Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics
Benidorm (Alicante, Spain), June 1st–6th, 2006

A Hidden Markov Dirichlet Process Model for Genetic Recombination in Open Ancestral Space

E. P. Xing and K-A. Sohn
School of Computer Science, Carnegie Mellon University
{epxing,ksohn}@cs.cmu.edu

Summary

We present a new statistical framework called hidden Markov Dirichlet process (HMDP) to jointly model the genetic recombinations among possibly infinite number of founders and the coalescence-with-mutation events in the resulting genealogies. The HMDP posits that a haplotype of genetic markers is generated by a sequence of recombination events that select an ancestor for each locus from an unbounded set of founders according to a 1st-order Markov transition process. Conjoining this process with a mutation model, our method accommodates both between-lineage recombination and within-lineage sequence variations, and leads to a compact and natural interpretation of the population structure and inheritance process. An efficient sampling algorithm based on a two-level nested Pólya urn scheme was also developed.

Keywords and Phrases: Dirichlet Process; HMM; MCMC; Statistical Genetics; Recombination; Population Structure; SNP.

1. INTRODUCTION

Recombinations between ancestral chromosomes during meiosis play a key role in shaping the patterns of linkage disequilibrium (LD)—the non-random association of alleles at different loci—in a population. Uneven occurrence of recombination events along chromosomal regions during genetic history can lead to "block structures" in molecular genetic polymorphisms such that within each block only low level of diversities are present in a population. The problem of inferring recombination hotspots is essential for understanding the origin and characteristics of genome variations; several combinatorial and statistical approaches have been developed for uncovering optimum block boundaries from single nucleotide polymorphism (SNP) haplotypes (Daly et al., 2001; Anderson and Novembre, 2003; Patil et al., 2001;

Eric P. Xing is an Assistant Professor of Computer Science at CMU, and Kyung-Ah Sohn is a CS Ph.D. student at CMU. This material is based upon work supported by the National Science Foundation under Grant No. 0523757, and and by the Pennsylvania Department of Health's Health Research Program under Grant No. 2001NF-Cancer Health Research Grant ME-01-739. E.P.X. is also supported by an NSF CAREER Award under Grant No. DBI-0546594.

Zhang et al., 2002). The deluge of SNP data also fuels the long-standing interest of analyzing patterns of genetic variations to reconstruct the evolutionary history and ancestral structures of human populations, using, for example, variants of admixture models on genetic polymorphisms (Rosenberg et al., 2002). These progress notwithstanding, the statistical methodologies developed so far mostly deal with LD analysis and ancestral inference separately, using specialized models that do not capture the close statistical and genetic relationships of these two problems. Moreover, most of these approaches ignore the inherent uncertainty in the genetic complexity (e,g., the number of genetic founders of a population) of the data and rely on inflexible models built on a pre-fixed, closed genetic space. Recently, Xing et al. (2004) have developed a nonparametric Bayesian framework for modeling genetic polymorphisms based on the Dirichlet process mixtures and extensions, which attempts to allow more flexible control over the number of genetic founders In this paper, we leverage this approach and present a unified framework to model complex genetic inheritance process that allows recombinations among possibly infinite founding alleles and coalescence-with-mutation events in the resulting genealogies.

We assume that individual chromosomes in a modern population are originated from an unknown number of ancestral haplotypes via biased random recombinations and mutations (Fig 1). The recombinations between the ancestors follow a a state-transition process we refer to as hidden Markov Dirichlet process (originated from the infinite HMM by Beal et al. (2001)), which travels in an open ancestor space. Our model draws inspiration from the HMM proposed in Greenspan and Geiger (2003), but we employ a two-level Pólya urn scheme akin to the hierarchical DP (Teh et al., 2006) to accommodate an open ancestor space, and allow full posterior inference of the recombination sites, mutation rates, haplotype origin, ancestor patterns, etc., conditioning on phased SNP data, rather than estimating them using information theoretic or maximum likelihood principles.

2. HIDDEN MARKOV DIRICHLET PROCESS FOR RECOMBINATION

Sequentially choosing recombination targets from a set of ancestral chromosomes can be modeled as a hidden Markov process (Niu et al., 2002; Greenspan and Geiger, 2003), in which the hidden states correspond to the index of the candidate chromosomes, the transition probabilities to the recombination rates between the recombining chromosome pairs, and the emission model to a mutation process that passes the chosen chromosome region in the ancestors to the descents. When the number of ancestral chromosomes is not known, it is natural to consider an HMM whose state space is countably infinite (Beal et al., 2001; Teh et al., 2006). In this section, we describe such an infinite HMM formalism, which we would like to call *hidden Markov Dirichlet process*, for modeling recombination in an open ancestral space.

