# Effective species count and motif efficiency: The value of comparative genomics in characterizing conserved sequence positions

Lee A. Newberg*

Technical Report 07-09
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York 12180-3590, USA

August 3, 2007

## Abstract

**Background:** The identification and characterization of functional, non-coding DNA sequence elements is key to the understanding of cell function, differentiation, and pathology because the elements affect when and to what extent nearby genes are expressed. The proliferation of completed genomic sequences during the past few years has provided impetus for numerous comparative-genomics efforts to identify such elements, while simultaneously underscoring the profound difficulty of accurate and exhaustive identification. In particular, when there is little evolutionary separation between the species, the data from phylogenetically related sites are significantly correlated, and the advantage of having multiple genomes is significantly diminished. Little work has been done to quantify the utility of obtaining additional genomes for the characterization of a DNA motif.

**Results:** We provide a mathematical formalism and an algorithm for evaluating a phylogenetic tree in terms of its utility for constructing a nucleotide equilibrium probability distribution for each multiply aligned DNA sequence position. "Motif efficiency" is measured via Fisher Information and the Cramér-Rao Inequality, and is scaled so that a set of indistinguishable genomes is deemed to have a 0% motif efficiency, and a set of well-separated genomes is deemed to have 100% motif efficiency. We analyze several standardized phylogenetic trees and several phylogenetic trees from the literature.

**Conclusions:** In our analysis of the standardized phylogenetic trees, we find that inadequate species separation is a particular matter for concern when the number of species is large or when the DNA sequence positions to be characterized have a nucleotide equilibrium probability distribution that is dominated by a pair of nucleotides. In our analysis of phylogenetic trees from the literature, we find that for a phylogenetic tree of nine mammals and for a phylogenetic tree of 45 vertebrates, motif efficiency is around 10%, and that, for a set of 14 prokaryotes, motif efficiency is around 33%.

**Availability:** A web server that analyzes phylogenetic trees for their effective species count and motif efficiency is available at `http://bayesweb.wadsworth.org/cgi-bin/Effective.pl`.

## Background

There is an explosion in the number of genomes being sequenced. While much effort has been focused on sequencing the genomes of widely divergent species, recently there has also been a focus on sequencing the genomes of closely related species, with the objective of comparing and contrasting them for subtle differences. There has been some work towards quantifying the utility of multiple genomes for the *detection* of conserved DNA regions. However, we know of no attempts to quantify the utility of multiple genomes for the purpose of *characterizing* a DNA motif. That is, existing analyses can tell us whether a set of related genomes will likely reveal the locations of DNA conserved regions, but these analyses do not indicate the accuracy to which we will be able to describe a position of a conserved motif in terms of a nucleotide equilibrium probability distribution or position frequency matrix.

### Our Previous Work in Conservation Detection

Because we build upon it in the current article, we give some detail of our previous work. A phylogenetic tree of nine mammalian genomes being sequenced was analyzed by us for its

*Also: The Center for Bioinformatics, Wadsworth Center, New York State Department of Health, Albany, NY 12208-3425, USA.

ability to reconstruct neutral multiply aligned DNA sequence positions [1]. Mathematically, our analysis addressed the following situation. Suppose that nucleotides are generated uniformly at random for the genome of the common ancestor of the nine mammals, with the nine mammalian genomes then being generated by use of the given phylogenetic tree and the nucleotide substitution model of Jukes & Cantor (1969) [2]. Suppose then, that we seek the nucleotide substitution model that best explains the generated data and, by some distance metric, measure how different that model is from the original model used to generate the genomes. Suppose further that this thought experiment is repeated many times, and that the resulting squared distances are averaged to give a model estimator variance. When the model estimator variance is high, then the phylogenetic tree is not efficient at reconstructing neutral sequence positions; when the model estimator variance is low, then the phylogenetic tree is efficient at reconstructing neutral sequence positions.

For the estimator model, in this earlier work we employed the nucleotide substitution model described by Felsenstein (1981) [3]. This is a model that is parameterized by a nucleotide equilibrium probability distribution, and it coincides with the model of Jukes & Cantor (1969) when it is parameterized with a uniform nucleotide equilibrium probability distribution. We defined the distance between the uniform distribution $\vec{\pi} = (0.25, 0.25, 0.25, 0.25)$, which was used to generate the genomes, and the nucleotide equilibrium probability distribution $\vec{\theta} = (\theta_A, \theta_C, \theta_G, \theta_T)$ that best explained the generated data, by

$$\text{distance} = \sqrt{\sum_{b \in A,C,G,T} (\theta_b - \pi_b)^2} \,. \qquad (1)$$

For a collection of distantly related genomes, the estimator variance of this distance measure is expected to be inversely proportional to the number of genomes; thus, we defined the "effective species count" in proportion to the reciprocal of the estimator variance, calibrated so that it gives the number of species in the limit where the species are very distantly related.

In a very similar manner, we also sought the nucleotide substitution model that best explained the generated data among those models that obey the Jukes & Cantor (1969) model, with an extra "conservation" parameter $\gamma$ that scaled all phylogenetic tree branch lengths. A phylogenetic tree was then judged by how well experiments would recover the value of $\gamma = 1$ that was used to generate random genomes of neutral DNA positions.

However, by computing only the ability to reconstruct neutral DNA positions, our analysis in did not describe how accurately functional DNA positions can be characterized; at best it quantified the detectability of conservation via the incompatibility of a conserved position's sequence data with the neutral/null model [1]. Furthermore, we employed simplistic nucleotide substitution models, Felsenstein (1981) [3] (which models non-uniform nucleotide equilibrium probability distributions in an *ad hoc* manner) and Jukes & Cantor (1969) [2] (with a parameter that models conservation in an *ad hoc* manner).

