

# Building mixture trees from binary sequence data

BY SHU-CHUAN CHEN

*Department of Statistics, The Pennsylvania State University, USA*

*National Health Research Institutes, Taiwan, ROC*

email: shu@stat.psu.edu

AND BRUCE G. LINDSAY

*The Pennsylvania State university, University Park, USA*

email: bgl@psu.edu

## SUMMARY

This article develops a new method for building a hierarchical tree from binary sequence data. It is based on an ancestral mixture model. The sieve parameter in the model plays the role of time in the evolutionary tree of the sequences. By sliding the sieve parameter, one can create a hierarchical tree that estimates the population structure at each fixed backward point in time. A case study of clustering the Mitochondrial DNA sequences of Griffiths and Tavaré (1994) is used to show that the approach performs well. In addition, theoretical and computational properties of the ancestral mixture model are further developed.

*Some key words:* Ancestral mixture model; Evolutionary tree; Hierarchical tree; Sieve parameter.

## 1. INTRODUCTION

One very natural method for clustering data is to build a mixture model for the data and then use the components of the fitted model to assign the data points to clusters. McLachlan and Basford (1988) provide an extensive treatment of this approach for the multivariate normal mixture model. Our interest here is to develop and extend the mixture model methodology with particular focus on binary sequence data because of its importance in biology.

Recently, single nucleotide polymorphisms (SNPs) have been gaining increasing attention. Here single nucleotide polymorphisms are single base pair positions in genomic DNA at which different sequence alternatives exist in normal individuals (Ott, 1999). Typically only two alternatives exist at any one position (either  $A$  and  $G$  at purine site or  $T$  and  $C$  at pyrimidine site) so that the data can

be thought of as binary. A map of 1.42 million single nucleotide polymorphisms (SNPs) over the human genome is described by the international SNP map working group (Sachidanandam, 2001). It is believed a map of these high density SNPs could be a public resource for defining nucleotide variation within the human genome that could help in identifying disease genes.

In this article, we show how to use mixture methodology to construct a hierarchical tree for binary sequence data. We first introduce an ancestral mixture model. This model is a discrete parallel to the multivariate normal mixture model that is used to cluster continuous multivariate data (McLachlan and Basford, 1988). We introduce a new methodology to build a hierarchical tree of clusters by mixture analysis, giving relationships between sequences that can be visually identified. After the clusters are identified, any individual can be assigned to a cluster in a probabilistic way. A case study indicates our approach performs quite well.

In previous literature, the ancestral mixture model was called a Bernoulli mixture. We have chosen a new name because of the close relationship of the model to a phylogenetic process. Govaert (1990) and Celeux and Govaert (1991) first used Bernoulli mixtures in clustering analysis. The same model was then used by Govaert and Nadif (1996) to compare different approaches to clustering analysis.

In this article, we propose a different rationale for using ancestral mixtures for cluster analysis. The structure and properties discovered in the article allow us to create a hierarchical tree. In addition, we perform a case study which shows our computer methodology works quite well.

The layout of this article is as follows. In section 2, we will describe the model we propose. In section 3, identifiability in the ancestral mixture model will be discussed. Point estimation by the maximum likelihood method will be described in section 4. Then, the induced tree structure will be discussed in section 5. Finally, some computing results on a case study will be shown in section 6.

## 2. THE MODEL AND ITS STRUCTURE

Suppose  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  are binary sequences of length  $L$  with  $(0, 1)$  coding. The goal of this article is to cluster these binary sequences in a probabilistic way. Some notations used in this article will be introduced as follows. The sampled binary sequences are denoted as  $\mathbf{X}$ , ancestor sequences are denoted by  $\mu$ , and the mutation rate is denoted by  $p$ . We will call each location in the sequence a *site* (or a *locus*), and each cluster a *component*, as it arises as a mixture component. The unicomponent model will be introduced first.

### 2.1 Unicomponent model

In the unicomponent model, we assume there is one component for the data. The simplest model, one ancestor with fixed mutation rate and one observation, will be introduced first.

We first build the model for a single binary sequence  $\mathbf{X}$  of length  $L$ . The sample space for  $\mathbf{X}$  can be represented as  $\{0, 1\}^L$ . In the single ancestor model, we assume there is a single ancestral sequence  $\mu_1 = (\mu_{11}, \mu_{12}, \dots, \mu_{1L})$  which is an unknown parameter. Here  $\mu_1$  itself is a binary sequence from  $\{0, 1\}^L$ . The observations  $(X_1, X_2, \dots, X_L)$  are then modelled as independent Bernoulli trials with

$$P(X_s \neq \mu_{1s}) = p \text{ and } P(X_s = \mu_{1s}) = \bar{p}$$

for  $s = 1, 2, \dots, L$ . If  $X_s \neq \mu_{1s}$ , we say a *mutation* occurred at site  $s$ , otherwise not. Here  $\bar{p} = 1 - p$ ;  $p$  will be called the *mutation probability*. It is constrained to be less than 0.5 as is needed for the identifiability of parameter  $\mu$ , as we will show in section 3.

Another way to develop this model is to create a binary sequence  $\varepsilon$ , where 0's represent no change and 1's represent change, where the  $\varepsilon_s$  are independent with

$$P(\varepsilon_s = 1) = p .$$

Then we can create the model by adding  $\varepsilon$  to  $\mu_1$ , using mod 2 arithmetic, *i.e.*

$$\mathbf{X} = \mu_1 + \varepsilon \pmod{2}.$$

That is, we can think of the  $\mathbf{X}'$ s as ‘‘ancestral type plus error’’ in modulo arithmetic. In this form we can see the close relationship to the additive normal model  $\mathbf{X} = \mu_1 + \varepsilon$ , where  $\varepsilon \sim N(0, \Sigma)$  implies  $\mathbf{X} \sim N(\mu_1, \Sigma)$ .

This model can also be thought of as arising from a process by which the ancestral sequence  $\mu$  evolves in one generation to  $\mathbf{X}$ , where each site has a probability  $p$  of mutation from the ancestral sequence type  $\mu_{1i}$  to the other type  $(1 - \mu_{1i})$ . That is to say, we move along the ancestral sequence and make independent decisions at each site about mutating the ancestor sequence; because  $p < 0.5$ , the probability of mutating the sequence at a site is less than leaving it as the same. As we will show later, the model is closed under multiple generations of this kind of mutation. Next, we will introduce the mutation kernel.

First, we can write the single ancestor model in a suggestive manner by using the following relationship. For variables  $a$  and  $b$  that are either 0 or 1,

$$(a - b)^2 = \begin{cases} 0 & , \text{ if } a \text{ and } b \text{ agree.} \\ 1 & , \text{ if } a \text{ and } b \text{ disagree.} \end{cases} \quad (1)$$

Since the observations  $\mathbf{X} = (X_1, \dots, X_L)$  are independent Bernoulli trials, using the device of equation (1), we can write the density for  $\mathbf{X}$  as:

$$\begin{aligned} \kappa(\mathbf{x}|\mu_1, p) &= \prod_{j=1}^L p^{(x_j - \mu_{1j})^2} \bar{p}^{1 - (x_j - \mu_{1j})^2} \\ &= p^{\mathbf{D}(\mathbf{x}, \mu_1)} \bar{p}^{L - \mathbf{D}(\mathbf{x}, \mu_1)} \\ &= \bar{p}^L \left( \frac{p}{\bar{p}} \right)^{\mathbf{D}(\mathbf{x}, \mu_1)} \end{aligned} \quad (2)$$

where

$$D(\mathbf{x}, \mu_1) = \sum_{j=1}^L (x_j - \mu_{1j})^2$$

is the number of *disagreements* between the elements of  $\mathbf{x}$  and the element of  $\mu_1$ . We will call  $\kappa(\mathbf{x}|\mu_1, p)$  the *mutation kernel*. When we use the mutation kernel with  $\mu_1$  as an unknown parameter,

we will call it the *single ancestor model*, with  $\mu_1$  as the ancestral sequence. Next, we will discuss the relationship to Normal density.

