Functional DNA Positions Change as Frequently as Do Neutral Ones

Lee A. Newberg*

Technical Report 05-08
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York 12180-3590, USA

March 29, 2005

Abstract

Much work has been done in statistically describing how DNA changes through evolution. Relative to models for neutral DNA sequence positions, models for functional DNA sequence positions usually include a far-from-uniform equilibrium probability distribution and a significantly reduced rate of change. We examine the mathematical basis for these modifications to the functional-DNA model.

We find that, although non-lethal selection pressures will skew the equilibrium probability distribution of alleles, thus making fitter alleles more common in the population, selection pressures do not significantly affect the rate of allele change.

Even beyond consideration of DNA sequences, the word "conservation" would more appropriately be used to indicate the non-uniformity of an equilibrium probability distribution of alleles, rather than to denote a reduced rate of change. The design of statistical models of substitution for functional alleles should reflect this characterization.

1 Introduction

It is generally believed that the rate of change in functional DNA positions is lower than that in neutral / non-functional DNA positions, and many statistical models have incorporated or rediscovered this accepted truth (*1–4*). We explore how this might have become the accepted wisdom and why, for soft / non-lethal selective pressures, it appears to be false.

Our interest is in models applicable to functional DNA positions, such as those within a *cis*-regulatory element (also termed a transcription factor binding site). Here we develop a statistical approach that can combine any nucleotide substitution model with the well-established Darwinian fitness formalism (*5–7*). We find that, although soft selection pressures will skew the equilibrium probability distribution of nucleotides, thus making fitter nucleotides more likely in the population, these pressures do not significantly affect the instantaneous rate of substitutions.

1.1 Mischaracterization of Conservation from the Counting of Mismatches

Although insertions and deletions are an important part of DNA evolution we do not, for the sake of clarity, consider them in the following. Instead, we focus on in-place nucleotide substitutions—those nucleotide mutations that survive selection.

As a thought experiment, suppose that we have a gapless, multiple sequence alignment of genes (*e.g.*, small ribosomal subunit RNA genes, which are present in all species) from a collection of very distantly related species. For a neutral DNA position of such an alignment, we expect a near-uniform equilibrium probability distribution of nucleotides across the sequences. In particular, if we choose two of the species at random, there is about a 25% chance that they have the same nucleotide at such a position, and about a 75% chance that there is a mismatch.

When the equilibrium probability distribution of alleles for a functional DNA position departs from uniform, the probability of mismatch in the joint probability distribution for two species descendant from a common ancestor will be capped at

$$\Pr[\text{mismatch}] \quad = \quad 1 - (\theta_A^2 + \theta_T^2 + \theta_C^2 + \theta_G^2)\,, \qquad (1)$$

which is below 75%. For example, if $\vec{\theta} = (70\%, 10\%, 10\%, 10\%)$, then the probability of a mismatch will not exceed 48%.

Given that $48\% \ll 75\%$, it is natural to label such a position as conserved. Notice that it is the skewed equilibrium probability distribution that caps the mismatch fraction, rather than some change in the rate of allele substitution, and that the cap would continue to hold even if the rate of substitution were *increased* at functional DNA sequence positions.

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

1.2 Mischaracterization of Conservation from Statistical Mixture Models

As another example that might lead us astray as to the true meaning of conservation, consider the collection of all protein first-codon positions. First-codon positions are evolutionarily conserved. As before, we imagine a gapless, multiple alignment of sequences from a collection of very distantly related species. For the sake of clarity, we suppose that the first-codon positions are of four types: A *predominant*, with equilibrium probability distribution $\vec{\theta} = (70\%,\ 10\%,\ 10\%,\ 10\%)$, and, analogously, T *predominant*, C *predominant*, and G *predominant*.

If we choose two of the sequences at random (with replacement), then the equilibrium joint probability distribution for a given A *predominant* first-codon position is:

	A	T	C	G
A	49%	7%	7%	7%
T	7%	1%	1%	1%
C	7%	1%	1%	1%
G	7%	1%	1%	1%

, (2)

and likewise for T *predominant*, C *predominant*, and G *predominant*, when rows and columns are appropriately permuted. This equilibrium joint probability distribution is what one would get from the FEL81 model (*8*) (parameterized with $\vec{\theta} = (70\%,\ 10\%,\ 10\%,\ 10\%)$), a degenerate form of the HKY85 model (*9*). In these models, or with other established techniques (*10, 11*), the evolutionary distance implied by this equilibrium joint probability distribution is infinite, as expected.

However, when we examine the *mixture* of all first-codon position data, and we suppose that each of the four kinds is equally likely, then the combined equilibrium joint probability distribution is:

	A	T	C	G
A	13%	4%	4%	4%
T	4%	13%	4%	4%
C	4%	4%	13%	4%
G	4%	4%	4%	13%

. (3)

This is the equilibrium joint probability distribution that one would get from the JC69 model (*12*), a degenerate form of the FEL81 and HKY85 models (*8, 9*). With this model, or with other established techniques (*10, 11, 13*), one finds that the evolutionary distance implied by this equilibrium joint probability distribution is approximately 0.7662 nucleotide substitutions.

Because this distance is significantly smaller than the near-infinite distance that we observe for neutral positions, it is natural to label such a position as conserved (*4*). Notice that it is the mixture of the equilibrium joint probability distribution models that serves to shorten the phylogenetic distance, rather than some change in the rate of allele substitution, and that the apparent shortening would occur even if the rate of substitution were increased at functional DNA sequence positions.

1.3 Mischaracterization of Conservation from Fixation within a Species

Building on earlier results (*14,15*), the HB98 model (*2*), a popular statistical approach for nucleotide substitutions in functional DNA, carefully considers fixation, the process in which a mutation in a single organism leads to a change in all organisms in the species. While this focus on the species scale is common, it may not be justified. Here we briefly explore why, in this context, species fixation may be a specious fixation.