## Other Previous Work in Conservation Detection

Cooper *et al.* (2003) measured "relative genome information content" for a new genome relative to genomes already sequenced as the ratio of (a) the total of the "relative lengths" for the edges connecting the new genome to the phylogenetic tree of the sequenced genomes, divided by (b) the total of the relative lengths for the edges within the phylogenetic tree of sequenced genomes [4]. They defined the relative length of an edge as its length, as a fraction of the total of all edge lengths in a phylogenetic tree of relevant genomes, averaged over several phylogenetic trees constructed from different data sets. Pardi & Goldman (2005) also measured efficiency in terms of the total of the phylogenetic tree branch lengths, and showed that a greedy approach for the sequencing ordering of genomes was optimal [5]. However, while the total of branch lengths is an indication of the dispersion of the genomes, the approach of these two articles fails to distinguish phylogenetic trees with similar total branch lengths that, nonetheless, have different abilities to characterize a nucleotide equilibrium probability distribution.

McAuliffe *et al.* (2005) did a conservation detectability analysis similar to our $\gamma$-based analysis (described above [1]), but quantified detectability in terms of the probability of the rejection of the $\gamma = 1$ null model, rather than in terms of the estimator variance of the $\gamma$ model parameter [6]. Eddy (2005) performed a similar detection analysis, on phylogenetic trees with a star topology (defined later), but rather than focusing on a single nucleotide position at a time, explored whether a conserved feature of larger width could be distinguished from the null model [7].

Margulies *et al.* (2005) describes an economically efficient approach to detecting conserved regions of DNA sequence [8]. It shows that nucleotide for nucleotide (*i.e.*, dollar for dollar) the low redundancy sequencing of additional genomes is a useful first step in locating conserved regions in the species of interest.

None of the above approaches quantifies the ability of the genomes of species related by a phylogenetic tree to yield an accurate nucleotide equilibrium probability distribution for each multiply aligned DNA sequence position. We address this goal in the following.

## Approach

We imagine the repeated experiment described in the Introduction (and in [1]). However, in that earlier work we used the Felsenstein (1981) nucleotide substitution model, which

provides an *ad hoc* approach to non-uniform nucleotide equilibrium probability distributions that arise from selection pressures, and evaluated model parameter estimator variance near a uniform nucleotide equilibrium probability distribution [3]. Instead, we now adopt the model of Halpern & Bruno (1998) [9], which explicitly incorporates selection pressures by analyzing the probability of fixation of a nucleotide within a population, and we examine model parameter estimator variance at a number of nucleotide equilibrium probability distributions.

Algorithmically, we do not conduct the repeated experiment but we arrive at the same results via a Fisher Information approach, employing the Cramér-Rao Inequality. Furthermore, in addition to computing an effective species count, we now compute a motif efficiency for a phylogenetic tree. This is calculated as

$$\text{motif efficiency} = \frac{(\text{effective species count}) - 1}{(\text{number of species}) - 1} \,. \quad (2)$$

This formula allows us to compare on equal footing phylogenetic trees with different numbers of species. A phylogenetic tree in the limit where all branch lengths are zero has an effective species count of 1.0, and a phylogenetic tree in the limit where all lengths approach infinity has an effective species count equal to the number of species. Thus, the motif efficiency is a measure of how much of the possible increase in effective species count is achieved by a phylogenetic tree.

Although the approach works equally well on any phylogenetic tree, for demonstration purposes we apply the algorithm to standardized phylogenetic trees with a star topology, parameterized by a pairwise distance. The parameter is the distance between the members of any pair of species; the phylogenetic tree is a star in that, from a shared common ancestor, each species is at a distance that is half the parameter value. We observe how the pairwise distance and the number of species affect the effective species count and motif efficiency.

We also apply the algorithm to phylogenetic trees from the literature, a phylogenetic tree of nine mammals from the Zoo project [10], a phylogenetic tree of 14 prokaryotes (7 *Enterobacteriales*, 4 *Vibrionales*, and 3 *Pasteurellales*) [11], and a phylogenetic tree of 45 vertebrates from the ENCODE project [12]. See Figures 1, 2, and 3.

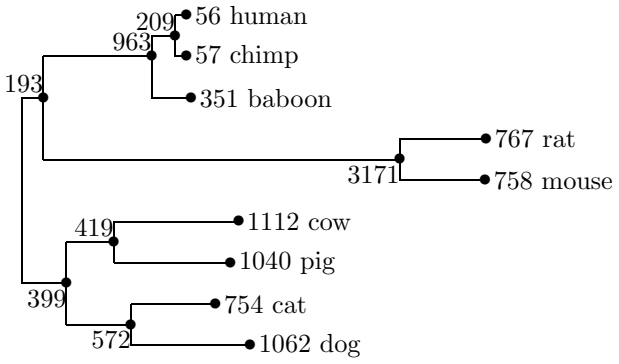


Figure 1: Phylogenetic tree of nine species from the Zoo project [10]. The numbers shown are the expected numbers of mutations that would occur in 10,000 neutral DNA sequence positions.