A simple mathematical relationship of the mutation kernel to the normal density can be derived as follows. Let  $\theta = \frac{p}{\bar{p}}$  be the odds of mutating, and let  $\psi = \log \frac{\bar{p}}{p}$ , so that  $\psi = -\log(\theta)$ . Since  $p < \frac{1}{2}$ , the range of  $\psi$  is  $(0, \infty)$ . Using this transformation, equation (2), the single ancestor model, can be rewritten as

$$\kappa(\mathbf{x}|\mu_1, \psi) = e^{-\psi \sum_{j=1}^L (x_j - \mu_{1j})^2} \cdot (1 + e^{-\psi})^{-L}. \quad (3)$$

Viewed from the structural perspective provided by equation (3), the single ancestor model looks like the independent normal density with means  $\mu_{1j}$  and variance  $\frac{1}{2\psi}$ . That is,  $\psi$  plays the role of  $\frac{1}{2 \times \text{variance}}$  in the normal density, and  $\mu_{1j}$  is either 0 or 1.

In addition to providing some intuition about the model, this relationship can be used to develop results about the modality of the mixture model.

## 2.2 Mixtures of mutation kernels

The single ancestor model can be extended to our mixture model as follows. We start building a mixture model with a random variable  $\vartheta$  that takes on values in the  $\mu$ -parameter space under a distribution  $\mathbf{Q}$ , where  $\mathbf{Q}$  will often be a discrete distribution with  $K$  points of support  $\mu_1, \mu_2, \dots, \mu_K$  and corresponding probabilities

$$P(\vartheta = \mu_{\mathbf{k}}) = \pi_{\mathbf{k}}.$$

Here  $\pi_{\mathbf{k}} \geq 0$  and  $\sum_{\mathbf{k}=1}^K \pi_{\mathbf{k}} = 1$ .

We then suppose the random variable  $\mathbf{X}$  is generated hierarchically by first generating  $\vartheta = \mu_{\mathbf{k}}$  from  $\mathbf{Q}$ , then generating  $\mathbf{X} = \mathbf{x}$  from  $\kappa(\mathbf{x}|\mu_{\mathbf{k}}, p)$ . We assume that  $\vartheta$  is unobserved, and so it will be called a *latent variable*.

If  $\mathbf{X}$  is generated in this fashion, then it will be said to have an *ancestral mixture model*, symbolically  $\mathbf{X} \sim \mathbf{A}(\mathbf{Q}, p)$ . The notation  $\mathbf{A}(\mu, p)$  will mean that  $\mathbf{Q}$  is degenerate at  $\mu$ . We can write the density of  $\mathbf{X}$  as

$$\mathbf{f}(\mathbf{x}; \mathbf{Q}, p) = \int \kappa(\mathbf{x}|\mu, p) d\mathbf{Q}(\mu).$$

This density function will be called a “ $\mathbf{Q}$ -mixture of mutation kernels”. When  $\mathbf{Q}$  is discrete, we can write this as

$$\begin{aligned} \mathbf{f}(\mathbf{x}; \mathbf{Q}, p) &= \sum_{\mathbf{k}=1}^K \pi_{\mathbf{k}} \kappa(\mathbf{x}|\mu_{\mathbf{k}}, p) \\ &= \sum_{\mathbf{k}=1}^K \pi_{\mathbf{k}} p^{\mathbf{D}(\mathbf{x}, \mu_{\mathbf{k}})} \bar{p}^{L - \mathbf{D}(\mathbf{x}, \mu_{\mathbf{k}})}. \end{aligned}$$

If we consider  $K$  fixed, we will call this the *K-component ancestral mixture model*. If we allow the distribution  $\mathbf{Q}$  to be arbitrary, we will call the model the *nonparametric ancestral model*. This corresponds to allowing an arbitrary number of components.

In the ancestral mixture model, we can create an observation  $\mathbf{X}$  by drawing  $\vartheta$  from  $\mathbf{Q}$ , then adding an error  $\varepsilon$ , using mod 2 arithmetic. Therefore we can write the model as  $\mathbf{X} = (\vartheta + \varepsilon) \pmod{2}$ .

The ancestral mixture model is a conditional independence model; that is, given  $\vartheta = \mu$ , the variables  $X_{is}$ ,  $s=1, \dots, L$  are independent. It is the variation in the latent variable  $\vartheta$  that “explains” the correlations between the observations. Next, we will discuss the nested structure of the ancestral mixture model.

A closure property of the normal mixture model, described in Lindsay (1995), is very helpful to learn the nested structure of the ancestral mixture model. It is described as follows.

**Proposition 1.** *Let  $N(\mathbf{Q}, \sigma^2)$  be the normal mixture model with mixing distribution  $\mathbf{Q}$ . Any mixture  $N(\mathbf{Q}, \sigma^2)$  can also be represented as a normal mixture by  $N(\mathbf{Q}^*, \sigma^2 - \sigma_1^2)$ , where  $\sigma_1^2 < \sigma^2$  and  $\mathbf{Q}^*$  is the convolution of  $\mathbf{Q}$  and  $N(0, \sigma_1^2)$  (Lindsay, 1995).*

As a consequence, the class of normal mixtures becomes richer and richer as the parameter  $\sigma^2 \downarrow 0$ . We can think of  $\sigma^2$  as a “sieve parameter”, in the sense that it can be used in a method of sieves approach to inference (Lindsay, 1995). A key here is that this sieve parameter cannot be estimated by maximum likelihood or other standard methods, as  $\sigma^2 = 0$  always provides the best fit.

The nested structure of the ancestral mixture model can be developed as follows. Suppose an ancestral sequence  $\mu$  goes through one generation of mutation at rate  $p_1 = 0.5 - \gamma_1$ , then goes through a second generation of mutation at rate  $p_2 = 0.5 - \gamma_2$ . We can calculate that the new distribution is also a single ancestor distribution with ancestor  $\mu$  and mutation rate

$$\begin{aligned} p &= p_1 \bar{p}_2 + \bar{p}_1 p_2 \\ &= 0.5 - 2\gamma_1 \gamma_2. \end{aligned} \tag{4}$$

We can write this symbolically as  $\mathbf{X} = (\mu + \varepsilon_1 + \varepsilon_2)$ , using mod 2 arithmetic, where  $\varepsilon_1 + \varepsilon_2$  is the two-generation mutation error.

This structure suggests a natural reparameterization of the model. Define  $\eta(p) = -\log(1 - 2p)$ . This provides a one-to-one increasing transformation of  $p \in [0, \frac{1}{2})$  into  $\eta \in [0, \infty)$ . We will write  $A(\mathbf{Q}, \eta)$  for the ancestral mixture model when using this parameter. Result (4) can be used to prove that if  $\varepsilon_1 \sim A(0, \eta_1)$  and  $\varepsilon_2 \sim A(0, \eta_2)$  are independent, then  $(\varepsilon_1 + \varepsilon_2) \pmod{2} \sim A(0, \eta_1 + \eta_2)$ . That is, the parameter  $\eta$  is additive in the same fashion as  $\sigma^2$  in the normal model.

Moreover, if a sequence  $\mu$  were to undergo  $T$  generations of mutation with constant rate  $p_0$  and  $\eta_0 = \log(1 - 2p_0)$ , then

$$\mathbf{X} = \mu + \varepsilon_1 + \dots + \varepsilon_T \pmod{2}$$

has distribution  $A(\mu, \eta_1)$  with  $\eta_1 = T \cdot \eta_0$ . For this reason we will call  $\eta$  the *time parameter*.

Finally, this structure gives us a result parallel to Proposition 1 for normal mixtures.