Although we frequently label a sequenced genome with the name of the species that it represents, what we have actually sequenced is usually a single individual organism within that species. Likewise, when we construct an evolutionary / phylogenetic tree from genome sequences, it is more precise to say that the tree relates the sequenced individuals, rather than their species. In the case of sequence positions within *cis*-regulatory elements, selection pressures are often soft, in that any nucleotide may be more or less fit than another, but substitutions are not lethal. In particular, it is entirely possible that the observed nucleotide at a given position is the result of a relatively recent, non-lethal mutation in this individual organism's ancestral line, and is *not* representative of the species as a whole.

Furthermore, the HB98 model assumes that nucleotide substitution is reversible (*i.e.*, phylogenetic trees can be re-rooted arbitrarily, so long as edge lengths are preserved); however there is evidence that substitution is not reversible (*4*).

2 Methods

In this section, we focus upon the elements of our statistical model that represent departures from the established, nucleotide phylogeny methodology. The analysis is most directly applicable to haploid organisms, which have a single copy of each chromosome in each cell. More details can be found in the appendix.

2.1 Substitution Models and Population Models

Traditionally, for a given DNA sequence position, an edge of a phylogenetic tree is described by a *substitution* matrix

$$M = \begin{pmatrix} \Pr[A|A] & \Pr[T|A] & \Pr[C|A] & \Pr[G|A] \\ \Pr[A|T] & \Pr[T|T] & \Pr[C|T] & \Pr[G|T] \\ \Pr[A|C] & \Pr[T|C] & \Pr[C|C] & \Pr[G|C] \\ \Pr[A|G] & \Pr[T|G] & \Pr[C|G] & \Pr[G|G] \end{pmatrix}, \quad (4)$$

where $\Pr[b_{\rm des}|b_{\rm anc}]$ is the probability that an arbitrarily chosen descendant will show nucleotide $b_{\rm des}$ when its ancestor shows nucleotide $b_{\rm anc}$. The matrix M is applied to a row vector representing a nucleotide probability distribution for the ancestor $\vec{\alpha} = (\alpha_A, \alpha_T, \alpha_C, \alpha_G)$ by multiplication on the right, to give the descendant's nucleotide probability distribution $\vec{\delta} = (\delta_A, \delta_T, \delta_C, \delta_G)$:

$$\vec{\alpha} M \quad = \quad \vec{\delta}\,. \tag{5}$$

The matrix M has the property that the sum of the elements in any row is one.

Instead, we here employ a *population* model, in which the elements of a row of M need not sum to one. That is, because of varying fitnesses, some alleles may have more progeny than do other alleles, and we indicate this by giving those alleles a larger row sum. For example, the model

$$M \quad = \quad \begin{pmatrix} 1.2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{6}$$

indicates that no mutations occur, and that allele A will have 20% more progeny than will the other alleles. If an ancestral individual is equally likely to have each allele, $\vec{\alpha} = (25\%,\ 25\%,\ 25\%,\ 25\%)$, then we compute $\vec{\delta} = (0.30,\ 0.25,\ 0.25,\ 0.25)$, indicating that allele A individuals benefit from their superior fitness. The chance that a randomly chosen descendant has the A allele is $0.30\ /\ (0.30 + 0.25 + 0.25 + 0.25) = 28.6\%$.

2.2 Generation Model

Approximating the biology of a haploid organism, we model a single generation as a possible mutation during DNA replication, followed by a mutationless selection effect between replication events. Mathematically, the consequence of this assumption is that the population model matrix for a generation is the matrix product of a substitution model, in which each matrix row sum *is* one, and a diagonal matrix, like that shown in Equation 6.

We also assume that the timescale for a generation is much shorter than both the timescale for mutations and the timescale for selection effects. In particular, we are assuming that mutations within *cis*-regulatory elements are not immediately lethal, but that they instead affect the fitness of a cell line over a timescale of multiple generations. Mathematically, the consequence of this assumption is that both the mutation matrix and the diagonal selection matrix for a generation are quite close to the identity matrix, which is the matrix having 1's on the main diagonal and 0's elsewhere.

2.3 Phylogenetic Model

To calculate the appropriate population model matrix for a phylogenetic tree edge representing a time of G generations, we take the single-generation matrix and raise it to the Gth power. This is efficiently evaluated via a matrix spectral decomposition.

To calculate the likelihood of a proposed phylogenetic tree, from observed nucleotides at a given DNA alignment position, we invoke the well-established Felsenstein algorithm (*8*), *twice*. The first invocation is the traditional approach, but using the population model matrices as if they were substitution model matrices.

We then scale this computed "unnormalized likelihood" by dividing it by the result of a second invocation of the algorithm, which is designed to efficiently compute the sum of these unnormalized likelihoods over all 4^s possible sets of observed nucleotides (where s is the number of sequences aligned). In an intuitive sense, this normalization process is a probabilistic conditioning upon the fact that all of the aligned sequences correspond to organisms that survived until the time of their sequencing.

3 Results

With the standard approach to nucleotide phylogeny, modified as described above, we discover that little change to the instantaneous rate of substitution results from soft selection pressures. Also, we evaluate existing models for their consistency with the model presented here. Additional details are available in the appendix.

3.1 Instantaneous Rate of Substitution

When a neutral DNA alignment position obeys the HKY85 model (*9*) with nucleotide equilibrium probability distribution $(\beta_A, \beta_T, \beta_C, \beta_G) = (30\%, 30\%, 20\%, 20\%)$ and transition / transversion ratio of $\kappa = 3$, the instantaneous rate of nucleotide substitutions will, in the presence of selection effects, be nearly unchanged from the rate for the neutral DNA positions. The rate ratio will fall in the interval $[0.901, 1.148]$, regardless of the soft selection fitnesses. Furthermore, when neutral DNA positions are described by the HKY85 model with $\vec{\beta}$ uniform and $\kappa = 1$ (*i.e.*, the simpler JC69 model (*12*)), then the instantaneous rate of nucleotide substitutions is not changed at all.

