Molecular function limits divergent protein evolution on planetary timescales

Mariam M. Konaté[a,b*], Germán Plata[a*], Jimin Park[a,c], 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 Dennis Vitkup:
1130 Saint Nicholas Ave. RM 803, New York, NY 10032
Tel: 212 851 5151
E-mail: dv2121@columbia.edu

Abstract

Sequence analyses demonstrate that even after billions of years of independent evolution, protein orthologs are still diverging from each other. Although the sequence and functional evolution of proteins have been studied in detail, it is currently unknown whether the requirement to continuously maintain the same molecular function imposes an effective limit on protein divergence. Here we investigate this fundamental question using a combined computational and experimental analysis of orthologous enzymes with the same molecular function. Interestingly, our results demonstrate that the mutual divergence rates of orthologous enzymes substantially decrease after ~1-2 billion years of independent evolution. As a result of this slowdown, which is observed well above the levels of detectible sequence homology, the sequence and structural similarities between ancient orthologs have not substantially decreased during the last billion years. We find that conservation of molecular function constrains the maximum divergence of all protein regions, including those distant from the active site residues. By experimentally characterizing the growth rates of all amino acid substitutions in the *E. coli* FolA protein, we delineate two stages of long-term sequence divergence with different contributions from sites with small and large fitness effects. Overall, our analysis suggests that the vast majority of ancient enzyme orthologs are unlikely to diverge, at least on planetary timescales, beyond ~25% sequence identity.

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 and sequence identity[3-5]. Specifically, characterizing the probability that two sequences share the same molecular function given their sequence similarity[4] has been important for understanding the emergence of new functions within protein families[6, 7], and for the functional annotation of protein sequences[3, 8]. In this work, we focused on a different and currently unaddressed question. Namely, how far can two sequences typically diverge in natural populations while continuously maintaining the same molecular function, for example the ability to catalyze the same enzymatic reaction. We note that the requirement for continuous conservation of molecular function is crucial in this context, as multiple examples of convergent evolution and protein engineering demonstrate that catalysis of the same chemical reactions can be accomplished by proteins with unrelated sequences and different folds[9-11].

It was previously demonstrated that proteins with the same fold frequently diverge to very low (~10%) levels of sequence identity[12]. Because only a relatively small fraction of protein residues (~3-5%) are usually directly involved in catalysis[13], it is plausible that divergent evolution, even when constrained by the same molecular function, will eventually lead to orthologs with essentially random levels of sequence identity. However, recent analyses have shown that even sites located far from catalytic residues may constrain protein evolution, as substitutions at these sites can have substantial effects on enzymatic activity[14]. In this study, we explore the long-term divergence limit of protein orthologs using several models of molecular evolution. Importantly, a divergence limit does not imply that the rate of amino acid substitutions slows down in evolution, but suggests that subsequent substitutions do not substantially further decrease the similarity between orthologous sequences and structures. We also investigate the correlation between experimentally measured fitness effects of amino acid substitutions and the contribution of corresponding protein sites to long-term sequence divergence.

Results

To study the evolution of proteins with the same molecular function, we used orthologous enzymes annotated in UniProt[15] with the same enzymatic activity (see Methods). We focused 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. In the present analysis, we used protein sequences representing 64 EC numbers from 22 diverse model organisms across the three domains of life (Table S1).

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 considered two simple models of long-term protein evolution, one without a limit of sequence divergence and the other with an explicit divergence limit. In the first model, all protein sites have equal and independent substitution rates; see Equation 1, where *y* represents global sequence identity, *t* represents divergence time, and R_0 represents the average substitution rate. In the second model, in addition to sites with equal and independent substitution rates, there is a minimum fraction of identical sites at long divergence times, represented by Y_0 in Equation 2.

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

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

We applied the two models to fit sequence divergence for each of the considered enzymatic functions. Species' divergence times were used to estimate the times of divergence between corresponding orthologous proteins. For each enzymatic activity, phylogenetic trees based on the orthologous protein sequences were constructed and compared to species' trees; proteins that showed clear differences in the tree topologies, suggesting horizontal gene transfer, were excluded from the analysis (see Methods).

The best model fits for four representative metabolic activities are shown in Figure 1 (black for the first model and red for the second); see Figure S1 for the other metabolic activities. 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, Table S2). Moreover, in 61 of the 64 cases, the maximum likelihood value of the parameter Y_0 is significantly higher than the average sequence identity of the optimal global alignment between random protein sequences (13.5%, shown in Figure 1 and Figure S1 by dashed black lines) (Wald test P-value <0.05). The distribution of the optimized parameter Y_0 suggests a long-term sequence identity higher than 25% between considered orthologs (Figure S2a), supporting the idea that conservation of enzymatic function significantly limits protein divergence.

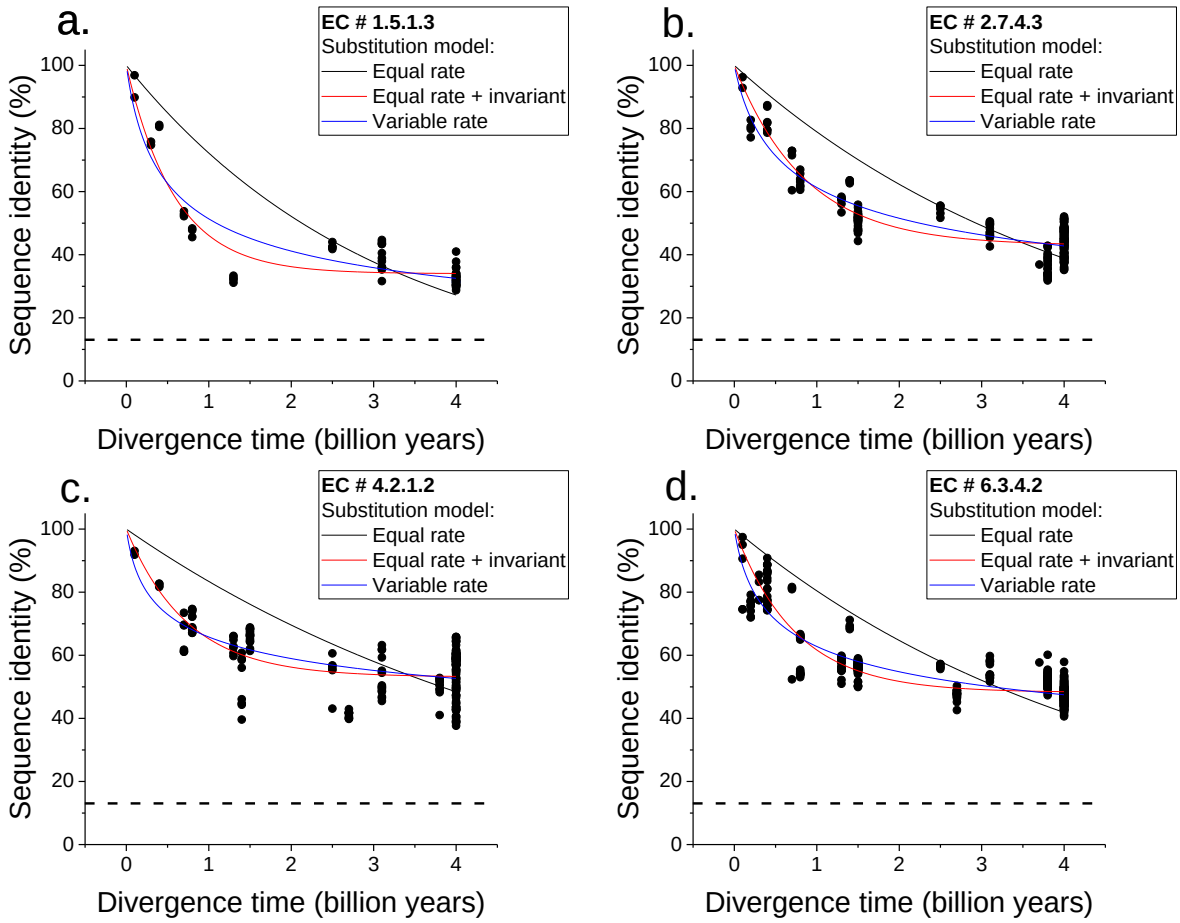

Figure 1. Sequence divergence of enzyme orthologs as a function of time. The global pairwise sequence identities between pairs of orthologous enzymes are shown as a function of divergence times between the corresponding species. Three models of amino acid substitution were used to fit the data. Model 1 (black line) assumes independent and equal substitution rates across all protein sites. Model 2 (red line) assumes that a given fraction of protein sites remains identical at long divergence times. Model 3 (blue line) 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 dashed black lines represent the average sequence identity for the alignment of unrelated protein sequences. The data and corresponding model fits for the other EC numbers considered in the analysis are given in Figure S1 and Table S2.

