Faster than Neutral Evolution of Constrained Sequences: The Complex Interplay of Mutational Biases and Weak Selection

David S. Lawrie[1], Dmitri A. Petrov[2], and Philipp W. Messer*[,2]

[1]Department of Genetics, Stanford University

[2]Department of Biology, Stanford University

*Corresponding author: E-mail: messer@stanford.edu.

Accepted: 30 March 2011

Abstract

Comparative genomics has become widely accepted as the major framework for the ascertainment of functionally important regions in genomes. The underlying paradigm of this approach is that most of the functional regions are assumed to be under selective constraint, which in turn reduces the rate of evolution relative to neutrality. This assumption allows detection of functional regions through sequence conservation. However, constraint does not always lead to sequence conservation. When purifying selection is weak and mutation is biased, constrained regions can even evolve faster than neutral sequences and thus can appear to be under positive selection. Moreover, conservation estimates depend also on the orientation of selection relative to mutational biases and can vary over time. In the light of recent data of the ubiquity of mutational biases and weak selective forces, these effects should reduce the power of conservation analyses to define functional regions using comparative genomics data. We argue that the estimation of true mutational biases and the use of explicit evolutionary models are essential to improve methods inferring the action of natural selection and functionality in genome sequences.

Key words: conservation, weak constraint, mutational biases, mutation-selection model.

Introduction

Identifying functionally important regions of genomes is a key challenge in evolutionary biology. Fueled by the availability of whole-genome sequence data for a constantly growing number of species, comparative genomics has emerged as the standard framework for the identification of functional regions (Hardison 2003; Pheasant and Mattick 2007).

The approach taken by comparative genomics is based on the assumption that mutations in functional regions would often be deleterious and thus filtered out by purifying selection, reducing the rate of evolution in functional regions relative to nonfunctional neutrally evolving regions. This signature is also commonly referred to as sequence conservation, which in the comparative genomics context is detected as regions of reduced divergence compared with neutrally evolving regions in sequence alignments. The more critical a functional region is, the greater the purifying selection to maintain it and the greater the signature of sequence conservation we expect to see. Most current comparative genomics approaches to identify and classify functional regions are built on this paradigm (Waterston et al. 2002; Cooper et al. 2005; Siepel et al. 2005; Margulies et al. 2007; Pheasant and Mattick 2007; Eory et al. 2010; Goode et al. 2010; Pollard et al. 2010).

The notion that functionality entails sequence conservation is rooted in Kimura's influential concept of the "neutral theory of molecular evolution" (Kimura 1983). Neutral theory surmises that positive selection is so infrequent that its contribution to the rate of evolution is negligible. The majority of sites in a genome are assumed to evolve neutrally, whereas some fraction, f_c, of functional sites are under strong selective constraint. If we define the rate of evolution, r, in terms of the rate at which new mutations fix in the population, then in functional regions $r = (1 - f_c)\, r_0$, where r_0 is the rate of evolution in the neutrally evolving regions. The common conception that increasing constraint can only decrease the rate of evolution, $r/r_0 < 1$, emerges as an immediate consequence.

An accepted shortcoming of this approach is that some or even many functional elements are affected by nonequilibrium processes such as positive selection, nonstationary mutation rates, and roving hot spots of biased gene conversion (BGC) (Pollard et al. 2006a; Galtier and Duret 2007; Pheasant and Mattick 2007; Berglund et al. 2009; Duret 2009). However, it is still a wide-held assumption that, in an equilibrium scenario, constraint necessarily entails sequence conservation. This notion is not justified—it has been shown more than a decade ago that the interplay of mutational biases and weak constraint can be quite complex; the rate of evolution at constrained sites can even be higher than that at neutral sites (McVean and Charlesworth 1999).

A bulk of recent evidence points to the ubiquity of the necessary ingredients—mutational biases and weak selective forces—for this effect to occur. Weak purifying selection appears to be acting in substantial parts of the genomes of many species (Bustamante et al. 2002, 2005; Ohta 2002; Lu and Wu 2005; Comeron 2006; Lipatov et al. 2006; Eyre-Walker and Keightley 2007; Eory et al. 2010). Furthermore, the process of BGC, which introduces fixation biases in a way similar to that of weak purifying selection (Nagylaki 1983), operates in most eukaryotic genomes and is the best candidate to explain observed genome-wide systematic biases in fixation probabilities (Galtier and Duret 2007; Duret and Arndt 2008; Duret and Galtier 2009).

Mutational biases have also been observed in many organisms, and weak selection-like forces often seem to be acting in opposition to the mutational biases, as indicated by the observation that genomic nucleotide contents are typically less biased than would be expected from the underlying mutational biases (Lynch et al. 2008; Hershberg and Petrov 2010; Hildebrand et al. 2010; Ossowski et al. 2010). The complexity of how sequences evolve under realistic scenarios of weak constraint and mutational biases has important ramifications for conservation analysis and its evolutionary interpretation.

Using a generalized Markov process to emulate the mutation-selection dynamics governing a genomic site, we investigate how complex interactions between weak selection and mutational biases affect different measurements of conservation under different scenarios. We first recapitulated the results of McVean and Charlesworth (1999) that weak constraint can accelerate the rate of evolution over that at neutral sites.

We demonstrate that this effect complicates the inference of constraint in the maximum-likelihood (ML) branch-length analysis that underlies many comparative genomics approaches such as GERP (Cooper et al. 2005) and PhyloP (Pollard et al. 2010). As a practical example, we investigate how GERP conservation scores are affected over a realistic (mammalian) species tree. We find that constrained sequence regions do not always show the signatures of sequence conservation. Furthermore, for the same strength of weak constraint, the measurement of conservation will typically vary with branch length and depend on the orientation of selection, that is, which particular bases are the preferred states in relation to the mutational biases. The power of ML methods to detect functional regions may thus be substantially reduced in regions of weak constraint. We discuss how inference methods that disentangle mutation and selection can improve such analyses.

Materials and Methods

Markov Models of Sequence Evolution

The evolution of DNA sequences can be modeled as a Markov process specified by a substitution rate matrix $\mathbf{R}$ (Lio and Goldman 1998). Its elements, $\mathbf{R}_{ij}$, denote the rates at which a nucleotide i is substituted by a nucleotide j; diagonal elements are $\mathbf{R}_{ii} = -\sum_{j\neq i}\mathbf{R}_{ij}$, the total rate away from i.

