Molecular function limits divergent protein evolution on planetary timescales

Mariam M. Konaté[1,2†], Germán Plata[1†], Jimin Park[1, 3], Harris H. Wang[1, 3], and Dennis Vitkup[1, 4]

[1]Department of Systems Biology, Columbia University, New York, NY, USA. [2]Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD, USA. [3]Department of Pathology and Cell Biology, Columbia University, New York, NY, USA. [4]Department of Biomedical Informatics, Columbia University, New York, NY, USA. [†]These authors contributed equally to this work.

Correspondence to DV at dv2121@columbia.edu

Abstract

Sequence analyses suggest that even after billions of years of independent evolution, protein orthologs are still diverging from each other[1]. 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 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. By experimentally characterizing the growth rates of all amino acid substitutions in the *E. coli* FolA protein, we show that the observed slowdown is reflected in different long-term divergence patterns at sites with relatively small and large contributions to fitness. Furthermore, we find that the vast majority of enzyme orthologs with the same function are unlikely to diverge, at least on planetary timescales, beyond ~25% sequence identity. The observed divergence limit is usually well above the levels of detectible sequence homology, and functional conservation significantly constrains all protein regions.

As proteins evolve from a common ancestor, their sequences and structures diverge from each other[2]. Multiple previous studies have investigated the relationship between the conservation of protein molecular function and sequence identity[3-5]. In this work, we focused on a different question. Specifically, how far two sequences can 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. 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[6-8]. It has been also previously demonstrated that proteins with the same fold frequently diverge to very low (~10%) levels of sequence identity[9]. As only a relatively small fraction of all protein residues (~3-5%) are usually directly involved in catalysis[10], 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.

To investigate the evolution of proteins with the same molecular function, we used orthologous enzymes annotated in UniProt[11] 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[12], 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.

We started the analysis by considering two simple models of long-term protein sequence evolution. The first one, a model in which 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 this model, the substitution probabilities at all protein sites are identical, and there is no limit on sequence divergence. The second model, in addition to sites with equal and independent substitution rates, explicitly includes a minimum fraction of identical sites at long divergence times; Equation 2, where Y_0 represents the fraction of identical protein sites at long divergence times.

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

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

We used the two models to fit the sequence divergence of orthologs for each of the considered enzymatic functions (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. The data and model fits demonstrate a substantial decrease in the mutual divergence rates between orthologs after 1-2 billion years of evolution. 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 (supplementary Figure S2a).

The two aforementioned models over-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[13-15] 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 supplementary Figure S1) show that the rates of mutual protein sequence divergence between orthologous enzymes have decreased by a median of ~50 times during the ~4 billion years of evolution (see Methods and Table S3). Notably, this decrease is more than an order of magnitude higher compared to the expected slow-down that would result simply from the decreasing fraction of non-mutated sites in a model with equal and independent substitution rates., i.e. due to a slower divergence simply due to multiple substitutions at the same sites. 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 ~3 billion years from the present[16]) (supplementary Figure S2b).

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

Importantly, the observed divergence limit is not due to an inability to detect remote homologs, as it is reached at relatively high sequence identities (Figure 1 and Figure S1), with corresponding orthologs easily identified by computational sequence comparison methods. Furthermore, the results remained similar when we considered, using publications referenced in the BRENDA database[17], a subset of orthologous enzyme pairs with experimentally validated molecular functions (Supplementary Figure S3).

