Published in final edited form as:
*Stat Appl Genet Mol Biol*. 2005 ; 4: Article13.

# The Relative Inefficiency of Sequence Weights Approaches in Determining a Nucleotide Position Weight Matrix[*]

**Lee A. Newberg**[†,‡,§], **Lee Ann McCue**[†], and **Charles E. Lawrence**[†,‡,¶]
[†]*The Center for Bioinformatics, Wadsworth Center, New York State Department of Health, Albany, NY 12208-3425, USA.*

[‡]*The Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA.*

## Abstract

Approaches based upon sequence weights, to construct a position weight matrix of nucleotides from aligned inputs, are popular but little effort has been expended to measure their quality.

We derive optimal sequence weights that minimize the sum of the variances of the estimators of base frequency parameters for sequences related by a phylogenetic tree. Using these we find that approaches based upon sequence weights can perform very poorly in comparison to approaches based upon a theoretically optimal maximum-likelihood method in the inference of the parameters of a position-weight matrix. Specifically, we find that among a collection of primate sequences, even an optimal sequences-weights approach is only 51% as efficient as the maximum-likelihood approach in inferences of base frequency parameters.

We also show how to employ the variance estimators to obtain a greedy ordering of species for sequencing. Application of this ordering for the weighted estimators to a primate collection yields a curve with a long plateau that is not observed with maximum-likelihood estimators. This plateau indicates that the use of weighted estimators on these data seriously limits the utility of obtaining the sequences of more than two or three additional species.

### Keywords

sequence weights; maximum likelihood; motifs; phylogeny; sequencing; consensus distribution

## Introduction

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

[*] Author version of work published as *Statistical Applications in Genetics and Molecular Biology* (2005) **4**(1), article 13.

[§] To whom correspondence should be addressed. Lee.Newberg's account at Wadsworth.ORG

[¶] Current Address: Division of Applied Mathematics, Brown University, Providence, RI 02919-2612, USA.

There are many sequence weighting approaches, and several ways in which sequence weights have been employed. Vingron & 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 et al. (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 & 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 & Sibbald (1993) created a methodology for evaluating sequence-weighting schemes, compared four, and concluded that the approach of Altschul et al. (1989) is best when sequences are phylogenetically related. Otherwise, the approach of Sibbald & Argos (1990) was considered best. Henikoff & 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 & 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. Altschul et al. (1997) build a position-specific scoring matrix as an early step in their popular PSI-BLAST algorithm. May (2001) applies six different sequence weighting schemes as it tackles protein classification.

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 nucleotides for each sequence position relevant to the transcription factor binding. (See the work of Lawrence & Reilly (1990) for a good description of the statistical model, and Benos et al. (2002) for an analysis of its effectiveness.)

A single transcription factor may play a role in multiple genes across multiple species; however it is infeasible that we will have the binding site sequences for all relevant genes in all living species. Since we 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, which is mathematically guaranteed to provide the best efficiency. For the test cases to which these methods are applied, we find that the optimal sequence weights approach is significantly inferior.

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 & Roberts (1995) indicated mutation at a slower rate than elsewhere.

For clarity of example the bulk of the following addresses the direct use of sequence weights in relation to the direct use of a maximum-likelihood approach. We then discuss the implications for the more complex Bayesian, mixture models currently in common use.

## System and Methods

For each algorithmic approach we calculate the relative efficiency of nucleotide-frequency estimates using the statistical approach described by Kendall & Stuart (1979), Sections 17.28–29. This efficiency is calculated via a total estimator variance (also termed, mean square error), which is a sum over the four possible nucleotides that could occur in a shared consensus distribution for a given set of sequence positions. The term of the sum for a given nucleotide is the expected square of the deviation of the estimator for the probability of that nucleotide from the underlying model value for that parameter. 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.

In our analysis we consider a single gene and its orthologs in the other species, and we assume that we have an alignment of the corresponding upstream intergenic regions. Within such an alignment, we consider a transcription factor binding site and its orthologs indicated by the alignment. Within such a *multi-species binding site*, we focus upon a position (also termed, column) and the nucleotides in such a position are modeled to be descendant from a single nucleotide in the common ancestral species. However, this analysis applies equally well to each position in the aligned binding site and thus to them all.

