\ell}, R_{\leq i}, C_{>i}\right) = f\left(x_{1:\ell}, R_{\leq i-1}, C_{>i-1}\right) \mathbb{P}(C_{>i} | C_{>i-1}) + f\left(x_{1:\ell}, T_\ell = i\right) \mathbb{P}(R_i, C_{>i} \mid T_\ell = i) + f\left(x_{1:\ell}, T_\ell > i\right) \mathbb{P}(R_i, C_{>i} \mid T_\ell > i), \tag{9}$$

with the base cases

$$f\left(x_{1:\ell}, T_\ell > d\right) = 0,$$

$$f\left(x_{1:\ell}, R_{\leq 1}, C_{>1}\right) = f\left(x_{1:\ell}, T_\ell > 1\right) \mathbb{P}(R_1, C_{>1} \mid T_\ell > 1) + f\left(x_{1:\ell}, T_\ell = 1\right) \mathbb{P}(R_1, C_{>1} \mid T_\ell = 1).$$

Hence, using the recursions (7), (8), and (9), it is possible to compute the entire forward DP table $\{f(x_{1:\ell}, S_\ell)\}_{\ell=1}^{L}$ exactly in $O(ndL)$ time.

2.2 A Linear-Time Backward Algorithm

The backward DP table $\{b(x_{\ell+1:L} \mid S_\ell)\}$ can be also computed in $O(ndL)$ time. Given the linear-time forward algorithm discussed in the previous section, the easiest way to compute the backward DP table is as follows: Let $x^{(r)} = x_1^{(r)}, x_2^{(r)}, \ldots, x_L^{(r)} = x_L, x_{L-1}, \ldots, x_1$ denote the reversed x and let $S_\ell^{(r)}$ denote the hidden states for the HMM generating $x^{(r)}$. Then, since the coalescent is reversible along the sequence,

$$b(x_{\ell+1:L}^{(r)} \mid s) = \frac{\mathbb{P}(x_{\ell+1:L}^{(r)}, S_\ell = s)}{\zeta(s)} = \frac{\mathbb{P}(x_{\ell:L}^{(r)}, S_\ell = s)}{\xi(x_\ell^{(r)} \mid s)\zeta(s)} = \frac{f(x_{1:L-\ell+1}^{(r)}, S_{L-\ell+1}^{(r)} = s)}{\xi(x_\ell^{(r)} \mid s)\zeta(s)}.$$

3 Linear-Time EM via an Augmented HMM

The primary application of PSMC and diCal is parameter estimation, specifically the estimation of demographic parameters such as changing population sizes. This is done through a maximum likelihood framework with the expectation maximization (EM) algorithm. In this section, we describe how to speed up the EM algorithm to work in linear time.

3.1 The Standard EM Algorithm with $O(d^2)$ Time Complexity

Let Θ denote the parameters we wish to estimate, and $\hat{\Theta}$ denote the maximum likelihood estimate:

$$\hat{\Theta} = \arg\max_{\Theta'} \mathcal{L}(\Theta') = \arg\max_{\Theta'} \mathbb{P}_{\Theta'}(x_{1:L}).$$

To find $\hat{\Theta}$, we pick some initial value $\Theta^{(0)}$, and then iteratively solve for $\Theta^{(t)}$ according to

$$\Theta^{(t)} = \arg\max_{\Theta'} \mathbb{E}_{S_{1:L};\Theta^{(t-1)}}[\log \mathbb{P}_{\Theta'}(x_{1:L}, S_{1:L}) \mid x_{1:L}],$$

where $S_{1:L} := S_1, \ldots, S_L$. The sequence $\Theta^{(0)}, \Theta^{(1)}, \ldots$ is then guaranteed to converge to a local maximum of the surface $\mathcal{L}(\Theta)$.

