Chapter 7
Mixture Tree Construction and Its Applications

Grace S.C. Chen, Mingze Li, Michael Rosenberg, and Bruce Lindsay

Abstract A new method for building a gene tree from Single Nucleotide Polymorphism (SNP) data was developed by Chen and Lindsay (Biometrika 93(4):843–860, 2006). Called the mixture tree, it was 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 varying the sieve parameter, one can create a hierarchical tree that estimates the population structure at each fixed backward point in time. In this chapter, we will review the model and then present an application to the clustering of the mitochondrial sequences to show that the approach performs well. A simulator that simulates real SNPs sequences with unknown ancestral history will be introduced. Using the simulator we will compare the mixture trees with true trees to evaluate how well the mixture tree method performs. Comparison with some existing methods including neighbor-joining method and maximum parsimony method will also be presented in this chapter.

7.1 Introduction

There are two major families of methods for building phylogenetic trees: character-based and distance-based. For the character-based methods, the Maximum Parsimony (MP), the method of Maximum Likelihood (ML), and Bayesian methods are the most well-known ones.

G.S.C. Chen (✉) and M. Li
School of Math & Stat, Arizona State University, Tempe, U.S.A.
e-mail: scchen@math.asu.edu, mingzeli@asu.edu

M. Rosenberg
School of Life Sciences, Arizona State University, Tempe, U.S.A.
e-mail: msr@asu.edu

B. Lindsay
Department of Statistics, Penn State University, University Park, U.S.A.
e-mail: bgl@psu.edu

H. Horng-Shing Lu et al. (eds.), *Handbook of Statistical Bioinformatics*,
Springer Handbooks of Computational Statistics, DOI 10.1007/978-3-642-16345-6_7,

Among these methods, the Parsimony method was introduced by Edwards and Cavalli-Sforza [3], and is one of the first methods to be used to infer phylogeny. A phylogeny having fewer changes to account for the way a group of sequences has evolved is preferable. In other words, the most parsimonious explanation for the observed data is sought. In the method of Maximum Parsimony [2] the tree with the shortest branch lengths is the best. The steps to create this tree are as follows. First, informative sites, or sites where at least two different states occur in at least two taxa, are identified. A subset of trees (or all trees for less than a dozen taxa) is evaluated using a heuristic approach, and the tree with the shortest branch length is chosen.

For cases where there are large amounts of evolutionary changes in different branches of a tree, the method of Maximum Likelihood (ML) is to be preferred. Maximum Likelihood was created by Ronald A. Fisher [6–8] and later applied to gene frequency data for phylogenies by Edwards and Cavalli-Sforza [4] and to nucleotide sequences by Felsenstein [5]. This computationally intensive but flexible method searches for the tree with highest probability of producing the observed data. The likelihood of each residue in an alignment is calculated based on some model of the substitution process.

Unlike ME and MP, the ML and Bayesian methods make use of all of the information contained within an alignment of DNA sequences. Both ML and Bayesian methods rely on a likelihood function, L(Parameter) = Constant × Prob [Data—Parameter(s)], where the constant is arbitrary and the probability of observing the data conditioned on the parameter is calculated using stochastic models [10]. In ML, the combination of parameters that maximizes the likelihood function is the best estimate. In Bayesian analysis, the joint probability distribution of the parameters is calculated. The posterior probability distribution for the parameters is the likelihood function times the prior probability distribution of the parameters divided by a function of the data. However, unlike ML, Bayesian methods treat parameters as random variables.

Minimum Evolution (ME) is a distance-based approach. In this method, the tree is fit to the data, and the branch lengths are determined using the unweighted least squares method. In this method, distance measures that correct for multiple hits at the same sites are used, and a topology showing the smallest value of the sum of all branches is chosen as an estimate of the correct tree.

When there are a large number of taxa, ME is time consuming, so the neighbor-joining method can be used instead. The Neighbor Joining (NJ) method [17] is a clustering method that minimizes the sum of the branch lengths (this is an approximation to the ME method). The algorithm begins with a star-like structure. Pairwise comparisons are made to determine the most closely related sequences that are connected by a single node, called neighbors. Neighbors form a clade, and the process repeats until the topology is complete.

