Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 17

Model Selection for Mixtures of Mutagenetic Trees

Junming Yin, *Department of EECS, University of California, Berkeley*
Niko Beerenwinkel, *Department of Mathematics, University of California, Berkeley*
Jörg Rahnenführer, *Max-Planck-Institute for Informatics, Saarbrücken, Germany*
Thomas Lengauer, *Max-Planck-Institute for Informatics, Saarbrücken, Germany*

Recommended Citation:
Yin, Junming; Beerenwinkel, Niko; Rahnenführer, Jörg; and Lengauer, Thomas (2006) "Model Selection for Mixtures of Mutagenetic Trees," *Statistical Applications in Genetics and Molecular Biology*: Vol. 5: Iss. 1, Article 17.
DOI: 10.2202/1544-6115.1164

Model Selection for Mixtures of Mutagenetic Trees

Junming Yin, Niko Beerenwinkel, Jörg Rahnenführer, and Thomas Lengauer

Abstract

The evolution of drug resistance in HIV is characterized by the accumulation of resistance-associated mutations in the HIV genome. Mutagenetic trees, a family of restricted Bayesian tree models, have been applied to infer the order and rate of occurrence of these mutations. Understanding and predicting this evolutionary process is an important prerequisite for the rational design of antiretroviral therapies. In practice, mixtures models of K mutagenetic trees provide more flexibility and are often more appropriate for modelling observed mutational patterns.

Here, we investigate the model selection problem for K-mutagenetic trees mixture models. We evaluate several classical model selection criteria including cross-validation, the Bayesian Information Criterion (BIC), and the Akaike Information Criterion. We also use the empirical Bayes method by constructing a prior probability distribution for the parameters of a mutagenetic trees mixture model and deriving the posterior probability of the model. In addition to the model dimension, we consider the redundancy of a mixture model, which is measured by comparing the topologies of trees within a mixture model. Based on the redundancy, we propose a new model selection criterion, which is a modification of the BIC.

Experimental results on simulated and on real HIV data show that the classical criteria tend to select models with far too many tree components. Only cross-validation and the modified BIC recover the correct number of trees and the tree topologies most of the time. At the same optimal performance, the runtime of the new BIC modification is about one order of magnitude lower. Thus, this model selection criterion can also be used for large data sets for which cross-validation becomes computationally infeasible.

KEYWORDS: model selection, mixtures of mutagenetic trees, BIC, empirical bayes

Author Notes: Part of this work has been performed in the context of the BioSapiens Network of Excellence (EU contract no. LSHG-CT-2003-503265). Financial support was provided by an IMPRS scholarship (J.Y.), by Deutsche Forschungsgemeinschaft under grant No. HO 1582/1-3 and BE 3217/1-1 (N.B.) and by BMBF grant No. 01GR0453 (J.R.). We would like to thank Michael I. Jordan and Simon Lacoste-Julien for helpful discussion.

1 Introduction

Drug resistance presents a major obstacle to successful treatment of HIV infected patients. The emergence of resistant escape mutants in the virus population results in therapy failure and limits future treatment options. Understanding this evolutionary process is important for the design of effective antiviral treatment strategies. The development of drug resistance is characterized by the ordered accumulation of amino acid changes in the viral genome (Boucher et al., 1992; Molla et al., 1996). We have previously developed a probabilistic graphical model, namely the mutagenetic trees mixture model (Beerenwinkel et al., 2005a), to describe this accumulative evolutionary process.

The basic building block of this model is a weighted directed tree, which was introduced by Desper et al. (1999) in the context of oncogenesis. Vertices of the tree represent binary random variables that indicate the occurrence of genetic events (mutations). Each edge is weighted with the conditional probability of the child event given that the parent event has occurred. Thus, the model aims at identifying directed dependencies between mutational events.

Mixture models of mutagenetic trees can identify multiple evolutionary pathways acting on the same genetic events (Beerenwinkel et al., 2005a). Each such process is represented in one specialized component of the mixture model. A K-mutagenetic trees mixture model consisting of K tree components can be learned from cross-sectional data by an Expectation Maximization (EM) algorithm.

However, the number K of tree components is usually unknown. If K is chosen too small, it is impossible to detect all mutational pathways present in the data; if K is chosen too large, overfitting or redundancy results. In this paper, we develop and evaluate methods that address this model selection problem for mixture models of mutagenetic trees.

Besides the ability to generalize well, we are particularly interested in the interpretability of the estimated model. Indeed, the mutational pathways represented in the trees might indicate functional constraints of the molecule and even suggest the choice of specific therapies. Thus, among different models with similar predictive power on unseen data, we prefer the most concise model. In general, this means that models with few tree components are preferred over models with many components if they have similar estimated extra-sample performance.

A standard method for model selection is based on cross-validation and on picking the least complex model within one standard error of the best performing model (Hastie et al., 2001). However, this approach is too computationally expensive for most real world applications. Here, we pursue a more practical approach and consider variations of Bayesian model selection. We investigate the empirical Bayes method, the Bayesian Information Criterion (BIC), and the Akaike Information Criterion (AIC) for mutagenetic tree mixture models. In particular, we show how to approximate the marginal likelihood and compute the effective number of parameters of this model.

We also present a new modified version of the BIC that extends the standard BIC heuristically by penalizing redundancy in the mixture model. Experiments on simulated and real world HIV data show that empirical Bayes, standard BIC, and AIC tend to select models that contain repetitive and redundant tree structures. Both cross-validation and the modified BIC perform significantly better than those methods, and this result holds for three different measures of the performance of recovering the true model structure. Thus, unlike standard BIC the modified BIC provides a competitive and much faster alternative to cross-validation with sizeable real world data sets. In addition, the modified BIC yields models with increased interpretability, since the selected model tends to avoid repetitive structure among different tree components, even if the repetition would not penalize generalizability of the model.

The program code for the modified BIC and all other tested criteria has been added to the `Mtreemix` software package for statistical inference with mutagenetic tree models (Beerenwinkel et al., 2005b).

2 Mutagenetic Trees Mixture Models

In this section, we recall the definition and basic properties of mixtures of mutagenetic trees. We start with the single tree model.

2.1 Mutagenetic Trees

A *mutagenetic tree* $T = (V, E)$ for ℓ genetic events $\{1, \ldots, \ell\}$ is a connected branching on the vertices $V = \{0, 1, \ldots, \ell\}$ rooted at 0. The set $E \subseteq V \times V$ denotes the edges of T. For each vertex $v \in V$, we denote by $\mathrm{pa}(v)$ its parent and by $\mathrm{ch}(v)$ its children in T. The vertices of T correspond to binary random variables $X_0, X_1, \ldots, X_\ell$. The event $\{X_v = 1\}$ indicates the occurrence of mutation v. The variable X_0 serves only to simplify notation, and we set $\Pr(X_0 = 1) = 1$.

Let $p \in [0, 1]^\ell$ be a parameter vector. We write p_v $(v \in V \setminus \{0\})$ for the

DOI: 10.2202/1544-6115.1164

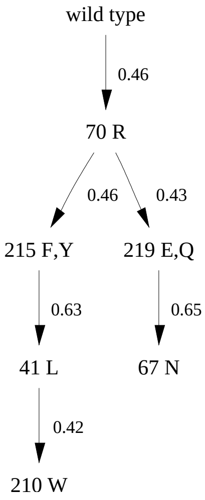


Figure 1: Mutagenetic tree for the development of zidovudine resistance. Nodes are labeled with resistance-associated mutations in the HIV-1 reverse transcriptase (sequence position followed by substituted amino acids), edge labels represent conditional probabilities between mutational events.

coordinates of p, and regard p_v as the weight of the edge $(\mathrm{pa}(v), v) \in E$. The *mutagenetic tree model* $\mathcal{T}$ induced by T is defined as the Bayesian tree model (Jordan, 2004) on (V, E) with transition matrices

$$\big(P(X_v = b \mid X_{pa(v)} = a)\big)_{a,b=0,1} = \begin{pmatrix} 1 & 0 \\ 1-p_v & p_v \end{pmatrix}, \quad v = 1, \dots \ell.$$

It follows that p_v is the conditional probability of event v given that its parent event $\mathrm{pa}(v)$ has occurred. Furthermore, the first row of the transition matrix implies that an event can only occur if its predecessor in the tree has occurred. Figure 1 shows a mutagenetic tree for the development of resistance to zidovudine, the first approved antiretroviral drug.

