Statistical Applications in Genetics and Molecular Biology

Volume 6, *Issue* 1 2007 *Article* 11

A Bayesian Model of AFLP Marker Evolution and Phylogenetic Inference

Ruiyan Luo* Andrew L. Hipp†

Bret Larget‡

*University of Wisconsin - Madison, rluo@stat.wisc.edu

†The Morton Arboretum, ahipp@mortonarb.org

‡University of Wisconsin - Madison, brlarget@wisc.edu

A Bayesian Model of AFLP Marker Evolution and Phylogenetic Inference*

Ruiyan Luo, Andrew L. Hipp, and Bret Larget

Abstract

Amplified Fragment Length Polymorphism (AFLP) markers are formed by selective amplification of DNA fragments from digested total genomic DNA. The technique is popular because it is a relatively inexpensive way to produce large numbers of reproducible genetic markers. In this paper, we describe a Bayesian approach to modeling AFLP marker evolution by nucleotide substitution and an MCMC approach to estimate phylogeny from AFLP marker data. We demonstrate the method on species in *Carex* section *Ovales*, a group of sedges common in North America. We compare the results of our analysis with a clustering method based on Nei and Li's restriction-site distance and a two-state Bayesian analysis using MrBayes.

KEYWORDS: amplified fragment length polymorphism, Markov chain Monte Carlo, phylogeny, restriction site, statistical phylogenetics

---

*R.L and B.L were supported in part by NIH grant R01 GM068950-01. A.H.'s work was supported in part by NSF Dissertation Improvement Grant 0308975.

INTRODUCTION

The amplified fragment-length polymorphism (AFLP) technique, first developed by Vos *et al.* (1995), is a powerful tool to produce DNA fingerprints of organismal genomes. The generation of AFLP markers involves breaking the whole genome into fragments with restriction enzymes. A small fraction of the fragments, selected by a specific primer pair, are amplified using PCR and visualized using gel electrophoresis. The fragments are measurable in units of nucleotides. Bands for which there is some variability among the separate individuals under study are the genetic markers. A single primer pair can produce twenty to one hundred or more separate markers in a single individual, and several times this number in a population sample of individuals. Additional markers can be found using different primer pairs. While the method has been calibrated in some well studied organisms to distinguish individuals that are heterozygous at a given locus, the resulting data are usually recorded as a 0/1 matrix–allele absent or allele present, with no information about whether the individual is homozygous or heterozygous–with a row for each individual in the study and a column for each marker. The specific primer pair and fragment length associated with each marker are known with relative certainty (though see discussion below).

One increasingly common use of AFLP marker data is as a source of genetic information for phylogenetic inference, the estimation of evolutionary trees from genetic data. In addition to low cost, there are additional characteristics of AFLP markers that make them suitable for phylogenetic inference in many situations. AFLP markers are highly reproducible (Powell *et al.*, 1996; Jones *et al.*, 1997) and easily detected using automated sequencers and software. They are less prone to homology problems than are other anonymous DNA fragment methods such as randomly amplified polymorphic DNA fragments (RAPD) or inter-simple sequence repeat (ISSR) polymorphisms (Wolfe and Liston, 1998). Moreover, as a multilocus method, AFLPs have the benefit of integrating phylogenetic signals from loci distributed throughout the genome, reducing the degree to which lineage sorting and reticulate evolution (hybridization) are expected to confound efforts to reconstruct phylogenies among rapidly radiating taxa (Albertson *et al.*, 1996). Because of these qualities, AFLPs have come into increasingly frequent use in phylogenetic studies among closely-related species.

The most common method of inferring phylogenetic relationships from AFLP and other anonymous molecular fragment data (Landry and Lapointe, 1996) is to convert the binary data matrix into a pairwise distance matrix using one of several possible distance measures and then to construct a tree

using a clustering algorithm such as neighbor joining (Saitou and Nei, 1987) or the unweighted pair-group method using arithmetic average (UPGMA). A more rigorous but less frequently used method uses the same pairwise distance matrix, but searches for an optimal tree using a least-squares or minimum evolution criterion (Rzhetsky and Nei, 1992). Innan *et al.* (1999) have developed likelihood-based distance methods to estimate genetic divergence between two individuals from AFLP marker data, based on the Jukes-Cantor nucleotide substitution model (Jukes and Cantor, 1969), which assumes equal base composition and equal rates of substitution specifically. Vekemans *et al.* (2002) extend this approach for general genomic GC content. These methods both build on Nei and Li (1979)'s work on inferring nucleotide diversity from restriction fragment polymorphism (RFLP) data.

Distance methods are generally considered to be less desirable than character-based methods because reduction of discrete characters to pairwise distances entails the loss of information. Maximum parsimony, a character-based method that specifies that the optimal tree is that which requires the least changes in character states, can be used to infer phylogenetic trees from a wide range of quantitative or categorical characters. Parsimony, however, poses several problems for AFLP data. First, because the presence of an AFLP band depends on the presence of two specific recognition sites (Mueller and Wolfenbarger, 1999), parallel losses of a band should be more common than parallel gains. This asymmetry violates the reversibility assumption of Wagner parsimony. This asymmetry cannot be altogether gotten around using Dollo or weighted parsimony, because parsimony also does not account for the expectation that parallel gains should be more frequent in closely-related taxa than in distantly-related taxa.

Mau and Newton (1997) present a Bayesian model for phylogenetic inference from binary data. A similar model is implemented in the software MrBayes (Huelsenbeck and Ronquist, 2001). This two-state Markov model for marker presence and absence is an over-simplification of the underlying biological process by which AFLP fragments come into and out of existence. Stated another way, a Markov process acting on the hidden genetic states would not, in general, result in a Markovian observable process.

Felsenstein (1992) describes a model of restriction site evolution that more closely approximates the evolution of AFLP fragments (Smouse and Li, 1987; Felsenstein, 2004). However, the evolution of a given AFLP fragment is likely to be more complex than the evolution of the restriction sites that flank it. For example, in the typical double-digest AFLP method with a fluorescent label on only one of the primer pairs, the presence of a band is contingent on the presence of either of two combinations of restriction sites. Moreover,

DOI: 10.2202/1544-6115.1152

mutations, insertions, and deletions within the band itself can cause that band to "disappear" from one generation to the next, even if no mutations affect the restriction sites on each side of it. Accurate phylogenetic inference based on AFLP data is likely to benefit from the development of more accurate models of AFLP evolution.

In this paper, we develop an explicit likelihood model for AFLP marker evolution based on the underlying genetic changes that cause marker gain and loss. The second section of the paper provides a detailed description of the molecular basis of AFLP marker measurement and the model on which it is based. In the third section we describe novel Markov chain Monte Carlo methodology specifically tailored for our model to implement a Bayesian approach to phylogenetic inference from AFLP marker data. We conclude the paper with an application of the new methodology to analyze AFLPs from several taxa in *Carex* section *Ovales*, a group of sedges common in North America, with a comparison of our method with alternative methods of analysis, and with a discussion of the computational issues and modeling assumptions associated with our method.

MODEL FOR AFLP MARKER EVOLUTION

Genetic Background of AFLPs

The model we present for AFLP marker evolution is based on the explicit mechanism through which nucleotide sequence data leads to observable marker data. While previously published papers such as Vos *et al.* (1995) describe the experimental method thoroughly, we expect most statistical readers will be unfamiliar with the details. Furthermore, to better understand the basis of our model, we need to present a more thorough description of the underlying genetic processes associated with AFLP markers to show how specific genomic sequences are related to the presence of AFLP markers of specific lengths, a connection that is not at all clear in the literature.

The process of generating AFLP marker data begins by digesting whole genomic DNA, typically with two restriction enzymes. The original protocol (Vos *et al.*, 1995) uses *Eco*RI, which cleaves DNA whenever the sequence "GAATTC" appears in the 5' to 3' direction, and *Mse*I, which cleaves DNA at the four-base recognition sequence "TTAA". (Notice that each of these recognition sequences is a complementary palindrome, so that the recognition sequence on the opposite strand read in the opposite direction is identical.) We will describe our method assuming these are the two restriction enzymes, but our method is easily modified to accommodate other choices. The cleaved

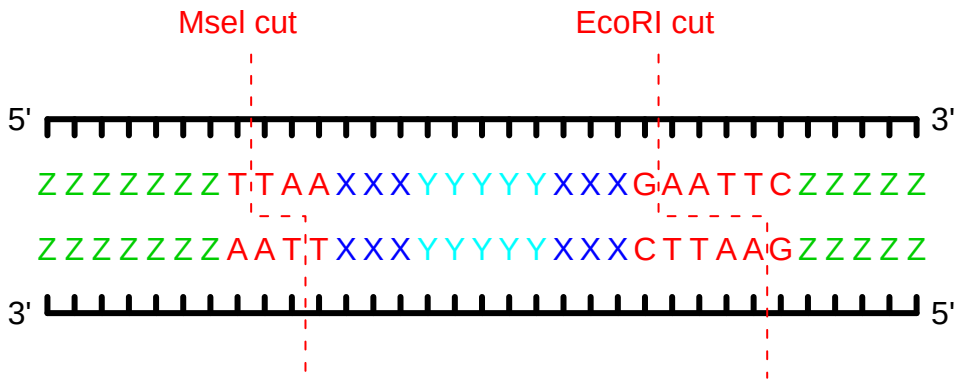

Figure 1: **The restriction of genomic DNA.** In the typical application of the AFLP method (and in the application as described by Vos *et al.* (1995)), total genomic DNA is digested with two restriction enzymes, *Eco*RI and *Mse*I, which cleave the DNA at recognition sequences 5'-GAATTC-3' and 5'-TTAA-3', respectively, along the red dashed line. In the figure, the DNA bases denoted with Xs and Ys are arbitrary as long as there are no additional restriction sites between the two sites shown. The Xs are important in a later step. The DNA bases marked Z extend in each direction. These bases will become parts of different fragments, which may or may not be detected in the final analysis.