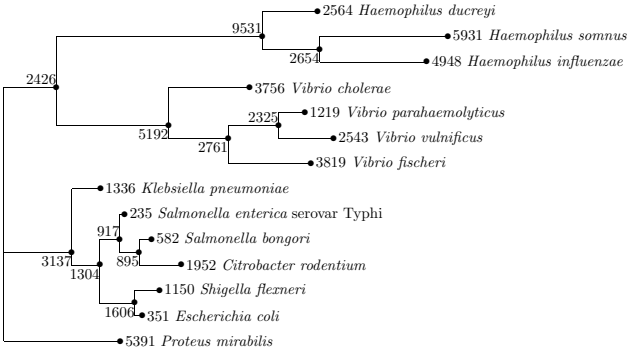


Figure 2: Phylogenetic tree of 14 prokaryotic species, which is reported to be realistic, although not definitive [11]. The numbers shown are the expected numbers of mutations that would occur in 10,000 neutral DNA sequence positions.

# Methods

## The Halpern & Bruno (1998) Model

The "HB98" approach [9] computes a nucleotide substitution model from (1) a background nucleotide substitution model that is appropriate for neutral sequence positions and (2) a foreground nucleotide equilibrium probability distribution that describes the equilibrium that results from selection/fitness pressures.

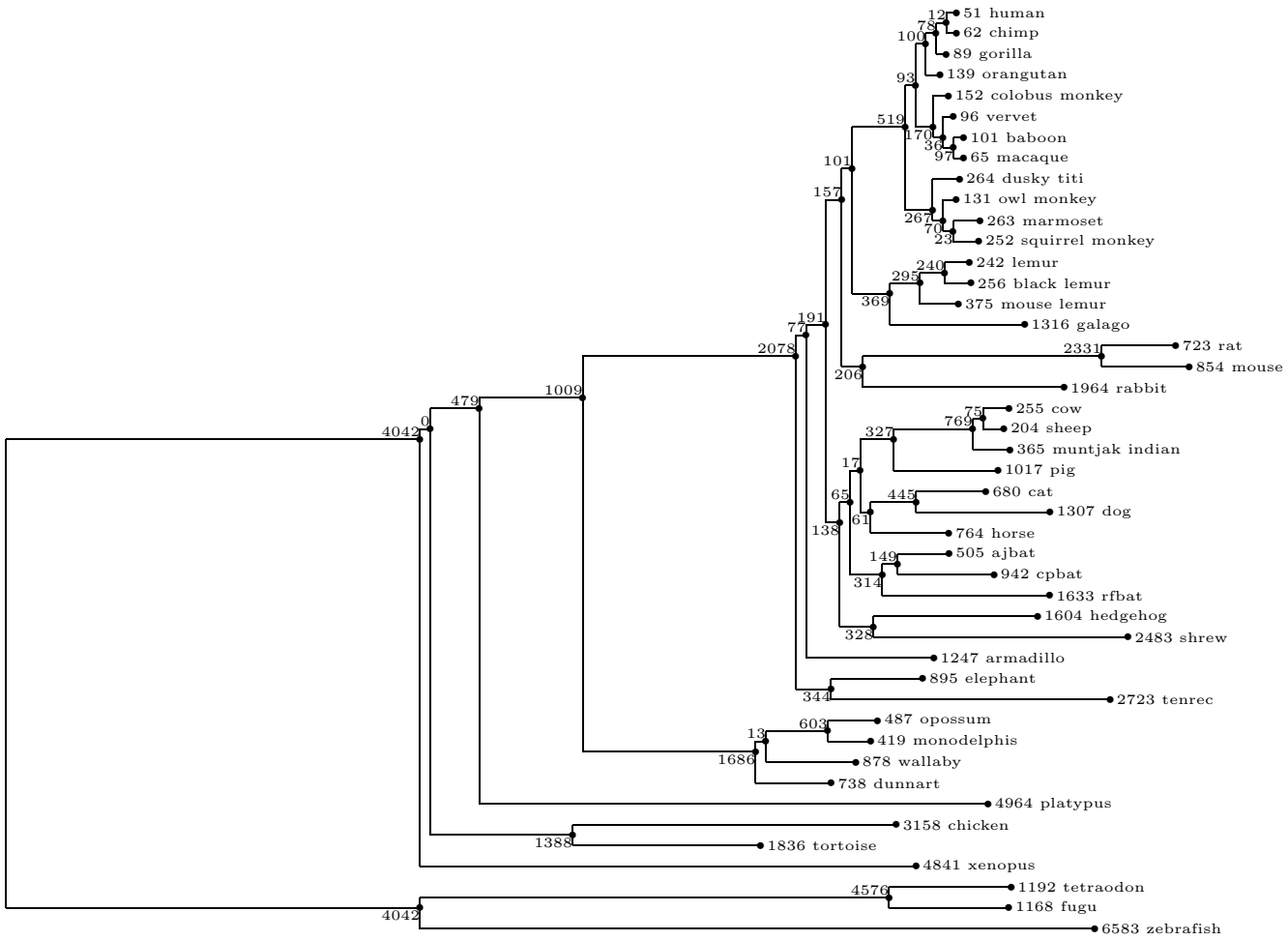


Figure 3: Phylogenetic tree of 45 species from the ENCODE project [12]. This is the August 2006 version generated by Elliott Margulies. It is built from four-way-degenerate third codon positions in the ENm001 (*i.e.*, the greater CFTR) ENCODE region. The numbers shown are the expected numbers of mutations that would occur in 10,000 neutral DNA sequence positions.

