New Methods for Detecting Lineage-Specific Selection

Adam Siepel[1⋆], Katherine S. Pollard[1⋆⋆], and David Haussler[1,2]

[1] Center for Biomolecular Science and Engineering, U.C. Santa Cruz,
Santa Cruz, CA 95064, USA
[2] Howard Hughes Medical Institute, U.C. Santa Cruz,
Santa Cruz, CA 95064, USA

Abstract. So far, most methods for identifying sequences under selection based on comparative sequence data have either assumed selectional pressures are the same across all branches of a phylogeny, or have focused on changes in specific lineages of interest. Here, we introduce a more general method that detects sequences that have either come under selection, or begun to drift, on any lineage. The method is based on a phylogenetic hidden Markov model (phylo-HMM), and does not require element boundaries to be determined *a priori*, making it particularly useful for identifying noncoding sequences. Insertions and deletions (indels) are incorporated into the phylo-HMM by a simple strategy that uses a separately reconstructed "indel history." To evaluate the statistical significance of predictions, we introduce a novel method for computing P-values based on prior and posterior distributions of the number of substitutions that have occurred in the evolution of predicted elements. We derive efficient dynamic-programming algorithms for obtaining these distributions, given a model of neutral evolution. Our methods have been implemented as computer programs called DLESS (Detection of LinEage-Specific Selection) and phyloP (phylogenetic P-values). We discuss results obtained with these programs on both real and simulated data sets.

1 Introduction

In recent years, abundant sequence data has led to widespread interest in methods for detecting genomic sequences that are evolving faster, slower, or by different patterns of substitution than would be expected under neutral drift. While some such sequences could result from non-uniformities in mutational and repair processes, the majority are thought to be subject to pressure by natural selection, and to have evolutionarily important biological functions. The genomes of most species of interest are too vast, and laboratory assays are still too labor-intensive, to permit exhaustive wet-laboratory searches for functional elements. Computational screens based on comparative sequence data allow whole genomes to be reduced to much smaller sets of candidate functional elements, which can more feasibly be tested in the lab (e.g., [1, 2]).

In the comparative genomics community, much attention has focused on two problems in particular: (1) identifying (especially noncoding) sequences that are unusually conserved across species, and thus are likely to be subject to negative selection (e.g., [3–6]); and (2) identifying protein-coding genes that show unusually high d_N/d_S ratios, and thus might be subject to positive selection (e.g., [7–12]). Methods focused on problem (1) generally have made the assumption (explicitly or implicitly) that selectional pressures are the same across all branches of a phylogeny—i.e., that each candidate sequence is under selection in all species or not under selection in any species. This assumption is sometimes relaxed in methods focused on problem (2) (e.g., [8, 10, 11]), but these methods generally can be used only with protein-coding sequences, whose boundaries are predetermined by annotations of known genes. In addition, most methods that allow for lineage-specific selection have required *a priori* specification of the branches of the tree on which the mode of selection may change [8, 10].

In recent work, we have developed methods for identifying sequences (coding or noncoding) that are significantly changed in the human lineage (K. Pollard, S. Salama, B. King, et al., submitted). These methods are efficient enough to be applied at the scale of complete vertebrate genomes, given the locations of candidate sequences. Our aim here is to develop more general methods capable of detecting sequences that have been subject to lineage-specific selection on any (unspecified) branch of a phylogeny, and that do not

⋆ Current address: Dept. of Biological Statistics & Computational Biology, Cornell University, Ithaca, NY 14853
⋆⋆ Current address: U.C. Davis Genome Center and Dept. of Statistics, Davis, CA 95616

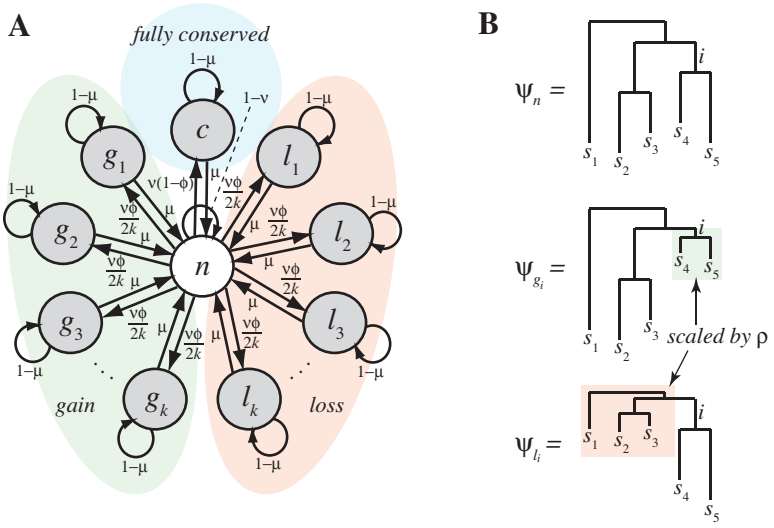

Fig. 1. (A) State-transition diagram for DLESS. The probability of beginning with each state (not shown) is taken to be that state's probability at stationarity. (B) Neutral phylogenetic model ($\boldsymbol{\psi}_n$), with a branch i indicated, and derived phylogenetic models for a "gain" ($\boldsymbol{\psi}_{g_i}$) and "loss" ($\boldsymbol{\psi}_{l_i}$) of a conserved element on branch i.

require predefined element boundaries. These methods must remain highly efficient and suitable for use with noncoding sequences. We focus on the case of negative selection, although our methods can be extended to positive selection (see Discussion). We describe two programs, called DLESS (Detection of LinEage Specific Selection) and phyloP (phylogenetic P-values), that address the problem of detecting lineage-specific selection, and show good power and low false positive rates in simulation experiments. These programs are fast enough to run in a few minutes on multiple alignments for the ENCODE regions [13] (which span ∼1% of the human genome), using a small compute cluster. We describe our methods in detail, and discuss results for both real and simulated data.

2 Methods

2.1 The Model

HMM and Phylogenetic Models DLESS is based on a phylogenetic hidden Markov model (phylo-HMM), an HMM that emits columns of a multiple alignment according to probability distributions that are defined by phylogenetic models associated with its states [14, 15] (reviewed in [16]). DLESS's model is a generalization of the two-state phylo-HMM used by the phastCons program [6]. PhastCons has a state c for conserved sequences and a state n for nonconserved sequences; these states are associated with two phylogenetic models, $\boldsymbol{\psi}_c$ and $\boldsymbol{\psi}_n$, respectively, which are identical except that the branch lengths of $\boldsymbol{\psi}_c$ are scaled by a factor $\rho \in (0, 1)$. Based on this two-state model, phastCons parses an alignment into likely "conserved" and "nonconserved" segments. DLESS works by the same principle, but also allows for conserved elements that have been "gained" or "lost" on any branch of the phylogeny. The new model has $2k+2$ states, labeled c (the "fully conserved" state), n ("nonconserved"), $g_1, \ldots, g_k$ ("gain"), and $l_1, \ldots, l_k$ ("loss"), where k is the number of branches in the tree in question (Fig. 1A). (For a phylogeny of N present-day species, $k = 2N - 3$, assuming a reversible model and an unrooted tree.).

To limit the number of parameters, the states are arranged in a "hub and spokes" configuration (Fig. 1A). As a result, predicted conserved elements are required to be separated from one another by at least one base of nonconserved sequence. In practice, this is not a severe limitation, because, conserved elements in vertebrates are relatively sparse. In addition, conserved elements of all classes are assumed to have the same (geometric) length distribution, and all lineage-specific elements are assumed to occur with the same (prior) probability. Three parameters—μ, ν, and ϕ—define all transition probabilities in the HMM (Fig. 1A). For

interpretability, it is useful to reparameterize μ and ν as $\omega = \frac{1}{\mu}$, the expected length of conserved elements, and $\gamma = \frac{\nu}{\mu+\nu}$, the expected fraction of bases in conserved elements [6]. The third free parameter, ϕ, is the probability that an element is lineage-specific given that it is conserved. Note that this model fails to allow for scenarios in which single elements undergo multiple "gain" and "loss" events over evolutionary time, even if these events occur on separate lineages (see Discussion).

