Subtree power analysis and species selection for comparative genomics

Jon D. McAuliffe[*] Michael I. Jordan[*†] Lior Pachter[‡§]

April 4, 2005

Departments of [*]Statistics and [‡]Mathematics and [†]Division of Computer Science, University of California, Berkeley, CA 94720

[§]To whom correspondence should be addressed at Department of Mathematics, University of California, 970 Evans Hall #3840, Berkeley, CA 94720-3840.
Phone: (510) 642–2028
Fax: (510) 642–8204
Email: lpachter@math.berkeley.edu

Classification 1: Biological Sciences/Genetics
Classification 2: Physical Sciences/Statistics

Number of text pages: 17
Number of figures: 2 + 1 supplemental
Number of tables: 2 + 1 supplemental
Words in abstract: 112
Characters in paper: 28,521

Abbreviations: SST, symmetric star topology; FOSST, fully-observed symmetric star topology; HASST, hidden-ancestor symmetric star topology; MPSS, most-powerful Steiner subtree; MDSS, most-divergent Steiner subtree.

Abstract

Sequence comparison across multiple organisms aids in the detection of regions under selection. However, resource limitations require a prioritization of genomes to be sequenced. This prioritization should be grounded in two considerations: the lineal scope encompassing the biological phenomena of interest, and the optimal species within that scope for detecting functional elements. We introduce a statistical framework for optimal species subset selection, based on maximizing power to detect conserved sites. Analysis of a phylogenetic star topology shows theoretically that the optimal species subset is not in general the most evolutionarily diverged subset. We then demonstrate this empirically in a study of vertebrate species. Our results suggest that marsupials are prime sequencing candidates.

1 Introduction

Comparative genomic methods can reveal conserved regions in multiple organisms, including functional elements undetected by single-sequence analyses [1, 2]. Individual studies have demonstrated the effectiveness of genomic comparison for specific regions and elements [3, 4, 5, 6, 7]. Such successes indicate that comparative considerations should play a major role in decisions about what unsequenced species to sequence next. For comparative purposes, sequencing choices must first of all be guided by specification of the widest range of species sharing the function or character in question, which we call the lineal scope [8]. Boffelli et al. [10] discuss the utility of comparisons in lineal scopes ranging from the primate clade to the vertebrate tree.

Most lineal scopes selected in practice will include far more extant species than can be sequenced with today's resources. Thus, sequencing prioritization is an unavoidable issue, both for smaller-scale efforts targeting particular regions and for whole-genome projects, whose focus should reflect in part the aggregate needs of comparative analyses. Few studies on comparative methods provide a quantitative framework for decision-making about what to sequence. An exception is the work of Sidow and others [9, 11]: given a set of sequenced organisms and an inferred phylogeny, Cooper et al. [9] argue that decisions should be based on maximizing additive evolutionary divergence in a phylogenetic tree.

While additive divergence captures part of the problem underlying organism choice, it fails to reflect the inherent tradeoff that characterizes the problem. On the one hand, the success of procedures for assessing conservation does depend on sufficient evolutionary distance among the sequences [5, 4, 12]. On the other hand, a given set of species may have diverged too far from one another to be useful, even when orthology is preserved: in the limit of large evolutionary distance, conservation and nonconservation are just as indistinguishable as at distance zero [13]. Furthermore, phylogenetic topology affects the power of comparative methods in counterintuitive ways.

Here, we present a decision-theoretic framework which captures these issues, providing a procedure for making systematic, quantitative choices of species to sequence. Statistical power is our optimality criterion for species selection. Thus, we measure the effectiveness of a species subset directly in terms of error rates for detecting and overlooking conservation at a single orthologous site. Measuring power disentangles effects due to the number of species used from effects due to relative evolutionary distances in the phylogeny. We illustrate these ideas theoretically, in an analysis of a star phylogeny, and practically, with an empirically-derived phylogeny on 21 representative vertebrate species. The results indicate that adding the dunnart or a closely-related marsupial to finished and underway vertebrate sequences would most increase the power to detect conservation at single-nucleotide resolution.

2 Decision-Theoretic Setting

We study conservation detection in the following decision-theoretic setting. The data $\mathbf{x}$ are the nucleotides at an orthologous site across a set of species, i.e., an ungapped alignment column. We view these bases as corresponding to the leaves of a phylogeny with unobserved ancestral bases. We take as given the phylogenetic topology, the Markov substitution process along the branches, and the branch lengths. The phylogeny induces the observed-data probability distribution $p(\mathbf{x}; r)$ as the marginal distribution on its leaves, which can be evaluated efficiently for any $\mathbf{x}$ and r [14]. The parameter $r > 0$ is an unknown global mutation rate shared among all branches. We choose two threshold values $r_N > r_C$ for r: an actual mutation rate of at least r_N corresponds by definition to a nonconserved site, whereas a rate no more than r_C means the site is strongly conserved. When $r_N > r > r_C$, the conservation is too weak to interest us.