DNA fragments have jagged edges with the 5' end of one strand overhanging the 3' end of the other. Each fragment cleaved by *Eco*RI has a four-base single-strand extension with bases "AATT" and the fragments cleaved by *Mse*I have two-base single-strand extensions with bases "TA". (See Figure 1.)

The second step is adaptor ligation, in which double-stranded adaptors specific to each restriction enzyme attach to the end of each fragment, forming caps. Each adaptor is designed so that ligation of a fragment to an adaptor does not reconstitute the restriction site. Here only the *Eco*RI adaptors are fluorescently labelled to make the fragments visible. After digestion and ligation, the complete genomic DNA has been partitioned into many fragments of differing lengths. Each fragment consists of two adaptor end caps and a central portion from the original DNA. Fragments may have identical end caps of either type or one end cap of each type. (See Figure 2.)

The next step is selective DNA amplification using polymerase chain reaction (PCR). The reaction uses a primer pair, one for each adaptor, that matches a portion of the corresponding adaptor and restriction site plus three additional bases (specified by Xs in Figure 2). Fragments that complement one of the two primers on each end double in number in each PCR cycle (Hartl

DOI: 10.2202/1544-6115.1152

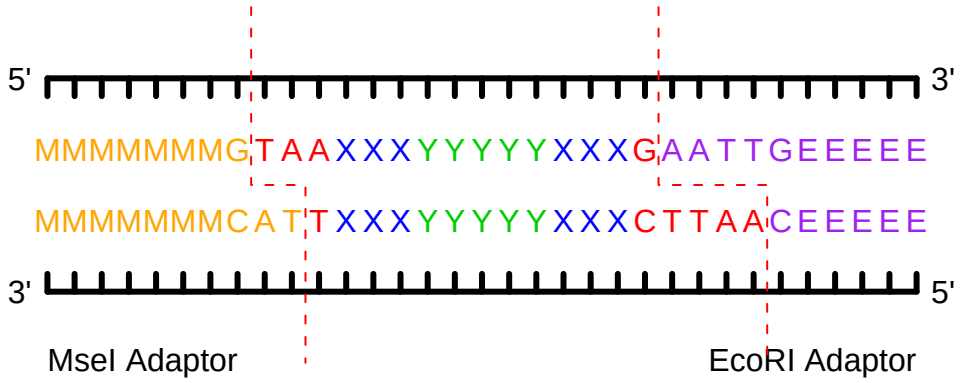

Figure 2: **The ligation of adaptors complimentary to the restriction sites.** Fragments are capped with adaptors specific to the restriction sites on each end. The end sequences of each adapted fragment now consist of the adaptor sequence and the remaining part of the restriction sequence. Adaptor sequences do not recreate the restriction site (e.g., in this case the *Eco*RI restriction site 5'-GAATTC-3' becomes 5'-GAATTG-3' after adaptor ligation).

---

and Jones, 1998). Using primers that require matching three additional bases adjacent to each restriction site causes only a small fraction (approximately $1/4^6 = 1/4096$) of fragments to be amplified. (Depending on genome size, using primers that match other numbers of additional bases can be preferable to achieve appropriate numbers of markers.) The amplified fragments are run through denaturing polyacrylamide gels by electrophoresis, separating the fragments by length. The length of each amplified fragment is the sum of the lengths of the primers plus the length of the sequence in between. Furthermore, as only the *Eco*RI adaptors are fluorescently labelled, the fragments between neighboring *Mse*I restriction sites do not form visible bands. As *Mse*I restriction sites match only four sites instead of six, they are approximately 16 times as frequent. Thus, roughly $(16/17)^2$ or 88.6% of all amplified fragments do not form visible bands.

In the applied example we consider, each primer is 19 base pairs long, so the length of the intermediate region (the Ys in Figures 1, 2, and 3) would be 38 less than the measured fragment length. However, with the commonly used *Taq* DNA polymerase, PCR fragments typically also have an extra adenine (A) appended to the 3' end. To account for this PCR artifact, we subtract 39 from each measured length to find the length of the intermediate region. In limited testing, our inferences appear to be robust to small changes in length adjustments.

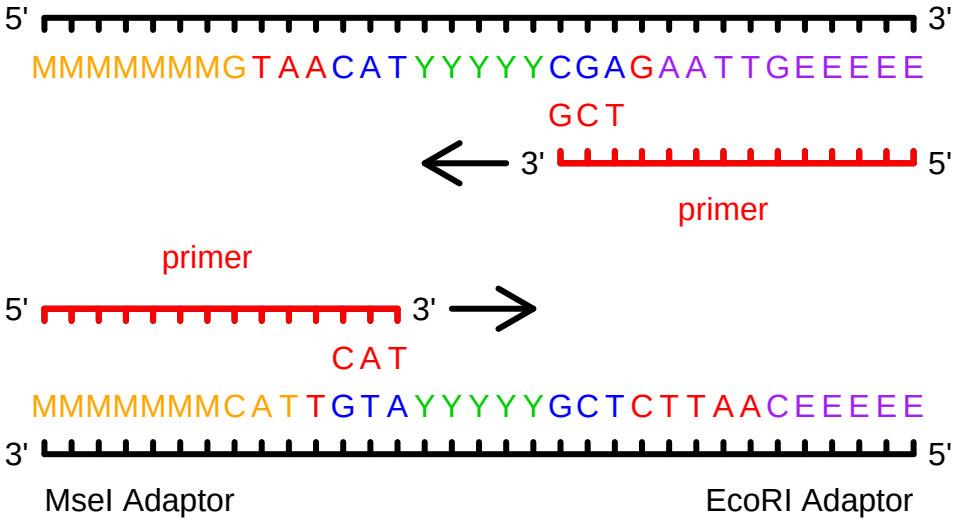

Figure 3: **Selective amplification.** Only a subset of the adapted fragments is amplified by selecting primers which match three bases beyond the restriction site on each end of the fragment. In this case, the end of the *Mse*I primer is complementary to "CAT" and the end of the *Eco*RI primer is complementary to "TCG", from 5′ to 3′.

In addition, only fragments with lengths between 50 and about 600 bases can be sized reliably using most equipment and sizing standards in common use, so only fragments with 11 or more bases in the intermediate region have the possibility of being measured. The fragment in Figure 3 would amplify but would not be scored in the data matrix, because it is too short to size reliably using sizing standards composed of fragments ranging from 50 to 625 nucleotides in length.

Several genetic changes can cause gain or loss of AFLP markers. Nucleotide substitution can eliminate a marker for a specific primer pair by either removing a restriction site or changing one of the six additional bases necessary for amplification. A substitution that creates a new restriction site in the interior of a fragment will also cause a marker loss. Mutational processes other than nucleotide substitution, such as insertion, deletion, and inversion, can also result in sequence changes that affect AFLP markers. Specifically, indel processes could either cause the loss of a marker through removing part of a restriction site or the neighboring amplification sites. Indel events in the intermediate region have the potential to cause a single homologous locus to

DOI: 10.2202/1544-6115.1152

result in two or more markers of different lengths. In this paper, we concentrate on AFLP marker evolution by nucleotide substitution only and consider additional genetic processes in future work.

Model Assumptions

For the purposes of the model developed in this paper, we partition a typical fragment into three parts (see Figure 4). The first and third parts include bases necessary for each end of the fragment to be cut plus additional bases necessary for amplification. We call these two parts together the *end region*. The second part is an *intermediate region* where the specific sequence is unimportant provided no additional restriction sites are present and assuming the Jukes-Cantor model for nucleotide substitution.

Assuming only nucleotide substitution as a mutational process, band loss is due either to mutation in the end region or by gain of a restriction site in the intermediate region. In particular, nucleotide substitution at the end region causes either loss of one of the restriction sites at one end of the fragment merging it with a neighboring fragment or causes a change in the recognition sites causing the fragment not to be amplified. A nucleotide substitution in the intermediate region usually has no effect, but can create a new restriction site resulting in the marker fragment being broken into two smaller fragments. We will model these two processes in the paper.