As with phastCons, the phylogenetic models associated with the states are identical, except that certain branches are scaled by the parameter $\rho \in (0, 1)$. The neutral model, $\boldsymbol{\psi}_n$, is assumed to be given (it can be estimated, e.g., from fourfold degenerate sites in coding regions), and all other models are derived from it. The model for a gain event on branch i, $\boldsymbol{\psi}_{g_i}$, is equal to $\boldsymbol{\psi}_n$, except that branch i and all branches in the subtree beneath it are scaled by the factor ρ. Similarly, the model for a loss event on branch i, $\boldsymbol{\psi}_{l_i}$ is equal to $\boldsymbol{\psi}_n$, except that all branches outside the subtree beneath (and including) branch i are scaled by the factor ρ (Fig. 1B). The model parameters can be estimated by maximum likelihood or treated as "tuning" parameters to be set according to some other principle (see below).

Model for Indels Most efforts to use phylogenetic models in the identification of functional elements have finessed the issue of indels (which have long been a thorny problem in phylogenetic analysis), by treating alignment gaps as missing data (e.g., [6]), treating indels like substitutions (e.g., [17, 18]), or using other heuristics (e.g., [19, 5]). Previous approaches, however, are inadequate for the problem of identifying elements under lineage-specific selection. In some cases, alignment gaps are the strongest indication that an element has been "lost" or "gained" (consider an element that was completely deleted on some branch of the tree), so they cannot be treated as missing data. On the other hand, methods that assume site-independence of gaps tend to be too sensitive to occasional indels of moderate length. Other methods (e.g., [19]) cannot be readily applied.

Ideally, one would sample over indel scenarios conditional on an alignment, and possibly sample over alignments as well (e.g., [20, 21]), but we take a short-cut here, which is simpler, faster, and adequate for our purposes. Briefly, we reconstruct an "indel history" (a history of insertion and deletion events on all branches of the tree) by parsimony, using a slightly modified version of the inferAncestors program [22]. We then compute emission probabilities of indels for a phylo-HMM conditional on this history. Given an alignment and indel history, probabilities of indels can be computed using well-known pair-HMM methods, and indel parameters can easily be estimated by maximum likelihood. In addition, it turns out to be straightforward to integrate this indel model into a phylo-HMM (see detailed description in Appendix). This approach, of course, is only as good as the accuracy of the alignment and the indel history, but simulation experiments suggest that their accuracy is quite good, at least for mammalian genomes at modest evolutionary distances [22].

2.2 Assessing Significance

The significance of predicted conserved elements is summarized by a P-value, indicating how surprising the aligned sequences (within the region of the prediction) would be under neutral evolution. We introduce a novel method for computing P-values that is based on counts of substitutions. It appears to have nearly as much power as the likelihood ratio tests (LRTs) more commonly used in statistical phylogenetics (e.g., [7, 8, 11]), and it has certain advantages over LRTs (see Discussion).

The Distribution for One Branch We first derive a solution to the problem of finding the distribution of the number of substitutions along a single branch of a phylogenetic tree, under a general continuous-time Markov model of substitution. To our knowledge, a general solution for this problem has not been published, although methods for computing the mean and variance of such a distribution have been developed [23].

Consider a branch of length t in a phylogenetic tree, connecting a "child" sequence and a "parent" sequence. Assume that substitutions occur by a continuous-time Markov model, defined by a rate matrix $\mathbf{Q} = \{q_{a,b}\}$, where $q_{a,b}$ is the instantaneous rate at which base a changes to base b. Thus, the probability of a base b in the child species given an orthologous base a in the parent species, denoted $P(b|a, t)$, is given by element (a, b) of the matrix $\mathbf{P}(t) = \exp(\mathbf{Q}t) = \sum_{i=0}^{\infty} \frac{(\mathbf{Q}t)^i}{i!}$. We assume that $\mathbf{Q}$ is scaled such that t has units of expected substitutions per site.

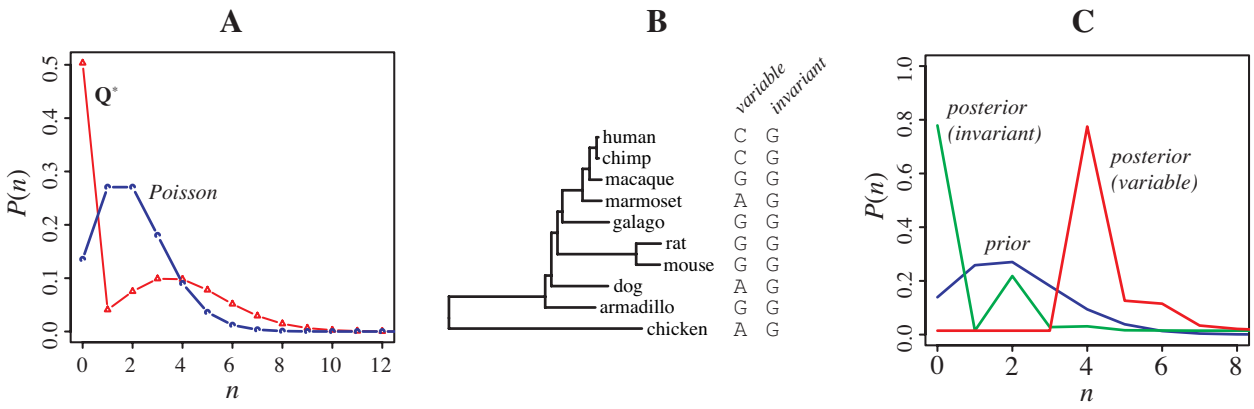

Fig. 2. (A) Distribution of the number of substitutions per site (n) for a rate matrix $\mathbf{Q}^*$ and a single branch of length $t = 2$, obtained with phyloP, and the Poisson distribution of the same mean. $\mathbf{Q}^*$ has very high rates of substitution between bases A and G, but much lower rates (1000×) between other pairs of bases. The stationary distribution is uniform, and the model is reversible. The left peak in the bimodal distribution reflects a starting base of C or T and the right peak reflects a starting base of A or G. (B) Phylogeny of 10 vertebrates, estimated from fourfold degenerate sites in the ENCODE regions [13], and two alignment columns selected from the data set, one highly variable and one invariant. (C) Prior and posterior distributions for the phylogeny and alignment columns in (B). The prior distribution is indistinguishable from a Poisson distribution. The posterior distributions have modes corresponding to maximum parsimony solutions but also give considerable weight to nonparsimonious scenarios. Note that some numbers of substitutions have zero probability in the posterior.

The probability mass function of interest is $P(n|t)$, where n is the number of substitutions per site, allowing for so-called "multiple hits"—i.e., substitutions that obscure other substitutions. The expected value of this distribution, $E[n|t]$, is equal to t, but the entire distribution is not known; it depends not only on t but on the particular process by which sequences of substitutions occur, as defined by the matrix $\mathbf{Q}$. This distribution is sometimes assumed to be Poisson with rate t, and indeed, under certain conditions (e.g., when all substitutions occur at the same rate, as in the Jukes-Cantor model [24]) this assumption is correct. In general, however, the Poisson postulates are violated by the dependency of substitution rates on the starting base in the continuous-time Markov chain. For example, this state-dependency causes the numbers of events in disjoint time intervals to be dependent. It is possible to come up with matrices $\mathbf{Q}$ that cause $P(n|t)$ to be quite unlike a Poisson distribution (Fig. 2A). The mean and variance of $P(n|t)$, for general $\mathbf{Q}$, are of interest in computing the widely used index of dispersion [25, 23].

Solving directly for $P(n|t)$ appears to be difficult (for general $\mathbf{Q}$), but the distribution can be obtained fairly easily by working with the embedded discrete Markov process associated with $\mathbf{Q}$. We decompose the substitution process into a "jump process," which does obey the Poisson postulates, and a substitution process conditional on jumps. The construction is such that every substitution follows a jump, but not every jump is followed by a substitution. Let λt be the rate of the jump process, and let $\mathbf{R} = \{r_{a,b}\}$ be a matrix of conditional probabilities of substitution given a single jump; i.e., $r_{a,b} = P(b|a, 1 \text{ jump})$. Both λ and $\mathbf{R}$ can be derived from $\mathbf{Q}$ (see Appendix). The desired distribution can now be written as:

$$P(n|t) = \sum_{j=0}^{\infty} P(n|j)\text{Pois}(j|\lambda t), \tag{1}$$

where $P(n|j)$ is the probability of n substitutions given j jumps and $\text{Pois}(j|\lambda t)$ is the probability of j jumps in time t. $P(n|j)$ is a function of $\mathbf{R}$ only, and can be precomputed for all n and j less than some adequately large $j_{\max}$ and stored in a table. This computation can be done efficiently by dynamic programming (see Appendix). Subsequently, $P(n|t)$ can be approximated arbitrarily closely, for any n and t of interest, by taking the sum of the first $j_{\max}$ terms of the RHS of equation 1. Using similar methods, it is also possible to obtain the posterior distribution, $P(n|a, b, t)$, and the distribution in the presence of rate variation (see Appendix).