The decision-theoretic goals are now twofold. First, fixing a set of species, we wish to select a decision rule $\delta(\mathbf{x})$ which declares the site either nonconserved ($\delta(\mathbf{x}) = 0$) or conserved ($\delta(\mathbf{x}) = 1$) using only data from those species. Every nontrivial $\delta(\mathbf{x})$ will have positive probability of making two mistakes: when $r \geq r_N$, $P_r(\delta(\mathbf{X}) = 1)$ is the probability it erroneously detects conservation, and when $r \leq r_C$, $P_r(\delta(\mathbf{X}) = 0)$ is the probability it overlooks conservation. Minimizing these probabilities guides our choice of $\delta(\mathbf{x})$. We formulate a Neyman-Pearson hypothesis test [15] of the null hypothesis $H_0 : r \geq r_N$ versus the alternative hypothesis $H_A : r \leq r_C$, stipulating a maximum allowed probability α of falsely rejecting H_0 (falsely declaring conservation). While control of this error probability is a central concern [9], we also want to find a test $\delta(\mathbf{x})$ with large power to detect conservation, or equivalently small probability of overlooking conservation.

The second goal is to maximize the power of $\delta(\mathbf{x})$ over the choice of species subset in the larger phylogeny determined by the chosen lineal scope. This amounts to choosing a subtree in the phylogeny, with the chosen species as its leaves. The choice of subtree determines the distribution of $\mathbf{x}$ and hence the power of $\delta(\mathbf{x})$. For example, we might optimize over all subtrees on k existing species within the anthropoid clade, where k is determined by sequencing resource limitations.

3 Symmetric Star Topology

We first develop intuition for the species selection problem in a phylogenetic setting called the symmetric star topology (SST). Here, k existing species are connected to a single ancestor by branches of common length $t > 0$. Choosing k and t in the SST is like choosing k existing species within a larger phylogeny, such that each pair of chosen species is at a distance of approximately $2t$. We consider the fully-observed SST (FOSST), where the ancestral base x_0 is known, and the hidden-ancestor SST (HASST), where it is not. The HASST is of some practical interest, and the FOSST is useful because it approximates the HASST for small to moderate t. This follows because there is little uncertainty about the ancestral base at short evolutionary distances: with high probability, it equals the most-occurring base among the descendants.

We use the Jukes-Cantor substitution process along each branch in the SST. For each k and t, the probability associated with the FOSST observation $(x_0, \mathbf{x})$ is

$$p(x_0, \mathbf{x}; r) = \frac{1}{4} \left(\frac{1 + 3e^{-4rt}}{4} \right)^{n(x_0,\mathbf{x})} \left(\frac{3(1 - e^{-4rt})}{4} \right)^{k-n(x_0,\mathbf{x})} . \tag{1}$$

Here, r is the unknown mutation rate at the site, and $n(x_0, \mathbf{x})$ counts the number of descendant bases that agree with the ancestral base. We use the Jukes-Cantor equilibrium distribution (the uniform distribution) on x_0.

The FOSST likelihood-ratio statistic for testing $H_0 : r \geq r_N$ vs. $H_A : r \leq r_C$ therefore has the form

$$\frac{(1 + 3e^{-4r_C t})^{n(x_0,\mathbf{x})} (1 - e^{-4r_C t})^{k-n(x_0,\mathbf{x})}}{(1 + 3e^{-4r_N t})^{n(x_0,\mathbf{x})} (1 - e^{-4r_N t})^{k-n(x_0,\mathbf{x})}} . \tag{2}$$

The likelihood-ratio test $\mathcal{T}_\alpha(x_0, \mathbf{x})$ rejects H_0 and declares conservation for large values of Eq. 2, with the rejection threshold chosen to insure a false-positive probability of at most α. As detailed in section 5, this test is uniformly most powerful. In other words, the likelihood-ratio test has the largest possible power to detect conservation in the FOSST, no matter what the unknown mutation rate r. This answers the question raised in section 2 of finding an optimal decision rule.

We derive two more properties of the FOSST likelihood-ratio test in section 5. First, its power is lower against the alternative $r = r_C$ than against any other alternative $r < r_C$, so studying the case $r = r_C$ provides conservative power bounds. Second, it is equivalent to the intuitive test which declares conservation when $n(x_0, \mathbf{x})$ is large, that is, when many descendant bases agree with the ancestral base.

The power of the FOSST and HASST likelihood-ratio tests can be computed exactly, as described in section 6. Figure 1 shows an example for fixed (r_C, r_N, α), as t and k vary. For each t, power increases monotonically in k, as one would expect. However, for each k, power does not increase monotonically in t. Instead, there is a unique power-maximizing branch length $t^*(k)$. The existence of $t^*(k)$ implies that maximizing additive divergence, as in [9], is suboptimal: for each k, the optimal tree has finite divergence $k \cdot t^*(k)$, rather than arbitrarily large divergence. On the other hand, as k increases, $t^*(k)$ stabilizes at a positive value (Figure 2), so the optimal divergence

$k \cdot t^*(k)$ does grow without bound as a function of k.

We can explain the existence of $t^*(k)$ in the SST, using the Jukes-Cantor process or any other substitution process which is stationary, Markov, and continuous, as follows. Fix k. At $t = 0$, the distribution $p(\mathbf{x}; r)$ in an SST is the same for every r. Thus H_0 and H_A coincide. In this circumstance, the power is equal to α. As $t \rightarrow \infty$, the distribution of each descendant base approaches the process's stationary distribution, independent of the ancestral base. Since the stationary distribution does not involve r, all distributions in H_0 and H_A converge to the same limit. The limiting power in t is therefore again α. The fact that power begins at α when $t = 0$ and approaches α as $t \rightarrow \infty$, plus the fact that power is continuous in t and greater than α on $(0, \infty)$, implies a maximal power $t^*(k)$ must be attained.