**Lemma 1.** *If  $\eta_1 < \eta_2$ , then  $\mathbf{X} \sim A(\mathbf{Q}, \eta_2)$  implies  $\mathbf{X} \sim A(\mathbf{Q}^*, \eta_1)$ , where  $\mathbf{Q}^*$  is the convolution (mod 2) of  $\mathbf{Q}$  and  $A(0, \eta_2 - \eta_1)$ .*

*Proof* : First, if  $\varepsilon_1 \sim A(0, \eta_1)$  and  $\varepsilon_2 \sim A(0, \eta_2 - \eta_1)$  are independent, then

$$(\varepsilon_1 + \varepsilon_2) \pmod 2 \sim A(0, \eta_2) .$$

Hence, if  $\vartheta \sim \mathbf{Q}$ , then we can write  $\mathbf{X} \sim A(\mathbf{Q}, \eta_2)$  as

$$\mathbf{X} = \vartheta + (\varepsilon_1 + \varepsilon_2) \pmod 2 ,$$

but we can also write it as

$$\mathbf{X} = (\vartheta^*) + \varepsilon_1 \pmod 2$$

where  $\vartheta^* = \vartheta + \varepsilon_2$  has the  $\mathbf{Q}^*$  distribution. □

Following from above, if we can write a density  $\mathbf{g}$  as a  $(\mathbf{Q}, \eta)$  ancestral mixture, then we can also write  $\mathbf{g}$  as a  $(\mathbf{Q}^*, \eta_1)$  mixture, for any  $\eta_1 < \eta$ , where  $\mathbf{Q}^*$  corresponds to the distribution of  $\mu + \varepsilon_1$ . This means that if we let  $\mathcal{M}_\eta$  be the set of all mixture densities for a fixed  $\eta$ , then  $\mathcal{M}_\eta \subset \mathcal{M}_{\eta_1}$ , so that the models are *nested* as  $\eta$  varies, become richer as  $p$  (or  $\eta$ ) goes to 0, with  $\mathcal{M}_0$  containing all possible densities on  $\{0, 1\}^L$ .

As a consequence, we will be able to use  $\eta$  as a sieve parameter, where as  $\eta \downarrow 0$ , the class of mixture models becomes richer and richer. For the model selection, we will consider the problem of selecting a suitable value of  $\eta$  from this sieve of models. Next, the identifiability in the ancestral mixture model will be discussed.

### 3. IDENTIFIABILITY IN THE MODEL

In this section, we will verify the identifiability of the parameters in the nonparametric ancestral mixture model.

A parameter  $\theta$  for a family of distributions  $f_\theta$  is said to be identifiable if distinct values of  $\theta$  correspond to distinct distributions; that is,  $\theta$  is identifiable if  $\theta_1 \neq \theta_2$  implies  $f_{\theta_1} \neq f_{\theta_2}$ . Note that identifiability of the parameters in a model can depend on the choice of parameter space. Here, we will seek the largest parameter space in which the ancestral mixture model is identifiable. First identifiability in the unicomponent model will be discussed.

#### *3.3 Identifiability in the unicomponent model*

We first note that the parameter  $\mu$  in the unicomponent model is not identifiable when  $p = 1/2$ . This is clear because the density is a constant when  $p = 1/2$ : for any  $\mu$ ,

$$f(\mathbf{x}; \mu, 1/2) = (1/2)^L .$$

Hence  $f(\mathbf{x}; \mu_1, 1/2) = f(\mathbf{x}; \mu_2, 1/2)$  for every  $\mu_1$  and  $\mu_2$ . Hereafter we exclude  $p = 1/2$  from the parameter space.

Secondly, we have that  $f(\mathbf{x}; \mu, p) = f(\mathbf{x}; \bar{\mu}, \bar{p})$ , where  $\bar{\mu} = (1 - \mu_1, 1 - \mu_2, \dots, 1 - \mu_L)$  is the complement of  $\mu$ . Thus identifiability will fail unless we make some further constraint on the parameters. To do so, we restrict  $p$  to  $[0, \frac{1}{2})$ . With this done, the parameters are identifiable.

**Proposition 2.** *The parameters  $(\mu, p)$  in the unicomponent model are identifiable in  $\Omega_1 = \{0, 1\}^L \times [0, 1/2)$ .*

*Proof:* We show how to solve for the parameters from the density. Suppose

$$P(X_{ij} = 1) = q .$$

If  $q < \frac{1}{2}$ , set  $\mu_j = 0$  and if  $q > \frac{1}{2}$ , then  $\mu_j = 1$ . Clearly  $p = \min\{q, 1 - q\}$ . □

Next, the identifiability in the ancestral mixture model will be discussed.

### 3.4 Identifiability in the nonparametric ancestral mixture model

In this section, some basic results about noncentral moments of the joint distribution of  $n$  binary variables from Settini and Smith (2000) will be reviewed. Starting from these, we can prove identifiability of the distribution  $\mathbf{Q}$  in the nonparametric ancestral mixture model when  $0 \leq p < 0.5$  is a fixed parameter but  $\mathbf{Q} \in \mathcal{P}$ , the class of all possible mixing distributions on the parameter space  $\mu \in \{0, 1\}^L$ .

Suppose  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$  where  $Y_1, Y_2, \dots, Y_n$  are binary variables taking values in  $\{-1, 1\}$ . Given a vector of nonnegative integers,  $\mathbf{a} = (a_1, a_2, \dots, a_n)$ , we define the notation  $\mathbf{Y}^{\mathbf{a}} = \prod_{i=1}^n Y_i^{a_i}$ . Since  $Y_i$  is 1 or  $-1$ ,  $Y_i^2 = 1$ , and therefore,

$$\begin{aligned} \mathbf{Y}^{2\mathbf{a}} &= \prod_{i=1}^n Y_i^{2a_i} \\ &= \prod_{i=1}^n (Y_i^2)^{a_i} \\ &= 1. \end{aligned}$$

Let the elements of the vector  $\mathbf{b}(\mathbf{a})$  be the elements of the vector  $\mathbf{a}$  reduced to 0 or 1 using  $\pmod{2}$  arithmetic. It follows that

$$\mathbf{Y}^{\mathbf{a}} = \mathbf{Y}^{\mathbf{b}(\mathbf{a})} .$$

As a result the noncentral moments,  $m_{\mathbf{Y}}(\mathbf{a}) = E(\mathbf{Y}^{\mathbf{a}})$ , have the property

$$m_{\mathbf{Y}}(\mathbf{a}) = m_{\mathbf{Y}}(\mathbf{b}(\mathbf{a})).$$

Thus the moments corresponding to strictly  $\{0, 1\}$  sequences  $\mathbf{a}$  determine all the moments. We now apply the results above to prove that when  $\Omega = \{0, 1\}^L \times [0, 1/2)$ , the parameters in the ancestral mixture model are identifiable. We will use the fact that a finite discrete distribution is completely determined by its moments (Settini and Smith, 2000).

**Proposition 3.** *The parameters  $(\mathbf{Q}, p)$  in the ancestral mixture model are identifiable in  $\Omega = \mathcal{P} \times \{p\}$ , where  $p$  is any fixed value in  $[0, \frac{1}{2})$ . (If  $p = \frac{1}{2}$ , the model is not identifiable.)*

*Proof* : We can change our sample and parameter space from  $\{0, 1\}^L$  to  $\{-1, 1\}^L$  by setting all 0 values to -1. We can then represent the ancestral mixture model symbolically as  $Y_i = \theta_i \varepsilon_i$ , where  $\theta$  is from the  $Q$  distribution transformed to  $\{-1, 1\}^L$ , and  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_L$  are *i.i.d.* with  $P(\varepsilon_i = -1) = p$  and  $P(\varepsilon_i = 1) = 1 - p$ . Further,  $\theta$  and  $\varepsilon$  are independent. This follows because

$$Y_i = \begin{cases} \theta_i & \text{if } \varepsilon_i = 1 \text{ (no mutation)} \\ -\theta_i & \text{if } \varepsilon_i = -1 \text{ (mutation)}. \end{cases}$$

Therefore, the noncentral moment of  $\mathbf{Y}$  can be expressed as

$$\begin{aligned} m_{\mathbf{Y}}(\mathbf{a}) &= E(X^{\mathbf{a}}) \\ &= E\left(\prod_{i=1}^L (\theta_i \varepsilon_i)^{a_i}\right) \\ &= E_{\mathbf{Q}}\left(\prod_{i=1}^L \theta_i^{a_i}\right) E\left(\prod_{i=1}^L \varepsilon_i^{a_i}\right) \\ &= m_{\mathbf{Q}}(\mathbf{a}) \prod_{i=1}^L E(\varepsilon_i^{b_i(a_i)}), \end{aligned}$$

where  $m_{\mathbf{Q}}(\mathbf{a})$  is the  $\mathbf{a}$ th moment of  $\theta$  and under  $\mathbf{Q}$

$$E(\varepsilon_i^{b_i(a_i)}) = \begin{cases} 1 & \text{if } b_i(a_i) = 0 \\ 1 - 2p & \text{if } b_i(a_i) = 1. \end{cases}$$