We observe that when we have a single equilibrium distribution that governs a fixed number of sequence positions from each of $S$ statistically independent sequences, the total estimator variance for that equilibrium will be $1/S$ of the total estimator variance we would have had had we used just the data from a single sequence. Thus, even when we are considering phylogenetically-related (*i.e.*, statistically dependent) multi-species aligned sequence data, we report the *efficiency* (or *effective number of independent observations*, or *effective species count*) of the data to be the total estimator variance if we use the sequence of just a single species, divided by the total estimator variance derived from the complete set of sequences:

$$\mathit{effective\ species\ count} = \frac{\mathit{total\ \ variance}(\mathit{single\ species})}{\mathit{total\ \ variance}(\mathit{all\ species})} \tag{1}$$

The effective species count gives a comparative measure of the ability of the data at one multi-species binding site to yield a good estimate of its equilibrium distribution, relative to that of multiple sites from a single species. For instance, suppose there is a set of sixteen species that has an effective species count of 3.1. This finding indicates that the discovery of a multi-species binding site (*i.e.*, an aligned set of sixteen transcription factor binding sites descendant from a transcription factor binding site in the common ancestral species) is slightly more statistically useful than 3 distinct binding sites from a single-species data set. Again, this analysis applies to each of the positions in the aligned site and thus to all of them.

### Phylogenetic Data Model

We use the statistical phylogenetic model for the likelihood of aligned phylogenetic sequence data first described by Neyman (1971) and Felsenstein (1981). As is common, we track nucleotide mutations/substitutions from an ancestral sequence to a descendant sequence as a matrix *M*, in which the rows of the matrix correspond to the different possibilities for the nucleotide in the ancestral sequence, and the columns of the matrix correspond to the same nucleotide possibilities in the descendant sequence. For instance, with the nucleotides ordered as (*A*, *T*, *C*, *G*), the matrix is written:

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

where, for example, Pr[*A*|*C*] is the probability that the descendant sequence has an *A* where the ancestral 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 successive mutations at a single position are always included in the count) as the sum of the edge lengths along the path that connects those two sequences.

We choose the nucleotide substitution 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 nucleotides) as allowed by the model of Hasegawa et al. (1985). (But, see the Discussion Section.) The nucleotide substitution matrix between two sequences separated by a path of length *x* is:

$$M_x = e^{-kx} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \left(1 - e^{-kx}\right) \begin{pmatrix} \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \end{pmatrix} \qquad (2)$$

$$k = \frac{1}{1 - \left(\theta_A^2 + \theta_T^2 + \theta_C^2 + \theta_G^2\right)} \geq \frac{4}{3}. \qquad (3)$$

This model has the necessary features that as $x \to 0$, the substitution matrix is the identity matrix; that as $x \to +\infty$, the substitution matrix gives an equilibrium distribution independent of which nucleotide we started with (*i.e.*, all of the rows are equal); and that $M_{a+b} = M_a M_b$, correctly modeling that the substitution 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*. We choose *k* according to the convention of Lanave et al. (1984). By this convention, when the ancestral sequence begins in the equilibrium distribution the expected number of nucleotide substitutions per position between ancestor and descendant sequences implied by the substitution matrix $M_x$ is *x*.

Without this choice of normalization, one proposed distribution might be penalized relative to another—not because the equilibrium is less reflective of 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.

## Total Estimator Variance of the Sequence Weights Approach

With a single aligned sequence position (*i.e.*, a column from a sequence alignment), sequence-weighting estimates are obtained as follows:

$$\hat{\theta}_b = \sum_s w_s D_{sb} \tag{4}$$

where $D_{sb}$ is 1 if sequence $s$ has nucleotide $b$ at the position 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 et al. (1999), and depicted in Figure 1, and an assumption for the equilibrium distribution $\vec{\theta}^*$ to define $k^*$ via Equation 3, 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 4. We measure average squared distance of these sampled $\hat{\theta}_b$ values to the model mean $\theta_b^*$ and repeat the experiment for each nucleotide $b$. 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 4, the total variance of these estimators is computed as

$$\begin{aligned}\sum_b \mathrm{Var}\left[\hat{\theta}_b\right] &= \sum_b \mathrm{E}\left[\left(\hat{\theta}_b - \theta_b^*\right)^2\right] \\ &= \sum_{s,s',b} w_s w_{s'} \left(\mathrm{E}\left[\left(D_{sb} - \mathrm{E}\left[D_{sb}\right]\right)\left(D_{s'b} - \mathrm{E}\left[D_{s'b}\right]\right)\right]\right) \\ &= \vec{w}^T C \vec{w}\end{aligned} \tag{5}$$

