The Relative Inefficiency of Sequence Weights in Determining a Nucleotide Consensus Distribution

Lee A. Newberg[*][†][‡] Lee Ann McCue[*][§] Charles E. Lawrence[*][†][¶]

Technical Report 03-10
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York 12180-3590, USA

August 7, 2003

Abstract

The use of sequence weights to estimate a consensus distribution of nucleotides at any position of an alignment of nucleic acid sequences is found to perform poorly in comparison to a maximum-likelihood method based upon phylogenetic relationships. We derive optimal sequence weights for sequences related by a phylogenetic tree but find that, among a collection of primate sequences, the sequences-weights approach is 51% as efficient as the maximum-likelihood approach in making use of the alignment data.

Preferable to the use of sequence weights to estimate the distribution of bases at each aligned sequence position is the use of an alternative recipe. First, seek a phylogenetic tree that connects the species in question, perhaps employing the aligned sequence data set itself for that purpose. Second, use the tree topology and edge lengths to calculate a maximum-likelihood distribution of bases at each position in the aligned sequence.

Introduction

Sequence weights are frequently used to combine data from aligned sequences into a single consensus sequence model, in which each position is described by a probability distribution over the range of possible nucleotides or amino acid residues. These consensus models have proven useful for

[*] The Wadsworth Center, New York State Department of Health, Albany, New York 12201-0509, USA

[†] Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York 12180-3590, USA.

[‡] Corresponding author. E-mail: lnewberg at wadsworth.org. Phone: (518) 473-8426. Fax: (518) 473-2900.

[§] Supported by DOE grant DEFG0201ER63204

[¶] Supported by NIH (NHGRI) grant 5R01HG00125707 and DOE grant DEFG0201ER63204

describing functional sections of sequence and in database searches for additional similar sequences. Much effort has been put into the design of weighting schemes that will ensure that the consensus model is close to true under a wide variety of conditions. A typical extenuating circumstance is one in which a set of sequences exhibiting a particular feature overwhelms the data from another set of sequences having a different feature, merely because the latter set has fewer representatives available for analysis.

There are many sequence weighting approaches, and several ways in which sequence weights have been employed. Vingron and Argos [1989] calculated the weight of an amino acid sequence to be the sum of the Hamming distances from that sequence to the other sequences, given a proposed alignment; they used this information for multiple sequence alignment. Altschul, Carroll, and Lipman [1989] computed weights for pairs of sequences based upon a variance-minimizing condition among such pairs, and used this information in multiple sequence alignment in which the quality of a multiple alignment is the weighted sum of the quality of each implied pairwise alignment. They also used weights for the estimation of a continuous characteristic at the root of a tree. Sibbald and Argos [1990] computed the weight for a sequence as its Voronoi volume, using a Hamming distance metric within the space of sequences generated by sampling each position's value randomly from the values for that position among the input sequences (akin to bootstrapping). Vingron and Sibbald [1993] created a methodology for evaluating sequence-weighting schemes, compared four, and concluded that the approach of Altschul and colleagues is best when sequences are phylogenetically related. Otherwise, the approach of Sibbald and Argos [1990] was considered best. Henikoff and Henikoff [1994] computed the weight of a sequence as the sum of weights assigned to each of its positions; with the weight of a position equal to the reciprocal of the number of sequences that have the same amino acid residue or nucleotide at that position. Krogh and Mitchison [1995] chose weights that maximize a sum over the aligned positions, with each term being the entropy of the distribution at that position implied by the weights.

Our interest in sequence weights arises from the task of locating transcription factor binding sites. These binding sites are short sections of DNA (6–30 base pairs long), often upstream of the first exon of a gene, which play a critical role in transcription and the overall regulation of gene expression. Variations in the sequence recognized by a particular transcription factor (protein) are common, and it is natural to describe these binding-site sequences by choosing the strand of the DNA encoding the gene and, for that strand, giving a motif, a consensus distribution of bases for each sequence position relevant to the transcription factor binding. (See the work of Lawrence and Reilly [1990] for a good description of the statistical model.)

A single transcription factor may play a role in multiple genes across multiple species; however since we usually have only a non-random subset of the genes in a non-random subset of the species, we need an approach that can find an accurate consensus distribution in the presence of these biases.