The instantaneous rate matrix for the HB98 nucleotide substitution model is computed from their Equation 13. With rearrangement of terms, and in our notation, the formula becomes

$$R'_{rc} = \begin{cases} R_{rc} \left( \frac{\log\left(\frac{\pi_r R_{rc}}{\pi_c R_{cr}}\right)}{\frac{\pi_r R_{rc}}{\pi_c R_{cr}} - 1} \right) & \text{if } \pi_r R_{rc} \neq \pi_c R_{cr} \\ R_{rc} & \text{if } \pi_r R_{rc} = \pi_c R_{cr} \end{cases} \quad (3)$$

where $r$ and $c$ are the row and column indices of matrices, respectively; $r$ and $c$ range over $\{A, C, G, T\}$; the background model is described by $R_{rc}$, the instantaneous rate of substitution from nucleotide $r$ to nucleotide $c$ when selective pressures do not apply; $\vec{\pi}$ is the foreground nucleotide equilibrium probability distribution resulting from selective pressures; and the foreground model is described by $R'_{rc}$, the instantaneous rate of substitution from nucleotide $r$ to nucleotide $c$ that arises when selective pressures do apply.

For the background nucleotide substitution model, we use the model of Jukes & Cantor (1969) [2], for which

$$R = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix} . \quad (4)$$

(Although this background nucleotide substitution model is simplistic, the foreground model is the important one.) The value of $\vec{\pi}$, which determines the foreground nucleotide substitution model, is varied by example in the Results section to demonstrate its effect on effective species count and motif efficiency.

## Probability of the Nucleotide Data

The instantaneous rate matrix $R'$ is converted to the nucleotide substitution model matrix $M$ for a branch of length $x$ via matrix exponentiation [13],

$$M = \exp(xR) . \quad (5)$$

The branches' nucleotide substitution model matrices are used in the tree-likelihood algorithm of Felsenstein (1981) [3] to compute the probability of the nucleotide data, one nucleotide per species, at a single multiply aligned DNA sequence position.

## Root Mean Square Estimator Variance

The effective species count is defined in terms of an error bar (or confidence limit) size. When a nucleotide equilibrium probability distribution can be estimated with a tight error bar, this is reflected as a high effective species count and high motif efficiency, in a manner that will become apparent in the following. We define the distance between the actual nucleotide equilibrium probability distribution and an estimate of it by Equation 1, and we deem the size of the error bar to be the root mean square of the distance values that would be collected should the experiment described in the Introduction be run multiple times.

## Fisher Information and the Cramér-Rao Inequality

We can calculate the root mean square distance without repeated simulation, via a Fisher Information matrix as described below. According to the Cramér-Rao Inequality (see, *e.g.*, [14]), this approach in general guarantees only a lower bound on the root mean square value (and thus an upper bound on the effective species count and motif efficiency). In practice, however, we can make a stronger statement. Consider the situation in which we have several multiply aligned DNA sequence positions that are subject to the same selective pressures, *e.g.*, this might be the case for several DNA positions each believed to be the first nucleotide of a common *cis*-regulatory element (although see [15]). Although the nucleotides at any one of these multiply aligned DNA sequence positions may be closely correlated due to close phylogenetic relationships of the aligned genomes, it is often reasonable to posit that the paralogous relationship between any two of these multiply aligned DNA sequence positions is significantly less close. When data from these nearly statistically independent multiply aligned DNA sequence positions are combined, to produce a nucleotide equilibrium probability distribution estimate, the bound provided by the Cramér-Rao Inequality will be quite tight. (Mathematically, when a large number of statistically independent data sets are combined, the Cramér-Rao Inequality approaches equality.)

The effective species count is computed via an expected log-likelihood, defined as

$$\mathrm{LL}(\vec{\theta}|\vec{\pi}) = \sum_{D \in 4^{(\text{species})}} \log(\Pr[D|\vec{\theta}]) \Pr[D|\vec{\pi}] , \quad (6)$$

where $D$ denotes the nucleotide data, one nucleotide per species, at a single multiply aligned DNA sequence position; the summation is over all $4^{(\text{number of species})}$ possibilities for $D$; $\vec{\pi}$ is the foreground nucleotide equilibrium probability distribution parameter of the HB98 nucleotide substitution model; $\Pr[D|\vec{\pi}]$ is the probability of generating nucleotides $D$, given the phylogenetic tree and the foreground nucleotide equilibrium probability distribution $\vec{\pi}$, computed via Felsenstein's tree-likelihood algorithm; and $\Pr[D|\vec{\theta}]$ is the same probability if the nucleotide data are instead explained by the HB98 nucleotide substitution model parameterized with the foreground nucleotide equilibrium probability distribution $\vec{\theta}$.

When the number of species exceeds 8, we instead randomly generate genomes of length $4^8$ according to the phylogenetic tree and foreground nucleotide substitution model

parameterized by $\vec{\pi}$, and we calculate the approximation

$$\mathrm{LL}(\vec{\theta}|\vec{\pi}) = 4^{-8} \sum_{D \in 4^8} \log(\Pr[D|\vec{\theta}]) \,, \tag{7}$$

where the sum is over all $4^8$ randomly generated multiply aligned DNA sequence positions.

If it were not for the fact that the components of $\vec{\theta}$ are constrained to sum to 1.0, our approach to applying the Cramér-Rao Inequality would be to compute the matrix of pure and mixed second derivatives of LL with respect to the nucleotide equilibrium probability distribution components $\theta_A$, $\theta_C$, $\theta_G$, and $\theta_T$; to then negate this matrix and evaluate it at $\vec{\theta} = \vec{\pi}$ so as to get the Fisher Information matrix; to take the matrix inverse; and finally to sum the values along the resulting diagonal

$$\text{Variance[distance]} = \text{Trace}\,(V) \tag{8}$$

$$V = \left( - \left. \frac{\partial^2 \mathrm{LL}(\vec{\theta}|\vec{\pi})}{\partial \theta_r \partial \theta_c} \right|_{\vec{\theta}=\vec{\pi}} \right)^{-1} \tag{9}$$

where $r$ and $c$ range over $\{A, C, G, T\}$ and are the row and column indices of the matrix, respectively.

