bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Molecular function limits divergent protein evolution on planetary timescales

Mariam M. Konaté[a,b*], Germán Plata[a*†], Jimin Park[a,c], Dinara R. Usmanova[a], Harris H. Wang[a,c], and Dennis Vitkup[a,d†]

[a] Department of Systems Biology, Columbia University, 1130 Saint Nicholas Ave., New York, NY, 10032, USA. [b] Division of Cancer Treatment and Diagnosis, National Cancer Institute, 31 Center Dr., Bethesda, MD, 20892, USA. [c] Department of Pathology and Cell Biology, Columbia University, 3960 Broadway, New York, NY, 10032, USA. [d] Department of Biomedical Informatics, Columbia University, 1130 Saint Nicholas Ave., New York, NY, 10032, USA.

[*]These authors contributed equally to this work.

[†]Correspondence to DV (dv2121@columbia.edu) or GP (gap2118@columbia.edu) :
1130 Saint Nicholas Ave. RM 803, New York, NY 10032
Tel: 212 851 5151

Abstract

Functional conservation is known to constrain protein evolution. Nevertheless, the long-term divergence patterns of proteins maintaining the same molecular function and the possible limits of this divergence have not been explored in detail. We investigate these fundamental questions by characterizing the divergence between ancient protein orthologs with conserved molecular function. Our results demonstrate that the decline of sequence and structural similarities between such orthologs significantly slows down after ~1-2 billion years of independent evolution. As a result, their sequence and structural similarities have not substantially decreased for the past billion years. The effective divergence limit (>25% sequence identity) is not primarily due to protein sites universally conserved in all linages. Instead, less than four amino acid types are accepted, on average, per site in orthologs strictly conserving their molecular function. Our analysis also reveals different divergence patterns for protein sites with experimentally determined small and large fitness effects of mutations.

Introduction

As proteins evolve from a common ancestor, their sequences and structures diverge from each other[1, 2]. Multiple previous studies have investigated the relationship between the conservation of protein molecular function, sequence identity[3-5] and structural similarity[1, 6]. For example, the likelihood that two proteins share the same molecular function, given their sequence[4] or structural[6] similarity, has been used to investigate the emergence of new protein functions [7, 8], and to perform

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

functional annotations of protein sequences[3, 6]. In this work, we focused on a different and currently unaddressed set of questions. Namely, how far can two sequences diverge while continuously maintaining the same molecular function? what are the temporal patterns of this divergence across billions of years of evolution? and how different protein sites contribute to the long-term divergence between orthologs with the same molecular function? We note that the requirement for the continuous conservation of molecular function is crucial in this context, as multiple examples of convergent evolution and protein engineering demonstrate that the same molecular function, such as catalysis of the same chemical reaction, can in principle be accomplished by proteins with unrelated sequences and different folds [9-11].

It was previously demonstrated that proteins with the same structural fold frequently diverge to very low (~10%) levels of sequence identity[12]. These results suggest that conservation of protein folds, i.e. the overall arrangement and topological connections of protein secondary structures [13], exerts relatively minor constraints on how far protein sequences can diverge. In contrast to protein folds, it is possible that conservation of specific molecular functions will significantly limit the long-term divergence of protein orthologs. While only a relatively small fraction of protein residues (~3-5%) are often directly involved in catalysis[14], recent analyses have demonstrated that even sites located far from catalytic residues may be significantly constrained in evolution. Because substitutions at these sites can have substantial effects on molecular function[15], it is likely that sequence constraints due to functional conservation extend far beyond catalytic residues.

In this study, we explored the long-term divergence patterns of protein orthologs by characterizing their pairwise sequence and structural similarity as a function of their divergence time. We used several models of molecular evolution to calculate the divergence rates, defined as the decrease in pairwise sequence identity or structural similarity per unit time, between orthologous proteins with the same molecular function. We also characterized the long-term divergence patterns at protein sites with different levels of evolutionary conservation, different locations in protein structures, and different experimentally measured fitness effects of amino acid substitutions. Finally, we explored how the limits of sequence and structural divergence after billions of years of evolution depend on the degree of functional conservation between orthologs.

Results

To study the evolution of proteins with the same molecular function, we initially focused our analysis on enzymes because their molecular function is usually well defined. The Enzyme Commission (EC) classifies enzymatic functions using a hierarchical 4-digit code[16], such that two enzymes that share all four EC digits catalyze the same biochemical reaction. We used protein sequences representing 64 EC numbers from 22 diverse model organisms across the three domains of life (Supplementary file 1). The considered activities include members of all 6 major enzyme classes: oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases.

To investigate whether the conservation of enzymatic function limits the divergence between orthologous sequences, we first calculated global pairwise sequence identities between orthologs as a

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

function of their divergence times (Figure 1, Figure 1- figure supplement 1). The pairwise divergence times reported in the literature[17] between the considered 22 species (Supplementary file 1) were used as a proxy for the divergence times between corresponding orthologous proteins. For each enzymatic activity, we constructed phylogenetic trees based on the orthologous protein sequences and compared them to the corresponding species' trees. Protein sequences showing clear differences in phylogenetic tree topologies, suggesting cases of horizontal gene transfer, were excluded from the analysis (see Methods).

We next considered two simple models of long-term protein evolution, one without a limit of sequence divergence and the other with an explicit divergence limit. The first model corresponds to sequence divergence with equal and independent substitution rates across all proteins sites [18, 19]; see Equation 1, where *y* represents global sequence identity, *t* represents divergence time, and R_0 represents the average substitution rate[18]. Under this model, back substitutions are not allowed, and sequence divergence slows down with time simply due to multiple substitutions at the same protein sites and progressively fewer non-mutated sites. The second model corresponds to sequence divergence where, in addition to sites with equal and independent substitution rates, there is a minimal fraction of identical sites at long divergence times; the fraction of identical sites is represented by Y_0 in Equation 2.

$$y = 100 * e^{-R_0 * t} \quad (1)$$

$$y = Y_0 + (100 - Y_0) * e^{-R_0 * t} \quad (2)$$

We applied the two models to fit the sequence divergence of each of the considered enzymatic functions. The best model fits for four representative metabolic activities are shown in Figure 1 (black for the first model and red for the second); the fits for the remaining metabolic activities are shown in Figure 1 – figure supplement 1. In 62 of the 64 cases, the second model fits the divergence data significantly better than the first model (F-test P-value <0.05, Supplementary file 2a). Moreover, in 95% of the cases (61/64) the maximum likelihood value of the parameter Y_0 is significantly higher (Wald test P-value <0.05) than the average sequence identity between random protein sequences based on their optimal global alignment (~13.5%, shown in Figure 1 and Figure 1 – figure supplement 1 by dashed black lines). The distribution of the fitted parameter Y_0 suggests a long-term sequence identity >25% (with average ~40%) between considered orthologs (Figure 2a); this demonstrates that conservation of a specific enzymatic function significantly limits long-term protein sequence divergence. Notably, model 2 is mathematically equivalent (see Methods) to a divergence model with equal substitution rates across sites, a limited number of amino acid types accepted per site, and allowed back substitutions [20-22]. In this model, parameter Y_0 represents the inverse of the effective number of acceptable amino acid types per site during protein evolution. Our results thus suggest that, on average, only 2 to 4 amino acids are acceptable per site for proteins that strictly conserve their molecular function (Figure 2a, top blue X axis); we note

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that this low average number does not imply that more than four amino acid types can never be observed at a given protein site[23].

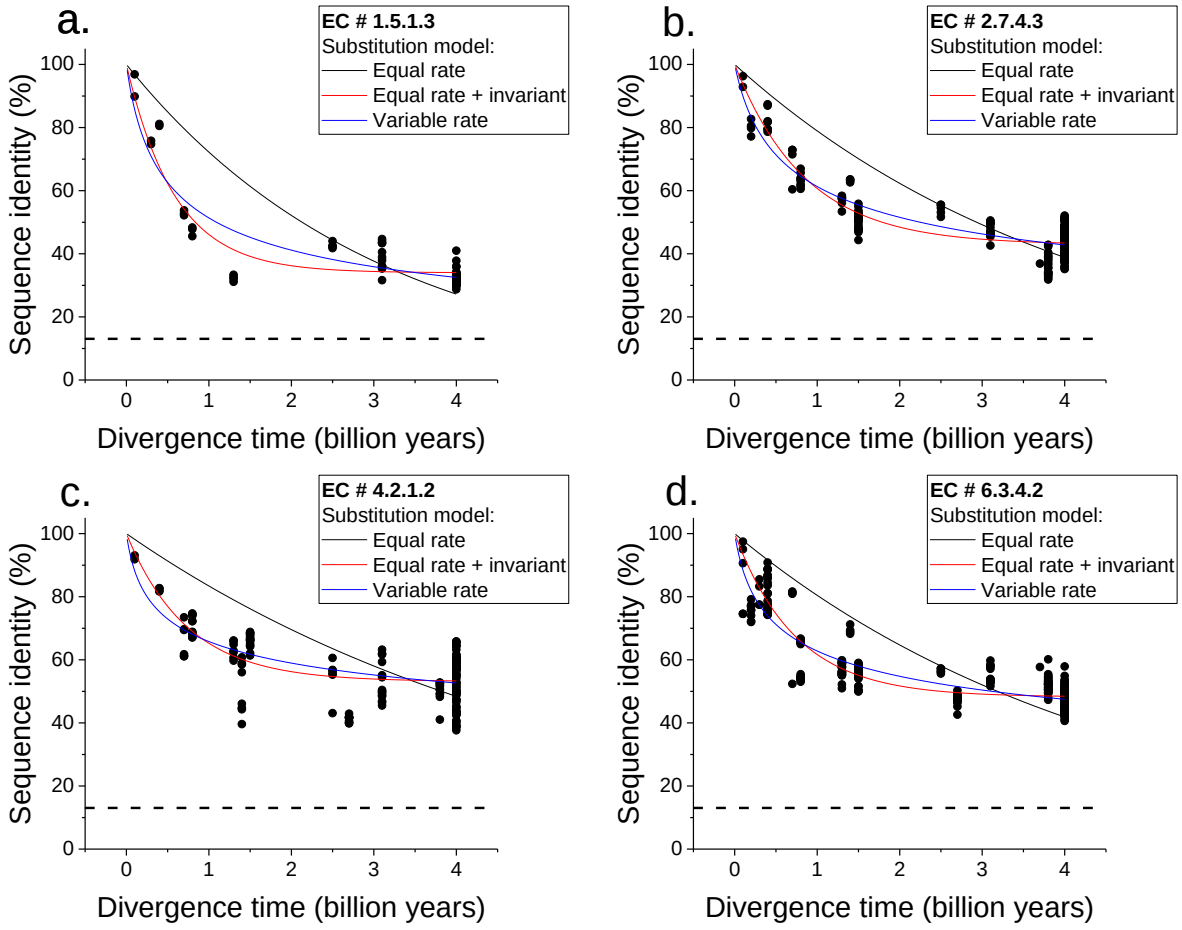

Figure 1. Sequence divergence of enzyme orthologs as a function of time. The global pairwise sequence identities between pairs of orthologous enzymes are shown as a function of divergence times between the corresponding species. Three models of amino acid substitution were used to fit the divergence data. Model 1 (black lines) assumes independent and equal substitution rates across all protein sites. Model 2 (red lines) assumes that a given fraction of protein sites remains identical at large divergence distances. Model 3 (blue lines) assumes a gamma distribution of substitution rates across sites. Best fits of the models are shown for 4 representative EC numbers: **a.** 1.5.1.3**, b.** 2.7.4.3**, c.** 4.2.1.2**, d.** 6.3.4.2. The horizontal dashed black lines represent the average sequence identity for the global alignment of unrelated protein sequences. The data and corresponding model fits for the other EC numbers considered in the analysis are given in Figure 1 – figure supplement 1 and Supplementary file 2a.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

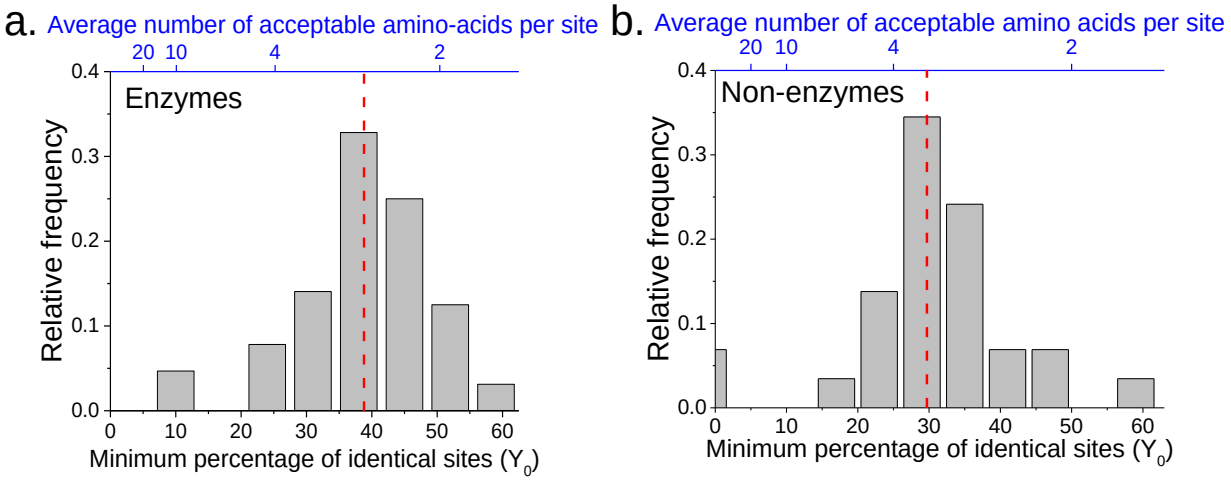

Figure 2. The limit of long-term protein sequence divergence between orthologous proteins. **a.** The distribution of Y_0 parameter values across 64 EC numbers for Model 2 fits (Equation 2). The Y_0 parameter represents the minimum percentage of protein sites that remain identical at long divergence times. The parameter Y_0 (considered as a fraction) can also be interpreted as the inverse of the average number of amino acids accepted per protein site during long-term protein evolution (top blue X axis). **b.** Similar to panel a, but for 29 protein families annotated with non-enzymatic functions. In panels a and b, the vertical red dashed lines represent the median values of the distributions (39% and 30%, respectively).