Comparing Figures 1A and 1B shows that the likelihood-ratio test's power in the FOSST closely matches its power in the HASST, in a large interval around $t^*(k)$. For a given t and k, no HASST testing procedure can have higher power than the FOSST likelihood-ratio test, because the latter is optimal in the FOSST and uses more data (namely, x_0). These facts suggest the likelihood-ratio test should have very good power properties in the HASST. By analogy, we expect it to have good power in general phylogenies as well. We therefore proceed with likelihood-ratio testing in our empirical analysis.

4 Empirical Power Analysis

We explored subtree power maximization empirically, using the previously-reported CFTR sequence data [6] on 21 representative vertebrates (see Table 3, which is published as supporting information on the PNAS website). We constructed a multiple alignment using MAVID [16]. We then used maximum likelihood [17, 14] to fit a phylogenetic tree topology and branch lengths to the alignment. The fitted branch lengths were initially measured as expected substitutions at a neutrally-evolving site. However, we wanted the tree to have unscaled branch lengths (branch lengths under $r = 1$) corresponding to a typical evolutionary rate for exons in vertebrate genomes, which is around half the neutral rate. This allows us to use $r_C = 1$ as an exonic conserved rate threshold, with $r_N = 2$ corresponding to neutral evolution and $r_N > 2$ to positive selection. We therefore divided all branch lengths in the initial tree by two. The result is shown in Figure 3, which is published as supporting information on the PNAS website. Both the phylogeny estimation and subsequent power analysis employed the nucleotide substitution process of Felsenstein [18], using a transition-transversion ratio of 2:1 and a uniform equilibrium nucleotide distribution.

There are two assumptions required to apply the testing procedure globally. First, we assume we can restrict the candidate phylogenetic distributions for a previously-unobserved alignment column to those with a certain fixed topology and relative branch lengths. Second, we assume the fixed topology and relative branch lengths can be inferred from data in a region which might not contain the column being tested. The first assumption validates the formulation of the test in terms of the scaling parameter r. It can be checked, for example, by fitting one tree for each of several rate categories, in cases where the appropriate rate category for a collection of sites is known with confidence. If the trees from each rate category happen to have the same topology and relative branch lengths, the first assumption appears reasonable. The second assumption validates the use of a tree estimated in one region for a test in another. It can be checked by fitting trees to different regions and comparing their topologies and relative branch lengths.

Cooper et al. [9] have checked both assumptions empirically in two regions, with three different rate categories: codons, UTRs, and non-exonic DNA. Fixing the consensus topology for their eight species, they found strong evidence that relative branch lengths are stable across rates and between regions. Similarly, Yap and Pachter [19] studied the stability of relative branch lengths in a whole-genome human/mouse/rat alignment, where the topology is known. They considered four rate categories: ancient repeats, exons, rodent-specific insertions, and strongly conserved sequence. They too found stable relative branch lengths across most regions, genome-wide. These results tend to justify the assumptions behind the subtree-power testing procedure. The fact that topologies were not inferred in these studies reflects the reality that, in practice, much is usually known about topology in advance.

The goal is now to maximize the likelihood-ratio test's power over subsets of size k chosen from the 21 species, for various values of k. This entails searching for the maximal-power family subtree, or k-most-powerful Steiner subtree (k-MPSS), among the $\binom{21}{k}$ subtrees with k leaves. A Steiner subtree on k leaves is the unique smallest subtree rooted at their last common ancestor.

Finding the k-MPSS is a combinatorial optimization problem, which we solve in small to moderate-sized cases by evaluating the power of the likelihood-ratio test using every candidate Steiner subtree. We can also solve the problem for larger k, by constraining the species at many of the leaves in the subtree. The power computation for a particular subtree is described in section 6.3.

Table 1 shows the k-MPSS (starred) in comparison to the subtree on k leaves with largest additive divergence (the k-most-divergent Steiner subtree, or k-MDSS, daggered). The latter has been the focus of previous work [4, 20, 9]. These two subtree selection criteria do not coincide. For instance, at $r_N = 2$, the 5-MPSS includes the dunnart, whereas the 5-MDSS instead uses the platypus. The t-statistic on the difference in power is 2.06, so variability in the power estimate is not a likely explanation. A more extreme example is $r_N = 10$: the 4-MPSS and 4-MDSS have only one species in common, and the absolute loss in power that results from using the 4-MDSS is nearly 8.5% (t-statistic 105.7). Here, more than 4,400 subtrees have higher power than the 4-MDSS. The disagreement at larger values of k, where subtree topology becomes more complicated, highlights the importance of including a realistic phylogenetic topology in the species selection procedure.

We carried out a similar comparison, under the constraint that the nine completely or partially sequenced vertebrates in the data set appear in the subtree (Table 2). This reveals the species whose addition to the current sequencing mix would most improve the power to detect single-site conservation. As in Table 1, the most-powerful and most-divergent subtrees generally differ. The differences in power are smaller than in Table 1. This may be because forcing half of the phylogeny's leaves to appear in the subtree limits the possible power increase from any subtree selection method.