The mutagenetic tree model $\mathcal{T} = (T, p)$ defines the following factorization of the joint probability $\Pr(X_1, \dots, X_\ell)$ of mutational patterns $x \in \{0,1\}^\ell$. Let $V[x] = \{v \in V \mid x_v = 1\}$ be the set of occurred events specified by the pattern x. If there is a subset $E' \subseteq E$ such that $V[x]$ is the set of all vertices reachable from 0 in the induced subtree $T' = (V[E'], E')$, then

$$\Pr(X = x \mid \mathcal{T}) = f_x^{(T)}(p) := \prod_{v \in V[x]} p_v \prod_{\substack{\mathrm{pa}(v) \in V[x] \\ v \notin V[x]}} (1 - p_v). \tag{1}$$

If there is no such edge subset, the topology of T does not allow for generating x and hence $\Pr(x \mid \mathcal{T}) = f_x^{(T)}(p) := 0$. We call the mutational patterns $x \in \{0,1\}^\ell$ with non-zero probability the *compatible states* of $\mathcal{T}$. Equation 1 can be computed efficiently by performing a breadth-first search in T.

The structure and the parameters of a mutagenetic tree can be learned efficiently from cross-sectional data by solving the maximum weight branching problem in the complete graph on $\ell + 1$ vertices (Desper et al., 1999).

Since trees can only represent a limited set of acyclic dependency structures, this model class is too small for most applications. For example, the tree displayed in Figure 1 does not capture all of the known pathways to zidovudine resistance. Therefore, we consider the broader class of mixtures of mutagenetic trees.

2.2 Mixture Models

Consider K mutagenetic trees $M_K = (T_1, \dots T_K)$ for the same genetic events $\{1, \dots, \ell\}$. The *K-mutagenetic trees mixture model* $\mathcal{M}_K = (\mathcal{T}_1, \dots, \mathcal{T}_K)$ is defined as the familiy of distributions of $X = (X_1, \dots, X_\ell)$ of the form

$$\Pr(X = x \mid \mathcal{M}_K) = f_x^{(M_K)}(\lambda, p) := \lambda_1 \, f_x^{(T_1)}(p_1) + \cdots + \lambda_K \, f_x^{(T_K)}(p_K). \qquad (2)$$

The mixture model has parameters $\theta := (\lambda, p)$ comprising the tree parameters $p = (p_k)_{k=1,\dots,K} = (p_{k,v})_{k=1,\dots,K,\, v=1,\dots,\ell}$ and the mixing parameters $\lambda = (\lambda_1, \dots, \lambda_K)$ such that $\sum_{k=1}^{K} \lambda_k = 1$. In this model, in general, not all possible mutational patterns have positive likelihood. Thus, when applied to real data, we typically fix the first tree component $\mathcal{T}_1$ to be a star with uniform edge weights (Figure 2). This *noise component* models events as being independent of each other and as occurring with the same probability. If $p_{1,v} \neq 0$ for all v, then all 2^ℓ possible mutational patterns have non-zero probability in the model. If not stated otherwise, a general mixture model refers to the unrestricted model without enforced star topology or uniform weights.

The mixture model $\mathcal{M}_K$ can be regarded as a Bayesian network with a hidden K-ary random variable Z that is connected by an edge to all roots of the tree components. The value of Z determines which mutagenetic tree to use for generating a particular pattern.

The mixture model can be learned by an EM-like algorithm (Beerenwinkel et al., 2005b). In the E step, the current parameters and structure are used to compute the responsibilities of the different tree components for the data. In the M step, the structure and parameters of the tree models are re-estimated from the weighted data. These two steps are iterated until the increase in

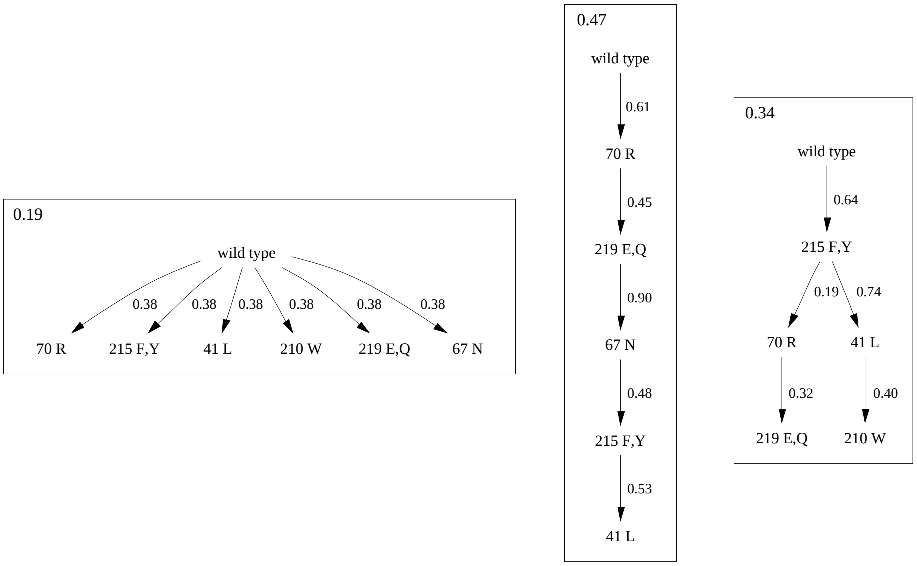


Figure 2: 3-Mutagenetic trees mixture model for the development of zidovudine resistance. Each tree component is displayed within a box, and its weight appears in the upper left corner of the box.

likelihood is negligible. In addition to the training patterns, this algorithm requires as input the number K of tree components.

3 Model Selection Criteria

The determination of the number of tree components in the mixture model is a model selection problem. Model selection aims at identifying models that provide an accurate fit to the data, generalize well, and are no more complex than necessary for explaining the data. The problem can be viewed as an optimization problem involving two basic components (Ghahramani, 2004): (1) a strategy for searching through the family of possible models efficiently, and (2) a measure (criterion) for scoring the models.

In the context of mixtures of mutagenetic trees, the model space is divided by the number of tree components K. To each K corresponds the K-fold product of the space of all trees on the vertices $\{0, 1, \ldots, \ell\}$ rooted at 0. Since the number of trees grows doubly exponentially in ℓ, we will avoid searching tree space for efficiency in our model selection criterion that will be introduced in the following Section 4. There, we restrict ourselves to considering one model for each K.

Addressing the second issue of a model selection criterion, a simple and widely used approach is cross-validation in conjunction with the one-standard-error rule (Hastie et al., 2001). The score implicitly used by cross-validation is the estimated extra-sample performance. The motivation for this criterion lies in the attempt to fit a model, which not only accounts for the training data, but also generalizes well to unseen data. In order to obtain an accurate and approximately unbiased estimate of the generalization error, the number of partitions used in cross-validation needs to be large enough and this can make the procedure very time-consuming.

3.1 Bayesian Model Selection

The Bayesian method provides an alternative approach to model selection. The centerpiece of Bayesian model selection is the marginal likelihood, obtained by marginalizing out the parameters of the model. Let the given data be denoted by $\mathcal{D} = \{x^{(1)}, \ldots, x^{(N)}\}$. Regarding the model space as indexed by the number of tree components K, the marginal likelihood of a mixture model with K tree components can be written as

$$\Pr(\mathcal{D} \mid K) = \iint \Pr(\mathcal{D} \mid M_K, \theta) \Pr(\theta \mid M_K) \Pr(M_K)\, d\theta\, dM_K, \tag{3}$$

where $\Pr(M_K)$ denotes the prior over the K tree topologies and $\Pr(\theta \mid M_K)$ is the conditional prior over tree parameters given tree topologies. We also write $\Pr(\mathcal{D} \mid M_K) = \int \Pr(\mathcal{D} \mid M_K, \theta) \Pr(\theta \mid M_K)\, d\theta$, and hence $\Pr(\mathcal{D} \mid K) = \int \Pr(\mathcal{D} \mid M_K) \Pr(M_K)\, dM_K$.

