Bayesian Haplotype Inference via the Dirichlet Process

Eric Xing[†] EPXING@CS.BERKELEY.EDU
Roded Sharan[†] RODED@ICSI.BERKELEY.EDU
Michael I. Jordan[†♯] JORDAN@CS.BERKELEY.EDU
Computer Science Division[†] and Department of Statistics[♯], University of California, Berkeley, CA 94720-1776

Abstract

The problem of inferring haplotypes from genotypes of single nucleotide polymorphisms (SNPs) is essential for the understanding of genetic variation within and among populations, with important applications to the genetic analysis of disease propensities and other complex traits. The problem can be formulated as a mixture model, where the mixture components correspond to the pool of haplotypes in the population. The size of this pool is unknown; indeed, knowing the size of the pool would correspond to knowing something significant about the genome and its history. Thus methods for fitting the genotype mixture must crucially address the problem of estimating a mixture with an unknown number of mixture components. In this paper we present a Bayesian approach to this problem based on a nonparametric prior known as the Dirichlet process. The model also incorporates a likelihood that captures statistical errors in the haplotype/genotype relationship. We apply our approach to the analysis of both simulated and real genotype data, and compare to extant methods.

1. Introduction

The availability of a nearly complete human genome sequence makes it possible to begin to explore individual differences between DNA sequences on a genome-wide scale, and to search for associations of such genotypic variation with disease and other phenotypes (Risch, 2000). The largest class of individual differences in DNA are the *single nucleotide polymorphisms (SNPs)*. Millions of SNPs have been detected thus far out of an estimated total of ten million common SNPs (Sachidanandam et al., 2001).

A SNP commonly has two variants, or *alleles*, in the population, corresponding to two specific nucleotides chosen from $\{A, C, G, T\}$. A *haplotype* is a list of alleles at contiguous sites in a local region of a single chromosome. Assuming no recombination in this local region, a haplotype is inherited as a unit. Recall that for diploid organisms (such as humans) the chromosomes come in pairs. Thus two haplotypes go together to make up a *genotype*, which is the list of *unordered* pairs of alleles in a region. That is, a genotype is obtained from a pair of haplotypes by omitting the specification of the association of each allele with one of the two chromosomes—its *phase*. Common biological methods for assaying genotypes typically do not provide phase information; phase can be obtained at a considerably higher cost (Patil et al., 2001). It is desirable to develop automatic methods for inferring haplotypes from genotypes and possibly other data sources (e.g., pedigrees). With a set of inferred haplotypes in hand, associations to disease can be explored.

From the point of view of population genetics, the basic model underlying the haplotype inference problem is a finite mixture model. That is, letting $\mathcal{H}$ denote the set of all possible haplotypes associated with a given region (a set of cardinality 2^k in the case of binary polymorphisms, where k is the number of heterozygous SNPs), the probability of a genotype is given by:

$$p(g) = \sum_{h_1, h_2 \in \mathcal{H}} p(h_1, h_2) \mathbb{I}(h_1 \oplus h_2 = g) \qquad (1)$$

where $\mathbb{I}(h_1 \oplus h_2 = g)$ is the indicator function of the event that haplotypes h_1 and h_2 are consistent with g. Under the assumption of Hardy-Weinberg equilibrium (HWE), an assumption that is standard in the literature and will also be made here, the mixing proportion $p(h_1, h_2)$ is assumed to factor as $p(h_1)p(h_2)$.

Given this basic statistical structure, the simplest

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

methodology for haplotype inference is maximum likelihood via the EM algorithm, treating the haplotype identities as latent variables and estimating the parameters $p(h)$ (Excoffier & Slatkin, 1995). This methodology has rather severe computational requirements, in that a probability distribution must be maintained on the (large) set of possible haplotypes, but even more fundamentally it fails to capture the notion that small sets of haplotypes should be preferred. This notion derives from an underlying assumption that for relatively short regions of the chromosome there is limited diversity due to population bottlenecks and relatively low rates of recombination and mutation.

One approach to dealing with this issue is to formulate a notion of "parsimony," and to develop algorithms that directly attempt to maximize parsimony. Several important papers have taken this approach (Clark et al., 1998; Gusfield, 2002; Eskin et al., 2003) and have yielded new insights and algorithms. Another approach is to elaborate the probabilistic model, in particular by incorporating priors on the parameters. Different priors have been discussed by different authors, ranging from simple Dirichlet priors (Niu et al., 2002) to priors based on the coalescent process (Stephens et al., 2001) to priors that capture aspects of recombination (Greenspan & Geiger, 2003). These models provide implicit notions of parsimony, via the implicit "Ockham factor" of the Bayesian formalism.

We also take a Bayesian statistical approach in the current paper, but we attempt to provide more explicit control over the number of inferred haplotypes than has been provided by the statistical methods proposed thus far, and the resulting inference algorithm has commonalities with the parsimony-based schemes.