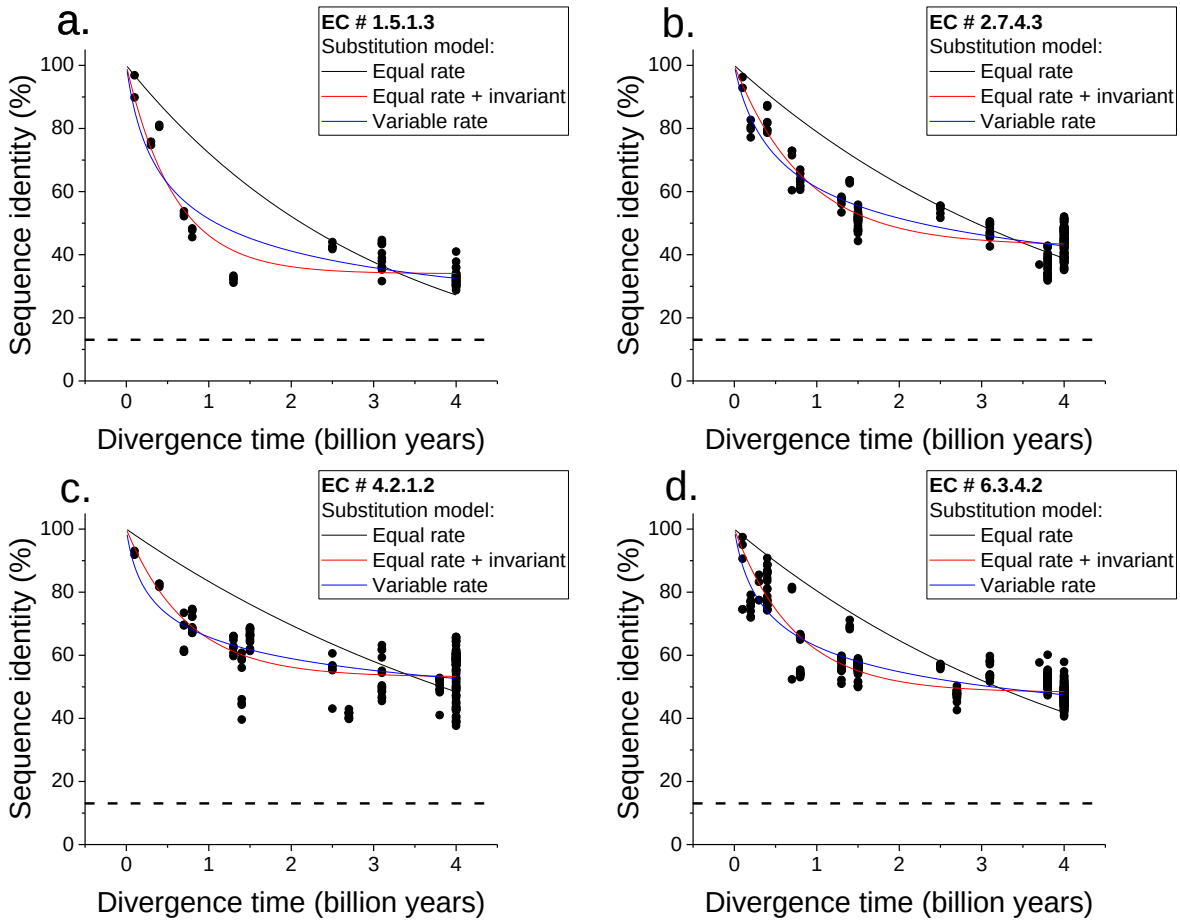


Figure 1. Sequence divergence of enzyme orthologs as a function of time. The global pairwise sequence identities between pairs of orthologous enzymes are shown as a function of divergence times between the corresponding species. Three models of amino acid substitution were used to fit the data. Model 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 all 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 supplementary Figure S1 and Table S2.

Next, to understand the contribution of different protein sites to long term 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[18]. Also, the long term protein sequence identity between FolA orthologs (~32%, see Figure 1a) is similar to other analyzed enzymes (Supplementary Figure S2a). To introduce every possible amino acid substitution at each protein site, we used the Multiplex Automated Genome Engineering (MAGE) approach[19]. We then

measured fitness effects (relative growth rates) of each variant by performing growth competition experiments, followed by next generation sequencing (see Methods, Supplementary Dataset 1).