The two aforementioned models simplify the process of sequence divergence by considering similar substitution rates across protein sites. A more realistic and commonly used model of protein evolution assumes a gamma distribution[17-19] of substitution rates across sites; see Equation 3, 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 S1) show that the rates of mutual protein sequence divergence between orthologous enzymes have decreased by more than 10 times during ~4 billion years of evolution (see

Methods and Table S3). Across enzymes, the sharpest decrease in divergence rates occurred within 1-2 billion years of divergence from a common ancestor (Figure 1 and Figure S1). Notably, the observed decrease in divergence rates is substantially higher (by >5 fold) than the one expected simply due to progressively fewer non-mutated sites in models with equal and independent substitution rates. The slowdown in divergence does not imply that mutation rates have decreased in molecular evolution. Instead, a substantial level of conservation is likely necessary for the continuous preservation of molecular function, leading to amino acid substitutions that do not further decrease sequence identity[20]. The third model does not explicitly consider a long-term divergence limit. Nevertheless, the projected sequence identity also shows that the vast majority of orthologous enzymes will remain above 20-25% sequence identity on the timescales in which Earth environments will remain hospitable to life (estimated at 1-2.8 billion years from the present[21]) (Figure S2b).

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

Interestingly, following the previously introduced metaphor of the expanding protein universe[2], we can use the third model to express the divergence rate between orthologs as a function of protein distance (*D = 1 – y*, where *y* is sequence identity), see Equation 4. This equation, similarly to Hubble's law of universe expansion[22], describes how the divergence rate changes with increasing distance between protein orthologs.

$$\frac{\partial D}{\partial t} = 100 * R_0 * (1 - D)^{(\alpha + 1/\alpha)} \qquad (4)$$

We note that 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 S1) for which corresponding orthologs can be easily identified by computational sequence comparison methods. Furthermore, the results remained similar when we restricted the analysis to orthologous enzyme pairs with experimentally validated molecular functions (Figure S3), using publications referenced in the BRENDA database[23].

Next, to understand how different protein sites contribute to long-term sequence divergence constraints, 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 chose FolA for the experiments due to its small size (159 amino acids) and essential role in the *E. coli* metabolism[24]. Also, the long-term protein sequence identity between FolA orthologs (~32%, see Figure 1a) is similar to other analyzed enzymes (Figure S2a). Following a recently described strategy[25], we used the Multiplex Automated Genome Engineering (MAGE) approach[26], to introduce every possible amino acid substitution at each FolA site in *E. coli*. We then measured fitness effects (relative growth rates) of each variant by performing growth competition experiments, followed by next generation sequencing (see Methods, Dataset S1).

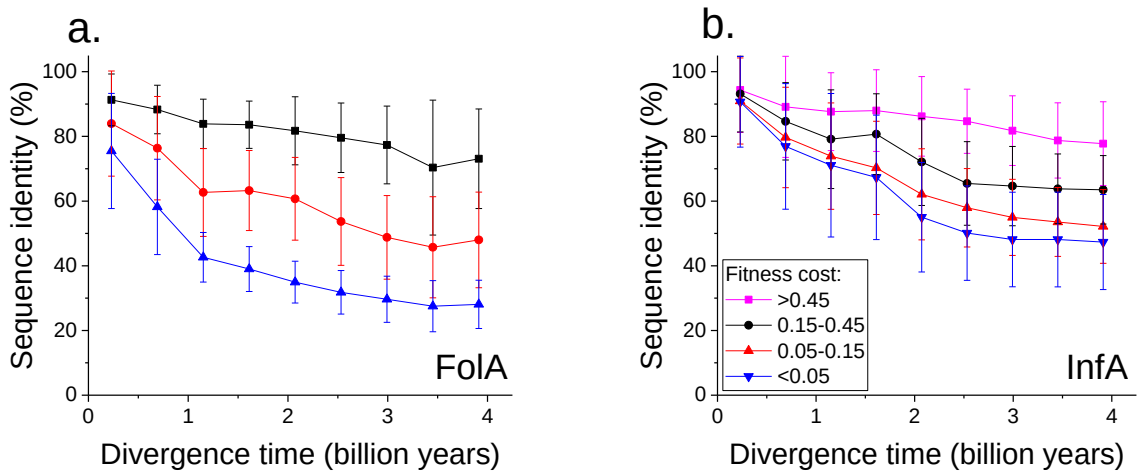

Figure 2. Sequence divergence of protein sites with different mutant fitness effects in *E. coli*. a. Sequence identity between pairs of FolA orthologs as a function of divergence time. Results are shown for subsets of protein sites (represented by different colors) with different average fitness effects of mutations in *E. coli*. The average sequence identity values were calculated using ~300 bacterial FolA orthologs in the COG database[27]; divergence times were estimated using bacterial 16S rRNA sequences (see Methods). **b.** Similar to panel a, but for the sequence divergence of *E. coli* translation initiation factor InfA orthologs (~500 sequences). Error bars represent the standard deviation of sequence identity in each bin.

Using the growth measurements in *E. coli*, we investigated the long-term sequence divergence of protein sites with different contribution to fitness. To that end, we sorted FolA protein sites into several groups according to the experimentally measured average fitness effects. We then used a large multiple sequence alignment of FolA orthologs (~300 sequences) from the COG database[27] (see Methods) to estimate sequence identity as a function of corresponding species' divergence times for sites with different average fitness effects (Figure 2a, different colors). Similar to the global sequence identity, after ~1.5 billion years of evolution, we observed a significant decrease in mutual divergence rates for sites in the least deleterious fitness bin (Figure 2a, blue). In contrast, sites with the most deleterious mutations in *E. coli* (Figure 2a, black) display approximately constant divergence rate throughout evolutionary history.

To understand the generality of the FolA results, we used a similarly obtained fitness dataset[25] for the *E. coli* translation initiation factor InfA (Figure 2b). Using the InfA data we observed, on average, lower mutant growth rates and lower rates of sequence divergence than for FolA. Nevertheless, the observed long-term divergence patterns were qualitatively similar (Figure 2b). For sites in the least deleterious fitness bin (Figure 2b, blue) we observed a substantial decrease in divergence after ~2 billon years of evolution, while sites with most deleterious mutations in *E. coli* (Figure 2b, pink) display an approximately constant but slow divergence rate.

The results described above suggest a correlation between substitution fitness effects at a protein site and the site's long-term sequence identity. To further quantify this relationship, we calculated the probability that each site is occupied by the same amino acid in pairs of orthologs at large phylogenetic distances, when mutual divergence rates have significantly slowed down (~2 billion years for FolA and ~2.5 billion years for InfA). To this end, we used the TARGETING package[28] to identify phylogenetically independent pairs of species with FolA and InfA orthologs (Figure 3a), and calculated the fraction of independent pairs where a site was identical (Figure 3b). Notably, the probability that a protein site is identical, and thus contributes to the divergence limit, first increases linearly with increasing fitness effects at the site, and then begins to saturate for sites with high (>30%) fitness effects (Figure 3b).

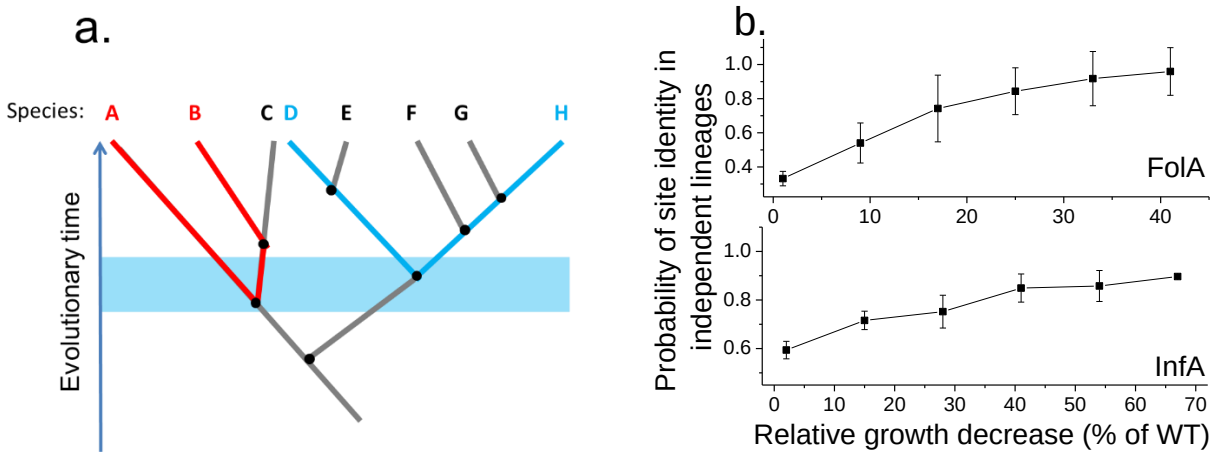

Figure 3. Probability of protein sites conservation in phylogenetically independent lineages. a. Illustration of species pairs representing phylogenetically independent lineages. In the figure, A-B and D-H are pairs of species that diverged within the specified time window (illustrated by the shaded region), and do not share more recent edges in the phylogenetic tree. **b.** 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 sites in *E. coli*. The probability represents the fraction of phylogenetically independent pairs of orthologs in which sites are identical at long divergence distances (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 with a given fitness effect.

