Wagner and Dollo: A Stochastic Duet by Composing Two Parsimonious Solos

ALEXANDER V. ALEKSEYENKO,1 CHRISTOPHER J. LEE,2 AND MARC A. SUCHARD1,3,4
1Department of Biomathematics and 3Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, California 90095, USA; E-mail: alexander.alekseyenko@ucla.edu (A.V.A.); msuchard@ucla.edu (M.A.S.)
2Department of Chemistry and Biochemistry, Institute for Genomics and Proteomics, Molecular Biology Institute, UCLA, Los Angeles, California 90095, USA
4Department of Biostatistics, UCLA School of Public Health, Los Angeles, California 90095, USA

Abstract.—New contributions toward generalizing evolutionary models expand greatly our ability to analyze complex evolutionary characters and advance phylogeny reconstruction. In this article, we extend the binary stochastic Dollo model to allow for multi-state characters. In doing so, we align previously incompatible Wagner and Dollo parsimony principles under a common probabilistic framework by embedding arbitrary continuous-time Markov chains into the binary stochastic Dollo model. This approach enables us to analyze characters that exhibit both Dollo and Wagner characteristics throughout their evolutionary histories. Utilizing Bayesian inference, we apply our novel model to analyze intron conservation patterns and the evolution of alternatively spliced exons. The generalized framework we develop demonstrates potential in distinguishing between phylogenetic hypotheses and providing robust estimates of evolutionary rates. Moreover, for the two applications analyzed here, our framework is the first to provide an adequate stochastic process for the data. We discuss possible extensions to the framework from both theoretical and applied perspectives. [Alternative splicing evolution; Bayesian phylogenetic inference; immigration-mutation-death process; intron conservation; stochastic Dollo.]

Reconstruction of species phylogenies relies heavily on our ability to describe the underlying evolutionary processes. The simplest form of phylogenetic inference is that governed by parsimony principles. The most prominent and commonly used form of parsimony is Wagner parsimony (Kluge and Farris, 1969), which assumes that all character states are reversible with equal (or at least similar) rates of transitions. Dollo parsimony (Le Quene, 1974; Farris, 1977) assumes a different model of evolution, where transitions to some states have negligible probability of occurrence. Wagner and Dollo parsimony principles are commonly exploited for two-state character traits. These states are canonically labeled as 0 and 1. Wagner parsimony assumes an unknown ancestral state and allows both transitions from 0 to 1 and the reverse. Dollo parsimony, however, enforces that the ancestral state is 0 and only allows a single transition from 0 to 1 over evolutionary time. After state 1 is attained, transitions to 0 are possible; however, no reverse transitions are then allowed after these final 1 to 0 events.

Clearly some phylogenetic characters are better modeled by one or the other of these parsimony models. For instance, a highly mutable character follows Wagner parsimony more closely. In the case of molecular sequence evolution, the characters are individual nucleotides or amino acids of a molecular sequence. Each character undergoes mutation, thereby changing its state. For DNA, a nucleotide mutates between four states (A,C,G,T). A rare or a more complex character better yields itself to Dollo parsimony principles. For example, presence or absence of a morphological character may be the evolutionary character of interest. Gain of a complex morphological character is a rare event that is often assumed to occur only once in the history of life. On the other hand, loss of such a character is a more common event than gain. Morphological characters and similar complex traits are historically assumed to follow the Dollo parsimony principle (see Gould, 1970, for motivation).

Dollo and Wagner parsimony principles both entertain their stochastic counterparts. The state changes of Wagner characters are most naturally modeled via continuous-time Markov chain (CTMC) transitions (Suchard et al., 2001; Minin and Suchard, 2008, for instance). A recent stochastic Dollo (SD) model (Nicholls and Gray, 2006, 2008) implements Dollo parsimony principles by envisioning an immigration-death stochastic process. The benefits of stochastic models for evolution are many-fold. First, probabilistic models yield themselves to more detailed analysis of rates governing evolutionary processes. Second, stochastic models better account for uncertainty in parental states of characters. Third, the models easily integrate into phylogenetic inference, especially in the Bayesian framework. Finally, stochastic models are relatively easy to extend and combine in order to produce a more precise mathematical representation of reality. We briefly review CTMC and stochastic Dollo methods later in this article.

Although characters generally follow either Wagner or Dollo transitions, some characters exhibit both types of behaviors. For these irregular characters, some state transitions are Dollo-like, allowing for essentially a single event throughout the whole evolutionary history. Other state transitions are more volatile, allowing for multiple changes to be observed over the evolutionary history, just like in the case of Wagner parsimony. For instance, imagine that the presence and the color of an animal tail is the character of interest. Whereas color shades change easily throughout evolution, gain and loss of a tail is a much rarer event. In fact, independent gains might be easily ruled out. Therefore, this character does not conform to either of the parsimony principles. Other examples of such irregular characters are varying gene alleles. Alleles readily change via mutations; at the same time, consider one of the alleles to be due to a complete absence of the gene in the organism. Here gene gain is a
complex and rare event, whereas its loss and mutation are much simpler in nature.

Currently, no model allows characters to exhibit both types of state transitions. In this article, we bridge this gap by developing the multi-state stochastic Dollo (MSSD) model that aligns previously incompatible Wagner and Dollo parsimony principles under a common probabilistic framework. We provide a description of the stochastic process underlying MSSD, details of likelihood calculations under the model, and two applications of our generalized framework to analyses of intron pattern conservation and evolution of alternatively spliced exons.

Materials and Methods

Assumptions about the Evolution of Character Traits

Our main assumption about the evolution of discrete characters is that each of the characters is uniquely identifiable from the data. The existence of an identifiable character and its realized character state are governed by separate evolutionary processes that together detail the probability of observing a given collection of realized character states or their absence. Each character identifies a homology statement about the exhibited character states. Specifically, if we identify a character with a specific character state in one organism, then the same character in another organism, although potentially exhibiting a different state, is homologous, deriving from a common ancestor at some point backwards in evolutionary time. For example, let an orthologous gene be such a character. Although mutations accumulate to make the gene sequence different in two organisms, the overall homology tells us that we are dealing with the same gene. Thus here the gene is the character and its sequence in each organism is its realized character state.

We begin with a description of the SD model of character evolution and pay particular attention to the subtle differences between the underlying character gain-loss and mutation of the character state processes. Infinitely far back in time, no characters exist. As time moves forward, new characters come into existence at a constant rate. However, we assume that these characters are sufficiently complex such that the same character can only arise once. Once gained, a character can subtly mutate between a variety of realized character states or can be lost. Some lucky characters survive to the present and become observable (Fig. 1).

Given a set of organisms, both the presence or absence of a character and, when present, its character state tell us about the evolutionary tree relating these organisms and the underlying gain-loss and mutation processes. If a character is present in two organisms, then that character must have first evolved in a common ancestor that the organisms share in the tree. Alternatively, if one organism possesses a character and another does not, then either the character first evolved after these organisms diverged in the tree or the character has been lost in the latter organism since they shared an ancestor.