Because a consensus distribution is our goal, we define the efficiency of an approach by how well it can estimate a consensus distribution. In this paper, we derive optimal sequence weights that minimize estimator variance for sequences related by a phylogenetic tree. Further, we calculate the efficiency of estimates using these optimal weights, relative to the efficiency of a maximum-likelihood method based upon phylogenetic relationships. For the two test cases to which these methods are applied, we find that the latter approach is superior.

Previous work in the field of incorporation of multiple species data into the location of transcription factor binding sites can be found in the literature. McCue et al. [2002] modeled the sequences as if they were independent in calculating the maximum *a posteriori* probability (MAP) of a proposed motif. However, based upon simulations incorporating phylogenetic relationships they raised the hurdle that such a MAP must clear to be deemed significant. Rajewsky et al. [2002] looked at pairs of sequences, locating functional DNA by detecting where the observed mutation rate is lower than that expected from the overall mutation rate between the species. Boffelli et al. [2003] looked at all of the sequences together, seeking for functional DNA by observing where the phylogenetic model of Yang and Roberts [1995] indicated mutation at a slower rate than elsewhere.

Materials and Methods

We calculate the relative efficiency of base-frequency estimates for each approach in terms of the method's effective number of additional independent sequences. "Effective number of observations" is a common statistical measure that allows us to benchmark data sets not composed of independent observations, against a common ladder calibrated by data sets in which the observations are independent.

In this case, the effective number of independent sequences is calculated via a total estimator variance, which is a sum over the four possible bases that could occur at a given binding site position on one strand of the DNA. The term of the sum for a given base is the expected variance of the estimator for the probability of that base. For this measure, a small sum of variances is indicative of an efficient set of estimators, and a larger sum of variances is indicative of a poorer set of estimators. When we have S independent sequences, the total estimator variance will be $1/S$ of the total estimator variance of a single sequence; thus, we calibrate our ladder by defining the effective number of independent sequences for the data to be the total estimator variance if we use a single sequence, divided by the total estimator variance for the complete set of dependent sequences.

Phylogenetic Data Model

We use the standard phylogenetic model for the likelihood of aligned phylogenetic sequence data as computed from base mutations, first described by Neyman [1971] and Felsenstein [1981]. As is common, we track base mutations/transitions from a parent sequence to a child sequence as a matrix M, in which the rows of the matrix correspond to the different possibilities for the base in the parent sequence, and the columns of the matrix correspond to the same base possibilities in the child sequence. For instance, with the bases ordered as (A, T, C, G), the matrix is written:

$$M = \begin{pmatrix} \Pr[A|A] & \Pr[T|A] & \Pr[C|A] & \Pr[G|A] \\ \Pr[A|T] & \Pr[T|T] & \Pr[C|T] & \Pr[G|T] \\ \Pr[A|C] & \Pr[T|C] & \Pr[C|C] & \Pr[G|C] \\ \Pr[A|G] & \Pr[T|G] & \Pr[C|G] & \Pr[G|G] \end{pmatrix} \tag{1}$$

where, for example, $\Pr[A|C]$ is the probability that the child sequence has an A where the parent sequence has a C.

We use phylogenetic tree topologies and edge lengths such as those depicted in Figure 1. A tree describes the expected number of mutations per sequence position between any two sequences in the tree (henceforth always including in the count successive mutations at a single position) as the sum of the edge lengths along the path that connects those two sequences.

We choose the base transition model of Felsenstein [1981] because of its direct connection to an underlying equilibrium distribution, even though it does not directly model features such as the difference between transitions ($A \leftrightarrow G$ and $C \leftrightarrow T$) and transversions (other changes in the bases) as was first done with the model of Kimura [1980]. (See the Discussion section.) The base transition matrix between two sequences separated by a path of length x is:

$$\begin{aligned} M_x &= e^{-kx} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + (1 - e^{-kx}) \begin{pmatrix} \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \end{pmatrix} \\ k &> 0 \,. \end{aligned} \tag{2}$$

This model has the necessary features that as $x \to 0$, the transition matrix is the identity matrix; that as $x \to +\infty$, the transition matrix gives an equilibrium distribution independent of which base we started with (*i.e.*, all of the rows are equal); and that $M_{a+b} = M_a M_b$, correctly modeling that the transition resulting from evolution described by an evolutionary distance a, followed by evolution described by an evolutionary distance b, is equal to the evolution described by the sum of the evolutionary distances.

