Novel Distances for Dollo Data

MICHAEL WOODHAMS\textsuperscript{1}, DOROTHY A. STEANE\textsuperscript{2,3,4}, REBECCA C. JONES\textsuperscript{3}, DEAN NICOLLE\textsuperscript{5}, VINCENT MOULTON\textsuperscript{6}, AND BARBARA R. HOLLAND\textsuperscript{1}\textsuperscript{*}

\textsuperscript{1}School of Mathematics and Physics; \textsuperscript{2}CRC for Forestry; \textsuperscript{3}School of Plant Science, University of Tasmania, Private Bag 55, Hobart 7001, Australia; \textsuperscript{4}Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, Maroochydore DC, Queensland 4558, Australia; \textsuperscript{5}Currency Creek Arboretum, P.O. Box 808, Melrose Park, South Australia 5039, Australia and \textsuperscript{6}School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK.

\textsuperscript{*}Correspondence to be sent to: School of Mathematics and Physics, University of Tasmania, Private Bag 37, Hobart, Tasmania 7001, Australia; E-mail: barbara.holland@utas.edu.au.

Received 29 February 2012; reviews returned 24 April 2012; accepted 8 August 2012

Associate Editor: Lars Fjérnin

Abstract.—We investigate distances on binary (presence/absence) data in the context of a Dollo process, where a trait can only arise once on a phylogenetic tree but may be lost many times. We introduce a novel distance, the Additive Dollo Distance (ADD), that applies to data generated under a Dollo model and show that it has some useful theoretical properties including an intriguing link to the LogDet/paralinear distance. Simulations of Dollo data are used to compare a number of binary distances including ADD, LogDet, a restriction-site-based distance, and some simple, but to our knowledge previously unstudied, variations on common binary distances. The simulations suggest that ADD outperforms other distances on Dollo data. Interestingly, we found that the LogDet distance performs poorly in the context of a Dollo process; this may have implications for its use in connection with conditioned genome reconstruction. We apply the ADD to two Diversity Arrays Technology data sets, one that broadly covers Eucalyptus species and one that focuses on the Adnataria phylogeny; conditioning genomes; Diversity Arrays Technology; Dollo process; Eucalyptus phylogeny; gene content phylogeny; LogDet/paralinear distances.

A fundamental idea in evolutionary biology is that when two species share a complex trait, then the most likely explanation of the similarity is that both species have inherited the trait from a common ancestor. However, the absence of a particular trait carries far less information. For instance, wings and eyes are complex traits that have been lost many times independently in different parts of the evolutionary tree of life. As long ago as 1893 (Dollo 1893), Louis Dollo captured this idea in what is now known as Dollo’s Law, it states that complex traits can be gained only once somewhere in evolutionary history, but may be subsequently lost independently many times.

In this article, we will only consider Dollo models that generate binary data, recording the presence or absence of some trait. For example, does an organism have any genes in a particular gene family? Does it have some skeletal feature, such as the mammalian inner ear? When its DNA is digested by a mix of restriction enzymes, is a particular DNA fragment produced? In reality, determining whether a trait is present or absent can be less clear cut; for example, paralogous genes can confound gene presence/absence decisions. In a stochastic Dollo model, the gain and loss of such traits is treated as arising from a simple probability model. Although few situations match Dollo’s Law exactly, it provides a useful model in many evolutionary scenarios of interest. Dollo models have been used to understand gene families (Huson and Steel 2004; Dagan and Martin 2007) and complex morphological traits (Gould 1970), and stochastic Dollo models have been used to study cognates in language evolution (Nicholls and Gray 2008; Ryder and Nicholls 2011).

In some of these scenarios, such data can have another interesting property: only traits that are present in particular reference taxa are visible; or, in other words, the data are censored. This happens, for example, with array-based studies where a small set of taxa is used to create a set of traits (i.e., a set of DNA fragments that make up an array) to which other taxa can be compared. The idea of a reference taxon also has parallels in the gene-content setting, where some authors have proposed “conditioned genome reconstruction” (Lake and Rivera 2004), here one genome is selected as a reference and, for the remaining genomes, only gene families present in the reference genome are analyzed. Data thought to follow Dollo’s Law traditionally have been analyzed using a parsimony approach (Queens 1974; Farris 1977). As is normally the case with parsimony approaches, branch length information is not taken into account. The use of stochastic Dollo models is relatively recent (Alekseyenko et al. 2008; Nicholls and Gray 2008; Ryder and Nicholls 2011) and, so far, they have only been implemented in a Bayesian framework. Bayesian methods are computationally intensive, so there is a need for an approach that is both computationally efficient and statistically consistent.

This motivated us to develop a distance-based approach to Dollo data. Our initial motivation was to derive a distance suitable for phylogenetic analysis of Diversity Array Technology (DArT) data (Jaccoud
et al. 2001) which, by its nature, is censored. On further consideration, we realized that the same formula can be derived directly from the mathematics of the stochastic Dollo process, or as a limiting case of the LogDet distance (Lockhart et al. 1994) and the essentially equivalent paralinear distance (Lake 1994).

In the following sections, we begin by deriving the Additive Dollo Distance (ADD) in a general Dollo context and then show why it also applies to censored Dollo data. We then describe an intriguing link to the popular LogDet distances. After introducing a few other binary distances, we present a simulation study that compares the performance of the new ADD with other binary distances when applied to Dollo data under a range of censoring schemes. As an illustration of our approach, we apply the new distance to three case studies, two involving DArT data for Eucalyptus species and one using gene-content information. We conclude with a discussion in which we point out some potential future directions.

METHODS

Deriving an Additive Distance for the Stochastic Dollo Process

Our description of a stochastic Dollo process follows that of Huson and Steel (2004), whose discussion is in the context of the gene content of a genome. We note that the Dollo process when viewed from the point of view of a particular trait is certainly not a Markov process, as it has zero probability, whereas the sequence of transitions from present→absent→present has nonzero probability. However, the model of Huson and Steel (2004) applies at the level of the total number of genes and can be described as a constant-birth, proportional-death Markov process. New markers are acquired (e.g., genes added to the genome) at a rate λ, and existing markers are independently deleted at a rate μ. G(t) is the set of markers present at time t. We make an initial observation of the set of markers G(s) at start time s. It is assumed that the system is at equilibrium at time s (i.e., it has been evolving by the stochastic Dollo process for long enough to be independent of initial conditions). The genome then evolves for a further time t and we observe the marker set G(s+t). We then put

\[ n_{11} = |G(s) \cap G(s+t)| \] (1)
\[ n_{10} = |G(s) - G(s+t)| \]
\[ n_{01} = |G(s+t) - G(s)| , \]
where \( n_{11} \) is the number of shared presences (markers present in the genome at both time points), \( n_{10} \) is the number of markers present at time s but not at time s + t, and \( n_{01} \) is the number of markers present at time s + t but not at time s.