Therefore,

$$\begin{aligned} m_{\mathbf{Y}}(\mathbf{a}) &= m_{\mathbf{Q}}(\mathbf{a}) \prod_{i=1}^L (1 - 2p)^{b_i(a_i)} \\ &= m_{\mathbf{Q}}(\mathbf{a}) (1 - 2p)^{\sum_{i=1}^L b_i(a_i)} \end{aligned}$$

where  $b_i(a_i)$  is either 0 or 1. Because the moments of  $\mathbf{Y}$  are identifiable,  $\mathbf{Q}$  is identifiable for fixed  $p$  if

$$m_{\mathbf{Q}_1}(\mathbf{a}) (1 - 2p)^{\sum_{i=1}^L b_i(a_i)} = m_{\mathbf{Q}_2}(\mathbf{a}) (1 - 2p)^{\sum_{i=1}^L b_i(a_i)} \quad (5)$$

implies  $\mathbf{Q}_1 = \mathbf{Q}_2$ . If  $p < 1/2$ , (but possibly 0), then we can cancel  $(1 - 2p)^{\sum_{i=1}^L b_i(a_i)}$  from (5) to get  $m_{\mathbf{Q}_1}(\mathbf{a}) = m_{\mathbf{Q}_2}(\mathbf{a})$ , so  $\mathbf{Q}_1 = \mathbf{Q}_2$ . If  $p = 1/2$ , then  $m_{\mathbf{Y}}(\mathbf{a}) = 0$  for all  $\mathbf{a}$ , regardless of  $\mathbf{Q}$ , so we cannot determine  $\mathbf{Q}$  from  $\mathbf{Y}$ 's distribution.  $\square$

So far we have discussed the identifiability in the model when the number of components is not fixed. We will discuss other identifiability questions next.

### 3.5 Other identifiability questions

The nested structure of the ancestral mixtures implies that given any  $A(\mathbf{Q}, p)$  and any  $q < p$ , there exists  $\mathbf{Q}^*$  such that  $A(\mathbf{Q}, p) = A(\mathbf{Q}^*, q)$ . Therefore,  $(p, \mathbf{Q})$  jointly are not identifiable in the non-parametric sense.

A modelling technique one might use to overcome the identifiability problem in this case would be to fix the number of components in the distribution  $\mathbf{Q}$  at  $K$ . If we denote the mixing distribution as  $\mathbf{Q}_K$  for  $K$  fixed, then it can be proved that  $(p, \mathbf{Q}_K)$  are jointly identifiable in the reduced parameter space, provided  $K < 2^L$ . Note that  $\mathbf{Q}^*$  above has  $K = 2^L$  support points, so that  $K = 2^L$  must be excluded. We will discuss this modelling approach further in the next section.

#### 4. POINT ESTIMATION BY MAXIMUM LIKELIHOOD METHOD

The nested structure of the nonparametric ancestral mixture model implies that  $(p, \mathbf{Q})$  are not jointly identifiable. To overcome the identifiability problem, two possible approaches will be discussed here.

The first one is to use a fixed number of components  $K$  in the model, then jointly estimate the resulting parameters and the mutation rate  $p$ . However, compared with the second approach we develop later, this approach lacks uniqueness of the likelihood solutions (as the likelihood is potentially multimodal). In addition, only the second approach leads to a hierarchical tree.

The second approach is to use a fixed  $p$ , and model the mixing distribution nonparametrically. If we do so, the nonparametric maximum likelihood estimator (NPMLE) will automatically produce an estimate of the number of components in the ancestral model. This approach will have a unique solution, and as we show, will produce a tree structure as we vary the fixed  $p$ . Therefore, we will focus on the second approach.

Next we will show how to determine the NPMLE of the mixing distribution  $\mathbf{Q}$  with  $p$  fixed, then apply the properties of the NPMLE summarized in the monograph of Lindsay (1995).

##### 4.1 One ancestor with fixed $p$

We will start with maximum likelihood estimation in the single ancestor model with fixed  $p$ .

Following equation (2), a random sample  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  from the single ancestor model has the likelihood function

$$\begin{aligned} \mathbf{L}(\mu, p; \mathbf{x}) &= \prod_{i=1}^n \kappa(\mathbf{x}_i | \mu_1, p) \\ &= \bar{p}^{nL} \left( \frac{p}{\bar{p}} \right)^{\sum_{i=1}^n \mathbf{D}(\mathbf{x}_i, \mu_1)}. \end{aligned}$$

Since  $p$  is fixed and  $0 < p < 0.5$ , we have

$$0 < \left( \frac{p}{\bar{p}} \right)^{\sum_{i=1}^n \mathbf{D}(\mathbf{x}_i, \mu_1)} < 1.$$

If follows that maximizing  $\mathbf{L}$  is equivalent to minimizing  $\sum_{i=1}^n \mathbf{D}(\mathbf{x}_i, \mu_1)$ . But one can minimize  $\sum_{i=1}^n \sum_{j=1}^L (x_{ij} - \mu_{1j})^2$  for each  $j$  separately by choosing  $\mu_{1j}$  to minimize the number of disagreements. That is to say,  $\hat{\mu}_{1j}$ , the estimate of  $\mu_{1j}$ , is equal to the *majority vote* of the elements of the sequence  $x_{1j}, x_{2j}, \dots, x_{nj}$ . That is, the MLE for  $\mu_1$  is

$$\hat{\mu}_{1j} = \begin{cases} 1 & , \text{ if } \frac{1}{n} \sum_{i=1}^n x_{ij} > \frac{1}{2} \\ 0 & , \text{ if } \frac{1}{n} \sum_{i=1}^n x_{ij} < \frac{1}{2} \\ \text{either} & , \text{ if } \frac{1}{n} \sum_{i=1}^n x_{ij} = \frac{1}{2} \end{cases}$$

where  $j = 1, 2, \dots, L$ .

#### 4.2 Multiple ancestors with fixed $p$

In this section, we consider the NPMLE of the multiple ancestors mixture model with fixed  $p$ . For  $p = 0$ , an explicit solution can be found. For  $0 < p < 0.5$ , we will use the  $K$ -component EM algorithm as a tool in computing the NPMLE. For the case of multiple ancestors with fixed  $p$ , we have the following result.

**Lemma 2.** *When  $p = 0$ , the NPMLE  $\hat{\mathbf{Q}}$  of  $\mathbf{Q}$  for the ancestral mixture model is the empirical distribution of the sample,  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ . That is,  $\hat{\mathbf{Q}}$  has as support points the set of distinct sequences  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_D$ , with masses  $\pi_k = \frac{n_k}{n}$ , where  $n_k$  is the sample frequency of sequence  $\mathbf{y}_k$ .*

*Proof :* As  $p \rightarrow 0$ ,

$$\begin{aligned} \kappa_p(\mathbf{x}; \mu) &= \bar{p}^{2L} \left( \frac{p}{\bar{p}} \right)^{\mathbf{D}(\mathbf{x}, \mu)} \\ &\rightarrow \begin{cases} 1 & \text{if } \mathbf{D}(\mathbf{x}, \mu) = 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Thus the distribution of  $\mathbf{X}$  becomes degenerate at  $\mu$ . Thus, we can write the likelihood kernel as

$$\kappa_0(\mathbf{x}; \mu) = \mathbf{I}(\mathbf{x} = \mu) .$$

To check that  $\hat{\mathbf{Q}}$  is the NPMLE, it suffices to check a gradient inequality (Lindsay, 1995). Using the notation of Lindsay (1995, p. 115), it states that  $Q$  is a NPMLE if and only if

$$D_{\mathbf{Q}}(\mu) \leq 0, \quad \forall \mu \in \Omega. \tag{6}$$

Here

$$D_{\mathbf{Q}}(\mu) = \sum_{k=1}^D n_k \left[ \frac{\kappa(\mathbf{y}_k; \mu)}{\sum_j \pi_j \kappa(\mathbf{y}_k; \mu_j)} - 1 \right] .$$

We have

$$\begin{aligned}
L_j(\hat{Q}) &= \sum_k \hat{\pi}_k \kappa_0(\mathbf{y}_j; \hat{\mu}_k) \\
&= \sum_k \frac{n_k}{n} \mathbf{I}(\mathbf{y}_j = \mu_k) \\
&= \frac{n_j}{n}.
\end{aligned}$$