The value of k serves to calibrate the units of x, and a proper choice for k requires some discussion. We choose k according to the convention of Lanave et al. [1984] and Rodríguez et al. [1990] so that the expected number of base mismatches per position between closely related parent and child sequences implied by the transition matrix M_x is x when the parent sequence begins in the equilibrium distribution. Expansion of Equation 2 in a power series about $x = 0$ gives:

$$M_x \approx \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + kx \begin{pmatrix} \theta_A - 1 & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T - 1 & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C - 1 & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G - 1 \end{pmatrix} , \tag{3}$$

and the implied joint distribution of bases when the parent sequence is in the equilibrium distribution is:

$$J_x \approx \begin{pmatrix} \theta_A & 0 & 0 & 0 \\ 0 & \theta_T & 0 & 0 \\ 0 & 0 & \theta_C & 0 \\ 0 & 0 & 0 & \theta_G \end{pmatrix} + kx \begin{pmatrix} \theta_A^2 - \theta_A & \theta_A\theta_T & \theta_A\theta_C & \theta_A\theta_G \\ \theta_T\theta_A & \theta_T^2 - \theta_T & \theta_T\theta_C & \theta_T\theta_G \\ \theta_C\theta_A & \theta_C\theta_T & \theta_C^2 - \theta_C & \theta_C\theta_G \\ \theta_G\theta_A & \theta_G\theta_T & \theta_G\theta_C & \theta_G^2 - \theta_G \end{pmatrix} . \tag{4}$$

Because we wish the sum of the off-diagonal elements to be x (or, equivalently, the sum of the diagonal elements to be $1 - x$), we choose

$$k = \frac{1}{1 - (\theta_A^2 + \theta_T^2 + \theta_C^2 + \theta_G^2)} . \tag{5}$$

Without this choice of normalization, one proposed distribution might be penalized relative to another — not because the equilibrium is less reflective of the the underlying biological process, but because, via poor normalization, the two distributions imply a different total number of mismatches between some pair of closely related sequences.

By making the matrices of Equation 2 have dimensions 20 × 20 and by using the obvious generalization of Equation 5, we can use this mutation process for amino acid residues at each position in a polypeptide chain.

Total Estimator Variance of the Sequence Weights Approach

Sequence-weighting estimates are obtained as follows:

$$\hat{\theta}_b = \sum_s w_s D_{sb} \tag{6}$$

where D_{sb} is 1 if sequence s has base b and is 0 otherwise, and $\sum_s w_s = 1$.

To calculate the total estimator variance, we imagine the following experiment. We start with a phylogenetic model, such as the phylogenetic tree topologies and edge lengths calculated by Page, Chia, and Goodman [1999], and depicted in Figure 1, and an assumption for the equilibrium distribution $\vec{\theta}^*$ to define k^* via Equation 5, and M_x via Equation 2. We imagine generating instances of a sequence position's data according to the model, in this case a single nucleotide for each primate species, and from that sample we calculate $\hat{\theta}_b$ according to Equation 6. We measure average squared distance of these sampled $\hat{\theta}_b$ values to the model mean θ_b^* and repeat the experiment for each base b. The sum of these measured variances is the total estimator variance that we seek.

We can find the total estimator variance analytically, without the repeated sampling just described. From Equation 6, the total variance of these estimators is computed as

$$\sum_b \mathrm{Var}[\hat{\theta}_b] = \sum_b \mathrm{E}[(\hat{\theta}_b - \theta_b^*)^2] \tag{7}$$

$$
\begin{aligned}
&= \sum_s w_s \sum_{s'} w_{s'} (\sum_b \mathrm{E}[(D_{sb} - \mathrm{E}[D_{sb}])(D_{s'b} - \mathrm{E}[D_{s'b}])]) && (8)\\
&= \vec{w}^T C \vec{w} && (9)
\end{aligned}
$$

where $\vec{w}$ is the column vector of sequence weights, $\vec{w}^T$ is the corresponding row vector, C is the $S \times S$ covariance matrix with elements $C_{ss'} = \sum_b \mathrm{Cov}[D_{sb}, D_{s'b}]$, and S is the number of sequences.