The basic model of AFLP evolution in this paper rests on the following assumptions: (1) each AFLP marker is associated with a single genetic locus in each individual; (2) the loci in different individuals corresponding to the same AFLP marker are homologous (derived from a single locus in a common ancestor); (3) loci associated with visible markers are mutually independent; (4) bands are appropriately scored as present or absent; (5) each locus is represented by a band that is flanked either by an *Mse*I and an *Eco*RI site (with prior probability 32/33) or by two *Eco*RI sites (with prior probability 1/33); (6) a band is present for an *Mse*I/*Eco*RI fragment if there are zero mismatches among the 16 necessary bases and no restriction sites between the restriction sites corresponding to the fragment ends; (7) a band is present for an *Eco*RI/*Eco*RI fragment if there are zero mismatches among the 18 necessary bases and no restriction sites between the restriction sites corresponding to the fragment ends; and (8) all sites evolve independently with the same rate according to a Jukes-Cantor model (Jukes and Cantor, 1969). In a Jukes-Cantor model, all base pairs are equally likely, and given that a substitution occurs, the new base pair is equally likely to be any of the other three bases. Under these model assumptions, marker gain/loss due to changes at recognition

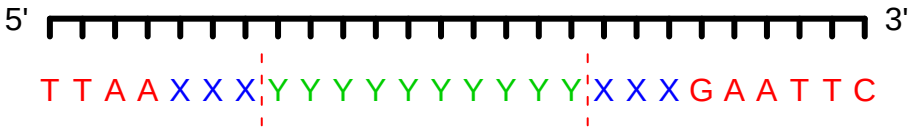

Figure 4: **The partition of a DNA fragment.** A typical fragment corresponding to an AFLP marker is partitioned into three parts. The first and third parts must match specific sequences exactly to be restricted, amplified, and measured. We refer to these sites as the *end region*. The sequence in the intermediate region, can vary provided that there are no restriction sites in the sequence.

---

sites is independent of fragment length, but changes due to the introduction of new restriction sites in the interior of a fragment depend on fragment length. Assumptions (1)–(3) are reasonable if indel rates are low compared to the substitution rate and if the chance of genomes containing multiple loci producing markers at the same lengths is small. Assumptions (4)–(7) will follow if the actual measurement process is accurate and consistent with the mechanism we described earlier. Assumption (8) deals with the likelihood model for nucleotide sequence evolution by substitution.

Notation

Let x_{ij} denote the AFLP marker value for the i-th taxon and the j-th band: $x_{ij} = 1$ indicates the presence of marker j in taxon i, and $x_{ij} = 0$ indicates the absence. The measured amplified fragment length (including lengths of the adaptors) is denoted as L_j for the j-th band. We let R_j be the length of the end region: $R_j = 16$ for *Mse*I/*Eco*RI fragments and $R_j = 18$ for *Eco*RI/*Eco*RI fragments. In addition, we let N_j be the number of nucleotide bases in the intermediate region where $N_j = L_j - 39$ if the sum of the primer lengths is 38 and *Taq* DNA polymerase is used.

Modeling AFLPs in a Single Lineage

Consider now the evolution in time in a single lineage of an AFLP marker corresponding to a locus with R bases in the end region and N intermediate bases. From the previous discussion, nucleotide substitution can lead to an AFLP marker loss either by producing a mismatch in the end region or by

DOI: 10.2202/1544-6115.1152

producing a new restriction site in the intermediate region. To measure these two genetic causes of marker loss by substitution, we introduce two processes. At time t, let $M(t)$ be the number of mismatches among the R recognition bases and let $C(t)$ be the number of cutters among the N middle bases, where we define a cutter to be an additional restriction site of either four or six bases in the intermediate region of a marker. The marker indicator $X(t) = 1$ when $M(t) = C(t) = 0$ and is zero otherwise. The process $M(t)$ is a function of a Markov chain on the state space of all possible sequences of R bases. Under our model assumptions (Jukes-Cantor, independent bases), $M(t)$ itself is a continuous-time Markov process on the state space $0, 1, 2, \ldots, R$ where the only positive infinitesimal transition rates are between states that differ by exactly one. $C(t)$ is also a function of a Markov chain on the state space of all possible sequences of N bases and is independent of $M(t)$, but $C(t)$ is not itself a Markov process due to the overlap of possible restrictions sites. However, we can approximate the process $Z(t) = 1_{\{C(t)>0\}}$ that indicates presence of at least one cutter accurately with a two-state continuous-time Markov chain.

First consider the process for $M(t)$. This part of the model is identical to the model for restriction sites in Felsenstein (1992) except that we have a larger number of states. Under the Jukes-Cantor model, the probability that the base at a single site is different at time t than the base at time 0 is

$$p(t \mid u) = \frac{3}{4}\left(1 - \mathrm{e}^{-\frac{4}{3}ut}\right) \tag{1}$$

where u is the rate of substitutions per unit time.

If there are $R = r$ bases, the probability of changing from i mismatches to j mismatches in time t can be computed as a sum of a product of two binomial probabilities, summing over the number of matches that become mismatches, as in Felsenstein (1992):

$$P_{ij}^{(r)}(t) = \sum_{k=\max(0,i-j)}^{\min(i,r-j)} \left(\binom{r-i}{j-i+k} p^{j-i+k}(1-p)^{r-j-k}\right)\left(\binom{i}{k}\left(\frac{p}{3}\right)^{k}\left(1-\frac{p}{3}\right)^{i-k}\right)$$

where $p = p(t \mid u)$. The stationary probability of i mismatches among r independent sites is

$$\pi_i^{(r)} = \binom{r}{i}\frac{3^i}{4^r} \quad . \tag{2}$$

The infinitesimal transition rate from state i to state j for the $M(t)$ process

conditional on r is:

$$Q_{ij} = \begin{cases} iu/3 & \text{if } j = i-1 \\ (r-i)u & \text{if } j = i+1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } 0 \le i, j \le r$$

and $P(t) = \mathrm{e}^{Qt}$.

Now consider the process $C(t)$ that counts the number of restriction sites in the intermediate region. We let $Z(t) = 1_{\{C(t)>0\}}$ indicate the presence of interior *Mse*I or *Eco*RI cutters. For a fragment with N bases in the intermediate region, there are $N-4+1$ and $N-6+1$ possible locations for *Mse*I and *Eco*RI sites, cutting at sequences "TTAA" and "GAATTC", respectively. Each potential interior restriction site corresponds to an indicator variable, and $Z(t) = 1$ if any one of the $2N-8$ indicator variables equals one. The exact stochastic process for $Z(t)$ is complicated due to the overlap in potential restriction sites and corresponding lack of independence. However, an approximation of $Z(t)$ with a two-state Markov process on state space $\{0,1\}$ is very accurate. Intuitively, this is the case because each individual indicator variable has quite small success probability and most pairs of potential sites are independent as they do not overlap. Furthermore, we show that the dependence among overlapping sites is weak.

In a fragment with N nucleotides in the intermediate region, there are $N-6+1$ possible 6 bp sequences and $N-4+1$ possible 4 bp sequences. Consider the former first and let $I_i^{(6)}$ indicate the presence of the sequence "GAATTC" for the six bases starting at position i for $i = 1, 2, \cdots, N-6+1$. For some location away from the left end where $i > 5$, we will calculate the probability of a restriction site given that the previous overlapping 6 bp regions do not contain restriction sites.

$$\begin{aligned} & \mathbb{P}\left(I_i^{(6)} = 1 \,|\, I_{i-1}^{(6)} = 0, I_{i-2}^{(6)} = 0, \ldots, I_{i-5}^{(6)} = 0\right) \\ = & \frac{\mathbb{P}\left(I_i^{(6)} = 1, I_{i-1}^{(6)} = 0, I_{i-2}^{(6)} = 0, \ldots, I_{i-5}^{(6)} = 0\right)}{\mathbb{P}\left(I_{i-1}^{(6)} = 0, I_{i-2}^{(6)} = 0, \ldots, I_{i-5}^{(6)} = 0\right)} \\ = & \frac{\mathbb{P}\left(I_i^{(6)} = 1\right)}{\mathbb{P}\left(I_{i-1}^{(6)} = 0, I_{i-2}^{(6)} = 0, \ldots, I_{i-5}^{(6)} = 0\right)} \\ \approx & \mathbb{P}\left(I_i^{(6)} = 1\right) . \end{aligned}$$

DOI: 10.2202/1544-6115.1152

The final approximation holds since

$$\begin{aligned} & \mathbb{P}\left(I_{i-1}^{(6)}=0, I_{i-2}^{(6)}=0, \ldots, I_{i-5}^{(6)}=0\right) \\ = \; & 1-\mathbb{P}\left(\text{at least one of } I_{i-j}^{(6)}=1, j=1, \ldots, 5\right) \\ = \; & 1-\frac{5}{4^6} \approx 1 \; . \end{aligned}$$

So the indicator variables $\{I_i^{(6)}\}$ are weakly dependent for different i. A similar argument shows that $\{I_i^{(4)}\}$ are also weakly dependent, and we argue as well that $\{I_i^{(6)}\}$ and $\{I_i^{(4)}\}$ are mutually weakly dependent. This is a theoretical argument that a two-state approximation can be expected to be accurate. We note that Innan *et al.* (1999) gives the distribution of N under the assumption that the indicator random variables above are independent.