Since $(x_{1:L}, S_{1:L})$ forms an HMM, the joint likelihood $\mathbb{P}(x_{1:L}, S_{1:L})$ can be written as

$$\mathbb{P}_{\Theta'}(x_{1:L}, S_{1:L}) = \zeta_{\Theta'}(S_1) \left[\prod_{\ell=1}^{L} \xi_{\Theta'}(x_\ell \mid S_\ell)\right] \left[\prod_{\ell=2}^{L} \phi_{\Theta'}(S_\ell \mid S_{\ell-1})\right].$$

Letting $\mathbb{E}[\#\ell : E \mid x_{1:L}]$ denote the posterior expected number of loci where event E occurs, and $\pi(x) := \mathbb{P}(x) = \sum_s f(x_{1:L}, s)$ denote the total probability of observing x, we then have

$$\begin{aligned}
&\mathbb{E}_{S_{1:L};\Theta}\left[\log \mathbb{P}_{\Theta'}(x_{1:L}, S_{1:L}) \middle| x_{1:L}\right] \\
=&\sum_s (\log \zeta_{\Theta'}(s))\, \mathbb{P}_{\Theta}(S_1 = s | x_{1:L}) \\
&\quad + \sum_{(h,i)} \sum_{a,b \in \mathcal{A}} (\log \xi_{\Theta'}(a|b,i))\, \mathbb{E}_{\Theta}\left[\#\ell : \{S_\ell = (h,i), h_\ell = b, x_\ell = a\} \middle| x_{1:L}\right] \\
&\quad + \sum_{s,s'} (\log \phi_{\Theta'}(s' \mid s))\, \mathbb{E}_{\Theta}\left[\#\ell : \{S_{\ell-1} = s, S_\ell = s'\} \middle| x_{1:L}\right] \\
=&\frac{1}{\pi_\Theta(x)} \Bigg[\sum_s (\log \zeta_{\Theta'}(s))\, f_\Theta(x_1, s) b_\Theta(x_{2:L}|s) \\
&\quad + \sum_{(h,i)} \sum_{a,b \in \mathcal{A}} (\log \xi_{\Theta'}(a|b,i)) \sum_{\substack{\ell : x_\ell = a \\ h_\ell = b}} f_\Theta(x_{1:\ell}, (h,i)) b_\Theta(x_{\ell+1:L}|h,i)
\end{aligned}$$

$$+\sum_{s,s'}(\log\phi_{\Theta'}(s' \mid s))\left(\sum_{\ell=1}^{L-1} f_\Theta(x_{1:\ell}, s)\phi_\Theta(s' \mid s)\xi_\Theta(x_{\ell+1} \mid s')b_\Theta(x_{\ell+2:L} \mid s')\right)\Bigg]. \tag{10}$$

Note that we have to compute the term $\sum_\ell f_\Theta(x_{1:\ell}, s)\phi_\Theta(s' \mid s)\xi_\Theta(x_{\ell+1} \mid s')$ $b_\Theta(x_{\ell+2:L} \mid s')$ for every pair of states s, s', which makes computing the EM objective function quadratic in the number d of discretization time intervals, despite the fact that we computed the forward and backward tables in linear time.

3.2 A Linear-Time EM Algorithm

By augmenting our HMM to condition on whether recombination occurred between loci ℓ and $\ell+1$, the EM algorithm can be sped up to be linear in d. We now describe this augmented HMM. Let $\mathcal{M}$ denote our original HMM, with states $S_{1:L}$ and observations $x_{1:L}$. Between loci ℓ and $\ell+1$, define

$$\mathcal{R}_{l,l+1} = \begin{cases} \overline{R}, & \text{if no recombination,} \\ R_i, & \text{if recombination occurred at time } i. \end{cases}$$