The model of Equation 2 gives us the formula for the joint probability distribution for two sequences s and s' separated by a distance x:

$$
J_x = e^{-k^* x} \begin{pmatrix} \theta^*_A & 0 & 0 & 0 \\ 0 & \theta^*_T & 0 & 0 \\ 0 & 0 & \theta^*_C & 0 \\ 0 & 0 & 0 & \theta^*_G \end{pmatrix} + (1 - e^{-k^* x}) \begin{pmatrix} \theta^*_A\theta^*_A & \theta^*_A\theta^*_T & \theta^*_A\theta^*_C & \theta^*_A\theta^*_G \\ \theta^*_T\theta^*_A & \theta^*_T\theta^*_T & \theta^*_T\theta^*_C & \theta^*_T\theta^*_G \\ \theta^*_C\theta^*_A & \theta^*_C\theta^*_T & \theta^*_C\theta^*_C & \theta^*_C\theta^*_G \\ \theta^*_G\theta^*_A & \theta^*_G\theta^*_T & \theta^*_G\theta^*_C & \theta^*_G\theta^*_G \end{pmatrix} . \quad (10)
$$

It follows that

$$
C_{ss'} = \sum_b \mathrm{Cov}[D_{sb}, D_{s'b}] = \sum_b [(J_x)_{bb} - (\theta^*_b)^2] = \sum_b e^{-k^* x}(\theta^*_b - (\theta^*_b)^2) = \frac{e^{-k^* x}}{k^*} . \quad (11)
$$

Setting zero equal to the gradient of the right-hand side of Equation 9 with respect to the vector $\vec{w}$ (while using a LaGrange multiplier to ensure that $\sum_s w_s = 1$), we find optimal sequence weights:

$$
\begin{aligned}
\vec{w} &= \frac{C^{-1}\vec{1}}{\vec{1}^T C^{-1} \vec{1}} && (12)\\
\min_{\vec{w}} \ \vec{w}^T C \vec{w} &= \frac{1}{\vec{1}^T C^{-1} \vec{1}} && (13)
\end{aligned}
$$

where $\vec{1}$ is the column vector of all ones. Note that Equation 12 is identical in form to equations which appeared in the work of Altschul, Carroll, and Lipman [1989], Vingron and Sibbald [1993], and Arvestad and Bruno [1997]. In the first, C was instead a matrix of tree path lengths between between sequences. In the second, C was instead a matrix of "(dis)similarity" values between sequences. In the third, C was instead a matrix of covariances of distance estimates computed from the spectral components of the transition matrix, and the formula was used to compute a precise consensus distance.

Total Estimator Variance of the Maximum Likelihood Approach

As for sequence weights, we wish to evaluate the total estimator variance as if we had sampled the observed data from an underlying model, and we desire an approach that allows us to integrate out the data so that we get the exact solution, rather than an approximation from sampling.

The standard approach via the Fisher information matrix will work. The data samples are assumed to occur with frequency proportional to their probability $\Pr[D|\vec{\theta}^*]$; assuming an underlying model based upon some $\vec{\theta}^*$; and the likelihood of a model based upon an equilibrium $\vec{\theta}$ is calculated as

$$\log L(\vec{\theta}) \quad = \quad \sum_D \log(\Pr[D|\vec{\theta}]) \Pr[D|\vec{\theta}^*] \ . \tag{14}$$

(If the number of terms in the sum is too large, we can always revert to sampling D proportionately to $\Pr[D|\vec{\theta}^*]$.) Intuitively, our confidence in the maximum-likelihood estimate depends on the shape of $\log L(\vec{\theta})$ at its maximum $\vec{\theta}^*$; the more steeply $\log L(\vec{\theta})$ falls off from this maximum, the more confident we are of the estimate's accuracy.