Hence, the gradient is

$$\begin{aligned}
D_{\hat{Q}}(\mu) &= \sum_j n_j \left[ \frac{\mathbf{I}(\mathbf{y}_j = \mu)}{\frac{n_j}{n}} - 1 \right] \\
&= n \sum_j \mathbf{I}(\mathbf{y}_j = \mu) - \sum_j n_j.
\end{aligned}$$

This function equals  $n - n = 0$  when  $\mu$  is one of the support points  $(\mathbf{y}_j)$  and equals  $-n < 0$  when it is not. This verifies the gradient characterization.  $\square$

As a computing approach to find the NPML for other values of  $p$ , we will follow the lead of Laird (1978), who suggested using the  $K$ -component EM algorithm with a large number of support points  $K$ . However, we will also utilize the sieve parameter  $p$  in order to get good starting values for the algorithm. From Lemma 2, we know the solution for  $p = 0$ . We might suppose that this solution is also a good set of starting values for  $p = 0.001$ . Similarly, the solution at  $p = 0.001$  obtained by the EM algorithm might be a good starting value for  $p = 0.002$ , and so forth. In this way, we will compute a linked family of estimates  $\hat{Q}_p$  for  $p$  on a grid. We will consider the consequences of this linkage after introducing the EM algorithm.

Then, we will show how to compute estimates of parameters using the EM algorithm introduced by Dempster et al. (1977). We will use multinomial indicator vectors to construct the ‘‘complete data’’ likelihood function.

For fixed  $p$ , consider  $K$  ancestor sequences  $\mu_{\mathbf{k}}$  with weights  $Pr(\theta = \mu_{\mathbf{k}}) = \pi_{\mathbf{k}}$ , giving a density

$$f(\mathbf{x}; Q_K) = \sum_{j=1}^K \pi_j \times \kappa(\mathbf{x} | \mu_j, p).$$

Let the multinomial indicator vector be defined for  $i = 1, \dots, n$  and  $j = 1, \dots, K$  as

$$Z_{ij} = \begin{cases} 1 & , \text{ if } \mathbf{x}_i \text{ originated from component } j \\ 0 & , \text{ otherwise.} \end{cases}$$

Then, the augmented-data likelihood function for individual  $i$  is

$$\begin{aligned}
L_i &= P(\mathbf{X}_i = \mathbf{x}_i, \mathbf{Z}_i = \mathbf{z}_i) \\
&= P(\mathbf{X}_i = \mathbf{x}_i | \mathbf{Z}_i = \mathbf{z}_i) \times P(\mathbf{Z}_i = \mathbf{z}_i) \\
&= \kappa(\mathbf{x}_i | \mu_1)^{Z_{i1}} \kappa(\mathbf{x}_i | \mu_2)^{Z_{i2}} \cdots \kappa(\mathbf{x}_i | \mu_K)^{Z_{iK}} \pi_1^{Z_{i1}} \pi_2^{Z_{i2}} \cdots \pi_K^{Z_{iK}} \\
&= \prod_{j=1}^K \pi_j^{Z_{ij}} \kappa(\mathbf{x}_i | \mu_j)^{Z_{ij}}.
\end{aligned}$$

Here

$$\begin{aligned}\kappa(\mathbf{x}_i|\mu_j)^{Z_{ij}} &= \left[ p^{\mathbf{D}(\mathbf{x}_i, \mu_j)} \times \bar{p}^{L-\mathbf{D}(\mathbf{x}_i, \mu_j)} \right]^{Z_{ij}} \\ &= \left[ \left( \frac{p}{\bar{p}} \right)^{\mathbf{D}(\mathbf{x}_i, \mu_j)} \times \bar{p}^L \right]^{Z_{ij}}\end{aligned}$$

and

$$\mathbf{D}(\mathbf{x}_i, \mu_j) = \sum_{s=1}^L (x_{is} - \mu_{js})^2.$$

The augmented-data likelihood function is therefore

$$\begin{aligned}L_a &= \prod_{i=1}^n L_i \\ &= \prod_{i=1}^n \prod_{j=1}^K \kappa(\mathbf{x}_i|\mu_j)^{Z_{ij}} \times \pi_j^{Z_{ij}}.\end{aligned}$$

Taking logarithms, we get

$$\log L_a = \sum_{i=1}^n \sum_{j=1}^K Z_{ij} [\log \kappa(\mathbf{x}_i|\mu_j) + \log \pi_j].$$

The EM algorithm involves two steps, the E-step and the M-step. The E-step is to calculate a function  $H$  that is defined as the expectation under current parameter values of the augmented data log-likelihood function conditional upon the observed data. The M-step is to maximize the function  $H$  with respect to the parameter over the parameter space.

In summary, the estimates for the parameters at  $(t+1)$ th iteration of EM algorithm are

$$\hat{\mu}_{js}^{(t+1)} = \begin{cases} 1 & , \text{ if } \frac{\sum_{i=1}^n \delta(\pi^{(t+1)}, \mu_j^{(t)} | \mathbf{x}_i) \times x_{is}}{\sum_{i=1}^n \delta(\pi^{(t+1)}, \mu_j^{(t)} | \mathbf{x}_i)} > \frac{1}{2} \\ 0 & , \text{ if } \frac{\sum_{i=1}^n \delta(\pi^{(t+1)}, \mu_j^{(t)} | \mathbf{x}_i) \times x_{is}}{\sum_{i=1}^n \delta(\pi^{(t+1)}, \mu_j^{(t)} | \mathbf{x}_i)} < \frac{1}{2} \\ \text{either} & , \text{ if } \frac{\sum_{i=1}^n \delta(\pi^{(t+1)}, \mu_j^{(t)} | \mathbf{x}_i) \times x_{is}}{\sum_{i=1}^n \delta(\pi^{(t+1)}, \mu_j^{(t)} | \mathbf{x}_i)} = \frac{1}{2}. \end{cases} \quad (7)$$

and

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^n \delta(\pi^{(t)}, \mu_j^{(t)} | \mathbf{x}_i)}{n}$$

where

$$\delta(\pi^{(t)}, \mu_j^{(t)} | \mathbf{x}_i) = \frac{\pi_j^{(t)} \times \kappa(\mathbf{x}_i|\mu_j^{(t)})}{\sum_{j=1}^K \pi_j^{(t)} \times \kappa(\mathbf{x}_i|\mu_j^{(t)})}.$$

Note that the structure of the model makes the third case, the ties, in (7) extremely rare. It also makes no difference in the EM likelihood which  $\mu_j$  we use in this case. We arbitrarily programmed  $\hat{\mu}_j^{(t+1)}$  to be 0 if  $\frac{1}{2}$  occurred.

## 5. INDUCED TREE STRUCTURE

In this section, we will discuss how the NPMLE of the ancestral model under varying  $p$  induces a tree structure.

If we take a set of data to compute the NPMLE for fixed  $p$ , we will find some random number of components between 0 and  $D$  (Lindsay, 1995). If we compute the NPMLE  $\hat{\mathbf{Q}}_p$  for each  $p$  between 0 and 0.5, then we will find a linked sequence of mixture estimates. If we equate  $p$  (actually  $\eta = -\log(1 - 2p)$ ) with “number of generations of mutation”, it is intuitively clear that the further back in time we go (larger  $\eta$ ) the fewer the ancestors.

From Proposition 3, we know the parameters in the model are not identifiable when  $p = 0.5$ . When  $p$  is near 0.5, no matter what the ancestral sequences are, the sequences we observe look very similar to sequences generated by randomly tossing a coin to decide success or failure at each site, with nearly equal probability of choosing 0 and 1. Thus we might expect a single ancestor sequence to fit the data adequately. At the other extreme, when  $p = 0$ , we have already seen in Lemma 2, that each distinct observed sequence becomes a support point. These  $D$  sequences are the ancestral sequences with weights equal to  $n_k/n$ .