These results indicate that for reasonable, neutral DNA models, even with arbitrary, soft selection fitnesses, it may be appropriate to approximate the nucleotide substitution rate for functional DNA positions as being unchanged from the neutral rate.

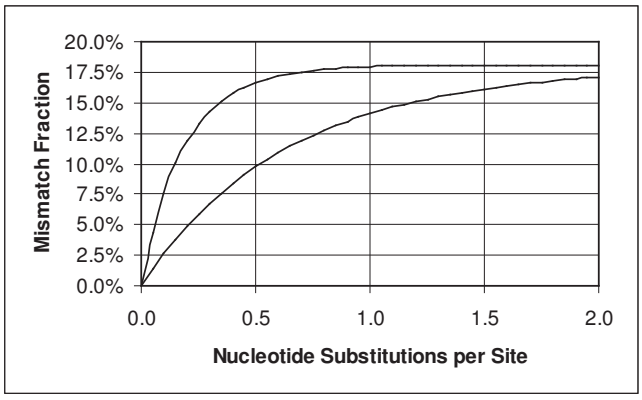

Figure 1: The probability of mismatch in the case that selection favors one nucleotide, to the extent that the equilibrium probability distribution is $(28/31,\ 1/31,\ 1/31,\ 1/31)$. The x-axis is the expected number of nucleotide substitutions per neutral DNA sequence position, between two individuals. The y-axis shows the probability of mismatch according to the model presented here (upper curve), and according to the HB98 model (lower curve). As a side-effect of its focus on species fixation, the HB98 model underpredicts the mismatch probability.

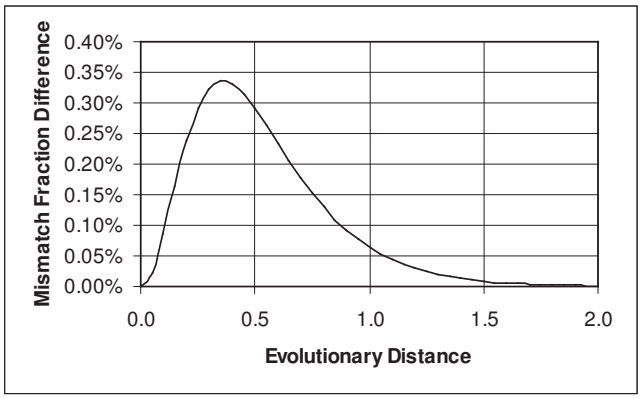

Figure 2: The *difference* between the probabilities of mismatch derived from each of two models, in the case that selection favors one nucleotide, to the extent that the equilibrium probability distribution is $(28/31,\ 1/31,\ 1/31,\ 1/31)$. The x-axis is the expected number of nucleotide substitutions per neutral DNA sequence position, between two individuals. The y-axis shows the amount by which the HKY85 model overpredicts the probability of mismatch, relative to our model. Although the HKY85 model is designed to reflect an arbitrary equilibrium probability distribution in the *absence* of selection effects, the extent of its overprediction of the mismatch probability is only slight.

3.2 Evolutionary Timescales

In examining the case that selection favors one nucleotide, to the extent that the equilibrium probability distribution is $(28/31,\ 1/31,\ 1/31,\ 1/31)$, we find that the HKY85 model approximates the effects of selection better than does the HB98 model.

Figure 1 shows that the HB98 model deviates significantly from the model presented here. Although the models agree in the limit of infinite evolutionary separation, they disagree significantly for distances commonly found in a phylogenetic tree. The HB98 model significantly underpredicts the expected number of mismatches in functional DNA positions as a function of the expected number of substitutions in neutral DNA positions.

Figure 2 shows the *difference* in the number of mismatches predicted by the model presented here and the HKY85 model parameterized with the appropriate equilibrium probability distribution. Although there is a difference, it is small; accordingly, in situations where a traditional model must be used, it appears that the HKY85 model functions well.

4 Discussion

4.1 Reduction of Substitution Rate Is Inappropriate

Most current algorithms for locating *cis*-regulatory elements use a separate "foreground" equilibrium probability distribution $\vec{\theta}$ for each position within a *cis*-regulatory element. In such a situation, when a foreground equilibrium probability distribution $\vec{\theta}$ is known (or estimated) but a corresponding nucleotide fitness matrix is not, we can use, to some advantage, the similarity of the HKY85 model (*9*) (parameterized with $\vec{\theta}$ and the neutral ratio of transitions to transversions) to our model. In particular, for evaluation of a phylogenetic tree, it may be perfectly adequate to use the HKY85 model rather than the less efficient foreground model that we present here, given that the latter requires computation of the relative nucleotide fitnesses. On the other hand, the use of the HB98 model gives results that appear to be significantly less accurate than the results from the use of the HKY85 model.

4.2 Existing Algorithms Can Be Adapted

Because of the likelihood normalization that is inherent in our population model matrix approach, the final likelihood values do not change if we multiply all nucleotide fitnesses by a constant. In order to remove this extra degree of freedom, it may make sense if we restrict those fitness matrices that are allowed. This can be achieved by (i) scaling the fitnesses so that the maximum fitness is 1.0; (ii) scaling so that the generation model gives a population level that is stable (*i.e.*, neither asymptotically growing or shrinking); or (iii) employing other

criteria. Generally, with the use of a fixed background model for neutral DNA positions (such as the HKY85 model (*9*) with a specified $\vec{\beta}$ and κ) and one such restriction on fitnesses, we obtain a one-to-one correspondence between foreground equilibria $\vec{\theta}$ and selection fitness matrices. Thus, if within any algorithm we have a hypothesis for $\vec{\theta}$, then we can numerically determine the fitnesses. This allows adaptation of existing $\vec{\theta}$-based algorithms for use with the model presented here.