The two aforementioned models simplify the process of sequence divergence by considering the same substitution rates across protein sites. A more realistic and commonly used model of protein evolution assumes a gamma distribution[24] of substitution rates across sites; see Equation 3[25], where α represents the shape parameter of the gamma distribution. The best fits of such a variable-rate model (blue in Figure 1 and Figure 1 – figure supplement 1) show that the rates of protein sequence divergence between orthologous enzymes have decreased by more than 10 times during ~4 billion years of evolution (see Methods and Supplementary file 2b). Although the third model does not explicitly consider a long-term divergence limit, the obtained model fits also show that the vast majority of orthologous enzymes with the same function will remain above 25% sequence identity on the timescales when Earth environments will be hospitable to life (1-3 billion years from the present[26]) (Figure1 – figure supplement 2).

$$y = 100 * \left(\frac{R_0 * t}{\alpha} + 1 \right)^{-\alpha} \qquad (3)$$

The observed divergence limit is not due to an inability to detect remote protein homologs, as it occurs at relatively high sequence identities (Figure 1 and Figure 1 – figure supplement 1), for which corresponding orthologs can be easily identified by computational sequence comparison methods.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Furthermore, the results remained similar when we restricted the analysis to orthologous enzyme pairs with experimentally validated molecular functions (Figure 1 – figure supplement 3), based on publications referenced in the BRENDA database[27]. The results also remain robust towards the variance in the estimates of divergence times between considered species (see Methods). We note that the divergence limit between orthologs with the same molecular function does not imply that the rates of molecular substitutions decrease in evolution. It is also not simply due to the curvilinear relationship between time and sequence identity caused by multiple mutations at the same sites; specifically, the observed decrease in divergence rates is substantially higher (by >10 fold) than the one expected under model 1 simply due to multiple substitutions with equal probabilities at the same protein sites. Instead, the effective limit is reached when, due to a small number of amino acids accepted per protein site and back substitutions, additional amino acid replacements do not lead to a substantial further increase in protein sequence and structural divergence[28].

Interestingly, following the previously introduced metaphor of the expanding protein universe[2, 29], we can use the third model (Equation 3) to express the divergence rate between orthologs as a function of protein distance ($D = 1 - y$, where y is the fractional sequence identity ranging from 0 to 1), see Equation 4. This equation, similarly to Hubble's law of universe expansion[30], describes how the divergence rate depends on the distance between protein orthologs. According to our analysis, the divergence rate between orthologs decreases, on average, to only ~2% per billion years when their mutual sequence identity reaches 30% (corresponding to protein distance of 70%; Figure 1 – figure supplement 4).

$$\frac{\partial D}{\partial t} = R_0 * (1 - D)^{(\alpha+1)/\alpha} \qquad (4)$$

The analyses described above focused on the divergence of enzymes with the same molecular function. In order to investigate whether the observed divergence patterns are not specific to enzymes, we repeated the same analysis for non-enzymatic ancient orthologs (Figure 1 – figure supplement 5, Supplementary file 2c). The set of analyzed 29 protein families included ribosomal proteins, heat shock proteins, membrane transporters, and electron transfer flavoproteins (Supplementary file 2d). Using the same 22 species and this set of non-enzymatic orthologs, we found that model 2 fitted the data significantly better than model 1, and that the parameter Y_0 was >25% for the majority (23/29) of the protein families (Figure 2b, Supplementary file 2c). Interestingly, we also identified 19 additional orthologous groups showing two clearly different divergence patterns (Figure 1 – figure supplement 6), with pairs of eukaryotic orthologs diverging faster and farther than prokaryotic orthologs in the same protein family. The orthologous groups with this behavior included mitochondrial ribosomal proteins and initiation factors of mitochondrial translation (Supplementary file 2e). It has been previously postulated that mitochondrial ribosomal proteins diverged significantly faster in eukaryotes, compared to the divergence between their bacterial orthologs, due to compensatory protein substitutions following the accumulation of slightly deleterious substitutions in the mitochondrial ribosomal RNA[31].

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Having established, in the first half of the manuscript that conservation of molecular function significantly limits long-term sequence evolution, we investigated, in the second half, how different protein sites contribute to the observed divergence constraints. Specifically, whether the same protein sites are conserved between ancient orthologs in different phylogenetic lineages, how sites with different fitness effects of amino acid substitutions contribute to the divergence limit, and how structural locations of protein sites affect their long-term divergence patterns. We also explored how different levels of functional specificity constrain sequence and structural divergence.

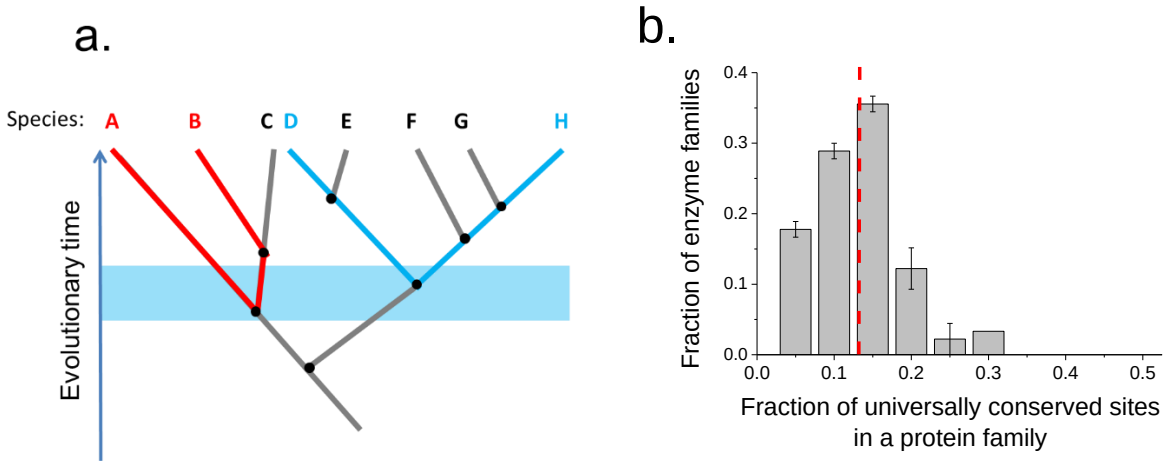

Figure 3. Conservation of protein sites in phylogenetically independent lineages. To identify the fractions of protein sites that are universally conserved — defined as sites that are identical in at least 90% of orthologs — we considered phylogenetically independent lineages. **a.** Illustration of pairs of species (e.g. A-B and D-H) representing phylogenetically independent lineages. In the figure, A-B and D-H are pairs of species that diverged within a certain time window (illustrated by the blue shaded region); the pairs do not share more recent edges in the phylogenetic tree. **b.** The distribution of the fraction of universally conserved sites across 30 enzymatic families. The analysis was performed using 30 enzymatic families for which at least 20 independent pairs of orthologs with the same function could be identified based on annotations in the KEGG database[32] (see Methods); pairs of orthologs were selected to have divergence times >2 billion years. Error bars represent the S.E.M. based on three replicates using different sets of orthologous pairs. The dashed red line indicates the median of the distribution (~13%).

To investigate whether the same protein sites are conserved between orthologs in different phylogenetic lineages, we aligned the sequences of ancient enzyme orthologs with the same molecular function (see Methods). We then quantified how often each protein site was occupied by identical amino acids across pairs of orthologs from phylogenetically independent linages (Figure 3 – figure supplement 1). Orthologous protein pairs from independent lineages were obtained from species pairs that do not share any edges in the phylogenetic tree[33] (Figure 3a); for example, in Figure 3a the pair D-H is independent of the pair A-B but not of the pair E-F. We performed the above analysis using 30 enzymatic activities for which at least 20 independent pairs of orthologs with the same function could be identified

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

based on annotations in the KEGG database[32] (see Methods). The results demonstrated that only a relatively small fraction of protein sites (10-20%) are universally conserved, i.e. they are identical in a majority (>90%) of independent lineages (Figure 3b). Therefore, the observed long-term divergence limit between orthologs is not primarily due to sets of universally conserved protein sites; instead, different sites usually contribute to the limit in independent phylogenetic lineages. By comparing the fractions of universally conserved sites to the average sequence identity between distant orthologs (~40%, Figure 2a) we found that, on average, these sites account for only ~35% of the observed sequence identity at long divergence distances. The analysis also revealed that different protein families show different probability distributions of identical sites (Figure 3 – figure supplement 1). This is likely a consequence of diverse structural and functional requirements across protein families, leading to protein-family specific constraints on protein sites.

We next investigated the long-term divergence patterns at protein sites with different fitness effects of amino acid substitutions. To that end, we experimentally measured the fitness effects of all possible single amino acid substitutions in a representative enzyme, the *Escherichia coli* dihydrofolate reductase (FolA, EC 1.5.1.3). We selected FolA for the experiments due to its small size (159 amino acids) and essential role in the *E. coli* metabolism[34]; also, the long-term protein sequence identity between FolA orthologs (~32%, see Figure 1a) is similar to other analyzed enzymes (Figure 2a). Following a recently described strategy[35], we used the Multiplex Automated Genome Engineering (MAGE) approach[36] to introduce every possible amino acid substitution at each FolA site in *E. coli*. To evaluate the fitness effects of protein substitutions we measured the relative growth rate of strains containing each protein variant compared to the "wild type" (WT) strain into which substitutions were introduced. Relative growth rates were measured in parallel by performing growth competition experiments between the pooled mutants. Amplicon sequencing of the *folA* gene was then used to measure the relative changes of mutant and WT abundances as a function of time (see Methods, Supplementary file 3).