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

$$C_{ss'} = \sum_b \mathrm{Cov}\left[D_{sb}, D_{s'b}\right],$$

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 $x$:

$$\begin{aligned}\mathcal{J}_x &= e^{-k^* x}\begin{pmatrix}\theta_A^* & 0 & 0 & 0 \\ 0 & \theta_T^* & 0 & 0 \\ 0 & 0 & \theta_C^* & 0 \\ 0 & 0 & 0 & \theta_G^*\end{pmatrix} \\ &+\left(1 - e^{-k^* x}\right)\begin{pmatrix}\theta_A^*\theta_A^* & \theta_A^*\theta_T^* & \theta_A^*\theta_C^* & \theta_A^*\theta_G^* \\ \theta_T^*\theta_A^* & \theta_T^*\theta_T^* & \theta_T^*\theta_C^* & \theta_T^*\theta_G^* \\ \theta_C^*\theta_A^* & \theta_C^*\theta_T^* & \theta_C^*\theta_C^* & \theta_C^*\theta_G^* \\ \theta_G^*\theta_A^* & \theta_G^*\theta_T^* & \theta_G^*\theta_C^* & \theta_G^*\theta_G^*\end{pmatrix}\end{aligned} \tag{6}$$

It follows that

$$\begin{aligned} C_{ss'} &= \sum_b \mathrm{Cov}\left[D_{sb}, D_{s'b}\right] = \sum_b \left[\left(\mathcal{J}_x\right)_{bb} - \left(\theta_b^*\right)^2\right] \\ &= \sum_b e^{-k^* x}\left(\theta_b^* - \left(\theta_b^*\right)^2\right) = \frac{e^{-k^* x}}{k^*}. \end{aligned} \tag{7}$$

Setting zero equal to the gradient of the right-hand side of Equation 5 with respect to the vector $\vec{w}$ (while using a La-Grange multiplier to ensure that $\sum_s w_s = 1$), we find optimal sequence weights:

$$\vec{w} = \frac{C^{-1}\vec{1}}{\vec{1}^T C^{-1}\vec{1}} \tag{8}$$

$$\min_{\vec{w}} \vec{w}^T C \vec{w} = \frac{1}{\vec{1}^T C^{-1}\vec{1}} \tag{9}$$

where $\vec{1}$ is the column vector of all ones. Note that Equation 8 is identical in form to equations which appeared in the work of Altschul et al. (1989), Vingron & Sibbald (1993), and Arvestad & Bruno (1997). In the first, *C* was instead a matrix of tree path lengths 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 substitution matrix, and the formula was used to compute a precise consensus distance.

In our model we have *n* aligned sequence positions (*e.g.*, the first position in each of *n* multi-species binding sites), which are assumed to be independent given their shared nucleotide consensus distribution. The total estimator variance in this situation differs from Equation 9 by a factor of 1/*n*:

$$\frac{1}{n}\left(\frac{1}{\vec{1}^T C^{-1}\vec{1}}\right). \tag{10}$$

## Total Estimator Variance of the Maximum Likelihood Approach

As we did 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 asymptotic approach via the Fisher information matrix will work. (See Kendall & Stuart (1979), Section 17.39.) A data sample *D* for a multi-species sequence position (*i.e.*, the specification of a nucleotide for each of the species) is assumed to occur with frequency proportional to its probability $\Pr\left[D \mid \vec{\theta}^*\right]$; assuming the underlying phylogenetic model of Equation 2, based upon some $\vec{\theta}^*$. The expected log-likelihood of a model based upon an equilibrium $\vec{\theta}$ is calculated as

$$\mathrm{LL}(\vec{\theta}) = n\sum_D \log(\Pr[D \mid \vec{\theta}])\Pr\left[D \mid \vec{\theta}^*\right]$$

where *n* is the number of independent multi-species sequence positions sharing the consensus distribution. (If the number of terms in the sum is too large, we can always revert to sampling *D* proportionately to $\Pr\left[D \mid \vec{\theta}^*\right]$.) Intuitively, our confidence in the maximum-likelihood estimate depends on the shape of $\mathrm{LL}(\vec{\theta})$ at its maximum $\vec{\theta}^*$; the more steeply $\mathrm{LL}(\vec{\theta})$ 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 $\mathrm{LL}(\vec{\theta})$ 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. As with Expression 10, the dependence of the total estimator variance on *n* will be through a factor of 1/*n*.

