Nonadaptive Amino Acid Convergence Rates Decrease over Time

Richard A. Goldstein,[1] Stephen T. Pollard,[2] Seena D. Shah,[2] and David D. Pollock*[,2]

[1]Division of Infection & Immunity, University College London, London, United Kingdom

[2]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora

***Corresponding author:** E-mail: david.pollock@ucdenver.edu.

Associate editor: Tal Pupko

Abstract

Convergence is a central concept in evolutionary studies because it provides strong evidence for adaptation. It also provides information about the nature of the fitness landscape and the repeatability of evolution, and can mislead phylogenetic inference. To understand the role of adaptive convergence, we need to understand the patterns of nonadaptive convergence. Here, we consider the relationship between nonadaptive convergence and divergence in mitochondrial and model proteins. Surprisingly, nonadaptive convergence is much more common than expected in closely related organisms, falling off as organisms diverge. The extent of the convergent drop-off in mitochondrial proteins is well predicted by epistatic or coevolutionary effects in our "evolutionary Stokes shift" models and poorly predicted by conventional evolutionary models. Convergence probabilities decrease dramatically if the ancestral amino acids of branches being compared have diverged, but also drop slowly over evolutionary time even if the ancestral amino acids have not substituted. Convergence probabilities drop-off rapidly for quickly evolving sites, but much more slowly for slowly evolving sites. Furthermore, once sites have diverged their convergence probabilities are extremely low and indistinguishable from convergence levels at randomized sites. These results indicate that we cannot assume that excessive convergence early on is necessarily adaptive. This new understanding should help us to better discriminate adaptive from nonadaptive convergence and develop more relevant evolutionary models with improved validity for phylogenetic inference.

Key words: **convergence, coevolution, epistasis, epistatic interactions, thermodynamics, adaptation, selection, evolutionary process, molecular evolution, phylogenetics, amino acid propensities, Stokes shift, Stokes–Fisher model.**

Introduction

Although evolution mostly proceeds by accumulation of differences between groups, numerous examples of convergent evolution exist, where similar solutions are found to similar evolutionary problems. Well-known morphological examples include eyes and wings, but an increasing number of examples are known at the molecular level, including proteins involved in echolocation in bats and cetaceans (Liu et al. 2010; Shen et al. 2012; Parker et al. 2013), foregut fermentation proteins in monkeys and cows (Stewart et al. 1987), transcription factors in mammals and birds (Yokoyama and Pollock 2012), and mitochondrial proteins among different snakes (Castoe et al. 2008), and mitochondrial proteins between snakes and agamid lizards (Castoe et al. 2009).

Such convergence at the molecular level can both confound and inform evolutionary analyses. Convergent evolution can result in erroneous phylogenetic trees by showing strong support for incorrect topologies (Castoe et al. 2009). However, replicated evolution to the same trait or amino acid in different lineages provides convincing evidence of adaptation (Castoe et al. 2008). In addition, convergent evolution can provide important information about the adaptive landscape; the relationship among genotype, phenotype, and fitness; the constraints acting on evolutionary processes; and the role of chance and necessity in evolution.

Statistically meaningful analyses of adaptive convergence rely on estimates of the likelihood that such convergence could occur by chance in the absence of adaptation. Such analyses generally rely on standard models of evolution (Rokas and Carroll 2008; Parker et al. 2013), but it is now clear that these models are woefully inadequate, drastically underestimating the levels of nonadaptive convergence (Castoe et al. 2009). We need to improve our ability to predict the amount of expected nonadaptive convergence if we want to avoid errors in phylogenetic relationships, make accurate inferences of adaptive evolution, and investigate what convergence tells us about the fitness landscape and evolutionary process.

Two assumptions common to most evolutionary models are that evolutionary processes are homogeneous among sites in an alignment, and over time. It is becoming increasingly clear that both assumptions are unjustified. Different distributions of amino acids are found in buried locations in the protein structure, exposed locations, tight turns, transmembrane helices, disordered regions, and locations of

Article

Fast Track

functional significance, indicating different selective constraints at these different types of locations. These differences are embodied in some mutation selection models (Halpern and Bruno 1998; Tamuri et al. 2011) and mixture models (Koshi and Goldstein 1998, 2001; Lartillot and Philippe 2004).

The evolutionary process at individual sites can also vary as a result of changes in structure, function, physiological role, or context of the corresponding location in the protein structure (Robinson et al. 2003; Blackburne et al. 2008; Tamuri et al. 2009; Kleinman et al. 2010; Pollock et al. 2012). In addition, in the presence of epistatic or coevolutionary interactions between sites, the process at one site will change due to substitutions that occur at other coupled sites (Pollock et al. 2012; Pollock and Goldstein 2014). There has been increasing evidence for the importance of epistatic interactions. For instance, Bloom et al. performed measurements on influenza proteins and observed that the effect of a substitution on the thermodynamic properties depended on the amino acids found in other positions (Ashenberg et al. 2013; Pollock and Goldstein 2014). Pollock et al. (2012) demonstrated how amino acid propensities at a site will adjust over time after a substitution, such that the resident amino acid (and others with similar physicochemical properties) tends to be the most favorable amino acid at that site, an effect they termed an "evolutionary Stokes shift." As a result, the selective constraints at each site will shift to follow the changing occupant at that site.

The amount of amino acid variation in a protein can be decomposed into the variation allowed due to the site- and time-specific constraints, plus the effect of variation in those constraints among sites and over time. As a result, models that neglect variation in evolutionary constraints over sites and time tend to underestimate the magnitude of instantaneous selective constraints at individual sites, resulting in an underestimation of the expected amount of neutral convergence. In addition, temporal heterogeneity in selective constraints may induce time dependence to the neutral rate of convergence. We therefore set out to quantify the frequency of convergence in a data set of mitochondrial proteins and investigate changes in convergence patterns over time. We then compare these results with predictions from standard models, as well as simulated proteins evolving under purifying selection for thermodynamic stability, similar to simulations used in Pollock et al (2012). We then consider what the results indicate about the process of protein evolution.