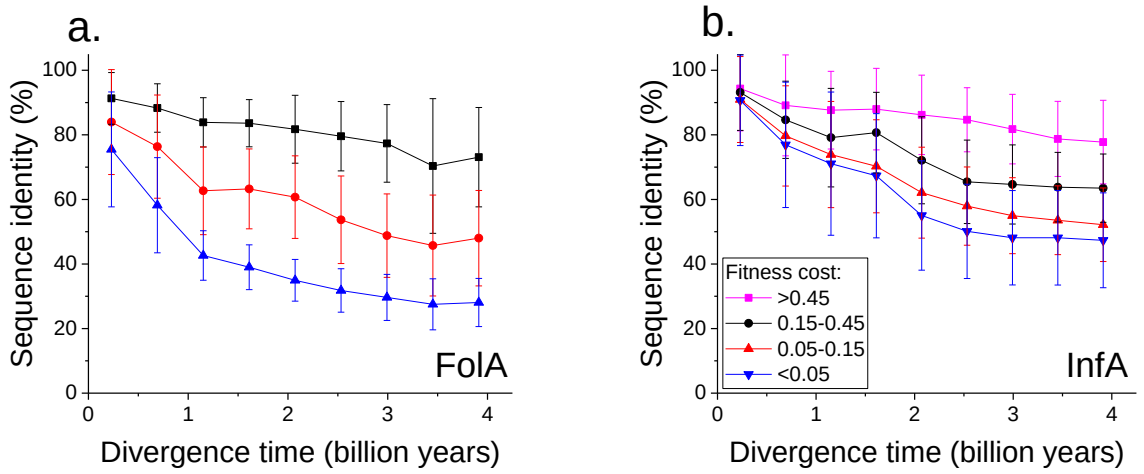


Figure 2. Sequence divergence of protein sites with different mutant fitness effects in *E. coli*. a. Sequence identity between pairs of FolA orthologs as a function of divergence time. Results are shown for subsets of protein sites (represented by different colors) with different average fitness effects of mutations in *E. coli*. The average sequence identity values were calculated using bacterial FolA orthologs in the COG database[20]; divergence times were estimated using bacterial 16S rRNA (see Methods). **b.** Similar to panel a, but for the sequence divergence of the *E. coli* translation initiation factor InfA orthologs. Error bars represent the standard deviation of sequence identity in each bin.

Using the experimental growth measurements in *E. coli*, we investigated how the divergence constraints vary between protein sites with different contribution to fitness. To that end, we used a large multiple sequence alignment of FolA orthologs from the COG database[20] (see Methods) to examine sequence divergence 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 how general the results obtained for FolA are, we used another comprehensive fitness dataset[21] for the *E. coli* translation initiation factor InfA (Figure 2b). In InfA we observed, on average, lower mutant growth rates and lower rates of sequence divergence. 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 significant 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.

To quantify the contribution of protein sites with different fitness effects to the long-term divergence constraints in both proteins, we calculated the probability that each site is occupied by the same amino acid in independent pairs of species after the significant decrease in mutual divergence rates (~2 billion years for FolA and ~2.5 billion years for InfA). To this end, we used the TARGETING package[22]

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 saturates for sites with high (>30%) fitness effects (Figure 3b).

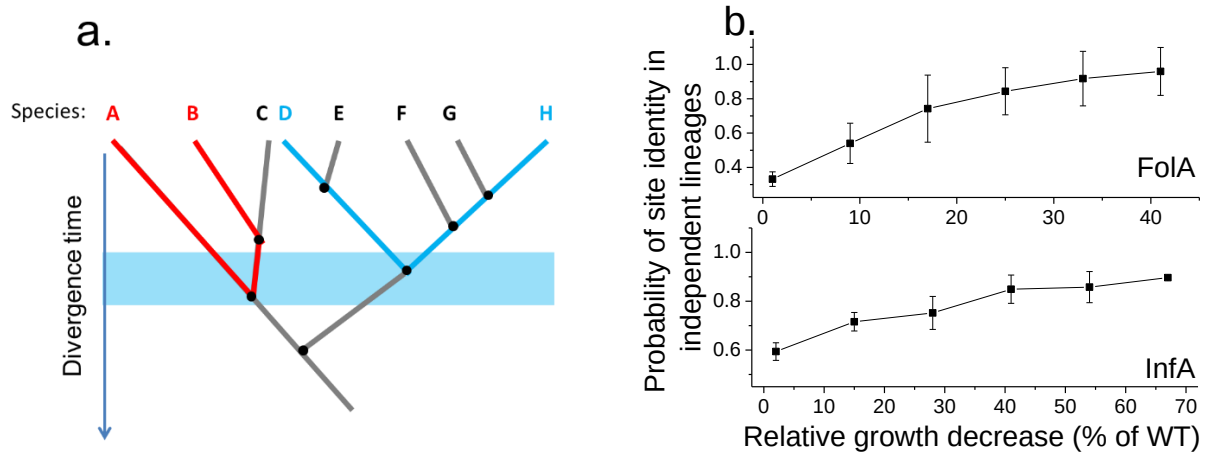


Figure 3. Probability of protein sites conservation in phylogenetically independent lineages. a. Illustration of species pairs representing phylogenetically independent lineages. In the figure, A-B and D-H are pairs of species that diverged within the specified time window (illustrated by the shaded region) and thus 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.