Given that some protein sites are more likely to be conserved between ancient pairs of orthologs, it is interesting to understand to what extent the same sites contribute to the divergence limit in different lineages. Towards that end, we calculated the distribution quantifying how frequently protein sites are occupied by identical amino acids in phylogenetically independent linages (Figure 4). The results, for both FolA and InfA orthologs, demonstrate that only a relatively small fraction of protein sites (10-20%) are identical in a majority (>90%) of independent lineages. Therefore, the observed divergence limit is not due to a uniform set of conserved protein sites between all pairs of orthologs. Instead, heterogeneous sets of sites contribute to the limit in independent lineages. Analysis of the sites that contribute to sequence identity between orthologs at two consecutive divergence times (black and gray in Figure 4), revealed only minor differences in the corresponding probability distributions. Therefore, the observed slow-down of sequence divergence after 1.5-2 billion years of evolution is accompanied by the convergence of the probability distribution characterizing the relative contribution of different protein sites to the limit.

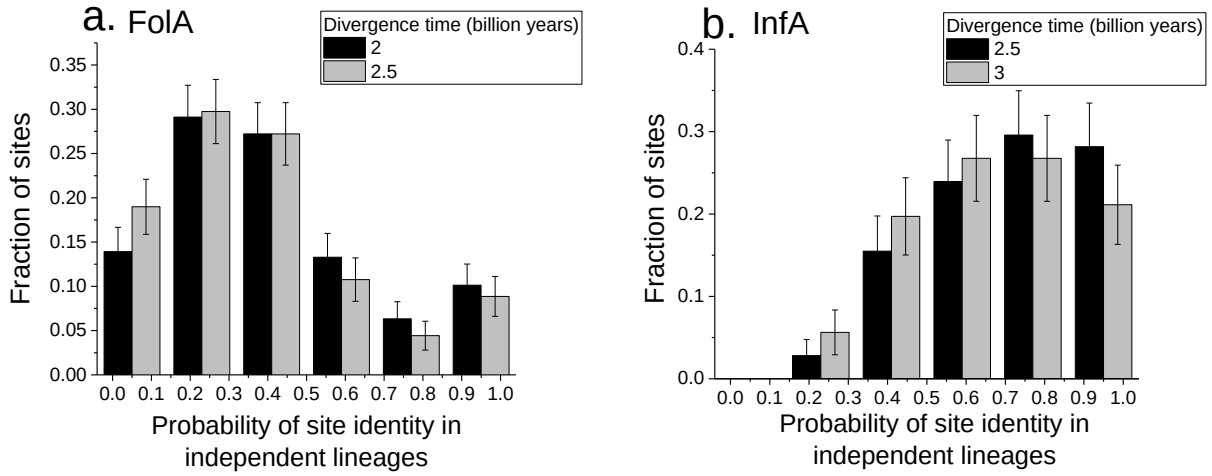

Figure 4. Contribution of different protein sites to the observed divergence limit. a. Distribution of FolA protein sites according to the probabilities of being identical in phylogenetically independent lineages. Probabilities were calculated as the proportion of phylogenetically independent pairs of FolA orthologs with identical amino acids. Results are shown for divergence times of 2 (black) and 2.5±0.25 (grey) billion years between pairs of orthologs. **b.** Similar to panel a. but for InfA orthologs. Results are shown for divergence times of 2.5 (black) and 3±0.25 (grey) billion years. Error bars represent the standard error of the calculated fractions across sites in each probability bin.

Conservation of molecular function should also significantly constrain protein structural evolution[8, 29]. To investigate the long-term structural divergence of proteins with the same molecular function, we structurally aligned orthologous enzymes using the TM-align algorithm[30]. In this analysis,

we used more than 1000 orthologous pairs of enzymes (sharing all 4 EC digits) with known 3D structures (see Methods). The average root mean square deviation (RMSD) between C-alpha atoms of the orthologous enzymes increases (Spearman's r=0.44, P-value<1e-20) with divergence time between the corresponding species (Figure 4a). Nevertheless, the C-alpha RMSD rarely diverges beyond 3 Å even at long evolutionary distances. Consistent with sequence evolution (Figure 1), we observed a substantial decrease in the rate of structural divergence after ~1.5 billion years of evolution.

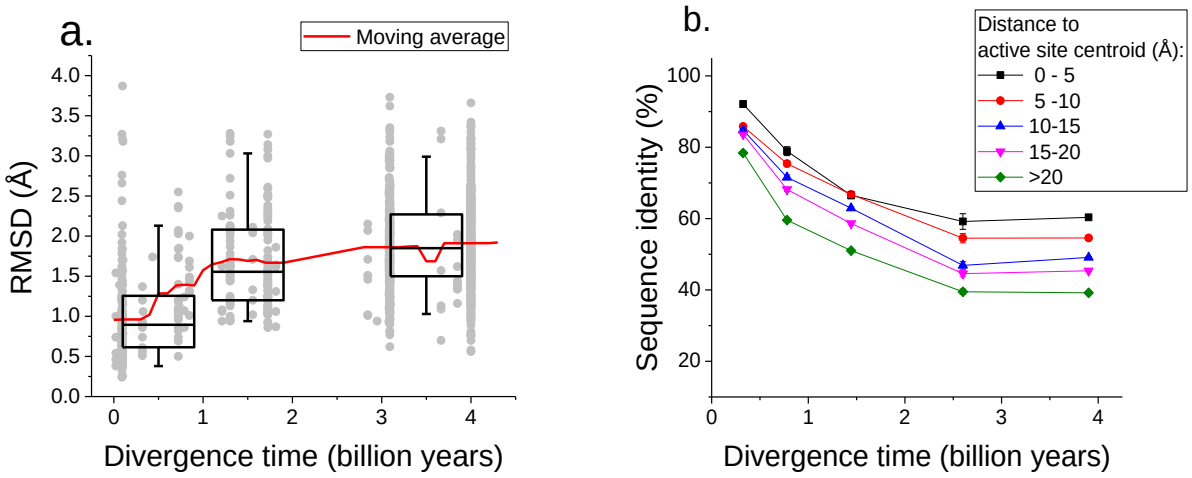

Figure 5. Long-term structural evolution of proteins with the same molecular function. a. The pairwise C-alpha root mean square deviation (RMSD) as a function of the divergence time between pairs of orthologs annotated with the same EC number and experimentally determined structures (gray dots). RMSD values were calculated based on structural alignments using the TM-align algorithm[30]. Boxes indicate the median and 25-75 percentiles at the corresponding divergence times, the vertical lines indicate the 5-95 percentile range, and the red line shows a moving average of the data. **b.** Divergence of sequence identity 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 S1); the average sequence identities at each distance shell were calculated across all pairs of orthologs annotated with the same EC number (see Methods). Error bars show the S.E.M. across ortholog pairs.

Only a small fraction of all protein residues usually forms an enzyme active site. Therefore, it is interesting to investigate how the sequence divergence limit depends on the spatial proximity to active site residues. While it was recently demonstrated that evolutionary rates correlate with the sites' spatial distance to catalytic residues[31], it is not clear whether conservation of molecular function would limit the maximum possible divergence of corresponding protein sites. To address this question, we considered catalytic site annotations available from the Protein Data Bank[32], UniProt-KB[15] and the Catalytic Site Atlas[33] and quantified the average divergence of sequence identity at various distances from catalytic residues (see Methods, Figure 5b). For this analysis, we considered the same set of enzymatic activities

259 used to study global sequence divergence (Figure 1 and Figure S1). Although, as expected, residues close
260 to the active site are most highly conserved[31, 34], even residues distant from the active site display an
261 effective divergence limit at long evolutionary distances. This result suggests that the spatial constraints
262 required to conserve molecular function usually propagate throughout the entire protein structure.