If all sites in a sequence are assumed to evolve according to the same model and independently of each other, we can specify the equilibrium nucleotide composition, $\rho = \{\rho_A,\rho_C,\rho_G,\rho_T\}$, as the eigenvector corresponding to the largest eigenvalue of $\mathbf{R}$, which is guaranteed to be zero:

$$\rho\mathbf{R} = 0. \quad (1)$$

The stationary rate of evolution, that is, the average rate of substitution per site in equilibrium, is then:

$$r = \sum_i \rho_i \sum_{j\neq i} \mathbf{R}_{ij}. \quad (2)$$

The probability $\mathbf{P}_{ij}(t)$ of observing nucleotide j after time t if the ancestral state at that site was i can be calculated from the matrix exponential:

$$\mathbf{P}(t,\mathbf{R}) = e^{\mathbf{R}t}. \quad (3)$$

Notice that $\mathbf{P}(t,\mathbf{R})$ allows for the possibility of any number of substitutions having occurred during the time interval t. Diagonal elements $\mathbf{P}_{ii}(t,\mathbf{R})$ specify the probabilities of observing no change. When two sequences separated by time t are compared, their expected observed divergence $d(t)$, that is, the fraction of sites at which the sequences differ, can be calculated as the total probability of observing different base pairs at homologous sites:

$$d(t) = \sum_i \rho_i[1-P_{ij}(t,\mathbf{R})]. \quad (4)$$

To disentangle the processes of mutation and selection, we first define a mutation rate matrix $\mathbf{Q}$ with its elements μ_{ij} specifying the rates at which new mutations i to j occur in individuals. Substitution rates can then be decomposed into the product of mutation rates μ_{ij}, the effective (haploid)

population size N and the probability of such mutations eventually fixing in the population, ν_{ij}:

$$\mathbf{R}_{ij} = \mu_{ij} \times N \times \nu_{ij}. \quad (5)$$

In this framework, specific biases in the raw rates at which new mutations occur can be incorporated into the mutation rates μ_{ij}. The fixation probabilities ν_{ij} can account for selective advantages or disadvantages of a nucleotide j over a different nucleotide i. If we specify with $s(i,j)$ the selection coefficient of nucleotide j relative to i, the fixation probabilities are approximately $\nu_{ij} = 2s(i,j)/[1-e^{-\gamma(i,j)}]$ (Kimura 1983). Note that $\mathbf{R}_{ij}$ then only depends on μ_{ij} and the amount of effective selection, $\gamma(i,j) = 2Ns(i,j)$. The particular scenario of purifying selection at a given site is specified in terms of the $\gamma(i,j)$ for all $i,j \in \{A,C,G,T\}$. For example, if nucleotide a is the preferred nucleotide at that site and it is preferred at an equal strength γ_{func} over all other nucleotides, whereas mutations between each two unpreferred nucleotides are selectively neutral, then: $\gamma(i,a) = \gamma_{\text{func}} = -\gamma(a,i)$ for all $i \neq a$ and $\gamma(i,j) = 0$ for all $i,j \neq a$.

We model BGC analogously to purifying selection preferring C/G alleles over A/T alleles at specific strength γ_{BGC}. Mutations C $\leftrightarrow$ G and A $\leftrightarrow$ T are not affected. In the scenarios where both BGC and purifying selection are acting, the resulting effective selection is the sum: $\gamma(i,j) = \gamma_{\text{BGC}}(i,j) + \gamma_{\text{func}}(i,j)$.

Under neutrality and in the absence of BGC ($\gamma = 0$), we have $\nu_{ij} = 1/N$ and thus $\mathbf{R} = \mathbf{Q}$. The mutational equilibrium composition, π, is then determined by $\pi\mathbf{Q} = 0$. The total rate of evolution yields $r = \sum_i \pi_i \sum_{j \neq i} \mu_{ij}$, and the expected divergence between two neutrally evolving sequences separated by time t is given by $d(t) = \sum_i \pi_i [1-\mathbf{P}_{ii}(t,\mathbf{Q})]$.

We always assume reverse complement symmetry ($\mu_{AT} = \mu_{TA}$ and $\mu_{CG} = \mu_{GC}$). Under this assumption, it is convenient to describe the raw mutational bias in terms of the equilibrium A/T content in the absence of BGC and selection, $\pi_{A+T} = \pi_A + \pi_T$. Note that detailed balance yields $\pi_{A+T} = (\mu_{C/G \to A/T})/(\mu_{A/T \to C/G} + \mu_{C/G \to A/T})$. A mutational bias of $\pi_{A+T} = 0.8$, for example, implies that mutations from C or G to A or T occur four times more often than the reverse process.

The scale in which time is measured can be chosen freely in the above framework. By convention, $\mathbf{R}$ is normalized such that time is measured in units of expected number of substitutions at neutrally evolving sites, that is, it is determined by the condition $\sum_i \pi_i \sum_{j \neq i} \mu_{ij} = 1.2$

ML Branch-Length Inference

Let $\mathbf{D}$ be the observed pair-count matrix in a two-sequence alignment of length n, that is, elements $\mathbf{D}_{ij}$ denote at how many of the n positions in the alignment nucleotide i is observed in the first sequence while the second sequence has nucleotide j. The total observed divergence between two sequences, d, is the sum of the nondiagonal elements of $\mathbf{D}$. Equation (3) specifies a probabilistic model for $\mathbf{D}$ given a rate matrix $\mathbf{R}$ and a divergence time t. For an empirical observation $\mathbf{D}$ and assuming an inference model $\mathbf{M}$, we can then define a likelihood function for the divergence time t assuming that the sequences have evolved under $\mathbf{M}$ and that nucleotide content is in equilibrium:

$$L(t) = \Pr[\mathbf{D}|\mathbf{M}, t] \quad (6)$$

This is the multinomial probability to observe, in n trials, counts $\mathbf{D}_{ij}$ under the normalization condition $\sum_{ij}\mathbf{D}_{ij} = n$ and given individual probabilities $p_{ij}(t,\mathbf{M}) = \rho_i \mathbf{P}_{ij}(t,\mathbf{M})$ per trial. Therefore:

$$\log L(t) = \text{const} + \sum_{ij} \mathbf{D}_{ij} \log p_{ij}(t,\mathbf{M}). \quad (7)$$

For an inference model $\mathbf{M}$ and an observed count matrix $\mathbf{D}$, the ML estimate t^* is obtained by maximizing $\log L(t)$ with respect to t. The constant does not depend on t and can be omitted from the maximization. Note that we implicitly assumed a reversible mutation model here. The above framework can be extended to include nonreversible models by using the definition $p_{ij}(t,\mathbf{M}) = \sum_k \rho_k \mathbf{P}_{ki}(t/2,\mathbf{M})\mathbf{P}_{kj}(t/2,\mathbf{M})$.

To ascertain whether a test region is evolutionarily conserved, one typically compares the ML branch-length estimate t^* of this test region with an estimate t_0^* inferred from a neutrally evolving reference region. Conservation then corresponds to $t^*/t_0^* < 1$, that is, a reduction in the branch-length inferred from the test region compared with the reference region. Note that one has to specify the particular inference model $\mathbf{M}$ used in equation (7).

In figures 2 and 3, we investigate how the choice of $\mathbf{M}$ affects conservation estimates c when the true substitution model for the test sequence is $\mathbf{R}$ and the true substitution model for the reference sequence is $\mathbf{R}_0$. To obtain the expected ML branch-length estimates for given $\mathbf{R}$ and $\mathbf{R}_0$, we first calculate the expected count matrices $\langle\mathbf{D}\rangle$ and $\langle\mathbf{D}_0\rangle$ under these models:

$$\langle\mathbf{D}\rangle_{ij} = n \times p_{ij}(t,\mathbf{R}) \quad \text{and} \quad \langle\mathbf{D}_0\rangle_{ij} = n \times p_{ij}(t,\mathbf{R}_0). \quad (8)$$

For the scenario of figure 3 where purifying selection in the test sequence is randomly oriented, there are four different $\mathbf{R}$ models, one for each of the four possible preferred states. Because all bases have equal probability of being the preferred base at a given site, all four $\mathbf{R}$ models contribute equally to the count matrix $\mathbf{D}$.