It is interesting to understand whether similar protein sites contribute to the divergence limit in different lineages. Towards that end, we calculated the distribution of the probabilities that protein sites tend to be occupied by identical amino acids in pairs of orthologs in phylogenetically independent linages (Figure 4). The results, both for 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. Instead, heterogeneous sets of protein 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.

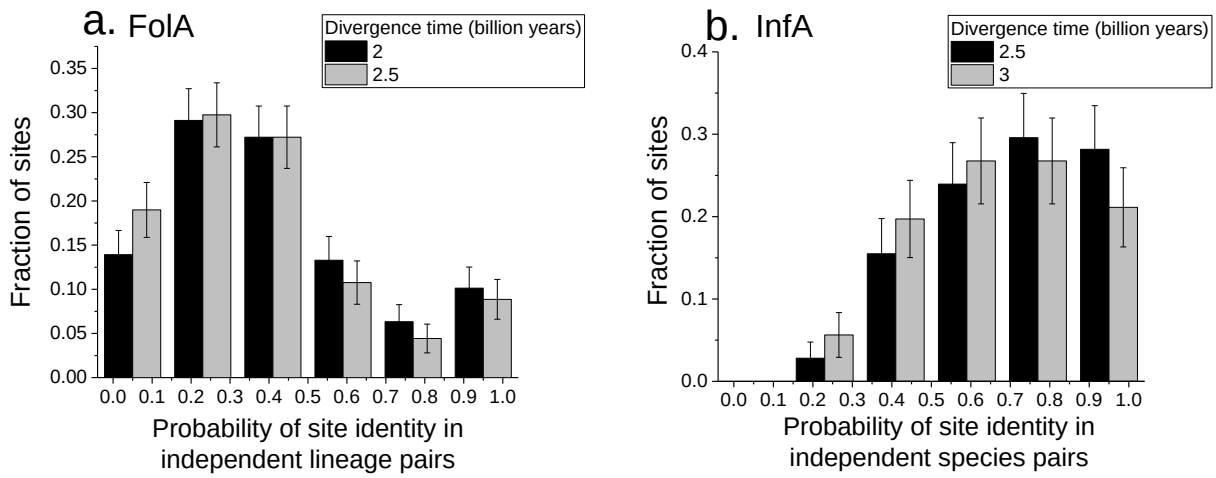


Figure 4. Contribution of different protein sites to the observed divergence limit. a. Distribution of FolA protein sites according to the probabilities of being identical in phylogenetically independent lineages. Probabilities were calculated as the proportion of phylogenetically independent pairs of 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[23,24]. To investigate the long-term structural divergence of proteins with the same molecular function, we structurally aligned orthologous enzymes using the TM-align algorithm[25]. 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.

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. To that end, we considered catalytic site annotations available from the Protein Data Bank[26], UniProt-KB[11] and the Catalytic Site Atlas[27] 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 used to study global sequence divergence (Figure 1 and Supplementary Figure S1). Although, as expected, residues close to the active site are most highly conserved[28], even residues distant from the active site display an effective divergence limit at long evolutionary distances. This result suggests that the spatial constraints required to conserve molecular function usually propagate throughout the entire protein structure.