263 **Discussion**

264 The presented results demonstrate that, in contrast to proteins with the same fold[12], the
265 requirement to maintain the same molecular function significantly constrains the long-term divergence
266 of protein orthologs. Moreover, our analysis delineates two different stages of divergence for ancient
267 orthologs with the same function. During the first stage, substitutions at sites with mild fitness effects
268 lead to a relatively fast initial divergence. However, after 1-2 billion years of independent evolution,
269 divergence at these sites effectively saturates. Notably, the long-term sequence identity between these
270 sites at saturation (30-40%) is still substantially higher than the expected identity between two random
271 protein sequences (~10%). This suggests that, even at sites with small fitness effects, not all combinations
272 of amino acids are compatible with maintaining the same molecular function. In contrast, the divergence
273 rate at sites with large fitness effects is slow but approximately constant throughout ~4 billion years of
274 evolution. The saturation at sites with small fitness effects and the slow divergence at sites with large
275 effects, lead to a substantially slower overall sequence and structural divergence during the second stage,
276 i.e. after ~1-2 billon years of independent evolution. As a result of this slowdown, for the past billion years
277 there has not been a substantial divergence in sequence or structure between ancient orthologs with the
278 same molecular function, i.e. no significant further decrease in their sequence and structural similarity.

279 Our analysis also demonstrates that different protein sites are usually conserved between
280 orthologs in independent lineages. This result is likely due to the epistatic nature of protein sequence
281 landscapes, where mutations neutral in one lineage are often prohibitively deleterious in another[35, 36].
282 Despite epistatic interactions, our experimental results indicate that there is a correlation between the
283 average fitness effect of mutations at a protein site, measured in a single species, and the long-term
284 conservation of the site across multiple species. This result is consistent with the recent study describing
285 correlated fitness landscapes between distantly related TIM barrel proteins with the same enzymatic
286 activity[37].

287 A previous analysis by Povolotskaya *et al.*[2] demonstrated that ancient protein orthologs are still
288 diverging from each other, but that study did not explicitly considered the effect of conservation of
289 molecular function. Our analysis shows that the requirement to conserve a specific molecular function
290 significantly constraints sequence and structural divergence of orthologs. For example, enzymes sharing

291 only the first three digits of the EC annotation, i.e. only a general class of substrates or cofactors[16],
292 display a significantly lower sequence identity at large phylogenetic distances compared orthologs sharing
293 all four EC digits (Figure S4). Similarly, significant sequence identity between homologous restriction
294 endonucleases is usually limited to isoschizomers, i.e. proteins specific to the same target sequence[38].
295 These patterns suggest that the effective divergence limit is likely a consequence of functional pressure
296 to closely preserve the relative spatial positions and dynamical properties of protein residues participating
297 in ligand binding and catalysis[13]. The conservation of the chemical structures of enzyme substrates,
298 coupled with the requirement for efficient catalysis, likely makes the long-term evolutionary drift in
299 sequence space increasingly slow. Further analyses of the biochemical and biophysical constraints
300 affecting orthologs' evolution will improve our understanding of what structural protein features are
301 essential for catalysis, and how catalytic efficiency may be compromised by deleterious mutations[39].

302 **Methods**

303

304 **Considered enzyme activities and corresponding protein orthologs**. We selected for analysis the
305 sequences annotated in UniProt[15] with EC numbers associated to the following metabolic pathways (as
306 defined in the KEGG database[40]): Glycolysis and gluconeogenesis, pentose phosphate pathway, TCA
307 cycle, purine metabolism, and pyrimidine metabolism. Using protein sequences from 22 diverse
308 organisms (Table S1) we identified clusters of orthologous groups (COGs) using the EdgeSearch
309 algorithm[41]. Following previous studies, COGs were defined so that any two proteins from different
310 species and members of the same COG were considered orthologs[42]. COGs were obtained using the
311 COGsoft software[41]. Specifically, starting from an all-against-all psi-blast[43] search, setting the
312 database size at 10^8, and using a maximum considered E-value of 0.1. In order to obtain the largest
313 number of likely orthologs we did not apply a filter on low complexity or composition-based statistics.
314 Only proteins sharing the same EC number and assigned to the same COG were compared, and only COGs
315 with sequences in 10 or more of the 22 species were used.

316 In order to exclude proteins showing evidence of Horizontal Gene Transfer (HGT), we constructed
317 a maximum likelihood phylogenetic tree of the 12 prokaryotes used in our analysis using a concatenated
318 alignment of marker genes[44]. The tree was manually compared to the individual trees of the prokaryotic
319 sequences sharing EC annotations within each COG; COG-specific trees were built using the GAMMA
320 model of amino-acid substitution implemented in the RAxML software[45]. Proteins that showed clear
321 differences in tree topologies, suggesting HGT, were excluded from the analysis.

322

323 **Models of long-term protein sequence evolution.** Divergence times between organisms were obtained
324 from the TimeTree database[46]. Specifically, we used the mean divergence times between pairs of
325 species across published studies (November, 2015). Global sequence identities for pairs of proteins
326 annotated with the same molecular function in the same COG were calculated from pairwise alignments
327 with ClustalW2[47]. Sequence identity was computed as the number of identical sites at aligned positions,
328 divided by the total number of aligned sites, i.e. excluding gaps. In the analysis we only considered EC
329 numbers with pairs of orthologs showing at least 5 different divergence times distributed across 4 billion
330 years.

331 Models were fitted to equations 1 to 3 using the least-squares minimization algorithm
332 implemented in the MATLAB R2017a fitnlm function (The MathWorks, Inc., Natick, MA). The best fits of
333 the model 1 and model 2 were compared using the F-test. The minimum sequence identity parameter in
334 the model 2 (Y_0, from equation 2) was compared, for each enzymatic activity, to the average global
335 sequence identity between unrelated protein pairs using the Wald test.

336 To assess the effect of computational annotations on the observed divergence results, we
337 repeated the analysis using only sequences with experimentally validated molecular functions. To keep
338 only sequences with validated molecular functions, we manually reviewed published references for
339 enzyme annotations in the BRENDA database[23], and discarded any annotations that were based
340 exclusively on computational methods or high-throughput studies. After filtering the annotations, we only
341 considered EC numbers with pairs of orthologs showing at least 4 different divergence times distributed
342 across 4 billion years.

343

344 **Calculation of the divergence rate**. Based on Model 3, we determined the divergence rate at a given
345 divergence time *t* by solving for the derivative of Equation 3 with respect to time:

346

347 $$\frac{dy}{dt} = \frac{d\left(100*\left(\frac{R_0*t}{\alpha}+1\right)^{-\alpha}\right)}{dt} = -100 * R_0 \left(\frac{R_0*t}{\alpha}+1\right)^{-\alpha-1} \quad (5)$$

348

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

351

352 **FolA competition experiment in *E. coli*.** To perform competition experiments we used the EcNR2 strain
353 derived from *E. coli* K12 MG1655. Mutagenesis was performed using Multiplex Automated Genomic

354 Engineering (MAGE), as previously described[26]. 90 bp DNA oligomers were designed around each folA
355 codon using the MG1655 wild type sequence as reference. For each codon, all possible nucleotide variants
356 were synthesized. To avoid simultaneous mutations of multiple codons, cells were transformed targeting
357 ten consecutive codons at a time. After four rounds of electroporation, cells were recovered and pooled
358 together at approximately the same concentration based on cell counts. Two competition growth
359 experiments were carried out, one for each half of the protein. For the competition experiments, cells
360 were grown in LB media in a turbidostat while maintaining constant volume and cell density. Samples
361 were taken every 2 hours for a period of 16 hours, spun down, washed in PBS, spun down again and stored
362 at -20°C until all samples were collected. For each competition, the corresponding FolA region was
363 amplified through PCR while assigning a specific DNA barcode for each time point. PCR products were
364 then pooled and paired-end sequenced using the MiSeq Reagent Kit 2 from Illumina. Sequence reads were
365 deposited to the SRA database with accession number: SRP152339.