The pattern of disagreement between MDSS and MPSS in Table 2 is not systematic: when $r_N = 5$, for example, they disagree at 10 and 11 species, agree at 12 and 13, and disagree at 14. Table 1 exhibits similar properties. Thus, the MDSS is not a reliable approximation to the MPSS. Table 2 reveals that the single most beneficial species to sequence next is the dunnart (improving power by a relative 12.5%), whereas the species which adds the most evolutionary divergence is the platypus.

Our fitted phylogenetic topology differs slightly from estimates based on considerations of large-scale indel mutations and morphology, for example in its placement of the chicken and platypus. At issue here, however, is its suitability for a single-site power analysis under a substitutional mutation model. We chose our tree estimation procedure to obtain a phylogeny directed to this goal.

5 Monotone Likelihood Ratio in the FOSST

Here we derive properties of the FOSST likelihood-ratio test, using the notion of a monotone likelihood ratio [15]. Fix k and t. Let $\mathcal{P} = \{p(x_0, \mathbf{x}; r) : r > 0\}$ be the family of FOSST probability mass functions using the Jukes-Cantor substitution process, indexed by rate parameter r. Then $\mathcal{P}$ has a monotone likelihood ratio in the statistic $n(x_0, \mathbf{x})$: for each pair of rate parameters $r_N > r_C > 0$, Eq. 2 is an increasing function of $n = n(x_0, \mathbf{x}) \in \{0, 1, \ldots, k\}$. This follows upon observing that, because $r_N > r_C$,

$$\frac{1 + 3e^{-4r_C t}}{1 + 3e^{-4r_N t}} > 1 \quad \text{and} \quad \frac{1 - e^{-4r_C t}}{1 - e^{-4r_N t}} < 1 \ .$$

It is now a standard result of monotone likelihood-ratio theory that the likelihood-ratio test is uniformly most powerful. The theory also implies that the power function $r \mapsto P_r\big\{T_\alpha(X_0, \mathbf{X}) \text{ rejects}\big\}$ is monotonic in r. From this we conclude that the size α is attained at the null distribution $r = r_N$, and the lowest power is attained at the alternative distribution $r = r_C$. As a further consequence of the monotone likelihood ratio, the likelihood-ratio test is equivalent to rejecting for large values of $n(x_0, \mathbf{x})$. This is an intuitive procedure, which declares conservation when few descendant bases have mutated.

6 Power Calculations

6.1 FOSST

In this section we employ the Jukes-Cantor substitution process. The power of $\mathcal{T}_\alpha(x_0, \mathbf{x})$ against the particular alternative $r = r_C$ can be written explicitly as a function of k and t:

$$\rho(k,t) = G_A(n_\alpha + 1; k) + \left(\frac{\alpha - G_0(n_\alpha + 1; k)}{f_0(n_\alpha; k)} \right) f_A(n_\alpha; k) \ . \tag{3}$$

The notation in Eq. 3 is defined as follows. $f_0(\cdot; k)$ is the probability mass function of a binomial random variable with k trials and success probability $d(r_N, t) = (1 + 3\exp(-4r_N t))/4$. $f_A(\cdot; k)$ is the same, but using $d(r_C, t)$. $G_0(\cdot; k)$ and $G_A(\cdot; k)$ are the corresponding cumulative binomial right-tail probabilities, and n_α is a known critical value. To derive Eq. 3, recall from section 5 that $\mathcal{T}_\alpha(x_0, \mathbf{x})$ is equivalent to the test which rejects H_0 when the statistic $n(x_0, \mathbf{x})$ exceeds a corresponding n_α. Both tests thus have the same power $\rho(k, t)$. Let P_0 and P_A denote the distribution of $n(X_0, \mathbf{X})$ under $r = r_N$ (the size-determining distribution) and $r = r_C$, respectively. Because $n(x_0, \mathbf{x})$ can take on only finitely many values, we use randomized rejection to achieve level exactly α. The critical value is $n_\alpha = \min\{n : P_0(n(X_0, \mathbf{X}) > n) \leq \alpha\}$. When $n(x_0, \mathbf{x}) > n_\alpha$, we reject. When $n(x_0, \mathbf{x}) = n_\alpha$, we reject with probability $\gamma(\alpha)$ satisfying

$$P_0(n(X_0, \mathbf{X}) > n_\alpha) + \gamma(\alpha) P_0(n(X_0, \mathbf{X}) = n_\alpha) = \alpha \ . \tag{4}$$

This implies that setting

$$\gamma(\alpha) = \frac{\alpha - P_0(n(X_0, \mathbf{X}) > n_\alpha)}{P_0(n(X_0, \mathbf{X}) = n_\alpha)} \tag{5}$$

guarantees a test with size α. It now follows that

$$\rho(k,t) = P_A(n(X_0,\mathbf{X}) > n_\alpha) + \gamma(\alpha)P_A(n(X_0,\mathbf{X}) = n_\alpha) \ . \tag{6}$$

Each descendant nucleotide X_i has probability $d(r,t)$ of differing from X_0, independent of all other descendants. Thus $n(X_0,\mathbf{X})$ is a binomial random variable with k trials and success probability $d(r,t)$. Eq. 3 follows upon substituting $G_0(n_\alpha+1;k)$ for $P_0(n(X_0,\mathbf{X}) > n_\alpha)$, $f_0(n_\alpha;k)$ for $P_0(n(X_0,\mathbf{X}) = n_\alpha)$, and similarly for P_A.