To compute P-values of conservation for conserved elements, we need the distribution of the number of substitutions in an interval consisting of m sites. As long as m is not too large, this distribution can be obtained by taking a convolution of the individual-site distributions, assuming site independence. In the case of the prior distribution, these individual-site distributions are identical, but in the case of the posterior distribution, they differ according to the bases observed at each site.

The Distribution for a Full Phylogeny The distribution of the number of substitutions per site for a general phylogeny can be computed by an algorithm that takes a convolution of the distributions for each branch, using the recursive structure of the tree.

Let X be a (possibly observed) column in an alignment, let u be a node in the tree, let t_u be the length of the branch above node u, let X_u be a random variable representing the base at node u, and let $x_{\underline{u}}$ indicate any observed data at the leaves beneath node u. In addition, let v and w be the children of node u. The key recurrence relation is:

$$P(n, x_{\underline{u}}|X_u = a) = \sum_{i=0}^{n} \left[\sum_b \sum_{j=0}^{i} P(j, x_{\underline{v}}|X_v = b) P(i-j, b|a, t_v) \right] \left[\sum_c \sum_{k=0}^{n-i} P(k, x_{\underline{w}}|X_w = c) P(n-i-k, c|a, t_w) \right] \quad (2)$$

where $P(n, x_{\underline{u}}|X_u = a)$ is the probability of n substitutions beneath node u and the data beneath node u, given that the base at node u is a. The terms $P(i-j, b|a, t_v)$ and $P(n-i-k, c|a, t_w)$ represent branch-specific distributions related to those of the previous section (see Appendix).

The algorithm for computing the distribution resembles Felsenstein's pruning algorithm [26]. It differs only slightly in the cases of the prior distribution and the posterior distribution. Details are given in the Appendix. As for a single branch, the distribution for m sites can be obtained by taking a convolution of the individual-site distributions.

The Joint Distribution for a Subtree and Supertree In the case of lineage-specific selection, the tree is partitioned at some branch B of interest into a subtree and its complementary "supertree." What is of interest in this case is the joint distribution of n_{sub}, the number of substitutions in the subtree beneath B, and n_{sup}, the number of substitutions in the supertree. If the substitution model is reversible, then the tree can be rerooted at the node above branch B, so that the original subtree becomes one subtree of the root, and the original supertree becomes the other subtree of the root. The joint distribution of interest can then be computed by a slight modification of the algorithm described in the previous section. Only the termination step of the algorithm, which is applied at the root of the tree, needs to be altered. Details are given in the Appendix.

As above, the distribution for m sites can be computed by taking a convolution of individual-site distributions. In this case, however, these distributions are bivariate.

***P*-Value Computation** The methods above allow a prior distribution $P(n|m, \boldsymbol{\psi}_n)$ and a posterior distribution $P(n|\mathbf{X}, \boldsymbol{\psi}_n)$ to be computed for any alignment fragment $\mathbf{X}$ of length m and neutral model $\boldsymbol{\psi}_n$. To compute a P-value, we interpret the prior distribution as a null distribution, reflecting the hypothesis of neutral evolution, and we take the mean of the posterior distribution as a proxy for an "observed" number of substitutions. With ample data and branches of modest length, the variance of the posterior distribution is fairly small (Fig. 2C), and it is reasonable to summarize the distribution by its mean. In computing the posterior distribution, the neutral model can be used as a prior, but this influences the posterior mean toward the prior mean, making the P-values conservative. To avoid this problem, we use an empirical Bayes approach: based on the alignment fragment of interest, we estimate a scale factor for the neutral model by maximum likelihood (using a numerical optimization algorithm), then use the scaled neutral model as a prior when computing the posterior distribution. A P-value for a posterior mean $E[n|\mathbf{X}, \hat{\rho}\boldsymbol{\psi}_n]$ is computed as:

$$P = \sum_{0 \leq i \leq E[n|\mathbf{X}, \hat{\rho}\boldsymbol{\psi}_n]} P(i|m, \boldsymbol{\psi}_n) \quad (3)$$

where $\hat{\rho}$ is the estimated scale factor and $\hat{\rho}\boldsymbol{\psi}_n$ denotes the scaled neutral model.

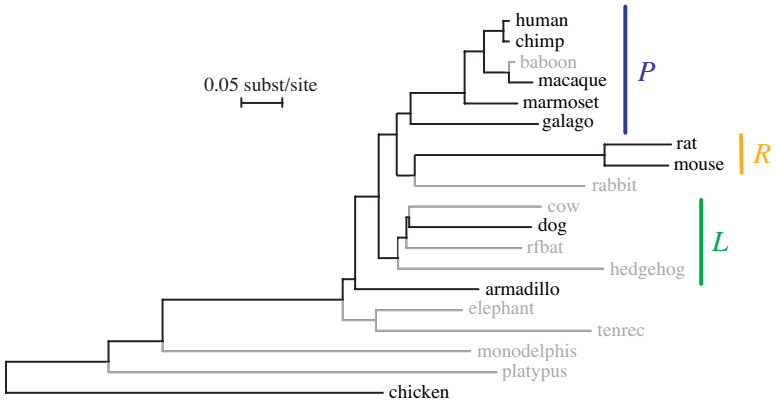

Fig. 3. Phylogenetic tree for the 19-species considered, with neutral branch lengths estimated from fourfold degenerate sites in the ENCODE regions. The 10-species subset is highlighted in black, and three subtrees of interest are indicated: the primates (P), rodents (R), and laurasiatherians (L).

In the case of lineage-specific selection, separate scale factors are estimated for the subtree and supertree, and four P-values are computed. The first two P-values are as described above, except that they are based on marginal distributions (prior and posterior, derived from the corresponding joint distributions) for the subtree and supertree in question. These P-values indicate whether, considered separately, the numbers of substitutions in the subtree and supertree are surprising in comparison to the null model. The other two P-values are conditional P-values, indicating how surprising are the numbers of substitutions in the subtree and supertree, given the total number of substitutions in the whole tree. These P-values allow for the possibility that the substitution rate across the whole tree does not fit the neutral model well, and focus attention more directly on differences between the subtree and supertree. Note that all four P-values are computed independently and do not account for correlation between tests. Adjustments for multiple hypothesis tests are needed when jointly interpreting the marginal P-values for a collection of elements.

2.3 Implementation and Experimental Design

The DLESS (Detection of LinEage Specific Selection) and phyloP (phylogenetic P-value) programs were implemented in C, as new modules in the PHAST (Phylogenetic Analysis with Space/Time models) package [6]. Simulation experiments were conducted to test the false positive rates and power of both programs. All experiments were based on a set of about 100,000 fourfold degenerate sites extracted from alignments of up to 19 species for the 44 ENCODE regions [13], and on a model of neutral evolution estimated from these sites using the REV substitution model (E. Margulies, pers. comm.). We looked at both the full 19-species set and a subset of 10 species (Fig. 3). We simulated neutral alignments using both a "parametric" method (generating sites from the estimated neutral model) and a "nonparametric" method (randomly drawing sites from the original alignment, with replacement). The phyloBoot program in PHAST was used. False positive rates were estimated by running the two programs on these neutral alignments.

To estimate power, we measured the ability of the programs to correctly identify simulated conserved elements, after controlling for false positive rates. Conserved alignment columns were generated parametrically using versions of the neutral model in which either all branches, or some subset of branches (in the case of lineage-specific conservation), were scaled by a factor $\rho \in (0, 1)$. In tests of the power to detect fully conserved elements, we also used a nonparameteric method in which columns were drawn randomly from protein-coding sites extracted from the ENCODE alignments. (Note that some of these sites are in reality not conserved.) Conserved elements of 15–200bp were generated. In tests of DLESS, conserved elements were embedded within neutral alignments of 300bp. Predictions of the correct types overlapping the embedded elements were counted as correct.