366 To determine, at each time point, the abundance of each mutant relative to wild type, we joined
367 paired-end reads using SeqPrep (v 1.1) and aligned the joined reads to the folA gene sequence using
368 Bowtie2[48]. We then counted the number of reads per mutant using a custom script. Reads with more
369 than a single mutated codon were discarded. Counts were median-normalized to control for noise due to
370 mutagenesis performed in batches of 10 codons. At each time point we calculated the ratio R_t of mutant
371 to wild type (WT) reads. In exponential growth, the growth rate difference between a given mutant and
372 WT was calculated based on the slope of $\ln(R_\tau)$ as a function of time:

373

374 $$\ln(R_t) = (m_i - m_{wt}) * t + \ln(R_0)$$

375

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

380 To obtain a single value characterizing the effect of all possible mutations at a protein site, we
381 first averaged the relative growth rates of mutants resulting in the same amino acid change. We then
382 calculated the average fitness effect of mutations at each protein site by averaging across 20 possible
383 amino acids substitutions (Dataset S1).

384

385 **Divergence of sites with different fitness effects.** To investigate divergence of sites with different fitness
386 effects, we obtained sequences of FolA and InfA bacterial orthologs from the COG database[27]. The FolA
387 orthologs annotated with the same EC number in UniProt (n=311) and the InfA orthologs annotated with
388 the same KO number in KEGG (n=514) were used to build multiple sequence alignments with
389 ClustalW2[47]. We then calculated the genetic distance between all pairs of species with aligned
390 sequences based on the corresponding 16S rRNA genes obtained from the GreenGenes database[49]
391 (October, 2016); the F84 model of nucleotide substitution implemented in the phylip package[50] was
392 used to compute the genetic distances. Divergence times were estimated from the 16S genetic distances
393 based on the linear regression of literature reported divergence times[46] (Table S1). In each divergence
394 bin (Figure 2), sequence identities of sites with different average fitness effects (represented by different
395 colors) were averaged across all pairs of orthologs.
396 To find phylogenetically independent pairs of species with specific divergence times, we used the
397 16S rRNA gene alignments for species with FolA and InfA orthologs. We built maximum likelihood
398 phylogenetic trees based on the 16S alignments using RAxML[45]. We then applied the Maximum Pairing
399 Problem approach by Arnold *et al*.[28] to find, at specific divergence times, the largest number of edge-
400 disjoint pairs of species.
401
402 **Analysis of global protein structural evolution***.* To study the divergence of protein structures as a function
403 of time, we obtained PDB codes for all proteins associated with EC numbers in the BRENDA database[23].
404 We then selected for the analysis species with solved enzyme structures for at least 10 different EC
405 numbers. Psi-blast searches with a conservative E-value cutoff of 10^{-6} were used to identify orthologs
406 (defined as bi-directional best hits) in the selected species. The 3D structures of orthologous pairs,
407 annotated with the same EC number, were aligned using the TM-align program[30] to obtain the C-alpha
408 RMSD values. Pairs of proteins were not considered if more than 70% of the residues of the shortest
409 protein could not be structurally aligned. We also removed from the analysis pairs of structures with
410 flexibility between domains, as they could result in large RMSD values despite significant structural
411 similarity. To identify such proteins we used the FATCAT[51] software to perform flexible structural
412 alignments of all structure pairs. We then specifically filtered the pairs that were split into two or more
413 domains by the FATCAT alignments.
414
415 **Analysis of the enzyme active sites**. To analyze divergence as a function of active site distance we used
416 protein sequences associated with the 64 EC numbers and 22 species considered in Figure 1 and Figure

S1. To that end, PDB[32] 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[33], PDB and UniProt-KB[15]. 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[47] 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.

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.
6. Rost, B. (2002). Enzyme function less conserved than anticipated. Journal of molecular biology *318*, 595-608.
7. 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.
8. 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.
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. Lehninger, A.L., Nelson, D.L., and Cox, M.M. (2013). Lehninger principles of biochemistry, 6th Edition, (New York: W.H. Freeman).
14. 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.

15. UniProt, C. (2015). UniProt: a hub for protein information. Nucleic acids research *43*, D204-212.
16. Bairoch, A. (1999). The ENZYME data bank in 1999. Nucleic acids research *27*, 310-311.
17. 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.
18. Fares, M.A., Byrne, K.P., and Wolfe, K.H. (2006). Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species. Molecular biology and evolution *23*, 245-253.
19. Fletcher, W., and Yang, Z. (2009). INDELible: a flexible simulator of biological sequence evolution. Molecular biology and evolution *26*, 1879-1888.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.
29. 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.
30. Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research *33*, 2302-2309.
31. 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.
32. 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.
33. 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.
34. Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. Cell *138*, 774-786.
35. 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.

505 36. Lunzer, M., Golding, G.B., and Dean, A.M. (2010). Pervasive cryptic epistasis in molecular
506 evolution. PLoS Genet *6*, e1001162.
507 37. Chan, Y.H., Venev, S.V., Zeldovich, K.B., and Matthews, C.R. (2017). Correlation of fitness
508 landscapes from three orthologous TIM barrels originates from sequence and structure
509 constraints. Nat Commun *8*, 14614.
510 38. Pingoud, A., Wilson, G.G., and Wende, W. (2014). Type II restriction endonucleases--a historical
511 perspective and more. Nucleic acids research *42*, 7489-7527.
512 39. Shendure, J., and Akey, J.M. (2015). The origins, determinants, and consequences of human
513 mutations. Science *349*, 1478-1483.
514 40. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference
515 resource for gene and protein annotation. Nucleic acids research *44*, D457-462.
516 41. Kristensen, D.M., Kannan, L., Coleman, M.K., Wolf, Y.I., Sorokin, A., Koonin, E.V., and Mushegian,
517 A. (2010). A low-polynomial algorithm for assembling clusters of orthologous groups from
518 intergenomic symmetric best matches. Bioinformatics *26*, 1481-1487.
519 42. Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families.
520 Science *278*, 631-637.
521 43. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997).
522 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic
523 acids research *25*, 3389-3402.
524 44. Wu, M., and Scott, A.J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with
525 AMPHORA2. Bioinformatics *28*, 1033-1034.
526 45. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
527 phylogenies. Bioinformatics *30*, 1312-1313.
528 46. Hedges, S.B., Dudley, J., and Kumar, S. (2006). TimeTree: a public knowledge-base of divergence
529 times among organisms. Bioinformatics *22*, 2971-2972.
530 47. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin,
531 F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0.
532 Bioinformatics *23*, 2947-2948.
533 48. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods
534 *9*, 357-359.
535 49. DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D.,
536 Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and
537 workbench compatible with ARB. Appl Environ Microbiol *72*, 5069-5072.
538 50. Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the
539 author: Department of Genome Sciences, University of Washington, Seattle.
540 51. Ye, Y., and Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs
541 allowing twists. Bioinformatics *19 Suppl 2*, ii246-255.

542

Table S1. Model species selected and pairwise average divergence times [a] (in billions of years)[46]

	Eukarya										Bacteria									Archaea		
	Homo sapiens	*Rattus norvegicus*	*Xenopus laevis*	*Xenopus laevis Danio rerio*	*Gallus gallus*	*Drosophila melanogaster*	*Saccharomyces cerevisiae*	*Caenorhabditis elegans*	*Arabidopsis thaliana*	*Oryza sativa*	*Bacillus subtilis*	*Thermotoga maritima*	*Pseudomonas aeruginosa*	*Mycobacterium tuberculosis*	*Yersinia pestis*	*Haemophilus influenzae*	*Salmonella enterica*	*Escherichia coli*	*Agrobacterium fabrum*	*Pyrococcus horikoshii*	*Sulfolobus solfataricus*	*Haloferax volcanii*
H.sapiens	0.0																					
R. norvegicus	0.1	0.0																				
X. laevis	0.4	0.4	0.0																			
D. rerio	0.4	0.4	0.4	0.0																		
G. gallus	0.3	0.3	0.4	0.4	0.0																	
D. melanogaster	0.8	0.8	0.8	0.8	0.8	0.0																
S. cerevisiae	1.3	1.3	1.3	1.3	1.3	1.3	0.0															
C. elegans	0.8	0.8	0.8	0.8	0.8	0.7	1.3	0.0														
A. thaliana	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	0.0													
O. sativa	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	0.2	0.0												
B. subtilis	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	0.0											
T. maritima	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	0.0										
P. aeruginosa	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.1	4.0	0.0									
M. tuberculosis	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.1	4.0	3.1	0.0								
Y. pestis	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.1	4.0	1.4	3.1	0.0							
H. influenzae	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.1	4.0	1.4	3.1	0.7	0.0						
S. enterica	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.1	4.0	1.4	3.1	0.4	0.7	0.0					
E. coli	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.1	4.0	1.4	3.1	0.4	0.7	0.1	0.0				
A. fabrum	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.1	4.0	2.5	3.1	2.5	2.5	2.5	2.5	0.0			
P. horikoshii	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	0.0		
S. solfataricus	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.8	0.0	
H. volcanii	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	3.7	3.8	0.0