However, because there are only three degrees of freedom in the four parameters $\theta_A$, $\theta_C$, $\theta_G$, and $\theta_T$, we re-parameterize the nucleotide equilibrium probability distribution. Because we are employing the matrix trace in Equation 8, any linear re-parameterization is equivalent. With

$$\theta_A = \psi_1 \tag{10}$$

$$\theta_C = \psi_2 \tag{11}$$

$$\theta_G = \psi_3 \tag{12}$$

$$\theta_T = 1 - \psi_1 - \psi_2 - \psi_3 \,, \tag{13}$$

Equation 9 becomes

$$V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix} \left( - \frac{\partial^2 \mathrm{LL}(\vec{\theta}(\vec{\psi})|\vec{\pi})}{\partial \psi_i \, \partial \psi_j} \right)^{-1} \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \tag{14}$$

This is computed numerically with $\Delta\psi_i = 10^{-5}$. Rather than using $T$ specially — Equation 13 is of a form different from that of Equations 10, 11, and 12 — in our implementation we instead treat specially a nucleotide that achieves the largest of the four values $\pi_A$, $\pi_C$, $\pi_G$, and $\pi_T$.

We compute

$$\text{effective species count} = \frac{\text{Trace}(V_1)}{\text{Trace}(V)} \tag{15}$$

where $V_1$ is the $V$ that arises on a degenerate phylogenetic tree with no branches and only a single species. (Note that, as for Trace$(V)$ generally, the value of Trace$(V_1)$ will depend on the foreground nucleotide equilibrium probability distribution $\vec{\pi}$.) This ratio is designed so as to yield a value of 1.0 in the limit when all phylogenetic tree branches have length zero, and it will yield a value that is the number of species in the phylogenetic tree in the limit that all the phylogenetic tree branches have length approaching infinity.

We compute the motif efficiency using Equation 2.

### Application to Synthetic and Real Phylogenetic Trees

We explored the effective species counts and motif efficiencies for star-topology phylogenetic trees with 2, 3, 4, 5, 8, 10, 16, and 25 species, with pairwise species separation ranging from $0.00$ to $5.00$, in increments of $0.05$. We also computed the effective species counts and motif efficiencies for the Zoo, prokaryote, and ENCODE phylogenetic trees depicted in Figures 1, 2, and 3.

For the star-topology phylogenetic trees, the notation S-EEEC used in Figures 4 and 5 and in Table 2 denotes an analysis of a phylogenetic tree with $S$ species and those multiply aligned DNA sequence positions that, due to selection pressures, have a nucleotide equilibrium probability distribution corresponding to EEEC (as described in Table 1).

Figure 4 shows the motif efficiency for star-topology phylogenetic trees with 2, 8, and 25 species. Figure 5 shows the motif efficiency for star-topology phylogenetic trees for the nucleotide equilibrium probability distributions 001S, 250X, and 001P.

The results for the Zoo, prokaryote, and ENCODE phylogenetic trees are in Table 2.

## Discussion

We find that, when the goal is the reconstruction of a nucleotide equilibrium probability distribution for a multiply aligned DNA sequence position (or a collection of such positions subject to the same selective pressures, *e.g.*, this may arise when the positions are all the same position of a common *cis*-regulatory element), it is important to get phylogenetically well-separated genomes. We find that nucleotide equilibrium probability distributions strongly dominated by a single nucleotide are more motif-efficient than is a uniform nucleotide equilibrium probability distribution; the latter is, in turn, more motif-efficient than are nucleotide equilibrium probability distributions strongly dominated by a pair of nucleotides (Figure 4). We find that star-topology phylogenetic trees with fewer species are more motif-efficient than those with more species (Figure 5).

Looking at, *e.g.*, the middle column of Table 2, we find the effective species counts of 1.73, 5.26, and 5.8, for the Zoo, prokaryote, and ENCODE phylogenetic trees. While there are

| Code | $1^{st}$ most common nucleotide | $2^{nd}$ most common nucleotide | $3^{rd}$ most common nucleotide | $4^{th}$ most common nucleotide |
|---|---|---|---|---|
| 001S | 0.997 | 0.001 | 0.001 | 0.001 |
| 010S | 0.970 | 0.010 | 0.010 | 0.010 |
| 100S | 0.700 | 0.100 | 0.100 | 0.100 |
| 250X | 0.250 | 0.250 | 0.250 | 0.250 |
| 100P | 0.400 | 0.400 | 0.100 | 0.100 |
| 010P | 0.490 | 0.490 | 0.010 | 0.010 |
| 001P | 0.499 | 0.499 | 0.001 | 0.001 |

Table 1: The three-digit code prefix indicates the frequency of the least common nucleotide. The one-letter code suffix indicates the number of nucleotides that dominate the nucleotide equilibrium probability distribution: the letter "S" means that the nucleotide equilibrium probability distribution is dominated by a single nucleotide, the letter "P" means that the nucleotide equilibrium probability distribution is dominated by a pair of nucleotides, and the letter "X" means that no nucleotide dominates the nucleotide equilibrium probability distribution.