We tested this approximation by simulating the true distribution of $Z(t)$, which is a function of a 4^N state Markov chain, by simultaneous simulation of N independent nucleotide bases under the Jukes-Cantor model. Specifically, we began with an initial sequence chosen from the stationary distribution. After an exponentially distributed amount of time, we picked a site uniformly at random and changed its base to one of the other three, uniformly at random. If this change in the sequence changed $C(t)$, we modified $Z(t)$ accordingly. From this procedure we obtained a series of alternating dwell-times. For a two-state Markov process, dwell-times would be mutually independent exponential random variables with odd- and even-indexed dwell-times possibly having different means. In the simulation, each marginal distribution is indistinguishable from an exponential distribution and the observed auto-correlation coefficient is bounded by 0.001. We conclude that $\{Z(t)\}$ can be approximated well by a two-state Markov process.

We next discuss how to estimate the probability transition matrix of the Markov process approximating $\{Z(t)\}$. Let $\pi_0^{(Z)}$ be the stationary probability that there is no cutter in the intermediate region. Assuming a Jukes-Cantor model and independence, the stationary distribution of no cutters in a fragment with $N = n$ bases in the intermediate region is as follows

$$\pi_0^{(Z)} = P(Z(0)=0) \approx \left(1-\frac{1}{4^4}\right)^{n-4+1}\left(1-\frac{1}{4^6}\right)^{n-6+1} . \qquad (3)$$

We can estimate the infinitesimal rate of moving from zero to at least one cutter in the following manner. Consider first a potential restriction site with four bases. With our assumptions including a substitution rate of u per site, the rate

of moving to a restriction site is $u/3$ if the four bases have a single mismatch (if the mismatched bases changes, there is a 1/3 chance of changing to the matching base) and 0 if there are two or more mismatches. The conditional probability of exactly one mismatch given at least one is $4\times(1/4)^3\times(3/4)/(1-(1/4)^4)$. The overall rate is the expected number of potential restriction sites with exactly one mismatch given none have zero mismatches times the rate of each, or

$$(n-4+1)\times\frac{4(1/4)^3(3/4)}{1-(1/4)^4}\times\frac{u}{3}=\frac{4(n-4+1)u}{4^4-1}\,. \tag{4}$$

There is a similar expression for the potential restriction sites with six bases and we have the following estimate for the infinitesimal rate

$$q_{01}^{(Z)}\approx\frac{4(n-4+1)u}{4^4-1}+\frac{6(n-6+1)u}{4^6-1}\,. \tag{5}$$

Equations 3 and 5 are sufficient to determine the approximate probability transition matrix for $\{Z(t)\}$ (as shown in the appendix)

$$P^{(Z)}(t)=\begin{pmatrix}\pi_0^{(Z)}+(1-\pi_0^{(Z)})\eta(t) & (1-\pi_0^{(Z)})(1-\eta(t))\\ \pi_0^{(Z)}(1-\eta(t)) & 1-\pi_0^{(Z)}(1-\eta(t))\end{pmatrix}\,, \tag{6}$$

where

$$\eta(t)=\exp\left(-\frac{q_{01}^{(Z)}ut}{1-\pi_0^{(Z)}}\right)\,. \tag{7}$$

Typically, we measure t in units of the expected number of substitutions per site and $u=1$.

Computing the likelihood

The processes $\{M(t)\}$ and $\{Z(t)\}$ characterize the AFLP data under the assumption that band losses and gains are due to nucleotide substitutions, but not to other events such as insertions and deletions. By definition, we compute the likelihood of a tree (tree topology and edge lengths) given data at the leaves by summing over all possible states at the internal nodes. In practice, this calculation is efficient using a dynamic programming algorithm known in the phylogenetics literature as Felsenstein's "pruning" algorithm (Felsenstein, 1981) which requires storage space and time linear, rather than exponential, in the number of nodes in the tree.

We modify the pruning algorithm slightly and define the "conditional likelihood" for a given node i on the tree and given location, j as the probability

DOI: 10.2202/1544-6115.1152

$L_i^{(j)}(z, m, r)$ that we would observe the marker data for location j at the descendants of node i, given that the *parent* of node i was in state (z, m) and that the recognition sequence is r bases long. This parameterization is similar to that in Larget and Simon (1999) and simplifies slightly the internal representation of the partial likelihood calculations. The resulting computed likelihood is identical to that found using Felsenstein's approach which conditions on the state at the node i rather than the parent of i. The parameter z indicates whether or not there are cutters in the interior region of the sequence at the parent node, and m counts the number of mismatches in $0, 1, \ldots, r$.

For a node i which is a tip of the tree, location j has a marker present ($x_{ij} = 1$) if and only if $z = 0$, $m = 0$. The conditional likelihood given the marker data, end region length $R = r$ and intermediate region length $N = n$, is

$$L_i^{(j)}(z, m, r) = \begin{cases} P_{z,0}^{(Z)}(t_i \,|\, n) \times P_{m,0}^{(M)}(t_i \,|\, r) & \text{if } x_{ij} = 1 \\ 1 - \left(P_{z,0}^{(Z)}(t_i \,|\, n) \times P_{m,0}^{(M)}(t_i \,|\, r) \right) & \text{if } x_{ij} = 0 \end{cases} \tag{8}$$

where t_i is the length of the branch leading to the tip i. The conditional likelihood for a non-root internal node i is a function of the edge length to the parent and the conditional likelihoods at the children nodes k and l.

$$L_i^{(j)}(z, m, r) = \sum_{z'=0}^{1} \sum_{m'=0}^{r} P_{z,z'}^{(Z)}(t_i \,|\, n) \times P_{m,m'}^{(M)}(t_i \,|\, r) \times L_k^{(j)}(z', m', r) \times L_l^{(j)}(z', m', r) \; . \tag{9}$$

Using the pruning algorithm, we compute the conditional likelihoods for leaves using Equation (8) and for non-root internal nodes using Equation (9) provided that conditional likelihoods at children nodes are computed before the parent nodes as in a post-order traversal of the tree (Drozdek, 2001).

As *Mse*I sites are expected *a priori* to occur 16 times as often as *Eco*RI sites and *Mse*I/*Eco*RI fragments are indistinguishable from *Eco*RI/*Mse*I fragments, we expect 32 times as many *Mse*I/*Eco*RI fragments as *Eco*RI/*Eco*RI fragments. We compute the overall likelihood at location j with intermediate region length n as the weighted sum of the conditional likelihoods at the two immediate children of root.

$$\begin{aligned} L^{(j)} = & \frac{32}{33} \sum_{z=0}^{1} \sum_{m=0}^{16} \pi_z^{(Z \,|\, n)} \times \pi_m^{(M \,|\, R=16)} \times L_k^{(j)}(z, m, 16) \times L_l^{(j)}(z, m, 16) \\ & + \frac{1}{33} \sum_{z=0}^{1} \sum_{m=0}^{18} \pi_z^{(Z \,|\, n)} \times \pi_m^{(M \,|\, R=18)} \times L_k^{(j)}(z, m, 18) \times L_l^{(j)}(z, m, 18) \end{aligned} \tag{10}$$

where $\pi_z^{(Z\,|\,n)}$ and $\pi_m^{(M\,|\,r)}$ are the stationary probabilities for Markovian processes $\{Z(t)\}$ and $\{M(t)\}$. The overall likelihood of the tree is the product of these over all locations

$$L = \prod_{j=1}^{S} L^{(j)} \,, \tag{11}$$

under the assumption that the S locations evolves independently.

Prior Specification

We need to specify the prior distribution in order to complete the model specification. In this paper we assume the typical uniform prior distribution over all possible unrooted tree topologies with the root selected uniformly at random from the tree edges. This vague prior distribution does not likely correspond to that of any individual interested in the particular analysis, but is conservative in the sense that each possible evolutionary relationship is small *a priori* so that high posterior probabilities for specific relationships must be strongly supported by the likelihood. We also assume that edge lengths $\{t_i\}$ on the rooted tree are mutually independent exponential random variables with common mean μ. This implies that the factor of the prior distribution due to edge lengths depends only on the sum of all edge lengths in the tree.

$$\prod_{i=1}^{E} \frac{1}{\mu} \mathrm{e}^{-t_i/\mu} = \left(\frac{1}{\mu}\right)^{E} \mathrm{e}^{-\sum_{i=1}^{E} t_i/\mu} \,,$$

where E is the number of edges. The likelihood does not distinguish among different rootings of the tree in this model, so we typically report summaries of unrooted trees.

MCMC APPROACH

Under this model, exact calculation of Bayesian posterior probabilities is intractable. We use Markov chain Monte Carlo (MCMC) (Metropolis *et al.*, 1953; Hastings, 1970; Green, 1995) to approximate the posterior distribution of the phylogenies. The idea is to construct a Markov chain that has as its state space the tree topologies and edge lengths and a stationary distribution that is the posterior probability distribution of them. In our MCMC approach, at each stage we select one of several possible updates with probability p_k. Taken in combination, these updates suffice to visit the entire state space. For a rooted tree with T taxa, there are $2T-1$ nodes including T leaves and $T-1$

DOI: 10.2202/1544-6115.1152

internal nodes, and $E = 2T - 2$ edges including T external edges and $T - 2$ internal edges.

For each update below we compute the prior ratio, the likelihood ratio, and the proposal ratio (or Hastings ratio). We follow Green (2003) to compute proposal ratios based on densities of the random variables used in the proposal and the Jacobian of a transformation. The acceptance probability is a function of these ratios.