[a] Divergence times were obtained from literature references in the TimeTree database[46]

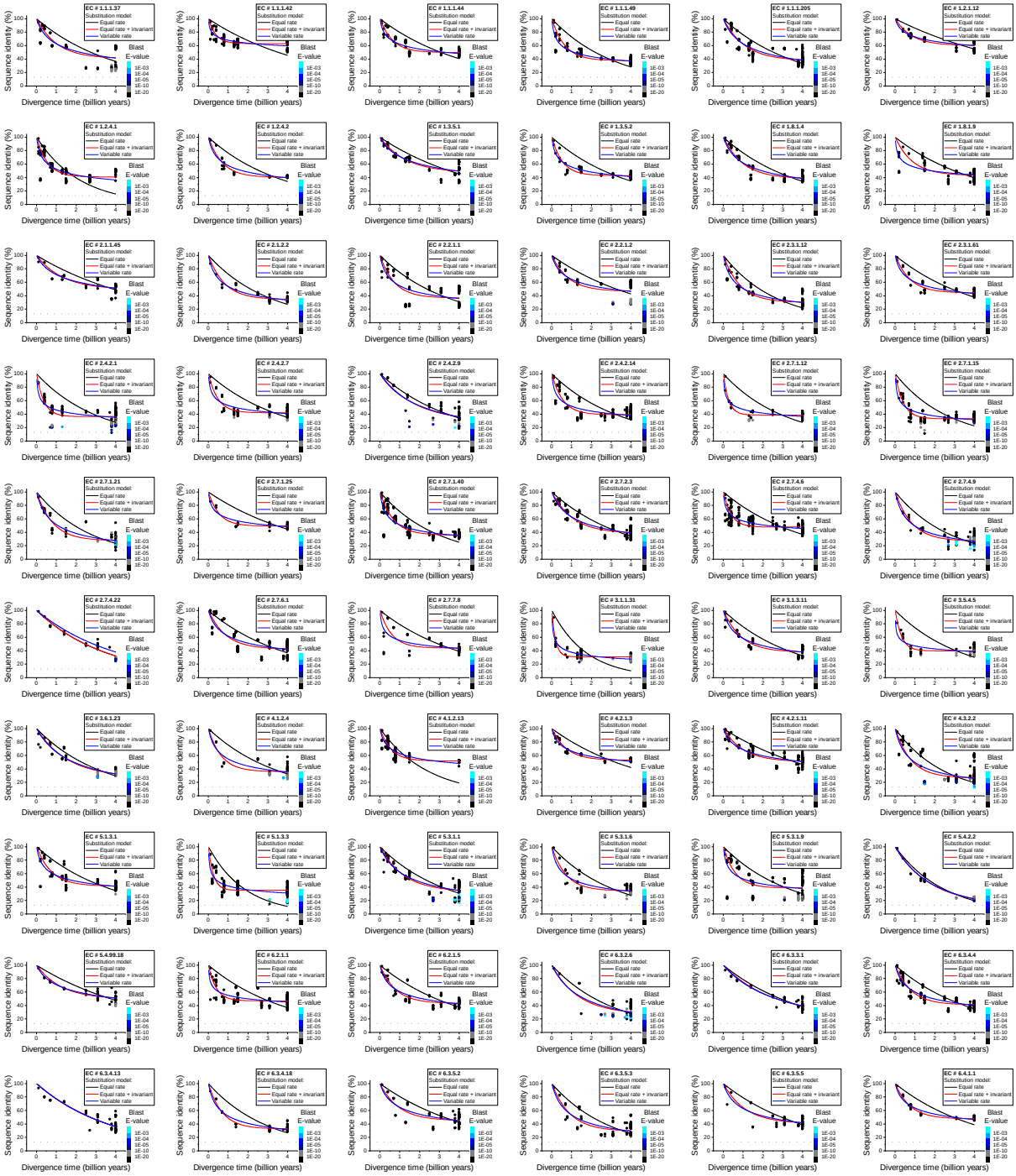

Figure S1. The global pairwise sequence identities between pairs of orthologous enzymes are shown as a function of divergence times between the corresponding species. The colored lines indicate fits to the data using the three models of amino acid substitution (Equations 1-3). The dashed black lines represent the average sequence identity for the alignment of unrelated protein sequences. Dot colors indicate the psi-blast E-value of the pairwise alignments for a database size of 3.8×10^5 sequences. Fit parameters are presented in Table S2.

Table S2. Fitted parameters and test results for 64 EC numbers.