| | 001S | 010S | 100S | 250X | 100P | 010P | 001P |
|---|---|---|---|---|---|---|---|
| Zoo/9 | 3.05 (26%) | 2.30 (16%) | 1.78 (10%) | 1.73 (9%) | 1.68 (9%) | 1.44 (6%) | 1.38 (5%) |
| prokaryote/14 | 9.31 (64%) | 7.32 (49%) | 5.27 (33%) | 5.26 (33%) | 4.74 (29%) | 3.30 (18%) | 3.04 (16%) |
| ENCODE/45 | 11.8 (25%) | 8.7 (17%) | 5.9 (11%) | 5.8 (11%) | 5.3 (10%) | 3.8 (6%) | 3.5 (6%) |

Table 2: The effective species count, with the motif efficiency in parentheses, is given for several different possible types of selective pressure. See Table 1 for an explanation of the column headings. The Zoo phylogenetic tree has nine species, the prokaryote phylogenetic tree has 14 species, and the ENCODE phylogenetic tree has 45 species. In this table, observe that, although the ENCODE project includes 44 genomes in addition to human, correlation of those genomes to human make them worth only 6%–25% of an "independent genome" each, in terms of the characterization of nucleotide equilibrium probability distributions at human DNA sequence positions. The phylogenetic tree is closest to a phylogenetic tree of 45 well-separated species for nucleotide equilibrium probability distributions that are very strongly dominated by a single nucleotide (*e.g.*, 001S and 010S), although the efficiency in this case is still far below 100%.

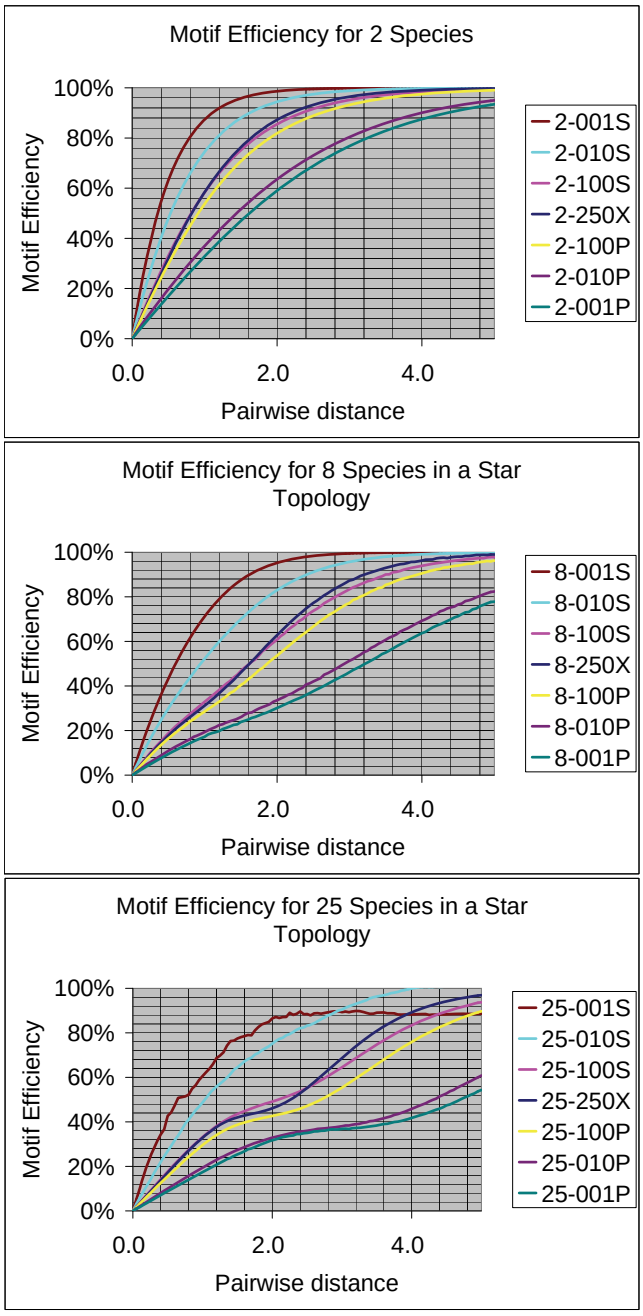


Figure 4: Motif efficiency for phylogenetic trees of 2, 8, and 25 species as a function of their pairwise phylogenetic distance, and the selective pressures. See Table 1 for an explanation of the labels 001S, 010S, 100S, 250X, 100P, 010P and 001P for selective pressures. In all three frames, we see that the nucleotide equilibrium probability distributions strongly dominated by a single nucleotide are more motif-efficient than is a uniform nucleotide equilibrium probability distribution; the latter is, in turn, more motif-efficient than are the nucleotide equilibrium probability distributions strongly dominated by a pair of nucleotides. That is, it is most important to get well-separated genomes when the goal is the characterization of DNA sequence positions that have a nucleotide equilibrium probability distribution dominated by a pair of nucleotides.