Given $\langle\mathbf{D}\rangle$ and $\langle\mathbf{D}_0\rangle$, ML estimates for t^* and t_0^* are then calculated from equation (7) using the specific inference model $\mathbf{M}$. Note that these calculations do not depend on the number of sites in the alignment; for convenience, we chose $n = 1$. Likelihoods were maximized numerically in Java.

Running GERP on Simulated Alignments Over a Tree

We simulate the evolution of a site over a species tree using a Markov model of nucleotide substitution specified by a rate matrix **R**. For each site, we first draw its state at the root of the tree from the equilibrium frequencies ρ of the rate matrix and then move down the tree drawing the states at each node. The probability of being in state *i* at node *k* where the ancestral node $k-1$ was assigned base *j* is given by equation (3), with the time *t* specifying the true neutral branch length between the two nodes in the tree. The states at the leaves of the tree then specify the multiple alignment column at that site.

We used the tree from the 44-way MULTIZ alignments from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/phastCons44way/placentalMammals.mod), restricted to its 32 eutherian mammalian members. The total number of expected neutral substitutions per site for the restricted tree is 4.74. The mutation model is an HKY85 model with mutational bias $\pi_{A+T} = 0.8$ and a transition/transversion ratio of 4.0.

In our evolutionary scenario, every site in the test sequence is modeled to be under purifying selection of strength γ_{func} with the preferred base being randomly chosen. BGC is acting uniformly over the sequence on top of purifying selection, favoring C/G over A/T alleles at strength γ_{BGC}. Hence, there are again four different substitution models, one for each preferred nucleotide. Multiple sequence alignments were generated by concatenating the resulting alignment columns of 10^5 individually simulated sites. The simulated sequence alignments together with the correct neutral tree were fed to GERP 2.1 (http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html), using its default parameters.

Results

Faster than Neutral Evolution of Constrained Sequences

The basic result that sequences evolving under weak constraint and biased mutation can evolve faster than neutral sequences was first observed by McVean and Charlesworth (1999) in their analysis of the rate of evolution at synonymous positions in the presence of mutational biases and selection on codon usage.

In figure 1, we recapitulate their result in a general equilibrium scenario where mutations are biased toward A/T, that is, the equilibrium A/T content in the absence of selection, π_{A+T}, is larger than 0.5, whereas weak purifying selection favors C/G alleles over A/T alleles at an effective strength $\gamma = 2Ns$ (the selection coefficient *s* rescaled by twice the haploid population size *N*). Mutations C ↔ G and A ↔ T are selectively neutral. Figure 1*A* shows the ratio of the average rate of substitution per site in equilibrium at constrained sites, *r*, and at neutral sites, r_0, in this scenario (Materials and Methods). When purifying selection is weak, the rate of evolution at constrained sites is indeed higher than that at the neutral sites ($r/r_0 > 1$). This acceleration becomes stronger as the mutational bias toward A/T becomes larger. For instance, when $\pi_{A+T} = 0.8$, the maximal rate of evolution at constrained sites is ~8% higher than that at the neutral sites.

The particular value of the effective strength of selection at which the rate of evolution is at its maximum increases as mutations become more biased. When the mutational bias is $\pi_{A+T} = 0.7$, for example, the maximal rate is observed when $\gamma \sim -0.6$, whereas for $\pi_{A+T} = 0.8$, the maximum is at $\gamma \sim -1$ (fig. 1*A*). As expected, when purifying selection becomes strong enough, the rate of evolution decreases again below the neutral rate, and for sufficiently strong purifying selection (e.g., $\gamma < -1.5$ for $\pi_{A+T} = 0.8$), one always observes conservation ($r/r_0 < 1$).

These results provide a clear illustration that even in fully equilibrium scenarios of constraint where selectively favored states do not change in time and all rates are stationary, constraint does not necessarily result in conservation. Intuitively, this rate acceleration can be explained by the weak purifying selection effectively lowering biases in the equilibrium nucleotide frequencies. This can be seen in figure 1*B* where the expected equilibrium A/T content at the constrained sites, ρ_{A+T}, is plotted for the strength of purifying selection yielding the maximal rate acceleration for a given mutational bias π_{A+T}. Common mutations drive substitutions away from the fitter states despite purifying selection, whereas selection favors fixation of uncommon mutations resulting in faster back substitutions to the fitter states. This allows for greater overall flux between states and thus a higher rate of substitution at the constrained sites compared with the neutrally evolving sites.

Rates of evolution are typically not measured directly. Instead, one observes sequence divergence. Figure 1*C* shows that the average expected divergence at constrained sites, *d*, can also be systematically higher than that at the neutral sites, d_0, and that this effect increases with divergence time. Similarly to figure 1*A*, the effect is again more pronounced when mutations are strongly biased away from the selectively preferred states.

The paradigm that constraint necessarily entails reduction in the rate of evolution does not hold in this scenario. In fact, given that the expected divergence at the constrained sites can be higher than that at the neutral sites, such sites might even be inferred to evolve under positive selection.

Effects on ML Branch-Length Inference

Today's toolbox of comparative genomics has expanded greatly from simply using the observed divergence in

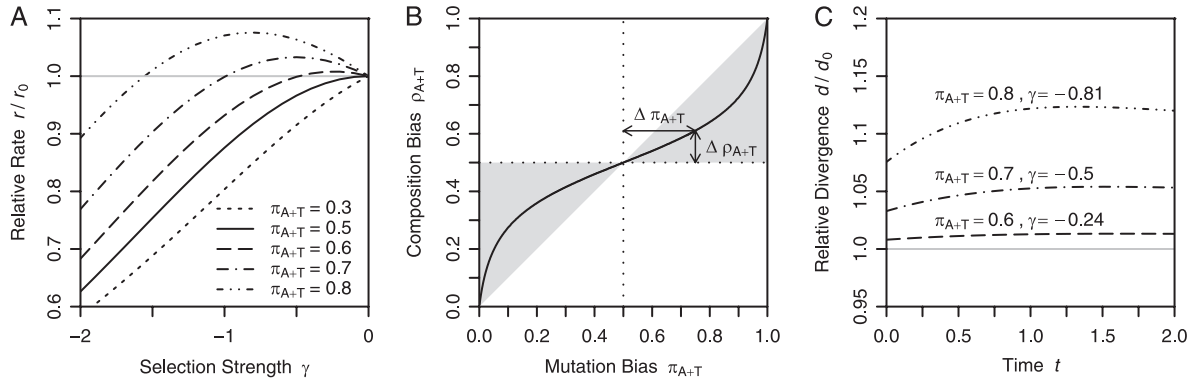

Fig. 1.—Accelerated rate of evolution when selection is counteracting a mutational bias. In the shown scenario, purifying selection favors C/G bases over A/T bases. Mutations occur according to a standard HKY85 model specified by the mutational bias, π_{A+T} and a transition/transversion ratio of four. Mutations G ↔ C and A ↔ T are neutral and mutational biases are symmetric ($\pi_A = \pi_T$, $\pi_C = \pi_G$). (*A*) Ratio of equilibrium substitution rate, *r*, over the equilibrium neutral rate, r_0, calculated according to equation (2). Values of $r/r_0 < 1$ indicate sequence conservation; values $r/r_0 > 1$ indicate rate acceleration. Rate acceleration is observed if weak purifying selection counteracts mutationally preferred states. (*B*) Comparison of mutational bias and actual composition bias for the selection coefficients that yield maximal rate acceleration for the respective mutational bias. Weak purifying selection counteracting the mutational biases effectively lowers the resulting composition bias compared with that expected if no selection were acting. Gray areas denote the respective regions where $|\Delta\rho_{A+T}| < |\Delta\pi_{A+T}|$. (*C*) Increase of divergence, *d*, over neutral divergence, d_0, for the selection coefficients that yield maximal rate acceleration for the respective mutational biases. Time is measured in units of the expected number of neutral substitutions per site.