Now let $S_1^* = S_1$, and $S_\ell^* = (\mathcal{R}_{\ell-1,\ell}, S_\ell)$ for $\ell > 1$. We let $\mathcal{M}^*$ be the HMM with hidden variables $S_{1:L}^* = S_1^*, \ldots, S_L^*$, observations $x_{1:L}$, transition probabilities $\mathbb{P}(S_\ell^* \mid S_{\ell-1}^*) = \mathbb{P}(S_\ell^* \mid S_{\ell-1})$, and emission probabilities $\mathbb{P}(x_\ell \mid S_\ell^*) = \mathbb{P}(x_\ell \mid S_\ell)$. Note that the probability of observing the data is the same under $\mathcal{M}$ and $\mathcal{M}^*$, i.e.,

$$\mathcal{L}(\Theta) = \mathbb{P}_\Theta(x_{1:L} \mid \mathcal{M}) = \mathbb{P}_\Theta(x_{1:L} \mid \mathcal{M}^*),$$

and so we may find a local maximum of $\mathcal{L}(\Theta)$ by applying the EM algorithm to the augmented HMM $\mathcal{M}^*$, instead of to the original HMM $\mathcal{M}$.

To compute the EM objective function for $\mathcal{M}^*$, we start by noting that the joint likelihood is

$$\begin{aligned}\mathbb{P}(x_{1:L}, S_{1:L}^*) = \zeta(S_1)\left[\prod_{\ell=1}^{L}\xi(x_\ell \mid S_\ell)\right]\left[\prod_{\ell:\mathcal{R}_{\ell,\ell+1}=\overline{R}}\mathbb{P}(\overline{R} \mid T_\ell)\right] \\ \times\left[\prod_{i=1}^{d}\prod_{\ell:\mathcal{R}_{\ell,\ell+1}=R_i}\mathbb{P}(R_i, T_{\ell+1} \mid T_\ell)\right]\left(\frac{1}{n}\right)^{\#\ell:\mathcal{R}_{\ell,\ell+1}\neq\overline{R}},\end{aligned} \tag{11}$$

where we decomposed the joint likelihood into the initial probability, the emission probabilities, the transitions without recombination, and the transitions with recombination. We note that the initial probability can be decomposed as

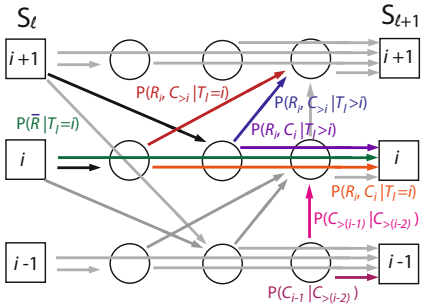


Fig. 2. This diagram illustrates the flow of transition probabilities through the augmented HMM. Lineages may transition between different coalescence times at loci ℓ and $\ell + 1$ by recombining and passing through the floating states represented by circles. Each interval contains three distinct floating states to capture the the dependence of recombination and coalescence probabilities on whether any of these events occur during the same time interval.

$$\zeta(S_1 = (h, j)) = \frac{1}{n}\left[\prod_{i=1}^{j-1} \mathbb{P}(C_{>i} \mid C_{>i-1})\right]\mathbb{P}(C_j \mid C_{>j-1}), \tag{12}$$

and from (3), we decompose the product of transition recombination probabilities as