Results

Convergence Decreases with Time in Vertebrate Mitochondrial Proteins

We examined convergence events occurring on distinct branches in a phylogenetic tree (supplementary fig. S1, Supplementary Material online) from a concatenated alignment of all 13 mitochondrial protein sequences from over 600 vertebrate mitochondrial genomes. A fixed amino acid substitution model mtMam (Yang et al. 1998) with site rate variation with five gamma distributed rates was used to infer the substitutions. When comparing the substitutions on two distinct branches, a pair of substitutions on each branch at the same site can be classified as either a convergence event (C) if the substitutions are to the same amino acid, or as a paired divergence event (D) if the substitutions resulted in different amino acids. Bayesian estimation was used to obtain the C and D totals for each branch pair considered. For short branches, C and D would both be roughly proportional to the product of the two branch lengths, suggesting that the branch length dependence could be minimized by considering the ratio of convergence and divergence events, C/D. This is supported by previous analyses showing that C and D are highly correlated and that D is a better predictor of C than branch lengths (Castoe et al. 2009). For display purposes, only substitutions with greater than 90% posterior probability were considered in calculating C/D for figures 1–4, although all significance and credible region estimates were obtained by integrating overall ancestral state uncertainty (see Methods and supplementary Methods, Supplementary Material online). Distances between branches were calculated as patristic distances along the phylogenetic tree, measured between the ancestral nodes on each branch. Note that we do not assess or make use of the state of the site in the more ancient common ancestor of both branches. Distances are given in units of expected number of replacements per site.

The observed C/D ratios depend strongly on the distance between branches (fig. 1), a result that might seem surprising in the context of standard time- and site-homogeneous models of substitution, simply because if the model does not change one might think the C/D ratio would not change either (we elaborate further on these expectations below). The ratio is extremely high (0.4) for the shortest distances between branches, falling to below 0.2 for the

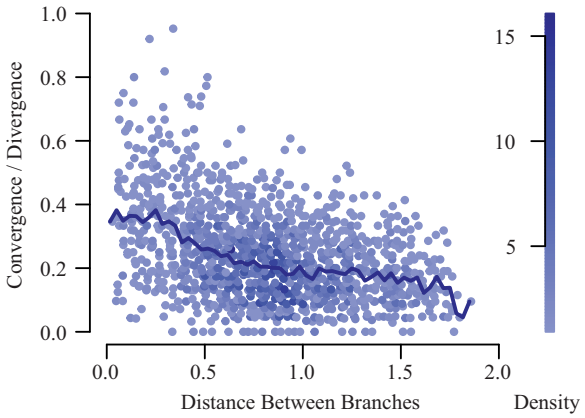

FIG. 1. Change in convergence over time in mitochondrial proteins. The convergence over branch-paired divergence ratio (C/D) was estimated for all eligible pairs of branches in the mitochondrial phylogeny. To help visualization of the data, overlapping data points were merged into single points with the color determined by the density of dots merged, with blue intensity gradient as shown in the scale to the right. We used a threshold of $D \geq 20$ for inclusion in this graph. The distance between branches shown is the patristic distance between the ancestral nodes of each branch, measured in average number of amino acid replacements per site. The blue line shown is a running average with window size 0.03.

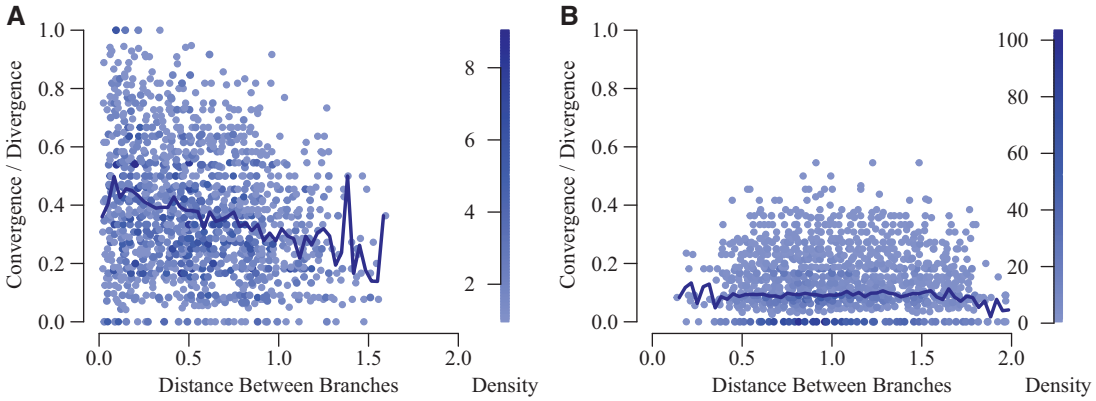

FIG. 2. Mitochondrial protein convergence for identical and different ancestral amino acids. The convergence over paired divergence ratios were estimated, merged, and colored as described in figure 1, except that events were separated into two categories depending on whether the ancestral amino acid at a site was the same (*A*) or different (*B*). We used a threshold of $D \geq 0$ for inclusion in these graphs.