## Results

Using a uniform distribution $\vec{\theta}^* = (0.25, 0.25, 0.25, 0.25)$, we find that the use of optimal sequence weights with the phylogenetic tree of primates from Figure 3 of Page et al. (1999), as depicted in Figure 1 herein, gives an effective species count (as defined by Equation 1) of 1.49. This means that, for the purpose of precisely determining a column of a position-weight matrix, each multi-species binding site located in aligned sequence data from these species will be as statistically useful as a multi-species binding site located in aligned sequence data of 1.49 independent species, or as statistically useful as 1.49 independent single-species binding sites located in single-species sequence data.

Looking at just human sequence data would give us an effective species count of 1.0. Thus, the information from the non-human primates adds an effective 0.49 independent sequences to our ability to determine the distribution of nucleotides. In contrast, the maximum-likelihood approach gives an effective species count of 1.96, an increase of 0.96 over using human alone. For sequence weights, the increase in the effective species count is 51% 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 species count for a consensus distribution. Figure 2 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 2 indicates that all of the sequences beyond the first two or three additional sequences contribute little to the estimates.

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 maximum-likelihood method of PHYLIP (http://evolution.genetics.washington.edu/phylip.html). This is a common technique employed in estimating phylogenetic tree topologies when little data is available—despite the general consensus that the applicability of standard nucleotide substitution models to the conserved nucleotides of a gene is suspect. 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 on average. Although the edge lengths in this tree, depicted in Figure 3, 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 species. In contrast, the maximum-likelihood approach gives the equivalent of 2.40 additional independent species. The sequence-weights approach performed relatively better for this tree, although still suboptimally; for the sequence-weights approach the effective number of additional independent species is 77% as large as that of the maximum-likelihood competitor.

### Bayesian Mixture Models for Sequence Weights

To this point in this analysis we have adopted a “frequentist” approach rather than a Bayesian approach in that we have have not incorporated an *a priori* distribution over $\vec{\theta}^*$ in the calculation of our estimates $\hat{\theta}$. For tractability purposes, in a Bayesian approach conjugate *a priori* distributions are usually employed. (These are distributions for which the *a priori* and *a posteriori* distributions are of the same mathematical form.) With such a choice, it is easy to speak of an *a priori* distribution as being implied by *pseudo* data—and the maximum *a posteriori* value of an estimator for the data can be computed as the maximum-likelihood estimator for the combined data and pseudo data. This interpretation of priors informs us of the their influence on our efficiency analysis.

In this case the pseudo data can be seen as pseudo sequence data that are mutually statistically independent of each other and the actual data. Each pseudo sequence represents an independent observation equivalent with either approach. With the addition of these sequences both approaches will yield reduced uncertainty in the estimators. For instance, in Figure 2, both curves will be shifted upwards; but, as a consequence of the Cramer-Rao Theorem, in no case will the size of the shift of the sequence weights curve be more than that of the maximum-likelihood curve.

With a mixture model of Dirichlet priors, which is used in many current sequence weights approaches, the analysis is more complicated, but the Cramer-Rao Theorem still applies; in no case will the size the shift of the sequence weights curve be more than that of the maximum-likelihood curve.

## Discussion

While we have picked the Felsenstein (1981) nucleotide substitution model, which does not directly recognize the differences between nucleotide transitions and nucleotide transversions, and have evaluated it with a uniform equilibrium distribution, we do not believe the results to be highly sensitive to these choices. For instance, in an extreme equilibrium example of the Felsenstein (1981) model, in which one nucleotide is modeled to occur 90% of the time and the occurrence of the other three nucleotides is 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. In another example, with the nucleotide substitution model of Hasegawa et al. (1985) (with a value of 10 for the ratio of a transition rate to a transversion rate), evaluated with a uniform equilibrium distribution of nucleotides, the sequence-weights approach is 48% as efficient as the maximum-likelihood approach.