The NJ and the ME tree are generally the same, but when the number of taxa is small the difference between the trees can be considerable [12]. If a long DNA or amino acid sequence is used, the ME tree is preferable. When the number of nucleotides or amino acids used is relatively small, the NJ method generates the correct topology more often than does the ME method [13, 18]. MEGA uses the close-neighbor-interchange search to examine the neighborhood of the NJ tree to find the potential ME tree.

Unlike NJ, the Unweighted Pair-Group Method with Arithmetic mean (UPGMA) assumes a molecular clock that is constant. This simple distance-based clustering algorithm is significantly less accurate than Neighbor Joining. Each sequence is assigned to its own cluster then new clusters are formed based on having a minimal distance between them. The UPGMA trees are always rooted, and the total branch length from the root to any tip is equal (i.e., the tree is ultrametric). Finding the root requires an outgroup or is given at the midpoint of the longest distance connecting two taxa in the tree.

In this chapter, we will review the mixture tree model and algorithm proposed by Chen and Lindsay [1] in Sect. 7.2 and then in Sect. 7.2.2 present an application to the clustering of the mitochondrial sequences to show that the approach performs well. A simulator that simulates real SNPs sequences with unknown ancestral history will be introduced. Using the simulator we will compare the mixture trees with true trees to evaluate how well the algorithm performs. Comparison with some existing methods including neighbor-joining method, and the maximum parsimony method will also be presented in Sect. 7.3.

7.2 Mixture Tree Algorithm

In this section, we will briefly reviewed the Ancestral mixture model and the Mixture Tree algorithm introduced in the paper Chen and Linsay [1].

7.2.1 Ancestral Mixture Model

The ancestral mixture model implements K-component mutation kernel mixture density to estimate the most common ancestor and the evolving history(phylogeny) of the observed binary DNA sequences. Suppose we observed a sample of binary DNA sequences $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$ of length L for a fixed mutation rate p. As all the sequences are binary, we can code one state 0 and the opposite 1. If we assume that they evolved from a single ancestor of length L, say $\boldsymbol{\mu}_1$, and we define μ_{1j} as the j^{th} site of $\boldsymbol{\mu}_1$, the mutation kernel density for $\mathbf{X}$ is defined as

$$\kappa(\mathbf{x}|\boldsymbol{\mu}_1, p) = \prod_{j=1}^{L} p^{(x_j-\mu_{1j})^2}(1-p)^{1-(x_j-\mu_{1j})^2} = p^{D(\mathbf{x},\boldsymbol{\mu}_1)}(1-p)^{L-D(\mathbf{x},\boldsymbol{\mu}_1)},$$

where $D(\mathbf{x}, \boldsymbol{\mu}_1) = \sum_{j=1}^{L}(x_j - \mu_{1j})^2$ is the number of disagreements between the site of $\mathbf{x}$ and the corresponding site of $\boldsymbol{\mu}_1$.

If the observed sample is evolving from K different ancestors, say $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$, and we consider ϑ as a random variable with distribution Q, where Q is a discrete distribution with K points of support which are the K ancestors, and $pr(\vartheta = \boldsymbol{\mu}_k) = \pi_k$, where $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$, then we suppose $\mathbf{X}$ is generated by first generating $\vartheta = \boldsymbol{\mu}_k$ from Q, and generating $\mathbf{X} = \mathbf{x}$ from $\kappa(\mathbf{x}|\boldsymbol{\mu}_k, p)$. ϑ is unobserved, and such $\mathbf{X}$ is said to have an ancestral mixture model: $\mathbf{X} \sim A(Q, p)$. The density of $\mathbf{X}$, when Q is discrete, is:

$$f(\mathbf{x}; Q, p) = \sum_{k=1}^{K} \pi_k p^{D(\mathbf{x},\mu_k)} (1-p)^{L-D(\mathbf{x},\mu_k)},$$

which is called a 'Q-mixture of mutation kernels'.