EC number	Model1		Model2			Model3			Model 1 Vs. Model 2	Y_0 Wald test
	R_0	R^2	Y_0	R_0	R^2	R_0	α	R^2	F test P-value	P-value[a]
1.1.1.37	0.25	0.30	40.95	0.97	0.51	0.96	0.35	0.51	6.0E-15	<1.0E-20
1.1.1.42	0.20	-2.41	63.12	2.70	0.46	5.32	0.09	0.48	9.2E-97	<1.0E-20
1.1.1.44	0.22	-0.68	49.99	1.59	0.71	2.12	0.19	0.62	3.4E-72	<1.0E-20
1.1.1.49	0.31	-0.36	37.81	1.52	0.58	3.51	0.24	0.76	1.6E-68	<1.0E-20
1.1.1.205	0.27	0.50	37.59	0.93	0.79	0.80	0.46	0.78	2.0E-76	<1.0E-20
1.2.1.12	0.15	-0.47	60.46	1.20	0.68	1.06	0.15	0.68	6.1E-57	<1.0E-20
1.2.4.1	0.46	-0.25	41.05	1.83	0.58	2.60	0.27	0.54	3.9E-49	<1.0E-20
1.2.4.2	0.27	0.09	40.46	1.43	0.92	1.55	0.30	0.86	1.5E-29	<1.0E-20
1.3.5.1	0.20	0.48	47.46	0.77	0.73	0.58	0.34	0.73	1.4E-28	<1.0E-20
1.3.5.2	0.26	-0.74	43.15	1.87	0.83	2.85	0.22	0.82	1.8E-53	<1.0E-20
1.5.1.3	0.33	0.24	33.99	1.69	0.88	1.91	0.36	0.84	2.1E-27	<1.0E-20
1.8.1.4	0.27	0.41	38.56	1.01	0.83	0.91	0.42	0.78	2.8E-68	<1.0E-20
1.8.1.9	0.23	-0.26	43.23	1.15	0.17	5.61	0.17	0.34	2.4E-16	<1.0E-20
2.1.1.45	0.18	0.50	49.39	0.77	0.63	0.62	0.30	0.65	7.2E-06	<1.0E-20
2.1.2.2	0.33	0.38	34.56	1.24	0.82	1.54	0.38	0.87	1.4E-13	<1.0E-20
2.2.1.1	0.29	0.12	36.80	1.38	0.44	1.67	0.34	0.43	4.8E-18	<1.0E-20
2.2.1.2	0.21	0.13	46.70	1.19	0.50	1.12	0.27	0.48	5.4E-14	<1.0E-20
2.3.1.12	0.38	0.23	30.07	1.19	0.76	1.68	0.43	0.79	7.1E-49	<1.0E-20
2.3.1.61	0.24	-0.05	44.35	1.35	0.81	1.53	0.26	0.76	1.9E-29	<1.0E-20
2.4.2.1	0.32	-1.75	36.56	2.77	0.43	9.70	0.20	0.34	8.6E-63	<1.0E-20
2.4.2.7	0.26	-1.45	42.23	2.12	0.62	5.11	0.19	0.48	7.0E-49	<1.0E-20
2.4.2.9	0.27	0.58	28.88	0.56	0.62	0.46	0.94	0.62	3.9E-03	6.50E-04
2.4.2.14	0.29	-1.00	39.38	1.82	0.60	3.90	0.23	0.51	1.1E-113	<1.0E-20
2.7.1.12	0.32	-2.50	38.83	3.73	0.81	10.78	0.18	0.61	2.6E-35	<1.0E-20
2.7.1.15	0.37	-2.29	33.36	2.52	0.71	7.85	0.23	0.50	9.7E-112	<1.0E-20
2.7.1.21	0.36	0.58	27.70	1.23	0.81	1.27	0.56	0.82	1.6E-19	<1.0E-20
2.7.1.25	0.20	-0.71	49.51	1.88	0.85	3.05	0.17	0.73	3.6E-22	<1.0E-20
2.7.1.40	0.34	-0.43	35.98	1.49	0.66	2.65	0.29	0.66	8.1E-179	<1.0E-20
2.7.2.3	0.27	0.50	38.49	0.94	0.82	0.82	0.44	0.79	1.3E-70	<1.0E-20
2.7.4.3	0.24	0.16	42.88	1.16	0.82	1.27	0.29	0.81	1.6E-77	<1.0E-20
2.7.4.6	0.25	-1.11	47.99	2.58	0.40	5.26	0.16	0.57	2.2E-123	<1.0E-20
2.7.4.9	0.36	0.27	26.95	1.19	0.68	1.50	0.52	0.66	9.4E-26	<1.0E-20
2.7.4.22	0.28	0.93	0.00	0.28	0.92	0.24	10.0	0.85	N.S.	N.S
2.7.6.1	0.25	0.13	42.55	1.15	0.60	1.11	0.32	0.54	2.8E-55	<1.0E-20
2.7.7.8	0.23	-0.18	44.23	1.70	0.24	4.87	0.18	0.40	3.5E-10	<1.0E-20
3.1.1.31	0.57	-3.32	31.23	5.33	0.67	15.70	0.23	0.67	1.6E-63	<1.0E-20
3.1.3.11	0.28	0.31	37.89	1.11	0.79	1.28	0.35	0.79	7.7E-32	<1.0E-20
3.5.4.5	0.30	-4.58	39.97	3.18	0.62	29.28	0.14	0.34	9.0E-52	<1.0E-20
3.6.1.23	0.30	0.77	27.28	0.64	0.85	0.55	0.92	0.85	2.1E-10	1.59E-12
4.1.2.4	0.28	0.32	36.16	1.55	0.63	1.56	0.36	0.63	9.6E-09	<1.0E-20
4.1.2.13	0.41	-0.69	52.02	1.80	0.68	1.79	0.20	0.66	1.6E-63	<1.0E-20
4.2.1.2	0.18	-0.50	53.09	1.35	0.47	1.85	0.17	0.43	4.1E-48	<1.0E-20
4.2.1.3	0.22	-0.26	52.94	1.35	0.92	1.32	0.20	0.90	3.7E-71	<1.0E-20
4.2.1.11	0.19	0.23	50.69	0.99	0.73	0.91	0.24	0.72	4.6E-103	<1.0E-20
4.3.2.2	0.40	0.21	26.73	1.25	0.53	1.38	0.57	0.48	7.8E-28	<1.0E-20
5.1.3.1	0.26	-0.16	41.62	1.33	0.56	2.00	0.25	0.57	1.2E-40	<1.0E-20
5.1.3.3	0.54	-2.23	35.37	4.01	0.35	12.77	0.21	0.45	9.3E-95	<1.0E-20
5.3.1.1	0.29	0.53	33.27	0.84	0.70	0.73	0.58	0.68	2.7E-27	<1.0E-20
5.3.1.6	0.32	-0.28	34.56	1.40	0.67	1.94	0.34	0.59	3.2E-31	<1.0E-20
5.3.1.9	0.30	-0.33	38.80	1.66	0.25	3.41	0.24	0.22	4.2E-38	<1.0E-20
5.4.2.2	0.40	0.94	13.27	0.56	0.96	0.52	2.30	0.96	1.7E-07	N.S
5.4.99.18	0.18	0.54	49.56	0.80	0.72	0.65	0.29	0.75	1.8E-07	<1.0E-20
6.2.1.1	0.25	-1.88	47.10	2.58	0.44	8.88	0.14	0.35	1.6E-82	<1.0E-20
6.2.1.5	0.24	0.09	43.31	1.28	0.74	1.37	0.28	0.70	4.3E-43	<1.0E-20
6.3.2.6	0.32	0.53	26.81	0.74	0.60	0.62	0.91	0.58	1.7E-04	4.25E-05
6.3.3.1	0.24	0.83	24.72	0.41	0.84	0.35	1.23	0.85	5.7E-03	2.34E-02
6.3.4.2	0.22	-0.22	48.23	1.35	0.81	1.65	0.22	0.81	1.1E-183	<1.0E-20
6.3.4.4	0.26	0.31	41.12	1.26	0.87	1.23	0.33	0.84	3.4E-156	<1.0E-20
6.3.4.13	0.25	0.77	0.00	0.25	0.76	0.26	20.84	0.76	N.S.	N.S
6.3.4.18	0.33	0.24	32.78	1.44	0.91	1.69	0.40	0.85	2.2E-21	<1.0E-20
6.3.5.2	0.21	0.32	44.93	1.06	0.59	0.95	0.31	0.57	9.7E-15	<1.0E-20
6.3.5.3	0.34	0.24	29.22	1.09	0.56	1.20	0.53	0.53	3.5E-18	<1.0E-20
6.3.5.5	0.23	0.40	40.73	0.92	0.63	0.94	0.36	0.63	4.3E-11	<1.0E-20
6.4.1.1	0.24	-0.18	47.69	1.34	0.96	1.28	0.25	0.88	9.1E-47	<1.0E-20

[a] P-values <0.05 indicate Y_0 is significantly higher than 13.5%, the average identity between unrelated sequences

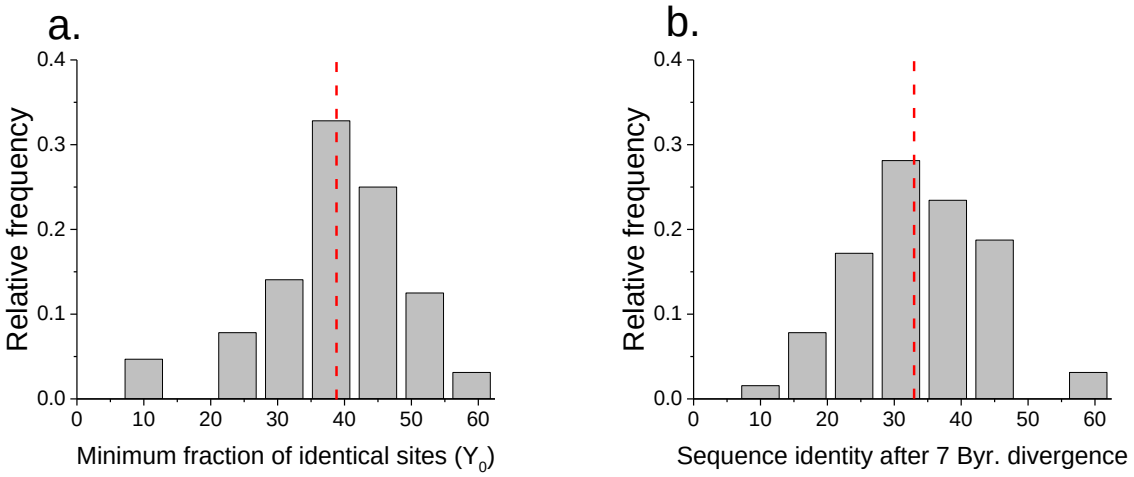

Figure S2. a. The distribution of Y_0 parameter values across 64 EC numbers for Model 2 fits (Equation 2). The Y_0 parameter represents the minimum fraction of protein sites that remain identical at long divergence times. **b.** The distribution across 64 EC numbers of the projected global sequence identity after 7 billion years of divergence according to Model 3 (Equation 3). In both panels the dashed red lines represent the corresponding median values (39% and 33% for a. and b., respectively).

Table S3. Estimated evolutionary rates for pairs of proteins according to Model 3 fits.