The programs were also run on a full set of ENCODE alignments, consisting of 19 species and about 35 million sites (including gaps in the human reference sequence), and produced by the TBA program [27]. The

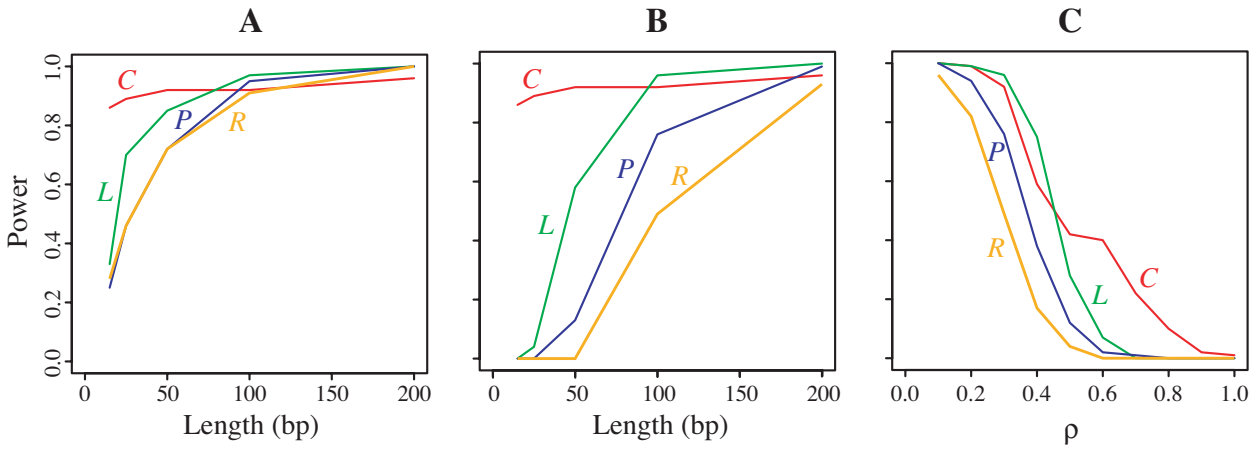

Fig. 4. Estimated power of DLESS to detect fully conserved elements (C), and elements gained or lost in the primate (P), rodent (R), or laurasiatherian (L) clades. Plots show the power to detect lineage-specific (A) losses and (B) gains as a function of element length, and (C) lineage-specific gains as a function of the scale parameter ρ. Results are for the 19-species phylogeny and 100 simulated data sets. In (A) and (B) a value of $\rho = 0.3$ was used, and in (C) elements were of length 100.

DLESS predictions are available as a track in the UCSC Genome Browser (http://genome.ucsc.edu/encode). Clicking on individual predictions causes the statistics computed by phyloP to be displayed.

3 Results

3.1 Simulation Results

Power of DLESS Based on simulated neutral data sets of 1 million columns, we adjusted the tuning parameters of DLESS to permit a false positive rate of approximately one base per thousand, as estimated by the parametric method. We found that γ and ϕ needed to be increased substantially from their maximum likelihood estimates (MLEs; based on the ENCODE data)—from $\gamma = 0.06$ to $\gamma = 0.35$, and from $\phi = 0.18$ to $\phi = 0.8$—to achieve a reasonable tradeoff between false positive and false negative rates. The MLEs led to high specificity but relatively weak sensitivity, especially for lineage-specific elements having short lengths, weak conservation, or small subtrees. This tendency to under-predict lineage-specific elements presumably results from these elements effectively being supported by less data than are fully conserved elements—i.e., it is primarily only the sequences in the subtree (in the case of a gain) or supertree (in the case of a loss) of interest that support the hypothesis of lineage-specific conservation, while all sequences support the hypothesis of full conservation. The parameter ω was set to 20 and the parameter ρ was set to 0.3, based on our experience with the phastCons program.

The power of DLESS to detect conserved elements depends on many factors, including the lengths of the elements, the sizes of the whole phylogeny and of the subtree and supertree in question (numbers of species and total branch length), and the degree of conservation (Fig. 4). The power is generally quite good when elements are of length 50bp or greater and the scaling parameter $\rho \leq 0.3$. The power for detecting fully conserved elements is excellent, even when element lengths are as small as 15bp. The power is also reasonably good for detecting elements gained or lost in subtrees with relatively large numbers of species and large total branch length (e.g., the laurasiatherian subtree; see Fig. 3), but it is significantly reduced when the number of species in a subtree is small (e.g., the rodent subtree), or when the total branch length is small (e.g., the primate subtree). Still, in these cases, longer and more conserved elements can be detected fairly reliably. Interestingly, the method has considerably more power to detect lineage specific "losses" than lineage-specific "gains," particularly for smaller subtrees (compare Fig. 4A & B). Apparently, there is more to be gained by switching states in the HMM when the supertree has short branches and the subtree has long ones (as in a loss) than when the supertree has long branches and the subtree has small ones (as in a

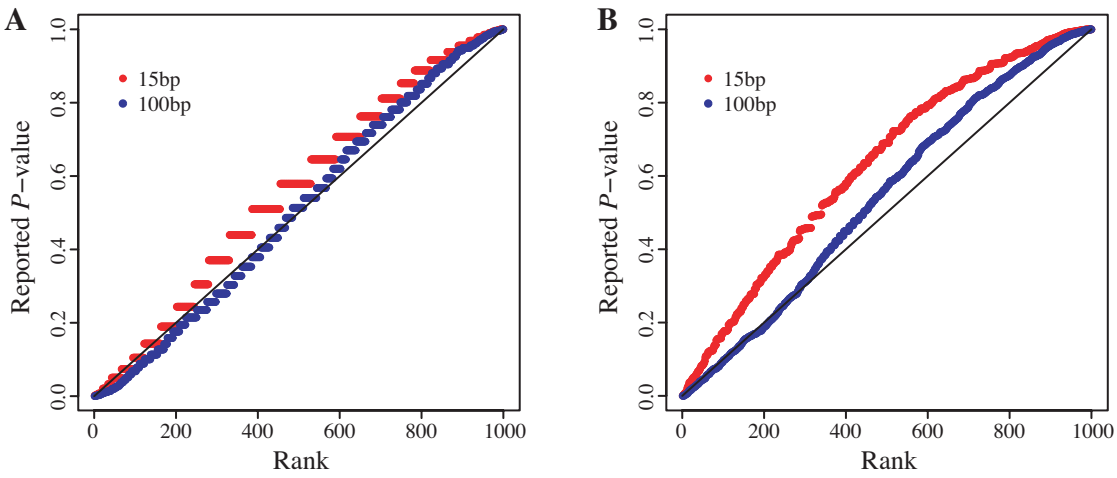

Fig. 5. P-values reported by phyloP versus rank for 1000 neutral data sets (i.e., alignments generated from the neutral model). In (A), P-values of complete conservation are shown, and in (B) conditional P-values of conservation in just the primate subtree are shown. In both cases, points for short elements (15bp) are shown in red and points for longer elements (100bp) are shown in blue. If the P-values were exact, they should fall close to the line shown in black. Instead they tend to fall slightly above this line, indicating that they are generally slightly conservative. PhyloP's P-values are more conservative in the subtree/supertree case than in the fully conserved case, and more conservative for short elements than for longer ones.

gain). This effect might be compensated for by alternative parameterizations of the transition probabilities of the model.

False Positive Rates and Power of PhyloP Because of the estimation of the scale factors and the use of the posterior mean as a proxy for an observed number of substitutions, the P-values reported by phyloP for data sets drawn from the null model are not guaranteed to be uniformly distributed. In practice, the reported P-values are nearly uniform but usually slightly conservative—i.e., the fraction of reported P-values below some p_0 is generally less than p_0, implying a false positive rate below the target value. The P-values are somewhat more conservative for short elements than for longer elements, and somewhat more conservative in the case of lineage-specific selection than in the case of fully conserved elements (Fig. 5).

Despite the conservative P-values, the method has good power. The power to detect fully conserved elements is excellent with $\rho \leq 0.5$, very good with $\rho = 0.7$, and respectable even for $\rho = 0.9$ at lengths of $\geq$ 100bp (Fig. 6A). With smaller values of ρ, elements as short as 15bp can reliably be detected. The nonparametric results, based on protein-coding sites ("CDS" curve in Fig. 6A), suggest that the method's performance in detecting these more conserved elements may be a reasonable indication of its ability to detect real functional elements. For lineage-specific elements (Fig. 6B), the power is reduced but still quite good as long as ρ is not too large, elements are not too short, and subtrees have adequate phylogenetic information. As with DLESS, the power to detect losses is greater than the power to detect gains, but the difference between losses and gains is less pronounced with phyloP than with DLESS. The primate (Fig. 6B) and rodent (not shown) subtrees had similar power curves, and power increased in the laurasiatherian subtree. The method has low power to detect elements in very small subtrees, such as the one consisting of just human (which is of obvious interest). Nevertheless, across a wide range of parameter values, the power of the method is nearly as good as that of an empirically calibrated likelihood ratio test (LRT; Fig. 6C), which is asymptotically most powerful, and is expected to be close to optimal in practice (see Discussion and Appendix). The gap between phyloP and the LRT is greatest for short elements, and is more evident with lineage-specific elements than with fully conserved elements.