7.2.1.1 Mixture Tree Algorithm

In order to find the MLE of π_j and $\boldsymbol{\mu}_j$, where $j = 1, \ldots, K$, an EM algorithm is employed. Give a value $Q^{(1)} = (\pi_1^{(1)}, \pi_2^{(1)}, \ldots, \pi_{k-1}^{(1)}, \boldsymbol{\mu}_1^{(1)}, \ldots, \boldsymbol{\mu}_K^{(1)})$for the mixture, standard EM calculations give

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} \delta(j|\mathbf{x_i}; \pi^{(t)}, \boldsymbol{\mu}^{(t)})}{n},$$

where

$$\delta(j|x; \pi^{(t)}, \boldsymbol{\mu}^{(t)}) = \frac{\pi_j \times \kappa(\mathbf{x}|\boldsymbol{\mu}_j)}{\sum_{j=1}^{K} \pi_j \times \kappa(\mathbf{x}|\boldsymbol{\mu}_j)}$$

We then reupdate the δ weights using the new π before update μ. During the E-step, the expected percentage of category 1 occurrences at site s in component j as

$$\upsilon_{js} = \frac{\sum_{i=1}^{n} \delta(j|\mathbf{x_i}; \pi^{(t+1)}, \mu^{(t)}) \times x_{is}}{\sum_{i=1}^{n} \delta(j|\mathbf{x_i}; \pi^{(t+1)}, \mu^{(t)})}$$

and in the M-step, we find the MLE of the parameter by 'voting' according to

$$\hat{\mu}_{js}^{(t+1)} = \begin{cases} 1 & \upsilon_{js} > \frac{1}{2}, \\ 0 & \upsilon_{js} < \frac{1}{2}, \\ either & \upsilon_{js} = \frac{1}{2}. \end{cases}$$

A tie in the third case in this structure of the model is extremely rare and it makes no difference in the EM likelihood.

7.2.1.2 An Alternative Revised Algorithm

The EM algorithm employed in the mixture models has computational problem such as small weight π_i problem. It is nature to propose an alternative revised EM that the weights π_i is not updated. We will call such revised EM the 'FixEM'. Later on, we will compare EM with FixEM in the simulation section.

7.2.2 An Example

In this section we will compare the mixture tree (MT) method with the Neighbor-joining tree and Maximum Parsimony tree in a visual way and give an example of the mixture tree structure by using the real data set in the paper [20]. This dataset can be downloaded from Genbank. There are 530 mtDNA sequences(population) in HVS1 region with different length and they are collected from people living in 17 locations(sub-populations) in East Asia who belong to two official ethnic groups, Miao and Yao, and the sample sizes within each location are different. Before constructing the trees using different methods, we did some necessary manipulations to the sequences:

1. Aligned all the sequences using MEGA4 with default setting.
2. Deleted those sites with gaps
3. Deleted those sites that are not binary
4. When applying mixture algorithm, deleted those sites that are identical

7.2.2.1 Trees Based on the Sample Contains One Random Sequence from Each Sub-population

After applying the above manipulations to all sequences, we constructed trees using four different methods: NJ, MP, ML and MixtureTree algorithm. It is time consuming and resulting tree structure is quite complex if we use all sequences. Therefore, one sequence from each location was randomly chosen and used when constructing trees. Note that the numbers of sequences in the locations are different and some sequences in the location have duplicates, however, sequences from different locations are different. After random selection of one sequence from each location, we have a sample which contains 17 different sequences. Base on the sample, we use MEGA4 to construct the NJ and MP trees which are presented in Figs. 7.1 and 7.2, respectively. Also, we use PHYLIP to construct the ML tree presented in Fig. 7.3. We then deleted all non-binary sites in all sequences then construct the mixture tree. The mixture method uses the frequency of a sequence in the population to assign a weight; here the weights are ones. The mixture tree is constructed and presented in Fig. 7.4.