Using the MAGE growth measurements in *E. coli*, we investigated the patterns of long-term sequence divergence at protein sites with different fitness effects of amino acid substitutions. Specifically, we sorted FolA protein sites into several groups according to their experimentally measured average fitness effects (Figure 4 – figure supplement 1), and explored the divergence of sequence identity for sites within each fitness group (Figure 4a, different colors). We evaluated sequence identity between FolA orthologs across divergence times using all pairwise comparisons between ~300 orthologous sequences from the COG database[37]. Although, as expected, sites with stronger fitness effects diverged more slowly, our analysis revealed interesting differences in temporal divergence patterns for sites with small and large fitness effects. For sites in the least deleterious fitness group (Figure 4a, blue) we observed, similar to the global sequence identity, a substantial decrease (~10-fold, see equation 5 in Methods) in mutual divergence rates after ~1.5 billion years of evolution. In contrast, sites with the most deleterious mutations (Figure 4a, black) displayed a much slower, but approximately constant average divergence rate throughout evolutionary history. This pattern suggests that, in contrast to divergence at sites with small fitness effects, the divergence at sites with large effects is not close to saturation. Furthermore, even for FolA sites with mild fitness effects, sequence identity remains above 25% at long divergence times.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

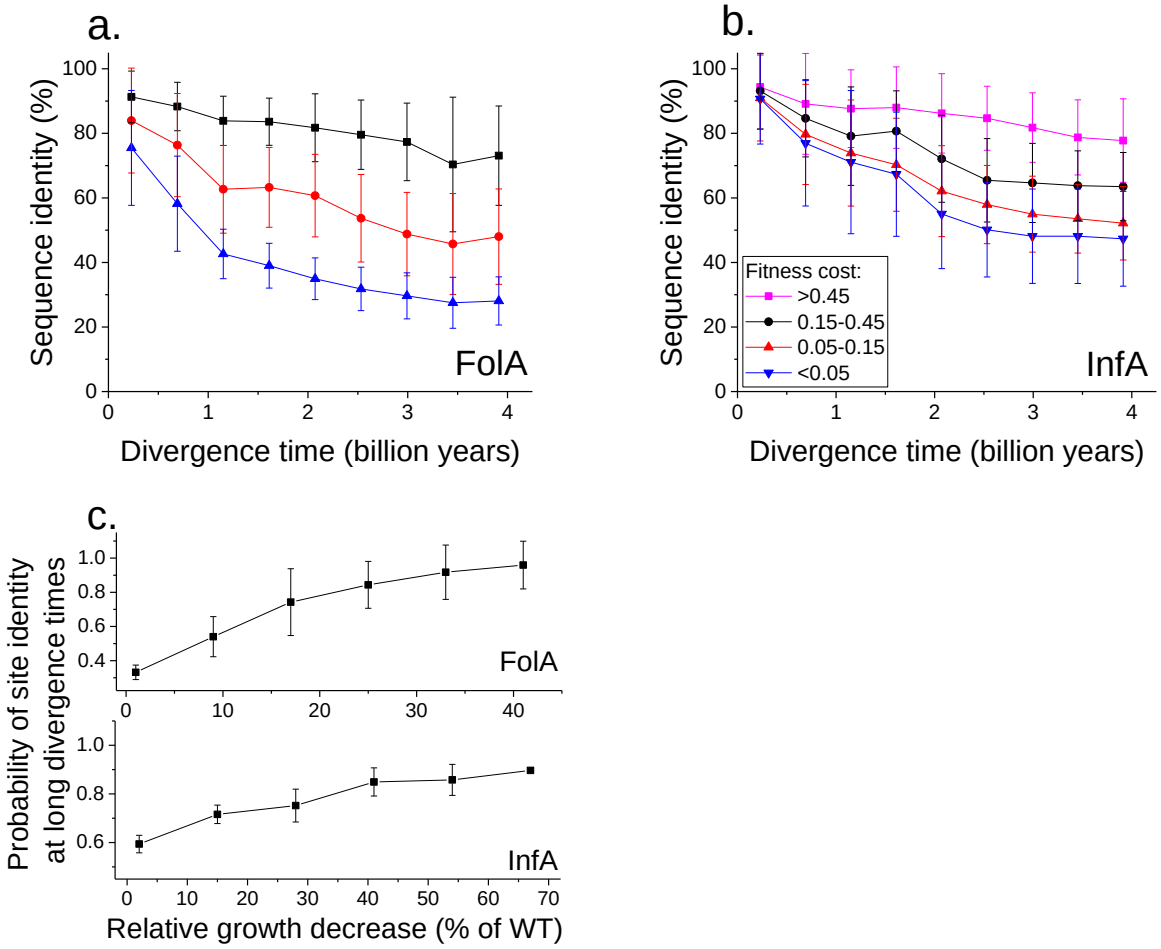

Figure 4. Sequence divergence of protein sites with different experimentally measured mutant fitness effects in *E. coli*. a. Sequence identity between pairs of FolA orthologs as a function of divergence time. Divergence at protein sites with different average fitness effects in *E. coli* are shown using different colors. The average sequence identities were calculated using bacterial FolA orthologs in the COG database[37]; divergence times were estimated using bacterial 16S rRNA sequences (see Methods). Error bars represent the S.D. of sequence identity in each bin. **b.** Similar to panel **a**, but for the sequence divergence between pairs of orthologs of *E. coli* translation initiation factor InfA. **c.** The probability that protein sites in FolA (upper panel) or InfA (lower panel) are occupied by identical amino acids as a function of the average mutant fitness at the corresponding sites in *E. coli*. The probability represents the fraction of phylogenetically independent pairs of orthologs in which sites are identical at long divergence times (2±0.25 billion years for FolA, and 2.5±0.25 billion years for InfA). Error bars represent the S.E.M. across sites.

To assess the generality of the FolA results we used a dataset[35], obtained using MAGE, of fitness values for all possible amino acids substitutions in the *E. coli* translation initiation factor InfA (Figure 4b). Consistent with the relatively higher level of sequence conservation of InfA, we observed lower average mutant growth rates and lower rates of sequence divergence in each fitness group. Nevertheless, the

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

long-term divergence patterns were qualitatively similar between the two proteins. For sites in the least deleterious InfA fitness group (Figure 4b, blue), we observed a substantial decrease in the divergence rate after ~2 billon years of evolution. In contrast, sites with strongest fitness effects (Figure 4b, pink) displayed a slower but approximately constant divergence rate.

Because the fitness effects of substitutions at a protein site may change in evolution[38, 39], it is interesting to investigate how fitness effects measured in one species, such as *E. coli*, correlate with the site conservation at the divergence limit. To explore this question, we calculated the probability that a protein site is occupied at large evolutionary distances (~2 billion years for FolA and ~2.5 billion years for InfA) by the same amino acid in phylogenetically independent lineages (Figure 3a). We then investigated how this probability changes as a function of the average fitness effects of substitutions at the site measured in *E. coli* (Figure 4c). For both FolA and InfA, the probability that a protein site is identical, and thus contributes to the observed divergence limit, first increases linearly with increasing average fitness effects, and then begins to saturate for sites with large (>30% growth decrease) fitness effects. Thus, the fitness effects at a protein site correlate with the site's conservation even after billions of years of evolution, at the divergence limit.

The sequence constraints revealed by our analysis are likely due to the conservation of corresponding protein structures required for efficient catalysis and molecular function[6, 40]. Therefore, in addition to sequence divergence, it is also interesting to investigate the long-term structural divergence of orthologous proteins with the same function. For this analysis we used >1000 orthologous pairs of enzymes (sharing all 4 EC digits) with known 3D structures in the PDB database[41] (see Methods); the orthologous enzymes were aligned using the TM-align algorithm[42]. The average root mean square deviation (RMSD) between C-alpha atoms of the orthologous enzymes significantly increases (Spearman's r=0.44, P-value<1e-20) with divergence time between the corresponding species (Figure 5a). Nevertheless, the C-alpha RMSD rarely diverged beyond 3 Å, even at long evolutionary distances. Consistent with sequence evolution (Figure 1), we also observed a substantial decrease in the rate of structural divergence after ~1.5 billion years of divergent evolution.

Only a small fraction of all enzyme residues forms an active site and directly participates in catalysis. Therefore, we investigated next how the sequence divergence depends on the spatial proximity of protein positions to active site residues. It was recently demonstrated that evolutionary rates of amino acid substitutions correlate with protein sites' spatial distance to catalytic residues[43]. The main goal of our analysis was different, i.e. to investigate the temporal patterns of the long-term divergence, and the effective divergence limit for sites at various distances to the active site. We considered catalytic site annotations available from the Protein Data Bank[41], UniProt-KB[44] and the Catalytic Site Atlas[45] and quantified the average divergence of sequence identity at various distances from catalytic residues (see Methods, Figure 5b). We based this analysis on the same set of enzymatic activities used to study global sequence divergence (Figure 1 and Figure 1 – figure supplement 1). Although, as expected, residues close to the active site were the most highly conserved[43, 46], even distant residues displayed an effective divergence limit at long evolutionary distances. This result suggests that the spatial constraints required to conserve specific molecular function usually propagate throughout the entire protein structure and significantly limit the long-term divergence even at sites distant from catalytic residues.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

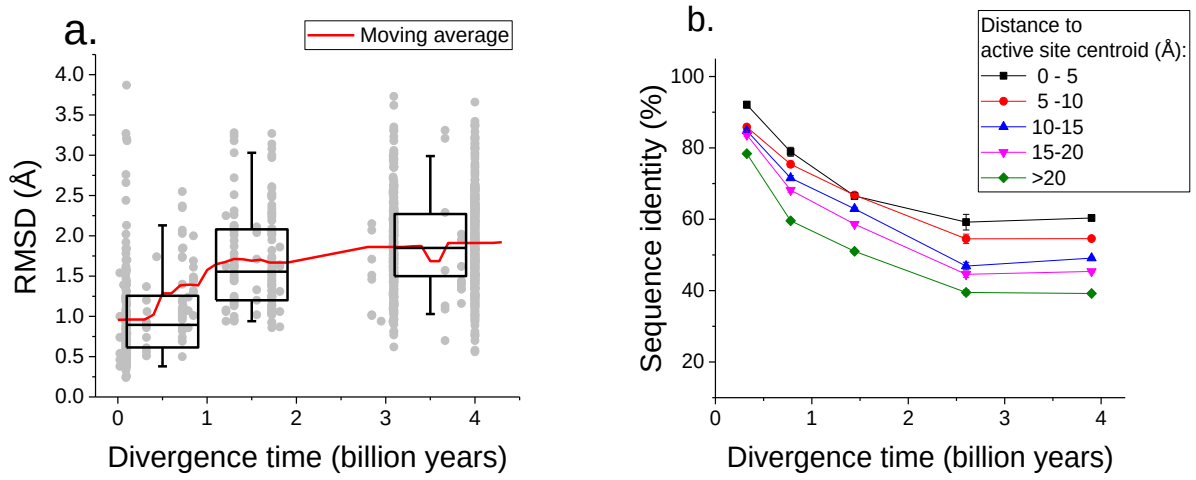

Figure 5. Long-term structural evolution of proteins with the same molecular function. a. The pairwise C-alpha root mean square deviation (RMSD) as a function of the divergence time between pairs of orthologs (shown by gray dots) annotated with the same EC number. RMSD values were calculated based on structural alignments using the TM-align algorithm[42]. Boxes indicate the median and 25-75 RMSD percentiles for the corresponding divergence times, the vertical lines indicate the 5-95 percentiles, and the red line shows the moving average of the data. **b.** Long-term divergence of sequence identity of protein sites located at different distances to enzymes' active sites. In this analysis we considered the same species and enzymatic activities used to explore the global sequence divergence (Figure 1 and Figure 1 – figure supplement 1); the average sequence identities within each distance shell (shown using different colors) were calculated across all pairs of orthologs annotated with the same EC number (see Methods). Error bars represent the S.E.M. across ortholog pairs.

Finally, we investigated how various degrees of functional conservation affect the long-term divergence between orthologs. To that end, we compared the long-term sequence and structural similarities of enzymes sharing their full EC classification to those sharing only the first three digits of their EC classification (Figure 6a, 6b); for this analysis we only used orthologs from species with divergence times >2 billion years (see Methods). In contrast to enzymes sharing all four EC digits, conservation of the first three digits indicates only a general class of substrates or cofactors[16]. This comparison revealed significantly lower sequence identities (27% vs. 37% identity, Mann-Whitney P-value $<10^{-20}$) and structural similarities (2.4 vs. 1.8 Å RMSD, P-value $2x10^{-18}$) between orthologs sharing only partial EC numbers. Notably, orthologs sharing only the first three EC digits are still substantially more conserved, both in sequence and structure (P-values $<10^{-20}$), than pairs of enzymes with the same structural fold but completely different enzyme classification (i.e. sharing no digits in the EC classification) [47].

We also investigated the sequence constraints at the same level of protein structural divergence for protein with different degrees of functional conservation. To that end, we calculated the sequence identity between orthologs, sharing either their full or partial EC numbers, at different bins of long-term structural similarity (Figure 6c). Interestingly, we observed that even at the same level of C-alpha RMSD

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

divergence, orthologs sharing full EC numbers usually have higher levels of sequence identity compared to orthologous pairs with the same level of structural divergence but sharing only three EC digits. This result indicates that functional conservation constrains sequence divergence even beyond the requirement to maintain a specific spatial structure.