For given tree topologies M_K, if we place a conjugate Dirichlet prior on the parameters θ, we obtain closed-form expressions for $\Pr(\mathcal{D} \mid M_K)$. Therefore, the marginal likelihood may be approximated by a simple Monte Carlo approach that samples n tree topologies $M_K^{(1)}, \ldots, M_K^{(n)}$ from $\Pr(M_K)$ to compute

$$\Pr(\mathcal{D} \mid K) \approx \frac{1}{n} \sum_{i=1}^{n} \Pr(\mathcal{D} \mid M_K^{(i)}). \tag{4}$$

Although this approximation is asymptotically correct, the rate of convergence will be so slow for most applications that it is of little practical value. The problem in our case is that in the absence of prior knowledge, an uninformative prior $\Pr(M_K)$ is chosen, placing little probability mass on tree topologies that are likely to generate the data $\mathcal{D}$. Hence most contributions to the sum in (4) will be very small and convergence tends to be slow.

DOI: 10.2202/1544-6115.1164

3.2 Empirical Bayes

Let us consider the empirical Bayes method for model selection (Robert, 2001). We regard the tree topologies M_K as hyperparameters and use the data $\mathcal{D}$ to obtain an estimate $\hat{M}_K$ by the learning algorithm described in Beerenwinkel et al. (2005a). The trees $\hat{M}_K$ may be interpreted as our prior belief regarding the topology of evolutionary pathways that lead to drug resistance. Thus, the marginal likelihood $\Pr(\mathcal{D} \mid K)$ in (3) is approximated by

$$\Pr(\mathcal{D} \mid \hat{M}_K) = \int \Pr(\mathcal{D} \mid \hat{M}_K, \theta) \Pr(\theta \mid \hat{M}_K)\, d\theta, \tag{5}$$

which is equivalent to putting all the probability mass on $\hat{M}_K$ in the prior $\Pr(M_K)$. Given $\hat{M}_K$, we now show how to place a conditional conjugate prior over θ and compute $\Pr(\mathcal{D} \mid \hat{M}_K)$ (Jordan, 2006).

In the absence of prior knowledge, it is common to resort to an uninformive prior; see (Jordan, 2006) for details. For the mode of this prior distribution, we define parameters $\tilde{\theta} = (\tilde{\lambda}, \tilde{p})$ such that all configurations of the random vector $(Z, X_1, \ldots, X_\ell)$ are equally probable. This is achieved by setting $\tilde{\lambda}$ and $\tilde{p}$ separately.

For a single tree $\hat{T}_k$, we set $\tilde{p}_{k,v} := (C_{k,v} - 1)/C_{k,v}$, where $C_{k,v}$ is the number of substates compatible with the subtree of $\hat{T}_k$ rooted at v. It is shown in the Appendix how to compute $C_{k,v}$, and a formal proof of the uniformity of this distribution is given. Under $\tilde{p}_{k,v}$ the probability of each compatible mutational pattern is $1/C_{k,0}$, where $C_{k,0}$ is the number of compatible states of $\hat{T}_k$. Figure 3 illustrates this definition with a tree T for five genetic events that has eleven compatible states. For example, $\Pr(10110 \mid T, \tilde{p}) = (10/11) \cdot (1 - 1/2) \cdot (4/5) \cdot (1/2) \cdot (1 - 1/2) = 1/11$.

For a mixture model, we set $\tilde{\lambda}_k = C_{k,0} / \sum_{k=1}^{K} C_{k,0}$. Together, the definitions of $\tilde{p}$ and $\tilde{\lambda}$ imply that the random vector $(Z, X_1, \ldots, X_\ell)$ is distributed uniformly with probability $\tilde{\lambda}_k \cdot (1/C_{k,0}) = 1/\sum_{k=1}^{K} C_{k,0}$.

Next, we define prior probability distributions for λ and p. Like for $\tilde{\lambda}$ and $\tilde{p}$, we define the distributions separately. Let $\tilde{N}$ be the number of "equivalent samples" that are assumed to give rise to the probability assessments of the mixture model $\hat{M}_K$. For each k, let $\pi_k = \tilde{N}\tilde{\lambda}_k$ be a hyperparameter, namely the number of equivalent samples underlying $\hat{T}_k$. Similarly, we define hyperparameters $\eta_{k,v} = \pi_k \tilde{p}_{k,v}$. Let $\pi = (\pi_1, \ldots, \pi_K)$ and $\eta = (\eta_{k,v})_{k=1,\ldots,K,\, v=1,\ldots,\ell}$. Placing a Dirichlet prior on λ we obtain

$$\Pr(\lambda \mid \hat{M}_K, \pi) = \frac{\Gamma(\tilde{N})}{\prod_k \Gamma(\pi_k)} \prod_{k=1}^{K} \lambda_k^{\pi_k - 1},$$

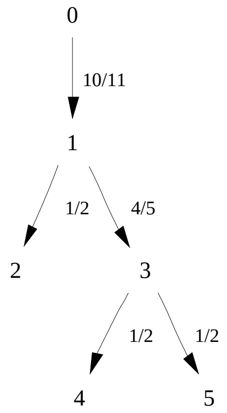


Figure 3: Mutagenetic tree with average parameters $\tilde{p}$ that induce the uniform distribution on all compatible patterns.

and defining the conjugate Beta distribution for $p_{k,v}$ yields

$$\Pr(p_{k,v} \mid \hat{M}_K, \pi, \eta) = \frac{\Gamma(\pi_k)}{\Gamma(\eta_{k,v})\Gamma(\pi_k - \eta_{k,v})}(1 - p_{k,v})^{\pi_k - \eta_{k,v} - 1} p_{k,v}^{\eta_{k,v} - 1}.$$

Furthermore, we assume that the random variables $p_{k,v}$ are independent for each value of k and v resulting in the following prior distribution for the parameters p:

$$\Pr(p \mid \hat{M}_K, \pi, \eta) = \prod_k \prod_v \frac{\Gamma(\pi_k)}{\Gamma(\eta_{k,v})\Gamma(\pi_k - \eta_{k,v})}(1 - p_{k,v})^{\pi_k - \eta_{k,v} - 1} p_{k,v}^{\eta_{k,v} - 1}.$$

We now turn to the computation of the empirical Bayes score $\Pr(\mathcal{D} \mid \hat{M}_K) = \prod_{x \in \mathcal{D}} \Pr(x \mid \hat{M}_K)$, for a given value K of the number of tree components. Using Equation 2 we find

$$\begin{aligned}
\Pr(x \mid \hat{M}_K) &= \iint \Pr(x \mid \hat{M}_K, p, \lambda) \Pr(p \mid \hat{M}_K, \pi, \eta) \Pr(\lambda \mid \hat{M}_K, \pi)\, dp\, d\lambda \\
&\overset{(2)}{=} \iint \sum_{k=1}^{K} \lambda_k \Pr(x \mid \hat{T}_k, p_k) \Pr(p \mid \hat{M}_K, \pi, \eta) \Pr(\lambda \mid \hat{M}_K, \pi)\, dp\, d\lambda \\
&= \sum_{k=1}^{K} \int \lambda_k \Pr(\lambda \mid \hat{M}_K, \pi) d\lambda_k \int \Pr(x \mid \hat{T}_k, p_k) \Pr(p_k \mid \hat{T}_k, \pi, \eta) dp_k.
\end{aligned}$$

The two integrals in the last term are identified as the expectations of λ_k and of $f_x^{(\hat{T}_k)}(p_k)$, respectively. Employing the Dirichlet and Beta distributions we obtain

$$\Pr(x \mid \hat{M}_K) = \sum_{k=1}^{K} \mathrm{E}[\lambda_k]\, \mathrm{E}[f_x^{(\hat{T}_k)}(p_k)] = \sum_{k=1}^{K} \tilde{\lambda}_k f_x^{(\hat{T}_k)}(\mathrm{E}[p_k]) = \sum_{k=1}^{K} \tilde{\lambda}_k f_x^{(\hat{T}_k)}(\tilde{p}_k).$$

Thus, $\Pr(x \mid \hat{M}_K) = f_x^{(\hat{M}_K)}(\tilde{\lambda}, \tilde{p})$ is computed by evaluating the mixture model at the "average parameters" $\tilde{\theta} = (\tilde{\lambda}, \tilde{p})$. In particular, this approximation of the marginal likelihood does not depend on the number $\tilde{N}$ of equivalent samples.

3.3 BIC and AIC

One common approximation to the marginal likelihood is the Bayesian Information Criterion (BIC) (Schwarz, 1978) This score consists of a likelihood term and a model complexity term reflecting the trade-off between goodness-of-fit and model complexity. This compromise can be regarded as an attempt to implement "Occam's Razor" (Jefferys and Berger, 1992), which states that the selected model should be no more complex than necessary for explaining the observed data.