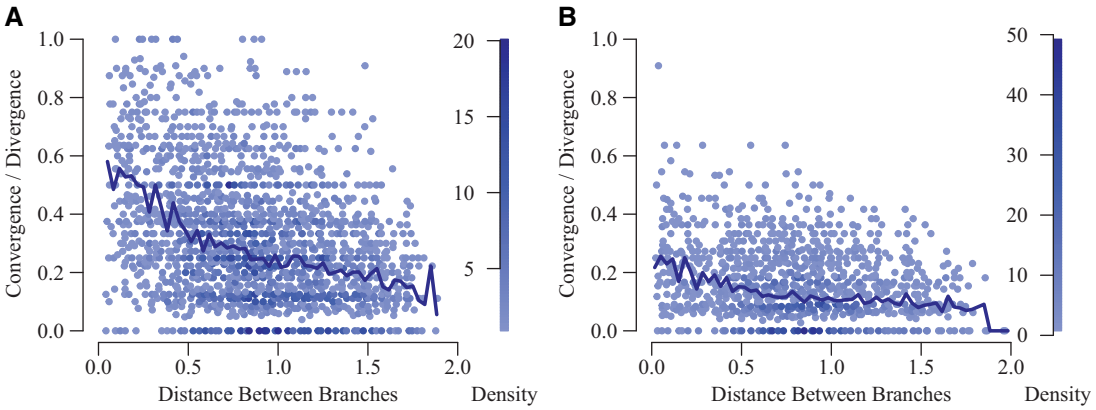

FIG. 3. Mitochondrial protein convergence for conserved and variable sites. The data and visualization are the same as in figure 1, except that ratios were estimated separately for conserved (*A*) and variable (*B*) sites. We used a threshold of $D \geq 7$ for *A* and $D \geq 10$ for *B*.

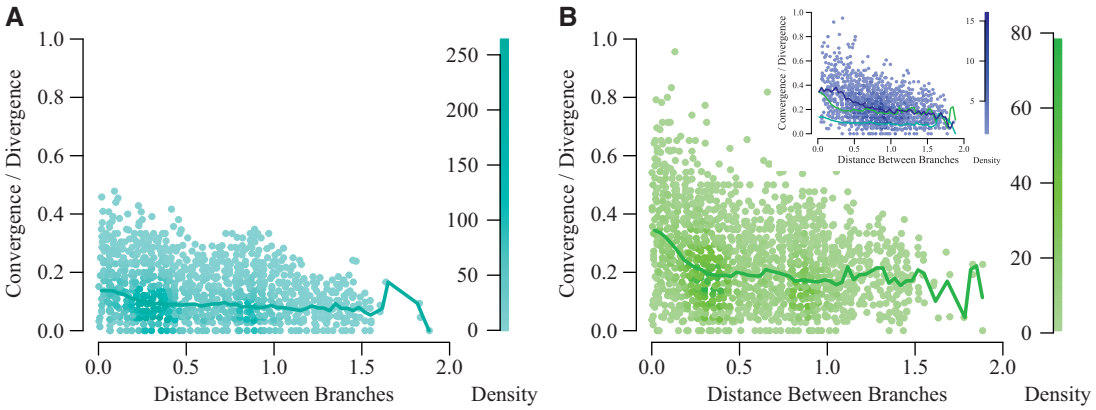

FIG. 4. Convergence in simulated data. Protein evolution was simulated along the mitochondrial tree using the WAG substitution model (*A*) and Stokes–Fisher protein evolution model (*B*). *C*/*D* ratios were calculated using the same method as with the mitochondrial data (figs. 1–3) and were visualized the same as in figure 1. The inset in (*B*) shows the SF and WAG averages along with the mitochondrial data average (in blue, as before), for comparison. We used a threshold of $D \geq 20$ for *A* and $D \geq 20$ for inclusion in *B*.

most separated branches. The 99% credible regions for the expected *C*/*D* ratios over time are shown in supplementary fig. S2, Supplementary Material online, and they are nonoverlapping until later times when the ratios fall below 0.2 (supplementary table S1 Supplementary Material online). The variation in ratios among branch pairs is high, with ratios for short to medium branch distances (<0.5 replacements per site) ranging from zero to nearly one. Notably, this high

variation mostly arises from biological variation in the expected ratio among branch pairs, not from poor estimation of ratios with few C or D counts (see model predictions below). There is also a strong dependence on whether or not the amino acid is different in the ancestral sequences of the two branches (fig. 2). When the ancestral amino acids are identical, the average ratio starts at about 0.45 and drops to 0.2, whereas when the ancestral amino acids are different the average ratio is approximately constant at 0.08.

These results strongly suggest that amino acid propensities and therefore substitution possibilities at each site are initially highly constrained. We can calculate an effective number of accessible residues by considering the size of the alphabet of states m that would result in a particular value of C/D if all substitutions were equally likely (i.e., a Jukes Cantor [JC] model; Jukes and Cantor 1969). As shown in the supplementary Material, Supplementary Material online, if there is only a short evolutionary distance between the branches so that neither amino acid has changed, $C/D = \frac{1}{m-2}$. An initial C/D ratio of 0.45 therefore indicates an effective number of only 4.2 accessible residues per site. (These initial measurements are taken over the period of time represented by the length of the branches, and given the falloff in the C/D ratio, the initial instantaneous ratio, prior to sequence divergence, may have been substantially higher). When the ancestral amino acids are known to differ, C/D is equal to $\frac{m-2}{m^2-3m+3}$, and therefore a ratio of 0.08 is equivalent to an effective number of 13.4 residues per site.

These results also strongly support the idea that amino acid constraints change over time, because the C/D ratios drop even when the ancestral amino acid is identical (fig. 2*A*). Although the changing convergence probability in the overall data set (fig. 1) can be understood by changing mixtures of sites with the same ancestral states (fig. 2*A*) and different ancestral states (fig. 2*B*), it does not appear possible to explain the drop in convergence seen in figure 2*A* based on changing site composition. If anything, as discussed below, the bias in composition due to removal of evolved sites should remove low constraint (low convergence probability) sites, which would increasingly produce a bias for sites with higher convergence probabilities. The 99% credible region for the slope of a linear model fit to the data from figure 2*B* shows a clear decrease in C/D ratios with divergence (-0.109, -0.094; supplementary fig. S3, Supplementary Material online). We therefore conclude that the constraints are likely changing over time.