This heuristic thinking gives us a feeling for how the ancestral mixture model might generate a tree structure. Let us treat the  $y$ -axis as  $\eta$ , where we can think of  $\eta$  representing time measured in amount of mutation. When  $p$  is close to 0.5, so  $\eta$  is very large, which is at the top of the tree, the NPMLE will estimate that the data are from one common ancestor  $\mu$ . In biological terms, this ancestor might be called the most recent common sequence (MRCS) of the sample sequences. Moreover, we know from the results of section that the MRCS is the majority winner of the sequences in each site. As time goes forward (*i.e.* when  $p$  decreases), we anticipate that the single ancestor will split into two ancestor sequences (*i.e.*  $\hat{\mathbf{Q}}_p$  has two support points), then three, and so forth until we reach  $D$  ancestral sequences at  $p = 0$ , where  $D$  is the number of observed distinct sequences.

From this argument, we conjecture that the NPMLE of the ancestral mixture model under varying  $p$  will induce a tree structure, something like an ancestral tree. We will call these linked estimators  $(p, \hat{\mathbf{Q}}_p)$  the *induced mixture tree*. When we compute the tree from the bottom up, our computational experience is that as we increase  $p$ , the number of components decreases monotonically, with two components that are distinct at one  $p$  merging at the next  $p$ .

After fitting the NPMLE for a fixed  $p$ , one can assign a sequence,  $\mathbf{x}$  to the component  $j$  that maximizes its posterior probability  $\delta(\pi, \mu_j | \mathbf{x})$  of being from this component. In the process we not only get a cluster assignment, but also a measure of the certainty with which the assignment is made. Different values of  $p$  will give a range of possible numbers of clusters to use.

## 6. CASE STUDY

In this section, we will present as a case study a computer analysis of a data set. We fit the ancestral mixture model to the Mitochondrial DNA sequences found in Griffiths and Tavaré (1994). The performance of our approach and some computing issues for our algorithm will be discussed in this section.

### 6.1 Mitochondrial sequences from Griffiths and Tavare (1994)

The data we used for our investigation is from the paper of Griffiths and Tavare (1994). These Mitochondrial DNA sequences first appeared in the paper of Ward et al. (1991). For studying the mitochondrial diversity within the Nuu-Chuah-Nulth, an Amerindian tribe from Vancouver Island, Ward et al. (1991) sequenced 360 nucleotide segments of the mitochondrial control region for 63 individuals from the Nuu-Chuah-Nulth. Because of the coalescent theory assumption that substitution at any nucleotide position can only occur once, Griffiths and Tavare (1994) eliminated some lineages and 8 of the pyrimidine segregating sites from the Mitochondrial DNA sequences of Ward et al.. This resulted in a subsample comprised of 55 of the 63 distinct sequences and 18 segregating sites including 13 pyrimidines ( $C, T$ ) and 5 purines ( $A, G$ ). The Mitochondrial DNA sequences are shown in Table 1. In the table, each lineage represents a distinct sequence. The frequency of a lineage represents the total number of individuals who have the same sequence. Next, we will apply the ancestral mixture model to these Mitochondrial DNA sequences.

Table 1: Mitochondrial data from Griffiths and Tavare (1994)

| Position | 1 | 1 | 2 | 2 | 3 | 8 | 9 | 2 | 4 | 6  | 6  | 9  | 3  | 6  | 7  | 7  | 3  | 3  | Lineage<br>freqs. |
|----------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|-------------------|
| Site     | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |                   |
| Lineage  |   |   |   |   |   |   |   |   |   |    |    |    |    |    |    |    |    |    |                   |
| a        | A | G | G | A | A | T | C | C | T | C  | T  | T  | C  | T  | C  | T  | T  | C  | 2                 |
| b        | A | G | G | A | A | T | C | C | T | T  | T  | T  | C  | T  | C  | T  | T  | C  | 2                 |
| c        | G | A | G | G | A | C | C | C | T | C  | T  | T  | C  | C  | C  | T  | T  | T  | 1                 |
| d        | G | G | A | G | A | C | C | C | T | C  | T  | T  | C  | C  | C  | T  | T  | C  | 3                 |
| e        | G | G | G | A | A | T | C | C | T | C  | T  | T  | C  | T  | C  | T  | T  | C  | 19                |
| f        | G | G | G | A | G | T | C | C | T | C  | T  | T  | C  | T  | C  | T  | T  | C  | 1                 |
| g        | G | G | G | G | A | C | C | C | T | C  | C  | C  | C  | C  | C  | T  | T  | T  | 1                 |
| h        | G | G | G | G | A | C | C | C | T | C  | C  | C  | T  | C  | C  | T  | T  | T  | 1                 |
| i        | G | G | G | G | A | C | C | C | T | C  | T  | T  | C  | C  | C  | C  | C  | T  | 4                 |
| j        | G | G | G | G | A | C | C | C | T | C  | T  | T  | C  | C  | C  | C  | T  | T  | 8                 |
| k        | G | G | G | G | A | C | C | C | T | C  | T  | T  | C  | C  | C  | T  | T  | C  | 5                 |
| l        | G | G | G | G | A | C | C | C | T | C  | T  | T  | C  | C  | C  | T  | T  | T  | 4                 |
| m        | G | G | G | G | A | C | C | T | T | C  | T  | T  | C  | C  | C  | T  | T  | C  | 3                 |
| n        | G | G | G | G | A | C | T | C | T | C  | T  | T  | C  | C  | T  | T  | T  | C  | 1                 |

### 6.2 Mixture tree by the ancestral mixture modelling

Using the Mitochondrial DNA sequences in Table 1, we first choose the majority of each site as the sequence type of MRCS, which is the sequence ( $G, G, G, G, A, C, C, C, T, C, T, T, C, C, C, T, T, C$ ). Next we re-code those Mitochondrial DNA sequences into sequences with (0, 1) coding, where 0 represents the estimated type of the MRCA sequence, and 1 represents the opposite. For  $p = 0$ , the number of the components is  $K = 14$  and the estimated ancestral types  $\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_{14}$  are the original coded lineages  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{14}$ .

As we move  $\eta$  slowly upward and run the EM algorithm at each  $\eta$ , we will see that the support points  $\hat{\mu}_1(\eta), \dots, \hat{\mu}_K(\eta)$  will stay constant for a period, then *evolve* into a new set at some  $\eta$ . That is, at some point in “time”, say  $\eta_0$ , a *merger* will occur in the sense that  $\hat{\mu}_k(\eta_0) = \hat{\mu}_j(\eta_0)$  for some

$(k, j)$  pair. The path taken by an individual  $\hat{\mu}_k(\eta)$  as  $\eta$  varies from 0 upward will be called the  $k$ th *lineage*. In the data of Table 1, we will be able to trace the lineages  $a$  to  $n$  backward in mutation time. At any point in time  $\eta_0$ , we can identify different lineages that have merged with each other as corresponding to *clusters* of the original sequences. The gradient check that enables us to check if we have found the NPMLE was discussed in Lindsay (1995).

In practice, there are two problems with using (6) to verify whether  $\hat{\mathbf{Q}}$  is an NPMLE. First,  $\hat{\mathbf{Q}}$  is usually obtained by an algorithm, so for any stopping rule in the algorithm there is limited accuracy, which means that (6) is inherently violated. Secondly, checking the inequality for a large number of values of  $\mu$  is time consuming.

As a solution to the second problem Lesperance and Kalbfleisch (1992) proposed to take a basic grid, then search the neighborhood of each grid point to see if there is gradient violation. That is, define a subset  $\Omega_s$  of  $\Omega$ , and check if the inequality (6) holds for all  $\mu \in \Omega_s$ . In our problem  $\Omega = \{0, 1\}^L$  has  $2^L$  elements; for  $L$  large, this is an infeasible space to search. Here, we will use the gradient stopping rule by defining the subset  $\Omega_s = \{\text{the support points } \hat{\mu} \text{ in } \hat{\mathbf{Q}}, \text{ and all the original distinct sequences } \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_D \}$ .

Even with the use of a grid basis, the algorithm will not stop in a finite number of steps unless we allow a tolerance in (6). Lindsay (1995) proposed that one allow a small positive tolerance value on the gradient function and stop the algorithm if

$$D_{\mathbf{Q}}(\mu) \leq tol, \quad \forall \mu \in \Omega_s$$

where  $tol$  is a small positive number. An important property of this rule is that it guarantees that the current log-likelihood is near its final maximized value ( $\ln(L(\hat{\mathbf{Q}})) - \ln(L(\mathbf{Q})) \leq tol$ , where  $\hat{\mathbf{Q}}$  is the maximizing mixture). Following Lindsay (1995, p. 131), we used  $tol = .005$ .