1. **Update all edge lengths.** Multiply all edge lengths by a common random factor $\gamma \sim \Gamma(\alpha, \alpha)$, where α is a tuning parameter. The prior ratio is
$$\prod_{i=1}^{E} \frac{\mathrm{e}^{-t_i\gamma/\mu}/\mu}{\mathrm{e}^{-t_i/\mu}/\mu} = \mathrm{e}^{-\sum_{i=1}^{E} t_i(\gamma-1)/\mu} ,$$
where E is the number of edges. The proposal ratio is
$$\frac{(\gamma^*)^{\alpha-1}\mathrm{e}^{-\alpha\gamma^*}}{\gamma^{\alpha-1}\mathrm{e}^{-\alpha\gamma}} = \frac{\mathrm{e}^{\alpha(\gamma-\frac{1}{\gamma})}}{\gamma^{2(\alpha-1)}}$$
since $\gamma^* = \frac{1}{\gamma}$. Note that new edge length $t_i^* = t_i\gamma$ for $i = 1, \ldots, E$ and $\gamma^* = \frac{1}{\gamma}$, the Jacobian (Green, 2003) is easily obtained as
$$\frac{\partial(t_1^*, \ldots, t_E^*, \gamma^*)}{\partial(t_1, \ldots, t_E, \gamma)} = -\gamma^{E-2} . \tag{12}$$
So, the acceptance probability for this proposal is
$$\min\left\{1, \exp\left(-\frac{1}{\mu}\sum_{i=1}^{E} t_i(\gamma-1) + \alpha(\gamma - \frac{1}{\gamma})\right)\gamma^{E-2\alpha} \times LR\right\} \tag{13}$$
where LR is the likelihood ratio.

2. **Update a single edge length.** Pick a random edge and multiply its length (denoted as t_i) by a random factor $C \sim \Gamma(\alpha, \alpha)$, where α is a tuning parameter. The acceptance probability is
$$\min\left\{1, \mathrm{e}^{-t_i(\gamma-1)/\mu+\alpha(\gamma-\frac{1}{\gamma})}\gamma^{1-2\alpha} \times LR\right\} . \tag{14}$$

3. **Reroot the tree.** Remove the root and treating the tree as unrooted, randomly pick an edge (with probability $\frac{1}{E-1}$) and select a new location for the root on this edge uniformly at random. Denoting the sum of the

two edge lengths from the old and new root as t_+ and t_+^*, respectively, the acceptance probability is

$$\min\left\{1, \frac{t_+^*}{t_+} \times LR\right\} . \tag{15}$$

4. **Local update.** A local update for unrooted trees is described in Larget and Simon (1999), but the derived acceptance probability was incorrect. A corrected formula is in Holder *et al.* (2005). In this version, we modify the proposal by updating the position of the root when the root is part of the local neighborhood under modification. The acceptance probability depends on the location of the root.

   In this update, we begin by temporarily disregarding the root and randomly picking an interior edge e from the unrooted tree topology. Uniformly at random designate the end nodes of this edge as u and v. One of the other nodes adjacent to u is selected at random and designated a. Similarly, we select a neighbor d adjacent to v. See Figure 5.

   Next, the lengths of the edges on the path from a to d are multiplied by a random amount $r = \exp(\lambda(U_1 - 0.5))$ where $U_1 \sim \text{Unif}(0, 1)$ and λ is a tuning parameter. This multiplier r has density $g(r \mid \lambda) = 1/(\lambda r)$ for $\exp(-\lambda/2) < r < \exp(\lambda/2)$. Then, a new location is selected for node u uniformly at random on the path from a to d. If this new location is further from a than v is from a, the tree topology changes. Edges on the path from a to d change length, but all other tree edge lengths remain unchanged. No node connections other than in this local neighborhood change.

   The proposal density is

$$\frac{1}{E-1} \times \left(\frac{1}{2}\right)^3 \times \frac{1}{\lambda r} \tag{16}$$

   If the distances from a to u, v, and d are x, y, and z, respectively, the new locations are $x^* = U_2 r x$, $y^* = ry$, and $z^* = rz$. The reverse proposal would require $r^* = 1/r$ and $U_2^* = x/z$. The Jacobian (Green, 2003) for the corresponding bijection is r and the proposal ratio is

$$\frac{(1/(E-1))(1/2)^3/(\lambda r^*)}{(1/(E-1))(1/2)^3/(\lambda r)} \times r = r^3 \tag{17}$$

   When the root is on the path from a to d, there is another parameter for the distance of the root from node a and the Jacobian and proposal ratio

DOI: 10.2202/1544-6115.1152

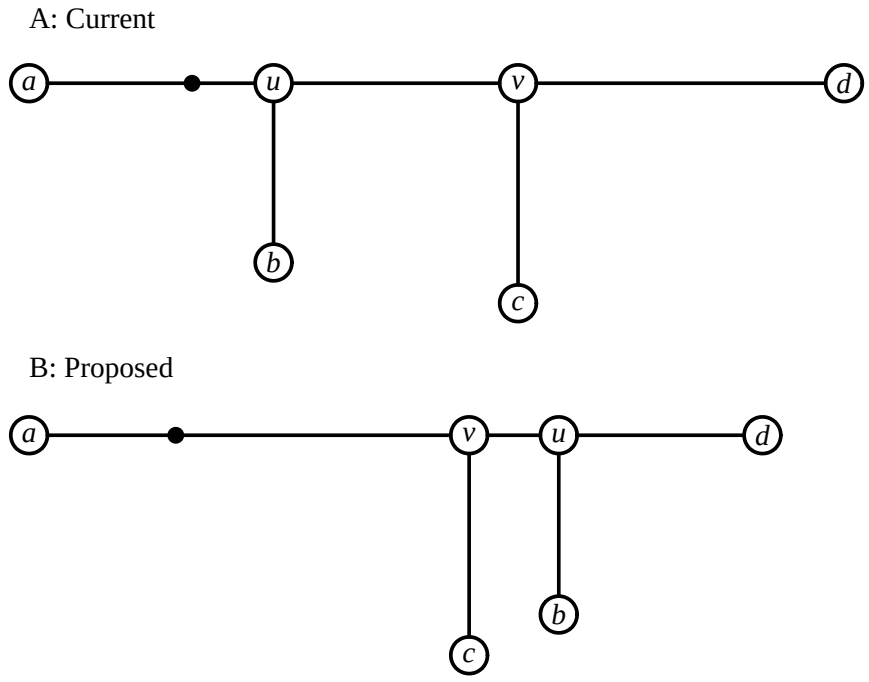

Figure 5: **Local Update.** A shows a local neighborhood of the tree before the proposal, B shows the same node after the proposal. The dark point represents a possible location of the root.

---

each have another factor of r. The prior ratio is a ratio of products of exponential densities which simplifies to $\exp(-(z^*-z)/\mu$ regardless of the position of the root. Putting this all together, we find the corresponding acceptance probability is

$$\begin{array}{ll} \min\{1, \exp((z-z^*)/\mu) r^3 \times LR\} & \text{if root is not on the path from } a \text{ to } d \\ \min\{1, \exp((z-z^*)/\mu) r^4 \times LR\} & \text{if root is on the path from } a \text{ to } d \end{array} \tag{18}$$

5. **Generalized local update.** This update generalizes local by picking a path between two leaves rather than to neighbors of an internal edge. Specifically, randomly pick two leaves, say a and b, and let $z = \text{dist}(a, b)$ be the distance of the path between them. Denote the number of nodes on the path from a and b not including nodes a and b as n_{ab}. Then, there are $n_{ab} + 1$ edges on the path with edge lengths $l_{ab}^{(i)}, i = 1, \cdots, n_{ab} + 1$. Stretch or shrink these edge lengths by a common multiplicative factor

$r = \mathrm{e}^{\lambda(U_1-0.5)}$, so that $z^* = z \times r$ with $(l_{ab}^{(i)})^* = l_{ab}^{(i)} \times r$. Randomly pick a non-root node between a and b and move it to a new location selected uniformly at random along the path, a distance $U_2 \times z^*$ from node a. The topology changes if the relative order of the nodes on the path changes. The acceptance probability is

$$\min\left\{1, \mathrm{e}^{-(z^*-z)/\mu}(r)^{n_{ab}+1} \times LR\right\} \ . \tag{19}$$

6. **SPR: subtree pruning and regrafting.** Pick an edge e whose parent node v is not the root uniformly at random (with probability $1/(E-2)$), remove it and the subtree rooted at its child node from the tree, joining the sibling edge of e and the parent edge of e into a single edge with length l_1. Next, pick an edge of the remainder of the tree uniformly at random and pick a location on this edge uniformly at random. Reconnect the subtree to the remainder of the tree by placing node v at this point, splitting the selected edge. Let l_2 be the length of the selected edge before it is split. As this proposal does not change the sum of all edge lengths in the tree, the prior ratio is 1. The proposal ratio is $\frac{1/l_1}{1/l_2} = \frac{l_2}{l_1}$. The acceptance probability is

$$\min\left\{1, \frac{l_2}{l_1} \times LR\right\} \ . \tag{20}$$

RESULTS AND DISCUSSION

We demonstrate the methodology presented in this paper on a data set consisting of 1394 AFLP markers from fourteen individuals from eight different sedge species. This data set is a subset of a larger data set published in Hipp *et al.* (2006). The taxa with number of individuals from each are: *Carex bebbii* (1), *Carex bicknellii* (1), *Carex festucacea* (2), *Carex normalis* (2), *Carex oronensis* (2), *Carex tenera* var. *echinodes* (2), *Carex tenera* var. *tenera* (2), and *Carex tincta* (2). The taxa chosen for this study represent a morphologically cohesive clade, with two closely-related taxa as outgroup (*C. bebbii* and *C. bicknellii*). Monophyly of the former is supported by neighbor joining (NJ) and minimum evolution (ME) analyses on an expanded dataset that includes all members of an eastern North American clade identified in a previous study using nuclear ribosomal DNA sequence data. Some of the relationships within the group, however, are not strongly supported using distance methods, which was one of the interests in exploring the phylogeny of this group using a more realistic model of character evolution.