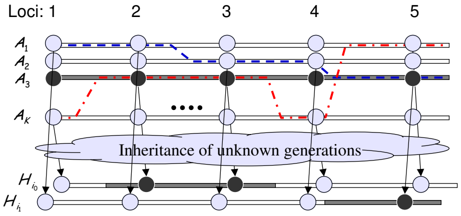

Figure 1: *An illustration of a hidden Markov Dirichlet process for haplotype recombination and inheritance .*

2.1. *Dirichlet Process mixtures*

Under a well-known genetic model known as *coalescence-with-mutation* (but without recombination), one can treat a haplotype from a modern individual—the joint allele configuration of a contiguous list of SNPs located on one of his/her chromosome (Fig 1)—as a descendent of an unknown ancestor haplotype (i.e., a founder) via random mutations. It can be shown that such a coalescent process in an infinite population leads to a partition of the population that can be succinctly captured by the following Pólya urn scheme. Consider an urn that 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. One can see that such a scheme leads to a partition of the balls according to their color. Letting parameter τ define the probabilities of the two types of draws, and viewing each (distinct) color as a sample from Q_0, and each ball as a sample from Q, Blackwell and MacQueen (1973) showed that this Pólya urn model yields samples whose distributions are those of the marginal probabilities under the *Dirichlet process.* One can associate mixture component with colors in the Pólya urn model, and thereby define a "clustering" of the data. The resulting model is known as a *DP mixture.* Note that a DP mixture requires no prior specification of the number of components. Back to haplotype modeling, following Xing et al. (2004), let $H_i = [H_{i,1}, \ldots, H_{i,T}]$ denote a haplotype over T SNPs from chromosome i; let $A_k = [A_{k,1}, \ldots, A_{k,T}]$ denote an ancestor haplotype (indexed by k) and θ_k denote the *mutation rate* of ancestor k; and let C_i denote an *inheritance variable* that specifies the ancestor of haplotype H_i. Then, under a DP mixture, we have the following Pólya urn scheme for sampling modern haplotypes:

- Draw first haplotype:

$a_1 \mid \mathrm{DP}(\tau, Q_0) \sim Q_0(\cdot)$,   sample the 1st founder;

$h_1 \sim P_h(\cdot|a_1, \theta_1)$,   sample the 1st haplotype from an inheritance model defined on the 1st founder;

- for subsequent haplotypes:

  – sample the founder indicator for the ith haplotype:

$$c_i|\mathrm{DP}(\tau, Q_0) \sim \begin{cases} p(c_i = c_j \text{ for some } j < i|c_1, ..., c_{i-1}) = \frac{n_{c_j}}{i-1+\tau} \\ p(c_i \neq c_j \text{ for all } j < i|c_1, ..., c_{i-1}) = \frac{\tau}{i-1+\tau} \end{cases}$$

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

  – sample the founder of haplotype i (indexed by c_i):

$$\phi_{c_i}|\mathrm{DP}(\tau, Q_0) \begin{cases} = \{a_{c_j}, \theta_{c_j}\} \text{ if } \ c_i = c_j \ \text{ for some } j < i \text{ (i.e., } c_i \text{ refers to an inherited founder)} \\ \sim Q_0(a, \theta) \ \text{ if } \ c_i \neq c_j \ \text{ for all } \ j < i \text{ (i.e., } c_i \text{ refers to a new founder)} \end{cases}$$

  – sample the haplotype according to its founder:

$h_i \mid c_i \sim P_h(\cdot|a_{c_i}, \theta_{c_i})$.

Notice that the above generative process assumes each modern haplotype to be originated from a single ancestor, this is only plausible for haplotypes spanning a short region on a chromosome. Now we consider long haplotypes possibly bearing multiple ancestors due to recombinations between an unknown number of founders.

2.2. *Hidden Markov Dirichlet Process (HMDP)*

In a standard HMM, state-transitions across a discrete time- or space-interval take place in a fixed-dimensional state space, thus it can be fully parameterized by, say, a K-dimensional initial-state probability vector and a $K \times K$ state-transition probability matrix. As first proposed in Beal et al. (2001), and later discussed in Teh et al. (2006), one can "open" the state space of an HMM by treating the now infinite number of discrete states of the HMM as the support of a DP, and the transition probabilities to these states from some source as the masses associated with these states. In particular, for each source state, the possible transitions to the target states need to be modeled by a unique DP. Since all possible source states and target states are taken from the same infinite state space, overall we need an open set of DPs with different mass distributions on the SAME support. In the sequel, we describe such a nonparametric Bayesian HMM using an intuitive hierarchical Pólya urn construction. We call this model a **hidden Markov Dirichlet process**.