Using the gradient stopping rule proposed the previous section, we did C++ programming of the EM algorithm. We estimated the ancestral mixture model by sliding  $p$  in increments of 0.01 and the 0.001. The clusters obtained by using a 0.01 sliding scale are listed in Table 2, and those obtained with a 0.001 sliding scale are listed in Table 3.

In Table 2, the first row shows that, at  $p = 0.09$ , lineages  $e$  and  $f$  merged due to a mutation occurring at site 5 in the  $\mu$  for lineage  $f$  so that the  $\mu$ -values became identical. Thus 13 clusters are identified between  $p = 0.09$  and  $p = 0.19$ ; lineages  $e$  and  $f$  form one cluster and the other 12 lineages are 12 distinct clusters. At  $p = 0.19$ , we see that lineages  $c$  and  $l$  merged due to a mutation in lineage  $c$  at site 2.

The goal of sliding  $p$  at a slow rate is twofold. First we wish to keep good starting values for the EM. Secondly, we wish to accurately capture the merging of support points. In Table 2, we see that three mergers happened “simultaneously” at  $p = 0.32$ . In order to capture the times and order of mergers, we did further sliding of  $p$  between 0.31 and 0.32 using a 0.001 scale. This showed that lineage  $i$  first merges with lineage  $j$  at  $p = 0.314$ , later lineage  $g$  merges with lineages  $c$  and  $l$  at  $p = 0.316$ , and finally lineage  $h$  merges with lineages  $c$ ,  $g$  and  $l$  at  $p = 0.320$ . Notice that the *structures* of the two trees in Tables 2 and 3 are the same in terms of the order of events. There are only small differences in the mutation times  $p$  between these two trees.

### 6.3 Performance of the tree algorithm

Table 2: *Clusters of Mitochondrial DNA data from Griffiths and Tavare (1994) using 0.01 scale*

| <b>p</b> | $-\log(1 - 2p)$ | <b>Clusters</b>   | <b>Mutated Site</b>             |
|----------|-----------------|---|---------------------------------|
| 0.09     | 0.198451        | { (e, <b>f</b> ), a, b, c, d, g, h, i, j, k, l, m, n }  | f: 5 (Site 5 of lineage f)      |
| 0.19     | 0.478036        | { (e, f), ( <b>c</b> , l), a, b, d, g, h, i, j, k, m, n }   | c: 2                            |
| 0.24     | 0.653926        | { ( <b>a</b> , e, f), (c, l), b, d, g, h, i, j, k, m, n }   | a: 1                            |
| 0.27     | 0.776529        | { (a, e, f), (c, l), ( <b>k</b> , <b>n</b> ), b, d, g, i, j, m }                                  | n: 7, 15                        |
| 0.28     | 0.820981        | { (a, e, f), (c, l), ( <b>k</b> , <b>m</b> , n), b, d, g, h, i, j }                               | m: 8                            |
| 0.29     | 0.867501        | { ( <b>a</b> , <b>b</b> , e, f), (c, l), (k, m, n), d, g, h, i, j }                               | b: 1, 10                        |
| 0.32     | 1.02165         | { (a, b, e, f), (c, <b>g</b> , <b>h</b> , l), ( <b>i</b> , j), (k, m, n), d }                     | i: 17; g: 11, 12; h: 11, 12, 13 |
| 0.35     | 1.20397         | { (a, b, e, f), (c, g, h, l), (i, j), ( <b>d</b> , k, m, n) }                                     | d: 3, 9                         |
| 0.42     | 1.83258         | { (a, b, e, f), ( <b>i</b> , <b>j</b> , c, g, h, l), (d, k, m, n) }                               | (i,j): 16                       |
| 0.45     | 2.30259         | { (a, b, e, f), ( <b>c</b> , <b>g</b> , <b>h</b> , <b>l</b> , <b>i</b> , <b>j</b> , d, k, m, n) } | (c, g, h, l, i, j): 18          |
| 0.49     | 3.91202         | { ( <b>a</b> , <b>b</b> , <b>e</b> , <b>f</b> , c, d, g, h, i, j, k, l, m, n) }                   | (a, b, e, f): 4, 6, 14          |

Table 3: *Clusters of Mitochondrial DNA data from Griffiths and Tavare (1994) using 0.001 scale*

| <b>p</b> | $-\log(1 - 2p)$ | <b>Clusters</b>   | <b>Mutated Site</b>        |
|----------|-----------------|---|----------------------------|
| 0.084    | 0.183923        | { (e, <b>f</b> ), a, b, c, d, g, h, i, j, k, l, m, n }  | f: 5 (Site 5 of lineage f) |
| 0.188    | 0.471605        | { (e, f), ( <b>c</b> , l), a, b, d, g, h, i, j, k, m, n }   | c: 2                       |
| 0.239    | 0.650088        | { ( <b>a</b> , e, f), (c, l), b, d, g, h, i, j, k, m, n }   | a: 1                       |
| 0.264    | 0.750776        | { (a, e, f), (c, l), ( <b>k</b> , <b>n</b> ), b, d, g, i, j, m }                                  | n: 7, 15                   |
| 0.272    | 0.785262        | { (a, e, f), (c, l), ( <b>k</b> , <b>m</b> , n), b, d, g, h, i, j }                               | m: 8                       |
| 0.284    | 0.83933         | { ( <b>a</b> , <b>b</b> , e, f), (c, l), (k, m, n), d, g, h, i, j }                               | b: 1, 10                   |
| 0.314    | 0.988861        | { (a, b, e, f), (c, l), ( <b>i</b> , j), (k, m, n), d, g, h }                                     | i: 17                      |
| 0.316    | 0.999672        | { (a, b, e, f), (c, <b>g</b> , l), (i, j), (k, m, n), d, h }                                      | g: 11, 12                  |
| 0.320    | 1.02165         | { (a, b, e, f), (c, g, <b>h</b> , l), (i, j), (k, m, n), d }                                      | h: 11, 12, 13              |
| 0.342    | 1.15201         | { (a, b, e, f), (c, g, h, l), (i, j), ( <b>d</b> , k, m, n) }                                     | d: 3, 9                    |
| 0.422    | 1.8579          | { (a, b, e, f), ( <b>i</b> , <b>j</b> , c, g, h, l), (d, k, m, n) }                               | (i,j): 16                  |
| 0.447    | 2.24432         | { (a, b, e, f), ( <b>c</b> , <b>g</b> , <b>h</b> , <b>l</b> , <b>i</b> , <b>j</b> , d, k, m, n) } | (c, g, h, l, i, j): 18     |
| 0.489    | 3.81671         | { ( <b>a</b> , <b>b</b> , <b>e</b> , <b>f</b> , c, d, g, h, i, j, k, l, m, n) }                   | (a, b, e, f): 4, 6, 14     |

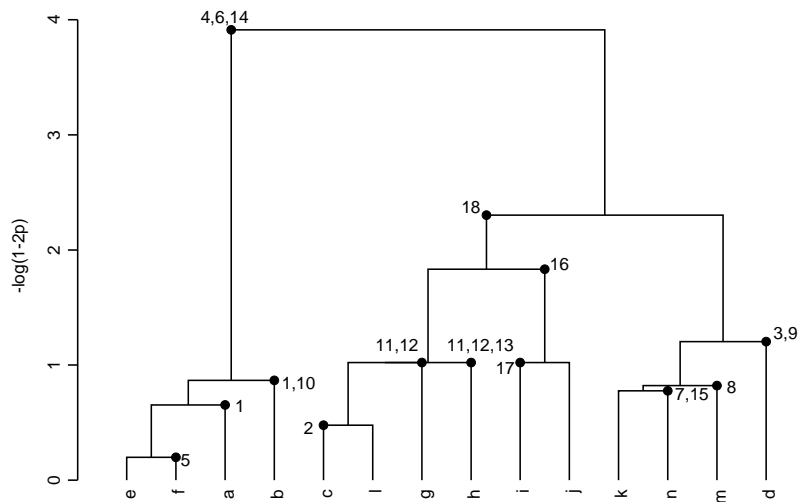


Figure 1: *Mixture tree of Mitochondrial DNA data from Griffiths and Tavare (1994) using 0.01 scale and natural time parameter.*

A graphical representation of the mixture tree for the sliding scale .001 using the natural time parameter  $\eta = -\log(1 - 2p)$  is shown in Figure 1. To illustrate, if one starts at the bottom of the tree and moves upwards along the branch of lineage  $f$ , one finds the merger at  $p = 0.09$ , where  $-\log(1 - 2p) = 1.98451$ . At this time, site 5 in lineage  $f$  mutates, resulting in the same ancestor  $\mu$  as in lineage  $e$ . Therefore lineages  $e$  and  $f$  merge.