DOI: 10.2202/1544-6115.1152

In the MCMC analysis, individuals from the same species were monophyletic (grouped together) with probability of at least 0.999. The tree topology with the single highest posterior probability, 0.688, is shown on the right of Figure 6. A 95% credible region included eight tree topologies.

The two most probable clades not found in our most probable tree had posterior probabilities greater than 0.08. Two additional clades had posterior probabilities between 0.06 and 0.08 while all other clades all had posterior probabilities less than 0.03. The most probable alternative clade includes *C. bebbii*, *C. bicknellii*, and *C. festucacea*, corresponding to the first four taxa in Figure 6, and has probability 0.156. We will call this clade Be/Bi/F. The next alternative clade includes *C. normalis*, *C. tenera* var. *echinodes*, and *C. tenera* var. *tenera*, the last six taxa in Figure 6. This clade, which we will denote as N/Te/Tt, has probability of 0.088. Each of these two alternative clades is more strongly supported by another method. We will compare and contrast our results with those from other methods next.

Comparison with Other Methods

The method we propose for analysis of AFLP marker data has some potential advantages over alternatives in that it is based on a molecular model for AFLP marker evolution which may promote more accurate phylogenetic reconstruction along with clearly interpretable measures of uncertainty.

Neighbor joining. Neighbor joining (NJ) is a general purpose hierarchical clustering method for inferring phylogenetic trees with edge lengths from pairwise distance data (Saitou and Nei, 1987). It is not based explicitly on any underlying assumptions of data generation, but will recover the correct tree topology when pairwise distances are additive, meaning that distances measured between pairs of taxa are equal to the sums of the lengths of branches in corresponding path connecting them in the tree (see also discussion in Gascuel and Steel (2006)).

To apply NJ to any 0/1 marker data including AFLP marker data, one first selects a distance measure. One such distance often used for restriction fragment data is the Nei-Li restriction site distance (Nei and Li, 1979) which is based on the Jukes-Cantor (1969) model of nucleotide substitution and assumes a fixed known size of restriction fragments.

To assess uncertainty in NJ analysis, one typically uses the bootstrap (Felsenstein, 1985). High bootstrap values indicate strong support, but bootstrap values are not directly comparable to Bayesian posterior probabilities (Alfaro *et al.*, 2003).

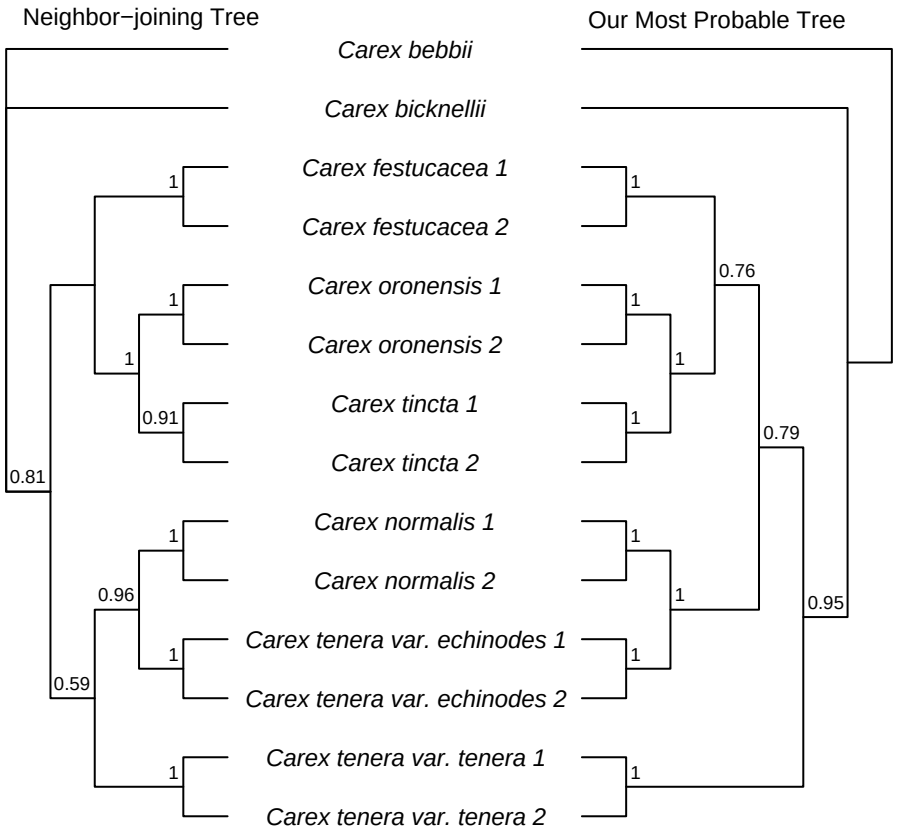

Figure 6: **Comparison of topologies.** The left tree topology is a neighbor joining tree based on Nei-Li restriction site distances (Hipp *et al.*, 2006). Numbers represent bootstrap support. The right tree topology has the greatest posterior probability using our Bayesian method, and numbers represent posterior probabilities of each clade. The trees are rooted such that *C. bebbii* and *C. bicknellii* are sister to the other species, based on previous work in the section.

DOI: 10.2202/1544-6115.1152

MrBayes. MrBayes implements a generic model for 0/1 marker data based on a continuous-time Markov process with two states. In particular, this model assumes alternating exponential dwell times in the two states and that the distributions are identical for all markers. While it may be reasonable (and in fact it is common) to model the substitution process of single nucleotides with a continuous-time Markov process, the genetic background in our paper illustrates that the process of AFLP marker gain and loss is complicated and depends on the interactions among many nucleotide sites. In the model we develop, the unobserved processes $M(t)$ and $Z(t)$ are Markov processes (approximately so for $Z(t)$), but the AFLP process $X(t)$ that indicates a marker presence is a function of these two independent processes, characterized by the relationship $X(t) = 1$ if and only if $M(t) = 0$ and $Z(t) = 0$. Thus the process $X(t)$ would be modeled more appropriately as a hidden Markov model (HMM). Functions of Markov chains or hidden Markov models are typically not Markovian. While we found the nominal HMM $Z(t)$ to be extremely well approximated by a two state Markov chain, the same result does not follow for $X(t)$. This observation implies that the 0/1 model implemented in MrBayes may be too simple to adequately model the actual AFLP marker evolution process.

Another implicit assumption in MrBayes is that all markers have the same stationary distribution for presence and absence. From our analysis based on the underlying genetic basis of AFLP marker evolution, we know that the distributions do, in fact, rely on the marker length and are not identical, and that distributions for very different lengths can be substantially different from each other. We note that the 0/1 model for marker data implemented in MrBayes is equivalent to the model in Mau and Newton (1997) for binary data.

Distinctive elements of our model. Our model generalizes the commonly used restriction site likelihood model (Smouse and Li, 1987; Felsenstein, 1992) by allowing the lengths of restriction sites to be among a set of values with known proportions *a priori* instead of fixed. In addition, our model includes substitutions in the intermediate region that can cause AFLP marker loss. Especially for long markers, such substitutions can be expected to cause a substantial proportion of events that lead to marker loss. The longest measurable fragments have an intermediate region length just under 600 while this length is 11 for the shortest measurable fragments. For an *Eco*RI/*Mse*I fragment with an end region of 16 bp, the total rate of substitutions that cause marker loss for the shortest fragments is approximately $16.1u$ where u is the

substitution rate per site and the rate of loss in the intermediate region is calculated according to Equation 5. For these short fragments, fewer than one percent of the substitutions that cause loss are expected to occur in the intermediate region. In contrast, for the longer fragments, the rate of loss would be about 50 percent higher, or about $25.1u$. For these fragments, over 35 percent of the substitutions causing marker loss are expected to be in the intermediate region.

Unlike other models for analyzing marker data, our method incorporates explicitly both the fragment length and the marker presence/absence information and so uses the available data more fully. If markers are gained and lost through the process of nucleotide substitution, it follows that markers associated with long fragments should be more readily lost as there are more possible sites for the introduction of new restriction sites in the interior of the fragment. Our method accounts for this and, in effect, gives greater weight to the information in long fragments than in shorter fragments. While our model makes use of this additional information, it is not clear how this use affects inference in general.