Suppose we know the true gain-loss history of a character, and we consider the mutation process giving rise to the different character states. This process acts only along a subtree of the original evolutionary tree relating the organisms, where the single gain and (potentially multiple) loss events truncate the original tree. In the trivial case, a character realizes just one value; i.e., existence. For instance, for a complex character, say an exon, the presence of the exon represents both the existence of the character and its realized state. For some complex characters, the character attains several states. For example, if we consider a particular DNA site as the character, then the character states are the nucleotides at that site (A, C, G, T). In pure mutation process modeling, a gap at a site represents absence of a corresponding letter, and therefore the absence of the character (Fig. 2). The transitions between different discrete values is commonly modeled as a CTMC process, defined through infinitesimal transition rates between states. The SD model induces the subtree of the tree of life where the character exists, whereas the CTMC model describes the changes in character states in this induced tree. In this light, our model is a generalization of the standardly used models for complex characters.

Binary Stochastic Dollo Model

The simplest case of a discrete character is the binary case. The character is either absent or present within each taxon of interest. In this case, we do not have to consider any character state changes, while it is extant. Nicholls and Gray (2008) previously consider the binary SD model for lexicostatistical analysis. Because their lexicostatistical model is not widely known and is a limiting case of our generalized MSSD model, we review some of the basic SD derivations here.

The binary SD model treats the gain of characters as a realization of a Poisson process over the evolutionary tree with instantaneous rate λ. Once a character has evolved, it undergoes a branching survival process down to the tips of the tree with an instantaneous per capita loss rate μ. To formalize the birth-death process, we start by examining the character birth on a single branch and then couple this birth with death over the descending lineages of the remaining tree.

Along a single branch of length t, the number of character births is Poisson distributed with expectation λt. However, we only observe the births that survive to the tips of the tree because some of the evolved characters...
The exact nucleotide residue at any point in time is indicated by lines corresponding to each of the four possible nucleotides (A, C, G, T). The tree where the nucleotide does not exist, either because it has not evolved yet (single direction shading) or it has been deleted (cross shading). The end of the branch is a branch of length 7 that arises more recently. Higher probability of being eliminated than the ones that as characters that arise in the more distant past have a specific time-point \( x \) on the phylogenetic times \( t_i \)'s, \( \forall i, j: j > i \rightarrow t_j \geq t_i \). We label the nodes \( i \in V \), such that the labels impose a partial ordering on the phylogenetic times \( t_i \)'s, \( \forall i, j: j > i \rightarrow t_j \geq t_i \). We label the edges \( (i, j) \in E \) of \( g \) such that the parental node \( j \) incident to the edge appears after the descendant node \( i \) in the pair; i.e., \( t_i < t_j \). Because it is hard to infer information from \( A \), we set its time \( t_A = \infty \). This has an additional benefit of bringing any process acting on edge \( (R, A) \) to equilibrium. A specific time-point \( x \) on \( g \) is represented by a tuple \( x = (\tau, i) \), where \( i \) uniquely specifies the edge directly ancestral to node \( i \) and \( \tau \) marks the time before the present. Let \( |g| = \bigcup_{(\tau, i) \in E} \{ (\tau, i) : t_i \leq \tau < t_j \} \) denote the space of all possible time-points on \( g \). If all of the tips are contemporary (i.e., measured at time \( 0 \) before the present), our description of the tree \( g \) implies an ultrametric constraint on edge lengths.

In general, the distribution of the number of observed unique characters \( Z_g \) along a tree follows a Poisson distribution. Let \( \Omega_g \) be the expected number of unique characters, then the probability mass function is

\[
\Pr(Z_g = k) = \frac{1}{k!} \exp(-\Omega_g)\Omega_g^k.
\]

Total probability mass over the tree \( \Omega_g \) can be computed by integrating the instantaneous observed birth rate \( \omega_g \) over the space of all time-points on \( g \).

\[
\Omega_g = \int_{y \in |g|} \omega_g(y) dy.
\]
births fail to be observed due to death or the specifics of data collection. Let \( S(x) \) denote the event that an observable character is present at time \( x \) on the tree, then \( \sigma_g(x) = \Pr(S(x) \mid x, g, \mu) \) is the probability that a character existing at time \( x \) on the phylogenetic tree is ultimately observed. Naturally, \( \sigma_g \) is the time-dependent birth thinning function over \( g \). Therefore, we express the observed birth rate as

\[
\omega_g(x) = \lambda \sigma_g(x).
\]

Ideally, we would like to express the integral in Equation 3 as a sum of contributions of individual edges of \( g \). To do so, we note that on a single branch \((i, j)\) the probability of observing a birth that occurred at time \( x = (\tau, i) \), can be expressed through the probability of surviving to the end of that branch

\[
\sigma_g(x) = \sigma_g[y = (t_i, i)] e^{-\mu(\tau - t_i)}.
\]

Integrating Equation 5 over \( \tau \) and substituting the result into the integral 3, we can express it as a sum of the edge contributions

\[
\Omega_g = \int_{x \in [g]} \omega_g(y) dy = \lambda \mu \sum_{(i, j) \in E} \sigma_g(t_i, i) (1 - e^{-\mu(t_i - t_j)}).
\]

Data Collection and Probability Mass of the Observations

As we mentioned in passing previously, the distribution of the number of observed characters along a tree depends largely on the observation process \( \sigma_g \). We explore this point further. Consider, for instance, that the observation process consists of picking a single tip taxon \( k \) and recording the conservation of the characters present in \( k \) in the remaining taxa. In this case, the valid character births can only lie along the directed path from \( A \) to \( k \). Consequently, the contribution of some edges to the sum in Equations 3 and 6 may be 0. Moreover, for this simple case, the total mass reduces to a simple equilibrium equation \( \Omega_\infty \) from Equation 1. Let path \((k, A) = ((k, k_1), (k_1, k_2), \ldots, (k_{n-1}, k), (k, A)) \) be the edges lying on the path from \( A \) to \( k \). Because trivially \( \sigma_g[x = (0, k)] = 1 \), we have that

\[
\Omega_g = \lambda \sum_{j=1}^{n-1} \int_{h_j}^{h_{j+1}} \sigma_g[x = (\tau, k j)] e^{-\mu(\tau - h_j)} d \tau,
\]

\[
= \lambda \int_0^{\infty} \sigma_g[x = (0, k)] e^{-\mu(\tau)} d \tau = \frac{\lambda}{\mu} = \Omega_\infty.
\]

In another data collection scenario, the observations initiate from any of the \( L \) tips. However, some observations are discarded as unreliable if the number of tips they occur in is less than a certain fixed number \( r \); e.g., a character has to be present in \( \geq 2 \) tips. In this instance, we need to consider the random variable \( M(x), \) defined as the number of distinct tips, in which a character evolved at time \( x \) appears. It is straightforward to see that observed intensity \( \omega_g \) is thinned by the probability that a character existing at time \( x \) survives into \( r \) or more tips, i.e., \( \sigma_g(x) = \Pr(M(x) \geq r \mid x, g, \mu) \). Nicholls and Gray (2008) gave a simple way to recursively compute these probabilities using Felsenstein’s pruning recursion (Felsenstein, 1981) for small values of \( r \).