Eq. 3 involves only known constants and binomial probabilities, which can be evaluated quickly to desired accuracy [21]. This allows us to compute $\rho(k,t)$ for many choices of k and t, leading to the power curves in Figure 1A. The kinks in each power curve correspond to values of t at which the critical value of the likelihood-ratio test changes. The locations of the kinks are easily determined, and the power curves are smooth between kinks. Thus, we can find $t^*(k)$ and $\rho^*(k)$ rapidly using numerical optimization (Figure 1A, Figure 2A).

6.2 HASST

We use the Jukes-Cantor substitution process in this section also. Here, the likelihood-ratio statistic has the form

$$\frac{\sum_{x_0}(1+3e^{-4r_C t})^{n(x_0,\mathbf{x})}(1-e^{-4r_C t})^{k-n(x_0,\mathbf{x})}}{\sum_{x_0}(1+3e^{-4r_N t})^{n(x_0,\mathbf{x})}(1-e^{-4r_N t})^{k-n(x_0,\mathbf{x})}} \ . \tag{7}$$

This is more difficult to deal with than Eq. 2. It is clear that Eq. 7 depends only on the occurrence counts of the four different bases, not on the leaf configuration which gives rise to the counts. Furthermore, Eq. 7 is invariant when the bases associated with the counts are permuted. This means that there are only as many distinct values of Eq. 7 as there are integer partitions of k into four parts (n_1,n_2,n_3,n_4), with partition values of zero allowed. The number of leaf configurations corresponding to each integer partition is the combinatorial quantity

$$\binom{4}{\tilde{n}_1\ \tilde{n}_2\ \tilde{n}_3\ \tilde{n}_4}\binom{k}{n_1\ n_2\ n_3\ n_4} \ , \tag{8}$$

where $(\tilde{n}_1, \tilde{n}_2, \tilde{n}_3, \tilde{n}_4)$ counts repetitions in (n_1, n_2, n_3, n_4). For example, if $(n_1, n_2, n_3, n_4) = (11, 6, 4, 4)$, then $(\tilde{n}_1, \tilde{n}_2, \tilde{n}_3, \tilde{n}_4) = (2, 1, 1, 0)$. We can generate all the required integer partitions quickly, even for k in the hundreds.

Multiplying the HASST probability mass at each integer partition by the partition's corresponding value of Eq. 8 results in the exact probability mass function of the likelihood-ratio statistic. Thus, we can compute the null distribution ($r = r_N$) and alternative distribution ($r = r_C$) of Eq. 7, for each required setting of (α, r_N, k, t). This yields the power of the HASST likelihood-ratio test, using Eq. 5 and Eq. 6 with the HASST distribution functions substituted for P_0 and P_A. We then maximize each curve $\rho(k, \cdot)$ numerically to determine $t^*(k)$ and $\rho^*(k)$ (Figure 1B, Figure 2B).

6.3 General Phylogenies

The empirical analysis of section 4 uses a general topology, with the Felsenstein substitution process. Let $x_1, \ldots, x_k$ be a leaf subset of size k in such a phylogeny, and let $a_{k+1}, \ldots, a_{2k-1}$ be the $k-1$ ancestral nodes in the corresponding Steiner subtree. The likelihood-ratio statistic based on the leaf subset has the form

$$\frac{\sum_{a_{k+1}} \cdots \sum_{a_{2k-1}} \prod_{i=k+1}^{2k-1} p(a_i |\, a_{\pi(i)}, r_C t_i) \prod_{j=1}^{k} p(x_j |\, a_{\pi(j)}, r_C t_j)}{\sum_{a_{k+1}} \cdots \sum_{a_{2k-1}} \prod_{i=k+1}^{2k-1} p(a_i |\, a_{\pi(i)}, r_N t_i) \prod_{j=1}^{k} p(x_j |\, a_{\pi(j)}, r_N t_j)} . \tag{9}$$

Here $\pi(m)$ is the index of node m's parent, t_m is the length of the branch coming into node m, and $p(y|\, x, rt)$ is the Felsenstein substitution probability for base y starting from base x over evolutionary distance rt [18]. The numerator and denominator can be computed efficiently using the Felsenstein pruning algorithm [14].

To compute the power of a test based on Eq. 9 in section 4, we used a Monte Carlo strategy. For each setting of r_N, with $\alpha = 0.05$, we generated 100,000 realizations from the null ($r = r_N$) and alternative ($r = r_C$) distributions on the leaves of the full phylogeny. This induced null and alternative empirical distributions on the leaves of every possible subtree. From these we obtained approximations to the true null and alternative distributions of the likelihood-ratio statistic. This yielded approximate critical values, as well as power estimates. We repeated the whole process ten times for each parameter setting, to assess variability in the Monte Carlo procedure.

7 Discussion

Our decision-theoretic point of view puts the focus on the important issue in detecting conservation: the two kinds of discrimination errors and their probabilities. The probability of falsely declaring conservation is controlled at a specified level α, and subject to this constraint the probability of overlooking conservation is minimized (power is maximized). Within a given lineal scope, species chosen according to the decision-theoretic power criterion generally differ from those chosen to maximize total evolutionary divergence. We have demonstrated this difference both theoretically and empirically.