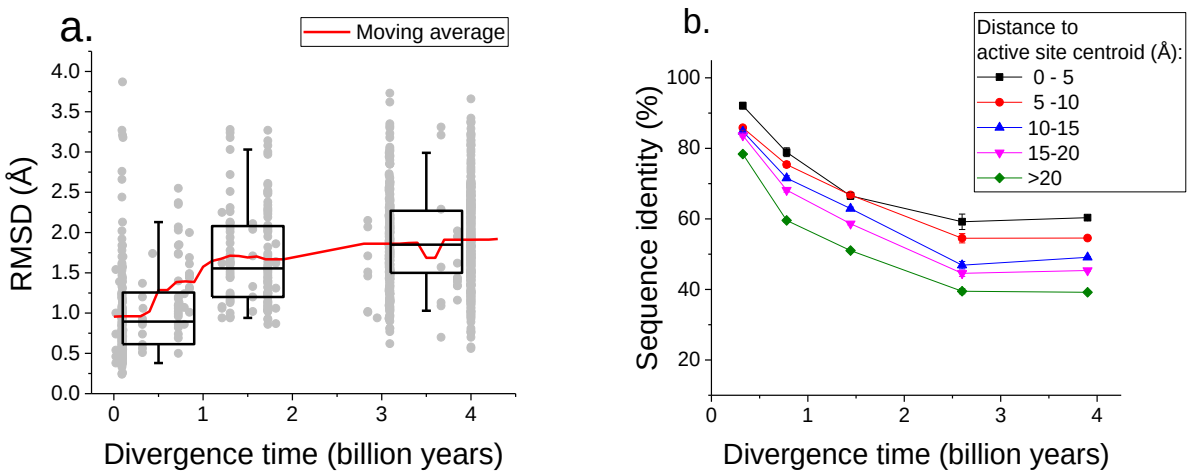


Figure 5. Long-term structural evolution of proteins with the same molecular function. a. The pairwise C-alpha root mean square deviation (RMSD) as a function of the divergence time between pairs of orthologs annotated with the same EC number and experimentally determined structures (gray dots). RMSD values were calculated based on structural alignments using the TM-align algorithm[25]. 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 Supplementary 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.

Overall, the presented results demonstrate that, in contrast to proteins with the same fold[9], the requirement to maintain the same molecular function significantly constrains the long-term divergence of protein orthologs. Moreover, our analysis delineates two general stages of divergence for ancient orthologs with the same molecular function. During the first stage, substitutions at sites with mild fitness effects lead to a relatively fast initial divergence. However, after 1-2 billion years of independent evolution, the divergence at these sites saturates. Notably, at saturation, the sequence identity between these sites is still substantially higher than the one expected between random protein sequences (~10%). This indicates that not all combinations of amino acids at the sites are compatible with maintaining the same molecular function, at least on planetary timescales. In contrast to sites with mild fitness effects, the divergence at sites with relatively high fitness effects is slow but approximately constant during ~4 billion years of evolution. The saturation at sites with mild fitness effects and slow divergence at sites with strong fitness effects, lead to significantly slower overall sequence and structural divergence during the second stage, i.e. after ~1-2 billon years of independent evolution.

We also demonstrate that the valley in sequence space explored in evolution by diverging enzyme orthologs is usually not defined by the strict conservation of the same protein sites in all lineages. In contrast, the results suggest that different sites are conserved in independent lineages. This result, at least partially, is likely due to the epistatic nature of protein sequence landscapes, in which mutations allowed

in one lineage are often prohibitively deleterious in another[29,30]. Notably, the sequence space valley accessible to diverging proteins is sensitive to the specificity of molecular function. Relaxation of functional specificity leads to a substantial increase in the level of sequence divergence. For example, orthologs sharing only the first three digits of the EC annotation, i.e. only a general class of substrates or cofactors[12], display a significantly lower long-term sequence identity than orthologs sharing all four EC digits (Figure S4). Similarly, in the evolution of restriction endonucleases, significant sequence identity between homologues is usually limited to isoschizomers, i.e. proteins specific to the same target sequence[31]. The observed divergence limit is likely a consequence of the requirement to preserve dynamical properties and relative spatial positions –usually to within a few angstroms– of protein residues participating in ligand binding and catalysis[10]. The conservation of the chemical structures of the substrates of orthologous enzymes, coupled with functional pressure for efficient catalysis, likely makes further drift in sequence space increasingly slow and inefficient. Further understanding of the biochemical and biophysical constraints responsible for limiting evolution of orthologs will improve protein design and engineering methods, which often allow to overcome these limits *in vitro*[32]. Analyses of function-specific drift in protein space will also contribute to understanding of how disease-associated mutations affect protein function[33].