Under these circumstances, Huson and Steel (2004) prove the following facts:

1. Defining \( l(s) = |G(s)| \), they show \( l(s) \) is Poisson distributed with mean \( m = \lambda/\mu \).
2. If \( l(0) \) is chosen according to this equilibrium Poisson distribution, then the process underpinning \( G(t) \) is a time-reversible Markov process.
3. \( N_{00} \) is Poisson distributed with mean \( m(1-e^{-\mu t}) \).
4. \( N_{10} \) is binomially distributed with \( l(s) \) trials each with probability \( 1-e^{-\mu t} \) of success, where a “success” in this context is the loss of a marker.

From 4, we derive the expected value \( E[N_{10}/(N_{11} + N_{10})] = (1-e^{-\mu t}) \) and we solve for \( t \):

\[ t = -\frac{1}{\mu} \log \left( 1 - E \left[ \frac{N_{10}}{N_{11} + N_{10}} \right] \right) \] (2)

and so, substituting the observed values \( N_{10} \) for the random variables \( N_{10} \), we can derive a distance \( d \) that is an estimate for \( \mu t \) by

\[ d = -\log \left( 1 - \frac{N_{10}}{n_{10} + n_{10}} \right) = \log \left( \frac{n_{10} + n_{10}}{n_{10}} \right) \] (3)

Note that we could equally well use \( d = \log((n_{11} + n_{10})/n_{11}) \). To make use of all available data (both \( n_{10} \) and \( n_{01} \)), we add these two distances to give

\[ d_{ADD} = \log \left( \frac{(n_{11} + n_{10})(n_{11} + n_{01})}{n_{11}} \right) \] (4)

and call this the Additive Dollo Distance.

Dollo Models with Censored Data

As mentioned in the “Introduction” section, some data sets of interest have an additional property whereby only markers that are present in the reference taxon or taxa can be detected. This is referred to as “censored data” or as an “ascertainment bias”. We want to extend the ADD distance to cases of both single and multiple reference taxa. As we will see below, the first case does not require any change in the formula.

We introduce our approach for analysis of censored data in the context of DArT data sets (Jacquod et al. 2001). DArT uses particular restriction enzymes to create a “genomic representation” (DNA fragments typically 300–1000 bp in length) from one or more reference taxa. The restriction enzymes used consist of a “rare cutter” (e.g., PstI, that cuts at the 6-bp sequence CTGCA) and a “frequent cutter” (e.g., BstNI, that cuts at the 4-bp sequence TCGA). Importantly, only fragments cut at both ends by the rare cutter are amplified and can become markers. These fragments are cloned and are
then arrayed onto a glass microscope slide. Genomic representations are prepared for study samples using the same restriction enzymes. The study samples are screened (via DNA–DNA hybridization) with the array. The presence or absence in the sample of the DArT markers on the slide is recorded to produce a binary data set.

A DArT marker can be lost during evolution by mutations that disrupt the “rare” 6-bp cutter target at either end of the marker, or by mutations that introduce a new “frequent” or “rare” target within the marker. Once lost, a marker can only be regained by reversing the mutation that caused the loss, before another loss-causing mutation occurs. This is a rare event, so we can model DArT marker gain/loss as a Dollo process.

The data from a DArT analysis suffer from ascertainment bias: only markers that are present in the reference taxa. Furthermore, the Nei–Li distance does not produce, but for fragments generated by a single data set.

We summarize the presence/absence data for the study samples using DArT. Furthermore, the Nei–Li distance does not produce, but for fragments generated by a single restriction enzyme. This is a rare event, so we can model DArT marker gain/loss as a Dollo process.

The data from a DArT analysis suffer from ascertainment bias: only markers that are present in the reference taxa. Furthermore, the Nei–Li distance does not produce, but for fragments generated by a single restriction enzyme. This is a rare event, so we can model DArT marker gain/loss as a Dollo process.

The data from a DArT analysis suffer from ascertainment bias: only markers that are present in the reference taxa. Furthermore, the Nei–Li distance does not produce, but for fragments generated by a single restriction enzyme. This is a rare event, so we can model DArT marker gain/loss as a Dollo process.

The data from a DArT analysis suffer from ascertainment bias: only markers that are present in the reference taxa. Furthermore, the Nei–Li distance does not produce, but for fragments generated by a single restriction enzyme. This is a rare event, so we can model DArT marker gain/loss as a Dollo process.

The data from a DArT analysis suffer from ascertainment bias: only markers that are present in the reference taxa. Furthermore, the Nei–Li distance does not produce, but for fragments generated by a single restriction enzyme. This is a rare event, so we can model DArT marker gain/loss as a Dollo process.

The data from a DArT analysis suffer from ascertainment bias: only markers that are present in the reference taxa. Furthermore, the Nei–Li distance does not produce, but for fragments generated by a single restriction enzyme. This is a rare event, so we can model DArT marker gain/loss as a Dollo process.

The data from a DArT analysis suffer from ascertainment bias: only markers that are present in the reference taxa. Furthermore, the Nei–Li distance does not produce, but for fragments generated by a single restriction enzyme. This is a rare event, so we can model DArT marker gain/loss as a Dollo process.

The data from a DArT analysis suffer from ascertainment bias: only markers that are present in the reference taxa. Furthermore, the Nei–Li distance does not produce, but for fragments generated by a single restriction enzyme. This is a rare event, so we can model DArT marker gain/loss as a Dollo process.

The data from a DArT analysis suffer from ascertainment bias: only markers that are present in the reference taxa. Furthermore, the Nei–Li distance does not produce, but for fragments generated by a single restriction enzyme. This is a rare event, so we can model DArT marker gain/loss as a Dollo process.

The data from a DArT analysis suffer from ascertainment bias: only markers that are present in the reference taxa. Furthermore, the Nei–Li distance does not produce, but for fragments generated by a single restriction enzyme. This is a rare event, so we can model DArT marker gain/loss as a Dollo process.

The data from a DArT analysis suffer from ascertainment bias: only markers that are present in the reference taxa. Furthermore, the Nei–Li distance does not produce, but for fragments generated by a single restriction enzyme. This is a rare event, so we can model DArT marker gain/loss as a Dollo process.
unbiased sample from the markers present in that reference taxon. In practice, this assumption may fail for two reasons. When constructing the DArT array, we attempt to eliminate redundancy, so a marker selected for the array from one reference taxon precludes the same marker being selected from a second reference taxon in which it may also be present. (Some redundancy is kept deliberately as an internal control, but omitted from the final character matrices.) Also, markers are chosen that show useful levels of polymorphism. A marker present in all taxa is not useful, nor is one that is present only in the reference taxon from which it was derived.