Comparison of results for sedges. To compare our method with NJ, we display the tree from pairwise Nei-Li distances on the left of Figure 6 opposite the most probable tree from our analysis. The NJ tree topology differs from ours in the placement of *C. tenera* var. *tenera* in clade N/Te/Tt. In our analysis, the complete NJ tree has posterior probability 0.013 and is the ninth most probable tree topology. Similarly, the consensus tree containing the most probable nonconflicting clades from MrBayes is shown in Figure 7. We find the posterior probability of the MrBayes consensus tree to be less than one percent.

Figure 7 (left) shows unrooted trees with estimated edge lengths from all three methods (NJ tree and consensus trees from both Bayesian methods). To make it easier to distinguish the three tree topologies, we also show the simplified trees on the right of the figure. All three trees group the individuals from the same taxa together, and all of them have clades {*C. oronensis* (2), *C. tincta* (2)} (denoted as O/Ti in Figure 7), {*C. normalis* (2), *C. tenera* var. *echinodes* (2)} (denoted as N/Te), {*C. tenera* var. *tenera* (2)} (denoted as Tt), {*C. festucacea* (2)} (denoted as F), and {*C. bebbii* (1), *C. bicknellii* (1)} (denoted as Be/Bi).

Relative edge lengths in each tree are very similar, but the topologies are different. The differences in the three tree topologies can be explained simply by different placement of the Be/Bi clade that contains the outgroup. The

DOI: 10.2202/1544-6115.1152

three topologies found by pruning the Be/Bi clade are identical. Our method finds fairly large differences in posterior probability for these alternative placements of the Be/Bi clade.

In their previous study of this group, Hipp *et al.* (2006) evaluated the strength of support for alternative outgroup placements by means of a paired sites test in a likelihood framework (Shimodaira and Hasegawa, 1999), using the standard model of restriction site evolution (implemented in Felsenstein's RESTML program of his PHYLIP package version 3.63; Felsenstein (1989)). This method of evaluating topologies fails to reveal statistically significant differences among among the three tree topologies in Figure 7, despite differences in the maximum log-likelihoods for each topology. Our Bayesian approach using a more sophisticated likelihood model describes the strength of evidence for and against alternative tree topologies with posterior probabilities, not p-values, and it is difficult to directly compare the two types of inference.

In the *Carex* section *Ovales* data set, our method does lead to different levels of quantitative support for some evolutionary relationships (clades) than do standard methods in the field. We can explain some of the differences among trees most strongly supported using the three methods by examining more closely the sites associated with clades that are supported differentially. For example, the only difference between the most probable tree topology in our analysis and the NJ tree is the relative placement of *C. tenera* var. *tenera*. We place the individuals from this species with *C. bebbii* and *C. bicknellii* in the unrooted tree with posterior probability 0.79. The NJ tree includes *C. tenera* var. *tenera* in a clade with *C. normalis* and *C. tenera* var. *echinodes*, albeit with bootstrap support 0.59. We find a posterior probability of 0.088 for this clade. A partial explanation is that the markers most favorable to a *C. tenera* var. *tenera*, *C. bebbii*, *C. bicknellii* clade are longer by about 30 bp, on average, than those that most favor a *C. tenera* var. *tenera*, *C. normalis*, *C. tenera* var. *echinodes* clade.

Computational Issues

Verification. We have implemented our new method in a program written in C++. We tested the implementation by running the MCMC process without data to simulate from the prior distribution on four taxon trees with independent and identically distributed exponential branch lengths. The sample was consistent with the prior distribution including a uniform distribution on topology and exponential branch lengths with the expected mean. Long simulations with posterior distributions based on observed data begun from disparate starting points are sufficiently consistent for us to be confident in

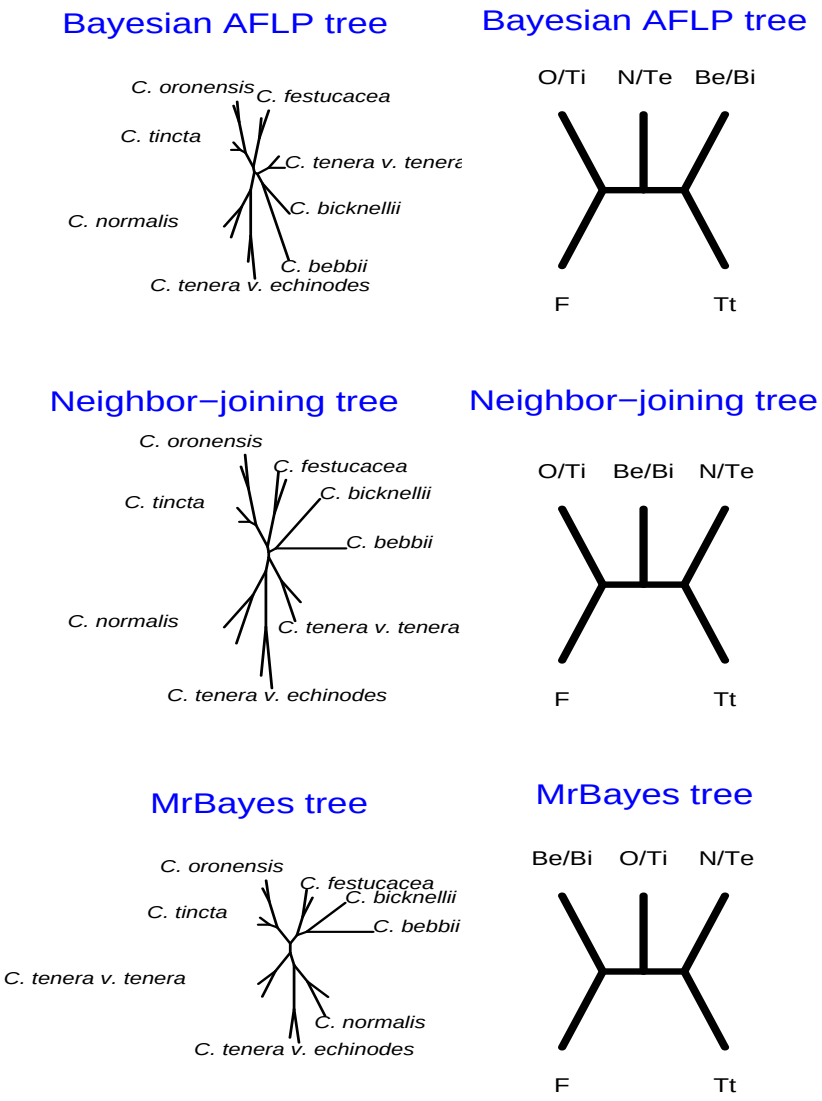

Figure 7: **Comparison of phylogenies.** The top phylogeny is the most probable tree topology using our Bayesian method. The middle phylogeny is the NJ phylogeny based on Nei-Li distances. The bottom phylogeny is a Bayesian phylogeny using a 0/1 model in MrBayes. The right part is a simplified illustration of trees on the left. See the explanation for notations from the text.

DOI: 10.2202/1544-6115.1152

the results presented.

Computational burden. The most important limitation of our approach is the enormous computational burden. The posterior probabilities we report in Figure 6 have Monte Carlo errors less than one percent, but these are based on averaging 20 MCMC runs, each of which ran for nearly two weeks on a standard desktop computer. Estimating a single tree using a distance approach is essentially instantaneous, and bootstrapping a fast method is also very fast. The analysis using MrBayes required less than ten minutes on a desktop computer. The primary reason for the discrepancy in time between MrBayes and our method is that the simpler model in MrBayes uses 2×2 matrices for each branch for likelihood calculations whereas our method requires calculations with both 34×34 and 38×38 matrices on each branch, for a combined total of over 600 times as many matrix elements. Furthermore, since our model uses marker lengths in the likelihood calculation, we must carry out a full computation for each marker whereas MrBayes takes advantage of sharing computations among markers with identical patterns.

So, while we feel that the statistical merit of our approach has much to recommend it, we have considerable work to do to improve its implementation to make it a practical tool for other scientists. Undoubtedly, a better implementation of our approach would result in substantially speedier calculations, but there is also clearly much room for improvement in the basic MCMC methods we have developed to this point.

Model assumptions. The assumption that different bands originate from different loci is a common simplification used in all current study, and we also take this simplification in our model. Actually, assuming only a substitution process, different loci in the ancestor having the same fragment length will produce different patterns of AFLP markers for the taxa. Our current method ignores the possibility that a marker with a given length is the superposition of bands from multiple loci. This possibility is especially problematic among plants where there has been recent polyploidy (although this is not the case in *Carex* section *Ovales*). We could potentially account for this by setting the probabilities that a particular band is from several different loci and calculating the total likelihood as a mixture of the likelihood for all loci.

In addition, insertion and deletion events have the potential to cause a single locus in an ancestral genome to appear as multiple separate AFLP markers with different fragment lengths in different taxa, hence causing the markers to be dependent. Thus, the first three assumptions about loci in our

model may not be met. If the indel rate is significant, ignoring it could cause us to overestimate the substitution rate because the substitution rate would need to account for changes due to both substitutions and indels. In addition, if multiple loci are superimposed on a single marker and we do not account for this, then the marker from combined loci could provide support for clades not in the true tree. We are currently extending the model described here to incorporate processes of insertion and deletion of bases.