In this section we have reviewed how to compute the total mass for the density of gain times, conditional on the observation process \( \int f(X \mid g, \mu, \lambda) dX \), where \( X \) is the set of all gain times. We have shown how to compute this quantity under several observation scenarios. In the next section we would like to use the density \( f \) in order to compute the data likelihood for a set of observations under this model conditional on the gain times \( X \).

Binary Character Data Likelihood

Suppose our observed data \( D = \{C^{(i)} \}_{1 \leq i \leq N} \) that consist of conservation patterns \( C = (c_1, \ldots, c_L) \) for \( L \) characters in \( L \) species. Let \( P(D \mid g, \mu, \lambda) \) be the likelihood of these data under the SD model. We compute this probability by first conditioning on the birth times of all \( N \) characters and then cleverly integrating out these times. If we let \( f(X \mid g, \mu, \lambda) \) denote the joint density of the gain points \( X = (x_1, \ldots, x_N) \), then naturally

\[
P(D \mid g, \mu, \lambda) = \int_{X \in |g|^N} P[D \mid X, g, \mu, S(X)] f(X \mid g, \mu, \lambda) dX.
\]

The joint density \( f(X \mid g, \mu, \lambda) \) follows as a product of independent Poisson process densities as we assume characters are conditionally independent,

\[
f(X \mid g, \mu, \lambda) = \frac{1}{N!} e^{-\lambda g} \prod_{c=1}^{N} \omega_g(x_c) dx_c.
\]

Substituting Equation 9 into the data likelihood yields

\[
P(D \mid g, \mu, \lambda) = \frac{1}{N!} e^{-\lambda g} \int \cdots \prod_{c=1}^{N} P[C^{(c)} \mid x_c, g, \mu, S(x_c)]
\]

\[
\times \omega_g(x_c) dx_c
\]

\[
= \frac{1}{N!} e^{-\lambda g} \prod_{c=1}^{N} P[C^{(c)} \mid x_c, g, \mu, S(x_c)]
\]

\[
\times \omega_g(x_c) dx_c.
\]

Substituting 4 for the intensity and using reasoning similar to Equation 6, we can write the likelihood in terms
of contributions of each individual edge in the tree:

\[
P(D|\mathbf{g}, \mu, \lambda) = \frac{1}{N!} \left( \frac{\lambda}{\mu} \right)^N e^{-\Omega_k} \times \prod_{c=1}^N \sum_{(i, j) \in E} P(C_c^{(c)}(t, i), \mathbf{g}, \mu, S(t, i)) \times \sigma_g(t, i)(1 - e^{-\mu(t_i - t)}).
\]

(11)

Similar to the example of our previous section where the data are ascertained in a single taxon, some of the edge-specific terms contributing to the sum in Equation 11 equal 0. In particular, the contribution of edges \((i, j)\) that are not ancestral to all of the tip taxa with character \(c\) present is zero. Therefore, we eliminate some of the summands by only considering the edges that do belong to the ancestral edge set for character \(c\):

\[
E^{(c)} = \cap_{i \in C_c^{(c)}} \text{Path}(i, A).
\]

(12)

Moreover, for the remaining edges (i.e., those that do lie on \(E^{(c)}\)),

\[
P(C_c^{(c)}|x, \mathbf{g}, \mu, S(x)) \times \sigma_g(x) \equiv P(C_c^{(c)}|x, \mathbf{g}, \mu, S(x)) \times \Pr[S(x)|x, \mathbf{g}, \mu],
\]

\[
= P(C_c^{(c)}, S(x)|x, \mathbf{g}, \mu),
\]

\[
= P(C_c^{(c)}|x, \mathbf{g}, \mu)
\]

\[
\times \Pr[S(x)|x, \mathbf{g}, \mu, C_c^{(c)}],
\]

\[
= P(C_c^{(c)}|x, \mathbf{g}, \mu),
\]

(13)

because \(\Pr[S(x)|x, \mathbf{g}, \mu, C_c^{(c)}] = 1\) trivially, as we must have observed these characters for them to contribute to the likelihood. This allows us to replace a hard-to-compute conditional probability \(P[C_c^{(c)}|S(x)]\) with a simpler expression for realization of the characters at the tips of the tree under survival model \(P(C_c^{(c)}|\cdot)\).

Computing the Likelihood

Substituting Equation 13 into the likelihood function 11 and changing the summation over \(E^{(c)}\) from Equation 12, the function attains a convenient computable form. The computational benefit is unraveled when we recognize that all we need to know to compute the likelihood are \(P(C_c^{(c)}(t_i, i), \mathbf{g}, \mu)\) for \(c = 1, \ldots, N\). A moments reflection reveals that these are the partial likelihoods that arise from carrying out a Felsenstein-style pruning recursion. Specifically, let \(L^{(c)}_+(i)\) and \(L^{(c)}_-(i)\) be the conditional likelihoods of pattern \(C_c^{(c)}\) given that the character at node \(i\) is in \(+\) (character present) or \(-\) (character absent), respectively; then \(P[C_c^{(c)}(t_i, i), \mathbf{g}, \mu] = L^{(c)}_+(i)\). Let \(L^{(c)}\) be the matrix of partial likelihoods for observation pattern \(C_c^{(c)}\), then

\[
L^{(c)} = \begin{pmatrix}
L^{(c)}_+(1) & L^{(c)}_+(1) \\
\vdots & \vdots \\
L^{(c)}_+(R) & L^{(c)}(R)
\end{pmatrix}.
\]

(14)

Further, for an ancestral set \(E^{(c)}\) of an observation pattern \(C_c^{(c)}\), let \(\Xi^{(c)} = \{\Xi_1^{(c)}, \ldots, \Xi_R^{(c)}\}\), where for \(r = 1, \ldots, R\)

\[
\Xi_r = \begin{cases}
1 - e^{-\mu(t_k - t)}, & \text{if } (r, k) \in E^{(c)} \text{ for some } k \in \mathcal{V}, \\
0, & \text{otherwise}.
\end{cases}
\]

(15)

Note that \(\Xi^{(c)}\)'s are uniquely defined because there can be only one edge in \(E^{(c)}\) that starts with \(i\). Further, let \(P = (1, 0)\). We call \(P\) the equilibrium vector for reasons that become obvious in consequent sections of this article. Given these definitions, we rewrite the likelihood 11 in a convenient matrix form

\[
P(D|\mathbf{g}, \mu, \lambda) = \frac{1}{N!} \left( \frac{\lambda}{\mu} \right)^N e^{-\Omega_k} \prod_{c=1}^N \Xi^{(c)} L^{(c)} P.
\]

(16)

Multi-state Stochastic Dollo Model

Following in the footsteps of Nicholls and Gray (2008), the previous sections consider the very limited case in which the characters occur in a single state. In many instances, however, complex characters, although remaining identifiably homologous, exhibit different distinguishable characteristics that change over evolutionary time. For example, if the character is a specific nucleotide position in a DNA sequence, then in addition to being present at that position at a particular point in time it attains any of the four different nucleotide states A, C, G, or T via sporadic mutations at that position (Fig. 2). In this section, we extend the SD model to allow for multi-state characters evolving according to a stochastic state transition model while the characters are in existence.