Specifically, our method is as follows. To calculate the total estimator variance at $\vec{\theta}^*$, we determine (via numerical differentiation) the Hessian matrix of pure and mixed second derivatives of $\log L(\vec{\theta})$ with respect to a set of three degrees of freedom implicit in the four components of $\vec{\theta}$; we invert the matrix to find the variances and covariances of these degrees of freedom; we adjust the covariance matrix to be 4×4 to represent the four natural parameters $\theta_A, \theta_T, \theta_C, \theta_G$; and we then take the trace of this matrix, *i.e.*, the sum of the expected estimator variances.

Results

We find that the use of optimal sequence weights with the phylogenetic tree of primates from Figure 3 of the article by Page, Chia, and Goodman [1999], as depicted in Figure 1 herein, gives an effective number of independent sequences of 1.49. That is, the information from the non-human primates adds an effective 0.49 independent sequences to our ability to determine the distribution of bases. In contrast, the maximum-likelihood approach gives an effective number of independent sequences of 1.96, an increase of 0.96 over using human alone. For sequence weights, the increase in the effective number of independent sequences is 51% as large as that of the maximum-likelihood competitor.

We tested the approaches on a phylogenetic tree for *Escherichia coli* K12 and some related bacteria. To build the tree, we retrieved DNA sequence data for the 16S rRNA gene for those species from public sources, aligned the data using ClustalW (`http://www.ebi.ac.uk/clustalw/`), and constructed a phylogenetic tree with the PHYLIP maximum-likelihood method (`http://evolution.genetics.washington.edu/phylip.html`). Because these data exhibit conservation, we scaled up the resulting edge lengths by a factor of nearly 14 to match the finding of McCue et al. [2002], in which aligned non-coding sequence from *Escherichia coli* K12 and *Salmonella enterica* serovar Typhi CT18 were 30% dissimilar. Although the edge lengths in this tree, depicted in Figure 2, should not be considered definitive, we find the tree to be a useful example. For this tree, the sequence-weights approach gives the equivalent of 1.84 additional independent sequences. In contrast, the maximum-likelihood approach gives the equivalent of 2.40 additional independent sequences. The sequence-weights approach performed relatively better for this tree, although still suboptimally; for the sequence-weights approach the effective number of additional independent sequences is 77% as large as that of the maximum-likelihood competitor.

Additionally, we explored the proper theoretical order in which to sequence species, if the sequence data are not yet available. Specifically, if we have some sequences, with a complete phylogenetic tree, we can ask which additional single species' sequence would most increase the effective number of independent sequences for a consensus sequence. Figure 3 shows the effective number of additional independent sequences, if we start with human and at each step add the single species whose sequence would most increase the efficiency at that step. We find that, with the use of human and the first two non-human primates, the maximum-likelihood approach is more efficient than is the use of all of the species and the sequence-weights approach. The long plateau for sequence weights in Figure 3 indicates that all of the sequences beyond the first two or three additional sequences contribute little to the estimates.

Discussion

While we have picked a simple phylogenetic model that does not directly recognize the differences between transitions and transversions, we do not believe the results to be highly sensitive to this choice. For instance, in an extreme equilibrium example in which one base is modeled to occur 90% of the time, with the occurrence of the other three bases equally likely among the remaining 10%, the sequence-weights approach is 59% as efficient as the maximum-likelihood approach in making use of the non-human primate sequences of Figure 1.

The choice of total estimator variance as a benchmark for evaluating the two approaches is somewhat arbitrary, and we can envisage alternatives. Even if we assume that a function of the $\hat{\theta}$ covariance matrix must be optimized, there are alternatives. The product of the (pure) variances and the determinant of the covariance matrix (*i.e.*, the volume of the confidence ellipsoid) are two obvious possibilities. (For more see, *e.g.*, Chapter 2 in Silvey [1980].) We settled upon the sum of the individual variances for several reasons:

- In describing a transcription factor binding site, or when describing a sequence pattern for database search, we frequently see each position in the sequence of the site described by a probability distribution of bases. Thus, it is reasonable to evaluate an approach's efficiency on the basis of how well it can determine a probability distribution of bases. That is, it is straightforward to use some function of the covariance matrix of the estimators $\hat{\theta}$.

- Unlike the case for some of the alternatives based upon the covariance matrix, for total estimator variance a zero variance in one of the dimensions does not hide the uncertainties in the other dimensions.