Our approach is based on a nonparametric prior known as the *Dirichlet process* (Ferguson, 1973). In the setting of finite mixture models, the Dirichlet process—not to be confused with the Dirichlet distribution—is able to capture uncertainty about the number of mixture components (Escobar & West, 2002). The basic setup can be explained in terms of an urn model, and a process that proceeds through data sequentially. Consider an urn which at the outset contains a ball of a single color. At each step we either draw a ball from the urn, and replace it with two balls of the same color, or we are given a ball of a new color which we place in the urn, with a parameter defining the probabilities of these two possibilities. The association of data points to colors defines a "clustering" of the data.

To make the link with Bayesian mixture models, we associate with each color a draw from the distribution defining the parameters of the mixture components. This process defines a *prior distribution* for a mixture model with a random number of components. Multiplying this prior by a likelihood yields a *posterior distribution.* Markov chain Monte Carlo algorithms have been developed to sample from the posterior distributions associated with Dirichlet process priors (Escobar & West, 2002; Neal, 2000).

The usefulness of this framework for the haplotype problem should be clear—using a Dirichlet process prior we in essence maintain a pool of haplotype candidates that grows as observed genotypes are processed. The growth is controlled via a parameter in the prior distribution that corresponds to the choice of a new color in the urn model, and via the likelihood, which assesses the match of the new genotype to the available haplotypes.

To expand on this latter point, an advantage of the probabilistic formalism is its ability to elaborate the observation model for the genotypes to include the possibility of errors. In particular, the indicator function $\mathbb{I}(h_1 \oplus h_2 = g)$ in Eq. (1) is suspect—there are many reasons why an individual genotype may not match with a current pool of haplotypes, such as the possibility of mutation or recombination in the meiosis for that individual, and errors in the genotyping or data recording process. Such sources of small differences should not lead to the inference procedure spawning new haplotypes.

In the current paper we present a statistical model for haplotype inference based on a Dirichlet process prior and a likelihood that includes error models for genotypes. We describe a Markov chain Monte Carlo procedure, in particular a procedure that makes use of both Gibbs and Metropolis-Hasting updates, for posterior inference. We present results of applying our method to the analysis of both simulated and real genotype data, comparing to the state-of-the-art PHASE algorithm (Stephens et al., 2001).

2. The Statistical Model

The input to a phasing algorithm can be represented as a *genotype matrix* G with columns corresponding to SNPs in their order along the chromosome and rows corresponding to genotyped individuals. $G_{i,j}$ represents the information on the two alleles of the i-th individual for SNP j. We denote the two alleles of a SNP by 0 and 1, and $G_{i,j}$ can take on one of four values: 0 or 1, indicating a homozygous site; 2, indicating a heterozygous site; and '?', indicating missing data.[1]

[1]Although we focus on binary data here, it is worth noting that our methods generalize immediately to non-

We will describe our model in terms of a pool of ancestral haplotypes, or *templates*, from which each population haplotype originates (Greenspan & Geiger, 2003). The haplotype itself may undergo point mutation with respect to its template. The size of the pool and its composition are both unknown, and are treated as random variables under a Dirichlet process prior. We begin by providing a brief description of the Dirichlet process and subsequently show how this process can be incorporated into a model for haplotype inference.

2.1. Dirichlet process mixtures

Rather than present the Dirichlet process in full generality, we focus on the specific setting of mixture models, and make use of an urn model to present the essential features of the process. For a fuller presentation, see, e.g., Ishwaran and James (2001). We assume that data x arise from a mixture distribution with mixture components $p(x|\phi)$. We assume the existence of a *base measure* $G(\phi)$, which is one of the two parameters of the Dirichlet process. (The other is the parameter τ, which we present below). The parameter $G(\phi)$ is not the prior for ϕ, but is used to generate a prior for ϕ, in the manner that we now discuss.

Consider the following process for generating samples $\{x_1, x_2, \ldots, x_n\}$ from a mixture model consisting of an unspecified number of mixture components, or *equivalence classes*:

- The first sample x_1 is sampled from a distribution $p(x|\phi_1)$, where the parameter ϕ_1 is sampled from the base measure $G(\phi)$.
- The ith sample, x_i, is sampled from the distribution $p(x|\phi_{c_i})$, where:

– The equivalence class of sample i, c_i, is drawn from the following distribution:

$$p(c_i = c_j \text{ for some } j < i | c_1, \ldots, c_{i-1}) = \frac{n_{c_j}}{i - 1 + \tau} \quad (2)$$

$$p(c_i \neq c_j \text{ for all } j < i | c_1, \ldots, c_{i-1}) = \frac{\tau}{i - 1 + \tau}, \quad (3)$$

where n_{c_i} is the *occupancy number* of class c_i—the number of previous samples belonging to class c_i.

– The parameter ϕ_{c_i} associated with the mixture component c_i is obtained as follows:

$$\phi_{c_i} = \phi_{c_j} \quad \text{if } c_i = c_j \text{ for some } j < i \text{ (i.e., } c_i \text{ is a populated equivalence class)}$$

$$\phi_{c_i} \sim G(\phi) \quad \text{if } c_i \neq c_j \text{ for all } j < i \text{ (i.e., } c_i \text{ is a new equivalence class)}$$

Eqs. (2) and (3) define a conditional prior for the equivalence class indicator c_i of each sample during a sequential sampling process. They imply a self-reinforcing property for the choice of equivalence class of each new sample—previously populated classes are more likely to be chosen.