Links to Conditioned Genome Reconstruction: The Additive Dollo Distance as a Limiting Case of LogDet

Another context in which Dollo models may be appropriate is gene-family presence/absence data. Such data are increasingly available as more and more genomes are studied. The COG database (Tatusov et al. 2003) sorts genes from 50 bacteria, 13 archaea, and 3 eukaryotes into nearly 5000 gene families. The gene family presence/absence data have been used for phylogenetic inference by several authors (Lake and Rivera 2004; Spencer et al. 2007; Cotton and McInerney 2008; Sangaralingam et al. 2010), but Dollo models have not, to our knowledge, been applied. A problem noted by previous analysts of these data is that the LogDet distance treats shared absences as if they needed to know the number of shared absences distance to data from the COG database, they realized biases, which is an issue the LogDet distance (Lake 1994; Lockhart et al. 1994) was designed to overcome. Indeed, it seems overwhelmingly likely that a significant amount of HGT must occur. For instance, using a Dollo model, Dagan and Martin (2007) showed that unless we are prepared to accept ancestral genome sizes much larger than observed modern genomes, HGT is required to explain patterns of gene presence and absence. However, that the LogDet distance treats shared presences and absences (\(n_00\) and \(n_{11}\)) symmetrically (i.e., swapping the values of \(n_{00}\) and \(n_{11}\) in the LogDet formula does not change its value), when physically they have very different meaning, seems to us to be a potential weakness in their method that has not been commented on previously.

Lake and Rivera’s (2004) assumption (that any gene family can be gained by any taxon at any time) can be thought of as one extreme. The opposite extreme is to discount the possibility of HGT completely and adopt a Dollo model. Consider the standard (non-Dollo) two-state Markov model: we have \(N_0\) characters evolving independently between two states (“present” and “absent”) by a continuous time Markov process. The rate of absent→present transitions is \(\lambda\) and the rate of present→absent transitions is \(\mu\), so the rate matrix is \((\begin{array}{cc} -\lambda & \lambda \\ \mu & -\mu \end{array}\)). As before, we sample this process at two different times and get counts \(n_{10}, n_{11}\) and \(n_{00}\) of shared presences and of presences at one time point but not the other. In addition, we get \(n_{00}\), the number of shared absences. In this two-state case, the LogDet distance formula (following the formulation of equation (3) from Lockhart et al. 1994) is

\[
d_{\text{LogDet}} = -\frac{1}{2} \log \left( \frac{n_{00}n_{11} - n_{01}n_{10}}{n_{00}n_{11} + n_{10}(n_{00} + n_{01}) + n_{11}(n_{10} + n_{11})} \right).
\] (9)

Now consider a Markov process with \(N\) characters and rate matrix \((\begin{array}{cc} -\lambda N/2 & \lambda N/2 \\ \mu N/2 & -\mu N/2 \end{array}\)). For \(N=N_0\) this is the process we considered earlier, and at equilibrium the mean number of present characters is \(N_0\lambda/N\mu+\lambda\). We now take the limit \(N \to \infty\), with the additional condition that the initial state of the process has a finite number of present characters. This is a Dollo process: the chance for any character to transition from absent to present is infinitesimal, so no character will undergo this transition twice in a finite period of time. The mean number of present characters at equilibrium is \(N_0\lambda/N\mu\), which is finite, and (by assumption) the initial number of present characters is finite. As \(n_{10} + n_{11}\) is the number of present

---

**FIGURE 1. Multiple reference taxa.** Each array of characters represents the 60 DArT markers that are present at either of the reference taxa R or S (note that these should be read as 60 characters on an array not as alignments). The character arrays at each node indicate the presence (1) or absence (0) of DArT markers. There is a 0.5 probability of marker loss along any edge. Each character (the 1s and 0s for a given marker) occurs a number of times proportional to its probability.
characters at one point in time, \( n_{11} \) and \( n_{10} \) (and by similar argument \( n_{01} \)) are finite (and in equilibrium distributed as for the stochastic Dollo process described above) so \( n_{00} \) must be infinite. Letting \( n_{00} \to \infty \) in equation (9).

\[
\lim_{n_{00} \to \infty} d_{\log \text{Det}} = \lim_{n_{00} \to \infty} \frac{1}{2} \log \left( \frac{n_{00}n_{11}}{\sqrt{n_{00}(n_{00} + n_{11})n_{00}(n_{10} + n_{11})}} \right) = \frac{1}{4} \log \left( \frac{(n_{11} + n_{10})(n_{11} + n_{01})}{n_{11}} \right) = d_{\text{ADD}} \tag{10}
\]

showing that in this limit, \( d_{\log \text{Det}} \) is proportional to \( d_{\text{ADD}} \). We note that the paralinear distance of Lake (1994) differs from the formula given in Lockhart et al. (1994) only by a factor of \( \frac{1}{2} \), where \( r \) is the number of states (here \( r = 2 \)), so that proportionality holds for the paralinear distance as well.

### Comparison of Binary Distances

Various distances have been defined in the literature for presence/absence data (see Choi et al. (2010) for a comprehensive review). We have picked a number of these to compare with \( d_{\text{ADD}} \) including:

- **Fractional Hamming distance**
  \[ d_{\text{Ham}} = \frac{n_{10} + n_{01}}{n_{00} + n_{01} + n_{10} + n_{11}}. \]

- **Jaccard distance**
  \[ d_{\text{Jac}} = \frac{n_{10} + n_{01}}{n_{00} + n_{01} + n_{10} + n_{11}}. \]

- **Dice Distance**
  \[ d_{\text{Dice}} = \frac{n_{01} + n_{10}}{n_{00} + n_{10} + 2n_{11}}. \]

- **LogDet distance**
  \[ d_{\log \text{Det}} = \frac{1}{2} \log \left( \frac{n_{00}(n_{11} - n_{01}n_{10})}{\sqrt{n_{00}(n_{00} + n_{11})(n_{00} + n_{01})(n_{10} + n_{11})}} \right). \]

- **Nei-Li distance**
  \[ d_{\text{NL}} = -\log(F), \]

  where \( F = \frac{p^4}{3 - 2p} \) and \( P = \frac{2n_{11}}{2n_{11} + n_{10} + n_{01}}. \)

The first three of these distances were introduced in the articles Hamming (1950), Jaccard (1901), and Dice (1945), respectively. The LogDet formula above is derived from equation (3) of Lockhart et al. (1994). The restriction site distance was introduced in Nei and Li (1979). Huson and Steel (2004) derived a maximum-likelihood distance for gene presence/absence data under a Dollo process

\[
d = -\log \left( \frac{\beta + \sqrt{\beta^2 + 4\alpha_2}}{2} \right) \tag{11}\]

where in our notation \( \beta = 1 - (n_{11} + n_{10} + n_{01})/m \), \( \alpha_2 = n_{11}/m \), and \( m = n/ \mu \) is the expected number of genes per genome. We do not know \( m \), but if we approximate it by the mean number of genes/markers at the two taxa \( m = (n_{11} + n_{10})/(n_{11} + n_{01})/2 \), then equation (11) simplifies to

\[
d = \log \left( \frac{2n_{11} + n_{01} + n_{10}}{2n_{11}} \right) \tag{12}\]