The choice of total estimator variance (also termed, mean square error) 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 nucleotides. Thus, it is reasonable to evaluate an approach's efficiency on the basis of how well it can determine a probability distribution of nucleotides. 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 $\vec{\theta}$.

Our analysis of the maximum-likelihood approach is based upon the asymptotic case in which the number of transcription factor binding sites is large. The Cramer-Rao Theorem (see, *e.g.*, Stuart et al. (1999)) guarantees that this asymptotic analysis provides a lower bound on the variance of the estimators.

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 nucleotides that accidentally coincide; this alignment bias may cause nucleotide-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 occurrences of $kx$ in Equation 2 by a positive factor $\gamma \leq 1$ to indicate an expected reduced rate of mutation, effectively shrinking the phylogenetic tree. 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 nucleotides.

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 $\gamma$ to vary by sequence position, but leaving $\vec{\theta}$ fixed across sequence positions.) It may prove to be the case that a combined approach, which maximizes the joint probability of the distribution $\vec{\theta}$ and the mutation rate $\gamma$, is 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.

Many sequence weights approaches are designed for speed; and a maximum-likelihood approach could be considerably slower. There may be applications where use of the rougher, faster method is more beneficial than the more precise, slower method.

The sequence-weights approach may be slightly less efficient than we have indicated here. Equation 8 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, with a corresponding decrease in the effective species count. 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 7. For additional accuracy when the result in any approach depends on $\vec{\theta}^*$, we might use the computed $\hat{\theta}$ vector as the value of $\vec{\theta}^*$ for a subsequent iteration, repeating until sufficient convergence is achieved.

## Conclusion

We have developed a procedure, based upon phylogenetic relationships, to determine optimal weights for a sequence weights approach to computing a consensus distribution of nucleotides at any position of an alignment of nucleic acid sequences.

We have shown that the use of optimal sequence weights performs significantly worse than a maximum-likelihood method based upon phylogenetic relationships. In particular, for aligned sequences from primates (as represented by the phylogenetic tree of Page et al. (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, for sequences from human and two well-chosen non-human primates, the maximum-likelihood approach is more efficient than is use of all twenty-two species with the sequence-weights approach. We also find that, aligned sequences from *Escherichia coli* K12 and related species of bacteria (as depicted in the phylogenetic tree of Figure 3) show a comparable 77% relative efficiency.

Generally, our procedure gives a means to estimate the loss in using a sequence weights approach for any given tree, and thus provides a measure, or at least a bound, of the efficiency cost of using a faster, weighting approach. In some cases the loss in efficiency may be a price worth paying for the added speed of computation. When this is not the case then, rather than the use of sequence weights to estimate the distribution of nucleotides at each position of the aligned sequences, we recommend an alternative recipe. First, obtain a phylogenetic tree that connects the species in question. Second, use the tree topology and edge lengths (as well as any position-specific variations implied by the more advanced models) to calculate the most likely consensus distribution of nucleotides at each position of the aligned sequence.

### Acknowledgments

The authors gratefully acknowledge the use of the Computational Molecular Biology Core of the Wadsworth Center. This work was supported by the National Institutes of Health (National Human Genome Research Institute) grant 5R01HG001257 to C. E. L. and by the Department of Energy grants DE-FG02-01ER63204 and DE-FG02-04ER63942 to C. E. L. and L. A. M.

## 1 References

Altschul SF, Carroll RJ, Lipman DJ. Weights for data related by a tree. J Mol Biol 1989;207(4):647–653. [PubMed: 2760928]PubMed 2760928

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17): 3389–3402. [PubMed: 9254694]PubMed 9254694

Arvestad L, Bruno WJ. Estimation of reversible substitution matrices from multiple pairs of sequences. J Mol Evol 1997;45(6):696–703. [PubMed: 9419247]PubMed 9419247

Benos PV, Bulyk ML, Stormo GD. Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res 2002;30(20):4442–4451. [PubMed: 12384591]PubMed 12384591

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. Phylo-genetic shadowing of primate sequences to find functional regions of the human genome. Science 2003;299 (5611):1391–1394. [PubMed: 12610304]PubMed 12610304

Bruno WJ. Modeling residue usage in aligned protein sequences via maximum likelihood. Mol Biol Evol 1996;13(10):1368–1374. [PubMed: 8952081]PubMed 8952081

Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 1981;17(6):368–376. [PubMed: 7288891]PubMed 7288891

Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 1985;22(2):160–174. [PubMed: 3934395]PubMed 3934395

Henikoff S, Henikoff JG. Position-based sequence weights. J Mol Biol 1994;243(4):574–578. [PubMed: 7966282]PubMed 7966282

Kendall, MG.; Stuart, A. *Inference and Relationship*, vol. 2, of *The Advanced Theory of Statistics*. Fourth. Macmillan Publishing Co, Inc; 1979.