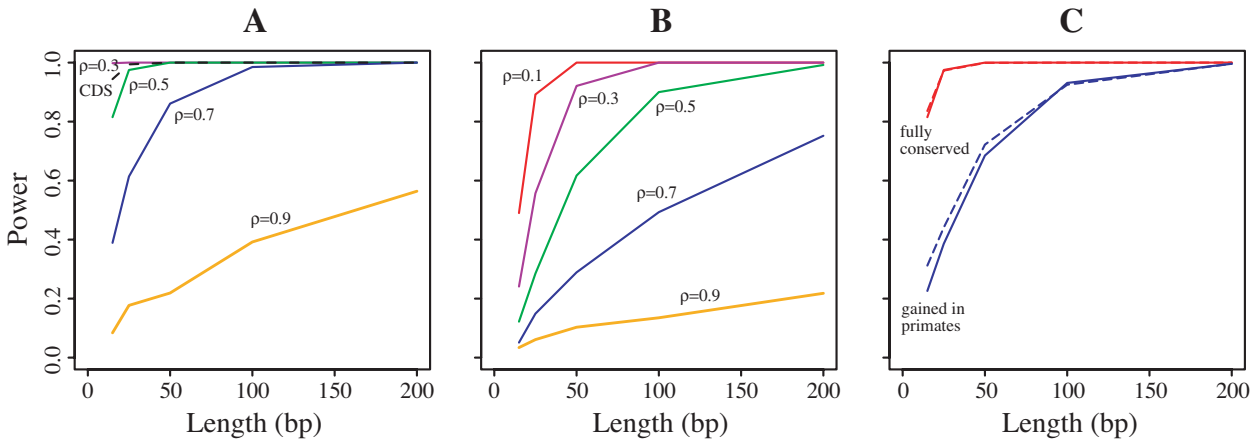

Fig. 6. Power of phyloP in simulation experiments to detect (A) fully conserved elements and (B) elements gained in primates; (C) power comparison between phyloP (solid lines) and a likelihood ratio test (dashed lines) for the representative case of $\rho = 0.5$. Power is estimated as the fraction of 1000 simulated data sets in which the null hypothesis was correctly rejected, with a P-value threshold of 0.05. Conditional P-values were used for lineage-specific elements. These experiments were based on the 10-species set. The results of the nonparametric test, based on sites from coding regions, are shown as the dashed "CDS" line in (A).

3.2 Results for ENCODE Data

DLESS predicted 24,011 elements covering 5.7% of the 29.9 million human bases in ENCODE regions. Retaining only fully conserved predictions with $P < 0.05$ (as reported by phyloP) and lineage-specific predictions with conditional $P < 0.05$ reduced these numbers to 20,959 and 4.8%, respectively. The resulting set of predictions is conservative because maximum-likelihood estimates were used for γ and ϕ, which, as noted above, produced high specificity (and relatively low sensitivity) in simulation experiments, especially for short elements, small subtrees, and weak conservation. In addition, phyloP does not consider indels in computing its P-values, so some of the discarded predictions (with $P \geq 0.05$) might be strongly supported by indel evidence.

About 76% of the predictions (covering 52% of bases) were of fully conserved elements, 14% (36%) were of lineage-specific losses, and the remaining 9% (12%) were of lineage-specific gains. These numbers are undoubtedly affected by differences in power in detecting lineage-specific versus conserved elements, and gains versus losses (Fig. 4). Still, because the method favors fully conserved elements, these results suggest that the number of elements under negative selection in any species is at least 30% higher than the number conserved in all species, and nearly twice as many bases are conserved in any species as are conserved in all species. The predictions covered 70% of bases in coding regions, of which 74% were fully conserved, 21% were losses, and only 5% were gains. In contrast, predictions in introns and intergenic regions covered about 3% of bases, and in these regions we saw more gains (18–20% of predicted bases) and fewer losses (29–30%) than average, while the fully conserved fraction was about average (52%). The most common type of lineage-specific prediction, by far, was a gain on the branch above the last common ancestor of the eutherian mammals, suggesting extensive gain-of-function evolution on this branch.

4 Discussion

In this paper, we have introduced DLESS, a phylo-HMM-based program for identifying sequences that are subject to lineage-specific selection, and phyloP, a new method for computing P-values of conservation (or acceleration) based on prior and posterior distributions of numbers of substitutions. These methods have performed quite well in simulation experiments and yielded promising results with real data. Nevertheless, much work remains to be done in this area.

DLESS currently allows only for neutral evolution and negative selection, and it allows conserved sequence to undergo at most a single "gain" or "loss" event on the branches of the tree. A straightforward extension to positive selection would be possible if the assumption of at most one change in selective "mode" per element were maintained. It is likely, however, that some sequences experiencing adaptive evolution undergo multiple changes in their selective mode. The codon model of Guindon et al. [11] allows for any number of such changes in continuous time, but assumes no correlation between codons. This approach may be reasonable for very deep alignments of protein coding sequences, but a similar model applied to current alignments of noncoding DNA would likely have weak power. A more promising approach for our purposes may be to use a generalization of a phylo-HMM with a separate state-transition Markov chain per node of the tree, rather than one shared chain for all nodes. Unfortunately, models of this type (like models for context-dependent substitution [16]) have "loops" of dependency and do not permit exact probabilistic inference; Markov chain Monte Carlo (MCMC) methods or variational methods would be needed for likelihood evaluation, parameter estimation, and prediction of lineage-specific elements.

A more standard way to compute P-values of conservation would be to use a likelihood ratio test (LRT) (e.g., [28]). LRTs have many appealing statistical properties, and have some advantages over our method. For example, they allow for the fact that some substitutions (e.g., transitions) are generally less surprising than others (e.g., transversions), while our test statistic (a count of substitutions) does not. Also, an LRT would avoid the problem that the statistic in question is not actually observed, which leads to some loss of power in our method. On the other hand, in our method, the exact null distribution of the test statistic can be computed from a model of neutral evolution, without needing to assume asymptotic behavior (e.g., [28]) or to conduct extensive simulation experiments (e.g., [12]), specific to each new set of model parameters. (Among other things, this means that very small P-values can be accurately computed—something which is important when ranking extreme cases, as in genome-wide screens for sequences of interest.) In addition, the test statistic—a count of substitutions—has a clear, intuitive meaning, unlike a likelihood ratio. The descriptions of the prior and posterior distributions produced by phyloP (mean, variance, 95% confidence interval, etc.) are easy to interpret and informative to the user. Moreover, the cost in power of using phyloP appears to be minimal (Fig. 6C). Interestingly, phyloP seems to have better power, in comparison with an LRT, than methods for the identification of positively selected amino acid sites that were also based on substitution counts [29]. This may be because the method is more similar than these methods to an LRT, the main difference being in the choice of the test statistic (R. Nielsen, pers. comm.). Given that relatively little power is sacrificed, interpretability and convenience of application may make our method an attractive alternative to an LRT for a variety of purposes.

While the number of substitutions is not Poisson-distributed for general $\mathbf{Q}$ matrices (Fig. 2A), it appears to have *essentially* a Poisson distribution for most realistic $\mathbf{Q}$ matrices. (Here we restrict ourselves to models of DNA substitution; the situation may be quite different with amino acid models [J. Felsenstein, pers. comm.].) We generated $\mathbf{Q}$ matrices under the HKY model [30] for a wide range of base compositions and transition/transversion ratios, then computed the distribution $P(n|t)$ for a range of values of t and compared each one to a Poisson distribution of the same mean. The symmetric KL divergence of the two distributions was never more than 0.053 bits, suggesting that in many cases it may be quite reasonable to assume a Poisson distribution (see related observations by Zheng [23]). We have not, however, considered rate variation in these experiments (either among sites, among lineages, or along individual branches), which is known to alter the distribution of the number of substitutions. Certain kinds of rate variation can be accommodated with our methods (see Appendix). With or without rate variation, of course, the posterior distribution of the number of substitutions is decidedly non-Poisson (Fig. 2C).