This is a simple transformation of the Dice distance, being \(-\log(1 - d_{\text{Dice}})\), so we name it the \( \log \text{Dice} \) distance. We can intuitively justify this transformation, arguing that logarithms correct for multiple events (e.g., gain, loss, mutation) on the same marker. We can also perform a similar transformation on the Jaccard distance to create the \( \log \text{Jaccard} \) distance. So in summary, in addition to the standard distances above, we introduce previously unstudied distances:

\[
d_{\log \text{Jac}} = -\log(1 - d_{\text{Jac}}) = \log \left( \frac{n_{11} + n_{01} + n_{10}}{n_{11}} \right), \]

\[
d_{\log \text{Dice}} = -\log(1 - d_{\text{Dice}}) = \log \left( \frac{2n_{11} + n_{01} + n_{10}}{2n_{11}} \right), \]

\[
d_{\text{ADD}} = \log \left( \frac{(n_{11} + n_{10})(n_{11} + n_{01})}{n_{11}} \right), \]

as well as the composite \( d_{\text{ADD}} \) distance method defined above.

### The Triangle Inequality and Additivity

Two important properties of distance functions (i.e., bivariate, non-negative functions \( d(x, y) \) with \( d(x, y) = 0 \) if and only if \( x = y \), and \( d(x, y) = d(y, x) \) for all \( x, y \)) are the triangle inequality and additivity. The triangle inequality states that \( d(x, y) \leq d(x, z) + d(z, y) \) must hold for all \( x, y, z \) (in which case \( d \) is known as a “metric”). Additivity states that if \( z \) was the last common ancestor of \( x \) and \( y \), and the sequences evolved independently, then \( d(x, y) = d(x, z) + d(z, y) \) should hold (on average). The desire for additivity accounts for the presence of the logarithm function in many phylogenetic distances.

Not all of the distances defined above satisfy the triangle inequality --- see Table 1 for counterexamples to the triangle inequality holding for some of the distances. In Table 2, we summarize which distances are additive and which obey the triangle inequality (Hamming 1950; Lipkus 1999). The additive Dollo distance \( d_{\text{ADD}} \) is additive by construction in the stochastic Dollo context, and it is a limiting case of the LogDet distance, which is additive (Lake 1994). Notably, the only two distances \( d_{\text{Ham}} \) and \( d_{\text{Jac}} \) known to obey the triangle inequality...
are not additive, and the only two distances known to be additive ($d_{\text{LogDet}}$ and $d_{\text{ADD}}$) violate the triangle inequality. Phylogeneticists appear to place greater value on additivity than on obeying the triangle inequality, as demonstrated by the popularity of LogDet.

It is worth noting that for $d_{\text{LogJac}}(x, y)$ and $d_{\text{LogDice}}(x, y)$ to be additive, it is necessary that they go to infinity as the evolutionary distance between $x$ and $y$ goes to infinity. For the stochastic Dollo process, $n_{11} = 0$ for infinitely separated $x$ and $y$, so $d_{\text{LogJac}}$ and $d_{\text{LogDice}}$ go to infinity as required, but for a Markov process where $E(n_{11}) > 0$ for unrelated $(x, y)$, $d_{11}$ and $d_{\text{LogDice}}$ will tend to a finite limit. We can formalize the formulae for correct this. In the case of LogJaccard, we have

$$d_{\text{LogJac}}(x, y) = -\log(b - d_{\text{Jac}}(x, y)),$$

where $b$ is the expected value of $d_{\text{Jac}}$ evaluated after infinite time (and a similar formula applies for $d_{\text{LogDice}}$). For example, for a Markov process where states 0 and 1 are equally likely at equilibrium, we have $b = 2/3$ for $d_{\text{LogJac}}$ and $b = 1/2$ for $d_{\text{LogDice}}$ when $x \neq y$. Note that we do not use such a correction here.

**Simulating Censored Dollo Data**

The general scheme for our simulations was to create a random tree, simulate a Stochastic Dollo process along it, select reference taxa, and, finally, select the markers that form the character matrix (on the basis of which markers are present at the reference taxa).

We generated clock-like and nonclock-like trees. For the nonclock-like trees, we generated the tree topology by a Yule process (Yule 1924), then branch lengths were set so that they were distributed uniformly between lengths 0.05 and 0.40. For clock-like trees, we generated a tree by a Yule process with mean branch length 0.1 and repeated this process until we obtained a tree whose shortest branch was no shorter than 0.01. (As short branch lengths are hard to resolve no matter how good the phylogenetic method, keeping such branches reduces the contrast between “good” and “poor” methods, this would make our simulation results harder to interpret.) Our simulated data were based on both 9- and 15-taxon trees.

For a given simulation run, we specified the expected number of markers per genome, $m$. The process is started at an arbitrary taxon, with the number of markers at that taxon drawn from a Poisson distribution with mean $m$, (i.e., we assume that the Markov process that describes the total number of markers is at equilibrium). Then, we propagated the set of markers through the tree. On each branch with length $b$, each existing marker was lost with probability $1 - e^{-b}$ and the number of new markers created has a Poisson distribution with mean $m(1 - e^{-b})$.

We used a number of different models for selecting the markers to be included in the character matrix, and (for the PADD method) how the markers are partitioned.

**incl1** (One reference taxon, included.) One taxon was chosen as a reference. Only markers present in that taxon were selected. There is only one partition of the markers.

**excl1** (One reference taxon, excluded.) As incl1, except we discarded the reference taxon from the character matrix.

**incl2** (Two reference taxa, included.) Two taxa were chosen as references. All markers present in either reference taxon were selected. For partitioning, a marker that was present in both reference taxa was assigned randomly to the partition of one of them. Markers that were present in only one reference taxon went into that taxon’s partition.

**excl2** (Two reference taxa, excluded.) As incl2, except we discarded both of the reference taxa from the character matrix.

**all** (All taxa are references.) All markers were included in the character matrix. Each marker was assigned randomly to the partition of one of the taxa at which it was present. (There are as many partitions as taxa.)

$p_{2inc}$ (Two reference taxa, included, predetermined partitioning.) Two reference taxa were chosen. Each marker was assigned randomly to the partition of one of the references. Only markers that were present in their partition’s reference taxon were included in the character matrix.

$p_{2exc}$ (Two reference taxa, excluded, predetermined partitioning.) As $p_{2inc}$, except the two reference taxa were discarded from the character matrix.

$p_{\text{all}}$ (All taxa are references, predetermined partitioning.) Each marker was assigned randomly to the partition of one of the taxa. Only markers that were present in their partition’s reference taxon were included in the character matrix.