Note that, generally, the model presented here is not reversible, and we cannot arbitrarily re-root a phylogenetic tree before performing likelihood calculations.

4.3 General Applicability

Although we have focused on models for nucleotide substitution, the results will be similar for the heritable alleles of any trait. The distinction between phenotype and genotype in diploid organisms, which have two copies of each chromosome in each cell, complicates the analysis in that setting, but we conjecture similar results.

4.4 Conclusion

We have shown that that there is reason to believe that soft / non-lethal selection pressures exhibit their effect through a skewing of the equilibrium probability distribution, but that the effect on the overall instantaneous rate of nucleotide substitution is small. In particular, the word "conservation" might more appropriately be used to indicate the non-uniformity of an equilibrium probability distribution, rather than a reduced rate of substitution.

We can calculate the likelihood of a phylogenetic tree in the presence of selection effects using the population model matrices presented here, via two passes of Felsenstein's algorithm (*8*).

When we approximate the probability of observed data at a sequence position by using the nucleotide substitution model for neutral DNA positions, and by parameterizing that model with a foreground equilibrium probability distribution $\vec{\theta}$ instead of the background equilibrium probability distribution $\vec{\beta}$, we have evidence that it is better for the modeler to leave the overall instantaneous rate of substitution unchanged than it is for the modeler to reduce that rate *ad hoc*. Similarly, in part because the HB98 model (*2*) changes the rate of substitution at functional DNA positions, our evidence is that the HB98 model does not describe nucleotide substitutions well.

A Appendix (or Supplemental Online Materials

We establish notation in Sections A.1 and A.2. In Section A.3 we set up the equations for evaluating the instantaneous rate of nucleotide substitution for a simple ancestor and descendant relationship, and we adapt it to a more realistic scenario of two descendants of a common ancestor in Section A.4. In Section A.5 we apply these equations to the HKY85 model for nucleotide substitutions at neutral DNA positions, to find that soft selection pressures do not significantly affect instantaneous rates of substitution. In Section A.6 we compare our model to established models, to evaluate the quality of the models over evolutionary timescales. In Section A.7 we explain how the consistency of observed data with population model matrices can be evaluated with two invocations of the well-established algorithm of Felsenstein (*8*). Finally, in Section A.8 we discuss why the soft selection model presented here is appropriate in the analysis of *cis*-regulatory elements.

A.1 Substitution in the Absence of Selection

We write a substitution matrix M for an edge of a phylogenetic tree as indicated by Equation 4. For closely related ancestral and descendant individuals, M is likely to be not very different from the identity matrix I. If the evolutionary distance is short enough that the probability of two substitutions at a single sequence position is negligible, then the distance of M from I is reasonably assumed to be proportional to the edge length in the phylogenetic tree. That is, there is an instantaneous rate matrix R, such that

$$M(\epsilon) \quad = \quad I + \epsilon R + \mathcal{O}(\epsilon^2)\,, \tag{7}$$

where ϵ is the short edge length, and where $\mathcal{O}(\epsilon^2)$ indicates that terms of size ϵ^2 or smaller have been omitted.

It has been shown that, since M is a substitution model, we must insist that the off-diagonal elements of R be non-negative, and that the diagonal elements of R be non-positive numbers, set so that the elements of each row of R sum to 0 (*11*). This last condition is represented by the equation:

$$R\vec{1}^T \quad = \quad \vec{0}^T\,, \tag{8}$$

where $\vec{1} = (1,1,1,1)$, $\vec{0} = (0,0,0,0)$, and where the superscript T indicates a matrix transposition.

For longer edge lengths, we apply the matrix $M(\epsilon)$ as many times as necessary to attain length x:

$$M(x) \quad \approx \quad (I + \epsilon R)^{x/\epsilon}\,, \tag{9}$$

where this expression is easier to understand when x/ϵ is a positive integer. The expression is exact in the limit as $\epsilon \to 0^+$, so we take the limit (proof omitted) and write

$$M(x) \quad = \quad \lim_{\epsilon \to 0^+} (I + \epsilon R)^{x/\epsilon} = \exp(xR)\,, \tag{10}$$

a formula that is efficiently evaluated via a matrix spectral decomposition. (See Section A.6.)

In the limit, as $x \to +\infty$, the descendant's probability distribution $\vec{\delta}$ will converge to an equilibrium probability distribution $\vec{\beta}$, independent of the ancestral probability distribution $\vec{\alpha}$. We define a diagonal matrix,

$$D_{\vec{\beta}} \quad = \quad \begin{pmatrix} \beta_A & 0 & 0 & 0 \\ 0 & \beta_T & 0 & 0 \\ 0 & 0 & \beta_C & 0 \\ 0 & 0 & 0 & \beta_G \end{pmatrix} , \qquad (11)$$

and compute

$$J(x) \quad = \quad D_{\vec{\beta}} M(x) , \qquad (12)$$

a matrix commonly referred to as the equilibrium joint probability distribution matrix for an edge of length x. It represents the joint probability distribution between the ancestor and the descendant, in the case when the ancestor starts in the equilibrium probability distribution $\vec{\beta}$.