It is important to emphasize that the process that we have discussed will be used as a *prior distribution*. We now embed this prior in a full model that includes a likelihood for the observed data. In Section 3 we develop Markov chain Monte Carlo inference procedures for this model.

2.2. The model

We present a probabilistic model for the generation of haplotypes in a population and for the generation of genotypes from these haplotypes. We assume that each individual's genotype is formed by drawing two random *templates* from an ancestral pool, and that these templates are subject to random perturbation. To model such perturbations we assume that each locus is mutated independently from its ancestral state with the same error rate. Finally, we assume that we are given noisy observations of the resulting genotypes. The model is displayed as a graphical model in Figure 1.

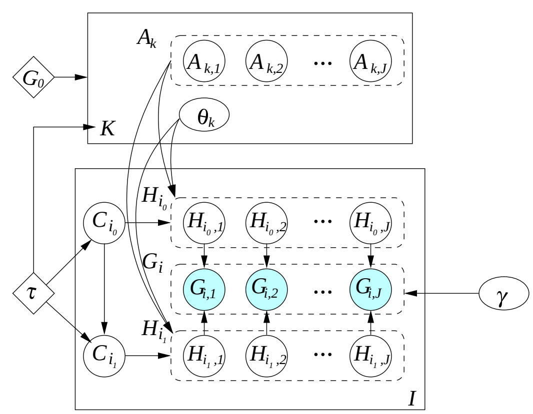

Figure 1. The graphical model representation of the haplotype model with a Dirichlet process prior. Circles represent the state variables, ovals represent the parameter variables, and diamonds represent fixed parameters. The dashed boxes denote sets of variables corresponding to the same ancestral template, haplotype, and genotype, respectively. The solid boxes correspond to i.i.d. replicates of sets of variables, each associated with a particular individual, or ancestral template, respectively.

Let J be an ordered list of loci of interest. For each individual i, we denote his/her paternal haplotype by $H_{i_0} := [H_{i_0,1}, \ldots, H_{i_0,J}]$ and maternal haplotype by $H_{i_1} := [H_{i_1,1}, \ldots, H_{i_1,J}]$. We denote a set of ancestral templates as $\mathbf{A} = \{A_1, A_2, \ldots\}$, where

binary data, and accommodate missing data.

$A_k := [A_{k,1}, \ldots, A_{k,J}]$ is a particular member of this set.

In our framework, the probability distribution of the haplotype variable H_{i_t}, where the sub-subscript $t \in \{0,1\}$ indexes paternal or maternal origin, is modeled by a mixture model with an unspecified number of mixture components, each corresponding to an equivalence class associated with a particular ancestor. For each individual i, we define the equivalence class variables C_{i_0} and C_{i_1} for the paternal and maternal haplotypes, respectively, to specify the ancestral origin of the corresponding haplotype. The C_{i_t} are the random variables corresponding to the equivalence classes of the Dirichlet process. The base measure G of the Dirichlet process is a joint measure on ancestral haplotypes A and mutation parameters θ, where the latter captures the probability that an allele at a locus is identical to the ancestor at this locus. We let $G(A,\theta) = p(A)p(\theta)$, and we assume that $p(A)$ is a uniform distribution over all possible haplotypes. We let $p(\theta)$ be a beta distribution, $\text{Beta}(\alpha_h, \beta_h)$, and we choose a small value for $\beta_h/(\alpha_h + \beta_h)$, corresponding to a prior expectation of a low mutation rate.

Given C_{i_t} and a set of ancestors, we define the conditional probability of the corresponding haplotype instance $h := [h_1, \ldots, h_J]$ to be:

$$\begin{aligned} & p(H_{i_t} = h | C_{i_t} = k, \mathbf{A} = \mathbf{a}, \boldsymbol{\theta}) \\ & \quad = \; p(H_{i_t} = h | A_k = a, \theta_k = \theta) \\ & \quad = \; \prod_j p(h_j | a_j, \theta), \end{aligned} \tag{4}$$

where $p(h_j|a_j,\theta)$ is the probability of having allele h_j at locus j given its ancestor. Eq. (4) assumes that each locus is mutated independently with the same error rate. For haplotypes, $H_{i_t,j}$ takes values from a set B of alleles. We use the following *single-locus mutation model*:

$$p(h_j|a_j,\theta) = \theta^{\mathbb{I}(h_j = a_j)} \Big(\frac{1-\theta}{|B|-1}\Big)^{\mathbb{I}(h_j \neq a_j)} \tag{5}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