In contrast to the strong apparent initial constraint, once the amino acids at a site diverge, the number of amino acids acceptable at a site is quite high, drastically reducing the chance of convergence. (Recall that a C/D ratio of 0.08 corresponds to an effective number of 13.4 accessible residues per site.) To determine if sites retain information about convergence probabilities in the case of different ancestral amino acids, we resampled the substitutions among all sites, maintaining the same ancestral amino acids for each substitution. For example, if a branch has a substitution at site 5 from alanine to glycine, we collected all the substitutions from alanine on all branches and at all sites, then replaced the glycine with an amino acid randomly chosen from the descendent amino acids of the collected substitutions. The C/D was then recalculated for every branch pair using these resampled substitutions. The results are shown in supplementary figure S4, Supplementary Material online. The C/D ratios for these resampled replacements are essentially the same as for the observed ratios (fig. 2*B*), indicating that, conditional on the different ancestral states, the sites provide no further detectable information about convergence probabilities.

To further understand these results, we partitioned the sites roughly evenly into three conservation classes. For recently diverged branch pairs, the average C/D ratio was highest for conserved sites, starting at about 0.6 and falling to below 0.2 (fig. 3*A*). In contrast, the average C/D ratio at variable sites was initially only slightly above 0.2 and fell quickly to near 0.1 (fig. 3*B*). As with the overall ratios, the fast- and slow-evolving sites may have started out with much higher C/D ratios, but the ratio dropped off too quickly to measure over finite branches. This would particularly affect the fast-evolving sites, and we cannot know for sure if the differences between figures 3*A* and *B* are due to an inherently higher convergence probability at more conserved sites or if they occur because highly variable sites reach equilibrium much faster. The results for identical and different ancestral states at each site for each conservation level (supplementary figs. S5 and S6, Supplementary Material online) are similar to the results for the complete data set (fig. 2), albeit noisier. It is worth noting, however, that the C/D ratio from identical amino acids at conserved sites (supplementary fig. S3*A*, Supplementary Material online) also falls off over time, indicating that the effect of fluctuating constraints over time on convergence probabilities is strong even for the most conserved sites.

Relationship of Convergence with Time under Different Evolutionary Models

Given the convergence results for the mitochondrial data, we wanted to know the degree that these results are predicted by existing substitution models. We first simulated data along the mitochondrial tree under two different models and then we inferred the ancestral sequences and calculated the C/D ratios using the same method as with the mitochondrial data. The two models we used were the following: An amino acid substitution matrix (Whelan and Goldman model, WAG) (Whelan and Goldman 2001), which neglects differences among sites and over time but accounts for differences in the rates of exchange among different amino acids; and the recent thermodynamic-based Stokes–Fisher (SF) model (Pollock et al. 2012), which allows for coevolution (epistasis) among sites, and thus allows for different processes among sites and over time. We wish to avoid the possible impression that the SF models we use are designed to accurately reflect the true model of evolution. Instead, the SF model was constructed to generate semirealistic simulations that have many salient aspects of evolution similar to real proteins and can therefore guide us to better interpret observations on real protein evolution.

The "WAG" model results in a low and slightly decreasing *C/D* ratio that is not highly variable (fig. 4*A*). In contrast, the SF model results (fig. 4*B*) are a remarkably good match to the mitochondrial data results (fig. 4*B* insert). This indicates that the variance observed in the mitochondrial data is not just the result of estimation error, and the general shape of the curve is a fundamental expectation for evolution of complex functional molecules such as proteins, and is not specific to mitochondrial proteins.

To further dissect the basis for the observed effect, we analyzed additional models of varying complexity, including the simple JC model (equal rates of amino acid exchange), a model with codon structure (Zihengian model, Z), WAG and Z with gamma-distributed rate parameters, and the CAT-60 model (CAT), which includes variation in constraint among sites (Jukes and Cantor 1969; Goldman and Yang 1994; Quang et al. 2008). These models, except the CAT-60 model and WAG, had their parameters fitted to data from SF simulations that used a star phylogeny, instead of the mitochondrial phylogeny. The form of these models allowed exact calculations of the expected mean *C/D* ratio, including the potentially higher initial instantaneous *C/D* ratio inaccessible to analyses on trees with finite branch lengths.

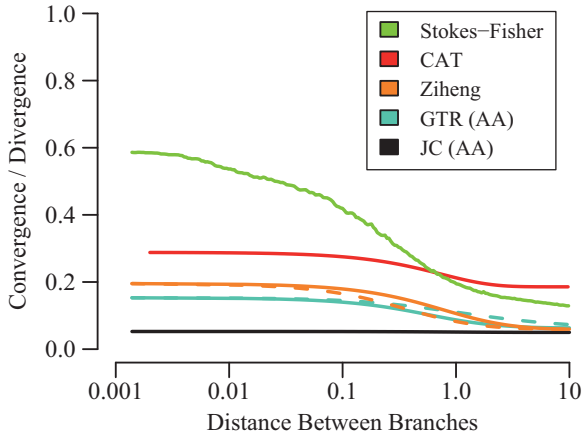

Fig. 5. Convergence in simulated data under models of different complexity. The *C/D* ratios shown are from exact calculations on specific models. The different models shown (see legend for line colors) are JC, WAG, WAG plus gamma, Z, Z plus gamma, and SF. The results for WAG and Ziheng with gamma rate variation are shown with a dashed line. Note that unlike previous figures, the distances are on a log scale.

There is a decreasing *C/D* with time in all models of evolution (fig. 5), although it is barely perceptible for the JC model, and it is a relatively small effect for the *WAG* and Z models. For models with a constant site-specific process over time (all models except SF), the change in ratio is mostly attributable to the difference in the number of available convergent states depending on whether the ancestral state is the same or different. It is interesting that the WAG and Z models both have small but slightly different responses to adding site-specific rate variation, with WAG somewhat delaying its drop in convergence levels and the Z model accelerating the drop. We speculate that convergence levels in the WAG model are more dependent on slower exchanges, whereas in the Z model the drop in faster sites takes precedence.