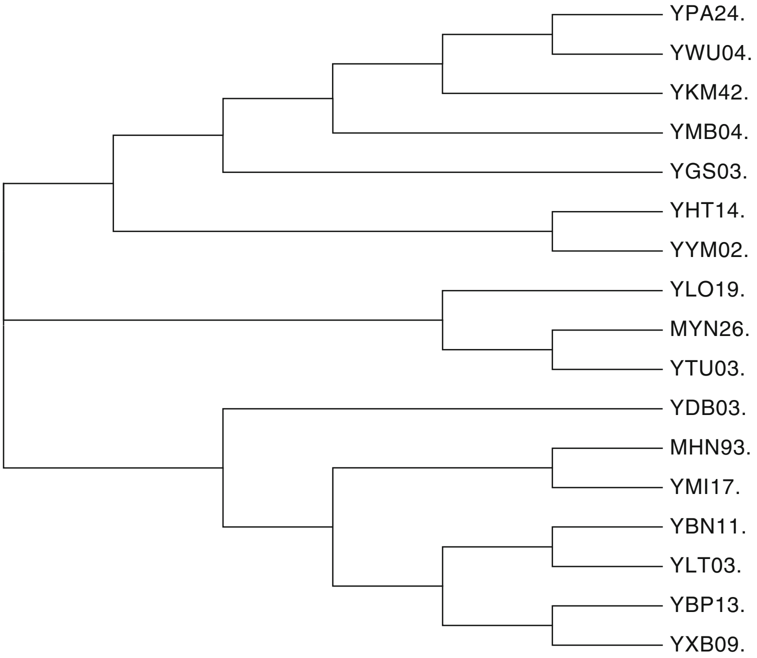

Fig. 7.1 The NJ tree for one sample of the data in Wen et al. [20]

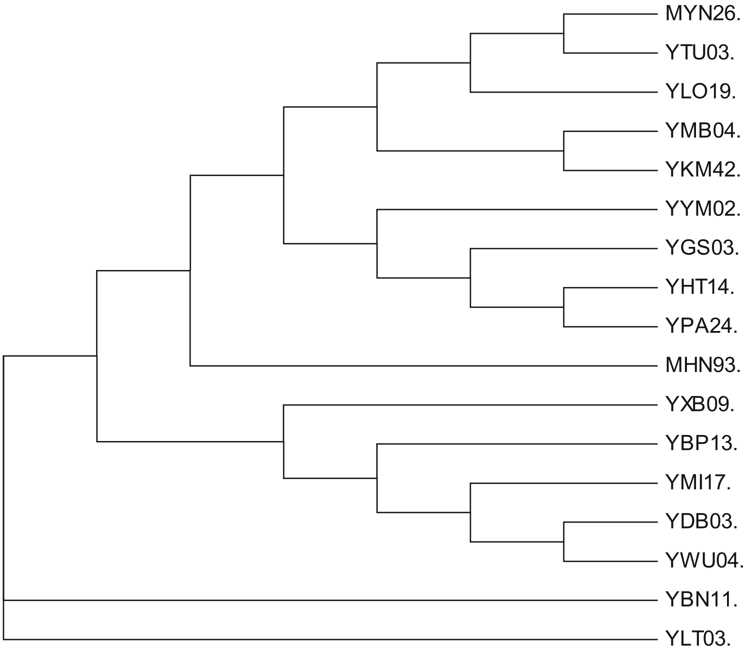

Fig. 7.2 The MP tree for one sample of the data in Wen et al. [20]