- Because it is the trace of the covariance matrix, and because the trace is a characteristic that is unchanged when we perform an orthogonal change of basis, the metric does not depend on the choice of orthogonal basis used to describe the equilibrium $\vec{\theta}$.

We need a phylogenetic tree if we are to use the maximum-likelihood approach for deriving a consensus distribution, but we need not use the aligned sequence data to generate the phylogenetic tree. If the sequence data are used to construct the tree, beware of the possibilities of alignment bias and sequence bias. Specifically, the “optimal” alignment may be assessed as optimal in part because it has matched up bases that accidentally coincide; this alignment bias may cause base-mutation rates and phylogenetic distances to be underestimated. Further, if the aligned sequence is in part conserved, this too may cause the mutation rates and phylogenetic distances to be underestimated.

If the goal is to find $\hat{\theta}$ at an aligned position that is believed to be significantly conserved, the mutation model of Equation 2 — for sequence positions not subject to natural selection — may not be directly applicable. In such a case, it may be reasonable to multiply kx by a small factor γ to indicate the expected reduced rate of mutation, effectively shrinking the phylogenetic tree:

$$\begin{aligned} M_x &= e^{-\gamma k x}\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + (1 - e^{-\gamma k x})\begin{pmatrix} \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \end{pmatrix} \qquad (15) \\ k &= \frac{1}{1 - (\theta_A^2 + \theta_T^2 + \theta_C^2 + \theta_G^2)} \\ \gamma &\in (0,1)\ . \end{aligned}$$

This is somewhat similar to the approach of Bruno [1996], where, in the context of amino acid residues, a lower mutation rate was desired when the number of residues that occur with significant probability is smaller. Translation of this to nucleotides and our conventions fixes $\gamma k = 1/((B - 1)\max\{\theta_A, \theta_T, \theta_C, \theta_G\})$, where $B = 4$ is the number of bases.

Boffelli et al. [2003] considered variations in the mutation rate as an indicator of the location of functional/conserved sequence. (In our notation, this would be equivalent to allowing the variable γ that appears in Equation 15 to vary by sequence position, but leaving $\vec{\theta}$ fixed across sequence positions.) We believe that a combined approach, which maximizes the joint probability of the distribution $\vec{\theta}$ and the mutation rate γ, is likely to be even better at estimating the consensus distribution $\vec{\theta}$ than is the simpler maximum-likelihood approach described here.

If there is no reason to believe that aligned sequence data are phylogenetically related, or if construction of a tree is not possible, then the maximum-likelihood approach will not be feasible. In this case, a sequence-weights approach may still be feasible.

The sequence-weights approach may be slightly less efficient than we have indicated here. Equation 12 can give some negative sequence weights. If we add constraints to force all weights to be non-negative, the total estimator variance can only increase. However, our evidence is that this effect is small.

For both the maximum-likelihood approach and the sequence-weights approach, the value of $\hat{\theta}$ at a particular position depends primarily on the sequence data for that aligned position; however, for the sequence-weights approach, $\hat{\theta}$ also weakly depends upon the assumed underlying $\vec{\theta}^*$, via the occurrences of k^* in Equation 11. For additional accuracy when the result in any approach depends on $\vec{\theta}^*$, we might use the average across the neutral positions of the resulting $\hat{\theta}$ vectors as the value of $\vec{\theta}^*$ for a subsequent iteration, repeating until sufficient convergence is achieved.

Conclusion

We have shown that the use of sequence weights to compute a consensus distribution of nucleotides at any position of an alignment of nucleic acid sequences does not perform as well as does a maximum-likelihood method based upon phylogenetic relationships. In particular, for aligned sequences from primates (as represented by the phylogenetic tree of Page, Chia, and Goodman [1999]), the sequence-weights approach is 51% as efficient as is the maximum-likelihood approach in making use of the data from the non-human primates. Furthermore, aligned sequences from *Escherichia coli* K12 and related species of bacteria (as depicted in the phylogenetic tree of Figure 2) show a comparable 77% relative efficiency. We also find that, for sequences from human and two well-chosen non-human primates, the maximum-likelihood approach is more efficient than is use of the entire tree with the sequence-weights approach.