The joint conditional distribution of haplotype instances $\mathbf{h} = \{h_{i_t} \;:\; t \in \{0,1\}, i \in \{1,2,\ldots,I\}\}$ and parameter instances $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\}$, given the ancestor indicator $\mathbf{c}$ of haplotype instances and the set of ancestors $\mathbf{a} = \{a_1, \ldots, a_K\}$, can be written explicitly as:

$$p(\mathbf{h}, \boldsymbol{\theta} | \mathbf{c}, \mathbf{a}) \propto \prod_k \theta_k^{m_k + \alpha_h - 1} \Big(\frac{1-\theta_k}{|B|-1}\Big)^{m'_k} \big[1-\theta_k\big]^{\beta_h - 1} \tag{6}$$

where $m_k = \sum_j \sum_i \sum_t \mathbb{I}(h_{i_t,j} = a_{k,j}) \mathbb{I}(c_{i_t} = k)$ is the number of alleles that were not mutated with respect to the ancestral allele, and $m'_k = \sum_j \sum_i \sum_t \mathbb{I}(h_{i_t,j} \neq a_{k,j}) \mathbb{I}(c_{i_t} = k)$ is the number of mutated alleles. The count $\mathbf{m}_k = \{m_k, m'_k\}$ is a sufficient statistic for the parameter θ_k and the count $\mathbf{m} = \{\mathbf{m}_k, \mathbf{m}'_k\}$ is a sufficient statistic for the parameter $\boldsymbol{\theta}$. The marginal conditional distribution of haplotype instances can be obtained by integrating out θ in Eq. (6):

$$p(\mathbf{h}|\mathbf{c},\mathbf{a}) = \prod_k R(\alpha_h, \beta_h) \frac{\Gamma(\alpha_h + m_k)\Gamma(\beta_h + m'_k)}{\Gamma(\alpha_h + \beta_h + m_k + m'_k)} \Big(\frac{1}{|B|-1}\Big)^{m'_k} \tag{7}$$

where $\Gamma(\cdot)$ is the gamma function, and $R(\alpha_h, \beta_h) = \frac{\Gamma(\alpha_h + \beta_h)}{\Gamma(\alpha_h)\Gamma(\beta_h)}$ is the normalization constant associated with $\text{Beta}(\alpha_h, \beta_h)$. (For simplicity, we use the abbreviation R_h for $R(\alpha_h, \beta_h)$ in the sequel).

We now introduce a *noisy observation model* for the genotypes. We let $G_i = [G_{i,1}, \ldots, G_{i,J}]$ denote the *joint genotype* of individual i at loci $[1, \ldots, J]$, where each $G_{i,j}$ denotes the genotype at locus j. We assume that the observed genotype at a locus is determined by the paternal and maternal alleles of this locus as follows:

$$\begin{aligned} & p(g_{i,j} | h_{i_0,j}, h_{i_1,j}, \gamma) \\ = \; & \gamma^{\mathbb{I}(h_{i,j} = g_{i,j})} [\mu_1(1-\gamma)]^{\mathbb{I}(h_{i,j} \overset{1}{\neq} g_{i,j})} [\mu_2(1-\gamma)]^{\mathbb{I}(h_{i,j} \overset{2}{\neq} g_{i,j})} \end{aligned}$$

where $h_{i,j} \triangleq h_{i_0,j} \oplus h_{i_1,j}$ denotes the unordered pair of two actual SNP allele instances at locus j; "$\overset{1}{\neq}$" denotes set difference by exactly one element (i.e., the observed genotype is heterozygous, while the true one is homozygous); "$\overset{2}{\neq}$" denotes set difference of both elements (i.e., the observed and true genotypes are different and both are homozygous); and μ_1 and μ_2 are appropriately defined normalizing constants. We place a beta prior $\text{Beta}(\alpha_g, \beta_g)$ on γ. Assuming independent and identical error models for each locus, the joint conditional probability of the entire genotype observation $\mathbf{g} = \{g_i \;:\; i \in \{1,2,\ldots,I\}\}$ and parameter γ, given all haplotype instances is:

$$\begin{aligned} p(\mathbf{g}, \gamma | \mathbf{h}) \; &= \; \prod_i p(g_i, \gamma | h_{i_0}, h_{i_1}) \\ &= \; \gamma^{\alpha_g + u - 1} \big[1-\gamma\big]^{\beta_g + u' + u'' - 1} \mu_1^{u'} \mu_2^{u''} \end{aligned} \tag{8}$$

where the sufficient statistics $\mathbf{u} = \{u, u', u''\}$ are computed as $u = \sum_{i,j} \mathbb{I}(h_{i,j} = g_{i,j})$, $u' = \sum_{i,j} \mathbb{I}(h_{i,j} \overset{1}{\neq} g_{i,j})$, and $u'' = \sum_{i,j} \mathbb{I}(h_{j,i} \overset{2}{\neq} g_{j,i})$, respectively. Note that $u + u' + u'' = IJ$. To reflect an assumption that the observational error rate is low we set $\beta_g/(\alpha_g + \beta_g)$ to a small constant (0.001). Again, the marginal conditional distribution of $\mathbf{g}$ is computed by integrating out γ.

Having described our Bayesian haplotype model, the problem of phasing individual haplotypes and estimating the size and configuration of the latent ancestral pool can be solved via posterior inference given the genotype data.

3. Markov chain Monte Carlo for Haplotype Inference

In this section, we describe a Gibbs sampling algorithm for exploring the posterior distribution under our model, including the latent ancestral pool. We also present a Metropolis-Hastings variant of this algorithm that appears to mix better in practice.