Related to detecting lineage-specific selection is the issue of the rate of "turnover" of functional elements. The rate of turnover is a critical factor in the relationship between the fraction of sites in a genome that are conserved (e.g., between human and mouse) and the fraction that are functional [31]. The methods described here may lead to improved estimates of the rate of turnover, but certain hurdles remain to be cleared. In particular, the strong dependency of the power of the method on element length, degree of conservation, properties of the subtree and supertree in question, and whether an event is a "gain" or a "loss," make it difficult to estimate the rate of turnover accurately. Obtaining good estimates of turnover rates remains an exciting challenge.

Acknowledgments

This project was inspired by questions raised by Bob Harris and Webb Miller about rates of turnover of functional elements, and by independent interest in the ENCODE working groups in detecting lineage-specific selection. The decomposition into a jump process and conditional substitution process was suggested by Rick Durrett. We thank Elliott Margulies for preparing the multiple alignments for the ENCODE regions and for providing us with phylogenetic models estimated from neutral sites; Brian Raney for his work distinguishing indels from missing data; Mathieu Blanchette for providing us with the inferAncestors program; and Greg Cooper, Arend Sidow, George Asimenos, Joe Felsenstein, and Rasmus Nielsen for helpful discussions.

Appendix

The Model for Indels

Let $\mathbf{X}$ be a multiple alignment of N rows (present-day species) and L columns, and let $\boldsymbol{\psi}$ be a corresponding phylogeny with N leaves and $M = 2N - 1$ nodes (Fig. A-1A). We define an *indel history* $\mathbf{Y}$ for $\mathbf{X}$ to be an alignment of M sequences (one for each node in $\boldsymbol{\psi}$), also of length L, each consisting of three characters: a *base* character (denoted 'b') representing any base, an *insertion gap* character ('^') serving as "padding" for insertions in other species, and a *deletion gap* character ('.') representing a previously deleted base (Fig. A-1B). If $\mathbf{Y}$ is *consistent* with $\mathbf{X}$—meaning that it has a 'b' at all positions corresponding to bases in $\mathbf{X}$, and a '^' or a '.' at all positions corresponding to gaps in $\mathbf{X}$—then, for any column of $\mathbf{X}$, a sequence of insertion and deletion events that have given rise to this column is evident from the corresponding column of $\mathbf{Y}$. For example, if the indel character for sequence s in column i of $\mathbf{Y}$ is a '^', and the indel character in the same column for sequence t, an immediate descendent of s, is a 'b', then an insertion is implied to have occurred on the branch between s and t. We will assume that the alignment is correct, in the sense that aligned bases are truly derived from a common ancestor. This imposes certain restrictions on $\mathbf{Y}$; namely, '.' characters can only give rise to '.' characters, 'b' characters cannot give rise to '^' characters, and there can be at most one insertion ('^' → 'b') event per column. The inferAncestors program [22] was modified to output an indel history obeying these rules as a file (in a compact format), which can be read by DLESS and used in predicting conserved elements.

An indel history induces a set of $M - 1$ pairwise alignments of indel strings, one for each branch of the tree. Moreover, given the history, these alignments are conditionally independent, so the probability of the indel history is the product of the probabilities of these $M - 1$ pairwise alignments. When computing the probability of a particular pairwise alignment, columns consisting of a '.' in both ancestor and descendant, or a '^' in both ancestor and descendant, can be ignored; these columns are artifacts of events that have taken place on other branches, and will be accounted for separately. In addition, in a parsimony framework we may assume that a '^' cannot give rise to a '.' (such a transition would require an intermediate base, of which no evidence remains). Thus, each column of interest in a pairwise alignment must be one of three types: (^,b), (b,b), or (b,.) (listed in ancestor, descendant order). These correspond to the insertion (I), match (M), and deletion (D) states of a pair HMM [32] (Fig. A-1C), and the probability of a pairwise alignment is simply the probability of the induced path through a pair HMM (Fig. A-1D). Note that each state of this HMM emits only one type of object (with probability 1), and these objects are all distinct and fully observed, so there is nothing "hidden" about the HMM; it is really just a first-order Markov model over the three types of columns. We describe it as a pair-HMM only because these models a familiar frameworks for pairwise alignment.

The pair HMM for branch i is defined in terms of an insertion rate parameter α, a deletion rate parameter β, an indel length parameter τ (all shared across branches), and the length of branch i, denoted t_i (Fig. A-1C). We assume that insertion and deletion rates are proportional to substitution rates, as appears approximately to be the case in mammalian genomes [17], and that α, β, and t_i are small enough that the insertion and deletion rates are linear in t_i. (For our data sets, estimates of α and β are approximately 0.01–0.05, similar to those estimated by Cooper et al. [17], and branch lengths are never more than 0.45 subst/site, so this assumption is reasonable.) Given an indel history and a tree with branch lengths, α, β, and τ are estimated by maximum likelihood.

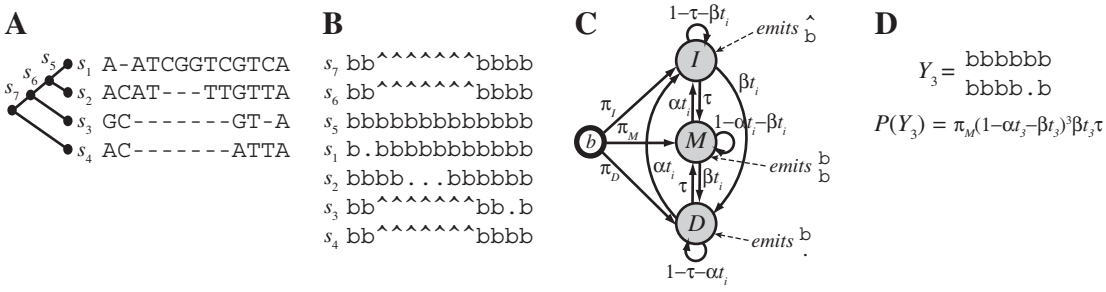

Fig. A-1. (A) An alignment and phylogeny for four present-day species and three ancestral sequences; (B) an indel history consistent with the alignment and phylogeny; (C) the pair-HMM for a branch i; (D) the pairwise alignment Y_3 for the branch above species s_3 (the irrelevant columns have been removed), and its probability under the pair-HMM for that branch.

The indel model is incorporated into the phylo-HMM as follows. Normally, the probability of an alignment $\mathbf{X}$ given a phylo-HMM (the likelihood of the model) is:

$$P(\mathbf{X}) = \sum_{\mathbf{Z}} \prod_{i=1}^{L} P(X_i|Z_i)P(Z_i|Z_{i-1}) \tag{A-1}$$

where L is the length of the alignment, X_i is the ith column, Z_i is the ith state visited in a path $\mathbf{Z}$, $P(Z_i|Z_{i-1})$ is the probability of a transition from state Z_{i-1} to state Z_i, and $P(X_i|Z_i) = P(X_i|\boldsymbol{\psi}_{Z_i})$ is the emission probability of X_i given Z_i. (All probabilities are implicitly conditioned on the model; to simplify notation, the probability of starting with state Z_1 is denoted $P(Z_1|Z_0)$.) To incorporate an indel history into this calculation, we write the joint probability of an alignment $\mathbf{X}$ and a (consistent) indel history $\mathbf{Y}$ as:

$$P(\mathbf{X}, \mathbf{Y}) = \sum_{\mathbf{Z}} \prod_{i=1}^{L} P(Y_i|Y_{i-1}, Z_i)P(X_i|Y_i, Z_i)P(Z_i|Z_{i-1}) \tag{A-2}$$

where $P(Y_i|Y_{i-1}, Z_i)$ is the probability of column i of the indel history given column $i-1$, which is a product over branches of the tree and is defined by the pair HMMs for each branch, and $P(X_i|Y_i, Z_i)$ is the probability of the bases in column X_i of the alignment, given the indel history. We define $P(X_i|Y_i, Z_i)$ to be the probability of just the bases in column $\mathbf{X}_i$, ignoring columns with gaps; it turns out to be the same as the probability of X_i treating gaps as missing data. The resulting model can be thought of as an HMM that jointly emits alignment columns and indel-history columns, with emission probabilities $P(Y_i|Y_{i-1}, Z_i)P(X_i|Y_i, Z_i) = P(X_i, Y_i|Y_{i-1}, Z_i)$. (Conditional independence of X_i and Y_{i-1} is assumed.) Thus, the only difference between the new model and an ordinary phylo-HMM (assuming alignment gaps are treated as missing data) is that the emission probabilities must be multiplied by an additional term. The standard Viterbi algorithm can still be used for prediction, and the standard forward algorithm can still be used for likelihood evaluation.