Krogh, A.; Mitchison, GJ. Maximum entropy weighting of aligned sequences of proteins or DNA. In: Rawlings, C.; Clark, D.; Altman, R.; Hunter, L.; Lengauer, T.; Wodak, S., editors. Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology. American Association for Artificial Intelligence AAAI Press.; Robinson College, Cambridge, United Kingdom: 1995. p. 215-221.PubMed 7584440

Lanave C, Preparata G, Saccone C, Serio G. A new method for calculating evolutionary substitution rates. J Mol Evol 1984;20(1):86–93. [PubMed: 6429346]PubMed 6429346

Lawrence CE, Reilly AA. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins 1990;7(1):41–51. [PubMed: 2184437]PubMed 2184437

May ACW. Optimal classification of protein sequences and selection of representative sets from multiple alignments: application to homologous families and lessons for structural genomics. Protein Eng 2001;14(4):209–217. [PubMed: 11391012]PubMed 11391012

McCue LA, Thompson W, Carmack CS, Lawrence CE. Factors influencing the identification of transcription factor binding sites by cross-species comparison. Genome Res 2002;12(10):1523–1532. [PubMed: 12368244]PubMed 12368244

Neyman, J. Molecular studies of evolution: a source of novel statistical problems. In: Gupta, SS.; Yackel, J., editors. Statistical Decision Theory and Related Topics. Academic Press; New York, NY: 1971. p. 1-27.

Page SL, Chiu C.-h. Goodman M. Molecular phylogeny of Old World monkeys (cercopithecidae) as inferred from γ-globin DNA sequences. Mol Phylogenet Evol 1999;13(2):348–359. [PubMed: 10603263]PubMed 10603263

Rajewsky N, Socci ND, Zapotocky M, Siggia ED. The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. Genome Res 2002;12(2):298–308. [PubMed: 11827949]PubMed 11827949

Sibbald P, Argos P. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. J Mol Biol 1990;216(4):813–818. [PubMed: 2176240]PubMed 2176240

Silvey, SD. Optimal Design: An Introduction to the Theory for Parameter Estimation. Chapman and Hall; 1980.

Stuart, A.; Ord, JK.; Arnold, S. *Classical Inference and the Linear Model* vol. 2A, of *Kendall's Advanced Theory of Statistics*. sixth. Arnold; London: 1999. p. 17.13-17.27.

Vingron M, Argos P. A fast and sensitive multiple sequence alignment algorithm. Comput Appl Biosci 1989;5(2):115–121. [PubMed: 2720461]PubMed 2720461

Vingron M, Sibbald P. Weighting in sequence space: a comparison of methods in terms of generalized sequences. Proc Natl Acad Sci U S A 1993;90(19):8777–8781. [PubMed: 8415606]PubMed 8415606

Yang Z, Roberts D. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol Biol Evol 1995;12(3):451–458. [PubMed: 7739387]PubMed 7739387

**Figure 1.**

Phylogenetic tree of primates from Figure 3 of Page et al. (1999). Each edge shows the number of nucleotide substitutions between an ancestral and descendant species (including multiple substitutions at a single position) that is expected in $10^4$ sequence positions. Edges are drawn to scale, except for the very shortest.

**Figure 2.**

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. The values are per multi-species site found; for example, with the maximum-likelihood approach, finding 5 multi-species sites in the ten best species would be as good as 5 × 0.868 = 4.34 additional single-species sites, or 9.34 single-species sites found in single-species data.

**Figure 3.**

Segment near *Escherichia coli* K12 of an unrooted phylogenetic tree based on 16S rRNA gene data (see text for details). Each edge shows the number of nucleotide substitutions between an ancestral and descendant species (including multiple substitutions at a single position) that is expected in $10^4$ sequence positions. These edge lengths are approximate and should not be considered definitive.