The results for that CAT model are especially notable when broken down into same and different ancestral states (fig. 6*A*). Although the *C/D* ratios for diverged ancestral states are somewhat higher than the equivalent results from the mitochondrial (fig. 2*B*) data, the truly notable observation is that the *C/D* ratio for sites with the same ancestral amino acid actually increases over time under the CAT model. In principle, if one observes *C/D* ratios for the same set of sites that change neither their ancestral amino acids nor their propensities over time (as in the CAT model), then their *C/D* ratio must remain constant. However, the set of amino acids is changing in this case because the sites that evolve more rapidly are more likely to have differing ancestral amino acids (about 80% of sites by the most divergent timepoint; see fig. 6*A* inset). Unsurprisingly, the sites that change tend to have higher entropy than the sites that do not, and the sites with unchanged amino acids have less average entropy over

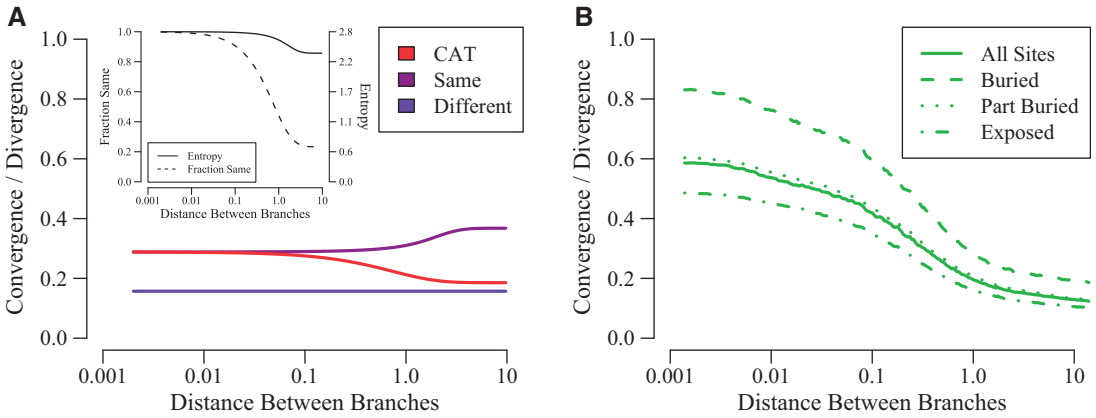

Fig. 6. Convergence for the CAT model and the SF model with sites segregated by burial in structure. (*A*) The *C/D* ratios are shown for the overall CAT model (red), as well as *C/D* ratios depending on whether the ancestral states are the same (purple) or different (blue). The inset shows the fraction of sites that have the same ancestral state (dashed line) and the entropy averaged over all sites that have the same ancestral state (solid line). (*B*) The *C/D* ratios shown are the same as the SF runs in figure 5, but sites were determined to be buried, partially buried, or exposed based on fraction exposed surface area for the corresponding amino acid in the protein structure.

time (fig. 6*A* inset). The increase in *C*/*D* ratios in the CAT model for sites with the same ancestral amino acid is thus explained by the lower average entropy (and thus greater constraint) over time at those sites that remain unchanged. This result is clearly exactly opposite the results from the mitochondrial data, in which sites with the same ancestral amino acids have clearly decreasing *C*/*D* ratios with time. Although it is possible to conceive the evolutionary models that are constant over time but still result in decreasing *C*/*D* ratios in this case (i.e., if low entropy sites all had extremely high mutation rates), such models would appear rather artificial and would have to overcome the naturally higher substitution rates of high entropy (low constraint) sites. It is much easier (and perhaps more natural) to explain these results with models that involve fluctuating constraints over time, of which the SF model is but one example.

Finally, we described above that sites with different levels of sequence conservation behaved differently in terms of their drop in *C*/*D* ratio. To understand this better, we separated out the instantaneous *C*/*D* ratio expectations for sites in the SF star phylogeny simulations corresponding to buried, partially buried, and exposed locations in the protein structure. From this we can see that indeed the buried sites start out with a higher *C*/*D* ratio of slightly over 0.8, and retain a higher ratio throughout the evolutionary simulation (fig. 6*B*). In contrast, the exposed sites start out with a lower ratio of about 0.5, and are always lower. This implies that the instantaneous site-specific constraints under the SF model are highest at buried (more slowly evolving) sites and lowest at exposed sites. By analogy, this suggests that similar factors are at play in producing the real mitochondrial protein differences in *C*/*D* ratios between slow- and fast-evolving sites observed in figure 3.

Discussion

Current treatment of convergent events generally assumes that nonadaptive convergence at the molecular level is well predicted by simple time-averaged and site-averaged models. However, our analysis of real proteins and model-based simulations demonstrates that the rate of convergence changes over time, and can be extremely high for recently diverged proteins. The convergence data presented here provide additional evidence that our understanding of how proteins evolve needs to be fundamentally revised. The patterns of convergent evolution observed may cause difficulties for phylogenetic reconstructions, but can also provide important information about adaptation and adaptive bursts, as well as allowing us to investigate the underlying topology of the fitness landscape and the nature of the substitution process.

Convergence probability is closely related to the number of amino acids that are acceptable at a given site at a given time. If a small hydrophobic amino acid is required, the probability that two acceptable substitutions in different lineages will result in the same small hydrophobic amino acid can be quite high. Constraints at another site requiring large flexible amino acids will result in a similarly high probability of convergence. If the substitution model is inferred by averaging over different sites, or the same site at different times, including instances where only small hydrophobic, or large flexible, or aromatic, or charged amino acids are required, the result is a model with few constraints that allows a wide variety of different amino acids. These simple models will overestimate the number of acceptable amino acid substitutions and underestimate the probability of convergence.