pairwise alignments for the inference of conservation. Modern approaches attempt to infer from a sequence alignment the true number of substitutions that have occurred since the divergence of the sequences (Felsenstein 2004). This can be achieved in a full ML framework by assuming a probabilistic model of the substitution process and correcting for "multiple hits" at the same site. The inferred total count of substitutions reflects the evolutionary time (branch-length) separating the two sequences. For conservation analysis, the ML branch-length estimates, t^*, of a test sequence region is compared with that of a presumably neutrally evolving reference sequence region, t_0^*. The ratio of the two branch-lengths, t^*/t_0^*, directly relates to the ratio r/r_0 of the substitution rates in the test region and in the reference region. Sequence conservation hence corresponds to $t^*/t_0^*<1$, whereas $t^*/t_0^*>1$ indicates rate acceleration.

ML branch-length inference requires the underlying neutral substitution model to be specified (Materials and Methods). Because the "true" neutral substitution model is generally unknown, one typically makes simplifying assumptions. In the simplest, the Jukes and Cantor (1969) (JC69) model, substitutions between different nucleotides are assumed to occur all at the same rate. The HKY85 model (Hasegawa et al. 1985) additionally allows for different equilibrium nucleotide frequencies and transition/transversion ratios.

The equilibrium nucleotide frequencies of the HKY85 model can be chosen in several ways: The HKY85:π model uses the nucleotide content inferred from a strictly neutrally evolving reference sequence for both the test and the reference sequence. In the HKY85:ρ model, nucleotide frequencies are the actual nucleotide contents of the two sequence regions and can thus differ between the test and the reference region.

In figure 2, we consider the performance of ML branch-length inference for the three inference models JC69, HKY85:π, and HKY85:ρ in an evolutionary scenario where mutation is uniformly biased in favor of A/T ($\pi_{AT} = 0.8$) in both the test and the reference region (Materials and Methods). The reference region is evolving neutrally, whereas in the test region, weak purifying selection is favoring C/G alleles over A/T alleles (fig. 2*A* and *C*) or A/T alleles over C/G alleles (fig. 2*B* and *D*). Mutations C ↔ G and A ↔ T are selectively neutral in all four cases.

Note that in this scenario, the HKY85:π inference model uses the true mutational bias $\pi_{C+G} = 0.2$ and $\pi_{A+T} = 0.8$ in both the reference and test sequence. The HKY85:ρ inference model uses π for the reference sequence, but for the test sequence, it assumes the resulting equilibrium frequencies from mutation and purifying selection, ρ.

When selection acts in favor of C/G with $\gamma = -1$ (fig. 2*A*), the JC69 and HKY85:ρ models both infer rate acceleration ($t^*/t_0^*>1$), whereas the HKY85:π model infers conservation ($t^*/t_0^*<1$). The inferred branch-length extensions under JC69 and HKY85:ρ are time dependent and become larger with the time of divergence. Both, branch-length extension and time dependence, are more pronounced under the JC69 model than under the HKY85:ρ model.

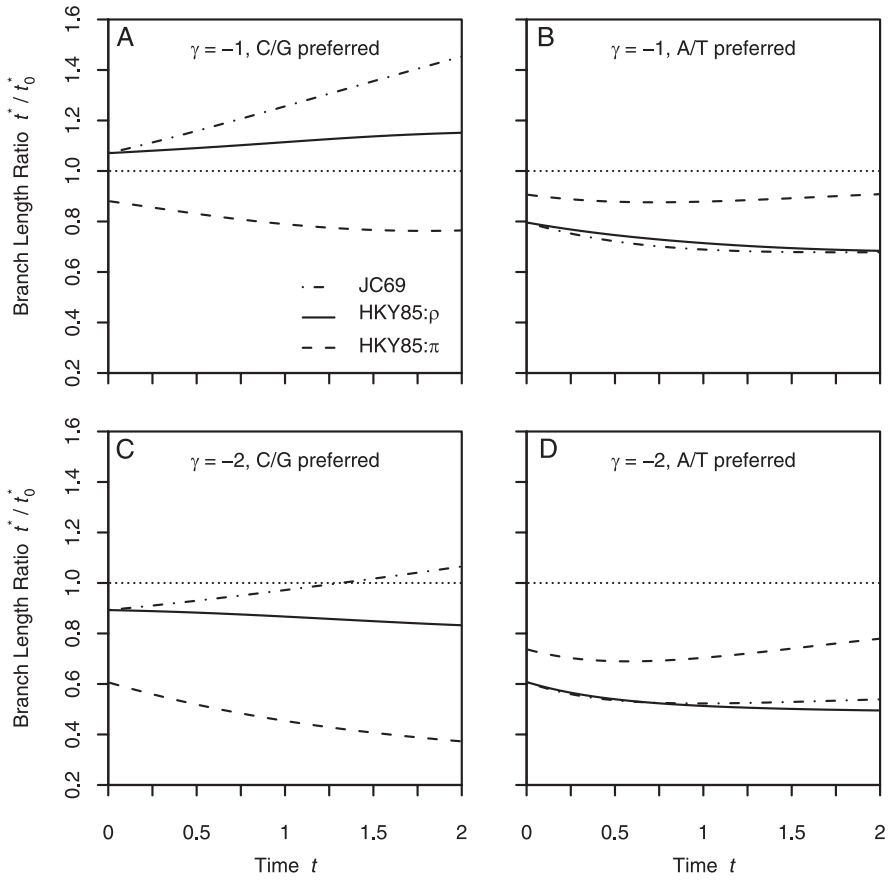

Fig. 2.—Performance of ML branch-length estimation in the presence of mutational biases and weak selective constraint using different neutral inference models. In all four evolutionary scenarios, the true mutation model is an HKY85 model specified by an equilibrium A/T content $\pi_{A+T} = 0.8$ at neutrally evolving sites and a transition/transversion ratio of four. In (*A*) and (*C*), C/G bases are preferred over A/T bases at constrained sites. Purifying selection is thus acting in opposition to the mutationally preferred states. In (*B*) and (*D*), A/T bases are preferred over C/G bases; purifying selection is acting in unison with the mutational biases. ML branch-length estimates at constrained sites, t^*, and at neutral sites, t_0^*, were calculated using three different neutral inference models: the Jukes-Cantor (JC69) model, the HKY85 model with its mutational biases estimated from the data (HKY85:ρ) and the HKY85 model using the true neutral mutation parameters (HKY85:π). Each used the default value of four for the transition/transversion ratio. Inferred branch-length ratios, t^*/t_0^*, are shown as a function of true divergence time t measured in units of the average number of substitutions per neutral site. Values $t^*/t_0^* < 1$ indicate sequence conservation, whereas $t^*/t_0^* > 1$ indicates faster than neutral evolution.