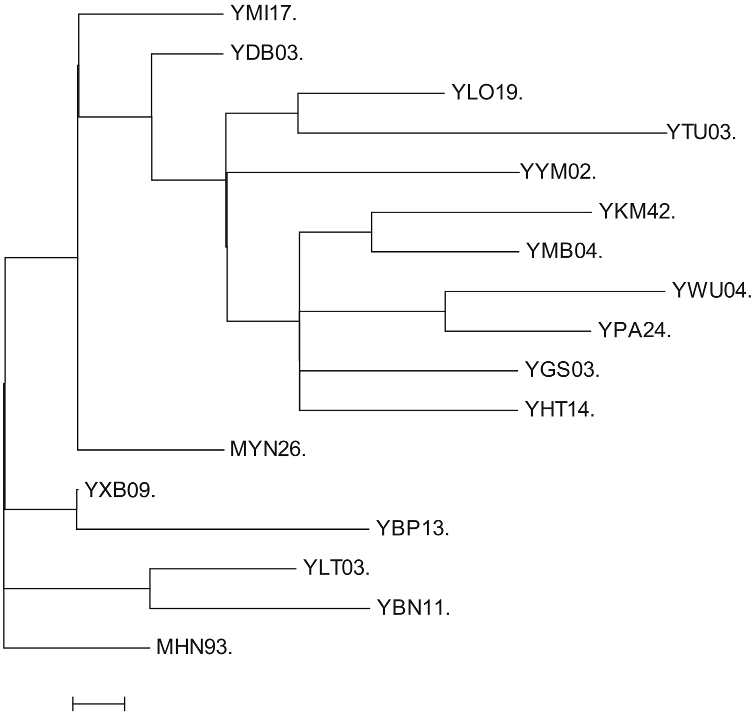

Fig. 7.3 The ML tree for one sample of the data in Wen et al. [20]

7.2.2.2 Trees Based on the Sample Contains all Sequences in the Population

We can also construct trees based on the full set of manipulated sequences in the population by using NJ, MP, ML, and the MT method. The NJ and MP trees can be constructed in MEGA4 and the ML tree can be constructed in PHYLIP. The resulting mixture tree is presented in Fig. 7.4.

7.3 Comparison

7.3.1 Simulator

The simulator we used in comparison of different tree reconstructed methods is ms [9], which is a program to generate samples under a variety of neutral models. A variety of assumptions about migration, recombination rate and population size can be set to generate the designated samples. The samples are generated using the standard coalescent approach in which the random genealogy of the sample is first

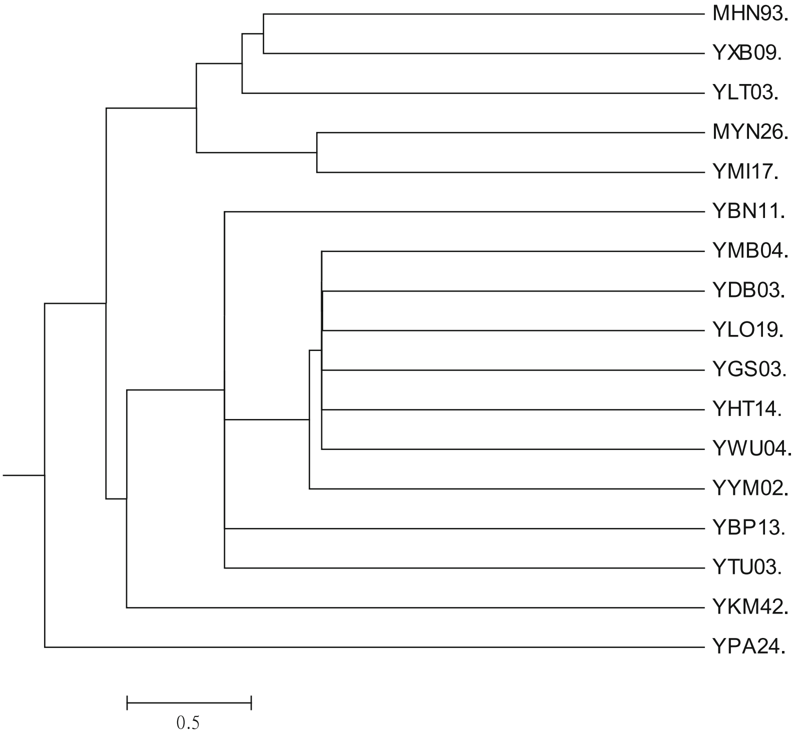

Fig. 7.4 The Mixture Tree for one sample of the data in Wen et al. [20]

generated and then mutations are randomly placed on the genealogy. The simulator can be run under the Unix-Like operating system like Linux.

The basic command line is:

ms nsam nreps -t θ

where

- **nsam** the number of copies of the locus in each sample;
- **nreps** the number of independent samples to generate.
- $\boldsymbol{\theta}$ the mutation parameter, $\theta = (4N_0\mu)$, where N_0 is the diploid population size and where μ is the neutral mutation rate for the entire locus.
- **-t** $\boldsymbol{\theta}$ set value of $4N_0\mu$.