Rather than the use of sequence weights to estimate the distribution of bases at each position of the aligned sequences, we recommend an alternative recipe. First, seek a phylogenetic tree that connects the species in question, perhaps using the aligned sequence data and the algorithm of Felsenstein, or a more complex algorithm such as that of Yang and Roberts; these algorithms have publicly available implementations. Second, use the tree topology and edge lengths (as well as any position-specific variations implied by a model such as that of Yang and Roberts) to calculate the most likely $\vec{\theta}$ at each position of the aligned sequence.

Acknowledgments

This work was supported by the National Institutes of Health (National Human Genome Research Institute) grant 5R01HG00125707 to C. E. L. and by the Department of Energy grant DEFG0201ER63204 to C. E. L. and L. A. M.

Literature Cited

S. F. Altschul, R. J. Carroll, and D. J. Lipman. Weights for data related by a tree. *Journal of Molecular Biology*, 207(4):647–653, June 20 1989.

L. Arvestad and W. J. Bruno. Estimation of reversible substitution matrices from multiple pairs of sequences. *Journal of Molecular Evolution*, 45(6):696–703, December 1997.

D. Boffelli, J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, and E. M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–1394, February 28 2003.

W. J. Bruno. Modeling residue usage in aligned protein sequences via maximum likelihood. *Molecular Biology and Evolution*, 13(10):1368–1374, December 1996.

J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

S. Henikoff and J. G. Henikoff. Position-based sequence weights. *Journal of Molecular Biology*, 243 (4):574–578, November 4 1994.

M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, December 1980.

A. Krogh and G. J. Mitchison. Maximum entropy weighting of aligned sequences of proteins or DNA. In C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak, editors, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 215–221, Robinson College, Cambridge, United Kingdom, July 16–19 1995. American Association for Artificial Intelligence, AAAI Press.

C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20(1):86–93, 1984.

C. E. Lawrence and A. A. Reilly. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7(1):41–51, 1990.

L. A. McCue, W. Thompson, C. S. Carmack, and C. E. Lawrence. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Research*, 12 (10):1523–1532, October 2002.

J. Neyman. Molecular studies of evolution: A source of novel statistical problems. In S. S. Gupta and J. Yackel, editors, *Statistical Decision Theory and Related Topics*, pages 1–27. Academic Press, New York, 1971.

S. L. Page, C.-h. Chia, and M. Goodman. Molecular phylogeny of Old World monkeys (cercopithecidae) as inferred from γ-globin DNA sequences. *Molecular Phylogenetics and Evolution*, 13(2): 348–359, November 1999.

N. Rajewsky, N. D. Socci, M. Zapotocky, and E. D. Siggia. The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. *Genome Research*, 12(2):298–308, February 2002.

F. Rodríguez, J. L. Oliver, A. Marín, and J. R. Medina. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142(4):485–501, February 22 1990.

P. Sibbald and P. Argos. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *Journal of Molecular Biology*, 216(4):813–818, December 20 1990.

S. D. Silvey. *Optimal Design: An Introduction to the Theory for Parameter Estimation.* Chapman and Hall, 1980.

M. Vingron and P. Argos. A fast and sensitive multiple sequence alignment algorithm. *Computer Applications in the Biosciences*, 5(2):115–121, April 1989.

M. Vingron and P. Sibbald. Weighting in sequence space: A comparison of methods in terms of generalized sequences. *Proceedings of the National Academy of Sciences of the USA*, 90(19): 8777–8781, October 1 1993.