At first glance, it might seem surprising that the HKY85:π model "correctly" infers conservation in figure 2*A*. Given that it uses the correct neutral substitution model one might expect that, similar to figure 1*A*, constrained sites would be inferred to evolve faster than neutral sites. The HKY85:π model is indeed the true substitution model for neutral sites and will therefore infer the correct ML branch-length estimate t_0^* at those sites. At constrained sites, however, it assumes the wrong equilibrium frequencies ($\pi \neq \rho$). In the test region, the HKY85:π model will overestimate the equilibrium A/T frequency ($\pi_{A+T} > \rho_{A+T}$) and the overall rate at which substitutions from C/G to A/T should occur. As a result, in the fashion of "two wrongs making it right," it infers branch-length reduction even though the rate of evolution is in fact higher at these sites than at the neutral sites. The HKY85:ρ model in contrast can be compared directly with the scenario of figure 1*A* because it uses the correct equilibrium frequencies in both the neutral and constrained cases. Consistent with the rate acceleration in figure 1*A*, the HKY85:ρ model infers branch-length extension in figure 2*A*.

One of the consequences of these patterns is that purifying selection of the same strength but operating in different orientations relative to the mutational bias leads to different estimations of conservation. For instance, when comparing the results of figure 2*A* with figure 2*B*, it can be seen that changing the favored nucleotide pair from C/G to A/T (while maintaining the same mutational bias toward A/T) markedly alters the level of inferred conservation, whereas the actual strength of purifying selection is the same in the two scenarios.

The conservation as measured by the HKY85:ρ model is lowered by 20% to 45%, whereas there is an increase up to 10% in the conservation as measured by the HKY85:π model.

In figure 2*C* and *D*, the strength of selection is increased to $\gamma = -2$. The discrepancies in the inferred strength of conservation between scenarios with different selection orientation but same selection strength do not vanish at this higher selection coefficient. This can be understood from figure 1*A* when comparing the curves for $\pi_{A+T} = 0.3$ and $\pi_{A+T} = 0.7$ (changing the mutational bias from π_{A+T} to $1 - \pi_{A+T}$ is comparable to changing the orientation of selection while keeping the mutational bias constant). The shift in r/r_0 between the two curves is almost the same for $\gamma = -1$ and -2.

Regardless of whether branch-length reduction or extension is observed, estimates of branch-length ratios are generally time dependent in figure 2. In the extreme, this can lead to situations where conservation will be estimated for short divergence times, whereas at larger divergence times, constrained sequences will appear to have evolved faster than neutral, as is the case in figure 2*C* for the JC69 model. These effects are a function of the mutational bias and become weaker as the mutational bias is reduced (supplementary fig. S1, Supplementary Material online). The time dependence is a consequence of the fact that selection does not uniformly decrease all individual substitution rates. Substitutions C ↔ G and A ↔ T, for example, are still effectively neutral in the constrained sequence. This cannot be accounted for by a constant, scalar reduction in branch lengths. The resulting error in the multiple-hits correction is compounded as time increases and thus, in principle, should be less pronounced for more closely related species.

Estimating Conservation in the Presence of BGC and Randomly Oriented Selection

In the above analysis, we assumed that all sites in the reference sequence are truly neutral and, in the case of the HKY85:ρ model, that we know the correct allele equilibrium frequencies at every site. Neither assumption is likely to hold in practice. Selection and selection-like forces, especially BGC, can operate in supposedly "neutral" regions. Estimating a true neutral model from any sequence would thus prove difficult without a priori knowledge of the strength of this effect. The underlying assumption of the HKY85:ρ model, that we actually know the correct equilibrium frequencies ρ at every site of the test sequence, is also unrealistic. Over a functional locus, preferred bases of selection will likely be in some jumble of orientations and different selective forces might be operating on the region and even on the same site.

To investigate how such complications affect ML branch-length inference, we created different reference and test models that incorporate BGC as well as heterogeneous models of selection in the test sequence. The mutational bias is again assumed to be $\pi_{A+T} = 0.8$ in the reference and test sequence. BGC is modeled as a selective preference for C/G alleles over A/T alleles, whereas changes G ↔ C and A ↔ T are not affected, similar to the selection scenario setup in figures 1 and 2. We assume BGC to operate in both the reference sequence and the test sequence with same strength γ_{BGC}. We investigate three scenarios: $\gamma_{BGC} = \{0, -1, -2\}$. In the test sequence, functional purifying selection is acting on top of BGC, whereas in the reference sequence, BGC is acting alone. We assume that at every site in the test sequence one base is selectively preferred over the three other bases. Which base is the preferred nucleotide at a given site is randomly chosen with all four states {A,C,G,T} having equal probability. Functional selection operates at all sites in the test sequence at the same strength, γ_{func}.

ML branch-length analysis is performed for the HKY85:ρ and HKY85:π inference models. As before, the HKY85:π model uses the true neutral substitution biases, π, without BGC or selection in both test and reference region. The HKY85:ρ model, in contrast, uses the expected nucleotide frequencies, ρ, of the sequences. In the reference sequence, ρ is the equilibrium nucleotide content resulting from mutational biases and BGC. In the test sequences, ρ incorporates mutational bias, BGC, and functional selection averaged over the test sequence.

Figure 3 shows the resulting conservation estimates. In the top two panels, the strength of functional selection is $\gamma_{func} = -1$, whereas in the bottom two panels $\gamma_{func} = -2$. The left panels show conservation estimates for the subset of sites where the preferred state is C and the right panels show the results for sites where A is preferred. Due to the symmetry in the specification of the mutation-selection models, conservation estimates are equivalent across A and T sites as well as across C and G sites.

Without BGC ($\gamma_{BGC} = 0$), the HKY85:π and the HKY85:ρ models behave similarly. Much like the HKY85:π model from figure 2, they always infer conservation. The reason for this is that weak selection that is randomly oriented has little effect on the overall content bias ($\rho_{A+T} = 0.78$ in the test sequence for that scenario), and thus $\rho \approx \pi$. As before, ML branch-length inference still suffers from asymmetries and time dependence.

At the sites where C is the preferred state by selection (fig. 3*A* and *C*), increasing the strength of BGC from $\gamma_{BGC} = -1$ to $\gamma_{BGC} = -2$ leads to more conservation. In this scenario, BGC and purifying selection are operating in the same direction, which can be interpreted as effective selection for C of strength γ_{BGC} in the reference sequence and of strength $\gamma_{func} + \gamma_{BGC}$ in the test sequence. Although the difference