The most common approach for modeling evolutionary state changes is through CTMC models. To gain intuition about a CTMC model, suppose a character can exist in \(K\) distinguishable states. State changes occur according to a memoryless process with infinitesimal rate matrix \(Q = \{q_{i,j}\}\), where \(q_{i,j}\) are the instantaneous rates of state transition from \(i\) to \(j\), for \(i \neq j\) and \(q_{i,i} = -\sum_{j \neq i} q_{i,j}\). Let \(X(t)\) be the random variable representing the character state at time \(t\), then \(Q(t) = \{q_{i,j}(t)\}\) is the finite-time transition matrix for this process, such that \(q_{i,j}(t) = \Pr[X(t) = j | X(0) = i]\) is the probability that the character is in state \(j\) at time \(t\), given that the character state was \(i\) at time 0. In general for CTMC models, the finite-time transition probabilities can be computed via
matrix exponentiation:
\[ Q(t) = \exp(tQ) = \sum_{k=0}^{\infty} \frac{(tQ)^k}{k!}. \]  

Now suppose, as in the binary case, the evolutionary process starts with the character absent from an ancestral organism. Again, births occur at constant rate \( \mu \). A character then arises in one of the \( K \) states according to a specific birth density \( P(i) \) for \( 1 \leq i \leq K \). The choice of \( P(i) \) is defined by the constraints of the specific application. For our purposes, we assume that the characters arise according to the equilibrium distribution of the CTMC \( \pi_i \), found by solving the equilibrium equations \( \pi_i q_{i,j} = \pi_j q_{j,i} \). After birth, the character continues to evolve according to the transition matrix \( Q \) until eventually the character dies with a constant per capita rate \( \mu \), where \( \mu \) is independent of character state. Note that for \( K = 1 \) this process is identical to the original binary SD process. Also, interestingly, given the subtree of the evolutionary tree where the character is in existence (see Fig. 3), the character does nothing more sophisticated than to follow the CTMC that defines its state changes, typical for phylogenetic models.

The likelihood derivation of Equation 11 considers only the presence and absence of a character, disregarding its state. In order to embed the CTMC into the SD model, we need to incorporate the state change information into this likelihood. To do so, we consider the evolutionary subtree descendant to the birth event that gives rise to the character in question (Fig. 3). At any time following this birth, the character exists in one of the \( K \) states of the underlying Markov chain, or it may not exist at all due to a death event occurring. By definition, the death state is absorbing, and conveniently we can handle death by augmenting the CTMC state-space with a special \((K+1)\)-st state from which no outgoing transitions occur. By doing so, we obtain a modified Markov chain process characterized by the infinitesimal transition matrix \( Q' \):
\[ Q' = \begin{pmatrix} Q - I & 1 \\ 0^T & 0 \end{pmatrix}, \]

where \( 0 \) (1) is a \( K \)-dimensional vector of zeros (ones). The finite-time transition matrix also has a very convenient form that is easily computable from the finite-time transition matrix \( Q(t) \):
\[ Q'(t) = \left[ Q(t) \ast e^{-\mu t} \begin{pmatrix} 1 & 1 - e^{-\mu t} \end{pmatrix} \right]. \]

Forcing the death state to be absorbing is important because otherwise we cannot guarantee that the homologous character observed in distinct organisms represents a single birth event in the evolutionary history. If we allow transitions back from the death state to the existing state(s), the reincarnation represents a re-evolution of a new character that must be classified as homologous with the preexisting one, contradicting the definition of homology. Also the possibility of reincarnation violates the Dollo parsimony principles.

A key step to deriving the likelihood for multi-state characters relies on a convenient way to evaluate the integral in Equation 10. Similar to before, we start by expressing the likelihood in terms of individual edge contributions:
\[ P(D|\mathbf{g}, \mu, \lambda) \propto \prod_{c=1}^{N} \int_{[\mathbf{x}]} P(C(c)|x_c, \mathbf{g}, \mu, S(x_c)) \omega_c(x_c) dx_c, \]
\[ = \prod_{c=1}^{N} \sum_{(i,j) \in E} \int_{\tau = t_i}^{t_j} P[C(c)|\tau, i, \mathbf{g}, \mu, S(\tau, i)] \times \omega_\tau(\tau, i) d\tau. \]

Unlike the likelihood expression in Equation 10, here \( P(C(c)|\tau, i) \) is the probability of the tip labels under the embedded Markov chain \( Q' \). Using this fact, we can rewrite this probability by conditioning on the state in which the character arises
\[ P[C(c)|\tau, i, \mathbf{g}, \mu, S(\tau, i)] \]
\[ = \sum_k \pi_k P[C(c)|\tau, i, k, \mathbf{g}, \mu, S(\tau, i)]. \]

Next, we condition on the character state at the internal node directly descendant to the time of the birth event. We do so by considering all possible transitions from state \( k \) in which the character arises at time \( \tau, i \) to state \( l \) in which the character ends up at time \( \tau, i \). The character must still exist at \( \tau, i \); therefore, we consider only transitions under the original CTMC \( Q \). This allows for simplification through the detailed balance identity
\[ \pi_i = \sum_k \pi_k q_{k,i}, \]  

such that  

\[ P[C^{(c)}(\tau, i), g, \mu, S(\tau, i)] = \sum_k \sum_l \pi_k q_{k,i}(\tau - t_i) P[C^{(c)}(t_i, l), g, \mu, S(t_i, i)], \]

\[ = \sum_l \pi_l P[C^{(c)}(t_i, i), l, g, \mu, S(t_i, i)]. \]  

(22)

The resulting expression for character state changes after character birth does not depend on \( \tau \). This is critical as it allows us to pull \( P[C^{(c)}(\tau, i), g, \mu, S(\tau, i)] \) out of the likelihood integral and arrive at a final expression that involves only finite sums:

\[ P(D, g, \mu, \lambda) \propto \prod_{c=1}^{N} \sum_{(i,j) \in E^{(c)}} \sum_l \pi_l P[C^{(c)}(t_i, i), l, g, \mu, S(t_i, i)] \]

\[ \times \frac{1}{\theta} g(t) \exp \left[ -\frac{1}{\theta} g(t) \mu(t) \right] \]

\[ = \prod_{c=1}^{N} \sum_{(i,j) \in E^{(c)}} \sum_l \pi_l P[C^{(c)}(t_i, i), l, g, \mu, S(t_i, i)] \]

\[ \times \exp \left[ -\frac{1}{\theta} g(t) \mu(t) \right] \]

\[ = \prod_{c=1}^{N} \sum_{(i,j) \in E^{(c)}} \sum_l \pi_l P[C^{(c)}(t_i, i), l, g, \mu] \]

\[ \times (1 - e^{-\mu(t)}), \]  

(23)

where the last step of this derivation follows from the fact that we can replace the summation over the ancestral set \( E^{(c)} \) and the fact that the characters must be observed in order for them to contribute to likelihood, the same reasoning as for Equations 12 and 13, respectively.