As indicated above, the high rate of convergence and the strong dependence of the convergence rate on evolutionary distance strongly suggest the importance of variation in the substitution rate across sites and over time. The idea of fluctuating amino acid substitution rates over time is an important feature of evolutionary Stokes-shift theory (Pollock et al. 2012). According to this theory, the fitness of an amino acid for any site, and therefore the propensities for the amino acid at that site, is dependent on how well suited it is to the environment formed by the amino acids at neighboring and interacting sites. As substitutions at neighboring sites alter the environment of a site, the amino acid propensities of that site will also be altered, resulting in fluctuating substitution rates at that site. Homologous but divergent proteins in other species will likely have fluctuated differently, meaning that the sets of acceptable amino acids at each position will diverge with evolutionary distance, causing a falloff in the convergence probability. In Stokes-shift theory, divergence in substitution models at a site is strongly coupled to substitutions at that site, so the convergence rate will also be significantly lower following a substitution, consistent with the data shown in figure 2.

The SF model makes three additional predictions. First, as the selection at different sites in the protein will be of different and fluctuating magnitude, there should be large differences and fluctuations in the convergence probability, as shown in figure 4*B*. Second, we would expect more buried locations to be under more stringent constraints, resulting in a higher convergence probability than exposed locations, as shown in figure 6*B*. Third, as also shown in figure 6*B*, we expect the selective constraints at buried locations to diverge slowly because the residues around such locations are also buried and evolve slowly, resulting in a slower decline in the convergence probability with increasing evolutionary distance. All these predictions are matched by the observations of mitochondrial proteins (figs. 1 and 3).

Both heterogeneity of selection at different sites in the protein and fluctuations in selection over evolutionary time can cause models that neglect these effects to underestimate convergence rates. In particular, the CAT model (Quang et al. 2008), which includes spatial variation and excludes temporal variation, generates initially high *C*/*D* ratios that decline over evolutionary distance in a similar manner as the SF model (fig. 6*A*). Similar drops in *C*/*D* ratios can also be seen in other highly parameterized site-specific models of spatial variation (data not shown). However, the effect of spatial versus temporal variation can be distinguished by considering the evolutionary distance dependence of *C*/*D* ratios from the same ancestral states. As shown in figure 6*A*, this ratio increases with evolutionary distance when a model is used (CAT) that includes only spatial variation. Sites with fewer constraints are more likely to undergo changes, and therefore less likely to

have the same ancestral states at longer divergence times. As a result, as shown in figure 6*A* inset, the sites with the same ancestral states become increasingly the highly constrained sites with lower sequence entropy. As more constrained sites have higher *C*/*D* ratios, this means that *C*/*D* for these sites will increase with evolutionary distance. In contrast, when there are temporal changes in selection, diverging sequences will increasingly be under different selective constraints. This can result in a decreasing *C*/*D* ratio with increasing evolutionary distance, as observed in figure 2*A*. A fluctuating temporal component is not surprising, as no plausible biophysical model would allow site-specific constraints to remain fixed in the face of divergence in the rest of the protein, and there is other strong evidence for coevolution (or epistasis) among residue positions (Pollock et al. 1999, 2012; Pollock and Goldstein 2014).

The effects of fluctuating and poorly estimated neutral convergence may have substantial effects on phylogenetic inference. Although truly neutral convergence is expected to be unbiased to any particular phylogenetic solution, it may well add considerable noise that would mask true phylogenetic signal. The distance dependence of the convergence probability may also interact in complex ways with the well-known phylogenetic problem of long-branch attraction (Felsenstein 2004), and we expect that extensive analyses will be necessary to sort out such interactions. Furthermore, it is clear that our new understanding of fluctuating substitution processes suggests a multitude of new questions about how protein evolution operates and the role of convergence analysis in understanding protein evolution. Can we use convergence to better estimate instantaneous constraints? Can we understand the role of interactions between different amino acid substitutions at different distances in a protein structure, and how substitutions at those positions affect the probability of convergence? Can we use convergence estimates over different lengths of time to better understand the rates of fluctuation in constraints both with and without substitution at a target site? The inclusion of variation in the substitution process across sites and over time—details that standard models currently lack—should be included in future evolutionary models to obtain more accurate descriptions of protein evolution.

Materials and Methods

Convergence calculations on mitochondrial proteins

Thirteen genes encoded in the mitochondrial genome were downloaded from GenBank for 641 tetrapod species. Separate alignments of amino acid sequences for every gene were made using ClustalX (Larkin et al. 2007). Aligning a selection of the sequences using PRANK (Löytynoja and Goldman 2005) yielded similar downward-sloping results, although there are differences in the height of the early curve (supplementary fig. S7, Supplementary Material online). The mutation pattern in genes across the mitochondrial genome has a complex pattern of changing asymmetry (Krishnan, Raina, et al. 2004; Krishnan, Seligmann, et al. 2004) that is not embodied in current phylogenetic reconstruction programs. We therefore made our phylogeny using only cytochrome oxidase 1 (CO1), which has the least asymmetric mutation rates among vertebrate mitochondrial genes (Krishnan, Raina, et al. 2004; Krishnan, Seligmann, et al. 2004). We partitioned the CO1 data by codon positions and determined the preferred model for the three data partition using the Akaike Information Criterion (Akaike 1973, 1983) in MrModeltest v2.2 (Nylander 2004). The Bayesian consensus tree was determined using the model for each partition (integrating over model parameters) and MrBayes 3.0b4 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003).