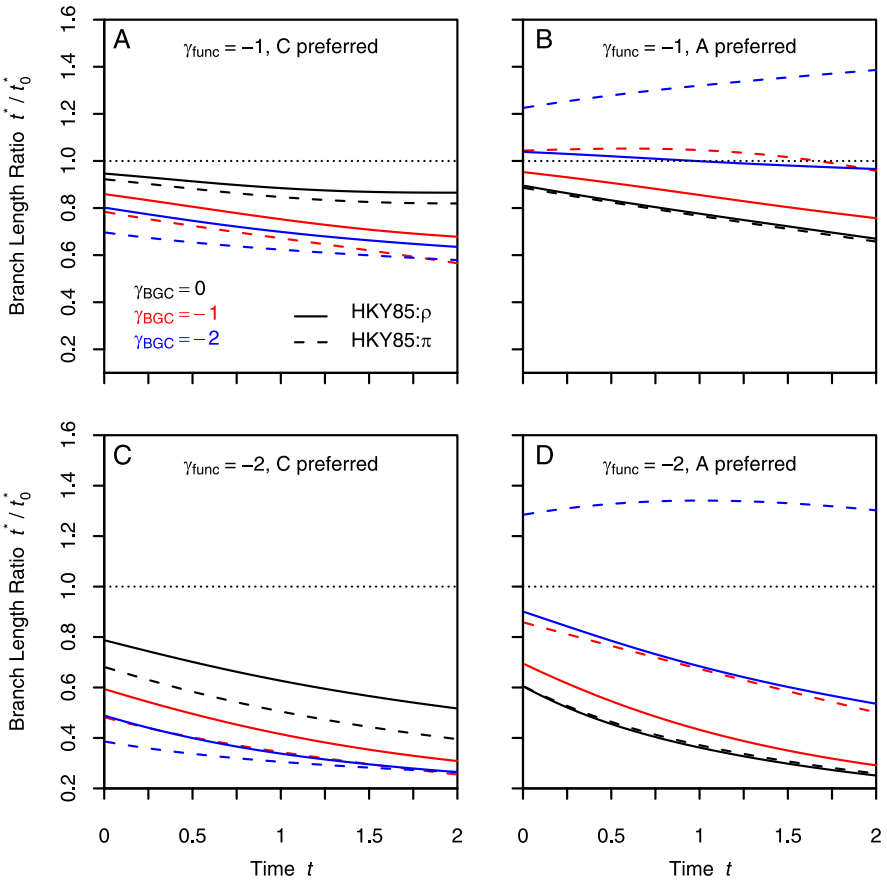

Fig. 3.—Performance of ML branch-length estimation in the presence of mutational biases, randomly oriented weak constraint, and BGC under the HKY85:ρ and the HKY85:π inference models. The mutational bias is always $\pi_{A+T} = 0.8$ with a transition/transversion ratio of four. (*A*) and (*C*) show the results where C is the preferred base, whereas (*B*) and (*D*) show the results where A is the preferred base. Different colors indicate different strengths of BGC, which uniformly favors C/G over A/T alleles in both the test and the reference sequence.

in the effective strength of selection between the two is independent of the strength of BGC, conservation estimates are not. This is a consequence of the fact that the rate of evolution is a concave function over the relevant range of effective selection coefficients, as can be seen in figure 1*A*. Similarly, at the sites where A is the preferred state by selection (fig. 3*B* and *D*), increasing the strength of BGC generally leads to less conservation. In this scenario, BGC and purifying selection are operating in opposite direction. Thus, there is effectively less selection on the test sequence than the reference sequence. In both cases, the increase/decrease in conservation estimates due to BGC becomes more profound in the HKY85:π inference model than in the HKY85:ρ model.

If the orientation of functional selection to mutational bias is random, then, in the absence of BGC, all functional sites will indeed show sequence conservation regardless of the orientation of selection to mutation. The presence of BGC, however, magnifies the asymmetries in the inferred amount of conservation between sites with different preferred bases with some sites appearing to evolve faster than neutral. These asymmetries arise even in the absence of strong mutational biases (supplementary fig. S2, Supplementary Material online). BGC could thus further reduce the power to detect functional regions on the basis of sequence conservation.

A Practical Application: GERP RS-Scores on a Realistic Species Tree

We have investigated the behavior of ML branch-length inference in an instructive generalized framework to elucidate how the link between conservation and purifying selection can be misleading when selective forces are weak and mutations are biased. ML branch-length inference underlies many popular methods for conservation analysis such as GERP (Cooper et al. 2005) and phyloP (Pollard et al. 2010), raising the question to what extent such tools will be affected by these problems.

In order to investigate the effects on an exemplary practical application, we tested the performance of GERP on simulated multiple sequence alignments over a realistic species tree (Materials and Methods). The sequences in the test alignment were modeled to have evolved in

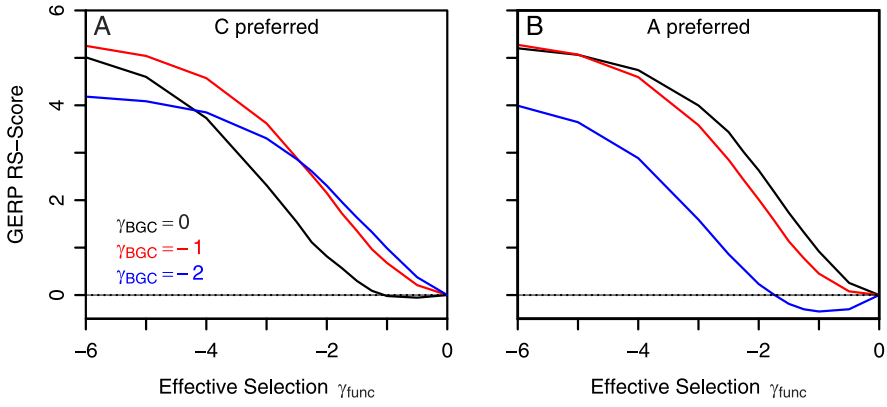

Fig. 4.—Performance of GERP on simulated sequence alignments over a realistic 32 mammalian species tree. The alignment sites were modeled to have evolved under mutational biases, randomly oriented weak constraint, and BGC. (*A*) Shows the mean RS-scores of all sites where C was the preferred state as a function of the strength of functional selection: γ_{func}. (*B*) Shows results for the sites where A was preferred state. The mutational bias was again $\pi_{A+T} = 0.8$ with a transition/transversion ratio of four. Positive RS-scores indicate branch-length reduction as the number of substitutions is lower than expected under neutrality (4.74)—the equivalent of $t^*/t_0^*<1$. Negative RS-scores indicate more substitutions having occurred than expected—he equivalent of $t^*/t_0^*> 1$. RS-scores were normalized such that the neutral class ($\gamma_{func} = 0$) has an RS-score of zero. Different colors indicate different strengths of BGC, which uniformly favors C/G over A/T alleles.

a scenario with mutational biases, BGC, and weak purifying selection. Mutations were again biased toward A/T with $\pi_{A+T} = 0.8$, and BGC was acting uniformly over the genome, favoring C/G over A/T bases at strength γ_{BGC}. In addition, purifying selection favored at every site a randomly selected base with the strength γ_{func}, analogous to the scenario in the previous section.

To obtain its site-wise rejected substitution scores (RS-scores), GERP calculates the difference between the number of substitutions expected to have occurred if the site was evolving neutrally and the actual number of substitutions inferred from the alignment at each site. GERP needs to be provided with the correct neutral tree. Underlying its ML inference is an HKY85 mutation model; the transition/transversion ratio needs to be specified by the user. Equilibrium nucleotide frequencies are estimated directly from the alignments (corresponding to our HKY85:ρ model from fig. 3). For our analysis, we provided GERP with the true tree that was used to generate the simulated sequence alignments and the correct transition/transversion ratio.

Figure 4 shows the mean RS-scores at sites where purifying selection favors C (fig. 4*A*) and at those sites where it favors A (fig. 4*B*), as a function of the strength of purifying selection (median RS-scores are shown in supplementary fig. S3, Supplementary Material online). Due to symmetry, A and T sites as well as C and G sites are again equivalent. Results are shown for three different BGC scenarios: $\gamma_{BGC} = \{0,-1,-2\}$.