Methods

Considered enzyme activities and corresponding protein orthologs. We selected for analysis the sequences annotated in UniProt[11] with EC numbers associated to the following metabolic pathways (as defined in the KEGG database[34]): Glycolysis and gluconeogenesis, pentose phosphate pathway, TCA cycle, purine metabolism and pyrimidine metabolism. Using protein sequences from 22 diverse organisms (Table S1) we identified clusters of orthologous groups (COGs) using the EdgeSearch algorithm[35]. Following previous studies, COGs were defined so that any two proteins from different species and members of the same COG were considered orthologs[36]. COGs were obtained using the COGsoft software[35]. Specifically, starting from an all-against-all psi-blast[37] search, setting the database size at 10^8, and using a maximum considered E-value of 0.1. In order 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 showing evidence of HGT, we constructed a maximum likelihood phylogenetic tree of the 12 prokaryotes used in our analysis using a concatenated alignment of marker genes[38]. The tree was manually compared to the individual trees of the prokaryotic sequences sharing EC annotations within each COG; COG-specific trees were built using the GAMMA model of amino-acid substitution implemented in the RAxML software[39]. Proteins that showed clear differences in tree topologies, suggesting HGT, were excluded from the analysis.

Models of long-term protein sequence evolution. Divergence times between organisms were obtained from the TimeTree database[40]. Specifically, we used the mean divergence times between pairs of species

across published studies (November, 2015). Global sequence identities for pairs of proteins annotated with the same molecular function in the same COG were calculated from pairwise alignments with ClustalW2[41]. 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. In the analysis we only considered EC numbers with pairs of orthologs showing at least 5 different divergence times distributed across 4 billion years.

Models were fitted to equations 1 to 3 using the Levenberg-Marquardt least-squares minimization algorithm implemented in the OriginPro software (OriginLab, Northampton, MA). The best fits of the model 1 and model 2 were compared using the F-test. The minimum sequence identity parameter in the 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 assess the effect of computational annotations on the observed divergence results, we repeated the analysis using only sequences with experimentally validated molecular functions. To keep only sequences with validated molecular functions, we manually reviewed published references for enzyme annotations in the BRENDA database[17], and discarded any annotations that were based exclusively on computational methods or high-throughput studies. After filtering the annotations, we only considered EC numbers with pairs of orthologs showing at least 4 different divergence times distributed across 4 billion years.

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

$$\frac{dy}{dt} = \frac{d\left(100*\left(\frac{R_0*t}{\alpha}+1\right)^{-\alpha}\right)}{dt} = -100 * R_0 \left(\frac{R_0*t}{\alpha}+1\right)^{-\alpha-1} \qquad (4)$$

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.

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[19]. 90 bp DNA oligomers were designed around each folA codon using the MG1655 wild type sequence as reference. For each codon, all possible nucleotide variants were 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 competitions 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.

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[42]. We then counted the number of reads per mutant using a custom script. 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 obtain 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 Dataset 1).

Divergence of sites with different fitness effects. To investigate divergence of sites with different fitness effects, we obtained sequences of FolA and InfA bacterial orthologs from the COG database[20]. The FolA orthologs annotated with the same EC number in UniProt and the InfA orthologs annotated with the same KO number in KEGG were used to build multiple sequence alignments with ClustalW2[41]. We then calculated the genetic distance between all pairs of species with aligned sequences based on the corresponding 16S rRNA genes obtained from the GreenGenes database[43] (October, 2016); the F84 model of nucleotide substitution implemented in the phylip package[44] was used to compute the genetic distances. Divergence times were estimated from the 16S genetic distances based on the linear regression of literature reported divergence times[40] (Table S1). In each divergence bin (Figure 2), sequence identities of sites with different average fitness effects (represented by different colors) were averaged across all pairs of orthologs.

To find phylogenetically independent pairs of species with specific divergence times, we used the 16S rRNA gene alignments for species with FolA and InfA orthologs. We built maximum likelihood phylogenetic trees based on the 16S alignments using RAxML[39]. We then applied the Maximum Pairing Problem approach by Arnold *et al.*[22] to find, at specific divergence times, the largest number of edge-disjoint pairs of species.

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[17]. We then selected for the analysis species with 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[25] 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 mobile domains, as they could result in large RMSD values despite significant structural similarity. To identify such proteins we used the FATCAT[45] software to perform flexible structural alignments of all structure pairs. We then specifically filtered the pairs that were split into two or more domains by the FATCAT alignments.

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 Supplementary Figure S1. To that end, PDB[26] 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[27], PDB and UniProt-KB[11]. 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[41] 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 Povolotskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465**, 922-926 (2010).