For the methods that omit reference taxa ($\text{excl1}$, $\text{excl2}$, $\text{p2exc}$), we simulated extra taxa at the tree generation stage to account for the discarded taxa.

In the $\text{incl2}$, $\text{excl2}$, and $\text{all}$ models, if a marker was present in any reference taxon, it was included in the analysis. This simulates the circumstance when all possible markers found in the reference taxa have been included on the DArT array. The $\text{p2inc}$, $\text{p2exc}$, and $p_{\text{all}}$ models simulate the situation where the number of possible markers is very much greater than the number we can put on the array, so we get an independent random sampling of markers from each reference.

The models do not all produce the same quantity of data. Compared with the expected number of markers present at each taxon, the expected number of markers analyzed is equal for the incl1, p2inc, and $p_{\text{all}}$ models; lower for excl1 and p2exc; higher for incl2; several times
TABLE 1. Counterexamples to the triangle inequality

<table>
<thead>
<tr>
<th>Distance</th>
<th>Dice</th>
<th>LogJaccard</th>
<th>LogDice</th>
<th>LogDet</th>
<th>ADD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character matrix:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B:01</td>
<td>A:0111</td>
<td>B:1101</td>
<td>B:0011</td>
<td>B:010</td>
<td></td>
</tr>
<tr>
<td>C:11</td>
<td>C:1111</td>
<td>C:01</td>
<td>C:0011</td>
<td>C:11</td>
<td></td>
</tr>
<tr>
<td>$d(A, B)$</td>
<td>1</td>
<td>log(5)≈1.61</td>
<td>log(2)≈0.69</td>
<td>log(6)≈1.79</td>
<td>log(4)≈1.39</td>
</tr>
<tr>
<td>$d(A, C) + d(B, C)$</td>
<td>2/3</td>
<td>2 log(5/3)≈1.02</td>
<td>log(1.5)+log(1.25)≈0.63</td>
<td>log(8/3)≈0.98</td>
<td>2 log(1.5)≈0.81</td>
</tr>
</tbody>
</table>

TABLE 2. Summary of the mathematical properties of the distances tested in this article

<table>
<thead>
<tr>
<th>Distance</th>
<th>Hamming</th>
<th>Jaccard</th>
<th>Dice</th>
<th>LogJaccard</th>
<th>LogDice</th>
<th>Nei-Li</th>
<th>LogDet</th>
<th>ADD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triangle ineq.</td>
<td>Yes*</td>
<td>Yes*</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Unknown</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Additive</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

*Hamming (1950).

Possibly $p_{all}$ performs poorly because there are so few markers in each partition. LogJaccard is not far behind the leading methods. Dice, Nei-Li, and LogDice round out the middle of the field. Table 4 shows rather different results for the clock-like trees, with the best distances being Jaccard, Dice, LogJaccard, PADD, then ADD. We were somewhat surprised that the Jaccard distance did so well here, given its poor performance on the nonclock-like trees. Variation between distance methods is smaller, as Yule trees have some very short branches, these are hard for any method to resolve. LogDet performs consistently poorly for both clock-like and nonclock-like trees.

RESULTS

Simulated Data

Table 3 shows the proportions of splits (i.e., edges) in the reconstructed trees that were incompatible with the true tree. (Additional tables for differing numbers of characters and number of taxa are provided in the Supplementary Material, Tables 1–10.) The LogDet, fractional Hamming, and Jaccard distances consistently perform very poorly. ADD has the best overall performance, producing either the best results or results that are not significantly different from the best results in six of the eight models; it also gives quite reasonable results in the remaining two cases. PADD has six near-best results, but fares worse on the remaining two. (The *all* model violates the assumptions of PADD. Possibly $p_{all}$ performs poorly because there are so few markers in each partition.) LogJaccard is not far behind the leading methods. Dice, Nei-Li, and LogDice round out the middle of the field. Table 4 shows rather different results for the clock-like trees, with the best distances being Jaccard, Dice, LogJaccard, PADD, then ADD. We were somewhat surprised that the Jaccard distance did so well here, given its poor performance on the nonclock-like trees. Variation between distance methods is smaller, as Yule trees have some very short branches, these are hard for any method to resolve. LogDet performs consistently poorly for both clock-like and nonclock-like trees.

Figure 2 plots the accuracy of branch length reconstruction against the average number of markers. The fractional Hamming, LogDet, Jaccard, and (to a lesser extent) Dice distances all show signs of ceasing to improve with increasing number of markers. This is expected when the method’s bias exceeds the sampling error. Only ADD improves at the optimal rate (lower dotted line).

In Figure 3, we investigate the possibility of bias in the distances due to tree shape. We divide the true trees according to how many cherries there are in the unrooted tree. (A “cherry” is an internal node directly connected to exactly 2 leaves.) The minimum number is 2 (a maximally unbalanced or “caterpillar” tree). For 15 taxa, the maximum is 7. Two-cherry trees were too rare to get reliable statistics, so Figure 3 shows results for 3–7 cherries. If a method is biased toward producing unbalanced trees, it may be more accurate when the true tree is unbalanced than when it is balanced.

The nonlogarithmic methods (Hamming, Jaccard, Dice) generally have high error rates on unbalanced trees, indicating a possible bias in favor of balanced trees. For the other methods, there is no obvious consistent bias. For example, Nei-Li, LogDice, and ADD are slightly biased toward unbalanced trees for the nonclock-like tree simulations and toward balanced trees for the clock-like tree simulations. More plots demonstrating this
Table 3. The proportion of incorrect splits in minimum evolution trees derived from the various distances on non-clock-like data.