We set up a single "stock" urn at the top level, which contains balls of colors that are represented by at least one ball in one or multiple urns at the bottom level. At the bottom level, we have a set of *distinct* urns (say, HMM-urns) which are used to define the initial and transition probabilities of the HMDP model. Specifically, one of HMM urns, u_0, is set aside to hold colored balls to be drawn at the onset of the HMM state-transition sequence. Each of the remaining HMM urns is used to hold balls to be drawn during the execution of a Markov chain of state-transitions. Now let's suppose that at time t the stock urn contains n balls of K distinct colors; the number of balls of color k in this urn is denoted by n_k. For urn u_0 and urns $u_1, \ldots, u_K$, let $m_{j,k}$ denote the number of balls of color k in urn u_j, and $m_j = \sum_{k \in \mathcal{C}} m_{j,k}$ denote the total number of balls in urn u_j. Suppose that at time $t-1$, we had drawn a ball with color k'. Then at time t, we either draw a ball randomly from urn $u_{k'}$, and place back two balls both of that color; or with probability $\frac{\tau}{m_j+\tau}$ we turn to the top level. From the stock urn, we can either draw a ball randomly and put back two balls of that color to the stock urn and one to $u_{k'}$, or obtain a ball of a new color $K+1$ with probability $\frac{\gamma}{n+\gamma}$ and put back a ball of this color to both the stock urn and urn $u_{k'}$ of the lower level. Essentially, we have a master DP (the stock urn) that serves as a base measure for infinite number of child DPs (HMM-urns).

2.3. *HMDP Model for Recombination and Inheritance*

For each modern chromosome i, let $C_i = [C_{i,1}, \ldots, C_{i,T}]$ denote the sequence of inheritance variables specifying the index of the ancestral chromosome at each SNP locus. When a recombination occurs, say, between loci t and $t+1$, we have $C_{i,t} \neq C_{i,t+1}$. We can introduce a Poisson point process to control the duration of non-recombinant inheritance. That is, given that $C_{i,t} = k$, then with probability $e^{-dr} + (1-e^{-dr})\pi_{kk}$, where d is the physical distance between two loci, r reflects the rate of recombination per unit distance, and π_{kk} is the self-transition probability of ancestor k defined by HMDP, we have $C_{i,t+1} = C_{i,t}$; otherwise, the source state (i.e., ancestor chromosome k) pairs with a target state (e.g., ancestor chromosome k') between loci t and $t+1$, with probability $(1-e^{-dr})\pi_{kk'}$. Hence, each haplotype H_i is a mosaic of segments of multiple ancestral chromosomes from the ancestral pool $\{A_{k,\cdot}\}_{k=1}^{\infty}$. Essentially, the model we described so far is a time-inhomogeneous infinite HMM. The emission process of the HMDP corresponds to an inheritance model from an ancestor to the matching descendent. For simplicity, we adopt the *single-locus mutation model* in Xing et al. (2004), which is widely used in statistical genetics as an approximation to a full coalescent genealogy (Liu et al., 2001).

The two-level nested Pólya urn schemes described in §2.2 motivates an efficient and easy-to-implement MCMC algorithm to sample from the posterior associated with HMDP. Details of this algorithms is available in Xing and Sohn (2007).

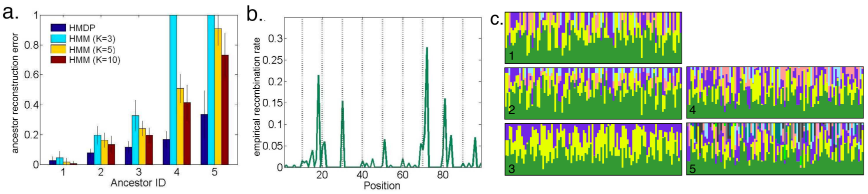

Figure 2: *(a)Ancestor reconstruction errors (the ratio of incorrectly recovered loci over all the loci). (b)The empirical recombination rates along 100 SNP loci with the pre-specified recombination hotspots (dotted lines). (c)The true (panel 1) and estimated (panel 2 for HMDP, and panel 3-5 for the HMMs with 3, 5, 10 states, repsectively.) population maps of ancestral compositions.*

3. EXPERIMENTS

We have applied the HMDP model to both simulated and real haplotype data. Our analyses focus on the three popular problems in statistical genetics of ancestral inference, LD-block analysis and population structural analysis.