For assessing model complexity, BIC uses the effective number of parameters, which is also known as the dimension of the model. Let d be the dimension of a K-mutagenetic trees mixture model $\mathcal{M}_K$. Then, the BIC score is defined as

$$\mathrm{BIC}(K) = \log \Pr(\mathcal{D} \mid \hat{\theta}, \hat{M}_K) - \frac{d}{2} \log N,$$

where $\hat{\theta}$ and $\hat{M}_K$ denote the maximum likelihood estimates of the parameters and the tree topologies of the model, respectively, and $\mathcal{D}$ denotes the data of size N. BIC is only asymptotically consistent, i.e., it will choose the true model if that is contained in the model space as $N \to \infty$. Thus in practice, BIC is often suboptimal for finite data sets.

A similar score, namely the Akaike Information Criterion (AIC) (Akaike, 1974) is defined as

$$\mathrm{AIC}(K) = \log \Pr(\mathcal{D} \mid \hat{\theta}, \hat{M}_K) - d.$$

Although both criteria reflect the trade-off described above, they have different motivations and rely on different assumptions (Hastie et al., 2001). We now turn to computing the model dimension d.

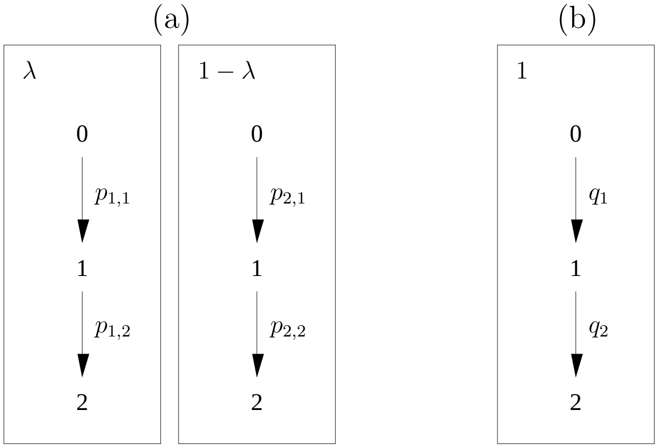


Figure 4: Two mixture models defining the same family of distributions: (a) mixture model of two trees that are identical in topology; (b) mixture model consisting of a single copy of the same tree.

4 Model Dimension and Redundancy

The single tree model ($K = 1$) has ℓ parameters and this is also the dimension of the model. If the edge weights are restricted to be identical, as in the case of the noise component, the dimension drops to one. The situation can be more deceiving for a K-mutagenetic trees mixture model $\mathcal{M}_K$. This model has $K + K \cdot \ell$ parameters $\theta = (\lambda, p)$ with the relation $\sum_{k=1}^{K} \lambda_k = 1$. However, the dimension of $\mathcal{M}_K$ can be much smaller than $K(\ell + 1) - 1$. For example, the mixture model in Figure 4(a) can be seen to define the same family of probability distributions as the single tree model displayed in Figure 4(b). Indeed, for a 2-paths mixture model with parameters $(\lambda, p_{1,1}, p_{1,2}, p_{2,1}, p_{2,2})$, the single path model with parameters (q_1, q_2) defines the same distribution if

$$q_1 = \lambda p_{1,1} + (1 - \lambda) p_{2,1} \quad \text{and} \quad q_2 = \frac{\lambda p_{1,1} p_{1,2} + (1 - \lambda) p_{2,1} p_{2,2}}{\lambda p_{1,1} + (1 - \lambda) p_{2,1}}.$$

Returning to the general case of a mixture model of mutagenetic trees $M_K = (T_1, \ldots, T_K)$, recall that this model can be regarded as a Bayesian network with the hidden variable Z that determines the choice of tree component. Consider the mapping from the space of parameters to the probability space of the model,

$$\mathbf{f}^{(M_K)} \colon \mathbb{R}^{K(\ell+1)-1} \longrightarrow \mathbb{R}^{2^\ell - 1}, \quad \theta \mapsto \left(f_x^{(M_K)}(\theta) \right)_{x \in \{0,1\}^\ell}.$$

DOI: 10.2202/1544-6115.1164

From Equations 1 and 2 we see that **f** is a polynomial map. Hence, its image is an algebraic variety, whose dimension equals the dimension d of the statistical model. Techniques from computational algebraic geometry can be used to compute d. For example, the dimension of the model displayed in Figure 4 can be derived directly from the set of algebraic invariants of this model (Beerenwinkel and Drton, 2005). However, in general, with an increasing number of genetic events these methods become too computationally expensive.

An alternative approach to computing the dimension d is based on the Jacobian matrix of the polynomial map **f**. In the case of a linear mapping $g : \mathbb{R}^m \to \mathbb{R}^n$ the dimension of the image of g equals the rank of the matrix that represents g. In general, if g is any smooth mapping, it can be approximated locally by a linear map that is given by the Jacobian matrix (Spivak, 1979). For Bayesian networks with hidden variables the rank of the Jacobian is constant with probability one and equals d (Geiger et al., 1996).

The Jacobian matrix $J(\theta)$ of **f** is obtained by computing the partial derivatives

$$\begin{aligned}
\frac{\partial f_x^{(M_K)}(\lambda, p)}{\partial \lambda_k} &= f_x^{(T_k)}(p_k) - f_x^{(T_K)}(p_K), \\
\frac{\partial f_x^{(M_K)}(\lambda, p)}{\partial p_{k,v}} &= \begin{cases} \frac{\lambda_k f_x^{(T_k)}(p_k)}{p_{k,v}} & \text{if } v \in V_k[x] \\ -\frac{\lambda_k f_x^{(T_k)}(p_k)}{1-p_{k,v}} & \text{if } v \notin V_k[x] \text{ and } \operatorname{pa}(v) \in V_k[x] \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}$$

where $V_k = V[T_k]$ denotes the vertices of the k-th tree.

Since the rank of $J(\theta)$ equals d for almost all parameter values θ, we can choose random vectors $\theta^{(1)}, \ldots, \theta^{(m)}$ and compute the rank numerically. The dimension d is obtained as the maximum of the resulting ranks of the matrices $J(\theta^{(1)}), \ldots, J(\theta^{(m)})$. In our experiments, we use samples of size $m = 5$. The dimension d is bounded from above by the minimum of the number of model parameters and the dimension of the ambient probability simplex:

$$d \leq \min\{K(\ell + 1) - 1,\ 2^\ell - 1\}. \tag{6}$$

In Figures 5(a), (b), and (c), three mixture models are displayed that share one path (the first model component) and that differ in their second tree component. The models have dimensions 5, 7, and 9, respectively. Thus, the BIC and AIC terms that penalize model complexity are lowest for (a) and highest for (c). However, when compared to alternative models consisting of a single tree, model (a) appears the least reasonable, because it repeats

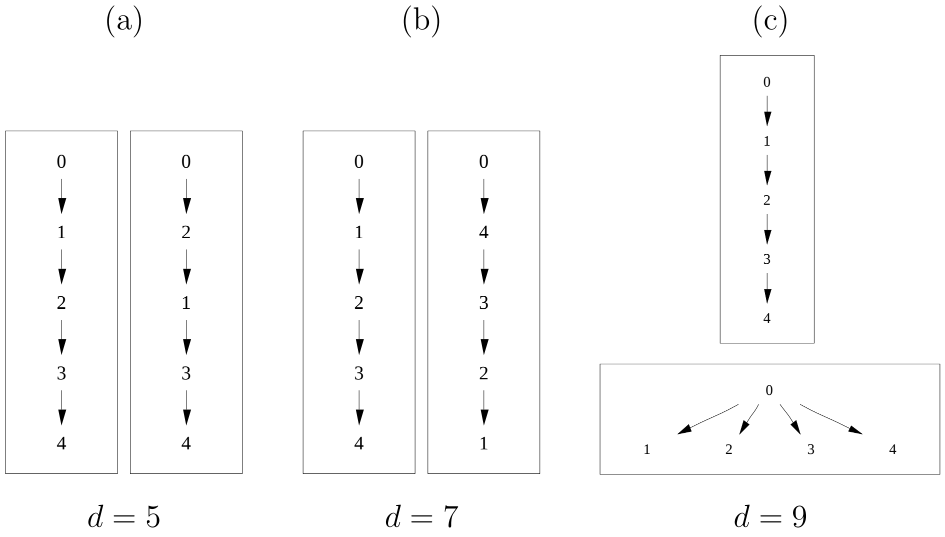


Figure 5: Three mixture models of size $K = 2$ for $\ell = 4$ mutations of dimensions $d = 5$, 7, and 9, respectively. The star in (c) is assumed to have non-uniform edge weights.

most of the structure of the first tree in its second tree. The model (c) has less redundant structure in the second tree, and intuitively it appears more difficult to replace this mixture by a single tree model.