<table>
<thead>
<tr>
<th>Method</th>
<th>incl1</th>
<th>excl1</th>
<th>incl2</th>
<th>excl2</th>
<th>all</th>
<th>p2inc</th>
<th>p2exc</th>
<th>p_all</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogDet</td>
<td>n/a</td>
<td>0.0624(15)</td>
<td>0.3472(227)</td>
<td>0.0136(64)</td>
<td>0.0024(273)</td>
<td>0.2616(121)</td>
<td>0.1242(42)</td>
<td>0.059(63)</td>
</tr>
<tr>
<td>Hamming</td>
<td>0.1204(51)</td>
<td>0.1047(35)</td>
<td>0.1098(59)</td>
<td>0.0857(37)</td>
<td>0.0370(121)</td>
<td>0.0966(36)</td>
<td>0.0931(28)</td>
<td>0.0457(26)</td>
</tr>
<tr>
<td>Jaccard</td>
<td>0.0649(22)</td>
<td>0.0699(19)</td>
<td>0.0506(25)</td>
<td>0.0579(21)</td>
<td>0.0665(220)</td>
<td>0.0600(17)</td>
<td>0.0679(16)</td>
<td>0.0512(30)</td>
</tr>
<tr>
<td>Dice</td>
<td>0.0353(10)</td>
<td>0.0464(7)</td>
<td>0.0246(7)</td>
<td>0.0288(5)</td>
<td>0.0391(123)</td>
<td>0.0347(4)</td>
<td>0.0402(3)</td>
<td>0.0231(9)</td>
</tr>
<tr>
<td>Nei-Li</td>
<td>0.0549(16)</td>
<td>0.0541(11)</td>
<td>0.0438(20)</td>
<td>0.0405(11)</td>
<td>0.0011(2)</td>
<td>0.0398(7)</td>
<td>0.0046(6)</td>
<td>0.0068(4)</td>
</tr>
<tr>
<td>LogDice</td>
<td>0.0497(14)</td>
<td>0.0495(9)</td>
<td>0.0363(15)</td>
<td>0.0337(8)</td>
<td>0.0006</td>
<td>0.0362(5)</td>
<td>0.0398(3)</td>
<td>0.0010(2)</td>
</tr>
<tr>
<td>LogJaccard</td>
<td>0.0406(9)</td>
<td>0.0416(5)</td>
<td>0.0232(6)</td>
<td>0.0222(1)</td>
<td>0.0074(1)</td>
<td>0.0299(2)</td>
<td>0.0329(0)</td>
<td>0.0117</td>
</tr>
<tr>
<td>ADD</td>
<td>0.0221</td>
<td>0.0308</td>
<td>0.0301(11)</td>
<td>0.0259</td>
<td>0.0006</td>
<td>0.0273(1)</td>
<td>0.0326</td>
<td>0.0139(2)</td>
</tr>
<tr>
<td>PADD</td>
<td>0.0241</td>
<td>0.0308</td>
<td>0.0146</td>
<td>0.0256</td>
<td>0.0082(25)</td>
<td>0.0260</td>
<td>0.0354(1)</td>
<td>0.0341(17)</td>
</tr>
</tbody>
</table>

Table 4. The proportion of incorrect splits in minimum evolution trees derived from the various distances on clock-like data.

<table>
<thead>
<tr>
<th>Method</th>
<th>incl1</th>
<th>excl1</th>
<th>incl2</th>
<th>excl2</th>
<th>all</th>
<th>p2inc</th>
<th>p2exc</th>
<th>p_all</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogDet</td>
<td>n/a</td>
<td>0.060(19)</td>
<td>0.209(139)</td>
<td>0.168(88)</td>
<td>0.023(4)</td>
<td>0.198(90)</td>
<td>0.1348(30)</td>
<td>0.0111(31)</td>
</tr>
<tr>
<td>Hamming</td>
<td>0.0432(8)</td>
<td>0.040(16)</td>
<td>0.0355(12)</td>
<td>0.0327(8)</td>
<td>0.0044(1)</td>
<td>0.0365(6)</td>
<td>0.0385(4)</td>
<td>0.0253(1)</td>
</tr>
<tr>
<td>Jaccard</td>
<td>0.0272</td>
<td>0.0287</td>
<td>0.0162</td>
<td>0.0189</td>
<td>0.0039</td>
<td>0.0248</td>
<td>0.0295</td>
<td>0.0233</td>
</tr>
<tr>
<td>Dice</td>
<td>0.0325(3)</td>
<td>0.0336(2)</td>
<td>0.029(3)</td>
<td>0.024(3)</td>
<td>0.0042(1)</td>
<td>0.028(2)</td>
<td>0.0337(2)</td>
<td>0.0209(1)</td>
</tr>
<tr>
<td>Nei-Li</td>
<td>0.0380(7)</td>
<td>0.0420(7)</td>
<td>0.0399(15)</td>
<td>0.0396(11)</td>
<td>0.009(4)</td>
<td>0.0356(6)</td>
<td>0.0408(5)</td>
<td>0.0314(4)</td>
</tr>
<tr>
<td>LogDice</td>
<td>0.0395(6)</td>
<td>0.0407(6)</td>
<td>0.0364(13)</td>
<td>0.034(9)</td>
<td>0.006(2)</td>
<td>0.0353(6)</td>
<td>0.0395(5)</td>
<td>0.0308(4)</td>
</tr>
<tr>
<td>LogJaccard</td>
<td>0.0340(3)</td>
<td>0.0357(3)</td>
<td>0.0254(6)</td>
<td>0.0275(5)</td>
<td>0.0049(1)</td>
<td>0.0302(3)</td>
<td>0.0352(3)</td>
<td>0.0269(2)</td>
</tr>
<tr>
<td>ADD</td>
<td>0.0336(3)</td>
<td>0.0364(4)</td>
<td>0.0366(13)</td>
<td>0.0346(9)</td>
<td>0.006(2)</td>
<td>0.0356(3)</td>
<td>0.0390(5)</td>
<td>0.0304(4)</td>
</tr>
<tr>
<td>PADD</td>
<td>0.0336(3)</td>
<td>0.0364(4)</td>
<td>0.0237(5)</td>
<td>0.0284(6)</td>
<td>0.018(9)</td>
<td>0.0238(4)</td>
<td>0.0596(5)</td>
<td>0.0364(7)</td>
</tr>
</tbody>
</table>

*These data are derived from simulations on non-clock-like trees using 15 taxa and a mean of 500 markers per taxon, with 500 random trials.

*Columns show the results for each of the various models of data censoring.

*Numbers in parentheses indicate how many standard deviations worse this result is than the best result for this model.

*Within each column, for the distance-based methods, the best (lowest) value is shown in bold, along with any that are not significantly worse than the best value.

*Also included for comparison are the Dollo parsimony results (DolloP).

Applications to Real Data

Case Study 1. Eucalyptus DArT data.—We analyzed the Eucalyptus DArT data set of Steane et al. (2011), which includes 94 species of *Corymbia* from across the full taxonomic range (excluding *Corymbia*). The data set comprised 7490 nonredundant DArT markers (newly acquired sequence data have allowed us to eliminate 964 markers as redundant, reducing the number of markers from the 8354 reported by Steane et al. 2011). This data set was generated during the development phase of the Eucalyptus DArT array and only about 32% of the markers in this data set are included on the final, publicly available Eucalyptus DArT array (Sansaloni et al. 2010).