In order to output the gene trees, the option $-T$ needs to be added in basic command. Also, **-s** j needs to be added, if one wants to make samples with fixed number of segregating sites, j.

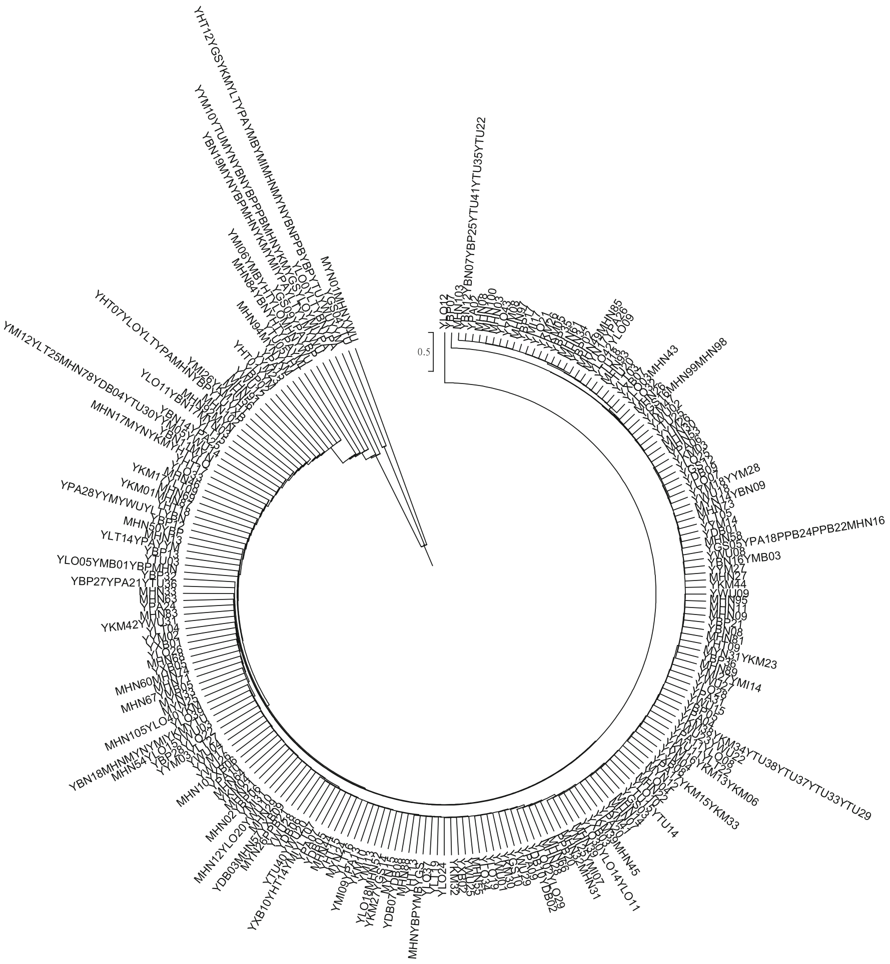

Fig. 7.5 The Mixture Tree for all samples of the data

7.3.2 Comparison

For a set of parameters($\boldsymbol{\theta}$, $\mathbf{s}$), we simulate a sample of size 200 with no identical sequences in each observation and no tie in the corresponding gene tree. Once we have the simulated distinct sequences (suppose it is saved in *tree1.fas*) and no tie in the gene tree (suppose it is saved in *tree1.nwk*), we do the following steps to complete the comparison:

- Change the format of *tree1.fas* to the format which can be used in the mixture tree algorithm and save it as *tree1.txt*;
- Run the mixture tree algorithm with sliding-scale 0.001 using *tree1.txt* and obtain the mixture tree *tree1mm.nwk*;

- Substitute A for 0, G for 1 in *tree1.fas* and reconstruct the Neighbor-joining (*tree1NJ.nwk*) and Maximum Parsimony tree(*tree1MP.nwk*) using MEGA4;
- Using the function *unroot*, *read.tree* and *dist.topo* in the package **ape** in R to compare the distance between *tree1.nwk* and *tree1mm.nwk*, *tree1NJ.nwk*, *tree1MP.nwk*, respectively. Record them.