It is common to calibrate tree edge lengths by having them represent the expected number of substitutions per sequence position (including multiple substitutions at a single sequence position), when the ancestral probability distribution $\vec{\alpha}$ is in the equilibrium probability distribution $\vec{\beta}$. This is achieved only if

$$\mathrm{ods}\left(\frac{\partial}{\partial x} J(x) \Big|_{x=0} \right) \quad = \quad 1 , \qquad (13)$$

where $\mathrm{ods}()$ is the sum of the off-diagonal elements.

To verify the desired calibration, we differentiate the matrix expression of Equation 12 to obtain

$$\frac{\partial}{\partial x} J(x) \Big|_{x=0} \quad = \quad D_{\vec{\beta}} R . \qquad (14)$$

We then exploit Equation 8, so as to prove that the sum of all the elements of $D_{\vec{\theta}} R$, which can be written as $\vec{1} D_{\vec{\theta}} R \vec{1}^T$, is exactly zero. This means that the sum of the off-diagonal elements of $D_{\vec{\theta}} R$ is exactly the negative of the sum of the diagonal elements, *i.e.*, the desired calibration is achieved in the case that the matrix trace (*i.e.*, the sum of the diagonal elements) of $D_{\vec{\beta}} R$ is negative one:

$$\mathrm{ods}\left(\frac{\partial}{\partial x} J(x) \Big|_{x=0} \right) \quad = \quad -\mathrm{trace}\left(D_{\vec{\beta}} R \right) . \qquad (15)$$

This equation is applicable for any instantaneous rate matrix R, and is not restricted to, *e.g.*, the HKY85 model (9), or reversible models.

A.2 Substitution in the Presence of Selection

Even when no allele is lethal, it is reasonable, if some nucleotides are more fit than others, to assume that some of them will produce more offspring than will others. For a short phylogenetic tree edge length, the relative fitnesses will not be very influential, and it is reasonable that the population of each nucleotide will be multiplied by a factor near 1. That is, we can reasonably assume that, for short tree edges, the nucleotide substitution matrix is:

$$N(\epsilon) \quad = \quad (I + \epsilon R + \mathcal{O}(\epsilon^2))(I + \epsilon S + \mathcal{O}(\epsilon^2)) . \qquad (16)$$

Here we are using N, instead of M, to indicate that selection pressures are present. S is a diagonal matrix and if, *e.g.*, nucleotide A is more fit than T, then $S_{AA} > S_{TT}$. Much as before, we see that a longer edge length is described by the matrix

$$\begin{aligned} N(x) \quad &= \quad \lim_{\epsilon \to \infty} (I + \epsilon Q + \mathcal{O}(\epsilon^2))^{x/\epsilon} \\ &= \quad \exp(xQ) , \end{aligned} \qquad (17)$$

where $Q = R + S$.

However, unless all fitnesses are exactly zero, we now have the case that the sum of the elements of $\vec{\alpha} N(x)$ is not necessarily equal to the sum of the elements of $\vec{\alpha}$. That is, the overall population may have shrunk or grown, and we have to compute the nucleotide frequencies of the descendant population with this in mind. Specifically,

$$\vec{\delta} \quad = \quad \frac{\vec{\alpha} N(x)}{\vec{\alpha} N(x) \vec{1}^T} . \qquad (18)$$

As in the case lacking selection, this will, regardless of $\vec{\alpha}$, have a limiting nucleotide equilibrium probability distribution as $x \to +\infty$. We will denote this equilibrium probability distribution as $\vec{\theta}$ so as to differentiate it from $\vec{\beta}$, which we used in the absence of selection. For a equilibrium joint probability distribution matrix, we write

$$K(x) \quad = \quad \frac{D_{\vec{\theta}} N(x)}{\vec{1} D_{\vec{\theta}} N(x) \vec{1}^T} . \qquad (19)$$

A question of central interest is: what does $K(x)$ look like? In particular, what is the instantaneous rate of substitution that it implies? We will show that, in cases relevant to *cis*-regulatory elements, it holds true that

$$\mathrm{ods}\left(\frac{\partial}{\partial x} K(x) \Big|_{x=0} \right) \approx -\mathrm{trace}\left(D_{\vec{\beta}} R \right) . \qquad (20)$$

That is, the instantaneous rate of substitution differs little from the rate for neutral DNA positions (Equation 15).

A.3 Ancestor-to-Descendant Substitution Rate, in the Presence of Selection

We can take the derivative $\partial / \partial x$ of the matrix $K(x)$ directly, using $N(x)|_{x=0} = I$ and $(\partial N(x)/\partial x)|_{x=0} = Q$:

$$\begin{aligned} &\frac{\partial}{\partial x} K(x) \Big|_{x=0} \\ &= \quad \frac{D_{\vec{\theta}} (\partial N(x)/\partial x)}{\vec{1} D_{\vec{\theta}} N(x) \vec{1}^T} - \frac{(D_{\vec{\theta}} N(x))(\vec{1} D_{\vec{\theta}} (\partial N(x)/\partial x) \vec{1}^T)}{(\vec{1} D_{\vec{\theta}} N(x) \vec{1}^T)^2} \end{aligned}$$

$$
\begin{aligned}
&= \frac{D_{\vec{\theta}}Q}{\vec{1}D_{\vec{\theta}}I\vec{1}^T} - \frac{(D_{\vec{\theta}}I)(\vec{1}D_{\vec{\theta}}Q\vec{1}^T)}{(\vec{1}D_{\vec{\theta}}I\vec{1}^T)^2} \\
&= D_{\vec{\theta}}Q - D_{\vec{\theta}}(\vec{1}D_{\vec{\theta}}Q\vec{1}^T) \\
&= D_{\vec{\theta}}R + D_{\vec{\theta}}S - D_{\vec{\theta}}(\vec{1}D_{\vec{\theta}}Q\vec{1}^T)\,. \qquad (21)
\end{aligned}
$$