In general, adding a similar tree component to a mixture model will result in only a small increase in dimension, but in a significant increase in redundancy. This effect is particularly pronounced for small values of ℓ and moderate to large values of K, because then model components are likely to share structural features and thus to define common model subvarieties. Figure 6 illustrates this phenomenon for $\ell = 4$ and $\ell = 6$ using 500 randomly generated mixture models with a noise component, for each $K = 2, \ldots, 6$ (see Section 5.2 for the exact simulation setup).

For $\ell = 4$ mutations, the dimension is bounded by $2^4 - 1 = 15$ according to (6). Overall, the dimension increases with the number of tree components K, but less so for larger values of K. In fact, many of the 4-mutagenetic trees mixture models already reach the upper bound of the dimension. For $\ell = 6$ mutations, the upper bound on the dimension of 63 is very unlikely to being attained. In our simulations, even dimension 36 is not exceeded, reflecting the larger size of tree space that reduces the chance of introducing a redundant tree component. In both cases, a large variety of redundancy values (to be defined formally in Section 4.1) is observed for all dimensions, suggesting that

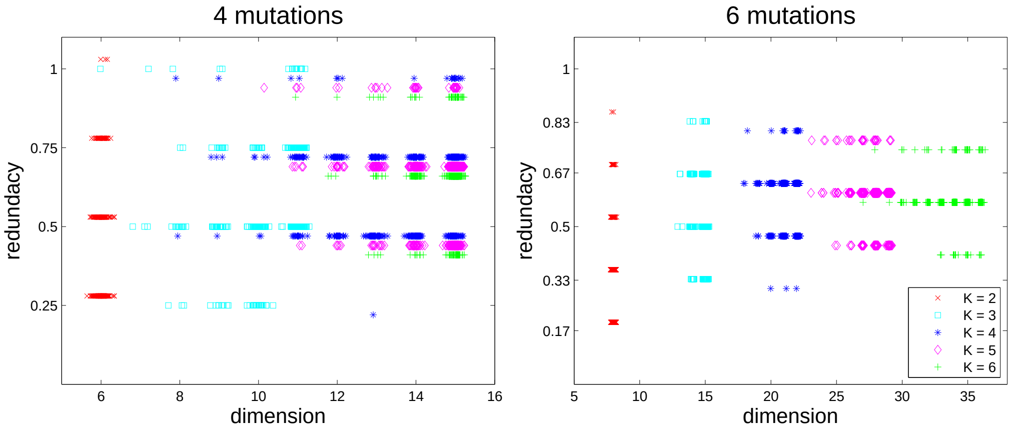


Figure 6: Relationship between dimension and redundacy. For $\ell = 4$ mutations (left panel) and $\ell = 6$ mutations (right panel) and $K = 2, \ldots, 6$, the dimension (x-saxis) and redundancy (y-axis) of 500 randomly generated models are displayed. Because of the discrete nature of both the dimension and the redundacy, a small amount of noise has been added to each dimension to visualize the density of points. In addition, points corresponding to different values of K have been untangled by small shifts of the redundancies that depend only on K. Hence, the differences in redundancy *within* each of the 4 (for $\ell = 4$) or 5 (for $\ell = 6$) discrete blocks are manufactured in order to fascilitate visual perception. All models are generated according to the procedure described in Section 5.2. In particular, due to the noise component the redundancy can only take the ℓ different values $R \in \{\frac{1}{\ell}, \frac{2}{\ell}, \ldots, \frac{\ell-1}{\ell}, 1\}$.

redundancy and dimension measure different features of the models.

Since redundant components do not decrease the log-likelihood of the data and increase the dimension only weakly, if at all, we expect BIC and AIC to fail in recovering the true model structure for moderate values of ℓ and K. In order to address this shortcoming we estimate the redundancy in a mixture model and penalize it.

4.1 Modified BIC

We define the similarity between two tree components T_k and T_l of a mixture model by

$$s_{kl} = s_{lk} = 1 - \frac{\|A_k - A_l\|_\infty}{\ell} \in [0, 1],$$

where A_k and A_l denote the adjacency matrices of T_k and T_l, respectively. The matrix norm $\|A\|_\infty = \max_i \sum_j |a_{ij}|$ is the maximum absolute row norm. Therefore, the term $\|A_i - A_j\|_\infty$ measures the maximum difference of outgoing edges between the two trees. We define the *redundancy* R of a mixture model as the maximum similarity among its tree components,

$$R = \max_{k \neq l} s_{kl}.$$

For large K and for models containing redundant structural parts, we want to incorporate the redundancy R into our model selection criterion. For a fixed K and data set $\mathcal{D}$, let $\hat{\mathcal{M}}_K$ be the K-mutagenetic trees mixture model estimated from $\mathcal{D}$ and let d be its dimension. Consider

$$\mathrm{BIC}_R(K) = \log \Pr(\mathcal{D} \mid \hat{\mathcal{M}}_K) - (1 + R) \cdot \frac{d}{2} \log N,$$

which doubles the penalty term for two identical tree components. We use the weighted average between standard BIC and BIC_R defined by

$$\mathrm{BIC}_w(K) = w \cdot \mathrm{BIC}(K) + (1 - w) \cdot \mathrm{BIC}_R(K),$$

with weight $w = \min(\frac{1}{\ell+1} \max(d_K - d_{K-1}, 0), 1)$.

The idea of this weighting is that a small increase in dimension with adding one tree component indicates repetitive model structure. In this case BIC_R is used to penalize redundancy. If the increase $d_K - d_{K-1}$ is large due to new model structure, standard BIC is applied. The min and max operators are used to bound w between 0 and 1, because in rare cases d or R may actually decrease with increasing K, because $\hat{\mathcal{M}}_K$ is re-estimated for each K.

5 Computational Experiments

In order to assess the performance of the different model selection criteria we define validation scores and conduct experiments on simulated and on real data. We compare cross-validation (XV), empirical Bayes (EB), AIC, standard BIC, and the modified BIC (BIC_w) introduced in the previous section.

5.1 Validation Scores

Three different scores are used to compare the structure of the true model $\mathcal{M}_K$ and the estimated model $\hat{\mathcal{M}}_{\hat{K}}$ based on different model selection criteria. Let

$S = (S_{kl})_{k=1,\ldots,K,\, l=1,\ldots,\hat{K}}$ be the similarity matrix of pairs of tree components of $\mathcal{M}_K$ and $\hat{\mathcal{M}}_{\hat{K}}$ defined by

$$S_{kl} = 1 - \frac{\|A_k - \hat{A}_l\|_\infty}{\ell},$$

where $\hat{A}_l$ is the adjacency matrix of the l-th tree in $\hat{\mathcal{M}}_{\hat{K}}$. We consider the following scores:

$$\begin{aligned} recov &= \frac{1}{K}\sum_k \max_l S_{kl}, \\ prec &= \frac{1}{\hat{K}}\sum_l \max_k S_{kl}, \\ dissim &= \sum_{(k,l)\in M} (1 - S_{kl}) + |K - \hat{K}|, \end{aligned}$$

where M denotes the maximum S_{ij}-weighted bipartite matching between the tree components of the respective models. The first score (*recov*) measures the success in recovering the true model structure, whereas the second score (*prec*) measures the preciseness in identifying the true model structure. Our main focus is on the third score (*dissim*), which provides a dissimilarity measure that combines both aspects and also accounts for unmatched trees if $\hat{K} \neq K$. In order to assess the significance of each score, we derive p-values from the empirical distribution of random scores as described in the following section.

5.2 Simulation Study

The simulation setup is similar to the experimental design in Van Allen and Greiner (2000). The following procedure was applied 500 times.