$$\begin{aligned}\prod_{i=1}^{d} \prod_{\ell:\mathcal{R}_{\ell,\ell+1}=R_i} \mathbb{P}(R_i, T_{\ell+1} \mid T_\ell) = \prod_{i=1}^{d} \Bigg\{ &\left[\prod_{\substack{\ell:\mathcal{R}_{\ell,\ell+1}=R_i \\ T_\ell = T_{\ell+1} = i}} \mathbb{P}(R_i, C_i \mid T_\ell = i)\right] \\ &\times \left[\prod_{\substack{\ell:\mathcal{R}_{\ell,\ell+1}=R_i \\ T_\ell > T_{\ell+1} = i}} \mathbb{P}(R_i, C_i \mid T_\ell > i)\right]\left[\prod_{\substack{\ell:\mathcal{R}_{\ell,\ell+1}=R_i \\ T_{\ell+1} > T_\ell = i}} \mathbb{P}(R_i, C_{>i} \mid T_\ell = i)\right] \\ &\times \left[\prod_{\substack{\ell:\mathcal{R}_{\ell,\ell+1}=R_i \\ T_\ell, T_{\ell+1} > i}} \mathbb{P}(R_i, C_{>i} \mid T_\ell > i)\right]\left[\prod_{\substack{\ell:T_{\ell+1}>i \\ \mathcal{R}_{\ell,\ell+1} \in R_{<i}}} \mathbb{P}(C_{>i} \mid C_{>i-1})\right] \\ &\times \left[\prod_{\substack{\ell:T_{\ell+1}=i \\ \mathcal{R}_{\ell,\ell+1} \in R_{<i}}} \mathbb{P}(C_i \mid C_{>i-1})\right]\Bigg\},\end{aligned} \tag{13}$$

where $R_{<i} := \cup_{j<i} R_j$. Figure 2 shows a graphical representation for the transitions of $\mathcal{M}^*$.

By plugging (12) and (13) into (11), then taking the posterior expected logarithm of (11), we obtain the EM objective function for $\mathcal{M}^*$:

$$\mathbb{E}_{S^*_{1:L};\Theta}\Big[\log \mathbb{P}_{\Theta'}(x_{1:L}, S^*_{1:L}) \mid x_{1:L}\Big] = -L \log n + \sum_{i=1}^{d} q_i(\Theta, \Theta'), \tag{14}$$

where

$$
\begin{aligned}
q_i(\Theta, \Theta') := & \sum_{a,b \in \mathcal{A}} \left[\frac{\log \xi_{\Theta'}(a \mid b, i)}{\pi_\Theta(x)} \sum_{\ell : x_\ell = a} \sum_{h : h_\ell = b} f_\Theta(x_{1:\ell}, (h,i)) b_\Theta(x_{\ell+1:L} \mid (h,i)) \right] \\
& + \left(\log \mathbb{P}_{\Theta'}(\overline{R} \mid T = i) + \log n \right) \mathbb{E}_\Theta \left[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = \overline{R}, T_\ell = i\} \mid x_{1:L} \right] \\
& + \left(\log \mathbb{P}_{\Theta'}(R_i, C_i \mid T = i) \right) \mathbb{E}_\Theta \left[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_\ell = T_{\ell+1} = i\} \mid x_{1:L} \right] \\
& + \left(\log \mathbb{P}_{\Theta'}(R_i, C_i \mid T > i) \right) \mathbb{E}_\Theta \left[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_\ell > T_{\ell+1} = i\} \mid x_{1:L} \right] \\
& + \left(\log \mathbb{P}_{\Theta'}(R_i, C_{>i} \mid T = i) \right) \mathbb{E}_\Theta \left[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_{\ell+1} > T_\ell = i\} \mid x_{1:L} \right] \\
& + \left(\log \mathbb{P}_{\Theta'}(R_i, C_{>i} \mid T > i) \right) \mathbb{E}_\Theta \left[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_\ell > i, T_{\ell+1} > i\} \mid x_{1:L} \right] \\
& + \left(\log \mathbb{P}_{\Theta'}(C_{>i} \mid C_{>i-1}) \right) \mathbb{E}_\Theta \left[\#\ell : \{\mathcal{R}_{\ell,\ell+1} \in R_{<i}, T_{\ell+1} > i\} \mid x_{1:L} \right] \\
& + \left(\log \mathbb{P}_{\Theta'}(C_i \mid C_{>i-1}) \right) \mathbb{E}_\Theta \left[\#\ell : \{\mathcal{R}_{\ell,\ell+1} \in R_{<i}, T_{\ell+1} = i\} \mid x_{1:L} \right] \\
& + \mathbb{P}_\Theta(T_1 > i \mid x_{1:L}) + \mathbb{P}_\Theta(T_1 = i \mid x_{1:L}). \qquad (15)
\end{aligned}
$$