Now, the second term of this last expression is the product of two diagonal matrices, $D_{\vec{\theta}}$ and S, and it will have no non-zero off-diagonal elements. Also, the final term of this last expression is the product of a diagonal matrix $D_{\vec{\theta}}$ and a scalar $\vec{1}D_{\vec{\theta}}Q\vec{1}^T$, and it will have no non-zero off-diagonal elements. This means that, when we are calculating the effect of S on the expected rate of substitutions, it suffices that we examine the sum of the off-diagonal elements of the first term, $D_{\vec{\theta}}R$. This is the formula of Equation 14, except that in the present case we use $D_{\vec{\theta}}$, from the selection-influenced, equilibrium probability distribution implied by Q, rather than $D_{\vec{\beta}}$, from the neutral-site equilibrium probability distribution implied by R alone. Note that, despite the switch from $D_{\vec{\beta}}$ to $D_{\vec{\theta}}$, the second factor does not need to be switched from R to Q.

As before, we can exploit the equality $R\vec{1}^T = \vec{0}^T$. The overall instantaneous rate of substitutions simplifies to

$$
\mathrm{ods}\left(\frac{\partial}{\partial x}K(x)\bigg|_{x=0}\right) = -\mathrm{trace}(D_{\vec{\theta}}R)\,, \qquad (22)
$$

a formula that can be evaluated so long as $\vec{\theta}$ and R are known (or estimated), even if the fitness matrix S is not known.

Because we usually have DNA of present-day individuals, with no guarantee that they will have plentiful progeny in the future, the equilibrium joint probability distribution $K(x)$ is not necessarily representative of actual data. However, we can repeat the calculations, looking at the equilibrium joint probability distribution between two present-day descendants of a common ancestor.

A.4 Equilibria Joint Probability Distribution for Descendants in the Presence of Selection

In the case that each of two individuals is an evolutionary distance $x/2$ from their common ancestor, their equilibrium joint probability distribution is

$$
L(x) = \frac{N(x/2)^T D_{\vec{\theta}} N(x/2)}{\vec{1}N(x/2)^T D_{\vec{\theta}} N(x/2)\vec{1}^T}\,. \qquad (23)
$$

Using a derivation along the lines of that which gave us Equation 22, we can conclude that

$$
\begin{aligned}
\mathrm{ods}\left(\frac{\partial}{\partial x}L(x)\bigg|_{x=0}\right) &= \frac{-\mathrm{trace}(R^T D_{\vec{\theta}} + D_{\vec{\theta}}R)}{2} \\
&= -\mathrm{trace}(D_{\vec{\theta}}R)\,. \qquad (24)
\end{aligned}
$$

A.5 Overall Instantaneous Rate of Nucleotide Subsection

We look at a specific example for the neutral-site substitution model, the HKY85 nucleotide substitution model (*9*). In this situation we have that

$$
R_{\vec{\beta}} = \mu_{\vec{\beta}} \begin{pmatrix} -\beta_Y - \kappa\beta_G & \beta_T & \beta_C & \kappa\beta_G \\ \beta_A & -\beta_R - \kappa\beta_C & \kappa\beta_C & \beta_G \\ \beta_A & \kappa\beta_T & -\beta_R - \kappa\beta_T & \beta_G \\ \kappa\beta_A & \beta_T & \beta_C & -\beta_Y - \kappa\beta_A \end{pmatrix},
$$

$$
\begin{aligned}
\beta_R &= \beta_A + \beta_G\,, \\
\beta_Y &= \beta_C + \beta_T\,, \\
\mu_{\vec{\beta}} &= \frac{1}{1 - \vec{\beta}\cdot\vec{\beta} + (\kappa-1)(2\beta_A\beta_G + 2\beta_T\beta_C)}\,, \\
M_{\vec{\beta}}(x) &= \exp(xR_{\vec{\beta}})\,, \text{ and} \\
J_{\vec{\beta}}(x) &= D_{\vec{\beta}}\exp(xR_{\vec{\beta}})\,, \qquad (25)
\end{aligned}
$$

where $\vec{\beta}\cdot\vec{\beta}$ is a simple dot product, and $\kappa > 0$ is the ratio of the transition rate to the transversion rate.

If a given position of a *cis*-regulatory element motif has a nucleotide equilibrium probability distribution of $\vec{\theta}$, then, even though we do not know S, we can compute the implied overall instantaneous rate of substitution in the presence of selection pressures, using Equation 24:

$$
\begin{aligned}
&\mathrm{ods}\left(\frac{\partial}{\partial x}L(x)\bigg|_{x=0}\right) \qquad (26) \\
&= \frac{1 - \vec{\theta}\cdot\vec{\beta} + (\kappa-1)(\theta_A\beta_G + \theta_G\beta_A + \theta_T\beta_C + \theta_C\beta_T)}{1 - \vec{\beta}\cdot\vec{\beta} + (\kappa-1)(2\beta_A\beta_G + 2\beta_T\beta_C)}\,.
\end{aligned}
$$

As an example, when the neutral-DNA equilibrium probability distribution (β_A, β_T, β_C, β_G) equals (30%, 30%, 20%, 20%) and κ equals 3, the overall instantaneous rate of Equation 26 will fall in the interval $[0.901, 1.148]$, regardless of the selection matrix S and the selection-influenced equilibrium probability distribution $\vec{\theta}$.

When the selection-neutral equilibrium probability distribution is described by $\kappa = 1$ and $\vec{\beta}$ uniform, Equation 26 reduces exactly to

$$
\mathrm{ods}\left(\frac{\partial}{\partial x}L(x)\bigg|_{x=0}\right) = 1\,, \qquad (27)
$$

regardless of the values of S and $\vec{\theta}$. These results indicate that for reasonable background models, even with an arbitrary soft selection matrix S and foreground equilibrium probability distribution $\vec{\theta}$, it may be appropriate to approximate with Equation 27.