3.1. A Gibbs sampling algorithm

The Gibbs sampler draws samples of each random variable from a conditional distribution of the variable to be sampled given (previously sampled) values of all the remaining variables of the model. The variables needed in our algorithm are: c_{i_t}, the index of the ancestral template of a haplotype instance t of individual i; $a_{k,j}$, the allele pattern at the j-th locus of the k-th ancestral template; $h_{i_t,j}$, the t-th allele of the SNP at the j-th locus of individual i; and $g_{i,j}$, the genotype at locus j of individual i (the only observed variables in the model). All other variables in the model—θ and γ—are integrated out. The Gibbs sampler thus samples the values of c_{i_t}, $a_{k,j}$ and $h_{i_t,j}$.

Conceptually, the Gibbs sampler alternates between two coupled stages. First, given the current values of the hidden haplotypes, we sample the c_{i_t} and subsequently $a_{k,j}$, which are associated with the Dirichlet process prior. Second, given the current state of the ancestral pool and the ancestral template assignment for each individual, we sample the h_{j,i_t} variables in the basic haplotype model.

In the first stage, the conditional distribution of c_{i_t} is:

$$p(c_{i_t} = k \,|\mathbf{c}_{[-i_t]}, \mathbf{h}, \mathbf{a}) \propto p(c_{i_t} = k \,|\mathbf{c}_{[-i_t]}) p(h_{i_t}|a_k, \mathbf{c}, \mathbf{h}_{[-i_t]}) \\ = \begin{cases} \frac{n_{[-i_t],k}}{n-1+\tau} p(h_{i_t}|a_k, \mathbf{m}_{[-i_t],k}) & \text{if } k = c_{i'_{t'}} \text{ for some } i'_{t'} \neq i_t \\ \frac{\tau}{n-1+\tau} \sum_{a'} p(h_{i_t}|a')p(a') & \text{if } k \neq c_{i'_{t'}} \text{ for all } i'_{t'} \neq i_t \end{cases} \tag{9}$$

where $[-i_t]$ denotes the set of indices excluding i_t; $n_{[-i_t],k}$ represents the number of $c_{i'_{t'}}$ for $i'_{t'} \neq i_t$ that are equal to k; n represents the total number of instances sampled so far; and $\mathbf{m}_{[-i_t],k}$ denote the m sufficient statistics associated with all haplotype instances originating from ancestor k, except h_{i_t}. This expression is simply Bayes theorem with $p(h_{i_t}|a_k, \mathbf{c}, \mathbf{h}_{[-i_t],})$ playing the role of the likelihood and $p(c_{i_t} = k \,|\mathbf{c}_{[-i_t]})$ playing the role of the prior. The likelihood $p(h_{i_t}|a_k, \mathbf{m}_{[-i_t],k})$ is obtained by integrating over the parameter θ_k, as in Eq. (7).

The conditional probability for a newly proposed equivalence class k that is not populated by any previous samples requires a summation over all possible ancestors: $p(h_{i_t}) = \sum_{a'} p(h_{i_t}|a')p(a')$. Since the gamma function does not factorize over loci, computing this summation takes time that is exponential in the number of loci. To skirt this problem we endow each locus with its own mutation parameter $\theta_{k,j}$, with all parameters admitting the same prior $\mathrm{Beta}(\alpha_h, \beta_h)$. This gives rise to a closed-form formula for the summation and also for the normalization constant in Eq. (9). It is also, arguably, a more accurate reflection of reality.

Now we need to sample the ancestor template a_k, where k is the newly sampled ancestor index for c_{i_t}. When k is not equal to any other existing index $c_{i'_{t'}}$, a value for a_k needs to be chosen from $p(A|h_{i_t})$, the posterior distribution of A based on the prior $p(A)$ and the single dependent haplotype h_{i_t}. On the other hand, if k is an equivalence class populated by previous samples of $c_{i'_{t'}}$, we draw a new value of a_k from $p(A|h_{i_t},\ \text{s.t. } c_{i_t} = k)$. If after a new sample of c_{i_t}, a template is no longer associated with any haplotype instance, we remove this template from the pool. The conditional distribution for this Gibbs step is therefore:

$$p(a_{k,j}|h_{i_t,j} \text{ s.t. } c_{i_t} = k) \propto \\ \begin{cases} p(h_{i_t,j}|a_{k,j}) = \left(\frac{\alpha_h}{\alpha_h+\beta_h}\right)^{\mathbb{I}(h_{i_t,j}=a_{k,j})} \left(\frac{\beta_h}{(|B|-1)(\alpha_h+\beta_h)}\right)^{\mathbb{I}(h_{i_t,j}\neq a_{k,j})} \\ \qquad \text{if } k \text{ is not previously instantiated} \\ p(h_{i_t,j} \text{ s.t. } c_{i_t} = k|a_{k,j}) = \frac{\Gamma(\alpha_h+m_{k,j})\Gamma(\beta_h+m'_{k,j})}{\Gamma(\alpha_h+\beta_h+n_k)\cdot(|B|-1)^{m'_{k,j}}} \\ \qquad \text{if } k \text{ is previously instantiated,} \end{cases} \tag{10}$$

where $m_{k,j}$ (respectively, $m'_{k,j}$) is the number of allelic instances originated from ancestor k at locus j that are identical to (respectively, different from) the ancestor, when the ancestor has the pattern $a_{k,j}$.