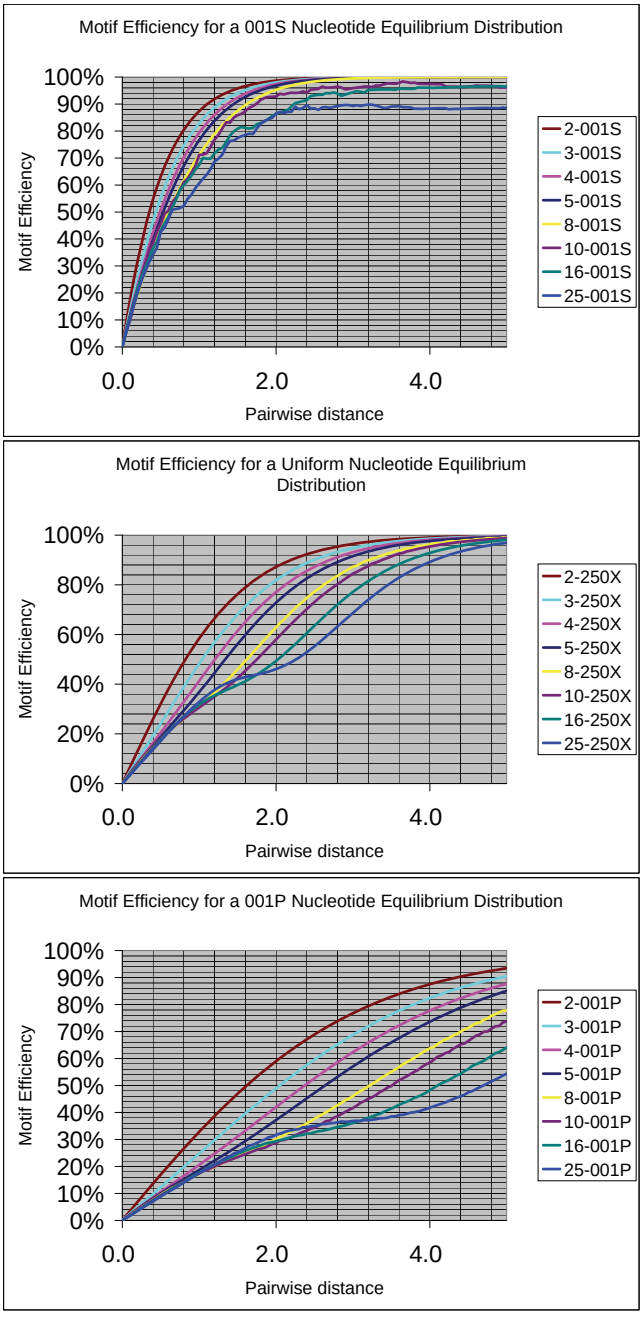


Figure 5: Motif efficiency for different nucleotide equilibrium probability distributions as a function of their pairwise phylogenetic distance, and the number of species. See Table 1 for an explanation of the labels 001S, 250X, and 001P for selective pressures. In all three frames, we see that the star-topology phylogenetic trees with fewer species are more motif-efficient than those with more species. Thus, inadequate species separation is more of a matter for concern when the number of genomes is large.

many other good reasons for large sequencing efforts, the sequencing of fewer genomes may be sufficient for the purpose of computing nucleotide equilibrium probability distributions. For the Zoo phylogenetic tree, the selection of just three distantly related genomes from the nine available genomes would have produced nearly the same effective species counts as the phylogenetic tree for all nine mammals; thus, characterization might have been achieved with fewer genomes, as long as the ability to make a multiple sequence alignments did not suffer appreciably. Likewise, for the prokaryote and ENCODE phylogenetic trees, the choice of 6–10 distantly related genomes might have been as effective as the full phylogenetic trees, so long as the sequence alignments were still obtainable. For the characterization of multiply aligned DNA sequence positions with the 001P nucleotide equilibrium probability distribution, even fewer genomes are required.

Whether effective species count or motif efficiency is the more appropriate measure depends upon the context. The web server available at `http://bayesweb.wadsworth.org/cgi-bin/Effective.pl` has the option of greedily adding one species at a time so as to maximize the effective species count, thus providing the researcher with a tool (1) for discovering a high effective species count achievable with a fixed number of genomes, or (2) for discovering a small set of genomes that provides a desired effective species count. (Note that, unlike with the work of Pardi & Goldman (2005) [5], we have no proof that this greedy approach is always optimal, although it has worked well in practice.)

If, instead, sequencing resources are not strictly limited and can be garnered for good causes, motif efficiency may be the more appropriate measure; a high motif efficiency demonstrates that the genomic sequencing is worth the effort.

When there are multiple DNA sequence positions within a genome posited to be subject to the same selective pressures, the effective species counts are additive if the paralogous relationships are believed to be distant. In particular, when the effective species count for each of $n$ positions is $X$, then having data for all these sites is as good as having just one such site with an effective species count of $nX$.

### Caveats

The above analyses presuppose that all of the species in the phylogenetic tree have DNA sequence positions that are orthologous with the relevant DNA sequence positions in a species of interest, and that these orthologous positions can be located and aligned. Margulies *et al.* (2005) finds evidence that even in low-redundancy sequencing efforts, locating and aligning orthologous sequences in mammals is not too difficult [8]. However, this is not always the case, and there is the obvious tradeoff: the set of genomes sought for the comparison should be diverse, but only if those genomes are likely to contain alignable, orthologous DNA sequence.

There is significant evidence that multiply occurring, functional DNA elements, such as transcription factor binding sites, are not necessarily subject to the same selection pressures at each genomic location [15]. Thus, except when there is evidence that the selection pressures are similar, the pooling of data from corresponding element positions should be done with care and the calculations here should be deemed approximate.