The computation time for each of the posterior expectations $\mathbb{E}_\Theta[\#\ell : * \mid x_{1:L}]$ and $\mathbb{P}_\Theta(T_1 \mid x_{1:L})$ does not depend on d; full expressions are listed in the supporting information (`http://www.eecs.berkeley.edu/~yss/publications.html`). Hence, the number of operations needed to evaluate (14) is linear in d.

We note another attractive property of (14). By decomposing the EM objective function into a sum of terms $q_i(\Theta, \Theta')$, we obtain a natural strategy for searching through the parameter space. In particular, one can attempt to find the $\arg\max_{\Theta'}$ of (14) by optimizing the $q_i(\Theta, \Theta')$ one at a time in i. In fact, for the problem of estimating changing population sizes, $q_i(\Theta, \Theta')$ depends on Θ' almost entirely through a single parameter (the population size λ'_i in interval i), and we pursue a strategy of iteratively solving for λ'_i while holding the other coordinates of Θ' fixed, thus reducing a multivariate optimization problem into a sequence of univariate optimization problems.

Although both the linear and quadratic EM procedures are guaranteed to converge to local maxima of $\mathcal{L}(\Theta)$, they may have different rates of convergence, and may converge to different local maxima. The search paths of the two EM algorithms may differ for two reasons: first, the intermediate objective functions (10) and (14) are not equal, and secondly, as discussed above, we use different search strategies to find the optima of (10) and (14). We have no proven guarantee that either search should perform better than the other, but our observations indicate that the linear-time EM algorithm typically converges to a value of Θ with a equal or higher value of $\mathcal{L}(\Theta)$ than the quadratic-time algorithm, in a fraction of the time (see Figure 5 for an example).

4 Results

To confirm the decrease in runtime, we ran the linear-time diCal method on simulated data with $L = 2$ Mb of loci and 2 haplotypes (in which case diCal is equivalent to PSMC), using $d = 2, 4, 8, 16, 32, 48, 64, 80, 96, 112, 128$ discretization intervals. To simulate the data, we used `ms` [14] with a population-scaled recombination rate $\rho = 0.0005$ to generate an ARG, and then added mutations using a population-scaled mutation rate of $\theta = 0.0029$ and a finite-sites mutation matrix described in Sheehan *et al.* [24]. Figure 3(a) shows the time required to compute the table of forward probabilities. We also measured the time required for one EM iteration and then extrapolated to 20 iterations to approximate the time required to estimate an effective population size history (Figure 3(b)). In both figures, the linear runtime of our new algorithm is apparent and significantly improves our ability to increase the number of discretization intervals.

To assess the gain in accuracy of population size estimates that is afforded by more discretization intervals, we ran both the linear- and quadratic-time methods on simulated data with 10 haplotypes and $L = 2$ Mb. The conditional sampling distribution was used in a leave-one-out composite likelihood approach [24] in this experiment. To run each method for roughly the same amount of time ($\approx$ 40 hours), we used $d = 9$ for the quadratic method and $d = 21$ for the linear method. For both methods, we ran the EM for 20 iterations and inferred $d/3$ size change parameters. As measured by the PSMC error function, which integrates the absolute value of the difference between the true size function and the estimated size function [15], larger values of d permit the inference of more accurate histories.