Z. Yang and D. Roberts. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Molecular Biology and Evolution*, 12(3):451–458, May 1995.

```
+-0.1206-+-0.0299-+-0.0251-+-0.0085-+-0.0027-+-0.0003-+-0.0057-+-0.0002-+-0.0079--Papio hamadryas cynocephalus
|        |        |        |        |        |        |        |        |
|        |        |        |        |        |        |        |        +-0.0037--Theropithecus gelada
|        |        |        |        |        |        |        |
|        |        |        |        |        |        |        +-0.0051--Lophocebus aterrimus
|        |        |        |        |        |        |
|        |        |        |        |        |        +-0.0055-+-0.0049-+-0.0026--Mandrillus sphinx
|        |        |        |        |        |                 |        |
|        |        |        |        |        |                 |        +-0.0095--Mandrillus leucophaeus
|        |        |        |        |        |                 |
|        |        |        |        |        |                 +-0.0092--Cercocebus galeritus
|        |        |        |        |        |
|        |        |        |        |        +-0.0068-+-0.0005-+-0.0070--Macaca nigra
|        |        |        |        |                 |        |
|        |        |        |        |                 |        +-0.0121--Macaca nemestrina
|        |        |        |        |                 |
|        |        |        |        |                 +-0.0038--Macaca mulatta
|        |        |        |        |
|        |        |        |        +-0.0030-+-0.0050-+-0.0080--Erythrocebus patas
|        |        |        |                 |        |
|        |        |        |                 |        +-0.0118--Chlorocebus aethiops
|        |        |        |                 |
|        |        |        |                 +-0.0137--Cercopithecus cephus
|        |        |        |
|        |        |        +-0.0156-+-0.0149-+-0.0084--Colobus species
|        |        |                 |        |
|        |        |                 |        +-0.0136--Colobus guereza
|        |        |                 |
|        |        |                 +-0.0120-+-0.0136--Nasalis larvatus
|        |        |                          |
|        |        |                          +-0.0208--Trachypithecus obscurus
|        |        |
|        |        +-0.0331--Homo sapiens
|        |
|        +-0.0529-+-0.0447--Cebus albifrons
|                 |
|                 +-0.0285--Ateles geoffroyi
|
+-0.1933-+-0.0190--Tarsius bancanus
|        |
|        +-0.0261--Tarsius syrichta
|
+-0.2546--Otolemur crassicaudatus
```

Figure 1: Phylogenetic tree of primates from Figure 3 of Page and colleagues.

```
+-0.3379-+-0.1778-+-0.1839--Klebsiella pneumoniae
|        |        |
|        |        +-0.2878-+-0.0118-+-0.0008--Escherichia coli O157:H7 EDL933
|        |                 |        |
|        |                 |        +-0.0017--Escherichia coli O157:H7
|        |                 |
|        |                 +-0.1006-+-0.0459--Escherichia coli CFT073
|        |                          |
|        |                          +-0.0267-+-0.0078-+-0.0089--Shigella flexneri 2a
|        |                                   |        |
|        |                                   |        +-0.0274--Escherichia coli K12
|        |                                   |
|        |                                   +-0.2581-+-0.0818-+-0.0092--Salmonella enterica
|        |                                            |        |          serovar Typhi CT18
|        |                                            |        +-0.0454--Salmonella enterica
|        |                                            |                   serovar Typhimurium LT2
|        |                                            +-0.0008--Salmonella paratyphi A
|        |
|        +-0.1892--Erwinia carotovora subsp. atroseptica
|
+-0.1619-+-0.9389--Photorhabdus asymbiotica
|        |
|        +-0.0867-+-0.0093--Yersinia pestis KIM
|                 |
|                 +-0.0017--Yersinia pestis CO-92
|
+-0.0803--Yersinia enterocolitica
```

Figure 2: Segment near *Escherichia coli* K12of a phylogenetic tree based on 16S rRNA gene data (see text for details). The edge lengths are approximate and should not be considered definitive.

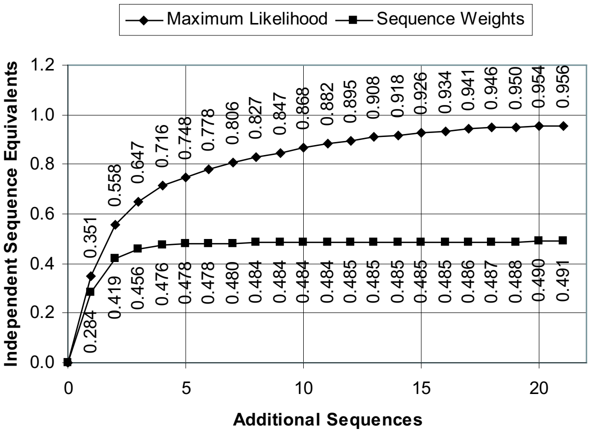

Figure 3: The effective number of additional independent sequences for the sequence-weights and maximum-likelihood approaches, as a function of the number of additional sequences. The sequences have been added to *Homo sapiens* so as to greedily maximize the efficiency at each addition.