2 Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *The EMBO journal* **5**, 823-826 (1986).

3 Lee, D., Redfern, O. & Orengo, C. Predicting protein function from sequence and structure. *Nature reviews. Molecular cell biology* **8**, 995-1005 (2007).

4 Tian, W. & Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of molecular biology* **333**, 863-882 (2003).

5 Worth, C. L., Gong, S. & Blundell, T. L. Structural and functional constraints in the evolution of protein families. *Nature reviews. Molecular cell biology* **10**, 709-720 (2009).

6 Bork, P., Sander, C. & Valencia, A. 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 (1993).

7 Galperin, M. Y., Walker, D. R. & Koonin, E. V. Analogous enzymes: independent inventions in enzyme evolution. *Genome Res* **8**, 779-790 (1998).

8 Omelchenko, M. V., Galperin, M. Y., Wolf, Y. I. & Koonin, E. V. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biol Direct* **5**, 31 (2010).

9 Rost, B. Protein structures sustain evolutionary drift. *Folding & design* **2**, S19-24 (1997).

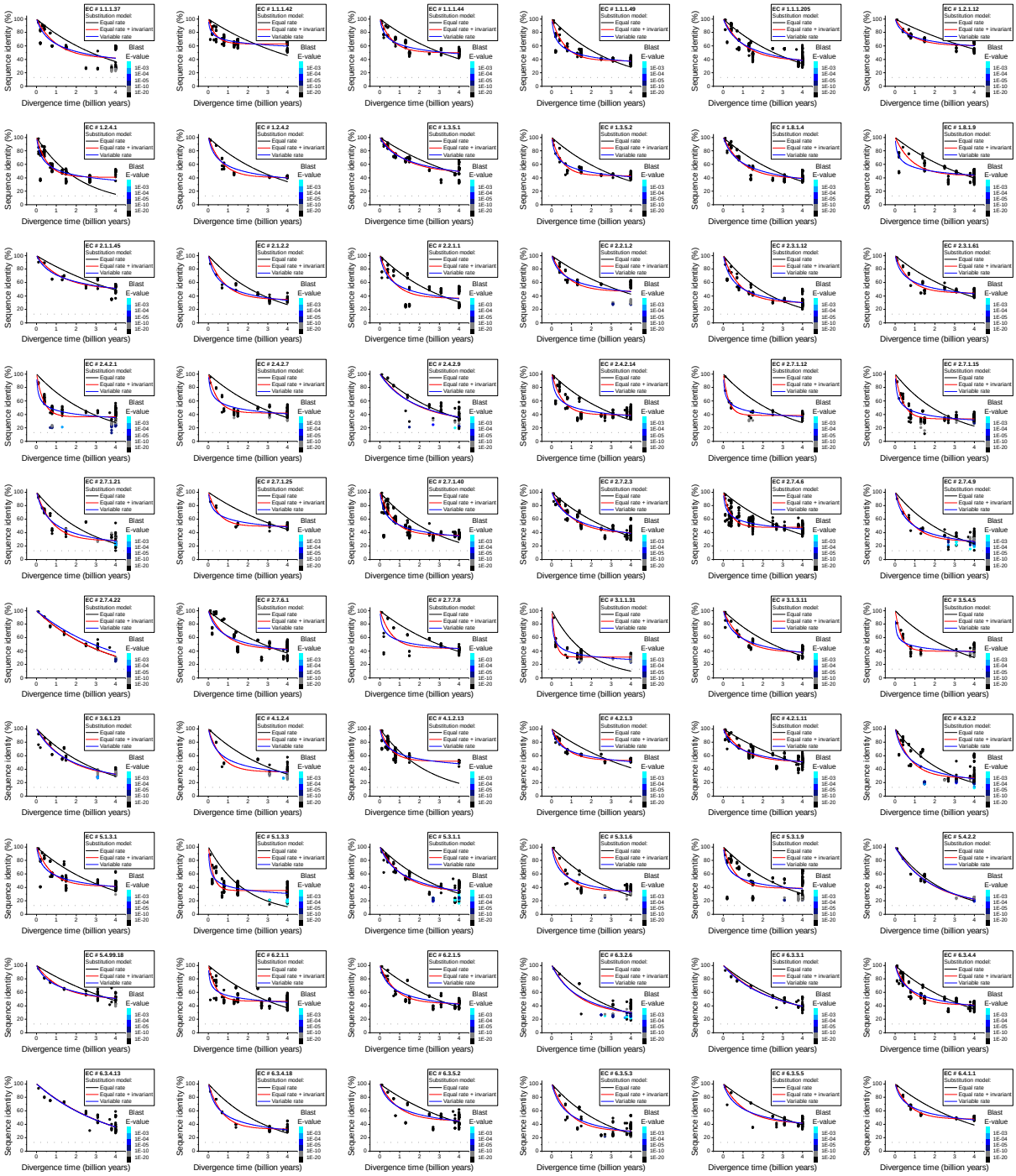
10 Lehninger, A. L., Nelson, D. L. & Cox, M. M. *Lehninger principles of biochemistry*. 6th edn, (W.H. Freeman, 2013).
11 UniProt, C. UniProt: a hub for protein information. *Nucleic acids research* **43**, D204-212 (2015).
12 Bairoch, A. The ENZYME data bank in 1999. *Nucleic acids research* **27**, 310-311 (1999).
13 Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431-449 (2000).
14 Fares, M. A., Byrne, K. P. & Wolfe, K. H. Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of Saccharomyces species. *Molecular biology and evolution* **23**, 245-253 (2006).
15 Fletcher, W. & Yang, Z. INDELible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution* **26**, 1879-1888 (2009).
16 O'Malley-James, J. T., Cockell, C. S., Greaves, J. S. & Raven, J. A. Swansong biospheres II: the final signs of life on terrestrial planets near the end of their habitable lifetimes. *Int J Astrobiol* **13**, 229-243 (2014).
17 Chang, A. *et al.* BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic acids research* **43**, D439-446 (2015).
18 Benkovic, S. J., Fierke, C. A. & Naylor, A. M. Insights into Enzyme Function from Studies on Mutants of Dihydrofolate-Reductase. *Science* **239**, 1105-1110 (1988).
19 Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894-898 (2009).
20 Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic acids research* **43**, D261-269 (2015).
21 Kelsic, E. D. *et al.* RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq. *Cell Syst* **3**, 563-571 e566 (2016).