We now proceed to the second sampling stage, in which we sample the haplotypes h_{i_t}. We sample each $h_{i_t,j}$, for all j, i, t, sequentially according to the following conditional distribution:

$$\begin{aligned} & p(h_{i_t,j}|\mathbf{h}_{[-(i,j)]}, h_{i_{\bar{t}},j}, \mathbf{c}, \mathbf{a}, \mathbf{g}) \\ & \quad \propto \quad p(g_i|h_{i_t,j}, h_{i_{\bar{t}},j}, \mathbf{u}_{[-(i,j)]}) p(h_{i_t,j}|a_{k,j}, \mathbf{m}_{[-(i_t,j)],k}) \\ & \quad = \quad R_g \frac{\Gamma(\alpha_g + u)\Gamma(\beta_g + (u' + u''))}{\Gamma(\alpha_g + \beta_g + IJ)} [\mu_1]^{u'} [\mu_2]^{u''} \times \\ & \qquad\quad R_h \frac{\Gamma(\alpha_h + m_{j,k})\Gamma(\beta_h + m'_{k,j})}{\Gamma(\alpha_h + \beta_h + n_k) \cdot (|B| - 1)^{m'_{k,j}}} \end{aligned} \tag{11}$$

where $[-(i_t, j)]$ denotes the set of indices excluding (i_t, j) and $m_{k,j} = m_{[-(i_t,j)],k} + \mathbb{I}(h_{i_t,j} = a_{k,j})$ (and similarly for the other sufficient statistics). Note that during each sampling step, we do not have to recompute the $\Gamma(\cdot)$, because the sufficient statistics are either not going to change (e.g., when the newly sampled $h_{i_t,j}$ is the same as the old sample), or only going to change by one (e.g., when the newly sampled $h_{i_1,j}$ results in a change of the allele). In such cases the new gamma function can be easily updated from the old one.

3.2. A Metropolis-Hasting sampling algorithm

Note that for a long list of loci, a uniform $p(A)$ of all possible ancestral template patterns will render the probability of sampling a new ancestor infinitesimal, due to the small value of the smoothed marginal likelihood of any haplotype pattern h_{i_t}, as computed from Eq. (9). This could result in slow mixing.

An alternative sampling strategy is to use a partial Gibbs sampling strategy with the following Metropolis-Hasting updates. For the proposal distribution for the equivalence class of h_{i_t} we use:

$$q(c^*_{i_t} = k | c_{[-i_t]}) = \begin{cases} \frac{n_{[-i_t],k}}{n-1+\tau} : \text{ if } & k = c_{i'_{t'}} \text{ for some } i'_{t'} \neq i_t \\ \frac{\tau}{n-1+\tau} : \text{ if } & k \neq c_{i'_{t'}} \text{ for all } i'_{t'} \neq i_t \end{cases} \tag{12}$$

Then we sample $a_{c^*_{i_t}}$ sequentially according to Eq. (10). For target distribution $p(c_{i_t} = k|\mathbf{c}_{[-i_t]}, \mathbf{h}, \mathbf{a})$, the proposal factor cancels when computing the acceptance probability ξ, leaving:

$$\xi(c^*_{i_t}, c_{i_t}) = \min\Big[1, \frac{p(h_{i_t}|a_{c^*_{i_t}}, \mathbf{h}_{[-i_t]})}{p(h_{i_t}|a_{c_{i_t}}, \mathbf{h}_{[-i_t]})}\Big]. \tag{13}$$

In practice, we found that the above modification to the Gibbs sampling algorithm leads to substantial improvement in efficiency for long haplotype lists, whereas for short lists, the Gibbs sampler remains better due to the high (100%) acceptance rate.

4. Experimental Results

We validated our algorithm by applying it to simulated and real data and compared its performance to that of the state-of-the-art PHASE algorithm (Stephens et al., 2001) and other current algorithms. We report on the results of both variants of our algorithm: The Gibbs sampler, denoted DP(Gibbs), and the Metropolis-Hasting sampler, denoted DP(MH). Throughout the experiments, we set the hyperparameter τ in the Dirichlet process to be roughly 1% of the population size, i.e., for a data set of 100 individuals, $\tau = 1$. We used a burn-in of 2000 iterations (or 4000 for datasets with more than 50 individuals), and used the next 6000 iterations for estimation.

4.1. Simulated data

In our first set of experiments we applied our method to simulated data ("short sequence data") from Stephens et al. (2001). This data contains sets of $2n$ haplotypes, randomly paired to form n genotypes, under an infinite-sites model with parameters $\eta = 4$ and $R = 4$ determining the mutation and recombination rates, respectively. We used the first 40 datasets for each combination of individuals and sites, where the number of individuals ranged between 10 and 50, and the number of sites ranged between 5 and 30.