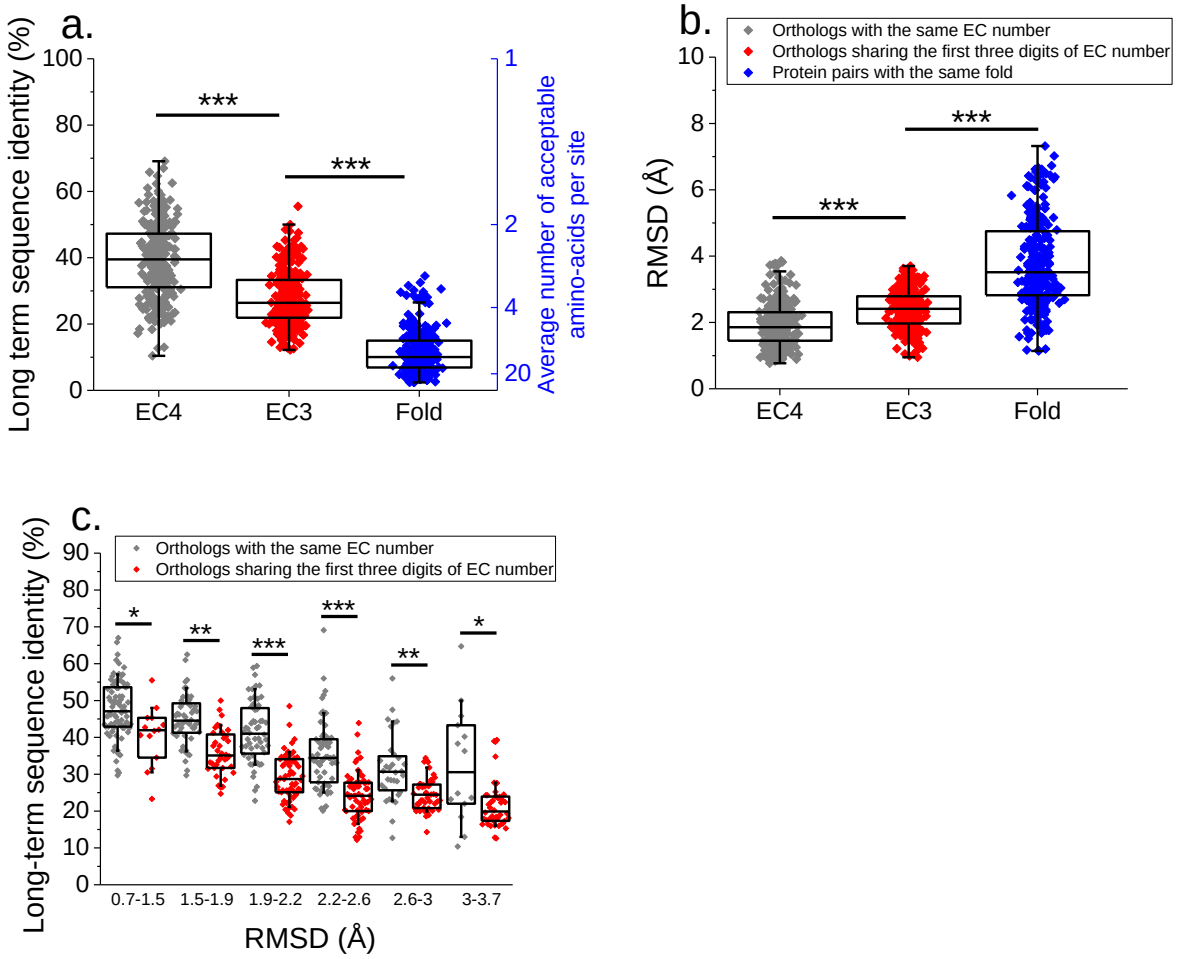

Figure 6. Effect of functional specificity on long-term sequence and structural similarity between orthologs. a. Sequence identities between orthologous pairs of enzymes from species that diverged over two billion years ago. Results are shown for pairs of orthologs sharing the same EC number (gray, n=272), and only sharing the first three digits of their EC numbers (red, n=265), i.e. enzymes conserving only a general class of substrates or cofactors. The results are based on enzyme COGs for the 22 species used to analyze global sequence divergence (Supplementary file 1). Blue points show the sequence identity between pairs of proteins with the same structural fold[47] but sharing no digits in the EC classification (n=298, see methods). The blue Y axis represents the average number of amino acids accepted per protein site during long-term protein evolution. **b.** Similar to panel a, but showing the corresponding C-alpha structural divergence (RMSD) between protein pairs. **c.** Sequence identities between orthologous enzyme pairs at the same level of long-term structural similarity. Results are shown for pairs of enzymes sharing their full EC classification (gray), or only sharing the first three digits of their EC classification (red). In all panels: *(p<0.05), **(p<1e-4), ***(p<1e-10) for the Mann-Whitney test.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

Our analysis demonstrates that, in contrast to proteins with the same fold[12], the requirement to strictly conserve the same molecular function significantly limits the long-term sequence and structural divergence of protein orthologs. Although we confirmed the result by Povolotskaya *et al.* [2] that ancient protein orthologs are still diverging from each other, our study reveals that the rate of this divergence becomes increasingly slow for orthologs that strictly conserve their function. Even a slight relaxation of functional specificity, for example from full to partial EC conservation (Figure 6a, 6b), leads to substantially more pronounced long-term sequence and structural divergence. Similarly, a significant sequence identity between homologous restriction endonucleases is usually limited to isoschizomers, i.e. proteins specific to the same target DNA sequence[48].

We believe that the observed divergence patterns can be explained by the following mechanistic model. Proteins with the same molecular function usually conserve the identity of their chemical and biological substrates and interaction partners. This conservation leads to functional pressure to closely preserve the spatial positions and dynamics of key protein residues necessary for efficient catalysis and function[14]. In turn, the requirement to continuously preserve structural properties and functional dynamics of key protein sites likely imposes a strict conservation on the overall protein structure, i.e. structural optimality is necessary for protein function. We note that the observed conservation may reflect the impact of amino acid substitutions on protein activity, due to changes in the identity, equilibrium positions and dynamics of protein residues, and on protein abundance, due to changes in overall protein stability[49, 50]. Nevertheless, direct and comprehensive biochemical experiments demonstrated that the deleterious effects of mutations primarily arise from changes in specific protein activity rather than decreases in protein stability and cellular abundance[15]. Our results are consistent with this model, demonstrating that conservation of functional specificity imposes substantially more stringent long-term sequence constraints than conservation of protein folds, and thus protein stability. The preservation of structural optimality (<3Å C-alpha RMSD) required for a given molecular function leads, in agreement with the results by Chothia and Lesk[1] and others[21], to substantial levels of sequence conservation and the observed divergence limit.

The presented results demonstrate that only about a third of the sequence conservation between distant orthologs with the same molecular function can be attributed to universally conserved protein sites, i.e. sites occupied by identical amino acids in almost all lineages. We observe that different protein sites are usually identical between orthologs from different lineages. This result is likely due, at least in part, to the epistatic nature of protein sequence landscapes, where mutations that are neutral in one lineage are often prohibitively deleterious in another[23, 38]. In the context of the aforementioned divergence model, the evolution of mitochondrial ribosomal proteins in eukaryotes (Figure 1 – figure supplement 6) provides an interesting example, suggesting that orthologs' divergence can be substantially accelerated by co-evolution with their interaction partners or relaxation of selection pressures.

Our experimental and computational analyses also delineate two distinct stages of the long-term divergence of orthologs with the same molecular function. During the first 1-2 billion years of divergence, substitutions at protein sites with mild fitness effects lead to a substantial (40-60%) decrease in sequence

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

identity. After the first stage, divergence at these sites effectively saturates. The saturation is due to the fact that less than four amino acid types, on average, are accepted per site for proteins strictly conserving their molecular function. The saturation at sites with small fitness effects, combined with very slow divergence rate at sites with large fitness effects (Figure 4), leads to a substantially slower sequence and structural divergence during the second stage. Interestingly, as a consequence of this slowdown, for the past billion years there has not been a substantial decrease in sequence and structural similarity between ancient orthologs with the same molecular function. Further analyses of biochemical, biophysical and cellular constraints will reveal how various structural and functional properties influence proteins' long-term evolution, and how protein functional efficiency may be compromised by deleterious mutations[51].

Methods

Key resources table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
strain, strain background (*Escherichia coli* EcNR2)	MG1655, bla, bio-, λ-Red+, mutS-::cmR	PMID: 19633652	Addgene #26931	
sequence-based reagent	90bp DNA oligos with phosphorothioated bases	This paper	See Supplementary file 4	100 nmole DNA Plate oligo, Integrated DNA Technologies
commercial assay or kit	Miseq Reagent Kit V2	Illumina	MS-102-2002	
commercial assay or kit	sybr green	ThermoFisher	S7567	
commercial assay or kit	Qubit HS DNA kit	ThermoFisher	Q32854	
commercial assay or kit	Q5 Hot Start High-Fidelity Mastermix	NEB	M0494S	
commercial assay or kit	DNA clean and concentration kit 5	Zymo Research	D4013	
commercial assay or kit	illustra bacteria genomicPrep Mini Spin kit	GE life sciences	28904259	
commercial assay or kit	Agilent DNA 1000 kit	Agilent Genomics	5067-1504	
software, algorithm	SeqPrep v1.1	John St. John	https://github.com/jstjohn/SeqPrep	
software, algorithm	Bowtie2	PMID: 22388286		
software, algorithm	Perl scripts to count mutant reads	This paper	https://github.com/platyias/count-MAGE-seq	
other	Turbidostat for growth competition assay	PMID: 23429717		

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Considered enzyme activities and corresponding protein orthologs.

We selected for analysis the sequences annotated in UniProt[44] with EC numbers associated with the following metabolic pathways (defined in the KEGG database[32]): Glycolysis and gluconeogenesis, pentose phosphate pathway, TCA cycle, purine metabolism, pyrimidine metabolism. Using the protein sequences from 22 diverse organisms (Supplementary file 1) we constructed clusters of orthologous groups (COGs) using the EdgeSearch algorithm[52]. Following previous studies, we considered any two proteins from different species in the same COG as orthologs[53]. COGs were obtained using the COGsoft software[52], starting from an all-against-all psi-blast[54] search, setting the database size at 10^8, and using a maximum considered E-value of 0.1. To obtain the largest number of likely orthologs we did not apply a filter on low complexity or composition-based statistics. Only proteins sharing the same EC number and assigned to the same COG were compared, and only COGs with sequences in 10 or more of the 22 species were used.

In order to exclude proteins clearly showing evidence of Horizontal Gene Transfer (HGT), we constructed a maximum likelihood phylogenetic tree of the 12 prokaryotes considered in our analysis using a concatenated alignment of marker genes[55]. The species tree was then manually compared to the individual trees of the prokaryotic sequences sharing the same molecular function within each COG; COG-specific trees were built using the GAMMA model of amino-acid substitution implemented in the RAxML software[56]. Proteins that showed clear differences in tree topologies, suggesting HGT, were excluded from further analysis. Ancient gene duplications, i.e. duplications occurring prior to the divergence between considered species, often lead to cases in which enzymes in the same COG but from different species have diverged for longer than the corresponding species' divergence times; thus, we did not consider COGs with tree topologies showing evidence of ancient gene duplications. Ancient gene duplications were defined as those occurring prior to the last common ancestor of 3 or more of the 22 species considered in the analysis.

The same procedure was used to select non-enzymatic COGs for analyses (Figure 1 – figure supplement 5). However, in this case we only considered COGs for which none of the proteins were annotated in UniProt with metabolic EC numbers. Naturally, UniProt functional annotations for non-enzymes vary in terms of their source and format. Therefore, it is difficult to ascertain the degree of functional specificity and conservation between non-enzymatic orthologs. To address this, we manually checked that the molecular functions associated with proteins in the same COG were related, although we could not ascertain perfect conservation of molecular function.

Models of long-term protein sequence evolution.

Global sequence identities for pairs of proteins annotated with the same molecular function in the same COG were calculated using pairwise alignments with ClustalW2[57]. Sequence identity was computed as the number of identical sites at aligned positions, divided by the total number of aligned sites, i.e. excluding gaps. Divergence times between organisms were obtained from the TimeTree database[17] (November, 2015) and used as a proxy for protein divergence times; in the analysis we used the mean divergence times across studies listed in the database. Divergence times between bacteria and

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

archaea were set to 4 billion years based on current estimates for the occurrence time of their Last Common Ancestor[58, 59] and existing evidence of an early origin of life on Earth[60]. It is likely that ancient eukaryotic genes originated through episodic endosymbiotic gene transfer events and vertical inheritance from bacterial and archaeal genomes[61, 62]. Because of the discrete nature of such transfer events, the vast majority of individual prokaryotic-eukaryotic orthologous pairs are likely to have diverged from each other long before the origin of eukaryotes (1.8 billion years ago[63]); specifically, because most ancient prokaryotic species would not have transferred genes to eukaryotes. Thus, based on the median divergence time between the considered prokaryotes (~4 billion years, Supplementary file 1), divergence times between eukaryotes and prokaryotes were set in our analyses at 4 billion years. The results presented in the paper remain insensitive to the exact value of this divergence estimate (within the 3-4 billion year interval). Based on the recently proposed affiliation of eukaryotes and members from the Lokiarchaeota[64], divergence times between *S. solfataricus* and eukaryotes were set at 2.7 billion years, i.e. the estimated age of the TACK superphylum[65, 66].

In order to study the long-term divergence patterns of orthologs, we only used COGs containing pairs of orthologs with at least 5 different divergence times distributed across 4 billion years. Sequence divergence were fitted with models 1 to 3 using the least-squares minimization algorithm implemented in the MATLAB R2017a fitnlm function (The MathWorks, Inc., Natick, MA). The best fits of the model 1 and model 2 were compared using the F-test. To test whether the conservation of molecular function limits protein sequence divergence, the minimum sequence identity parameter in model 2 (Y_0, from equation 2) was compared, for each enzymatic activity, to the average global sequence identity between unrelated protein pairs using the Wald test.