If there is a tie in the mixture tree, Neighbor-joining tree, and Maximum Parsimony tree during any steps above, we will discard the whole set of sequences.

In order to determine the extent of topological differences between the gene tree(*tree1.nwk*) and the trees created using the other methods (NJ, MP, and MT), Rzhestky and Nei [16] method is implemented. This method is based on the Penny and Hendy's [14] method of sequence partitioning, which provides equivalent numerical values to those obtained using the Robinson and Foulds' [15] method but is simpler to compute. For unrooted bifurcating trees, this distance is twice the number of interior branches at which sequence partitioning is different between the two trees compared. The topological distance can be thought of as the smallest number of transformations required to obtain the simulated tree topology from the tree constructed using the mixture algorithm. The Rzhestky and Nei method is a modification of this distance to take multichotomies into account. These values were standardized by dividing by twice the total number of internal branches. An unrooted bifurcating tree with n haplotypes has $n-3$ interior branches. Thus, the maximum possible value is $2(n-3)$. The topological distances were measured and standardized.

7.3.3 Summary of the Analysis

The maximum distance between two trees, given the number of lineage n, using Rzhestky and Nei [16] method, the maximum distance between two trees is $2(n-3)$. So it is reasonable to standardize the distances by dividing each distance by the maximum distance. With different number of different SNPs sequences, the maximum distance between two trees would vary under the Rzhestky and Nei method. The results of the analysis are summarized in Tables 7.1, 7.2, and 7.3.

In the summary Tables 7.1, 7.2, and 7.3, we will call the mixture trees reconstructed via the FixEM algorithm the 'FixMixture', the mixture trees reconstructed via the traditional EM algorithm the 'Mixture'. The 'NJ' means the trees are

Table 7.1 Comparison Results for simulated data with mutation rate 0.0000025 and sample size 200

Mutation rate 0.0000025	Length: 20	No. of Sequences: 5	Samplesize: 200	
	FixMixture	Mixture	NJ	MP
Sum of distance	142	136	156	124
Sum of std. distance	35.5	34	39	31

Table 7.2 Comparison Results for simulated data with mutation rate 0.00000375 and sample size 200

Mutation rate 0.00000375	Length: 10	No. of Sequences: 5	Samplesize: 200	
	FixMixture	Mixture	NJ	MP
Sum of distance	188	168	208	194
Sum of std. distance	47	42	52	48.5

Table 7.3 Comparison Results for simulated data with mutation rate 0.000005 and sample size 200

Mutation rate 0.000005	Length: 10	No. of Sequences: 5	Samplesize: 200	
	FixMixture	Mixture	NJ	MP
Sum of distance	198	146	192	184
Sum of std. distance	49.5	36.5	48	46

reconstructed by the ‘Neighbor-Joining’ algorithm. The ‘MP’ means the trees are reconstructed by the ‘Maximum Parsimony’ algorithm. The ‘Sum of Distance’ is the sum of the Rzhestky and Nei distance of all the units in the sample between mixture tree or Neighbor-joining tree or Maximum Parsimony tree and true gene tree, respectively. The ‘Sum of Std. Distance’ is the sum of Rzhestky and Nei distance of all units in the sample between three different kind trees and gene tree divided by the maximum distance of that unit in the sample, respectively. The ‘Length’ is the length of the simulated sequences in the sample. ‘No. of Sequences’ is the number of sequences in one sample. Please note again that the ‘sum of (Std.)distances’ are the sum of distance between the tree reconstructed by one of these three algorithms and the true gene tree of each unit in the sample. It is obvious that the smaller the distance between two types of trees, the more similar they are. So we can see that the ‘Mixture’ algorithm performed better than at least one algorithm among other tree algorithms in these tables. Sometimes ‘Fix Mixture’ algorithm performed equally better than ‘Mixture’ algorithm, sometimes not. Also, we can see that other two methods are more stable than ‘Mixture’ algorithm and ‘FixMixture’ algorithm, and it is probably due to the fact that ‘Mixture’ and ‘FixMixture’ algorithms embed a more complicated statistical model and take the frequency of each sequence into account when it constructs the tree.