Even when the most powerful set of species coincides with the most divergent set, the power calculation is more relevant: it measures the marginal benefit of additional sequenced species as an increase in detection probability. This lets us choose a species count k which optimizes the tradeoff between the benefit of detecting conservation and the cost of additional sequencing. Additive divergence, on the other hand, does not directly measure anything intrinsic to the problem of conservation detection.

Since the phylogeny and substitution process are parameters of our procedure, their choice can and should be tailored to particular investigations. Our emphasis on single-site detection of conservation will lead to conservative power estimates in situations where conservation is tested for simultaneously across multiple sites. However, modeling multiple-site detection requires additional assumptions on the form of across-site dependence, which we have avoided.

8 Acknowledgments

We thank Peter Bickel and Adam Siepel for helpful comments. M.J. was supported by a grant from the NIH (R33-HG003070). L.P. was supported by a grant from the NIH (R01-HG2362-3), a Sloan Foundation Research Fellowship, and an NSF Career award (CCF-0347992).

References

[1] Mouse Genome Sequencing Consortium. (2002) *Nature* **420**, 520–562.

[2] Rat Genome Sequencing Consortium. (2004) *Nature* **428**, 493–521.

[3] Flint, J, Tufarelli, C, Peden, J, Clark, K, Daniels, R, Hardison, R, Miller, W, Philipsen, S, Tan-Un, K. C, McMorrow, T, et al. (2001) *Hum. Mol. Genet.* **10**, 371–382.

[4] Boffelli, D, McAuliffe, J, Ovcharenko, D, Lewis, K. D, Ovcharenko, I, Pachter, L, & Rubin, E. M. (2003) *Science* **299**, 1391–1394.

[5] Dermitzakis, E. T, Reymond, A, Scamuffa, N, Ucla, C, Kirkness, E, Rossier, C, & Antonarakis, S. E. (2003) *Science* **302**, 1033–1035.

[6] Thomas, J. W, Touchman, J. W, Blakesley, R. W, Bouffard, G. G, Beckstrom-Sternberg, S. M, Margulies, E. H, Blanchette, M, Siepel, A. C, Thomas, P. J, McDowell, J. C, et al. (2003) *Nature* **424**, 788–793.

[7] Chapman, M. A, Donaldson, I. J, Gilbert, J, Grafham, D, Rogers, J, Green, A. R, & Göttgens, B. (2004) *Genome Res.* **14**, 313–318.

[8] O'Brien, S. J, Eizirik, E, & Murphy, W. J. (2001) *Science* **292**, 2264–2266.

[9] Cooper, G. M, Brudno, M, NISC Comparative Sequencing Program, Green, E. D, Batzoglou, S, & Sidow, A. (2003) *Genome Res.* **13**, 813–820.

[10] Boffelli, D, Nobrega, M, & Rubin, E. M. (2004) *Nature Rev. Genet.* **5**, 456–465.

[11] Sidow, A. (2002) *Cell* **111**, 13–16.

[12] Margulies, E. H, Blanchette, M, NISC Comparative Sequencing Program, Haussler, D, & Green, E. D. (2003) *Genome Res.* **13**, 2507–2518.

[13] Zhang, L. G, Pavlovic, V, Cantor, C. R, & Kasif, S. (2003) *Genome Res.* **13**, 1190–1202.

[14] Felsenstein, J. (1981) *J. Mol. Evol.* **17**, 368–376.

[15] Lehmann, E. (1986) *Testing Statistical Hypotheses*. (Springer-Verlag), 2nd edition.

[16] Bray, N & Pachter, L. (2004) *Genome Res.* **14**, 693–699.

[17] Olsen, G. J, Matsuda, H, Hagstrom, R, & Overbeek, R. (1994) *Comput. Appl. Biosci.* **10**, 41–48.

[18] Felsenstein, J & Churchill, G. A. (1996) *Mol. Biol. Evol.* **13**, 93–104.

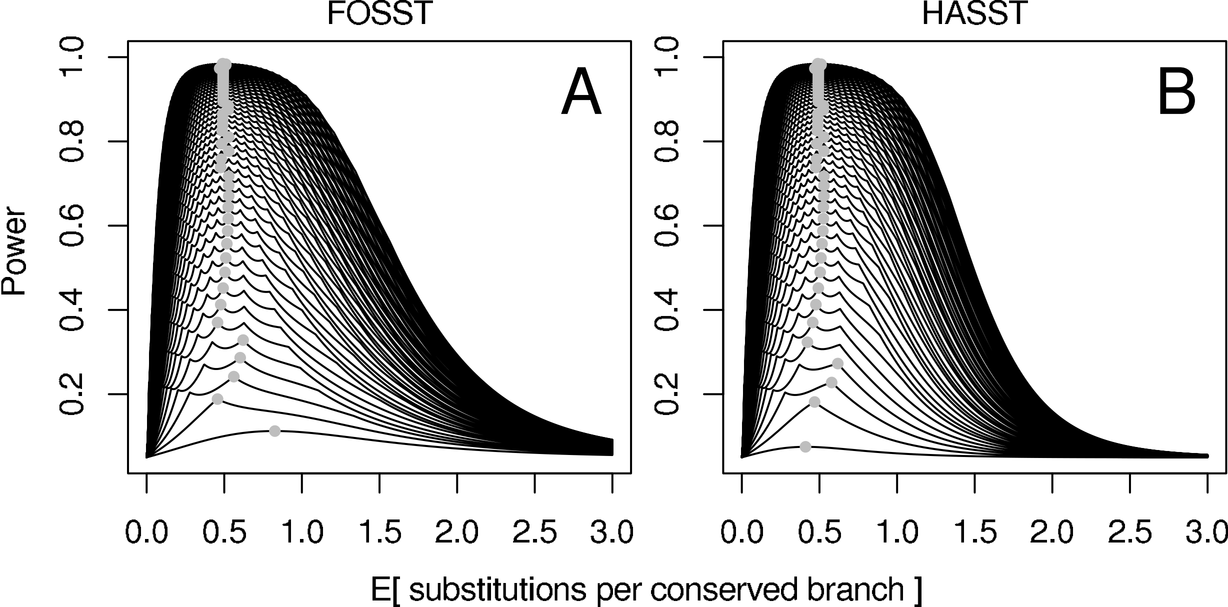
[19] Yap, V. B & Pachter, L. (2004) *Genome Res.* **14**, 574–579.