We also ran our method on 10 CEU haplotypes (Utah residents of European descent) sequenced during Phase I of the the 1000 Genomes Project [26] (Figure 4(b)). We can see that for the quadratic method with $d = 9$, we are unable

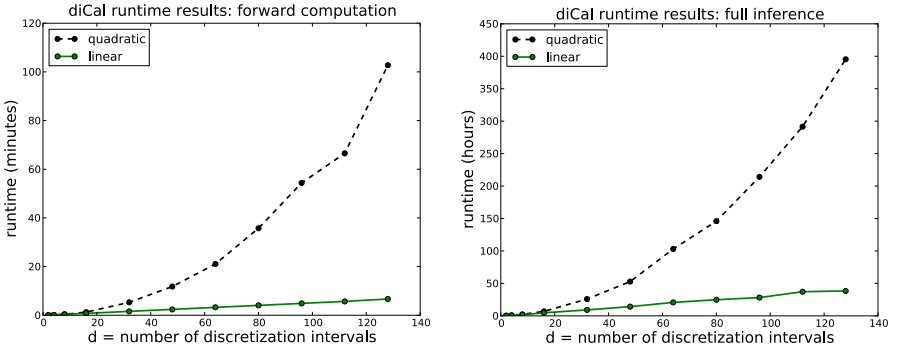


Fig. 3. Runtime results on simulated data with $L = 2$ Mb and 2 haplotypes, for varying number d of discretization intervals. (a) Runtime results (in minutes) for the forward computation. (b) Runtime results (in hours) for the entire EM inference algorithm (20 iterations) extrapolated from the time for one iteration.

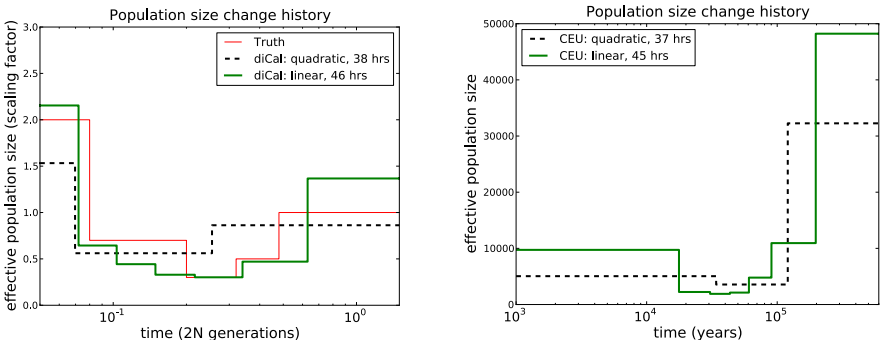


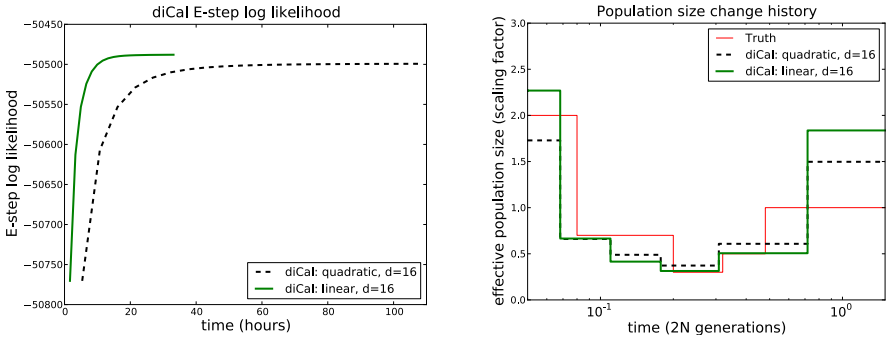
Fig. 4. Effective population size change history results. The speedup from the linear method allows us to use a finer discretization ($d = 21$) than the quadratic method ($d = 9$) for about the same amount of runtime. (a) Results on simulated data with $L = 2$ Mb and 10 haplotypes. Using the quadratic method with $d = 9$, the error was 0.148. Using the linear method with $d = 21$, the error dropped to 0.079. (b) Results on 10 European haplotypes over a 2 Mb region of chromosome 1. The out-of-Africa bottleneck is very apparent with $d = 21$, but is not as well characterized for $d = 9$.

to fully characterize the out-of-Africa bottleneck. In the same amount of computational time, we can run the linear method with $d = 21$ and easily capture this feature. The disagreement in the ancient past between the two methods is most likely due to diCal's lack of power in the ancient past when there are not many coalescence events. Using a leave-one-out approach with 10 haplotypes, the coalescence events in the ancient past tend to be few and unevenly spaced, resulting in a less confident inference.