7.4 Discussion

In this chapter we have given an overview of a new method for tree reconstruction called the mixture tree. It provides an estimator of population structure at each point in the past based on a mutational clock. The estimators, unlike some competing methods, are unique. Linking these estimators together over time provides a tree that describes how the population might have evolved. Such a tree can also be used to infer the likely coalescence of lineages, although indirectly. In this chapter we demonstrated how the output of this analysis creates a tree very similar to established methods in the phylogeny literature, and how it can provide a method that is competitive with, but not superior to those competitive methods. In fact, we believe the greater strength of the method lies not in tree construction for distinct phylogenies, but because it provides a clustering method, as well as density estimator, for studies of population structure based on samples from a single population. The theorems are developed in the paper of Lindsay et al. [11] and will be further investigated in the future. Moreover, the current algorithm is based on Bernoulli mixture, which only consider binary sequences. In the future, we will extend it to handle sequences with multiple category and different mutation rates for different types.

Acknowledgements We thank National Science Foundation for the funding – Division of Mathematical Sciences 0714949 to Shu-Chuan Chen and Michael Rosenberg, and National Science Foundation – Division of Mathematical Sciences 0714839 to Bruce Lindsay. In addition, we also acknowledge the support of National Center for Theoretical Sciences (South), Taiwan.

References

1. Chen, S. C., & Lindsay, B. (2006). Building mixture trees from binary sequence data. *Biometrika*, *93*(4), 843–860.
2. Czelsniak, J., Goodman, M., Moncrief, N. D., & Kehoe, S. M. (1990). Maximum parsimony approach to construction of evolutionary trees from aligned homologous sequences. *Methods in Enzymology*, *183*, 601–615.
3. Edwards, A. W. F., & Cavalli-Sforza, L. L. (1963). The reconstruction of evolution. *Annals of Human Genetics*, *27*, 105–106. (also published in Heredity 18:553)
4. Edwards, A. W. F., & Cavalli-Sforza, L. L. (1964). Reconstruction of evolutionary trees. In V. H. Heywood & J. McNeill (Ed.), *Phenetic and phylogenetic classification* (Vol. 6, pp. 67–76). London: Systematics Association Publ.
5. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum Liklihood Approach. *Journal of Molecular Evolution*, *17*, 368–376.
6. Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, *41*, 155–160.
7. Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32.
8. Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London A*, *222*, 309–368.
9. Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*, 337–338.
10. Huelsenbeck, J. P., & Ronquist, F. (2005). Bayesian analysis of molecular evolution using MrBayes. In Nielsen, R. (Ed.), *Statistical methods in molecular evolution*. New York: Springer.
11. Lindsay, B., Markatou, M., Ray, S., Kang, K., & Chen, S. C. (2008). Quadratic distances on probabilities: A unified foundation. *The Annals of Statistics*, *36*(2), 983–1006.
12. Nei, M., & Kumar, S. (2000). *Molecular evolution and phylogenetics*. New York: Oxford University Press.
13. Nei, M., Kumar, S., & Takahashi, K. (1998). The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. *Proceedings of the National Academy of Sciences of the United States of America*, *95*, 12390–12397.
14. Penny, D., & Hendy, M. D. (1985). The use of tree comparison metrics. *Systematic Zoology*, *34*, 75–82.

15. Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*, 131–147.
16. Rzhestky, A., & Nei, M. (1992). A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution*, *9*, 945–967.
17. Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular and Biological Evolution*, *4*, 406–425.
18. Takahashi, K., & Nei, M. (2000). Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Molecular Biology and Evolution*, *17*, 1251–1258.
19. Tamura, K., Dudley, J., Nei, M., & Kumar, S. (2007). MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, *24*, 1596–1599.
20. Wen, B., Li, H., Gao, S., et al. (2004). Genetic structure of Hmong-Mien speaking populations in east Asia as revealed by mtDNA lineages. *Molecular and Biological Evolution*, *22*(3), 725–734.