Analysis of this data set using FastME (Desper 2000) also agreed with the PADD distance: the tree shown in Figure 4. Branch support for this and subsequent trees were obtained using 1000 nonparametric (resampling randomly with replacement) bootstraps. Rooting on *Eucalyptus curtisii* (subgenus *Acroser*) was based on previous studies (Driinnan and Ladiges 1991; Steane et al. 2002). The topology of the PADD tree was highly concordant with the most recently published classification (Brooker 2000) and previous molecular studies using internal transcribed spacer (ITS) sequence data (Steane et al. 2002). The tree was also highly congruent with a cladistic analysis of the same data (Steane et al. 2011), but the PADD tree provided higher nonparametric bootstrap support at some key nodes. For example, sections *Latoangulatae* (SL), *Exsertaria* (SE), and *Racemosus* (SR) in the PADD tree form a cluster that is distinct from all other sections, largely in agreement with results from ITS sequence data of Steane et al. (2002). (For the non-botanical reader we note that a “section” is a taxonomic unit that sits between the subgenus and species levels in the taxonomic hierarchy.) The cladistic analysis of the DArT data (Steane et al. 2011) did not resolve the relationships between these sections and section *Maidenaria*. However, the position of section *Racemosus* in relation to sections *Latoangulatae* and *Maidenaria* remains equivocal, with cladistic analysis of the DArT data and cladistic analysis of ITS sequence data placing it close to section *Maidenaria*. The bootstrap values on the branches of the PADD tree are generally high compared with those obtained by the cladistic analysis. Bootstrap values tended to be higher on internal nodes where there were longer branches (i.e., splits with more character support), although bootstrap values were lack of consistency are provided in the Supplementary Material Online, Figures 1–6 (Dryad Digital Repository. doi:10.5061/dryad.2q77).
FIGURE 2. Root mean square (RMS) error in branch length estimation plotted as a function of the number of characters for the various distance methods. If the method is biased, the RMS error cannot fall below the error caused by the bias, and the line is approximately horizontal (e.g., Hamming distance.) The “1/sqrt(n)” line illustrates the expected slope of an optimal method, with variance inversely proportional to number of characters.

FIGURE 3. The effect of tree shape (measured by the number of cherries in the unrooted true tree) on split error rate. In each panel, unbalanced trees are on the left, balanced trees on the right. (Points are for 3-7 cherries.) Downward sloping lines indicate a bias toward constructing balanced trees, and upward sloping lines a bias for unbalanced trees. The vertical axis is linear and originates at zero, but is of different scale in each panel.
Case Study 2. Adnataria (Eucalyptus) DArT data.—We screened 90 species of Eucalyptus from section Adnataria (Brooker 2000) plus 3 outgroup taxa (E. cornuta, sect. Bissectae; E. torquata, sect. Dumaria; and E. staeri, sect. Longistylus) using the publicly available Eucalyptus DArT array (Sansaloni et al. 2010). Leaf samples were collected from Currency Creek Arboretum, South Australia; provenance details of the samples are given online (Dryad Digital Repository. doi:10.5061/dryad.2qt77). DNA was prepared as described previously (Steane et al. 2011). DArT analysis was conducted by DArT P/L (Canberra, Australia) using their standard protocol (Sansaloni et al. 2010).

Section Adnataria (subgenus Symphyomyrtus) includes 100–130 terminal taxa, of which 90 were included in the DArT analysis. Because of the large amount of potential homoplasys in DArT data, it is preferable to include as much genetic variation as possible from within the study group in order to minimize the risk of long-branch attraction and misleading results. Accordingly, the samples in the study represented 8 of the 9 series delineated by Brooker (2000) and represented the full geographic distribution of the section. (DArT data for E. dawsonii, the single species in series Dariosonianaee, were not available). Of the 7680 markers on the DArT array, 3707 provided potentially phylogenetically informative data (i.e., were not constant characters). Of these characters, 1230 were later found to be redundant, that is, extra copies of an existing fragment, and these were removed from the analysis, leaving 2477 markers to be analyzed. The FastME tree derived from PADD distances shown in Figure 5 is not entirely congruent with the established classification (Brooker 2000), but it does have some interesting features. Although at first glance it appears that—apart from series Aquilinoreae—most of the series within section Adnataria are polyphyletic, on closer inspection the series cluster into more-or-less discrete groups. Series Rhodoxylon and Siderophloiae form a distinct cluster (apart from the intrusion of E. rummeryi, series Buxales) even though these series are in different subsections (Brooker 2000). Most of series Buxales and Coalitae form a discrete cluster that is "sister" to a cluster comprising series Heterophloiae, Meliophorae, and Submelliodorae (and one species from series Buxales), none of which are considered by Brooker (2000) to be particularly close. Close examination of the data reveals that there may be a biogeographic aspect to the clusters identified by the PADD algorithm. This is being explored in a separate paper (Steane et al. in preparation).

Notable features of Figure 5 are the shortness of the internal branches and the poor bootstrap values. The internal edges of this tree have mean bootstrap support of 29% and ratio of internal to total edge lengths, or "stemminess," of 0.052 (Fiala and Sokal 1985). Compared with the Adnataria phylogeny (Figure 5), the Eucalyptus phylogeny has much higher bootstrap supports (mean 81%) and higher stemminess (0.138). The Adnataria data set differs from Eucalyptus in that it has fewer markers (2477 compared with 7490) but its taxa are much more closely related to one another (i.e., they span a much narrower taxonomic range). To test whether the structural and statistical differences between the trees were simply a function of sample size (i.e., the number of markers), we generated 100 samples of 2477 markers from the Eucalyptus data set (randomly chosen with replacement) and then bootstrapped each sample 1000 times. The resampled Eucalyptus data had mean bootstrap support of 60% and mean stemminess of 0.153 (SD 0.003), demonstrating that the differences are not simply due to sample size. We conclude that the short internal branch lengths in the Adnataria phylogeny may be indicative of the aftermath of a period of adaptive radiation that produced many species over a short period of time, or of a loss of phylogenetic signal as a result of hybridization (or both).

If a marker is monomorphic (all 0 or all 1) within a data set, the DArT data-processing cannot distinguish whether it is present or absent (Fig. 4 of Jaccooud et al. 2001). Monomorphic data are automatically excluded from the final data set. This is a potential problem with the Adnataria data set, as it is much less genetically diverse than the full Eucalyptus genus from which the array was developed. We are not concerned by markers that are absent from all samples being omitted from the data, as they are ignored by the ADD formula, but there is a prospect that some markers, which are present in all taxa, have been omitted. To test the effect of such an omission, we added to the data set 500 markers that were scored as present in all taxa and analyzed the revised data set. The distances were uniformly smaller; there was very little change to the topology, bootstrap values, or stemminess. We conclude that this potential for missing data does not materially affect our results. (The trees for this test are shown in the Supplementary Material Online, Figs. 7 and 8. Dryad Digital Repository. doi:10.5061/dryad.2qt77).

Case study 3. Gene Family Bacterial Phylogeny.—In this case study, we reanalyze the gene presence/absence data used by Spencer et al. (2007). A potential weakness of the conditioned genome reconstruction approach that has attracted recent attention is the influence of the choice of conditioning genome on the outcome. The most sophisticated attempt to address this, Spencer et al. (2007), used many conditioning genomes, produced a tree for each one, and finally constructed a consensus tree from these. The tree they derived for 40 of the bacterial genomes is shown in Figure 6.