The derivation above is valid when traits arise according to \( \pi \). In the more general case, Equation 23 takes the form

\[ P(D, g, \mu, \lambda) \propto \prod_{c=1}^{N} \sum_{(i,j) \in E^{(c)}} \sum_l \pi_l P[C^{(c)}(t_i, i), l, g, \mu] \]

\[ \times \frac{1}{\theta} g(t) \exp \left[ -\frac{1}{\theta} g(t) \mu(t) \right] \]

\[ \times \exp \left[ -\frac{1}{\theta} g(t) \mu(t) \right] \]

\[ = \prod_{c=1}^{N} \sum_{(i,j) \in E^{(c)}} \sum_l \pi_l P[C^{(c)}(t_i, i), l, g, \mu] \]

\[ \times (1 - e^{-\mu(t)}), \]  

(24)

that involves an inconvenient integral over the transitions along the branches. Although hard to evaluate analytically, in some specific instances closed-form solutions may exist. Of particular interest is the case where \( P(i) = 1 \) for some fixed \( i \) and 0 otherwise; i.e., when the character always evolves in a specific state \( i \).

**Computing Likelihoods under the Generalized Framework**

As with the binary SD likelihood, the most computationally intensive terms \( P[C^{(c)}(t_i, i), l, g, \mu] \) in the likelihood 23 are nothing more than partial likelihoods computed under the CTMC \( Q' \). To compute 23, we first let \( L^{(c)}(i) = P[C^{(c)}(t_i, i), k, g, \mu] \) for \( 1 \leq k \leq K + 1 \) denote the partial likelihood of character \( c \) at node \( i \) existing in state \( k \) given the node’s descendants. Then, similar to before, we define the matrix \( L^{(c)} \) as

\[ L^{(c)}(i) = \begin{pmatrix} L^{(c)}(1) & \cdots & L^{(c)}(K) & L^{(c)}(K+1) \\ \vdots & \ddots & \vdots & \vdots \\ L^{(c)}(1) & \cdots & L^{(c)}(K) & L^{(c)}(K+1) \end{pmatrix}. \]  

(25)

To complete likelihood calculations for multi-state characters, set \( P = (\pi_1, \ldots, \pi_K, 0) \) and apply Equation 16.

**Accounting for State-Dependent Death Rates**

In the basic MSSD model, character death rates are state-independent. This assumption may need to be relaxed, particularly when in some states death has higher probability than others. To allow for state-dependent death rates, we wish to model death rates \( \mu_i \) that depend on the state \( i \) in which the character realizes (Fig. 4). Let \( M = (\mu_1, \ldots, \mu_K)^T \) be the vector of per capita death rates, then the infinitesimal matrix for the embedded process becomes

\[ Q' = \begin{pmatrix} Q - IM & M \\ 0^T & 0 \end{pmatrix}. \]  

(26)

Unfortunately, unlike the embedded chains with state-independent death rates (see Equation 19), the matrix exponential of 26 does not yield a convenient general analytic solution. One, however, can still evaluate this matrix exponential model numerically, for instance, via computing the sum \( \sum_{k=0}^{\infty} \frac{1}{k!} (Q')^k \) to some sufficient level of accuracy (see Morel and van Loan, 1978, 2003, for review of numerical techniques to compute the matrix exponential). Once exponentiation is achieved, only a simple modification is necessary to the remaining likelihood calculation steps. In the final step of the derivation (Equation 23), we replace the death rate \( \mu \) by the average death rate \( \bar{\mu} = \sum_k \pi_k M_k \) to account for the differences in these rates. This modification is also necessary when computing the mass \( \Omega_k \) in Equation 6.

**Choice of Priors for the Generalized Framework**

The prior on branching times is induced by assuming an independent exponential distribution with parameter \( \theta \) for each branch length. Therefore, the joint prior on the branch lengths is

\[ f(g|\theta) \propto \theta^{|l|-1} e^{-\theta l(g)}, \]  

(27)
where \( l(g) = \sum_{(i,j) \in E \cap \langle \pi R \rangle} f_j - t_i \) is the total length of all tree branches. Further, as suggested by Nicholls and Gray (2008), we let the joint prior on parameter \( \lambda \) and hyper-parameter \( \theta \) be

\[
f(\lambda, \theta) = \frac{1}{\lambda \theta}, \quad \text{for } (\lambda, \theta) \in \mathbb{R}^2_+.
\]  

(28)

Although the prior 28 is improper, diligent readers may convince themselves that the posterior remains nonetheless proper, because the marginal likelihood of the data is finite for \( N > 0 \). Further, this prior choice allows us to integrate analytically \( \theta \) and \( \lambda \) out of the posterior. We do so in our empirical examples. To complete specification, we let the prior on \( \mu \) follow a diffuse gamma distribution with mean 1 and variance 1000. This choice is uninformative away from 0. We welcome comments on extending Jeffreys ideas to derive an appropriate conditional reference prior on \( \mu \).

The priors for CTMC rates can be obtained from other sources depending on the particular situation. For the case of a two-state Markov chain, we suggest parameterizing the CTMC model via stationary distribution parameter \( \pi \) and scale parameter \( \gamma \). Then we let the prior for \( \pi \) be uniform in the interval \([0, 1]\) and let \( \gamma \) follow gamma \([1/2, l(g)]\) distribution. The latter choice makes the prior scale of the CTMC be proportional to \( 1/\sqrt{\gamma} \) as \( \gamma \to 0 \) and \( \exp[\gamma/l(g)] \) as \( \gamma \to \infty \). Ferreira and Suchard (2008) discuss the details of this prior application.

As is common with phylogenetic likelihoods, the scale of parameters is arbitrary due to rate-time unidentifiability. Therefore, a proper scale, either relative or arbitrary, must be chosen, in case both the branch times and rate parameters need to be estimated simultaneously. One way to do so is to fix one of the scale parameters; for instance, let \( \mu = 1 \), and estimate the rest of the parameters relative to that choice. In some instances, the tree is assumed known, then the model is estimated conditional on the tree. We suggest scaling the tree such that \( l(g) = 1 \), which has an effect of increasing numerical stability since all of the model parameters are then on the scale from 0 to 1.

Conditioning on the tree is enough to be able to identify all other parameters of our model.

**Empirical Examples**

To demonstrate the potential of our generalized framework in distinguishing between phylogenetic hypotheses and providing robust estimates of evolutionary rates, we fit the MSSD model to intron conservation and alternative splicing datasets. We perform this inference through our implementation of MSSD in BEAST (Drummond and Rambaut, 2007), one of the leading phylogenetic analysis programs. Help on using this implementation and example project files are available from the first author.