To investigate the effect of the uncertainty of divergence times' estimates, we repeated the analysis of the 64 enzymatic activities while randomly assigning either the maximum or minimum value of the divergence times between lineages reported in the TimeTree database. This analysis was performed for a total of 1000 independent assignment runs. Across the independent assignment runs, the expected long-term sequence identity between orthologs was higher than 25% for at least 90% of enzymes (based on model 2), and the projected sequence identity after 7.8 billion years was above 25% (based on model 3) for at least 75% of enzymes (Figure 1 – figure supplement 7).

To assess the effect of computational functional annotations on the observed divergence results, we repeated the analysis using only sequences with experimentally validated molecular functions (Figure 1 – figure supplement 3). To keep only sequences with validated molecular functions, we manually reviewed published references for enzyme annotations in the BRENDA database[27], and discarded any functional assignments that were based exclusively on computational or high-throughput studies. After filtering for the experimentally validated annotations, we only considered EC numbers corresponding to pairs of orthologs with at least 4 different divergence times distributed across 4 billion years.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Calculation of the divergence rate.

Based on Model 3, we determined the divergence rate, i.e. the rate of the decrease in sequence identity per time, at a given divergence time *t* by solving for the derivative of Equation 3 with respect to time:

$$\frac{dy}{dt} = \frac{d\left(100*\left(\frac{R_0*t}{\alpha}+1\right)^{-\alpha}\right)}{dt} = -100 * R_0 \left(\frac{R_0*t}{\alpha}+1\right)^{-\alpha-1} \quad (5)$$

where *y* represents global sequence identity, *t* represents divergence time, R_0 represents the average substitution rate, and α represents the shape parameter of the gamma distribution.

Equivalency between model 2 and a Poisson divergence model with allowed back substitutions.

In the Jukes-Cantor model of nucleotide divergence [20, 67], the expected number of substitutions per site (δ) between two sequences after a divergence time *t* from a common ancestor is given by:

$$\delta = -\frac{a-1}{a} ln\left(1-\frac{a}{a-1}(1-y)\right) \quad (6)$$

where *y* is the proportion of identical sites and *a* is the number of allowed nucleotide types (usually 4). The same model can be applied to the divergence of protein sequences[22, 25], by setting *a* to the number of allowed amino acid types per protein site. Furthermore, $\delta = 2\lambda t$, where λ represents the substitution rate per site per unit time, which is assumed to be equal across all sites. Substituting δ, and solving the above equation for *y* yields:

$$y = \frac{1}{a} + \left(1-\frac{1}{a}\right)\exp\left(-\frac{2\lambda a}{a-1}t\right) \quad (7)$$

which is mathematically equivalent to model 2 (equation 2), with $R_0=\frac{2\lambda a}{a-1}$, and $Y_0 = \frac{1}{a}$. Thus, Y_0 can also be interpreted as the inverse of the average number of amino acids accepted per protein site during protein evolution.

FolA competition experiment in *E. coli*.

To perform competition experiments we used the EcNR2 strain derived from *E. coli* K12 MG1655. Mutagenesis was performed using Multiplex Automated Genomic Engineering (MAGE), as previously described[36]. 90 bp DNA oligomers were designed around each folA codon using the MG1655 wild type sequence as reference (Supplementary file 4). For each codon, all possible nucleotide variants were

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

synthesized. To avoid simultaneous mutations of multiple codons, cells were transformed targeting ten consecutive codons at a time. After four rounds of electroporation, cells were recovered and pooled together at approximately the same concentration based on cell counts. Two competition growth experiments were carried out, one for each half of the protein. For the competition experiments, cells were grown in LB media in a turbidostat while maintaining constant volume and cell density. Samples were taken every 2 hours for a period of 16 hours, spun down, washed in PBS, spun down again and stored at -20°C until all samples were collected. For each competition, the corresponding FolA region was amplified through PCR while assigning a specific DNA barcode for each time point. PCR products were then pooled and paired-end sequenced using the MiSeq Reagent Kit 2 from Illumina. Sequence reads were deposited to the SRA database with accession number: SRP152339.

To determine, at each time point, the abundance of each mutant relative to wild type, we joined paired-end reads using SeqPrep (v 1.1) and aligned the joined reads to the folA gene sequence using Bowtie2[68]. We then counted the number of reads per mutant using a custom script[69]. Reads with more than a single mutated codon were discarded. Counts were median-normalized to control for noise due to mutagenesis performed in batches of 10 codons. At each time point we calculated the ratio R_t of mutant to wild type (WT) reads. In exponential growth, the growth rate difference between a given mutant and WT was calculated based on the slope of $\ln(R_\tau)$ as a function of time:

$$\ln(R_t) = (m_i - m_{wt}) * t + \ln(R_0)$$

where m_i and m_{wt} represent the mutant and WT growth rates, respectively. Growth rate differences were calculated only for mutants with at least 5 time points with 20 or more reads. Relative growth rates were calculated by dividing the slopes in the equation above by the number of e-fold increases given the average dilution rate of the turbidostat (1.37/h).

To calculate a single value characterizing the effect of all possible mutations at a protein site, we first averaged the relative growth rates of mutants resulting in the same amino acid change. We then calculated the average fitness effect of mutations at each protein site by averaging across 20 possible amino acids substitutions (Supplementary file 3).

To estimate the sensitivity of our results to sequencing errors, we calculated the average fitness effect of substitutions at each FolA site using the relative growth rates of mutant strains carrying only 32 mutated codons selected at random out of 64 possible codons. We observed a high correlation (Pearson's r: 0.95, p-value < 1e-20, Figure 4 – figure supplement 2) between the average growth rate effects at each site calculated using two non-overlapping subsets of 32 codons. As expected, nonsense mutations and substitutions in the *folA* start codon had substantially stronger average effects on growth rates compared to other substitutions (26% versus 4% slower growth than WT, respectively. Mann Whitney U, p-value<10^{-20}). Also, the relative growth rates due to synonymous codon substitutions were usually very mild (0.2% higher growth compared to WT); 97% of synonymous substitutions had growth effects of less than 3%.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Contribution of different sites to the divergence limit.

In order to identify phylogenetically independent pairs of species, we aligned the 16S rRNA gene sequences of bacterial species corresponding to orthologs annotated with the target 30 EC numbers (Figure 3 – figure supplement 1). 16S rRNA sequences were obtained from the GreenGenes database[70] (October, 2016). We then built maximum likelihood phylogenetic trees based on the 16S alignments using RAxML[56]. Next, we used the Maximum Pairing Problem approach by Arnold *et al*.[33] to find the largest number of edge-disjoint pairs of species with 16S rRNA genetic distances corresponding to >2 billion years of divergence. Divergence times were estimated from the 16S genetic distances based on the linear regression of literature reported divergence times[17] (Supplementary file 1). The F84 model of nucleotide substitution implemented in the phylip package[71] was used to compute the genetic distances. Using the 16S alignment data, we calculated the probability that a protein site was identical across independent lineages. The probability was calculated as the fraction of orthologous pairs from phylogenetically independent species pairs with identical amino acid at the site. The amino acid identities at a given site were obtained based on the multiple sequence alignment of all orthologs associated with each EC number, obtained using ClustalW2[57]. A similar procedure was applied to analyze FolA and InfA orthologs from the COG database (Figure 3c).

To investigate the divergence of sites with different fitness effects, we used sequences of FolA and InfA bacterial orthologs from the COG database[37]. The FolA orthologs annotated with the same EC number in UniProt (n=311) and the InfA orthologs annotated with the same KEGG Orthology (KO) number in KEGG (n=514) were used to build multiple sequence alignments with ClustalW2[57]. Divergence times were estimated from the 16S genetic distances as described above. Within each divergence bin (Figure 4a, b), sequence identities of sites with different average fitness effects (represented by different colors in Figure 4a, b) were averaged across all pairs of orthologs at a given divergence time.

Analysis of global protein structural evolution.

To study the divergence of protein structures as a function of time, we obtained PDB codes for all proteins associated with EC numbers in the BRENDA database[27]. We then selected for the analysis species with experimentally solved enzyme structures for at least 10 different EC numbers. Psi-blast searches with a conservative E-value cutoff of 10^{-6} were used to identify orthologs (defined as bi-directional best hits) in the selected species. The 3D structures of orthologous pairs, annotated with the same EC number, were aligned using the TM-align program[42] to obtain the C-alpha RMSD values. Pairs of proteins were not considered if more than 70% of the residues of the shortest protein could not be structurally aligned. We also removed from the analysis pairs of structures with flexibility between domains, as they could result in large RMSD values despite significant structural similarity. To identify such proteins we used the FATCAT[72] software to perform flexible structural alignments of all structure pairs. We then filtered the structural pairs that were split into two or more domains by the FATCAT alignments.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Analysis of the enzyme active sites.

To analyze divergence as a function of active site distance we used protein sequences associated with the 64 EC numbers and 22 species considered in Figure 1 and Figure 1 – figure supplement 1. To that end, PDB[41] was searched for homologous sequences annotated with the same enzymatic activities and with known 3D structures. Annotations of active site residues for the corresponding structures were obtained from the Catalytic Site Atlas[45], PDB and UniProt-KB[44]. For each PDB structure with available active site information, protein sites were then stratified into different layers according to the distance between their alpha carbons and the centroid of the active site residues. Each pair of orthologs was then aligned using ClustalW2[57] with a homolog in PDB annotated with the same activity and with defined distance layers around the active site; the PDB sequence with the highest sequence identity to either member of the pair was used for the alignment. Sequence identities for different layers were calculated based on the structural positions in the corresponding PDB reference sequences.

Comparison of pairs of enzymes with the same structural folds.

We used structural classifications of protein domains from the CATH database (v4.2.0)[47]. For structural comparisons, we only considered PDB structures with a single classified domain per chain. Protein pairs classified in CATH in the same homologous structural superfamily were considered as having the same fold. To obtain pairs of proteins in the same fold but with different functions, we only considered PDB structures annotated with different EC numbers in BRENDA. For this analysis we randomly selected 300 pairs of structures with the same fold that do not share any digits of their EC classification.

Acknowledgements

We sincerely thank Dan Tawfik, Eugene Koonin, and Fyodor Kondrashov for very helpful discussions. This work was supported in part by the National Institute of General Medical Sciences grant GM079759 to DV.