The runtime of the full EM algorithm depends on the convergence of the M-step, which can be variable. Occasionally we observed convergence issues for the quadratic method, which requires a multivariate optimization routine. For the linear method, we used the univariate Brent optimization routine from Apache Math Commons (`http://commons.apache.org/proper/commons-math/`), which converges quickly and to a large extent avoids local maxima.

To examine the convergence of the two EM algorithms, we ran the linear and quadratic methods on the simulated data with 10 haplotypes and the same number of intervals $d = 16$. We examine the likelihoods in Figure 5(a). The linear method reaches parameter estimates of higher likelihood, although it is unclear whether the two methods have found different local maxima, or whether the quadratic method is approaching the same maximum more slowly. Figure 5(b) shows the inferred population sizes for each method, which although similar, are not identical.

We have also looked at the amount of memory required for each method, and although the difference is small, the linear method does require more memory to store the augmented forward and backward tables. A more thorough investigation of memory requirements will be important as the size of the data continues to increase.



(a) (b)

Fig. 5. Results on simulated data, using the same discretization for the linear and quadratic methods. Each method was run for 20 iterations. (a) The log likelihood of the EM algorithm, plotted against time, for both the linear and quadratic methods. (b) Population size change history results for the linear and quadratic methods, run with the same discretization using $d = 16$ and estimating 6 parameters.

5 Discussion

The improvement to diCal described in this paper will enable users to analyze larger datasets and infer more detailed demographic histories. This is especially important given that large datasets are needed to distinguish between histories with subtle or recent differences. By using samples of 10 haplotypes rather than 2, diCal v1.0 [24] was able to distinguish between histories that diverged from each other less than 0.1 coalescent time units ago, in which period PSMC tends to exhibit runaway behavior and hence cannot produce reliable population size estimates. The faster algorithm described here can handle samples of 30 haplotypes with equivalent computing resources. Our results indicate that this improves the method's ability to resolve rapid, recent demographic shifts.

In organisms where multiple sequenced genomes are not available, the resources freed up by $O(d)$ HMM decoding could be used to avoid grouping sites into 100-locus bins. This binning technique is commonly used to improve the scalability of PSMC, but has the potential to downwardly bias coalescence time estimates in regions that contain more than one SNP per 100 bp.

In general, it is a difficult problem to choose the time discretization that can best achieve the goals of a particular data analysis, achieving high resolution during biologically interesting time periods without overfitting the available data. Sometimes it will be more fruitful to increase the sample size n or sequence length L than to refine the time discretization; an important avenue for future work will be tuning L, n, and d to improve inference in humans and other organisms.

Another avenue for future work will be to develop augmented HMMs for coalescent models with population structure. Structure and speciation have been incorporated into several versions of CoalHMM and diCal, and the strategy presented in this paper could be used to speed these up, though a more elaborate

network of hidden states will be required. We are hopeful that our new technique will help coalescent HMMs keep pace with the number and diversity of genomes being sequenced and tease apart the demographic patterns that differentiated them.

Acknowledgments. We are grateful to Matthias Steinrücken and other members of the Song group for helpful discussions. This research was supported in part by NSF Graduate Research Fellowships to K.H. and S.S., and by an NIH grant R01-GM094402 and a Packard Fellowship for Science and Engineering to Y.S.S.