1. Draw a “true model” $\mathcal{M}_K$ at random:

   - Fix $K = 2$ or 3 and $\ell = 4$ or 6, and set $\lambda_1 = 0.1$ and $\lambda_k = 0.9/(K-1)$ for $k = 2, \ldots, K$.
   - Fix T_1 to be a star and let T_k $(k = 2, \ldots, K)$ be a random tree sampled uniformly from topology space. This is achieved by uniformly sampling $(\ell - 1)$-tuples of integers ranging between 0 and ℓ, which represent the Prüfer encoding of trees (Prüfer, 1918).

- Draw edge weights p_1 and $p_{k,v}$ ($k = 2, \ldots, K$, $v = 1, \ldots, \ell$) uniformly at random from the interval [0.2, 0.8].

2. Draw samples of size $N = 100$, 300, and 500 from $\mathcal{M}$.
3. Estimate $\hat{\mathcal{M}}_K$ from the sample for $K = 1, \ldots, 6$ and apply each model selection criterion (XV + one-standard-error rule, EB, AIC, BIC, and BIC$_w$) to obtain one estimate $\hat{K}$ for each criterion.
4. For each selected model, apply the validation scores for comparing it with the true model, and calculate the p-value of each score based on random scores obtained from 1000 randomly generated models. For the random models, we draw K from the empirical distribution obtained from 100 previous simulations of steps 1 to 3, while the topology of each tree component is randomly generated as above.

Figures 7 and 8 summarize the simulation results. For $K = 2$ and $\ell = 4$, all model selection criteria perform well as measured by recovery (*recov*) or preciseness (*prec*), and they improve with sample size. By contrast, the dissimilarity score (*dissim*) was less often significant (dark regions in Figure 8) for EB, AIC, and standard BIC. For $K = 3$ and $\ell = 4$, those criteria perform better according to the preciseness score (white regions in Figure 8), but still considerably worse when evaluated by the dissimilarity score. The picture is similar for $\ell = 6$ events, although EB, AIC, and BIC perform slightly better here, but the dissimilarity score is still more often significant for XV and BIC$_w$. The decreasing performance gap between BIC and BIC$_w$ for increasing ℓ may be explained by the reduced chance of adding redundant structural elements.

The improved dissimilarity performance of XV and BIC$_w$ over all other criteria appears to be due to the fact that EB, AIC, and BIC tend to select too many components for all sample sizes (Figure 7). We note that an increasing number of estimated tree components can improve both recovery and preciseness and this is the reason for the large number of significant *recov* and *prec* scores. The advantage of XV and BIC$_w$ over the other approaches becomes most apparent for the dissimilarity score, which penalizes both poor tree reconstruction and deviation of the estimated from the true number of trees.

The EB method is not competitive with the other approaches with respect to detecting the number of tree components. The reason for this might be the fact that we only consider $\hat{\mathcal{M}}_{\hat{K}}$ and ignore all the other tree topologies in computing the marginal likelihood $\Pr(\mathcal{D} \mid \hat{K})$, possibly resulting in a poor approximation.

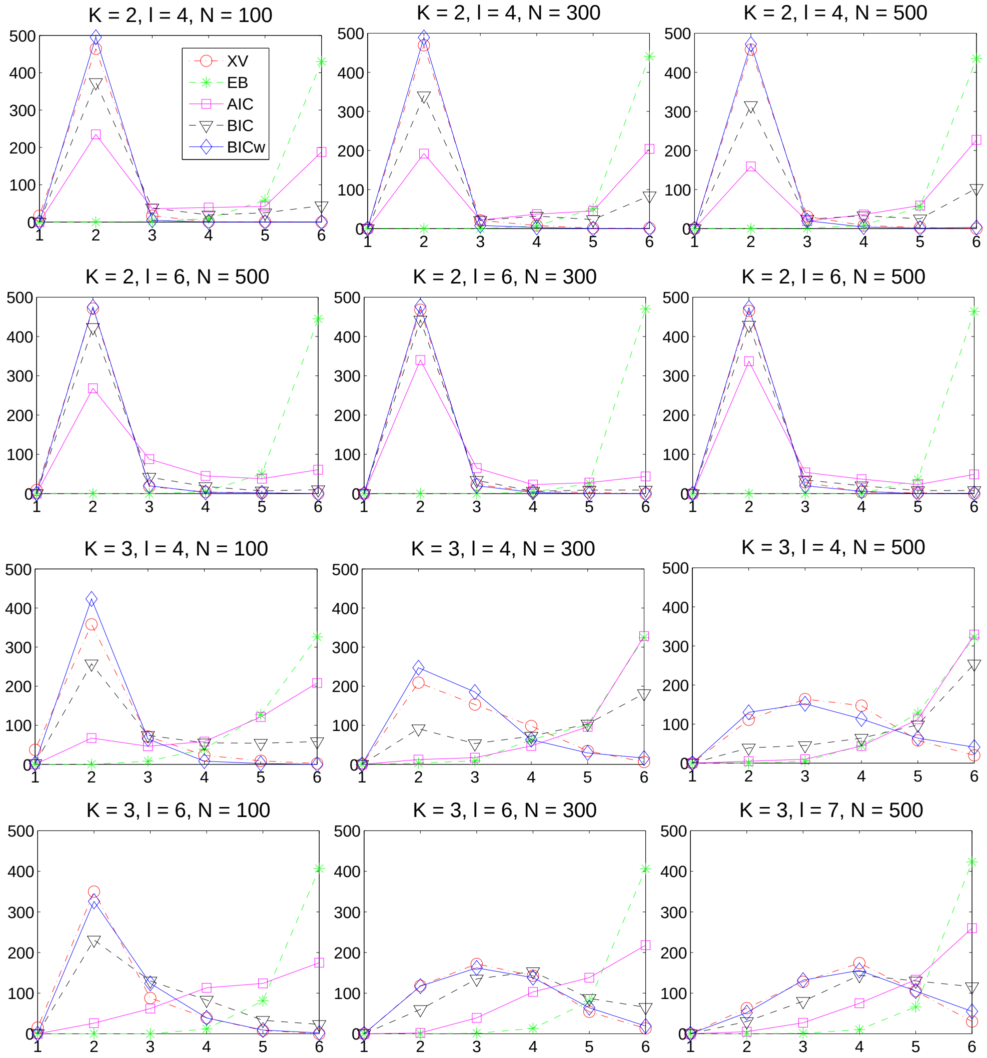
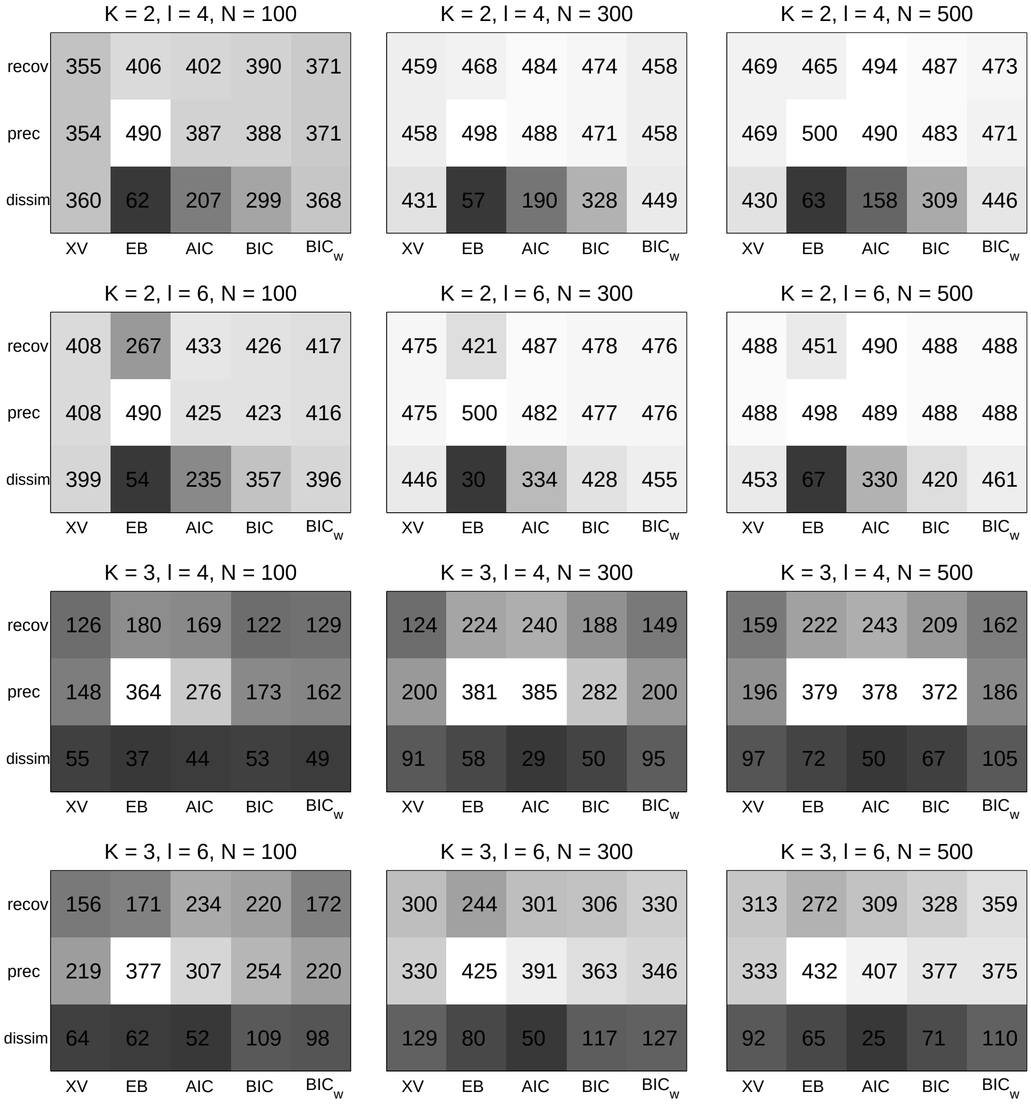