Also note that, for a given foreground nucleotide equilibrium probability distribution, the effective species count and the motif efficiency measure the error bar shrinkage relative to the same foreground nucleotide equilibrium probability distribution, on a degenerate phylogenetic tree that has all branch lengths equal to zero. In this article, we have not quantified the relative efficiency of characterizing one nucleotide equilibrium probability distribution versus another. That is, we have not discussed either Trace$(V)$ or Trace$(V_1)$ (from Equation 15) depends upon the nucleotide equilibrium probability distribution. However, as a general rule, the less uniform the nucleotide equilibrium probability distribution, the lower will be any phylogenetic tree's value of Trace$(V)$ and, hence, the lower will be its root mean square estimator variance and the smaller will be its error bar.

## Conclusions

When the goal is to precisely characterize the nucleotide equilibrium probability distribution for a multiply aligned DNA sequence position, it is important to have genomic sequences from a phylogenetically diverse set of species. We have provided a mathematical formalism and an algorithmic approach for quantification of the obtainable accuracy, in terms of an effective species count and a motif efficiency. We have applied the algorithm to several test cases, and have determined that large phylogenetic separation is most important when there are many genomes, and when nucleotide equilibrium probability distributions of interest are dominated by a pair of nucleotides. The algorithm is available on the web at `http://bayesweb.wadsworth.org/cgi-bin/Effective.pl`.

## Acknowledgments

Thank you to Charles E. "Chip" Lawrence for thought-provoking discussions on this topic, to C. Steven Carmack for the web site implementation, and to the Computational Molecular Biology and Statistics Core Facility at the Wadsworth Center. This research was supported by NIH/NHGRI grant 5K25HG003291.

# References


[1] Newberg LA, Lawrence CE: **Mammalian Genomes Ease Location of Human DNA Functional Segments but Not Their Description**. *Stat Appl Genet Mol Biol* 2004, **3**:article 23. [PubMed 16646802].

[2] Jukes TH, Cantor C: **Evolution of Protein Molecules**. In *Mammalian Protein Metabolism*, *Volume 3*. Edited by Munro HM, New York, NY: Academic Press. 1969:21–132.

[3] Felsenstein J: **Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach**. *J Mol Evol* 1981, **17**(6):368–376. [PubMed 7288891].

[4] Cooper GM, Brudno M, Green ED, Batzoglou S, Sidow A: **Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes**. *Genome Res* 2003, **13**(5):813–820. [PubMed 12727901].

[5] Pardi F, Goldman N: **Species Choice for Comparative Genomics: Being Greedy Works**. *PLoS Genet* 2005, **1**(6):e71. [PubMed 16327885].

[6] McAuliffe JD, Jordan MI, Pachter L: **Subtree power analysis and species selection for comparative genomics**. *Proc Natl Acad Sci U S A* 2005, **102**(22):7900–7905. [PubMed 15911755].

[7] Eddy SR: **A Model of the Statistical Power of Comparative Genome Sequence Analysis**. *PLoS Biol* 2005, **3**:95–102. [PubMed 15660152].

[8] Margulies EH, Vinson JP, Miller W, Jaffe DB, Lindblad-Toh K, Chang JL, Green ED, Lander ES, Mullikin JC, Clamp M: **An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing**. *Proc Natl Acad Sci U S A* 2005, **102**(13):4795–4800. [PubMed 15778292].

[9] Halpern AL, Bruno WJ: **Evolutionary Distances for Protein-Coding Sequences: Modeling Site-Specific Residue Frequencies**. *Mol Biol Evol* 1998, **15**(7):910–917. [PubMed 9656490].

[10] Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, Maskeri B, Hansen NF, Schwartz MS, Weber RJ, Kent WJ, Karolchik D, Bruen TC, Bevan R, Cutler DJ, Schwartz S, Elnitski L, Idol JR, Prasad AB, Lee-Lin SQ, Maduro VV, Summers TJ, Portnoy ME, Dietrich NL, Akhter N, Ayele K, Benjamin B, Cariaga K, Brinkley CP, Brooks SY, Granite S, Guan X, Gupta J, Haghighi P, Ho SL, Huang MC, Karlins E, Laric PL, Legaspi R, Lim MJ, Maduro QL, Masiello CA, Mastrian SD, McCloskey JC, Pearson R, Stantripop S, Tiongson EE, Tran JT, Tsurgeon C, Vogt JL, Walker MA, Wetherby KD, Wiggins LS, Young AC, Zhang LH, Osoegawa K, Zhu B, Zhao B, Shu CL, De Jong PJ, Lawrence CE, Smit AF, Chakravarti A, Haussler D, Green P, Miller W, Green ED: **Comparative Analyses of Multi-Species Sequences from Targeted Genomic Regions**. *Nature* 2003, **424**(6950):788–793. [PubMed 12917688].

[11] Carmack CS, McCue LA, Newberg LA, Lawrence CE: **PhyloScan: Identification of transcription factor binding sites using cross-species evidence**. *Algorithms Mol Biol* 2007, **2**:article 1. [PubMed 17244358].

[12] ENCODE Project Consortium: **The ENCODE (ENCyclopedia Of DNA Elements) Project**. *Science* 2004, **306**(5696):636–640. [PubMed 15499007].

[13] Lanave C, Preparata G, Saccone C, Serio G: **A New Method for Calculating Evolutionary Substitution Rates**. *J Mol Evol* 1984, **20**:86–93. [PubMed 6429346].

[14] Stuart A, Ord JK, Arnold S: *Classical Inference and the Linear Model*, London: Arnold., *Volume 2A of* Kendall's Advanced Theory of Statistics. Sixth edition 1999 chap. 17 Estimation and Sufficiency, :17.13–17.27.

[15] Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB: **Position Specific Variation in the Rate of Evolution in Transcription Factor Binding Sites**. *BMC Evol Biol* 2003, **3**:19. [PubMed 12946282].