The number of expected substitutions per site under neutrality and without BGC is 4.74 for our tree. According to figure 1, BGC of strength $\gamma_{BGC} = -1$ should then increase the average number of substitutions above the expectation of 4.74 substitutions per site, leading to positive RS-scores in the neutral scenario ($\gamma_{func} = 0$). In the scenario with $\gamma_{BGC} = -2$, fewer substitutions should occur and RS-scores should be negative in the neutral scenario. Because we want to compare RS-scores between constrained and neutral sites, RS-scores were shifted such that the class without functional selection ($\gamma_{func} = 0$) always has a score of zero. The magnitude of the resulting shift can be estimated from the difference between the limiting RS-score at large γ_{func} and the neutral expectation of 4.74 without BGC (at functional selection strengths of $\gamma_{func} < -6$ one obtains full conservation at each site).

As expected, GERP shows behavior similar to the theoretical HKY:ρ model in figure 3 at short timescales. In the absence of BGC, the sites where C is preferred appear to have evolved neutrally when functional selection is weaker than $\gamma_{func} \sim -1.5$. Sites where A is the preferred state, on the other hand, are inferred constrained for all $\gamma_{func} < 0$. When BGC is acting, this pattern switches: conservation will be inferred at sites where C is weakly preferred, whereas sites where A is weakly preferred now appear to have evolved neutrally. For strong BGC ($\gamma_{BGC} = -2$), RS-scores at the sites where A is preferred even drop below zero for $-2 < \gamma_{func} < 0$, indicating a faster than neutral rate of evolution in that range.

The complications emerging from the interplay of mutational biases, BGC, and weak functional selection thus also affect practical applications to infer sequence conservation from multiple sequence alignments, as exemplified here for the tool GERP. Depending on the orientation of functional selection, constrained sites will not always appear to be constrained and can even show less conservation than unconstrained sites. If conservation is inferred, the level of conservation will be highly asymmetric depending on which

base is the preferred state. These complications become even more profound when the mutational biases are weaker (supplementary fig. S4, Supplementary Material online). In functional regions, we typically expect the orientation of selection to be in some jumble over the different nucleotides. When averaged across such loci they should then appear less constrained than they actually are due to the lack of conservation at the sites where preferred bases and mutation oppose. As such, sitewise as well as regional conservation estimates may not always perfectly reflect the presence of functional selection.

Discussion

Interactions between mutational biases, BGC, and weak selection can have intricate effects on sequence conservation and its inference: the rate of evolution at constrained sites can actually be higher compared with neutrally evolving sites, and conservation estimates will typically depend on the orientation of selection and vary over time.

Integral to these complications is the presence of mutational biases. Recent studies point to the ubiquity of such mutational biases across a range of organisms. In the three classical genetic model organisms *Drosophila melanogaster*, *Saccharomyces cerevisiae*, and *Arabidopsis thaliana*, mutation accumulation studies revealed mutational biases pushing their genomes toward high equilibrium A/T contents. In Drosophila, the mutational equilibrium was calculated to be 67% A/T (Keightley et al. 2009); in yeast, the mutational bias should yield 74% A/T in equilibrium (Lynch et al. 2008); and in Arabidopsis, the mutational bias should push the genome toward 85% A/T (Ossowski et al. 2010). Likewise, the mutations in mitochondria of *S. cerevisiae*, *Caenorhabditiselegans*, and *D. melanogaster* appear to be highly biased (Haag-Liautard et al. 2008; Montooth and Rand 2008; Montooth et al. 2009). Mutations also seem to be generally biased toward A/T in prokaryotes (Hershberg and Petrov 2010; Hildebrand et al. 2010).

In addition to requiring mutational biases, selection needs to be weak for the complications in conservation estimation to arise. In order for substantial parts of a genome to be affected, many sites would have to be evolving under such weak selective forces. Is there evidence that weak selection is indeed common? Evidence of abundant weak purifying selection has been given by numerous population genetic studies (Bustamante et al. 2002, 2005; Ohta 2002; Lu and Wu 2005; Comeron 2006; Lipatov et al. 2006; Eyre-Walker and Keightley 2007; Eory et al. 2010). Furthermore, in genomic alignments, sites with intermediate GERP scores (those scores between neutrally evolving and totally conserved) are very common in constrained elements and coding sequences (Davydov et al. 2010; Goode et al. 2010). According to figure 4, for a site to appear completely conserved across our tree does not require strong selection. In fact, when $\gamma = -6$, almost all sites should appear completely conserved. Given the preponderance of sites with intermediate RS-scores, either constraint is weak at many sites or it varies over the tree such that the inferred constraint appears weak. If the former, then large parts of the genomes are indeed evolving in the parameter regime considered by our study.

There is also indirect evidence that weak selective forces are widespread across the genome in the form of discrepancies between mutational biases and genomic nucleotide contents. For instance, the A/T content of the yeast genome is only 60% as compared with its mutational equilibrium of 74% A/T (Lynch et al. 2008). In Arabidopsis, while the mutational bias should drive the genome toward 85% A/T, the actual intronic/degenerate codon A/T content bias is only 65/68% (Ossowski et al. 2010). If these differences are not assumed to reflect varying mutational biases over time, selective forces (i.e., either natural selection or BGC) need to be causing biases in fixation probabilities that partly compensate for the mutational biases. To yield the observed genomic content biases, the strength of such forces would correspond to effective selection of strength $\gamma \sim -0.7$ for fixation of A/T versus C/G in yeast and $\gamma \sim -1$ in Arabidopsis. For both cases, these points are almost exactly where r/r_0 has its maximum elevation for the respective mutational biases (fig. 1*A*). Many regions of the yeast and Arabidopsis genomes might then in fact be evolving more rapidly than they would if fixation probabilities were solely determined by random genetic drift.

In bacteria, the differences between mutational biases and genomic nucleotide composition are typically even more profound. Although recent evidence suggests that bacterial mutations is universally biased toward A/T (Hershberg and Petrov 2010; Hildebrand et al. 2010), genomic nucleotide contents can vary widely from ~20 to ~80% A/T. These examples illustrate various situations where weak selective forces seem to be acting systematically in opposition to existing mutational biases.

It thus seems that both necessary ingredients for the complications in conservation estimates to arise, mutational biases and weak selective forces, are likely to be common in nature. Consequently, a substantial fraction of genomic sites should suffer from misleading conservation estimates. What are the implications for conservation analysis?

Conservation clearly remains a consistent and useful measure for evolutionary constraint where selection is sufficiently strong and also for weak constraint if mutations are unbiased. After all, comparative genomics approaches based on sequence conservation have proven extremely successful in annotating functional regions. Regions evolving under weak constraint, however, could often be missed by current approaches as they will not always show the expected signature of conservation. In particular, sites where selection and mutational biases oppose might

actually be inferred to have evolved under positive selection. Such effects should generally reduce the power of approaches that identify functional regions from conservation signatures. This could be a contributing factor to why half of the functional regions characterized by the ENCODE project (Birney et al. 2007) do not show conservation signatures (Pheasant and Mattick 2007).

In addition, estimated conservation scores are difficult to relate to the effective strength of selection operating on a site. As we have shown, conservation estimates are highly asymmetric with regard to the preferred base and typically vary over time. Purifying selection of the same strength can thus give rise to very different conservation estimates depending on the particular scenario.