Because of the similarity of Equations 22 and 24, the results are identical for an ancestor and a descendant. In either case, the overall instantaneous rate of substitution in the presence of selection differs little from the overall neutral rate of substitution.

A.6 Mismatch Fraction as a Function of Distance

It is informative to look at the value of $\mathrm{ods}(L(x))$, the mismatch fraction of nucleotides between two individuals equally distant from their common ancestor, as a function of x, the evolutionary distance between the individuals as measured by the expected number of nucleotide substitutions per DNA position. We compare it to $\mathrm{ods}(J_{\vec{\theta}}(x))$ for $J_{\vec{\theta}}(x) = D_{\vec{\theta}}\exp(xR_{\vec{\theta}})$. $J_{\vec{\theta}}(x)$ is different from $J_{\vec{\beta}}(x)$ (from Equation 25), in that, although it is of the form of a background equilibrium joint probability distribution, it is parameterized by the foreground equilibrium probability distribution $\vec{\theta}$, rather than by the background equilibrium probability distribution $\vec{\beta}$. Additionally, we compare these to $\mathrm{ods}(\mathrm{HB98}_{\vec{\theta}}(x))$, the number of mismatches predicted by the HB98 model.

For illustrative purposes, suppose that we have the simplest of the background nucleotide substitution models, the JC69 model (*12*), in which $\kappa = 1$ and $\vec{\beta}$ is uniform:

$$R_{\vec{\beta}} = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix} . \quad (28)$$

We choose a selection fitness matrix S that favors one nucleotide:

$$\begin{aligned} S &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & Z & 0 & 0 \\ 0 & 0 & Z & 0 \\ 0 & 0 & 0 & Z \end{pmatrix} \quad (29) \\ Z &= -279/28 \approx -10 , \end{aligned}$$

where we have chosen a value of Z such that the following equations do not contain radicals.

An eigenvector (or spectral) decomposition of Q gives us

$$\begin{aligned} Q &= P^{-1}DP \\ &= \begin{pmatrix} 28 & 0 & 0 & 1 \\ 1 & -1 & -1 & -\frac{28}{3} \\ 1 & 1 & 0 & -\frac{28}{3} \\ 1 & 0 & 1 & -\frac{28}{3} \end{pmatrix} \\ &\times \begin{pmatrix} -\frac{27}{28} & 0 & 0 & 0 \\ 0 & -\frac{949}{84} & 0 & 0 \\ 0 & 0 & -\frac{949}{84} & 0 \\ 0 & 0 & 0 & -\frac{31}{3} \end{pmatrix} \\ &\times \begin{pmatrix} \frac{28}{787} & \frac{1}{787} & \frac{1}{787} & \frac{1}{787} \\ 0 & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ 0 & -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{3}{787} & -\frac{28}{787} & -\frac{28}{787} & -\frac{28}{787} \end{pmatrix} . \quad (30) \end{aligned}$$

We use this to determine

$$\begin{aligned} \exp(xQ) &= \sum_{k=0}^{\infty} \frac{x^k}{k!} Q^k = \sum_{k=0}^{\infty} \frac{x^k}{k!} (P^{-1}DP)^k \\ &= \sum_{k=0}^{\infty} \frac{x^k}{k!} P^{-1}D^kP = P^{-1}\left(\sum_{k=0}^{\infty} \frac{x^k}{k!} D^k\right)P \\ &= \begin{pmatrix} 28 & 0 & 0 & 1 \\ 1 & -1 & -1 & -\frac{28}{3} \\ 1 & 1 & 0 & -\frac{28}{3} \\ 1 & 0 & 1 & -\frac{28}{3} \end{pmatrix} \\ &\times \begin{pmatrix} e^{-\frac{27}{28}x} & 0 & 0 & 0 \\ 0 & e^{-\frac{949}{84}x} & 0 & 0 \\ 0 & 0 & e^{-\frac{949}{84}x} & 0 \\ 0 & 0 & 0 & e^{-\frac{31}{3}x} \end{pmatrix} \\ &\times \begin{pmatrix} \frac{28}{787} & \frac{1}{787} & \frac{1}{787} & \frac{1}{787} \\ 0 & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ 0 & -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{3}{787} & -\frac{28}{787} & -\frac{28}{787} & -\frac{28}{787} \end{pmatrix} , \quad (31) \end{aligned}$$

which we use to compute $\vec{\theta}$ and $\mathrm{ods}(L(x))$. Similarly we can compute $\mathrm{ods}(J_{\vec{\theta}}(x))$ and $\mathrm{ods}(\mathrm{HB98}_{\vec{\theta}}(x))$:

$$\begin{aligned} \vec{\theta} &= \frac{1}{31}(28,1,1,1) , \quad (32) \\ \mathrm{ods}\left(J_{\vec{\theta}}(x)\right) &= \frac{174}{961}\left(1 - e^{-\frac{961}{174}x}\right) , \quad (33) \\ \mathrm{ods}\left(\mathrm{HB98}_{\vec{\theta}}(x)\right) &= \frac{174}{961} - \frac{112}{961}e^{-\frac{31}{81}\log(28)x} \\ &\quad -\frac{2}{31}e^{-(1+\frac{28}{81}\log(28))x} , \quad (34) \\ \mathrm{ods}\left(L(x)\right) &= \frac{2}{31}\frac{\begin{pmatrix} +1910085\, e^{-\frac{27}{28}x} \\ -1898316\, e^{-\frac{949}{168}x} \\ +607600\, e^{-\frac{31}{3}x} \\ -619369\, e^{-\frac{949}{84}x} \end{pmatrix}}{\begin{pmatrix} +680605\, e^{-\frac{27}{28}x} \\ -122472\, e^{-\frac{949}{168}x} \\ +61236\, e^{-\frac{31}{3}x} \end{pmatrix}} \quad (35) \end{aligned}$$

The not-very-close relationship of $\mathrm{ods}(L(x))$ and $\mathrm{ods}\left(\mathrm{HB98}_{\vec{\theta}}(x)\right)$ is graphed in Figure 1. The closer relationship of $\mathrm{ods}(L(x))$ and $\mathrm{ods}\left(J_{\vec{\theta}}(x)\right)$ is graphed in Figure 2. These figures are shown and discussed in the main text.