Note that, because the branch lengths of the phylogeny depend on the state of the HMM (Fig. 1B), so too do the pair-HMM transition probabilities. It is through the dependency of indel probabilities on branch lengths that the phylo-HMM is able to use indels to help discriminate between conserved and nonconserved sequences. In practice, we use separate values of α, β, and τ for conserved and nonconserved sequences, estimating the "conserved" values from sites in previously annotated conserved elements, and the "nonconserved" values from the remaining sites. Indel rates seem to be even more suppressed in conserved regions than a linear relationship to substitution rates would imply, so having separate sets of parameters helps further in discrimination.

The Distribution of the Number of Substitutions for One Branch

Recall that λ is the rate of the jump process and $\mathbf{R}$ is a matrix of conditional probabilities of substitution given a single jump. For a short period of time $t = \epsilon$, during which it can be assumed that at most one jump

occurs, the probability of an $a \to b$ substitution is:

$$P(b|a,\epsilon) = \begin{cases} \lambda\epsilon r_{a,b} & \text{if } a \neq b \\ 1 - \lambda\epsilon + \lambda\epsilon r_{a,a} & \text{if } a = b \end{cases} \tag{A-3}$$

The same probability can be expressed in terms of the rate matrix $\mathbf{Q}$ as:

$$P(b|a,\epsilon) = \begin{cases} q_{a,b}\epsilon & \text{if } a \neq b \\ 1 + q_{a,a}\epsilon & \text{if } a = b \end{cases} \tag{A-4}$$

Equations A-3 and A-4 imply that $r_{a,b} = \frac{q_{a,b}}{\lambda}$ (for $a \neq b$) and $r_{a,a} = \frac{q_{a,a}}{\lambda} + 1$. We have some freedom in our choice of λ, but by setting $\lambda = \max_a(-q_{a,a})$, we ensure that the elements of $\mathbf{R}$ are probabilities, and minimize the probability of failing to make a substitution following a jump.

Note that if all terms on the main diagonal of $\mathbf{Q}$ are equal, then the main diagonal of $\mathbf{R}$ will consist of all zeroes, and a substitution will follow every jump, making $P(n|t)$ Poisson-distributed. Thus, as noted by Zheng [23], it is not necessary for all substitution rates to be equal to have Poisson-distributed numbers of substitutions, but only that the total rate of substitution not depend on the starting base. This is true for the Kimura two-parameter model [33] as well as for the Jukes-Cantor model [24].

As discussed in the text, the desired distribution is given by $P(n|t) = \sum_j P(n|j)\text{Pois}(j|\lambda t)$, where $P(n|j)$ is the probability of n substitutions given j jumps and $\text{Pois}(j|\lambda t)$ is the probability of j jumps in time t. It turns out to be more convenient, however, to work with the related distribution,

$$P(n,b|a,t) = \sum_j P(n,b|a,j)\text{Pois}(j|\lambda t) \tag{A-5}$$

$P(n,b|a,t)$ is the probability of n substitutions and final base b given starting base a and j jumps. Both the prior distribution $P(n|t)$ and the posterior distribution $P(n|a,b,t)$ can readily be obtained from $P(n,b|a,t)$ as

$$P(n|t) = \sum_{a,b} \pi_a P(n,b|a,t), \qquad P(n|a,b,t) = \frac{P(n,b|a,t)}{P(b|a,t)}$$

The quantity $P(n,b|a,j)$ can be computed recursively, using the recurrence relation:

$$P(n,b|a,j) = r_{b,b}P(n,b|a,j-1) + \sum_{c:c\neq b} r_{c,b}P(n-1,c|a,j-1) \qquad [n,j \geq 1] \tag{A-6}$$

with base case

$$P(n,b|a,j=0) = \begin{cases} 1 & \text{if } a = b \text{ and } n = 0 \\ 0 & \text{otherwise} \end{cases} \tag{A-7}$$

Note that $P(n,b|a,j) = 0$ for $n > j$ (at most one substitution can occur per jump). The values $P(n,b|a,j)$ can be precomputed and stored in a table, for all a, b and all n, j such that $0 \leq n, j \leq j_{\max}$, where $j_{\max}$ is sufficiently large that $\text{Pois}(j_{\max}|\lambda t)$ is not much larger than the machine precision. This table need only be computed once for a given rate matrix $\mathbf{Q}$. In practice, it does not have to be very large.

It is worth noting that in this framework certain types of rate variation simply imply alternative distributions for the number of jumps. For example, if we assume, in the manner of Yang [34], a random scaling factor R for t with some prior distribution $P(R = r|\theta)$, so that $P(b|a,t) = \int P(b|a,rt)P(R = r|\theta)\,dr$, then $P(n,b|a,t)$ is given by:

$$\begin{aligned} P(n,b|a,t,\theta) &= \int P(n,b|a,rt)P(R = r|\theta)\,dr \\ &= \sum_j P(n,b|a,j)P(j|\lambda t,\theta) \end{aligned} \tag{A-8}$$

where $P(j|\lambda t,\theta) = \int \text{Pois}(j|\lambda rt)P(R = r|\theta)\,dr$. If R is gamma-distributed, then $j|\lambda t$ has a negative binomial distribution.

The Distribution for a Full Phylogeny

The algorithm to compute the distribution $P(n|X, \boldsymbol{\psi})$, given tree model $\boldsymbol{\psi}$, visits the nodes of the tree in a post-order traversal. Like Felsenstein's pruning algorithm, it is defined by three cases: initialization (at the leaves of the tree), recursion (at internal nodes), and termination (at the root). The recursive case is defined by equation 2 in the text. The initialization depends on whether or not the alignment column X is observed. In the case of the posterior distribution, when X is observed, the base case is:

$$P(n, x_{\underline{u}}|X_u = a) = \begin{cases} 1 & \text{if } n = 0 \text{ and } X_u = a \\ 0 & \text{otherwise} \end{cases} \qquad \text{for leaf } u \tag{A-9}$$

In the case of the prior distribution, when X is unobserved, the base case is:

$$P(n, x_{\underline{u}}|X_u = a) = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{for leaf } u \tag{A-10}$$

The termination step, which is the same for the prior and posterior distributions, is to obtain the final distribution via the equation:

$$P(n|X, \boldsymbol{\psi}) = \frac{P(n, X|\boldsymbol{\psi})}{P(X|\boldsymbol{\psi})} = \frac{\sum_a \pi_a P(n, x_{\underline{r}}|X_r = a)}{P(X|\boldsymbol{\psi})} \tag{A-11}$$

where r is the root of the tree, $P(X|\boldsymbol{\psi})$ is the total probability of the observed data (the likelihood; 1 in the case of unobserved X), and π_a is the equilibrium frequency of base a. (Note that $x_{\underline{r}} = X$.)

The computational complexity of the algorithm is $O(Nd^3 n_{\max}^3)$, where N is the number of species, d is the size of the alphabet ($d = 4$ for DNA) and $n_{\max}$ is a "ceiling" on n, large enough that $P(n_{\max}|X, \boldsymbol{\psi})$ is nearly zero.

In the case of rate variation among sites but not among branches [34], with $P(R = r|\theta)$ being the distribution of a random scaling factor R (as described above), the distribution of n is simply:

$$P(n|X, \boldsymbol{\psi}, \theta) = \int P(n|X, r\boldsymbol{\psi}) P(R = r|\theta)\, dr. \tag{A-12}$$

This distribution may be approximated in the manner of Yang [34].

The Joint Distribution for a Subtree and Supertree

Let B be a branch of interest, dividing the phylogeny into a subtree and complementary supertree, and let u be the node above B. Assume that the substitution model is reversible. The tree can be rerooted at u and a zero-length branch can be added such that the former supertree becomes one subtree beneath u, and the former subtree becomes the other subtree beneath u (Fig. A-2). Let v and w be the new children of u, with the branch between v and u having length $t_v = 0$, as shown in Fig. A-2.