Figure 7: Estimated number of tree components. The x-axis represents the estimated number of tree components ($\hat{K} \in \{1, \dots, 6\}$) and the y-axis its count in 500 simulations. In each panel, the true number of tree components (K), the number of mutations (ℓ), and the sample size (N) are fixed.

5.3 Real World Data

We also applied the five model selection criteria to ten data sets consisting of HIV-1 reverse transcriptase mutation data collected under different drug therapies. In Table 1, the number of tree components chosen by the different criteria and their respective running times are compared. The computing time



K = 2, l = 4, N = 100

	XV	EB	AIC	BIC	BIC_w
recov	355	406	402	390	371
prec	354	490	387	388	371
dissim	360	62	207	299	368

K = 2, l = 4, N = 300

	XV	EB	AIC	BIC	BIC_w
recov	459	468	484	474	458
prec	458	498	488	471	458
dissim	431	57	190	328	449

K = 2, l = 4, N = 500

	XV	EB	AIC	BIC	BIC_w
recov	469	465	494	487	473
prec	469	500	490	483	471
dissim	430	63	158	309	446

K = 2, l = 6, N = 100

	XV	EB	AIC	BIC	BIC_w
recov	408	267	433	426	417
prec	408	490	425	423	416
dissim	399	54	235	357	396

K = 2, l = 6, N = 300

	XV	EB	AIC	BIC	BIC_w
recov	475	421	487	478	476
prec	475	500	482	477	476
dissim	446	30	334	428	455

K = 2, l = 6, N = 500

	XV	EB	AIC	BIC	BIC_w
recov	488	451	490	488	488
prec	488	498	489	488	488
dissim	453	67	330	420	461

K = 3, l = 4, N = 100

	XV	EB	AIC	BIC	BIC_w
recov	126	180	169	122	129
prec	148	364	276	173	162
dissim	55	37	44	53	49

K = 3, l = 4, N = 300

	XV	EB	AIC	BIC	BIC_w
recov	124	224	240	188	149
prec	200	381	385	282	200
dissim	91	58	29	50	95

K = 3, l = 4, N = 500

	XV	EB	AIC	BIC	BIC_w
recov	159	222	243	209	162
prec	196	379	378	372	186
dissim	97	72	50	67	105

K = 3, l = 6, N = 100

	XV	EB	AIC	BIC	BIC_w
recov	156	171	234	220	172
prec	219	377	307	254	220
dissim	64	62	52	109	98

K = 3, l = 6, N = 300

	XV	EB	AIC	BIC	BIC_w
recov	300	244	301	306	330
prec	330	425	391	363	346
dissim	129	80	50	117	127

K = 3, l = 6, N = 500

	XV	EB	AIC	BIC	BIC_w
recov	313	272	309	328	359
prec	333	432	407	377	375
dissim	92	65	25	71	110

Figure 8: Model selection performace. Reported are, for each criterion (cross-validation (XV), empirical Bayes (EB), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and modified BIC (BIC_w)) the number of times each score (recovery ($recov$), preciseness ($prec$), dissimilarity ($dissim$)) was among the top 5% of random scores in 500 simulations. In each panel, the true number of tree components (K), the number of mutations (ℓ), and the sample size (N) are fixed.

DOI: 10.2202/1544-6115.1164

Therapy	N	ℓ	XV	EB	AIC	BIC	BIC$_w$
ZDV	364	7	2 (524)	6 (32)	6 (55)	4 (55)	4 (52)
RTV	112	12	3 (350)	6 (18)	6 (37)	6 (40)	3 (42)
EFV	382	6	3 (550)	6 (56)	5 (53)	5 (60)	3 (67)
NVP	601	7	2 (873)	6 (197)	4 (204)	4 (212)	4 (223)
ABC	50	6	2 (134)	6 (17)	4 (14)	4 (13)	2 (15)
NFV	329	10	3 (550)	6 (21)	6 (20)	4 (26)	3 (23)
ZDV + 3TC	448	10	5 (858)	6 (80)	5 (85)	5 (95)	5 (83)
ZDV + ddI	286	9	3 (1237)	6 (123)	6 (103)	6 (98)	3 (146)
d4T + 3TC	177	10	2 (536)	6 (120)	5 (154)	5 (126)	4 (131)
ddI + d4T	106	9	2 (715)	6 (178)	6 (99)	5 (123)	2 (156)

Table 1: Performance on 10 HIV data sets. Reported are the therapy, the size of the data set (N), the number of mutations (ℓ), and for each model selection criterion the number of estimated tree components and the running time in seconds (in parentheses). Therapies are named by the drugs they comprise in a 3-letter code ('+' denotes treatment with two drugs).

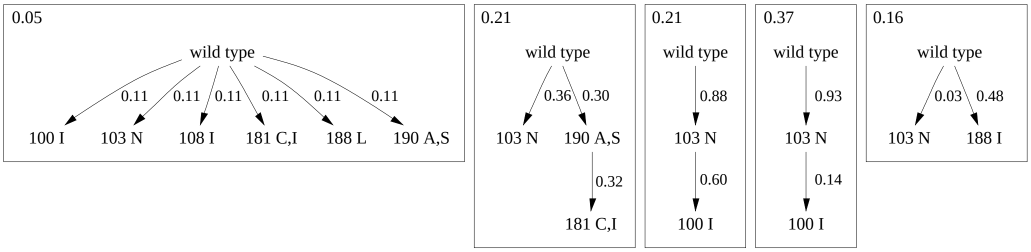


Figure 9: Optimal mixture model for the development of resistance to efavirenz in the HIV-1 reverse transcriptase gene as selected by BIC. See captions of Figures 1 and 2 for details.

for XV is about an order of magnitude higher than for the other methods. XV and BIC$_w$ tend to select similar numbers of trees (in fact, the same number in 6 out of 10 cases), whereas EB, AIC, and BIC consistently prefer models with more components.

For example, the evolutionary models for the development of resistance to the drug efavirenz (EFV) obtained with BIC and BIC$_w$ (or, equivalently, XV) differ significantly (Figures 9 and 10). The 5-mutagenetic trees mixture model obtained with BIC displays considerable structural redundancy. Indeed, all edges appearing in the second, fourth and fifth tree of the BIC model (Figure 9) are present in the second component of the BIC$_w$ model (Figure 10). The third

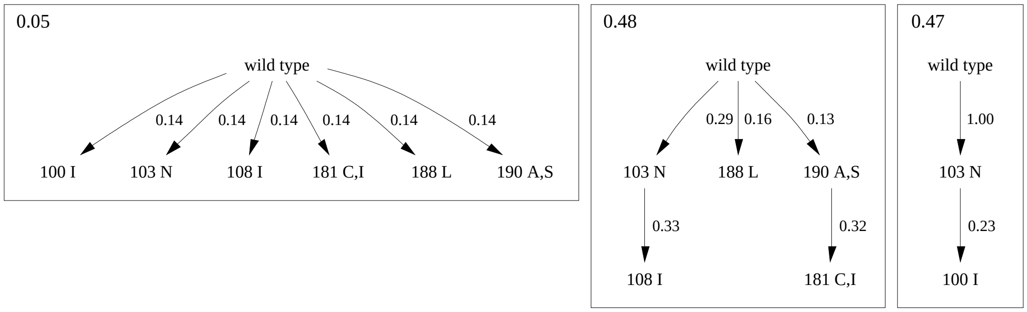


Figure 10: Optimal mixture model for the development of resistance to efavirenz in the HIV-1 reverse transcriptase gene as selected by BIC$_w$. See captions of Figures 1 and 2 for details.

components of both mixture models are topologically identical. Thus, both XV and the modified BIC criterion select a much more parsimonious model and BIC$_w$ does so in one minute as opposed to XV which requires nine minutes of computing time. We find the modified BIC to be the fastest and best performing model selection criterion among the five criteria we have tested.