References

1. Chothia, C., and Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. The EMBO journal *5*, 823-826.
2. Povolotskaya, I.S., and Kondrashov, F.A. (2010). Sequence space and the ongoing expansion of the protein universe. Nature *465*, 922-926.
3. Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. Nature reviews. Molecular cell biology *8*, 995-1005.
4. Tian, W., and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? Journal of molecular biology *333*, 863-882.
5. Worth, C.L., Gong, S., and Blundell, T.L. (2009). Structural and functional constraints in the evolution of protein families. Nature reviews. Molecular cell biology *10*, 709-720.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6. Wilson, C.A., Kreychman, J., and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. Journal of molecular biology *297*, 233-249.
7. Rost, B. (2002). Enzyme function less conserved than anticipated. Journal of molecular biology *318*, 595-608.
8. Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet *9*, 938-950.
9. Bork, P., Sander, C., and Valencia, A. (1993). Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. Protein science : a publication of the Protein Society *2*, 31-40.
10. Galperin, M.Y., Walker, D.R., and Koonin, E.V. (1998). Analogous enzymes: independent inventions in enzyme evolution. Genome Res *8*, 779-790.
11. Omelchenko, M.V., Galperin, M.Y., Wolf, Y.I., and Koonin, E.V. (2010). Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. Biol Direct *5*, 31.
12. Rost, B. (1997). Protein structures sustain evolutionary drift. Folding & design *2*, S19-24.
13. Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of molecular biology *247*, 536-540.
14. Lehninger, A.L., Nelson, D.L., and Cox, M.M. (2013). Lehninger principles of biochemistry, 6th Edition, (New York: W.H. Freeman).
15. Firnberg, E., Labonte, J.W., Gray, J.J., and Ostermeier, M. (2016). A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. Molecular biology and evolution *33*, 1378.
16. Bairoch, A. (1999). The ENZYME data bank in 1999. Nucleic acids research *27*, 310-311.
17. Hedges, S.B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics *22*, 2971-2972.
18. Dickerson, R.E. (1971). The structure of cytochromec and the rates of molecular evolution. Journal of Molecular Evolution *1*, 26-45.
19. Zuckerkandl, E., and Pauling, L. (1965). Evolutionary Divergence and Convergence in Proteins. In Evolving Genes and Proteins, V. Bryson and H.J. Vogel, eds. (Academic Press), pp. 97-166.
20. Tajima, F., and Nei, M. (1984). Estimation of evolutionary distance between nucleotide sequences. Molecular biology and evolution *1*, 269-285.
21. Gilson, A.I., Marshall-Christensen, A., Choi, J.M., and Shakhnovich, E.I. (2017). The Role of Evolutionary Selection in the Dynamics of Protein Structure Evolution. Biophys J *112*, 1350-1365.
22. Yang, Z. (2006). Computational molecular evolution, (Oxford: Oxford University Press).
23. Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C., and Kondrashov, F.A. (2012). Epistasis as the primary factor in molecular evolution. Nature *490*, 535-538.
24. Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics *155*, 431-449.
25. Ota, T., and Nei, M. (1994). Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. Journal of Molecular Evolution *38*, 642-643.
26. O'Malley-James, J.T., Cockell, C.S., Greaves, J.S., and Raven, J.A. (2014). Swansong biospheres II: the final signs of life on terrestrial planets near the end of their habitable lifetimes. Int J Astrobiol *13*, 229-243.
27. Chang, A., Schomburg, I., Placzek, S., Jeske, L., Ulbrich, M., Xiao, M., Sensen, C.W., and Schomburg, D. (2015). BRENDA in 2015: exciting developments in its 25th year of existence. Nucleic acids research *43*, D439-446.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28. Meyer, T.E., Cusanovich, M.A., and Kamen, M.D. (1986). Evidence against Use of Bacterial Amino-Acid-Sequence Data for Construction of All-Inclusive Phylogenetic Trees. Proceedings of the National Academy of Sciences of the United States of America *83*, 217-220.
29. Dokholyan, N.V., Shakhnovich, B., and Shakhnovich, E.I. (2002). Expanding protein universe and its origin from the biological Big Bang. Proceedings of the National Academy of Sciences of the United States of America *99*, 14132-14136.
30. Hubble, E. (1929). A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. Proceedings of the National Academy of Sciences of the United States of America *15*, 168-173.
31. Barreto, F.S., and Burton, R.S. (2013). Evidence for compensatory evolution of ribosomal proteins in response to rapid divergence of mitochondrial rRNA. Molecular biology and evolution *30*, 310-314.
32. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic acids research *44*, D457-462.
33. Arnold, C., and Stadler, P.F. (2010). Polynomial algorithms for the Maximal Pairing Problem: efficient phylogenetic targeting on arbitrary trees. Algorithms Mol Biol *5*, 25.
34. Benkovic, S.J., Fierke, C.A., and Naylor, A.M. (1988). Insights into Enzyme Function from Studies on Mutants of Dihydrofolate-Reductase. Science *239*, 1105-1110.
35. Kelsic, E.D., Chung, H., Cohen, N., Park, J., Wang, H.H., and Kishony, R. (2016). RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq. Cell Syst *3*, 563-571 e566.
36. Wang, H.H., Isaacs, F.J., Carr, P.A., Sun, Z.Z., Xu, G., Forest, C.R., and Church, G.M. (2009). Programming cells by multiplex genome engineering and accelerated evolution. Nature *460*, 894-898.
37. Galperin, M.Y., Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic acids research *43*, D261-269.
38. Lunzer, M., Golding, G.B., and Dean, A.M. (2010). Pervasive cryptic epistasis in molecular evolution. PLoS Genet *6*, e1001162.
39. Chan, Y.H., Venev, S.V., Zeldovich, K.B., and Matthews, C.R. (2017). Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. Nat Commun *8*, 14614.
40. Watson, J.D., Laskowski, R.A., and Thornton, J.M. (2005). Predicting protein function from sequence and structural data. Current opinion in structural biology *15*, 275-284.
41. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic acids research *28*, 235-242.
42. Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research *33*, 2302-2309.
43. Jack, B.R., Meyer, A.G., Echave, J., and Wilke, C.O. (2016). Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes. PLoS Biol *14*, e1002452.
44. UniProt, C. (2015). UniProt: a hub for protein information. Nucleic acids research *43*, D204-212.
45. Porter, C.T., Bartlett, G.J., and Thornton, J.M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic acids research *32*, D129-133.
46. Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. Cell *138*, 774-786.
47. Dawson, N.L., Lewis, T.E., Das, S., Lees, J.G., Lee, D., Ashford, P., Orengo, C.A., and Sillitoe, I. (2017). CATH: an expanded resource to predict protein function through structure and sequence. Nucleic acids research *45*, D289-D295.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

48. Pingoud, A., Wilson, G.G., and Wende, W. (2014). Type II restriction endonucleases--a historical perspective and more. Nucleic acids research *42*, 7489-7527.
49. Bershtein, S., Serohijos, A.W., Bhattacharyya, S., Manhart, M., Choi, J.M., Mu, W., Zhou, J., and Shakhnovich, E.I. (2015). Protein Homeostasis Imposes a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria. PLoS Genet *11*, e1005612.
50. Adkar, B.V., Manhart, M., Bhattacharyya, S., Tian, J., Musharbash, M., and Shakhnovich, E.I. (2017). Optimization of lag phase shapes the evolution of a bacterial enzyme. Nat Ecol Evol *1*, 149.
51. Shendure, J., and Akey, J.M. (2015). The origins, determinants, and consequences of human mutations. Science *349*, 1478-1483.
52. Kristensen, D.M., Kannan, L., Coleman, M.K., Wolf, Y.I., Sorokin, A., Koonin, E.V., and Mushegian, A. (2010). A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. Bioinformatics *26*, 1481-1487.
53. Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science *278*, 631-637.
54. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research *25*, 3389-3402.
55. Wu, M., and Scott, A.J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics *28*, 1033-1034.
56. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312-1313.
57. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.
58. Sheridan, P.P., Freeman, K.H., and Brenchley, J.E. (2003). Estimated minimal divergence times of the major bacterial and archaeal phyla. Geomicrobiol J *20*, 1-14.
59. Battistuzzi, F.U., Feijao, A., and Hedges, S.B. (2004). A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evol Biol *4*, 44.
60. Bell, E.A., Boehnke, P., Harrison, T.M., and Mao, W.L. (2015). Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. Proceedings of the National Academy of Sciences of the United States of America *112*, 14518-14521.
61. Ku, C., Nelson-Sathi, S., Roettger, M., Sousa, F.L., Lockhart, P.J., Bryant, D., Hazkani-Covo, E., McInerney, J.O., Landan, G., and Martin, W.F. (2015). Endosymbiotic origin and differential loss of eukaryotic genes. Nature *524*, 427-432.
62. Thiergart, T., Landan, G., Schenk, M., Dagan, T., and Martin, W.F. (2012). An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol Evol *4*, 466-485.
63. Parfrey, L.W., Lahr, D.J., Knoll, A.H., and Katz, L.A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proceedings of the National Academy of Sciences of the United States of America *108*, 13624-13629.
64. Spang, A., Saw, J.H., Jorgensen, S.L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A.E., van Eijk, R., Schleper, C., Guy, L., and Ettema, T.J.G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature *521*, 173-179.
65. Guy, L., and Ettema, T.J. (2011). The archaeal 'TACK' superphylum and the origin of eukaryotes. Trends Microbiol *19*, 580-587.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

66. Betts, H.C., Puttick, M.N., Clark, J.W., Williams, T.A., Donoghue, P.C.J., and Pisani, D. (2018). Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nat Ecol Evol.
67. Jukes, T.H., and Cantor, C.R. (1969). CHAPTER 24 - Evolution of Protein Molecules. In Mammalian Protein Metabolism, H.N. Munro, ed. (Academic Press), pp. 21-132.
68. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.
69. Plata, G. (2018). count-MAGE-seq. GitHub *https://github.com/platyias/count-MAGE-seq*, 4f827f620dbdde931c899f828f103fc305053f305024db.
70. DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol *72*, 5069-5072.
71. Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author: Department of Genome Sciences, University of Washington, Seattle.
72. Ye, Y., and Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics *19 Suppl 2*, ii246-255.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

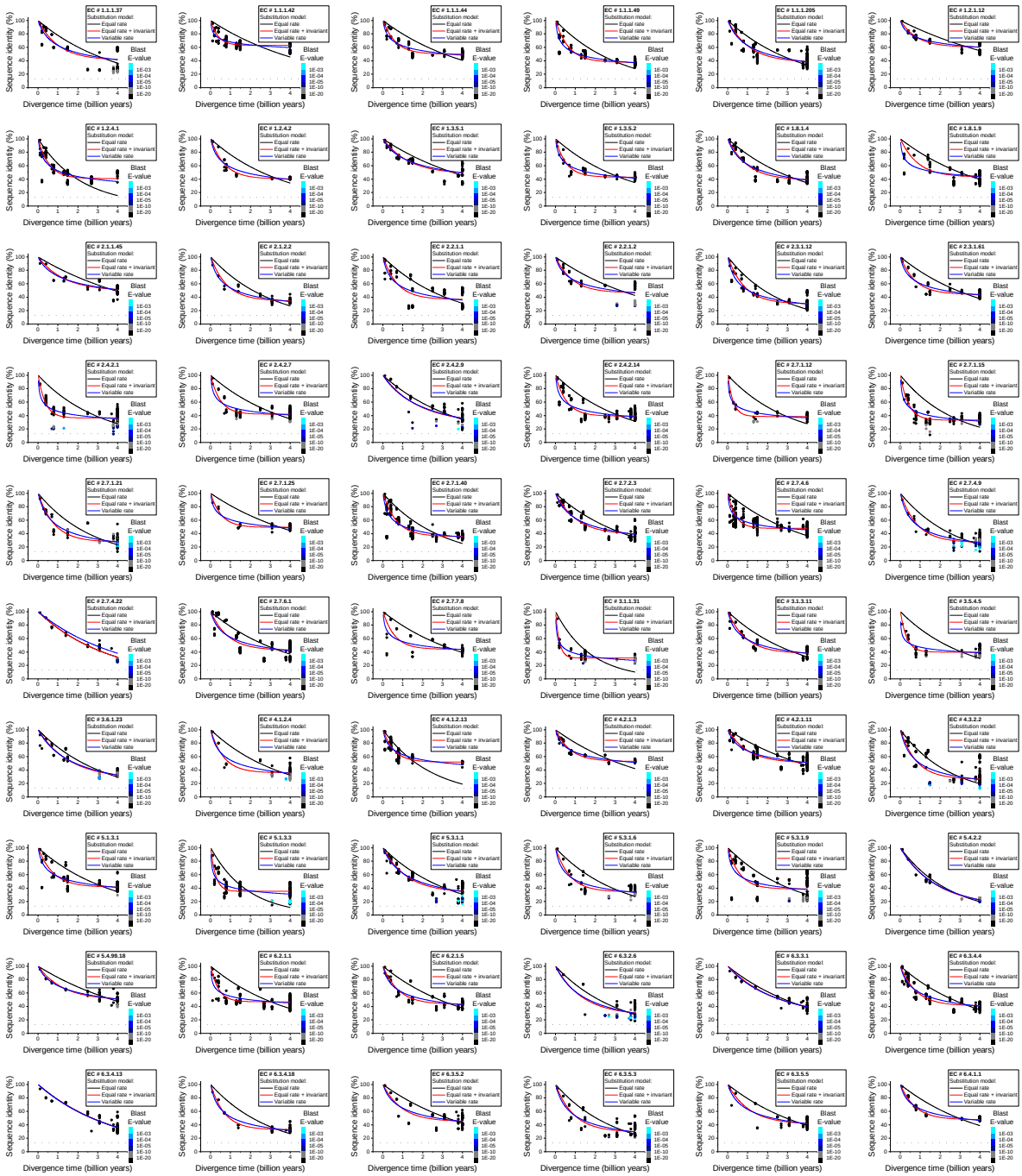

Figure 1 – figure supplement 1. **Sequence divergence of enzyme orthologs as a function of time.** The global pairwise sequence identities between pairs of orthologous enzymes as a function of divergence times between the corresponding species. The colored lines indicate fits to the data using the three models of amino acid substitutions (Equations 1-3). The horizontal dashed black lines show the average sequence identity for the global alignment of unrelated protein sequences. Dot colors indicate the corresponding psi-blast E-value of the pairwise alignments for a database size of 3.8×10^5 sequences. Parameter fits of the models are presented in Supplementary file 2a.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

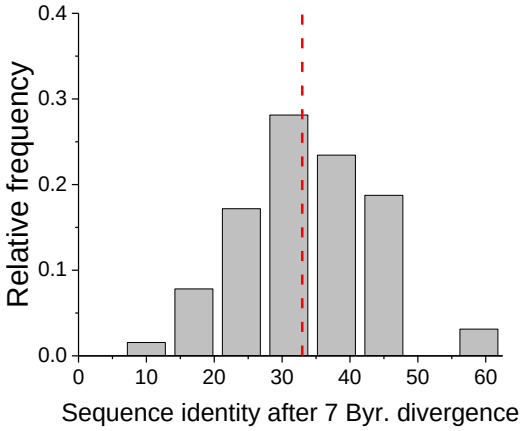

Figure 1 – figure supplement 2. Projected long-term sequence identity of metabolic orthologs. The distribution, across 64 EC numbers, of the projected global sequence identity after 7 billion years of divergence according to Model 3 (Equation 3). The vertical dashed red line represents the median value of the distribution (33%).