**Tree Inference Based on Intron Conservation Datasets**

Statistical modeling of evolutionary processes promotes the understanding of phylogenetic relationships between organisms. Starting from observable evolutionary data and a process to describe the generation of the data, some tree topologies are more likely than others, allowing one to delineate a subset of trees that have the largest probability of describing the true evolutionary history. Given a set of probable topologies, we can treat each individual tree as an individual hypothesis and test it against the others in order to capture our confidence in a specific tree. Alternatively, we can view topologies as indices of a family of evolutionary models. Here, each topology specifies a distinct model for the observed data, so that phylogenetic inference reduces to an instance of model selection. We evaluate the power of the SD model to distinguish between different topologies using the latter approach.

We apply our framework to the problem of resolving the phylogenetic relationship of arthropods, nematodes and deuterostomes. For over a decade, debate continues to rage over this relationship, providing arguments in support of two competing hypotheses: (i) arthropods are more closely related to nematodes (the “ecdysozoa” hypothesis, Fig. 5a; see, for instance, Aguinaldo et al., 1997; Philippe et al., 2005), as opposed to (ii) arthropods are more closely related to deuterostomes (the “coelomata” hypothesis, Fig. 5b; Hausdorf, 2000; Wolf et al., 2004). Recently, Roy and Gilbert (2005) have used intron conservation patterns to develop a series of invariant inequalities (reminiscent of Felsenstein’s invariant method; Cavender and Felsenstein, 1987) that give support to the ecdysozoa grouping. However, the results of this analysis are not completely satisfactory because the inference is based purely on empirical measures rather than modeling the evolutionary process of gain and loss of introns in a statistical framework. The empirical measures do not lend themselves to providing levels of
Netic models indexed by their underlying tree topologies (1995, for review). Given the data ability of trees with respect to each other, through a vide a more reliable result.

A full stochastic treatment is likely to pro-
sible. Therefore, a purely minimal change approach is
taken; multiple intron losses in different species is plau-
perficially, this suggests that Dollo parsimony is a good
different species represents a homologous character. Su-
thus every intron observed at a particular position in
intron insertion and loss rates are rather slow (Roy and
of intron conservation data (Rogozin et al., 2003). First,
statistical framework does.

We follow this approach because of its simplicity and
high computational time efficiency. Posterior draws \( \Theta^{(b)} \)
and their corresponding likelihoods \( P(D|g_j, \Theta^{(b)}) \)
are typically available as a part of posterior sampling and
evaluation of Equation 31 entails only within-model
sampling. A downside to the harmonic mean estimator
is its possibly infinite variance; we safeguard against this
limitation through extended posterior simulation and
detailed analysis, discussed later.

We compute the Bayes factor for the two competing hy-
theses: “ecdysozoa” versus “coelomata” by taking the
ratio of the harmonic mean estimates. In this analysis, we
use the plant Arabidopsis thaliana as a known outgroup.
Table 1 summarizes the particular species that we exploit
as representatives of the corresponding (super)phylum.
Because the splitting times and other parameters of the
model are unknown, we integrate them out numerically
using MCMC simulation.

We run the MCMC sampler for 1 million iterations
conditioning on each topology-based hypothesis tree.
We eliminate the initial 100,000 burn-in samples and
subsample every 100 iterations. We run 10 independent
MCMC chains to ensure that our choice of the number of
iterations is sufficient for reasonable convergence. The
Gelman-Rubin statistic (Brooks and Gelman, 1998) for
the 10 chains is below 1.05 for all parameters and effec-
tive sample sizes for the branch lengths are above 5000
per chain. Without loss of generality, we fix \( \mu = 1 \) to break
rate-time identifiability with the splitting times.

Figure 6 reports the cumulative integrated likelihood
estimates for the “ecdysozoa” and “coelomata” hypo-
theses. In Figure 6, we note that the stability of the harmonic
mean estimator of the integrated likelihood is not perfect.
Although we see a modest variation in the estimated
value of the integrated likelihood, as chains increase in
length, clear separation of the likelihood scales between
competing hypotheses is evident. This allows us to make
a reliable conclusion. Based on these data, we estimate
analysis, we use a very common technique involving a
harmonic mean estimator (Newton and Raftery, 1994).
Newton and Raftery (1994) show that the integrated like-
lihood \( P(D|g_i) \) can be estimated from \( B \) posterior draws of
the parameter values \( \Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(B)} \) via the harmonic
identity:

\[
P(D|g_i) \approx \left[ \frac{1}{B} \sum_{b=1}^{B} P(D|g_j, \Theta^{(b)}) \right]^{-1}.
\]

We follow this approach because of its simplicity and
high computational time efficiency. Posterior draws \( \Theta^{(b)} \)
and their corresponding likelihoods \( P(D|g_j, \Theta^{(b)}) \)
are typically available as a part of posterior sampling and
evaluation of Equation 31 entails only within-model
sampling. A downside to the harmonic mean estimator
is its possibly infinite variance; we safeguard against this
limitation through extended posterior simulation and
detailed analysis, discussed later.

We compute the Bayes factor for the two competing hy-
theses: “ecdysozoa” versus “coelomata” by taking the
ratio of the harmonic mean estimates. In this analysis, we
use the plant Arabidopsis thaliana as a known outgroup.
Table 1 summarizes the particular species that we exploit
as representatives of the corresponding (super)phylum.
Because the splitting times and other parameters of the
model are unknown, we integrate them out numerically
using MCMC simulation.

We run the MCMC sampler for 1 million iterations
conditioning on each topology-based hypothesis tree.
We eliminate the initial 100,000 burn-in samples and
subsample every 100 iterations. We run 10 independent
MCMC chains to ensure that our choice of the number of
iterations is sufficient for reasonable convergence. The
Gelman-Rubin statistic (Brooks and Gelman, 1998) for
the 10 chains is below 1.05 for all parameters and effec-
tive sample sizes for the branch lengths are above 5000
per chain. Without loss of generality, we fix \( \mu = 1 \) to break
rate-time identifiability with the splitting times.