[20] McAuliffe, J. D, Pachter, L, & Jordan, M. I. (2004) *Bioinformatics* **20**, 1850–1860.

[21] Abramowitz, M & Stegun, I. (1974) *Handbook of Mathematical Functions*. (Dover Publications, New York).

Figure 1. Power to detect conservation as a function of common branch length for the fully-observed (A) and hidden-ancestor (B) SSTs, using $r_C = 1$, $r_N = 2$, and $\alpha = 0.05$. Each power curve corresponds to an even number k of observed descendant species, from two (bottommost curve) to 100 (topmost). The maximum power attained for each k is indicated by a grey dot. The power against the alternative $r = r_C$ is shown; power against any other alternative is larger. Curves computed with other values of r_N and α remain qualitatively the same (not shown).

Figure 2. The optimal common branch length $t^*(k)$ in the fully-observed (A) and hidden-ancestor (B) SSTs, as a function of the number of descendant species k. Each black curve uses the indicated nonconserved rate $r_N = 2, 3, 5, 7$ with $\alpha = 0.05$; grey curves are analogous with $\alpha = 0.01$. As k increases, $t^*(k)$ stabilizes at a value depending on r_N but not α. For the larger r_N's, the curves are terminated when power reaches 99.9%.


FOSST
HASST
A
B
Power
1.0
0.8
0.6
0.4
0.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
E[substitutions per conserved branch]

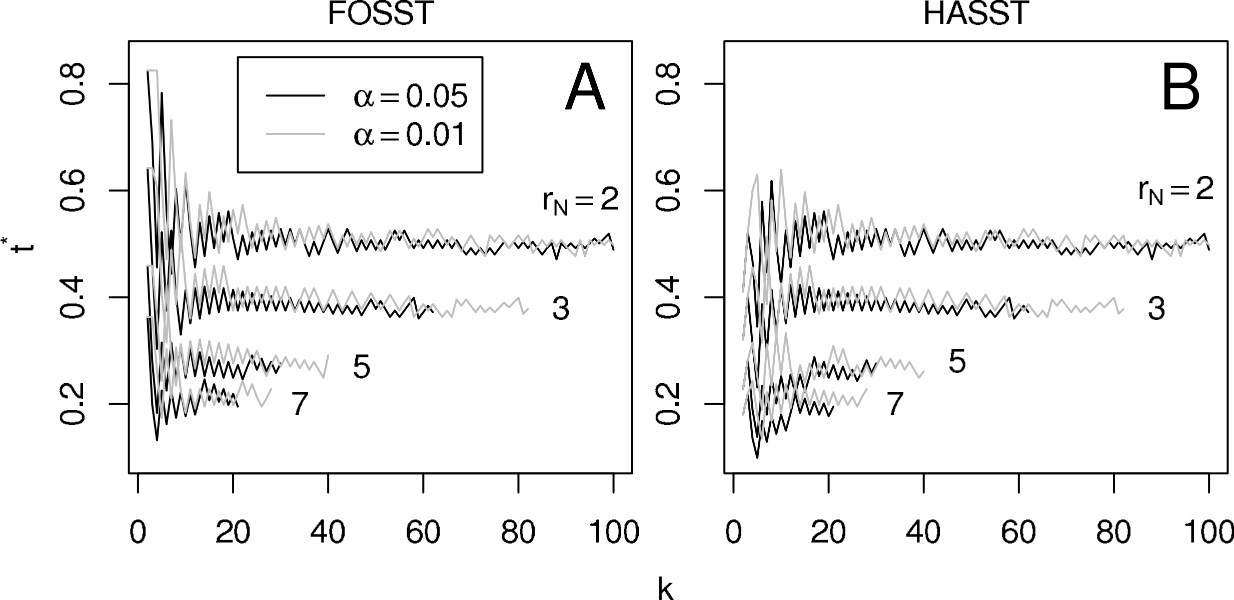

FOSST
HASST
A
B
α = 0.05
α = 0.01
$r_N = 2$
3
5
7
t^*
k
0.2
0.4
0.6
0.8
0
20
40
60
80
100