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

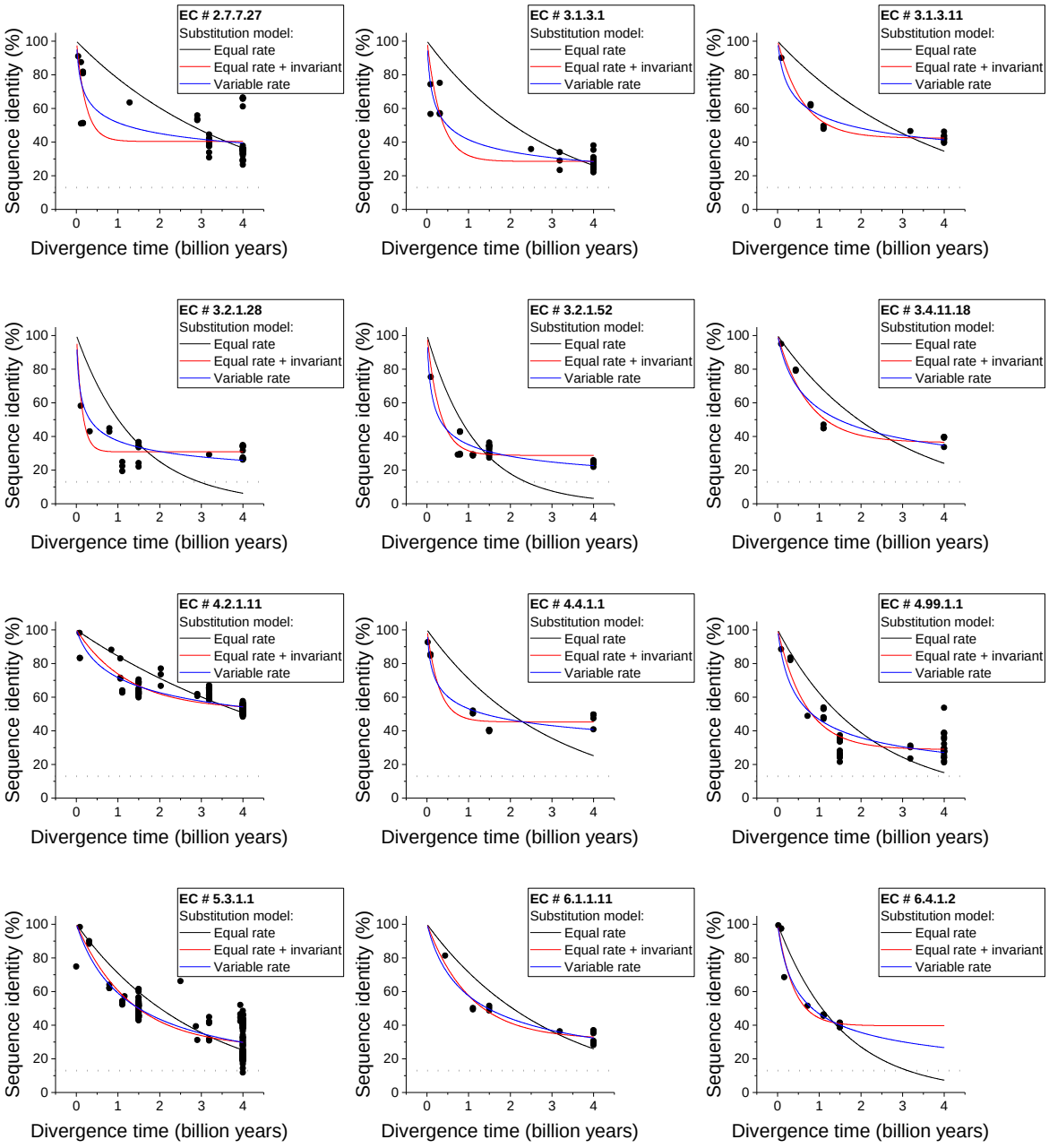

Figure 1 – figure supplement 3. Divergence of orthologs with experimentally validated functional annotations. Sequence identity as a function of divergence time for orthologous enzymes annotated with the same EC number and experimentally validated molecular functions. The color lines indicate fits to the data for the three models of amino acid substitution (Equations 1-3). Results are shown for 12 different EC numbers. For all EC numbers, Model 2 (red) gives a better fit to the data compared to Model 1 (black, F-test P-value < 0.05). Also, for all EC numbers the minimum sequence identity at long divergence times (Y_0, from equation 2) is higher than the average sequence identity between unrelated protein pairs (~13.5%, dashed black lines) (Wald-test P-value < 0.05).

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

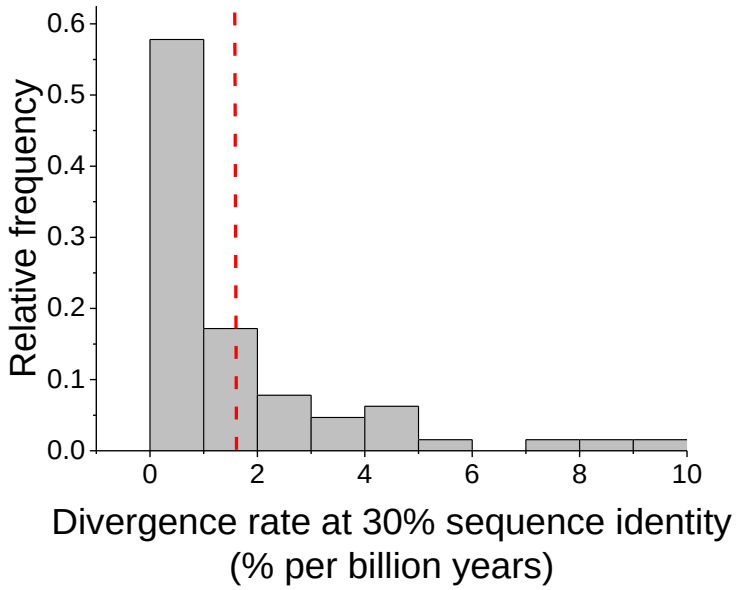

Figure 1 – figure supplement 4. Enzyme divergence rates at 30% sequence identity. The distribution of divergence rates between orthologous enzymes when their sequence identity reaches 30%. Divergence rates were defined as the decrease in percent sequence identity per billion years. Divergence rates were calculated based on Model 3 fits and equation 4. The dashed red line represents the mean value of the distribution across enzymes (1.6% per billion years).

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

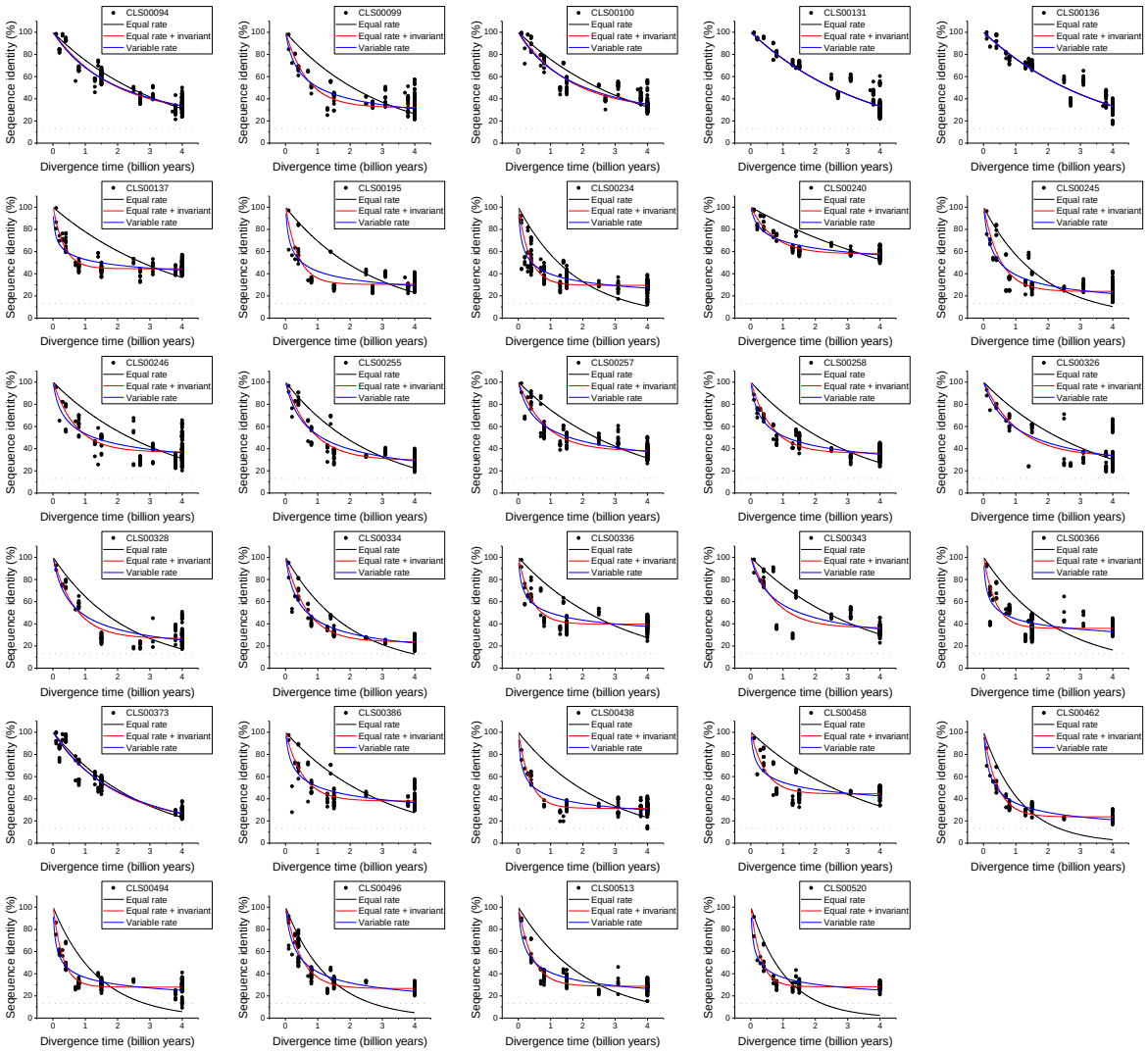

Figure 1 – figure supplement 5. Sequence divergence of non-enzyme orthologs as a function of divergence time. The global pairwise sequence identities between pairs of orthologous proteins that are not part of the EC nomenclature are shown as a function of divergence times between the corresponding species. The colored lines indicate fits to the data using the three models of amino acid substitution (Equations 1-3). The dashed black lines represent the average sequence identity for the global sequence alignment of unrelated protein sequences. Model parameter fits are presented in Supplementary file 2c.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

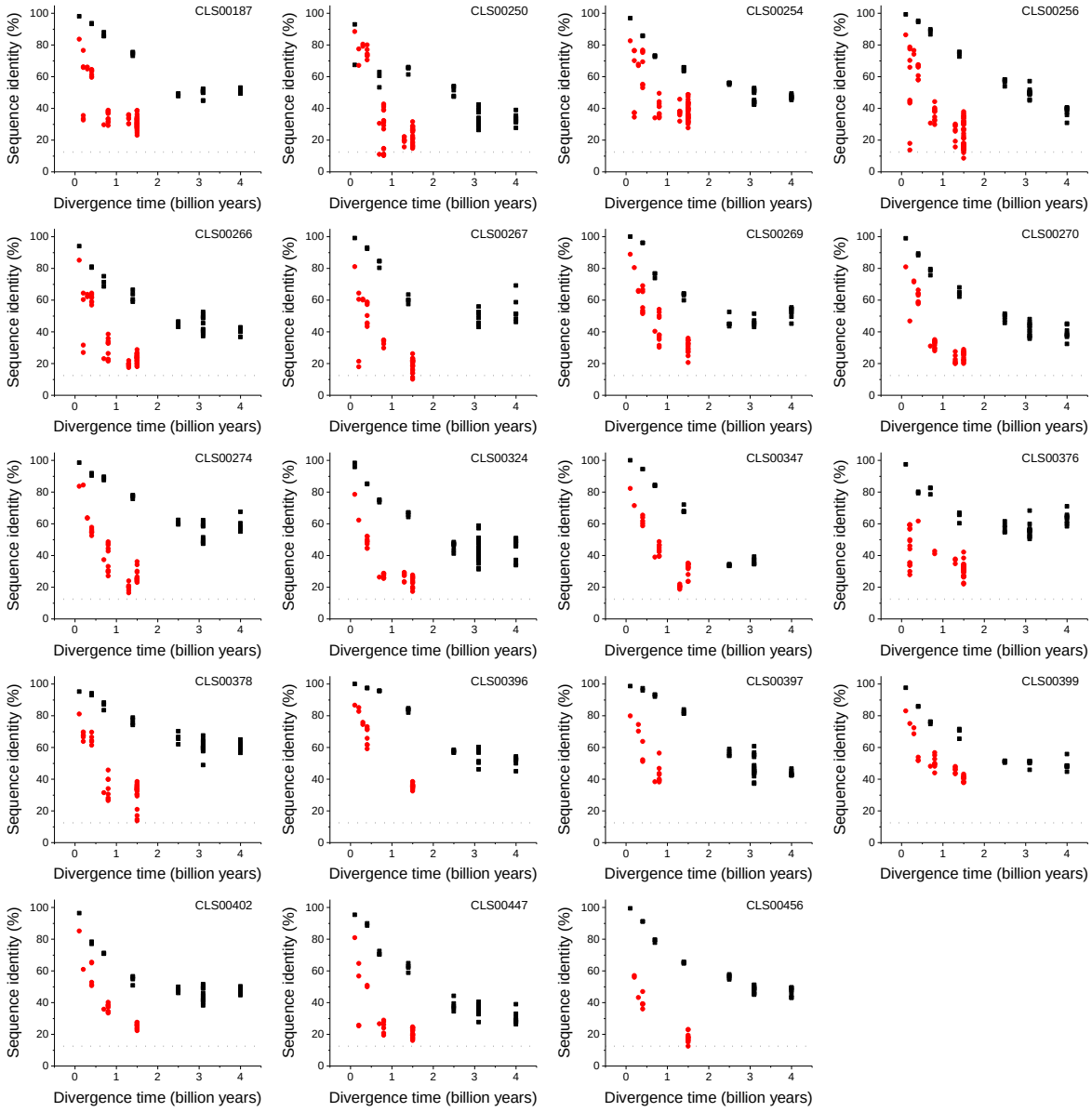

Figure 1 – figure supplement 6. Sequence divergence of mitochondrial ribosomal orthologs as a function of divergence time. Global sequence identity as a function of divergence time for orthologous proteins with substantially different divergence rates between eukaryotic orthologs (red) and pairs of orthologs involving at least one prokaryotic sequence (black). Eukaryotic proteins are mostly annotated as mitochondrial ribosomal proteins and translation initiation factors. The dashed black lines represent the average sequence identity for the global sequence alignment of unrelated protein sequences.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