To evaluate the performance of the algorithms we used the following error measures: err_s, the ratio of incorrectly phased SNP sites over all non-trivial heterozygous SNPs (excluding individuals with a single heterozygous SNP); err_i, the the ratio of incorrectly phased individuals over all non-trivial heterogeneous individuals; and d_s, the *switch distance*, which is the number of phase flips required to correct the predicted haplotypes over all non-trivial heterogeneous SNPs. The results are summarized in Table 1. Overall, we perform slightly worse than PHASE on the first two measures, and slightly better on the switch distance measure (which uses 100,000 sampling steps). Both algorithms provide a substantial improvement over EM.

4.2. Real data

We applied our algorithm to two real datasets and compared its performance to that of PHASE (Stephens et al., 2001) and other algorithms.

The first dataset contains the genotypes of 129 individuals over 103 polymorphic sites (Daly et al., 2001). In addition it contains the genotypes of the parents of each individual, which allows the inference of a large portion of the haplotypes as in Eskin et al. (2003). The results are summarized in Table 2. It is apparent that the Metropolis-Hasting sampling algorithm significantly outperforms the Gibbs sampler, and is to be preferred given the relatively limited number of sampling steps ($\sim$ 6000). The overall performance is comparable to that of PHASE and better than both HAP (Halperin & Eskin, 2002; Eskin et al., 2003) and HAPLOTYPER (Niu et al., 2002).

It is important to emphasize that our methods also provide a posteriori estimates of the ancestral pool of haplotype templates and their frequencies. We omit a listing of these haplotypes, but provide an illustrative summary of the evolution of these estimates during

	DP(MH)			PHASE			EM
#individuals	err_s	err_i	d_s	err_s	err_i	d_s	err_i
10	0.060	0.216	0.051	0.046	0.182	0.054	0.424
20	0.039	0.152	0.039	0.029	0.136	0.046	0.296
30	0.036	0.121	0.038	0.024	0.101	0.027	0.231
40	0.030	0.094	0.029	0.019	0.071	0.026	0.195
50	0.028	0.082	0.024	0.019	0.072	0.025	0.167

Table 1. Performance on data from Stephens et al. (2001). The results for the EM algorithm are adapted from Stephens et al. (2001).

		DP(Gibbs)			DP(MH)			PHASE			HAP	HAPLOTYPER
block id.	length	err_s	err_i	d_s	err_s	err_i	d_s	err_s	err_i	d_s	err_s	err_s
1	14	0.223	0.485	0.229	0	0	0	0.003	0.030	0.003	0.007	0.039
2	5	0	0	0	0.007	0.026	0.007	0.007	0.026	0.007	0.036	0.065
3	5	0	0	0	0	0	0	0	0	0	0	0.008
4	11	0.143	0.262	0.128	0	0	0	0	0	0	0.015	-
5	9	0.020	0.066	0.020	0.011	0.033	0.011	0.011	0.033	0.011	0.027	0.151
6	27	0.071	0.191	0.074	0.005	0.043	0.005	0	0	0	0.018	0.041
7	7	0.005	0.018	0.005	0.005	0.018	0.005	0.005	0.018	0.005	0.068	0.214
8	4	0	0	0	0	0	0	0	0	0	0	0.252
9	5	0.029	0.097	0.029	0.012	0.032	0.012	0.012	0.032	0.012	0.057	0.152
10	4	0.007	0.025	0.007	0.007	0.025	0.007	0.008	0.025	0.008	0.042	0.056
11	7	0.010	0.034	0.005	0.005	0.017	0.005	0.011	0.034	0.011	0.033	0.093
12	5	0.010	0.037	0.020	0	0	0	0	0	0	0	0.077

Table 2. Performance on the data of Daly et al. (2001), using the block structure provided by Halperin and Eskin (2002). The results of HAP and HAPLOTYPER are adapted from Halperin and Eskin (2002). Since the error rate in Halperin and Eskin (2002) uses the number of both heterozygous and missing sites as the denominator, whereas we used only the non-trivial heterozygous ones, we rescaled the error rates of the two latter methods to be comparable to ours.

sampling (Figure 2).

The second dataset contains genotype data from four populations, 90 individuals each, across several genomic regions (Gabriel et al., 2002). We focused on the Yoruban population (D), which contains 30 trios of genotypes (allowing us to infer most of the true haplotypes) and analyzed the genotypes of 28 individuals over four medium-sized regions (see below). The results are summarized in Table 3. All methods yield higher error rates on these data, compared to the analysis of the data of Daly et al. (2001), presumably due to the low sample size. In this setting, over all but one of the four regions, our algorithm outperformed PHASE for all three types of error measures. A preliminary analysis suggests that our performance gain may be due to the bias toward parsimony induced by the Dirichlet process prior. We found that the number of template haplotypes in our algorithm is typically small, whereas in PHASE, the haplotype pool can be very large (i.e., region 7b has 83 haplotypes, compared to 10 templates in our case and 28 individuals overall).