It has been noted previously (Lake and Rivera 2004; Spencer et al. 2007) that parallel loss of genes required for free living causes parasitic bacteria from all bacterial phyla to falsely group together in this analysis. It appears that the SHOT method of Korbel et al. (2002) is the only
As we have derived ADD as a special case of LogDet that is applicable to a Dollo process, we can analyze these COG data without the need for a conditioning genome. In effect, where Lake and Rivera (2004) use LogDet distances and a conditioning genome to find $n_{00}$, we use LogDet distances assuming $n_{00}$ is infinite (equation 10). The resulting phylogeny is shown in Figure 7. It differs from the Spencer et al. (2007) phylogeny on only three edges, two of which have $< 50\%$ bootstrap support in their tree.

**Discussion**

ADD is a simple, consistent distance suitable for use with binary data that have evolved under a stochastic Dollo process. We have shown that ADD is also consistent for data that have been censored in such a way that only traits that are present in a single reference taxon are observable. Censoring by multiple reference taxa is more complex but in the case where each trait is known to derive from a particular reference taxon (e.g., in the DART data presented above), ADD can be extended by a straightforward partitioning scheme (PADD).

Our simulation supports the expectation from theory that ADD and PADD should perform well on data generated under a stochastic Dollo model with various censoring schemes. By contrast, several other distances that are in common use for both DART data and for analysis of gene-content data do not perform well in the simulations. This suggests that they should probably not be used for data thought to be generated under a Dollo model. We used DART data and gene presence/absence data as illustrations of our new approach, but our
distances can be applied to any data derived from complex traits that are unlikely to evolve more than once independently.

More specifically, our simulations indicate that LogDet performs very poorly on stochastic Dollo data. This is a concern for the use of LogDet distances in conditioned genome reconstruction (Lake and Rivera 2004; Spencer et al. 2007). The use of LogDet on gene presence/absence data in deep bacterial phylogeny not only necessitates an extra layer of complication with conditioning genomes to find the number of shared absences but also depends critically on HGT being sufficiently common to justify the use of a Markov model. For the gene-content phylogeny based on the COG data, it is interesting that the tree found using ADD is very similar to the tree found by Spencer et al. (2007). Indeed, the two approaches have very different underlying assumptions — ADD ignores the possibility of HGT, which certainly does occur for these data, and the augmented conditioning genome reconstruction method of Spencer et al. (2007) ignores the Dollo aspect of the data, treating shared absences and shared presences equivalently. Both methods recover the (presumed artefactual) parasite clade (Sangaralingam et al. 2010). We could imagine an intermediate model for this type of data, where markers primarily evolve through a stochastic Dollo process, but are occasionally subject to horizontal gene transfer. In this situation, a suitable distance might be the LogDet distance, with \( r_{00} \) augmented, but still finite.

ADD also appears to be an appropriate distance to use for DArT data, and we have applied it to two closely related data sets. The results obtained using the P ADD algorithm for the Eucalyptus data are satisfyingly congruent with traditional morphology-based taxonomies (e.g., Brooker 2000) and phylogenies based on ITS sequence data. In addition, the ADD analysis of DArT data appears to somewhat improve on the parsimony-based analysis as it provides more resolution, possibly because the parsimony-based
approach does not take account of branch length information and thus can be more easily misled by homoplasy.

In pilot studies, we demonstrated that DArT data had the potential to resolve relationships among closely related Eucalyptus species (Steane et al. 2011). This level of resolution has always been elusive to eucalypt systematists for various reasons including recent/incomplete speciation and the high incidence of interspecific hybridization (Byrne 2008). To test the efficacy of PADD at this level of divergence, we applied it to a set of DArT data for section Adnataria. The PADD-derived Adnataria phylogeny suggests that the series delimited by Brooker (2000) are not robust groups. However, the lack of bootstrap support and the short branch lengths do not provide us with confidence that PADD analysis of the DArT data has uncovered the "true tree." Further analyses of this data set (currently underway) may reveal patterns of variation in other traits (e.g., morphology, physiology, biogeography) that will inform us about the plausibility of this DArT-based phylogeny derived using the PADD algorithm.

It is now reasonably well accepted that model-based phylogenetic inference is the ideal, and in putting forward this distance-based approach for Dollo data we are certainly not seeking to supplant the likelihood-based stochastic Dollo methods of Ryder and Nicholls (2011) and Alekseyenko et al. (2008). The Bayesian methods discussed in Ryder and Nicholls (2011) and Alekseyenko et al. (2008) also have the useful feature that they can incorporate censoring (or "thinning") of the data. Ryder and Nicholls (2011) can allow for data where singleton characters are not observed, and Alekseyenko et al. (2008) discussed the case that applies for the DArT data where characters are only observable if they appear in a particular taxon or set of taxa. These stochastic Dollo models provide powerful tools for inference and hypothesis testing; however, for
large data sets it is still very useful to have consistent distance-based methods. Approaches that give a score to each tree such as likelihood and Dollo parsimony cannot hope to search all of tree space for even moderately sized problems, so having a good starting point for the search is often critical. Another advantage in having a distance-based framework for Dollo data is that it can link in with methods such as Neighbour-Net (Bryant and Moulton 2004), which allow departures from tree-likeness to be detected.

With the increasing debate about the appropriateness of the Tree of Life metaphor for many domains of life, it would be timely to attempt to extend the basic Dollo model to incorporate borrowing of traits. Indeed, both the bacterial gene-content example and the *Adnataria* example are cases where Dollo models that include “borrowing” could be appropriate. Another avenue to pursue would be the development of consistent distances in the case of multi-state Dollo models (such as discussed by Alekseyenko et al. 2008). With the ever-increasing abundance of genomic data, finding good models for the evolution of complex traits is as appropriate now as it was in the time of Dollo.

**Supplementary Material**

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at http://datadryad.org, Dryad Digital Repository. doi:10.5061/dryad.2qf77.
This work was supported by BRH's ARC Future Fellowship FT100100031 and ARC [grant DP0986491 to Brad Potts], V.M. and B.R.H. thank the Royal Society and the New Zealand Marsden Fund [09-MAU-037] for supporting visit (V.M) to New Zealand where work started on this project.

ACKNOWLEDGEMENTS

This article was completed while D.A.S. was a CRN (Collaborative Research Network) research fellow at the University of the Sunshine Coast. Andrzej Kilian helped us understand details of the DArT data processing pipeline. Lars Jermiin suggested considering the triangle inequality and additivity properties of our distances. Bill Martin brought the link to conditioned genome construction to our attention. We also acknowledge Mark Robertson for assistance in the field, Sascha Wise for technical laboratory assistance, René Vaillancourt, Brad Potts for academic input, and the Central Science Laboratory at the University of Tasmania for use of their facilities.

REFERENCES