6 Conclusions

Mixture models of mutagenetic trees provide a family of biologically interpretable models for evolutionary processes that can be described as an accumulation of permanent genetic changes along multiple pathways. The conventional cross-validation approach to detecting the appropriate number of tree components in the mixture model is computationally too expensive such that it is infeasible for large data sets arising in applications. Here we have explored alternative model selection criteria, including empirical Bayes, the Bayesian Information Criterion, and the Akaike Information Criterion, and we have developed a new modified variant of BIC that penalizes redundant tree tolopogies. Empirical Bayes involved approximating the marginal likelihood, which was achieved in a closed form expression after construction of an uninformative prior. The speed-up of the other methods over cross-validation is based on measuring model complexity by the model dimension.

All five model selection criteria were applied to simulated and real world data sets. Only cross-validation and our modified BIC score showed satisfying performance in recovering true model structure over the investigated range of parameters. Given the considerable speed-up of the modified BIC, this criterion appears most useful for practical applications. As such it has been implemented in the `Mtreemix` software package (Beerenwinkel et al., 2005b).

Future work will include the application of more sophisticated similarity measures between trees, in order to better capture the redundancy of the mixture model. Further investigation of the relationship between the dimension and the structural redundancy of the mixture model from an algebro-geometric point of view is also a direction of future research.

7 Acknowledgments

Part of this work has been performed in the context of the BioSapiens Network of Excellence (EU contract no. LSHG-CT-2003-503265). Financial support was provided by an IMPRS scholarship (J.Y.), by Deutsche Forschungsgemeinschaft under grant No. HO 1582/1-3 and BE 3217/1-1 (N.B.) and by BMBF grant No. 01GR0453 (J.R.). We would like to thank Michael I. Jordan and Simon Lacoste-Julien for helpful discussion.

A Average Parameters

We provide here a formal derivation of the "average parameters" used in Section 3.2.

Let $T = (V, E)$ be a mutagenetic tree and denote by C_v the number of compatible substates of the subtree rooted at $v \in V$. In particular, C_0 is the number of compatible states of T. The cardinalities C_v can be computed as follows (cf. Beerenwinkel and Drton (2005), Algorithm 14.5). Visiting the vertices of T in reverse topological order, we have for all $v \geq 1$,

$$C_v = \begin{cases} 2 & \text{if } v \text{ is a leaf} \\ 1 + \prod_{w \in \mathrm{ch}(v)} C_w & \text{else} \end{cases}$$

and $C_0 = \prod_{w \in \mathrm{ch}(0)} C_w$, because the compatible states of any tree arise as the all zero state $(0, \ldots, 0)$ and the combinations of substates that are compatible with the subtrees rooted at each of the children of the root.

Proposition 1. *The parameters $\tilde{p}_v := (C_v - 1)/C_v$ define the uniform distribution, i.e.,* $\Pr(x \mid T, \tilde{p}) = 1/C_0$ *for all compatible states* $x \in \{0, 1\}^\ell$.

Proof. We proceed by induction on the number of vertices $\ell + 1$. If $\ell = 1$, then $\tilde{p}_1 = 1/2$ and $\Pr(X_1 = 0 \mid T, \tilde{p}) = \Pr(X_1 = 1 \mid T, \tilde{p}) = 1/2$.

Let $\ell > 1$. We decompose the joint distribution of X according to the first branching induced by the root and its children:

$$\Pr(x \mid T, \tilde{p}) = \prod_{\substack{w\in\mathrm{ch}(0)\\ x_w=1}} \tilde{p}_w \Pr(x[w] \mid T, \tilde{p},\, x[w] \neq (0,\ldots,0)) \cdot \prod_{\substack{w\in\mathrm{ch}(0)\\ x_w=0}} (1-\tilde{p}_w),$$

where $x[w]$ denotes the subpattern induced by the variables that appear in the subtree rooted at w. Now, $\Pr(x[w] \mid T, \tilde{p}) = 1/C_w$ by hypothesis. Hence $\Pr(x[w] \mid T, \tilde{p},\, x[w] \neq (0,\ldots,0)) = 1/(C_w - 1)$ and we have

$$\Pr(x \mid T, \tilde{p}) = \prod_{\substack{w\in\mathrm{ch}(0)\\ x_w=1}} \frac{C_w-1}{C_w}\frac{1}{C_w-1} \cdot \prod_{\substack{w\in\mathrm{ch}(0)\\ x_w=0}} \frac{1}{C_w} = \prod_{w\in\mathrm{ch}(0)} \frac{1}{C_w} = \frac{1}{C_0}.$$

□

References

Akaike H. (1974). A new look at statistical model identification. IEEE Transactions on Automatic Control 19:716–723.

Beerenwinkel N., Rahnenführer J., Däumer M., Hoffmann D., Kaiser R., Selbig J. and Lengauer T. (2005a). Learning Multiple Evolutionary Pathways from Cross-sectional Data. J. Comput. Biol. 12:584–598.

Beerenwinkel N., Rahnenführer J., Däumer M., Hoffmann D., Kaiser R., Selbig J. and Lengauer T. (2005b). Mtreemix: a software package for learning and using mixture models of mutagenetic trees. Bioinformatics 21:2106–2107.

Beerenwinkel N. and Drton M. (2005). Mutagenetic Tree Models. In Pachter L. and Sturmfels B. (eds) Algebraic Statistics for Computational Biology, chapter 14. Cambridge University Press, Cambridge, UK.

Boucher C. A. et al. (1992). Ordered appearance of zidovudine resistance mutations during treatment of 18 human immunodeficiency virus-positive subjects. J. Infect. Dis. 165:105–110.

Desper R., Jiang F., Kallioniemi O.-P., Moch H., Papadimitriou C.H. and Schäffer A.A. (1999). Inferring tree models for oncogenesis from comparative genome hybridization data, J. Comp. Biol. 6(1):37–51.

Dempster A.P., Laird N.M. and Rubin D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Statist. Soc. B 39:1–38.

Geiger D., Heckerman D. and Meet C. (1996). Asymptotic model selection for directed networks with hidden variables. Proc. 12th Conf. on Uncertainty in Artificial Intelligence (UAI '96), pp. 158-168.

Ghahramani Z. (2004). Unsupervised Learning, In Bousquet, O., Raetsch, G. and von Luxburg, U. (eds) Advanced Lectures on Machine Learning LNAI 3176, chapter 5. Springer, Heidelberg.

Hastie T., Tibshirani R. and Friedman J. (2001). The Elements of Statistical Learning. Springer, New York, NY.

Jefferys W.H. and Berger J.O. (1992). Ockham's razor and Bayesian analysis, American Scientist 80:64–72.

Jordan M. I. (2004). Graphical models. Statistical Science (Special Issue on Bayesian Statistics) 19:140–155.

Jordan M. I. (2006). An Introduction to Probabilistic Graphical Models. Forthcoming.

Molla A. et al. (1996). Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. Nat. Med. 2:760–766.

Prüfer H. (1918). Beweis eines Satzes über Permutationen. Arch. Math. Phys. 27:742–744.

Robert C. P.(2001). The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. Springer, New York, NY.

Schwarz G. (1978). Estimating the dimension of a model. Annals of Statistics 6:461–464.

Spivak M. (1979). A Comprehensive Introduction to Differential Geometry 1. Publish or Perish.

Van Allen T. and Greiner R. (2000). Model selection criteria for learning belief nets: an empirical comparison. Proc. 17th Inter. Conf. on Machine Learning (ICML '00), pp. 1047-1054.