EC number	Initial rate (R_0)	Rate at 4 Byr.	Fold-reduction in rate
1.1.1.37	0.96	0.03	28.3
1.1.1.42	5.32	0.01	377.0
1.1.1.44	2.12	0.02	94.5
1.1.1.49	3.51	0.02	158.5
1.1.1.205	0.80	0.04	20.7
1.2.1.12	1.06	0.02	48.9
1.2.4.1	2.60	0.02	106.2
1.2.4.2	1.55	0.03	54.5
1.3.5.1	0.58	0.04	15.6
1.3.5.2	2.85	0.02	126.3
1.5.1.3	1.91	0.03	67.7
1.8.1.4	0.91	0.04	25.1
1.8.1.9	5.61	0.02	307.5
2.1.1.45	0.62	0.03	18.3
2.1.2.2	1.54	0.03	50.7
2.2.1.1	1.67	0.03	58.0
2.2.1.2	1.12	0.03	38.3
2.3.1.12	1.68	0.03	55.6
2.3.1.61	1.53	0.03	56.5
2.4.2.1	9.70	0.02	561.5
2.4.2.7	5.11	0.02	264.6
2.4.2.9	0.46	0.06	8.3
2.4.2.14	3.90	0.02	182.9
2.7.1.12	10.78	0.02	643.2
2.7.1.15	7.85	0.02	426.1
2.7.1.21	1.27	0.03	36.9
2.7.1.25	3.05	0.02	151.4
2.7.1.40	2.65	0.02	107.4
2.7.2.3	0.82	0.04	21.5
2.7.4.3	1.27	0.03	43.2
2.7.4.6	5.26	0.02	289.7
2.7.4.9	1.50	0.03	46.7
2.7.4.22	0.24	0.09	2.6
2.7.6.1	1.11	0.03	35.3
2.7.7.8	4.87	0.02	256.4
3.1.1.31	15.70	0.02	996.1
3.1.3.11	1.28	0.03	40.8
3.5.4.5	29.28	0.01	2149.6
3.6.1.23	0.55	0.05	10.3
4.1.2.4	1.56	0.03	52.5
4.1.2.13	1.79	0.02	75.4
4.2.1.2	1.85	0.02	85.3
4.2.1.3	1.32	0.02	53.2
4.2.1.11	0.91	0.03	31.6
4.3.2.2	1.38	0.03	41.3
5.1.3.1	2.00	0.03	78.7
5.1.3.3	12.77	0.02	774.3
5.3.1.1	0.73	0.04	17.1
5.3.1.6	1.94	0.03	69.9
5.3.1.9	3.41	0.02	153.5
5.4.2.2	0.52	0.06	8.4
5.4.99.18	0.65	0.03	19.4
6.2.1.1	8.88	0.02	548.7
6.2.1.5	1.37	0.03	47.5
6.3.2.6	0.62	0.05	12.4
6.3.3.1	0.35	0.06	5.3
6.3.4.2	1.65	0.02	66.5
6.3.4.4	1.23	0.03	39.6
6.3.4.13	0.26	0.09	2.9
6.3.4.18	1.69	0.03	56.7
6.3.5.2	0.95	0.03	29.4
6.3.5.3	1.20	0.04	34.4
6.3.5.5	0.94	0.03	27.8
6.4.1.1	1.28	0.03	46.4

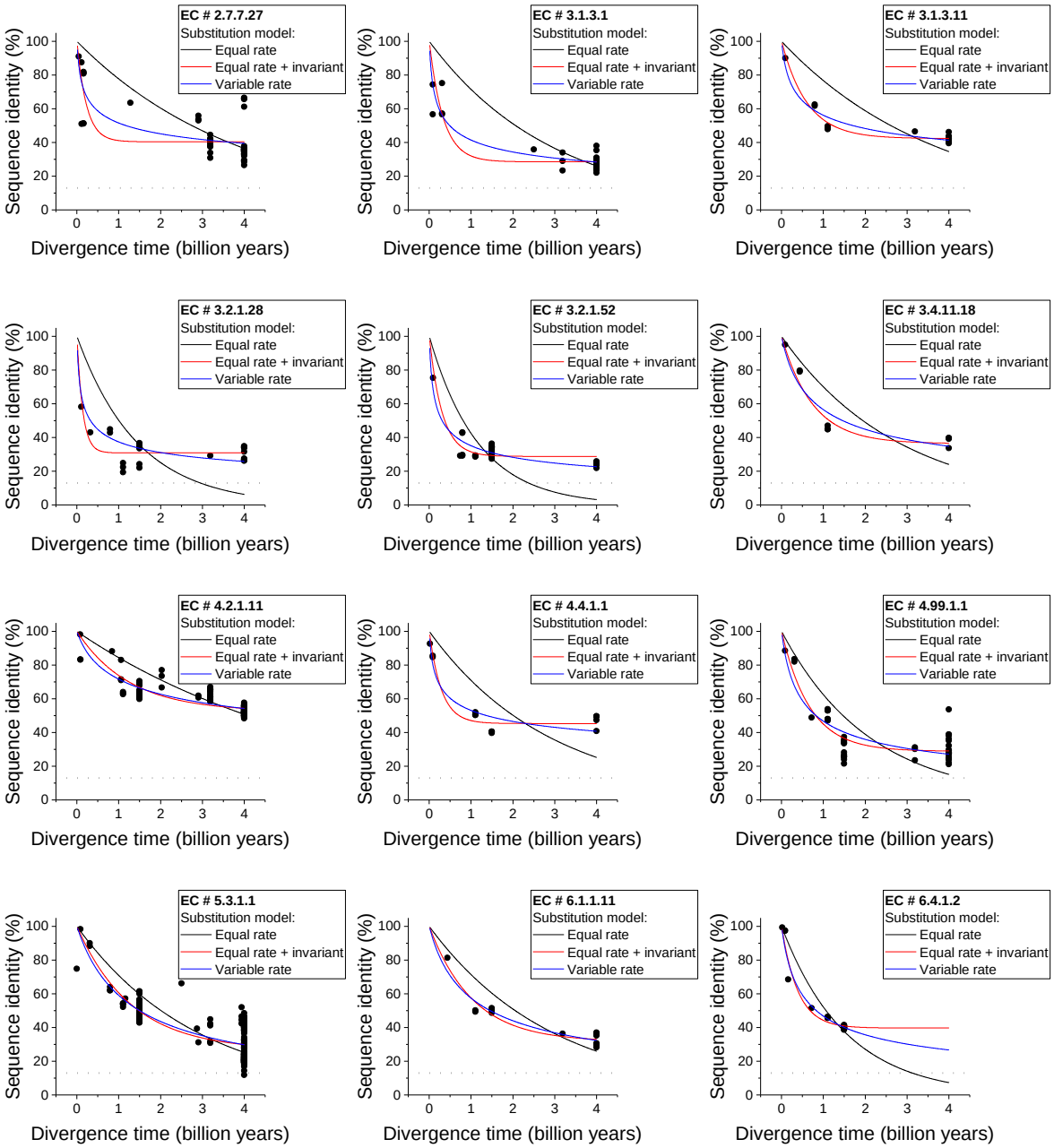

Figure S3. Sequence identity as a function of divergence time for orthologous enzymes annotated with the same EC number and experimentally validated molecular functions. The color lines indicate fits to the data for the three models of amino acid substitution (Equations 1-3). Results are shown for 12 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).

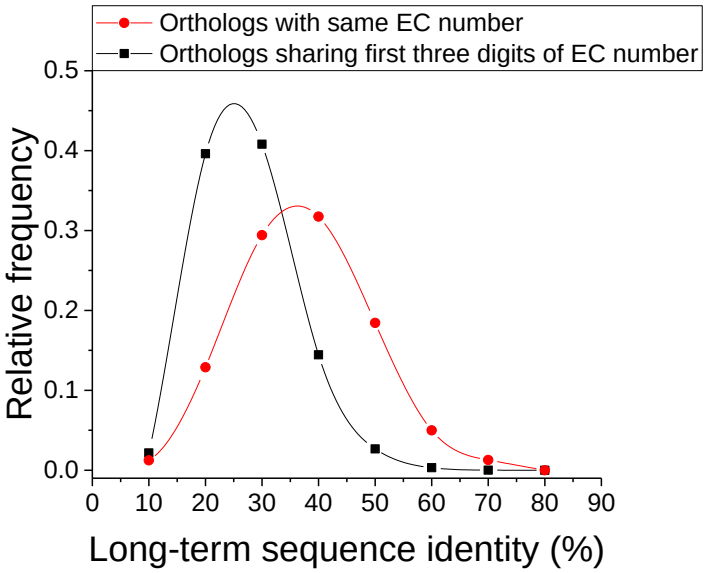

Figure S4. The distribution of sequence identities at long evolutionary distances between orthologous pairs of enzymes annotated (based on UniProt) with the same EC number (red), and only sharing the first three digits of an EC number (black); points in different sequence identity bins are connected using a spline function (solid lines). The distributions of sequence identities were obtained using pairs of orthologs from species that diverged over two billion years ago. Results are based on all enzyme COGs present in the 22 species used to analyze global sequence divergence (Table S1). To avoid functional biases, orthologs with the same EC number (red, n=8646 pairs) were restricted to the same set of COGs as the ones used for orthologs sharing the first three digits of an EC number (black, n=9123 pairs). The two distributions are significantly different (mean: 37.5 vs. 27.8%, Mann-Whitney P-value $<10^{-20}$).