		DP(MH)			PHASE		
region	length	err_s	err_i	d_s	err_s	err_i	d_s
16a	13	0.185	0.480	0.141	0.174	0.440	0.130
1b	16	0.100	0.250	0.160	0.200	0.450	0.180
25a	14	0.135	0.353	0.115	0.212	0.588	0.212
7b	13	0.105	0.278	0.066	0.145	0.444	0.092

Table 3. Performance on the data of Gabriel et al. (2002).

5. Conclusions

We have proposed a Bayesian approach to the modeling of genotypes based on a Dirichlet process prior. We have shown that the Dirichlet process provides a natural representation of uncertainty regarding the size and composition of the pool of haplotypes underlying a population. Using Markov chain Monte Carlo algorithms, we have shown that this model leads to effective inference procedures for inferring the ancestral pool and for haplotype phasing based on a set of genotypes. The model accommodates growing data collections and noisy and/or incomplete observations. The approach also naturally imposes an implicit bias toward small ancestral pools during inference, reminiscent of parsimony methods, doing so in a well-founded statistical framework that permits errors.

Our focus here has been on adapting the technology of the Dirichlet process in the setting of the standard haplotype phasing problem. But an important underlying motivation for our work, and a general motivation for pursuing probabilistic approaches to genomic inference problems, is the potential value of our model as a building block for more expressive models. In particular, as in Greenspan and Geiger (2003) and Lauritzen and Sheehan (2002), the graphical model formalism naturally accommodates various extensions, such as segmentation of chromosomes into

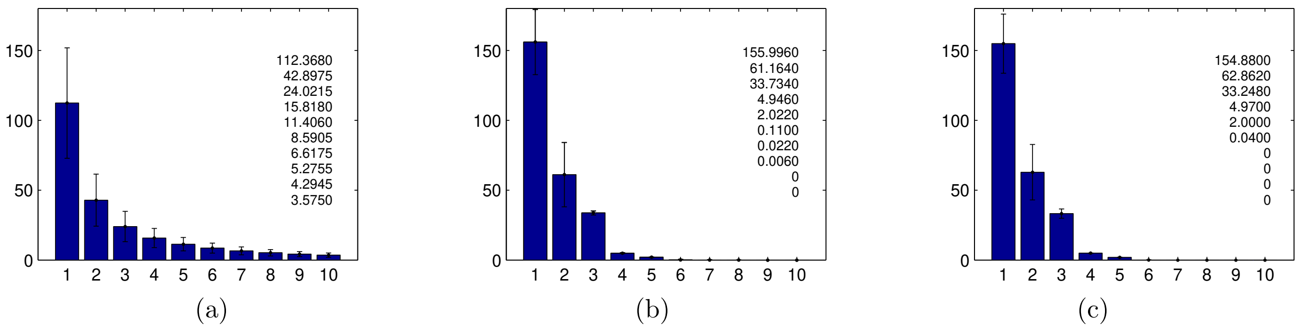

Figure 2. The top ten ancestral templates during Metropolis-Hasting sampling for block 1 of the data of Daly et al. (2001). (The numbers in the panels are the posterior means of the frequency of each template). (a) Immediately after burn-in (first 2000 samples). (b) 3000 samples after burn-in. (c) 6000 samples after burn-in.

haplotype blocks and the inclusion of pedigree relationships. The Dirichlet process parameterization also provides a natural upgrade path for the consideration of richer models; in particular, it is possible to incorporate more elaborate base measures G into the Dirichlet process framework—the coalescence-based distribution of Stephens et al. (2001) would be an interesting choice.

Acknowledgments

This research was supported in part by NSF ITR Grant CCR-0121555.

References

Clark, A., et al. (1998). Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *American Journal of Human Genetics*, *63*, 595–612.

Daly, M. J., et al. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics*, *29(2)*, 229–232.

Escobar, M. D., & West, M. (2002). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*, 577–588.

Eskin, E., Halperin, E., & Karp, R. (2003). Efficient reconstruction of haplotype structure via perfect phylogeny. *Journal of Bioinformatics and Computational Biology*, *1*, 1–20.

Excoffier, L., & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, *12*, 921–7.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*, 209–230.

Gabriel, S. B., et al. (2002). The structure of haplotype blocks in the human genome. *Science*, *296*, 2225–2229.

Greenspan, D., & Geiger, D. (2003). Model-based inference of haplotype block variation. *Proceedings of RECOMB 2003*.

Gusfield, D. (2002). Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. *Proceedings of RECOMB 2002* (pp. 166–175).

Halperin, E., & Eskin, E. (2002). Haplotype reconstruction from genotype data using imperfect phylogeny. Technical Report, Columbia University.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *90*, 161–173.

Lauritzen, S. L., & Sheehan, N. A. (2002). Graphical models for genetic analysis. TR R-02-2020, Aalborg University.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Computational and Graphical Statistics*, *9(2)*, 249–256.

Niu, T., Qin, S., Xu, X., & Liu, J. (2002). Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *American Journal of Human Genetics*, *70*, 157–169.

Patil, N., et al. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, *294*, 1719–1723.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, *405*, 847–56.

Sachidanandam, R., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, *291*, 1298–2302.

Stephens, M., Smith, N., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, *68*, 978–989.