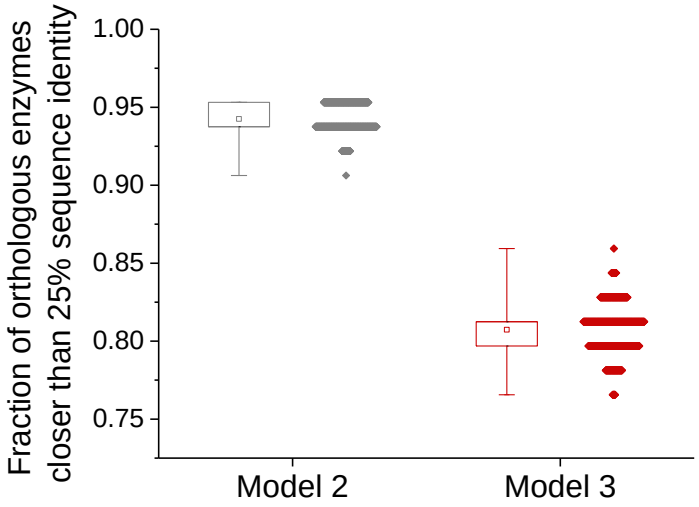

Figure 1 – figure supplement 7. Effect of uncertainty in the estimation of species divergence times on the model fits. Effect of different estimates of species' divergence times on the predicted long-term sequence identity between orthologous enzymes with the same molecular function. The data on the left shows the fraction of enzymes for which the model 2 Y_0 parameter is higher than 25% sequence identity; each dot corresponds to a different assignment of lineage divergence times (based on either minimum or maximum literature estimates; see Methods). The data on the right shows the fraction of orthologous enzymes for which the predicted sequence identity after 7.8 billion years of divergence is higher than 25% identity. The box plots show the minimum and maximum values and 25 and 75% percentiles based on 1,000 divergence time assignments.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

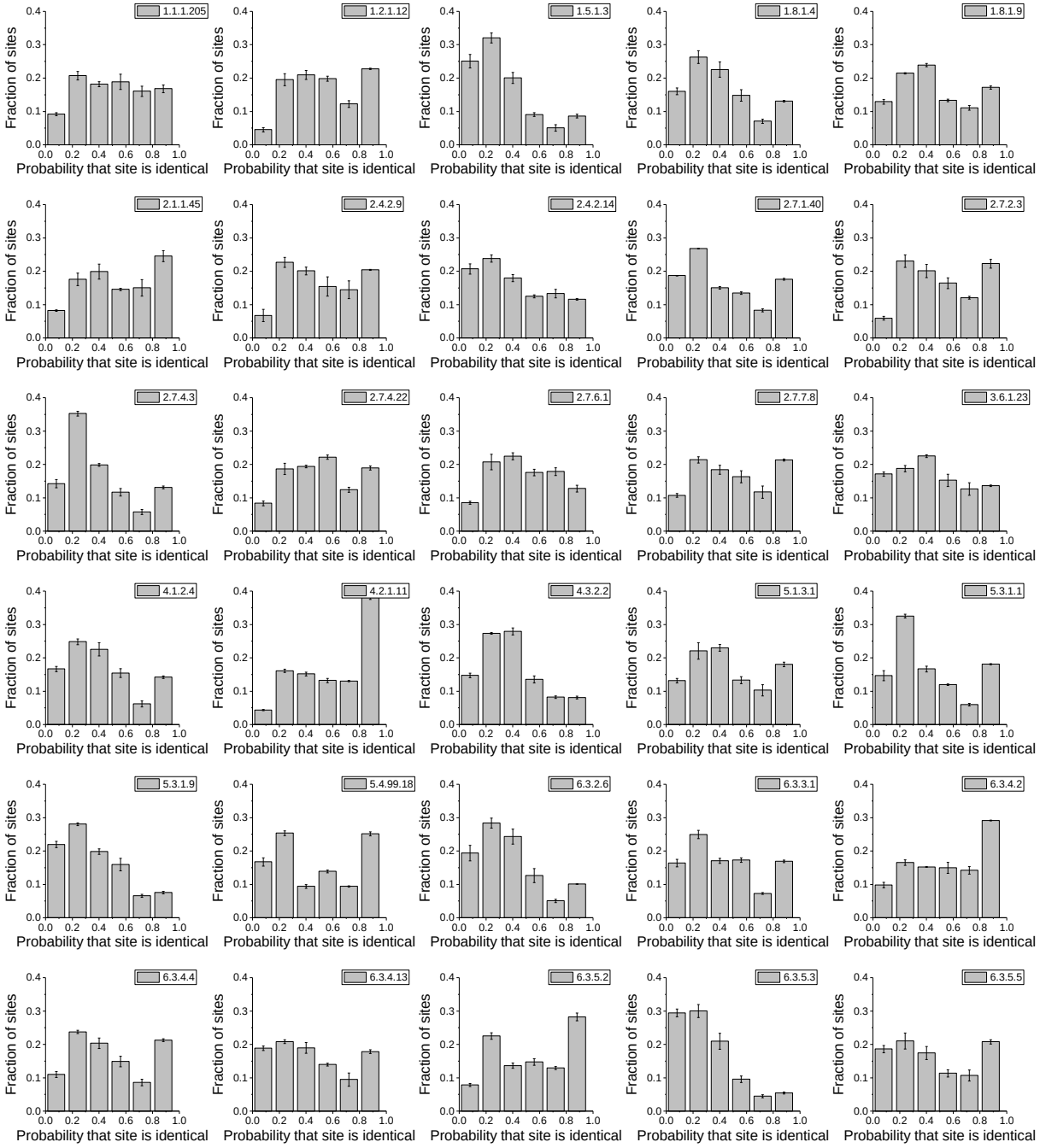

Figure 3 – figure supplement 1. Distribution of enzyme sites according to their conservation frequency. The panels show the distribution of protein sites according to their probability of being identical in phylogenetically independent lineages; the distributions are show for 30 different enzyme activities (EC numbers). The probabilities were calculated as the proportion of phylogenetyically independent pairs of orthologs with divergence times >2 billion years (Figure 2a) in which a protein site is occupied by identical amino acids. Error bars represent the S.E.M across three replicates with different assignments of independent orthologous pairs.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

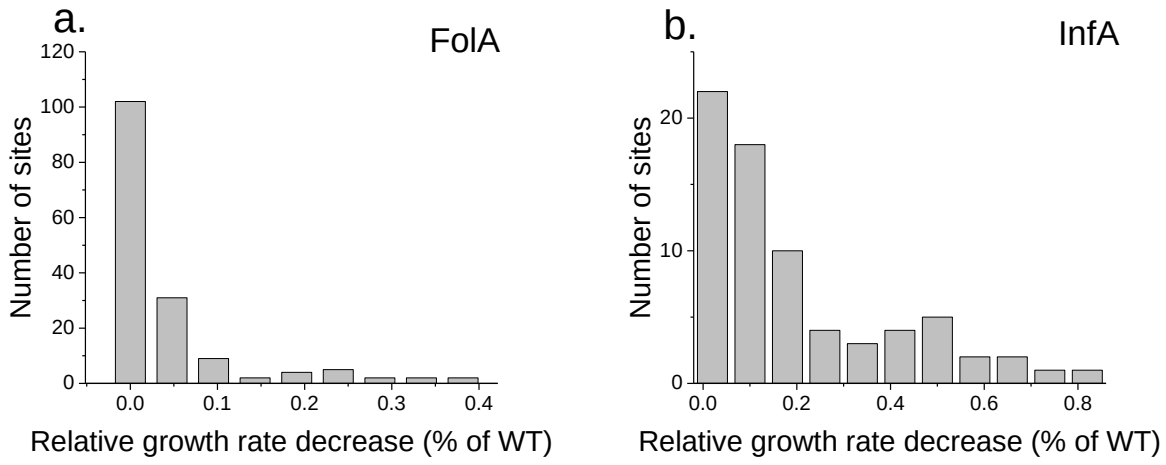

Figure 4 – figure supplement 1. Distribution of average fitness effects of amino acid substitutions. The panels show the distribution of average growth defects due to amino acid substitutions across FolA (**a**) and InfA (**b**) protein sites. The average growth (fitness) defects at each site was calculated as the mean values of experimentally measured growth phenotypes due to all possible amino acid substitutions at the site.

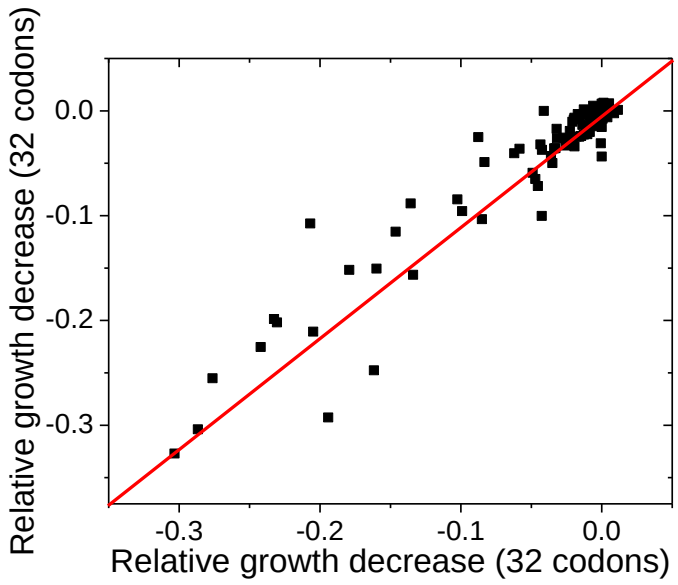

Figure 4 – figure supplement 2. Reproducibility of experimentally measured average fitness effects of amino acid substitutions across FolA sites. The similarity between the average fitness effects of substitutions across FolA sites was calculated using two non-overlapping sets of substitutions (see Methods). Each dot in the figure represents one FolA site. The relative growth decrease in the X axis was calculated based on the relative growth rates of mutants carrying 32 randomly selected codons; the relative growth decrease in the Y axis was calculated based on the relative growth rates of mutants carrying the remaining 32 codons. The line X=Y is shown in red. Pearson's r=0.95, p-value<1e-20.

bioRxiv preprint doi: https://doi.org/10.1101/198663; this version posted June 14, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary files legends

Supplementary file 1. Considered model species and pairwise average divergence times

Supplementary file 2A. Fitted model parameters and test results for the 64 considered activities (EC numbers)

Supplementary file 2B. Estimated rates of sequence divergence for pairs of orthologs according to Model 3 fits

Supplementary file 2C. Fitted model parameters and test results for 29 sets of orthologs not annotated with EC numbers

Supplementary file 2D. UniProt annotations of representative sequences from E. coli and H. sapiens for sets of orthologs not annotated with EC numbers

Supplementary file 2E. UniProt annotations of representative sequences from E. coli and H. sapiens for sets of orthologs not annotated with EC numbers and fast divergence rates in eukaryotes

Supplementary file 3. Average relative growth rate effect of amino acid substitutions in the FolA protein of E. coli

Supplementary file 4. DNA oligomers used to introduce amino acid substitutions along the FolA protein