Our analysis also highlights a notorious general problem of conservation analysis: the assumption of having a truly neutrally evolving reference sequence for comparison with the test region. If BGC or selection are the underlying causes for the observed discrepancies between mutational biases and genomic nucleotide contents, then these processes would likely be acting throughout the genome, making it very difficult to find truly neutrally evolving regions. Our results indicate that this should generally exaggerate the complications in conservation analysis. Moreover, BGC is likely to show regional variation along a genome. For example, it has been suggested that some fraction of the so-called human-accelerated regions, which show an increase of substitutions on the human lineage when compared with their homologous regions across a phylogeny (Pollard et al. 2006b), are in fact constrained sequences to which recently a BGC hot spot has moved (Pollard et al. 2006a; Galtier and Duret 2007; Berglund et al. 2009). Variation in the efficacy of BGC along genomes will lead to particular sequence regions evolving faster or slower than others, further obfuscating conservation estimates.

The limitations of conservation analysis presented here spotlight the need for more accurate inference methods in comparative genomics in order to also capture regions evolving under weak constraint. Substitution models more reflective of the actual substitution processes seem essential to these improvements. Such substitution models should disentangle the actions of mutation and selection, incorporating the true mutational biases and models of selection that explicitly account for fixation probabilities. Constraint can then be inferred directly by estimating the parameters of the mutation-selection model without the potentially misleading estimation of conservation with respect to a reference sequence.

Recent advances in experimental techniques render possible an unbiased estimation of mutation biases. With the advent of new sequencing technologies, resequencing large numbers of genomes has become a practical endeavor (Durbin et al. 2010). This opens up the possibility for whole-genome sequencing of mutation accumulation lines, as well as the analysis of deep polymorphism data from natural populations. Both approaches should allow for more accurate estimates of the mutational spectrum (Lynch et al. 2008; Messer 2009; Ossowski et al. 2010). Factors such as neighbor-dependent mutation rates or transcription-associated mutational asymmetry ideally should also be considered.

Substitution models that explicitly incorporate the action of selection do also already exist (Halpern and Bruno 1998; Moses et al. 2004; Doniger and Fay 2007; Yang and Nielsen 2008; Berglund et al. 2009; Rodrigue et al. 2010). The first attempt applied to codon substitution models (Halpern and Bruno 1998) suffered from a very high number of parameters as the fitness for each amino acid at every position in the protein had to be estimated. Recent work has ameliorated this issue by effectively reducing the number of free parameters while retaining some of the site-wise flexibility (Rodrigue et al. 2010). Other applications of mutation-selection models have focused on more tractable scenarios with intrinsically fewer parameters such as codon bias and BGC (Yang and Nielsen 2008; Berglund et al. 2009). Models for transcription factor binding sites have taken the concept one step further by additionally using functional information from position weight matrices as fitness parameters (Moses et al. 2004; Doniger and Fay 2007). For the scenarios to which they have been applied, mutation-selection models have been shown to generally outperform simple neutral models (Yang and Nielsen 2008; Rodrigue et al. 2010). The results presented in this paper suggest some of the reasons for this increase in performance.

With the seeming pervasiveness of strong mutational biases, weak selective constraint, and BGC, the substitution dynamics of any given genomic locus is likely to be poorly captured by neutral models. The ingredients for better inference models are 2-fold: 1) a precise measurement of the true underlying mutational biases and 2) the disentanglement of mutation and selection together with a more accurate modeling of the specific nature of the selective forces. Using a methodology with such ingredients built-in, constraint can be inferred directly, circumventing the complications arising from indirect inference via conservation. Much work is needed to independently measure mutational biases in many organisms and to improve the efficiency and fidelity of the selection models, while controlling model complexity and avoiding overfitting. Given the advantages, however, the design and implementation of such explicit mutation-selection models is highly desirable.

Supplementary Material

Supplementary figures S1–S4 are available at *Genome Biology and Evolution online* (http://gbe.oxfordjournals.org/).

Acknowledgments

This research was supported by grants to D.A.P. from the National Institute of Health (GM 077368) and the National Science Foundation (0317171). D.S.L. is supported by the Stanford Genome Training Program. P.W.M. is an HFSP post-doctoral fellow. We thank members of the Petrov Lab, Adam Siepel, Arend Sidow, Gill Bejerano, Nadia Singh, Guy Sella, and two reviewers for helpful feedback.

Literature Cited

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. PLoS Biol. 7:e26.

Birney E, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 447: 799–816.

Bustamante CD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature. 437:1153–1157.

Bustamante CD, et al. 2002. The cost of inbreeding in Arabidopsis. Nature. 416:531–534.

Comeron JM. 2006. Weak selection and recent mutational changes influence polymorphic synonymous mutations in humans. Proc Natl Acad Sci U S A. 103:6940–6945.

Cooper GM, et al. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15:901–913.

Davydov EV, et al. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 6:e1001025.

Doniger SW, Fay JC. 2007. Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol. 3:e99.

Durbin RM, et al. 2010. A map of human genome variation from population-scale sequencing. Nature. 467:1061–1073.

Duret L. 2009. Mutation patterns in the human genome: more variable than expected. PLoS Biol. 7:e1000028.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4:e1000071.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 10:285–311.

Eory L, Halligan DL, Keightley PD. 2010. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. Mol Biol Evol. 27:177–192.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet. 8:610–618.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet. 23:273–277.

Goode DL, et al. 2010. Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. Genome Res. 20: 301–310.

Haag-Liautard C, et al. 2008. Direct estimation of the mitochondrial DNA mutation rate in Drosophila melanogaster. PLoS Biol. 6:e204.

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol. 15:910–917.

Hardison RC. 2003. Comparative genomics. PLoS Biol. 1:E58.

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.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 6(9):e1001115.

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 6(9):e1001107.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.

Keightley PD, et al. 2009. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. Genome Res. 19:1195–1201.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Lio P, Goldman N. 1998. Models of molecular evolution and phylogeny. Genome Res. 8:1233–1244.

Lipatov M, Arndt PF, Hwa T, Petrov DA. 2006. A novel method distinguishes between mutation rates and fixation biases in patterns of single-nucleotide substitution. J Mol Evol. 62:168–175.

Lu J, Wu CI. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc Natl Acad Sci U S A. 102:4063–4067.

Lynch M, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A. 105:9272–9277.

Margulies EH, et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res. 17:760–774.

McVean GAT, Charlesworth B. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet Res. 74:145–158.

Messer PW. 2009. Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. Genetics. 182: 1219–1232.

Montooth KL, Abt DN, Hofmann JW, Rand DM. 2009. Comparative genomics of Drosophila mtDNA: novel features of conservation and change across functional domains and lineages. J Mol Evol. 69:94–114.

Montooth KL, Rand DM. 2008. The spectrum of mitochondrial mutation differs across species. PLoS Biol. 6:e213.

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.

Nagylaki T. 1983. Evolution of a finite population under gene conversion. Proc Natl Acad Sci U S A. 80:6278–6281.

Ohta T. 2002. Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci U S A. 99:16134–16137.

Ossowski S, et al. 2010. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science. 327: 92–94.

Pheasant M, Mattick JS. 2007. Raising the estimate of functional human sequences. Genome Res. 17:1245–1253.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20:110–121.

Pollard KS, et al. 2006a. Forces shaping the fastest evolving regions in the human genome. PLoS Genet. 2:e168.

Pollard KS, et al. 2006b. An RNA gene expressed during cortical development evolved rapidly in humans. Nature. 443:167–172.

Rodrigue N, Philippe H, Lartillot N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc Natl Acad Sci U S A. 107:4629–4634.

Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.
Waterston RH, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520–562.
Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol. 25:568–579.

Associate editor: Ross Hardison