The alignments for all genes were concatenated and taxa with a large number of gaps ($>$ 500 of 3,596 sites) were removed, leaving 629 taxa. PLEX (de Koning et al. 2012) was used to infer the ancestral sequences and substitutions along the (maximum likelihood) CO1 tree. A fixed amino acid substitution model mtMam (Yang et al. 1998) was used along with site rate variation with five gamma distributed rates. PLEX analyses were run for 400,000 Markov chain Monte Carlo generations after 100,000 generations of burn-in. All branch pairs except sister branches and branch pairs where one branch was the ancestor of the other were considered. For the significance calculations (supplementary Materials, Supplementary Material online), we sampled the complete set of ancestral node states on the phylogenetic tree every 100 generations. Double substitutions for each branch pair were determined by finding all sites that changed between ancestor and descendant on both branches in the pair in that generation. Double substitutions that ended at the same descendant amino acid in both branches were counted as convergent events, while the remaining double substitutions that ended at different descendant amino acids were counted as divergent events. For simplicity of display, for figures 1–4 the average *C*/*D* ratios were calculated using only sites with $>$ 90% posterior probability of having a substitution, and branch pairs were included in the average only if *D* was greater than a specified cutoff (see figure legends). The number of inferred substitutions along the tree were counted to classify sites as conserved (60 or fewer substitutions) or fast-evolving sites (75 or more substitutions). We estimated how well we could infer the *C*/*D* ratios using this method by simulating sequences evolving over the mitochondrial tree and then comparing the *C*/*D* ratios from the known ancestors with the inferred ancestors. The results are shown in supplementary figure S8, Supplementary Material online.

The Stokes–Fisher Model

The SF model used to simulate protein evolution in this study has been described previously (Williams et al. 2006; Goldstein 2011). It is based on modeling the evolutionary process where the fitness of the protein is the probability that the protein would be folded in a particular "native" structure under equilibrium conditions.

The free energy $G(S, C_k)$ of a protein sequence S in a particular conformation C_k was calculated based on the sum of pair-wise energies between amino acids that are in contact in that conformation (i.e., have their C_β atoms closer than 7 Å),

using the contact potentials determined by Miyazawa and Jernigan (1985) based on their analysis of protein structures. To calculate the free energy of folding $\Delta G_{\text{Fold}}(S)$, we calculated $G_{\text{NS}}(S)$, the free energy for the native state (the conformation of the 300-residue purple acid phosphatase, PDB 1QHW, Lindqvist et al. 1999) as well as a large ensemble of alternative folds. We assumed that the distribution of the free energies of the large ensemble of thermodynamically relevant unfolded and alternative conformations can be represented by a Gaussian distribution with sequence-dependent average $\overline{G}(S)$ and variance $\sigma(S)^2$, which we estimated by calculating the average free energy and variance of the free energies of the sequence in the conformation of the first 300 residues of 55 different structurally diverse protein structures. Assuming that a large set (10^{160}) of possible unfolded structures with free energies are drawn from that distribution, we can then calculate $\Delta G_{\text{Fold}}(S)$ and therefore the probability $P_{\text{Fold}}(S)$ that the protein would be folded at equilibrium. As in previous work, we considered the fitness of a sequence $\omega(S)$ to equal the probability that it folded to the native state.

For the star phylogeny simulations, we initialized a protein sequence by choosing 300 codons at random (ignoring stop codons), using the standard genetic code to determine the encoded amino acids. We then computed the codon substitution model at each site in the protein at each point in time. The mutation rate to all possible alternative codons Ω_{ij} was constructed using the K80 nucleotide model ($\kappa = 2$) (Kimura 1980), disallowing multiple nucleotide changes. For each nonsynonymous mutation, ω' of the resulting sequence was computed based on the value of $\Delta G_{\text{Fold}}(S')$, the free energy of folding for this sequence, and the corresponding folding probability $P_{\text{Fold}}(S')$. This fitness was then compared with the fitness of the premutated sequence ω; the mutation rate was multiplied by the acceptance probability calculated using the Kimura formula for diploid organisms (Kimura 1957, 1962; Crow and Kimura 1970):

$$Q_{ij} = \Omega_{ij} \frac{1 - \exp(-2s)}{1 - \exp(-4N_{\text{Eff}}s)} \quad (4)$$

where $s = \frac{\omega' - \omega}{\omega}$, with N_e, the effective population size set equal to 10^6.

The simulation proceeded for a sufficient number of generations such that the stability of the protein reached equilibrium (i.e., the average fitness was approximately constant over time and across independent runs). Equilibrium was reached due to mutation–selection balance, the point where stabilizing mutations are relatively uncommon and have smaller relative fitness benefits, while destabilizing (but marginally acceptable) mutations are greater in number.

For the star phylogeny simulations, 100 replicate sequences were evolved to approximate equilibrium and then split into 10 lineages diverging from one another to produce sequences related by a star phylogeny. Each lineage was evolved for a distance of 10.0 synonymous nucleotide substitutions per nucleotide site from the common ancestor (on average, 6.95 amino acid replacements per amino acid position). We estimated the expected *C*/*D* ratios from these data.

To calculate the expected *C*/*D* ratios, we considered the instantaneous codon–codon substitution rate matrices given the constraints at each site in the two proteins and the current codons at this site. We then calculated the rate at which a double transition to the same amino acid would be observed in both lineages, compared with the rate at which a double transition to different amino acids would be observed. *C* and *D* were summed over all sites, with the ratio of these quantities computed for that pair of proteins. We then averaged *C*/*D* over all pairs of proteins in each star phylogeny, and over all star phylogenies. The details are provided in the Supplementary Material, Supplementary Material online. The observed *C*/*D* ratios found in figure 4*B* were obtained from simulating SF over the mitochondrial tree and then inferring the ancestors and *C*/*D* ratios using the same method as for the mitochondrial data.

Phenomenological Substitution Models

We also considered the expected *C*/*D* ratio for a variety of phenomenological substitution models, as more fully described in the supplementary Material, Supplementary Material online. We again considered a site in two homologous proteins *i* and *j*. We then calculated the probability that every pair of amino acids (or codons) would be observed in proteins *i* and *j*. We then used the substitution model to calculate the rate at which these amino acids would undergo a double substitution to the same or different amino acids (or codons coding for the same or different amino acids). *C* and *D* were calculated by summing over all possible amino acids (or codons) for sequences *k*, *i*, and *j*. When a gamma distributed rate distribution was used, we also summed *C* and *D* over four different rate categories. The ratio then yielded the *C*/*D* ratio.