For computational simplicity, we assume a Jukes-Cantor model of nucleotide substitution. The incorporation of more realistic nucleotide substitution models that allow different base composition and unequal rates of substitution among different bases will require new computational approaches as the mismatch process $M(t)$ will no longer be a Markov chain. This model extension would complicate the calculation by increasing the state space that the MCMC would have to traverse.

CONCLUSION

We have developed the first Bayesian approach specifically to model AFLP marker evolution. Our model is based on an understanding of the genetic processes at the nucleotide level that directly cause marker gain and loss. Specifically, we model observable AFLP markers as the realization of a hidden Markov model where unobserved DNA sequences subject to nucleotide substitution constitute the underlying Markov chain. With an understanding of AFLP fragments consisting of an end region and an intermediate region and by assuming the Jukes-Cantor model for nucleotide substitution, we simplify the underlying Markov chain as a pair of independent processes, a mismatch process in the end region and a two-state process for the intermediate region. The result is a computationally tractable yet highly detailed probability model for AFLP marker evolution that provides a framework for the analysis of AFLP marker data. Alternative methods of analysis are generic for 0/1 marker data and are not based on the specific genetic underpinnings of AFLP markers.

One advantage that our method possesses is the ability to use additional information beyond simple marker presence/absence. Our model takes into account marker fragment lengths, which we have shown in one example can lead to different statistical inferences. While we expect in general that methods based on models that are more closely descriptive of the actual underlying biology may lead to more accurate inferences and we are hopeful that our method will enjoy this advantage in comparison to other existing alternatives, it remains an open question to assess the relative accuracy of our model in a variety of settings.

DOI: 10.2202/1544-6115.1152

There is a significant computational cost to using our method. In the current implementation, analysis of moderate sized data sets can require weeks of computer run time whereas alternative methods are much faster. There is clearly room for algorithmic development as well as improvements in software implementation of these ideas that could lower the computational burden in using our method. In addition, we are currently working on improving and extending the model to accounting for alternative genetic processes of AFLP marker evolution such as insertion and deletion processes as well as accounting for superposition of AFLP markers from multiple loci.

While there remain many interesting challenges, this paper describes a significant methodological advance in the analysis of AFLP marker data to infer phylogenetic trees.

APPENDIX

It's known that for a two-state Markov Chain, the transient transition matrix is of the form

$$Q = \begin{pmatrix} -q_{01} & q_{01} \\ q_{10} & -q_{10} \end{pmatrix}.$$

Q has eigenvalues 0 and $-(q_{01} + q_{10})$, with corresponding eigenvectors $(1, 1)^T$ and $(-q_{01}, q_{10})^T$. We want the decomposition of $Q = A\Lambda A^{-1}$, where A has the form

$$A = \begin{pmatrix} 1 & -x \\ 1 & y \end{pmatrix},$$

the first row of $A^{-1} = (\pi_0, 1 - \pi_0)$ is the stationary distribution, and where

$$\Lambda = \begin{pmatrix} 0 & 0 \\ 0 & -(q_{01} + q_{10}) \end{pmatrix}.$$

Computing A^{-1} directly we find

$$A^{-1} = \begin{pmatrix} \frac{y}{x+y} & \frac{x}{x+y} \\ -\frac{1}{x+y} & \frac{1}{x+y} \end{pmatrix}.$$

It follows that $\pi_0 = \frac{y}{x+y}$ and $\frac{x}{y} = \frac{q_{01}}{q_{10}}$, and so $y = q_{10} = \frac{\pi_0}{1-\pi_0} q_{01}$ and $x = q_{01}$. The probability transition matrix is

$$P = e^{Qt} = Ae^{\Lambda t}A^{-1} = \begin{pmatrix} \pi_0 + (1 - \pi_0)\eta(t) & (1 - \pi_0)(1 - \eta(t)) \\ \pi_0(1 - \eta(t)) & (1 - \pi_0) + \pi_0\eta(t) \end{pmatrix},$$

where

$$\eta(t) = \exp\left(-\frac{q_{01}t}{1 - \pi_0}\right).$$

REFERENCES

Albertson, R. C., J. A. Markert, P. D. Danley, and T. D. Kocher (1996). Phylogeny of a rapidly evolving clade: The cichlid fishes of Lake Malawi, East Africa. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 5107–5110.

Alfaro, M. E., S. Zoller, and F. Lutzoni (2003). Bayes or Bootstrap? a simulation study comparing the preformance of Bayesian Markov Chain Monte Carlo sampling and Bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution* **20**: 255–266.

Drozdek, A. (2001). *Data Structures and Algorithms in Java*, pages 214–215. Brooks/Cole, Pacific Grove, CA 93950.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Biology* **17**: 368–376.

— (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**: 783–791.

— (1989). PHYLIP: Phylogeny inference package (Version 3.2). *Cladistics* **5**: 164–166.

— (1992). Phylogenies from restriction sites: a maximum-likelihood approach. *Evolution* **46**: 159–173.

— (2004). *Inferring phylogenies*, chapter 15, pages 239–241. Sinauer Associates, Inc., Sunderland, Massachusetts.

Gascuel, O. and M. Steel (2006). Neighbor-Joining revealed. *Molecular Biology and Evolution* **23**: 1997–2000.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–732.

— (2003). Trans-dimensional Markov chain Monte Carlo. In P. J. Green, N. L. Hjort, and S. Richardson, editors, *Highly Structured Stochastic Systems*, pages 179–196. Oxford University Press.

Hartl, D. and E. Jones (1998). *Genetics: Principles and Analysis*, chapter 5, pages 207–208. Jones and Bartlett Publishers, fourth edition.

DOI: 10.2202/1544-6115.1152

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* .

Hipp, A. L., P. E. Rothrock, A. A. Reznicek, and P. E. Berry (2006). Chromosome number changes associated with speciation in sedges: A phylogenetic study in Carex section Ovales (Cyperaceae) using AFLP data. In J. T. Columbus, E. A. Friar, J. M. Porter, L. M. Prince, and M. G. Simpson, editors, *Monocots: Comparative biology and evolution*. Rancho Santa Ana Botanic Garden, Claremont, CA.

Holder, M., P. Lewis, D. Swofford, and B. Larget (2005). Hastings ratio of the local proposal used in bayesian phylogenetics. *Systematic Biology* **54**: 961–965.

Huelsenbeck, J. P. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Innan, H. R., G. Terauchi, G. Kahl, and F. Tajima (1999). A method for estimating nucleotide diversity from AFLP data. *Genetics* **151**: 1157–1164.

Jones, C. J., K. J. Edwards, S. Castaglione, M. O. Winfield, F. Sala, C. van de Wiel, G. Bredemeijer, B. Vosman, M. Matthes, A. Daly, R. Brettschneider, P. Bettini, M. Buiatti, E. Maestri, A. Malcevschi, N. Marmiroli, R. Aert, G. Volckaert, J. Rueda, R. Linacero, A. Vazquez, and A. Karp (1997). Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. *Molecular Breeding* **3**: 381–390.

Jukes, T. H. and C. R. Cantor (1969). *Mammalian protein metabolism*, volume 3, chapter Evolution of protein molecules, pages 21–132. Academic Press.

Landry, P. A. and F. J. Lapointe (1996). RAPD problems in phylogenetics. *Zoologica Scripta* **25**: 283–252.

Larget, B. and D. L. Simon (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* **16**: 750–759.

Mau, B. and M. A. Newton (1997). Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics* **6**: 122–131.

Metropolis, N. A., A. W. Rosenbluth, M. N. Rosenbluth, A. W. Teller, and E. Teller (1953). Equations of state calulations by fast computing machines. *Journal of Chemical Physics* **21**: 1087–1092.

Mueller, U. G. and L. L. Wolfenbarger (1999). AFLP genotyping and fingerprinting. *Trends in Ecology and Evolution* **14**: 389–394.

Nei, M. and W.-H. Li (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. In *Proceedings of the National Academy of Sciences*, volume 76, pages 5269–5273.

Powell, W., M. Morgante, C. Andre, M. Hanafey, J. Vogel, S. Tingey, and A. Rafalski (1996). The comparison of RFLP, RAPD, AFLP and SSR markers for germplasm analysis. *Molecular Breeding* **2**: 225–238.

Rzhetsky, A. and M. Nei (1992). A simple method for estimating and testing minimum-evolution trees. *Molecular Biology Evolution* **9**: 945–967.

Saitou, N. and M. Nei (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**: 406–425.

Shimodaira, H. and M. Hasegawa (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* **16**: 1114–1116.

Smouse, P. E. and W. H. Li (1987). Likelihood analysis of mitochondrial restriction-cleavage patterns for the human-chimpanzee-gorilla trichotomy. *Evolution* **41**: 1162–1176.

Vekemans, X., T. Beauwens, M. Lemaire, and I. Roldan-Ruiz (2002). Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. *Molecular Ecology* **11**: 139–151.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and M. Zabeau (1995). AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Research* **23**: 4407–4414.

Wolfe, A. D. and A. Liston (1998). *Molecular systematics of plants II*, chapter 2, pages 43–86. Kluwer Academic Pub.

DOI: 10.2202/1544-6115.1152