3.1. *Analyzing simulated haplotype population*

We simulated a population of 200 individual haplotypes from $K_s = 5$ (unknown to the HMDP model) ancestor haplotypes, via a $K_s = 5$-dimensional HMM.

Ancestral Inference. Using HMDP, we successfully recovered the correct number (i.e., $K = 5$) of ancestors in 21 out of 30 simulated populations; for the remaining 9 populations, we inferred 6 ancestors. From samples of ancestor states $\{a_{k,t}\}$, we reconstructed the ancestral haplotypes under the HMDP model. For comparison, we also inferred the ancestors under the 3 standard HMMs using EM (Fig 2a).

LD-block Analysis. From samples of the inheritance variables $\{c_{i,t}\}$ under HMDP, we can infer the recombination status of each locus of each haplotype. We define the empirical recombination rates λ_e at each locus to be the ratio of individuals who had recombinations at that locus over the total number of haploids in the population. Fig 2b shows a plot of the λ_e in one simulated population. We can identify the recombination *hotspots* directly from such a plot based on an empirical threshold λ_t (i.e., $\lambda_t = 0.05$). The inferred hotspots (i.e., the λ_e peaks) show reasonable agreement with the true hotspots shown as vertical dotted lines.

Population Structural Analysis. Finally, from samples of the inheritance variables $\{c_{i,t}\}$, we can also uncover the genetic origins of all loci of each individual haplotype in a population. For each individual, we define an empirical *ancestor composition vector* η_e, which records the fractions of every ancestor in all the $c_{i,t}$'s of that individual. Fig 2c displays a *population map* constructed from the η_e's (the thin vertical lines) of all individuals. Five population maps, corresponding to (1) true ancestor compositions, (2) ancestor compositions inferred by HMDP, and (3-5) ancestor compositions inferred by HMMs with 3, 5, 10 states, respectively, are shown in Fig 2c. The L_1 distance between the HMDP-derived population map and the true map is 0.190, whereas the distance between HMM-map and true map is 0.319.

3.2. *Analyzing two real haplotype datasets*

We also applied HMDP to two real haplotype datasets, the single-population Daly data (Daly et al., 2001), and the two-population HapMap data (Thorisson et al., 2005); our method was able to uncover known recombination hotspots and population structures underlying these data (see Xing and Sohn (2007) for details).

4. CONCLUSION

We have proposed a new Bayesian approach for joint modeling genetic recombinations among possibly infinite founding alleles and coalescence-with-mutation events in the resulting genealogies. By incorporating a hierarchical DP prior to the stochastic matrix underlying an HMM, our method can efficiently infer a number of important genetic variables, such as recombination hotspot, mutation rates, haplotype origin, and ancestor patterns, jointly under a unified statistical framework. HMDP can also be easily adapted to more complicated genetics problems (e.g., analyzing unphased genotype data) and many engineering and information retrieval contexts such as object and theme tracking in open space. Due to space limit, we leave out some details of the algorithms and more results of our experiments, which are available in the full version of this paper (Xing and Sohn, 2007).

REFERENCES

Anderson, E. C. and Novembre, J. (2003). Finding haplotype block boundaries by using the minimum- description-length principle. *Am J Hum Genet* 73, 336

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2001). The infinite hidden Markov model. *In Advances in Neural Information Processing Systems* 13

Blackwell, D. and MacQueen, J. B. (1973). The infinite hidden Markov model *Ann. Statist.* **1**, 353–355

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics* 29(2), 229–232.

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

Liu, J. S., Sabatti, C., Teng, J., Keats, B., and Risch, N. (2001). Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.* 11, 1716–1724.

Niu, T., Qin, S., Xu, X., and Liu, J. (2002). Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *Am J Hum Genet* 70, 157–169.

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

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science* 298, 2381–2385.

Teh, Y., Jordan, M. I., Beal, M., and Blei, D. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* (to appear).

Thorisson, G.A., Smith, A.V., Krishnan, L., and Stein, L.D. (2005). The international hapmap project web site. *Genome Research* 15, 1591–1593.

Xing, E. P., Sharan, R., and Jordan, M. (2004). Bayesian haplotype inference via the Dirichlet process. *In Proceedings of the 21st International Conference on Machine Learning*, New York, 2004.10. ACM Press.

Xing, E. P. and Sohn, K.-A. (2007). Hidden markov dirichlet process: Modeling genetic inference in open ancestral space. *Bayesian Analysis* (to appear).

Zhang, K., Deng, M., Chen, T., Waterman, M., and Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci. USA* 99(11), 7335–7339.