A.7 Likelihood of a Phylogenetic Tree

When a nucleotide fitness matrix is known (or estimated), we can exactly calculate the probability of the observed data for a sequence position, via two invocations of Felsenstein's algorithm. For each tree edge, we use the population model matrix of Equation 17, where x is the length of the tree edge, as if it were a substitution matrix. With these matrices we evaluate the observed data with a first invocation of Felsenstein's algorithm.

This gives us an unnormalized probability of the data, because we have not yet accounted for the fact that the leaf population sizes need to be scaled to 1.0. Although this may be

technically correct only when no element of S is positive, we might say that we have not yet statistically conditioned upon the fact that the leaf individuals have survived until the time of their sequencing.

To scale appropriately, we must divide the unnormalized probability by the sum of the unnormalized probabilities over all 4^s possible sets of observed nucleotides, where s is the number of sequences aligned. Fortunately, this value can be computed in a single use of Felsenstein's algorithm, by using $(1, 1, 1, 1)$ (instead of $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, or $(0, 0, 0, 1)$) as the "observed data" at each leaf.

We are currently attempting this in the "OrthoGibbs" version of our Gibbs Sampling software (*16*).

Notice that this approach, using Felsenstein's algorithm twice, generalizes to work with hard selection pressures, in which some nucleotides are lethal or nearly so. Whatever the appropriate component substitution and fitness matrices are for a generation, we can write them as M and T, respectively. Instead of Equation 17 we have

$$N(x) = (MT)^{x/\epsilon}\,, \tag{36}$$

exactly. This can be computed efficiently for any x, via a spectral decomposition of the matrix product MT, and used in the invocations of Felsenstein's algorithm, as above.

A.8 Soft Selection and *cis*-Regulatory Elements

Sequence positions within *cis*-regulatory elements are often ambiguous. In many cases, a single nucleotide will be strongly preferred at some position, but rarely is any allele excluded. This relates directly to the severity of the selection effects that give rise to the equilibrium probability distributions, and indicates that our assumption that selection effects are soft / non-lethal is reasonable.

Asymptotically, as the value of Z in Equation 29 becomes strongly negative, the resulting equilibrium probability distribution (Equation 32) has one part in $3|Z|$ for each of the nucleotides other than the favored one, with the remainder of the equilibrium probability distribution in the favored nucleotide. Thus, the fact that equilibrium probability distributions for element positions are relatively moderate is directly indicative of the softness of the selection fitnesses.

As an example, an equilibrium probability distribution of $\vec{\theta} = (99.7\%, 0.1\%, 0.1\%, 0.1\%)$ would arise from a value of $Z \approx -333$. Mutation rates have a typical time scale of 100 million years (*i.e.*, approximately 1% of nucleotides are mutated during each million years), thus a value of $Z = -333$ indicates that the timescale for selection effects is approximately 0.3 million years. This is much longer than the timescale for a generation, and it justifies the assumption that selection effects are soft.

Acknowledgments

The author is supported by The Office of The Provost of The Rensselaer Polytechnic Institute and by The Department of Energy grant DE-FG02-01ER63204.

References

1. Yang Z, Roberts D (1995) On the Use of Nucleic Acid Sequences to Infer Early Branchings in the Tree of Life. Mol Biol Evol 12: 451–458, PubMed 7739387.

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

3. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol 5: R98, PubMed 15575972.

4. Siepel A, Haussler D (2004) Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood. Mol Biol Evol 21: 468–488, PubMed 14660683.

5. Fisher R (1930) The Genetical Theory of Natural Selection. Clarendon.

6. Wright S (1931) Evolution in Mendelian populations. Genetics 16: 97–159.

7. Haldane JBS (1932) The Causes of Evolution. Longmans, Green & Co.

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

9. Hasegawa M, Kishino H, Yano T (1985) Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA. J Mol Evol 22: 160–174, PubMed 3934395.

10. Tajima F, Nei M (1982) Biases of the Estimates of DNA Divergence Obtained by the Restriction Enzyme Technique. J Mol Evol 18: 115–120, PubMed 6284946.

11. Lanave C, Preparata G, Saccone C, Serio G (1984) A New Method for Calculating Evolutionary Substitution Rates. J Mol Evol 20: 86–93, PubMed 6429346.

12. Jukes TH, Cantor C (1969) Evolution of Protein Molecules. In: Munro HM, editor, Mammalian Protein Metabolism, vol. 3, pp. 21–132, New York, NY: Academic Press.

13. Rodríguez F, Oliver JL, Marín A, Medina JR (1990) The General Stochastic Model of Nucleotide Substitution. J Theor Biol 142: 485–501, PubMed 2338834.

14. Kimura M (1962) On the Probability of Fixation of Mutant Genes in a Population. Genetics 4: 713–719, PubMed 14456043.

15. Golding B, Felsenstein J (1990) A Maximum Likelihood Approach to the Detection of Selection from a Phylogeny. J Mol Evol 31: 511–523, PubMed 2176699.

16. Thompson W, Rouchka EC, Lawrence CE (2003) Gibbs Recursive Sampler: Finding Transcription Factor Binding Sites. Nucleic Acids Res 31: 3580–3585, PubMed 12824370.