Table 1. The k-MPSS and k-MDSS as a function of the nonconserved rate r_N and the size k of the subtree, with $\alpha = 0.05$ throughout. Results are across 10 repetitions of the Monte Carlo power estimation procedure. The last three columns display the average power (and standard error), the t-statistic for the power difference between the k-MDSS and the k-MPSS (in cases where they differ), and the average power ranking (among all subtrees). Since r_C is calibrated to exonic conservation, the settings of r_N range from a neutral rate ($r_N = 2$) (19) towards extreme single-site mutability.

				t vs.	
r_N	Size	Species: $\star$ = MPSS, † = MDSS	Power% (SE)	MPSS	Rank
2	2	Rat, Zebrafish $\star$†	6.79 (0.01)		1.3
	3	Rat, Zebrafish, Chicken $\star$†	8.30 (0.01)		1.6
	4	Rat, Zebrafish, Chicken, Dog $\star$†	9.61 (0.02)		3.3
	5	Rat, Zebrafish, Chicken, Dog, Dunnart $\star$	10.88 (0.03)		4.4
		Rat, Zebrafish, Chicken, Dog, Platypus †	10.80 (0.02)	2.06	21.7
5	2	Rat, Zebrafish $\star$†	10.60 (0.02)		3.2
	3	Rat, Zebrafish, Chicken $\star$†	21.61 (0.06)		1.8
	4	Rat, Zebrafish, Chicken, Dog $\star$†	39.33 (0.17)		5.2
	5	Rabbit, Cat, Dunnart, Chicken, Hedgehog $\star$	49.96 (0.07)		12.2
		Rat, Zebrafish, Chicken, Dog, Platypus †	47.31 (0.07)	25.82	3894.4
10	2	Dunnart, Lemur $\star$	13.30 (0.03)		21.0
		Rat, Zebrafish †	12.67 (0.02)	16.67	153.0
	3	Dunnart, Cat, Zebrafish $\star$	37.53 (0.11)		10.4
		Rat, Zebrafish, Chicken †	36.83 (0.12)	4.13	77.2
	4	Dunnart, Chicken, Hedgehog, Opossum $\star$	64.69 (0.05)		4.4
		Rat, Zebrafish, Chicken, Dog †	56.21 (0.06)	105.70	4439.3
	5	Macaque, Lemur, Dog, Cow, Pig $\star$	69.75 (0.11)		8.6
		Rat, Zebrafish, Chicken, Dog, Platypus †	66.86 (0.07)	22.28	4867.4

Table 2. The k-MPSS and k-MDSS, under the constraint that the following nine species are included in the subtree: human, mouse, rat, chimpanzee, dog, chicken, fugu, zebrafish, and tetraodon. The scheme of the table is the same as Table 1.

				t vs.	
r_N	Size	New species: ⋆ = MPSS, † = MDSS	Power% (SE)	MPSS	Rank
2	9	{*clamped species only*}	12.81 (0.03)		
	10	Dunnart ⋆	14.42 (0.04)		1.1
		Platypus †	14.25 (0.04)	2.92	3.4
	11	Dunnart, Platypus ⋆	16.08 (0.05)		1.6
		Platypus, Hedgehog †	15.85 (0.04)	3.62	6.2
	12	Dunnart, Platypus, Hedgehog ⋆†	17.88 (0.06)		1.5
	13	Dunnart, Platypus, Hedgehog, Rabbit ⋆†	19.80 (0.08)		1.1
	14	Dunnart, Platypus, Hedgehog, Rabbit, Cow ⋆†	21.41 (0.08)		1.6
5	9	{*clamped species only*}	56.44 (0.16)		
	10	Dunnart ⋆	65.59 (0.20)		1.0
		Platypus †	64.74 (0.17)	3.18	3.0
	11	Dunnart, Opossum ⋆	71.05 (0.09)		2.3
		Platypus, Hedgehog †	70.54 (0.06)	4.74	14.6
	12	Dunnart, Platypus, Hedgehog ⋆†	72.77 (0.08)		1.2
	13	Dunnart, Platypus, Hedgehog, Rabbit ⋆†	76.02 (0.13)		1.0
	14	Dunnart, Platypus, Hedgehog, Rabbit, Opossum ⋆	80.41 (0.10)		2.2
		Dunnart, Platypus, Hedgehog, Rabbit, Cow †	80.08 (0.14)	1.88	2.1
10	9	{*clamped species only*}	86.61 (0.06)		
	10	Platypus ⋆†	91.67 (0.06)		1.3
	11	Dunnart, Opossum ⋆	94.07 (0.02)		3.3
		Platypus, Hedgehog †	93.96 (0.03)	2.66	10.7
	12	Dunnart, Platypus, Rabbit ⋆	95.84 (0.03)		2.4
		Dunnart, Platypus, Hedgehog †	95.79 (0.30)	1.30	4.4
	13	Dunnart, Platypus, Rabbit, Opossum ⋆	97.31 (0.02)		4.6
		Dunnart, Platypus, Rabbit, Hedgehog †	97.29 (0.02)	0.85	6.6
	14	Dunnart, Platypus, Rabbit, Hedgehog, Opossum ⋆	97.99 (0.01)		2.4
		Dunnart, Platypus, Rabbit, Hedgehog, Cow †	97.95 (0.02)	1.83	7.6