Figure 6 reports the cumulative integrated likelihood
estimates for the “ecdysozoa” and “coelomata” hypo-
theses. In Figure 6, we note that the stability of the harmonic
mean estimator of the integrated likelihood is not perfect.
Although we see a modest variation in the estimated
value of the integrated likelihood, as chains increase in
length, clear separation of the likelihood scales between
competing hypotheses is evident. This allows us to make
a reliable conclusion. Based on these data, we estimate

For phylogenetic models, the integral 30 typically does
not have a closed-form solution. Therefore, one is limited
to numerical estimation, usually via simulating the
posterior through Markov chain Monte Carlo (MCMC)
algorithms. Methods for computing the Bayes factors are
ubiquitous, with varying performance, precision, and
efficiency (see Han and Carlin, 2001, for a review). In our

The SD model properties are a reasonable description
of intron conservation data (Rogozin et al., 2003). First,
intron insertion and loss rates are rather slow (Roy and
Gilbert, 2005). Second, intron loss is virtually irreversible,
thus every intron observed at a particular position in
different species represents a homologous character. Su-
perficially, this suggests that Dollo parsimony is a good
description of the intron data. However, care needs to be
taken; multiple intron losses in different species is plau-
sible. Therefore, a purely minimal change approach is
unsuitable. A full stochastic treatment is likely to pro-
sible. Therefore, a purely minimal change approach is
unsuitable. A full stochastic treatment is likely to pro-

One common approach to phylogenetic inference us-
ing evolutionary models is to evaluate the posterior prob-
ability of trees with respect to each other, through a statist-
ic known as a Bayes factor (see Kass and Raftery, 1995, for review).
Given the data \( D \) and a set of phylo-
genic models indexed by their underlying tree topologies
\( g_i \), \( g_j \), and their prior probabilities \( P(g_i) \), the Bayes
factor estimates the ratio of posterior to prior odds in fa-
favor of a given model \( i \), relative to another model \( j \neq i \)
under the likelihood family \( P(D|g) \):

\[
B_{i,j} = \frac{P(g_j|D)P(g_i)}{P(g_i|D)P(g_j)} = \frac{P(D|g_j)}{P(D|g_i)}. \tag{29}
\]

The hard part in computing Bayes factors is the eval-
uation of \( P(D|g_j) \) because it entails integration of all the
parameters \( \Theta \) under the likelihood model \( P(D|g_j, \Theta) \) and
parameter priors \( p(\Theta) \):

\[
P(D|g_j) = \int P(D|g_j, \Theta)p(\Theta)d\Theta. \tag{30}
\]

For phylogenetic models, the integral 30 typically does
not have a closed-form solution. Therefore, one is limited
to numerical estimation, usually via simulating the
posterior through Markov chain Monte Carlo (MCMC)
algorithms. Methods for computing the Bayes factors are
ubiquitous, with varying performance, precision, and
efficiency (see Han and Carlin, 2001, for a review). In our

\[\begin{array}{l}
\text{(Super)Phyla} \\
\text{Species} \\
\text{Common name}
\end{array}\]

\[\begin{array}{lll}
\text{Arthropod (1)} & \text{Drosophila melanogaster} & \text{Fruit fly} \\
\text{Arthropod (2)} & \text{Anopheles gambiae} & \text{Mosquito} \\
\text{Nematode} & \text{Caenorhabditis elegans} & \text{Roundworm} \\
\text{Deuterostome} & \text{Homo sapiens} & \text{Human}
\end{array}\]
that the $\log_{10}$ Bayes factor in favor of the “ecdysozoa” hypothesis against the “coelomata” hypothesis lies between 26.3 and 29.4. Considering the most conservative estimate bound, we find decisive support (Kass and Raftery, 1995), consistent with the previous ad hoc conclusions (Roy and Gilbert, 2005).

Next we consider model parameter estimates under the two topology-based competing hypotheses and the complementary topology, previously considered by Roy and Gilbert (2005). The complementary topology represents a hypothesis that researchers believe implausible. We examine parameter estimates under all three trees to evaluate robustness of the SD model against topology misspecification. Table 3 summarizes the posterior median and 95% Bayesian credible intervals for each parameter. Essentially, estimates of the root and the first splitting time (the split between two insects) show consistent agreement between all three topologies. This indicates that, for uncontroversial tree branchings, consistent estimates of the splitting times are possible. All three topologies yield consistent estimates of $t_3$, the splitting time of the three major clades: nematodes, deuterostomes, and arthropods. The other two splitting times differ remarkably between the trees. While under the “ecdysozoa” hypothesis, separation between $t_2$ and $t_3$ is evident, the other two topologies exhibit significant overlap of the 95% credible intervals. This finding suggests an incorrect ordering of splitting events under these hypotheses. Such diagnostics provide further evidence in favor of the “ecdysozoa” hypothesis.

![Figure 6](https://academic.oup.com/sysbio/article-abstract/57/5/772/1618732)

**Figure 6.** Cumulative window estimates of integrated likelihood as a function of Markov chain Monte Carlo sample size for competing “ecdysozoa” (top line) and “coelomata” (bottom line) phylogenies.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>(a) “Ecdysozoa”</th>
<th>(b) “Coelomata”</th>
<th>(c) “Complement”</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>0.128</td>
<td>0.133</td>
<td>0.134</td>
</tr>
<tr>
<td></td>
<td>(0.117–0.140)</td>
<td>(0.121–0.145)</td>
<td>(0.122–0.146)</td>
</tr>
<tr>
<td>$t_2$</td>
<td>0.704</td>
<td>0.712</td>
<td>0.718</td>
</tr>
<tr>
<td></td>
<td>(0.678–0.730)</td>
<td>(0.701–0.757)</td>
<td>(0.695–0.753)</td>
</tr>
<tr>
<td>$t_3$</td>
<td>0.856</td>
<td>0.876</td>
<td>0.876</td>
</tr>
<tr>
<td></td>
<td>(0.828–0.886)</td>
<td>(0.848–0.905)</td>
<td>(0.847–0.905)</td>
</tr>
<tr>
<td>$t_4$</td>
<td>$-36.62$</td>
<td>$-65.47$</td>
<td>$-64.87$</td>
</tr>
<tr>
<td>log$_{10}$ $P(D</td>
<td>g)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.** Parameter estimates for intron conservation dataset for (a) “ecdysozoa,” (b) “coelomata,” and (c) “complement” phylogenies. See Figure 5 for mapping of time parameters $t_1$, $t_2$, $t_3$, and $t_4$ to edges of the phylogenetic trees (the mapping is similar for the “complement” phylogeny). Time estimates are given in units of expected number of losses per unit of time. The estimates are given as posterior medians and 95% Bayesian credible intervals, with the exception of the harmonic mean estimate of the log of the integrated likelihood of the data given the tree, $\log_{10} P(D|g)$.
a multi-state process may play an important role in exon gain.

In this section, we describe how a generalized MSSD model plays an important role in analyzing alternative splicing conservation datasets. We suppose that the gain of exon sequences is a slow process that involves a combination of many mutations over an extended period of time and view a specific exon sequence across organisms as the same character as long as sequence similarity exists up to an appreciable extent. Homologous exons may already possess necessary splicing signaling (i.e., the ability to be included in gene transcripts) or may lack necessary splice sites (AG/GT consensus sequence on the 3'/5' boundaries of the exon). We model changes in splice site signaling through a two-state CTMC model. We allow exon sequence similarity to be lost through any possible physical process; once lost, homology never reappears. Naturally, this construction of a character trait is somewhat subjective, as it depends on an operational definition of homology via sequence similarity. However, previous work suggests that orthologous exons and introns are identifiable even in distant genomes, if present, which in part validates our assumptions (Alekseyenko et al., 2007). Additionally, this model is simple enough to remain tractable while capturing the necessary characteristics of the data. Figure 7 graphically depicts the components of our model.