Supplementary Material

Supplementary figures S1–S8 and table S1, and material are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Todd Castoe and Jason de Koning for contributions to the alignments and tree reconstruction, and Nicolas Lartillot for providing us with the parameters for the CAT-60 model. Richard A. Goldstein and David D. Pollock designed the research, analyzed the data, and wrote the paper. Richard A. Goldstein, Seena D. Shah, Stephen T. Pollard, and David D. Pollock performed the research. This work was supported by the National Institutes of Health (grant number R01 GM083127) and the Medical Research Council (MRC) UK.

References

Akaike H, editor. 1973. Second international symposium on information theory. Budapest (Hungary): Akademia Kiado.

Akaike H. 1983. Information measures and model selection. *Bull Int Stat Inst.* 50:277–291.

Ashenberg O, Gong LI, Bloom JD. 2013. Mutational effects on stability are largely conserved during protein evolution. *Proc Natl Acad Sci U S A.* 110:21071–21076.

Blackburne BP, Hay AJ, Goldstein RA. 2008. Changing selective pressure during antigenic changes in human influenza H3. *PLoS Pathog.* 4: e1000058.
Castoe TA, Jiang ZJ, Gu W, Wang ZO, Pollock DD. 2008. Adaptive evolution and functional redesign of core metabolic proteins in snakes. *PLoS One* 3:e2201.
Castoe TA, de Koning APJ, Kim H-M, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc Natl Acad Sci U S A.* 106:8986–8991.
Crow JF, Kimura M. 1970. An introduction to population genetics theory. New York: Prentice Hall.
de Koning AP, Gu W, Castoe TA, Pollock DD. 2012. Phylogenetics, likelihood, evolution and complexity. *Bioinformatics* 28:2989–2990.
Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.
Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11:725–736.
Goldstein RA. 2011. The evolution and evolutionary consequences of marginal thermostability in proteins. *Proteins* 79:1396–1407.
Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol.* 15:910–917.
Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.
Jukes T, Cantor C. 1969. Evolution of protein molecules. In: Munro H, editor. Mammalian protein metabolism, Vol. III. New York: Academic Press. p. 21–132.
Kimura M. 1957. Some problems of stochastic processes in genetics. *Ann Math Stat.* 28:882–901.
Kimura M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* 47:713–719.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol.* 16:111–120.
Kleinman CL, Rodrigue N, Lartillot N, Philippe H. 2010. Statistical potentials for improved structurally constrained evolutionary models. *Mol Biol Evol.* 27:1546–1560.
Koshi JM, Goldstein RA. 1998. Models of natural mutations including site heterogeneity. *Proteins* 32:289–295.
Koshi JM, Goldstein RA. 2001. Analyzing site heterogeneity during protein evolution. *Pac Symp Biocomput.* 191–202.
Krishnan NM, Raina SZ, Pollock DD. 2004. Analysis of among-site variation in substitution patterns. *Biol Proced Online.* 6:180–188.
Krishnan NM, Seligmann H, Raina SZ, Pollock DD. 2004. Detecting gradients of asymmetry in site-specific substitutions in mitochondrial genomes. *DNA Cell Biol.* 23:707–714.
Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol.* 21:1095–1109.
Lindqvist Y, Johansson E, Kaija H, Vihko P, Schneider G. 1999. Three-dimensional structure of a mammalian purple acid phosphatase at 2.2 A resolution with a mu-(hydr)oxo bridged di-iron center. *J Mol Biol.* 291:135–147.
Liu Y, Cotton JA, Shen B, Han X, Rossiter SJ, Zhang S. 2010. Convergent sequence evolution between echolocating bats and dolphins. *Curr Biol.* 20:R53–R54.
Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A.* 102:10557–10562.
Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 18:534–552.
Nylander JAA. 2004. MrModeltest v2. Program distributed by the author. Uppsala (Sweden): Evolutionary Biology Centre, Uppsala University.
Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* 502:228–231.
Pollock DD, Goldstein RA. 2014. Strong evidence for protein epistasis, weak evidence against it. *Proc Natl Acad Sci U S A.* 111:E1450.
Pollock DD, Taylor WR, Goldman N. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol.* 287:187–198.
Pollock DD, Thiltgen G, Goldstein RA. 2012. Amino acid coevolution induces an evolutionary Stokes shift. *Proc Natl Acad Sci U S A.* 109: E1352–E1359.
Quang LS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.
Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol.* 20:1692–1704.
Rokas A, Carroll SB. 2008. Frequent and widespread parallel evolution of protein sequences. *Mol Biol Evol.* 25:1943–1953.
Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
Shen YY, Liang L, Li GS, Murphy RW, Zhang YP. 2012. Parallel evolution of auditory genes for echolocation in bats and toothed whales. *PLoS Genet.* 8:e1002788.
Stewart CB, Schilling JW, Wilson AC. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* 330:401–404.
Tamuri AU, dos Reis M, Goldstein RA. 2011. Using site-wise mutation-selection models to estimate the distribution of selection coefficients from phylogenetic data. *Genetics* 190(3):1101–1115.
Tamuri AU, dos Reis M, Hay AJ, Goldstein RA. 2009. Identifying changes in selective constraints: host shifts in influenza. *PLoS Comput Biol.* 5: e1000564.
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol.* 18:691–699.
Williams PD, Pollock DD, Blackburne BP, Goldstein RA. 2006. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol.* 2:e69.
Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol.* 15:1600–1611.
Yokoyama KD, Pollock DD. 2012. SP transcription factor paralogs and DNA-binding sites coevolve and adaptively converge in mammals and birds. *Genome Biol Evol.* 4:1102–1117.