To compute the joint distribution $P(n_{\text{sup}}, n_{\text{sub}}|X, \boldsymbol{\psi})$, we proceed exactly as above, computing the quantity $P(n, x_{\underline{u}}|X_u = a)$ recursively in a post-order traversal of the tree. In this case, however, we terminate the recursive procedure at nodes v and w, and compute the joint distribution as:

$$P(n_{\text{sup}}, n_{\text{sub}}|X, \boldsymbol{\psi}) = \frac{P(n_{\text{sup}}, n_{\text{sub}}, X|\boldsymbol{\psi})}{P(X|\boldsymbol{\psi})} \tag{A-13}$$

where $P(X|\boldsymbol{\psi})$ is the likelihood and $P(n_{\text{sup}}, n_{\text{sub}}, X|\boldsymbol{\psi})$ is given by

$$\begin{aligned} P(n_{\text{sup}}, n_{\text{sub}}, X|\boldsymbol{\psi}) &= \sum_a \pi_a \left[\sum_b \sum_{i=0}^{n_{\text{sup}}} P(i, x_{\underline{v}}|X_v = b) P(n_{\text{sup}} - i, b|a, t_v) \right] \left[\sum_c \sum_{j=0}^{n_{\text{sub}}} P(j, x_{\underline{w}}|X_w = c) P(n_{\text{sub}} - j, c|a, t_w) \right] \\ &= \sum_a \pi_a P(n_{\text{sup}}, x_{\underline{v}}|X_v = a) \sum_c \sum_{j=0}^{n_{\text{sub}}} P(j, x_{\underline{w}}|X_w = c) P(n_{\text{sub}} - j, c|a, t_w) \end{aligned} \tag{A-14}$$

where the simplification in the second step is possible because $t_v = 0$.

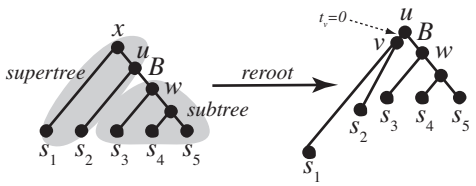

Fig. A-2. Rerooting of a phylogeny at the node u above a branch of interest B, so that the subtree beneath B becomes one subtree of the root and the complementary "supertree" becomes the other subtree of the root. Note that the branch B, on which a possible "loss" or "gain" event has occurred, is grouped with the subtree rather than the supertree. The node x is eliminated and a new node v is added, then connected to u by a zero-length branch.

Likelihood Ratio Tests

The likelihood ratio tests (LRTs) that were compared to phyloP were conducted as follows. In the case of fully conserved elements, we used the test statistic $R = \log P(\mathbf{X}|\hat{\rho}\boldsymbol{\psi}_n) - \log P(\mathbf{X}|\boldsymbol{\psi}_n)$, where $\mathbf{X}$ is the alignment in question, $\boldsymbol{\psi}_n$ is the neutral model, and $\hat{\rho}\boldsymbol{\psi}_n$ is a scaled version of $\boldsymbol{\psi}_n$, with $\hat{\rho}$ being the maximum likelihood estimate of the scaling parameter. This value was determined numerically (using the BFGS quasi-Newton method), subject to the constraint $0 \leq \hat{\rho} \leq 1$. The null distribution of R was assessed by simulation. For each element length, 1000 synthetic alignments were generated, R was computed for each one, and the 95th percentile of the observed distribution was taken as the threshold for rejecting the null hypothesis. This empirically determined threshold was then used in all power experiments for elements of that length. We used a similar procedure for the subtree/supertree experiments but in this case we compared the log likelihood of a version of $\boldsymbol{\psi}_n$ for which separate scale parameters were estimated for the subtree and supertree in question with the log likelihood of a version of $\boldsymbol{\psi}_n$ for which a single global scale parameter was estimated. (The scale parameter for the supertree was unconstrained, but the scale parameter for the subtree had to be less than one in the case of a gain and greater than one in the case of a loss.) The null distributions were separately examined for each subtree of interest and for the cases of losses and gains.

Note that no indels were present simulated data, and indels were not considered in any power tests.

References

1. Nobrega, M.A., Ovcharenko, I., Afzal, V., Rubin, E.M.: Scanning human gene deserts for long-range enhancers. Science **302** (2003) 413
2. Woolfe, A., Goodson, M., Goode, D., Snell, P., McEwen, G., Vavouri, T., Smith, S., North, P., Callaway, H., Kelly, K., et al.: Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol **3** (2005) e7
3. Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., Rubin, E.M.: Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science **299** (2003) 1391–1394
4. Margulies, E.H., Blanchette, M., NISC Comparative Sequencing Program, Haussler, D., Green, E.D.: Identification and characterization of multi-species conserved sequences. Genome Res **13** (2003) 2507–2518
5. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., Sidow, A.: Distribution and intensity of constraint in mammalian genomic sequence. Genome Res **15** (2005) 901–913
6. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al.: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res **15** (2005) 1034–1050
7. Nielsen, R., Yang, Z.: Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148** (1998) 929–936
8. Yang, Z., Nielsen, R.: Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol **19** (2002) 908–917
9. Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D., Kejariwal, A., Todd, M.A., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B., et al.: Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science **302** (2003) 1960–1963

10. Forsberg, R., Christiansen, F.B.: A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. Mol Biol Evol **20** (2003) 1252–1259
11. Guindon, S., Rodrigo, A.G., Dyer, K.A., Huelsenbeck, J.P.: Modeling the site-specific variation of selection patterns along lineages. Proc Natl Acad Sci U S A **101** (2004) 12957–12962
12. Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., et al.: A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol **3** (2005) e170
13. ENCODE Project Consortium: The ENCODE (ENCyclopedia Of DNA Elements) Project. Science **306** (2004) 636–640
14. Felsenstein, J., Churchill, G.A.: A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol **13** (1996) 93–104
15. Yang, Z.: A space-time process model for the evolution of DNA sequences. Genetics **139** (1995) 993–1005
16. Siepel, A., Haussler, D.: Phylogenetic hidden Markov models. In Nielsen, R., ed.: Statistical Methods in Molecular Evolution. Springer, New York (2005) 325–351
17. Cooper, G.M., Brudno, M., Stone, E.A., Dubchak, I., Batzoglou, S., Sidow, A.: Characterization of evolutionary rates and constraints in three mammalian genomes. Genome Res **14** (2004) 539–548
18. McAuliffe, J.D., Pachter, L., Jordan, M.I.: Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. Bioinformatics **20** (2004) 1850–1860
19. Siepel, A., Haussler, D.: Computational identification of evolutionarily conserved exons. In: Proc. 8th Int'l Conf. on Research in Computational Molecular Biology. (2004) 177–186
20. Holmes, I., Bruno, W.J.: Evolutionary HMMs: a Bayesian approach to multiple alignment. Bioinformatics **17** (2001) 803–820
21. Lunter, G., Miklos, I., Drummond, A., Jensen, J.L., Hein, J.: Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics **6** (2005) 83
22. Blanchette, M., Green, E.D., Miller, W., Haussler, D.: Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res **14** (2004) 2412–2423
23. Zheng, Q.: On the dispersion index of a Markovian molecular clock. Math Biosci **172** (2001) 115–128
24. Jukes, T.H., Cantor, C.R.: Evolution of protein molecules. In Munro, H., ed.: Mammalian Protein Metabolism. Academic Press, New York (1969) 21–132
25. Gillespie, J.: Lineage effects and the index of dispersion of molecular evolution. Mol Biol Evol **6** (1989) 636–647
26. Felsenstein, J.: Evolutionary trees from DNA sequences. J Mol Evol **17** (1981) 368–376
27. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al.: Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res **14** (2004) 708–715
28. Felsenstein, J.: Inferring Phylogenies. Sinauer Associates, Inc., Sunderland, Massachusetts (2004)
29. Nielsen, R., Huelsenbeck, J.P.: Detecting positively selected amino acid sites using posterior predictive P-values. Pac Symp Biocomput (2002) 576–588
30. Hasegawa, M., Kishino, H., Yano, T.: Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol **22** (1985) 160–174
31. Smith, N.G.C., Brandstrom, M., Ellegren, H.: Evidence for turnover of functional noncoding DNA in mammalian genome evolution. Genomics **84** (2004) 806–813
32. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, UK (1998)
33. Kimura, M.: A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol **16** (1980) 111–120
34. Yang, Z.: Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol **39** (1994) 306–314