Using above definitions, we examine the Vertebrate Exon Evolution Database (VEEDB) previously published by Alekseyenko et al. (2007). This dataset consists of conservation patterns of human exons in 17 vertebrate species. We assume a known phylogenetic tree relating these species (Alekseyenko et al., 2007). The tree is scaled such that the total time along all edges is 1.

In analyzing the VEEDB dataset, we are interested in estimating rates of exon gain and loss and splice site acquisition and deletion in exons stratified by their inclusion level into four categories. Let the th exon’s inclusion category be a function of its empirical inclusion proportion .

\[
S_i = \begin{cases} 
\text{constitutive, if } \rho_i \geq 0.99, \\
\text{major, if } 0.66 \leq \rho_i < 0.99 \\
\text{medium, if } 0.33 \leq \rho_i < 0.66, \text{ or} \\
\text{minor, if } \rho_i < 0.33.
\end{cases}
\] (32)

We are interested in evaluating the hypothesis that the rate of exon evolution is a function of exon mRNA transcript inclusion levels. Specifically, we ask if exons with lower inclusion rates undergo more frequent gains and losses. Unlike previous analyses involving this alternative splicing dataset (Alekseyenko et al., 2007), our approach considers the full phylogenetic likelihood of the observed conservation patterns. This allows us to generate reliable rate estimates of eukaryotic exon evolution. The new exon death rate estimates (Table 3) corroborate previous interpretations (Alekseyenko et al., 2007). Specifically, these estimates suggest that exon sequences are more volatile due to high death rates in exons with lower inclusion levels. Posterior densities clearly separate 95% Bayesian credible intervals of for each of the inclusion categories, suggesting a quantitative effect of inclusion rate on the sequence retention (Table 3).

The alternative splicing model provides insight into the selection pressures that act on exons through examining differences between splice site birth rate and death rate . When close to 1, the ratio provides evidence of a neutral fitness effect for splicing signaling gain or loss, whereas a ratio larger than 1 results from preferential retention of splicing signaling. Our results suggest that constitutive and major-form exons tend to retain their splice sites, medium-form exons experience nearly neutral splice site retention, while minor-form splice sites are lost at a higher rate (Table 3). Again, retention ratios appear to be quantitatively related to the inclusion level. The equilibrium probability of splice site retention gives another perspective on splicing evolution. Specifically, consider , the fraction of unexonized sequences that are present at any point in time. The fact that this quantity is highest in minor-form exons and their high sequence loss rate suggest that minor-form exons appear in the transcriptome highly transiently.

The analysis presented in this section describes a rather simplistic view on the evolution of alternatively spliced exons. This analysis can be improved in several respects. First, we have not incorporated any of the available exon sequence data, other than the splice sites. Incorporating the actual exon sequence might allow us to measure sequence-level effects like amino acid selection pressures and transition/transversion biases, etc., in addition to...
the larger scale edits considered here. Second, the simplicistic structure of the underlying Markov process is likely to overlook the possibility of a transient fixation period, through which newly evolved exons pass. During this period, the probability of exon sequence death and splicing gain and loss rates can be much higher than after a certain point in time. Accordingly, the exon sequence experiences a large shift towards higher selection pressure, associated with attaining a useful function. Existence of minor form exons that are differentially expressed in different tissues (Xing and Lee, 2005) and reduced $K_s$ in these exons furnishes evidence favoring this scenario. Future evolutionary analyses of alternatively spliced datasets should take these issues into consideration.

**DISCUSSION AND CONCLUSIONS**

We have extended the simple binary SD model to allow for multiple states by embedding an arbitrary CTMC model into the character birth-death process. This important extension enables Dollo characters to exist in one of the several states and to mutate between those states. We also demonstrate that the likelihood for the novel model is easily computable from partial likelihoods obtained via Felsenstein-style pruning recursion. The value of our approach is in generalizing previously incompatible Wagner and Dollo parsimony principles and bringing them both under a common probabilistic model. This common framework admits the analysis of a wide range of evolutionary characters.

We have applied this model to two distinct datasets. These applications demonstrate the potential that the MSSD model possesses to distinguish between phylogenetic hypotheses. Moreover, our model provides the first adequate probabilistic framework for analyses for these two datasets. Previously, researchers have analyzed these datasets in an ad hoc fashion using various heuristics. The probabilistic approach trumps these heuristics in providing statistically sound measures of hypothesis uncertainty. However, the analyses of datasets presented in this article serve only as demonstrations of the utility of our model, rather than to provide definite answers to the underlying questions. In particular, these analyses fall short in accounting for temporal rate variations and for variation among sites. More comprehensive analyses for these data need to be performed in the future.

One clear limitation of the proposed MSSD model and also of the original SD model is required absence of homoplasy in gain events. Both models assume that a gain event is so rare that it only happens once across the entire evolutionary history. This assumption might not hold in practice. For instance, consider the case where the gain event itself is composed of a (large) number of simpler events. Entire de novo gain remains rare; however, in some time intervals on the tree, a large enough number of these simple events might accumulate. In these intervals, the probability of accumulating the rest of the remaining simple changes necessary for gain to occur is rather small. If such an interval, for example, immediately precedes a branching event, then two “independent” gains following the branching point remain possible. These situations, however, are hard to detect unless additional data are present. Nonetheless, evidence for homoplasy evolution could derive from the literature (see Rogozin et al., 2007, for instance). One way to account for gain homoplasy in the context of MSSD is to introduce one or more additional proto-gain states. All gain events must then follow through these proto-gain states. These proto-gain states must also emit transitions into the lost state, reflecting the possibility of premature irreversible character loss. Clearly, this approach is only applicable given that the proto-gain states are identifiable (see Warnow et al., 2006, for discussion on homoplasy and identifiability issues).

Several extensions for MSSD are immediately obvious from our current formulation of the model. First, we can incorporate state-dependent death rates into the model. Such incorporation does not notably complicate calculations given a reliable numerical method for computing matrix exponentials. We have given a brief description of this extension in Materials and Methods. Second, we can relax the assumption of independent gains for each character. It is plausible that some characters arise in groups of random or fixed sizes. We can account for this by replacing the gain process of the MSSD with a multiple-immigrant Poisson process. Similarly, some applications require that death events remove several characters simultaneously. The general case for this extension is challenging to derive. However, it should be possible to conceive of simpler specific cases, in particular, when one has more information about the dependence between characters. For instance, combining correlated death with a multiple-immigrant gain process presents a natural way for describing long sequence insertions and deletions.
ACKNOWLEDGMENTS

We thank Bret Larget, Hirohisa Kishino, Jack Sullivan, and Laura Kubatko for valuable discussion and comments on our work. AVA was supported by an Alfred P. Sloan Research Fellowship. Grant DE-fC02-02ER63421, and a Dreyfus Foundation Teacher-Scholar Award. MAS was supported by an Alfred P. Sloan Research Fellowship.

REFERENCES


First submitted 6 February 2008; reviews returned 8 April 2008; final acceptance 25 August 2008

Associate Editor: Laura Kubatko