Our experience was that the tree algorithm performed quite reliably. The EM stopped in a few steps, each site mutated only once, and the two mixture trees using different sliding scales were consistent with each other. In our investigation, we expected that the  $p$  at which the lineages merged in 0.001 scale should be slightly smaller than those in 0.01 scale. We found one discrepancy at  $p = 0.422$ . So, we did some further investigation at this time point. Indeed the “votes” to decide  $\hat{\mu}_{j_t}$  were very close to 50/50; we believe that this discrepancy was caused by computing rounding error.

#### 6.4 Numerical questions

In this section, we investigate two important points regarding the success of our EM algorithm. We wish (1) to know whether our initial values for our EM algorithm were adequate to find the NPMLE and (2) to demonstrate that the EM converged adequately by the gradient stopping rule.

First, we consider the initial value problem. We have proposed that in the tree algorithm, the initial value of the EM at the current  $p$  should be given by the estimate  $\hat{\mathbf{Q}}_p$  obtained for the previous value of  $p$ . We did the following investigation of this method.

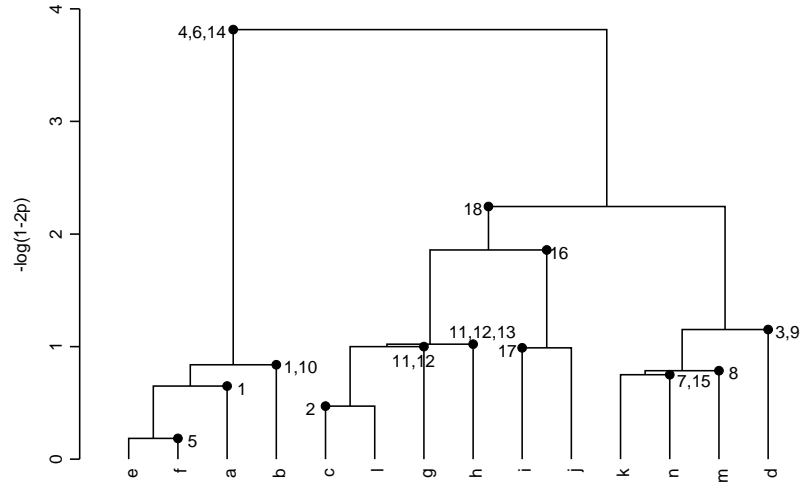


Figure 2: Mixture tree of Mitochondrial DNA data from Griffiths and Tavaré (1994) using 0.001 scale and natural time parameter.

For a given  $p$ , we compare the  $L(\hat{\mathbf{Q}}, p)$  and the total number of iterations for three different methods to select initial values which will be described in the following manner: (1) the empirical sequences,  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_{14}$ ; (2) the estimate obtained for the previous value of  $p$ , sliding by 0.02 until the fixed  $p$ ; (3) the estimate obtained for the previous value of  $p$ , sliding by 0.001 until the fixed  $p$ .

Log-likelihood values and total number of iterations are investigated for  $p = 0.08, 0.16, 0.22, 0.34, 0.40$  and the three methods. Not surprisingly, the estimate based on the smaller sliding scale is usually the best one in terms of likelihood. However, all values are very close, suggesting that the solutions are robust to the choice of starting methods.

Secondly, we would like to verify that the EM has adequately converged under the gradient stopping rule. To do so we compared the log-likelihood value from the gradient stopping rule with the log-likelihood value obtained by running 1000 iterations in each call of the EM algorithm, a technique that creates considerably more accuracy at the expense of long run times.

We estimated  $\mathbf{Q}$  at  $p = 0.34$  and the values of  $\mu$  are identical in the two runs, although the values of  $\pi$  differ slightly. The result was that the second likelihood is only 0.0000018 better than the first, much less difference than the target tolerance of .005. We also did comparisons with  $p = 0.07, 0.15, 0.23, 0.34, 0.38, \text{ and } 0.43$ . As predicted by the gradient stopping rule, no case occurred where the tolerance bound was exceeded by the difference in log-likelihood.

In conclusion, we found that our tree algorithm is reliable in finding the maximum and that the stopping rule provides high accuracy without excessive computation.

## 7. DISCUSSION

In this article, we proposed an ancestral mixture model for clustering high dimensional binary sequences, such as SNP data. We proved that the mixing distribution  $\mathbf{Q}$  in the nonparametric ancestral mixture model is identifiable for each fixed  $p$  in  $[0, 0.5)$ , and so we could use nonparametric estimation methods. By sliding the sieve parameter representing the mutation probability  $p$ , one can create a hierarchical tree which estimates the population structure at each fixed backward point in time. The parameter  $p$  was transformed to  $\eta = -\log(1 - 2p)$ , a natural time scale for the mutational process.

The nonparametric maximum likelihood estimate (NPMLE) in the ancestral mixture model was implemented via the EM algorithm using a gradient stopping rule. By computing the NPMLE for each fixed value of the sieve parameter  $p$ , we created a linked sequence of mixture estimates. Finally, a hierarchical mixture tree was constructed from these linked estimators, giving cluster relationships that can be visually identified. After the clusters are identified, any individual sequence can be assigned to a cluster in a probabilistic way. This technique was shown to be powerful in clustering bi-allelic genetic sequences.

After building up the hierarchical tree, we might be interested in selecting a value of  $\eta$  such that the model fits well. In doing so, we also obtain  $\hat{\mathbf{Q}}_\eta$ , which determines the number of components needed to fit the data well. In future work, we will propose a quadratic distance based approach for model selection. As for studying the variation properties of the ancestral mixture trees method, we will in the future report on a nonparametric bootstrap analysis.

Some extensions of the ancestral mixture model will be developed in the future. In many important genetics problems, the sequence data is polytomous in nature. For instance, in disease studies, the site variable  $X_s$  might indicate which one of many “marker alleles” is present; such a site is called a multi-allelic locus in biological terms. Also, in some biological sequences, different nucleotides seem to have different rates of mutation as well as different components of the population might have different mutation rates.

## ACKNOWLEDGMENTS

This research was partially supported by National Science Foundation, via Grant 0104443.

## References

- Celeux, G. and Govaert, G. (1991). Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8:157–176.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (C/R: p22-37). *Journal of the Royal Statistical Society, Series B, Methodological*, 39:1–22.
- Govaert, G. (1990). Classification binaire et modeles’. *Revue Statist. Appliquee*, 38:67–81.

- Govaert, G. and Nadif, M. (1996). Comparison of the mixture and the classification maximum likelihood in cluster analysis with binary data. *Computational Statistics and Data Analysis*, 23:65–81.
- Griffiths, R. C. and Tavaré, S. (1994). Ancestral inference in population genetics. *Statistical Science*, 9:307–319.
- Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73:805–811.
- Lesperance, M. L. and Kalbfleisch, J. D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association*, 87:120–126.
- Lindsay, B. G. (1995). *Mixture models: Theory, geometry and applications*. Institute of Mathematical Statistics.
- McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: inference and applications to clustering*. Marcel Dekker Inc.
- Ott, J. (1999). *Analysis of human genetic linkage*. Johns Hopkins University Press.
- Sachidanandam, R. e. a. T. i. S. m. w. g. (2001). A map of human genome sequence variation contain 1.42 million single nucleotide polymorphisms. *Nature*, 409:928–933.
- Settimi, R. and Smith, J. Q. (2000). Geometry, moments and conditional independence trees with hidden variables. *The Annals of Statistics*, 28(4):1179–1205.
- Ward, R. H., Frazier, B. L., Dew-Jager, K., and Paabo, S. (1991). Extensive mitochondrial diversity within a single amerindian tribe. *Proc. Natl. Acad. Sci.*, pages 8720–8724.