References

1. Browning, B.L., Browning, S.R.: A fast, powerful method for detecting identity by descent. Am. J. Hum. Genet. 88, 173–182 (2011)
2. Burge, C., Karlin, S.: Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78–94 (1997)
3. Cahill, J.A., Green, R.E., Fulton, T.L., et al.: Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. PLoS Genetics 9, e1003345 (2013)
4. Dutheil, J.Y., Ganapathy, G., Hobolth, A., et al.: Ancestral population genomics: the coalescent hidden Markov model approach. Genetics 183, 259–274 (2009)
5. Ernst, J., Kellis, M.: ChromHMM: automating chromatin-state discovery and characterization. Nature Methods 9, 215–216 (2012)
6. Groenen, M.A., Archibald, A.L., Uenishi, H., et al.: Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491(7424), 393–398 (2012)
7. Gronau, I., Hubisz, M.J., Gulko, B., et al.: Bayesian inference of ancient human demographic history from individual genome sequences. Nature Genetics 43, 1031–1034 (2011)
8. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., Bustamante, C.D.: Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics 5, e1000695 (2009)
9. Haddrill, P.R., Thornton, K.R., Charlesworth, B., Andolfatto, P.: Multilocus patterns of nucleotide variability and the demographic selection history of *Drosophila melanogaster* populations. Genome Res. 15, 790–799 (2005)
10. Hailer, F., Kutschera, V.E., Hallstrom, B.M., et al.: Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. Science 336, 344–347 (2012)
11. Harris, K., Nielsen, R.: Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genetics 9, e1003521 (2013)
12. Hobolth, A., Christensen, O.F., Mailund, T., Schierup, M.H.: Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genetics 3, 294–304 (2007)
13. Hudson, R.R.: Properties of the neutral allele model with intergenic recombination. Theor. Popul. Biol. 23, 183–201 (1983)
14. Hudson, R.R.: Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18(2), 337–338 (2002)

15. Li, H., Durbin, R.: Inference of human population history from individual whole-genome sequences. Nature 10, 1–5 (2011)
16. Mailund, T., Dutheil, J.Y., Hobolth, A., et al.: Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. PLoS Genetics 7, e1001319 (2011)
17. Mailund, T., Halager, A.E., Westergaard, M., et al.: A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genetics 8(12), e1003125 (2012)
18. Meyer, M., Kircher, M., Gansauge, M.T., et al.: A high-coverage genome sequence from an archaic Denisovan individual. Science 338, 222–226 (2012)
19. Miller, W., Schuster, S.C., Welch, A.J.: Polar and brown bear genomes reveal ancient admixture and demographic footprints of plast climate change. Proc. Natl. Acad. Sci. USA 109, 2382–2390 (2012)
20. Orlando, L., Ginolhac, A., Zhang, G., et al.: Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499, 74–78 (2013)
21. Palamara, P.F., Lencz, T., Darvasi, A., Pe'er, I.: Length distributions of identity by descent reveal fine-scale demographic history. Am. J. Hum. Genet. 91, 809–822 (2012)
22. Paul, J.S., Steinrücken, M., Song, Y.S.: An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. Genetics 187, 1115–1128 (2011)
23. Pritchard, J.: Whole-genome sequencing data offer insights into human demography. Nature Genetics 43, 923–925 (2011)
24. Sheehan, S., Harris, K., Song, Y.S.: Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. Genetics 194, 647–662 (2013)
25. Steinrücken, M., Paul, J.S., Song, Y.S.: A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. Theor. Popul. Biol. 87, 51–61 (2013)
26. The 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010)
27. Thornton, K., Andolfatto, P.: Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172, 1607–1619 (2006)
28. Wan, Q.H., Pan, S.K., Hu, L., et al.: Genome analysis and signature discovery for diving and sensory properties of the endangered chinese alligator. Cell Res. 23(9), 1091–1105 (2013)
29. Wiuf, C., Hein, J.: Recombination as a point process along sequences. Theor. Popul. Biol. 55, 248–259 (1999)
30. Zhao, S., Zheng, P., Dong, S., et al.: Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nature Genetics 45, 67–71 (2013)