22 Arnold, C. & Stadler, P. F. Polynomial algorithms for the Maximal Pairing Problem: efficient phylogenetic targeting on arbitrary trees. *Algorithms Mol Biol* **5**, 25 (2010).
23 Wilson, C. A., Kreychman, J. & Gerstein, M. 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 (2000).
24 Watson, J. D., Laskowski, R. A. & Thornton, J. M. Predicting protein function from sequence and structural data. *Current opinion in structural biology* **15**, 275-284 (2005).
25 Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**, 2302-2309 (2005).
26 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic acids research* **28**, 235-242 (2000).
27 Porter, C. T., Bartlett, G. J. & Thornton, J. M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic acids research* **32**, D129-133 (2004).
28 Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774-786 (2009).
29 Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535-538 (2012).
30 Lunzer, M., Golding, G. B. & Dean, A. M. Pervasive cryptic epistasis in molecular evolution. *PLoS Genet* **6**, e1001162 (2010).
31 Pingoud, A., Wilson, G. G. & Wende, W. Type II restriction endonucleases--a historical perspective and more. *Nucleic acids research* **42**, 7489-7527 (2014).
32 Huang, P. S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320-327 (2016).

33 Shendure, J. & Akey, J. M. The origins, determinants, and consequences of human mutations. *Science* **349**, 1478-1483 (2015).
34 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic acids research* **44**, D457-462 (2016).
35 Kristensen, D. M. *et al.* A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* **26**, 1481-1487 (2010).
36 Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631-637 (1997).
37 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389-3402 (1997).
38 Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* **28**, 1033-1034 (2012).
39 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
40 Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971-2972 (2006).
41 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).
42 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
43 DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**, 5069-5072 (2006).
44 Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author: Department of Genome Sciences, University of Washington, Seattle* (2005).
45 Ye, Y. & Godzik, A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* **19 Suppl 2**, ii246-255 (2003).

Table S1. Model species selected and pairwise average divergence times [a] (in billions of years)[40]

	Eukarya										Bacteria									Archaea		
	Homo sapiens	*Rattus norvegicus*	*Xenopus laevis*	*Xenopus laevis*	*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[40]

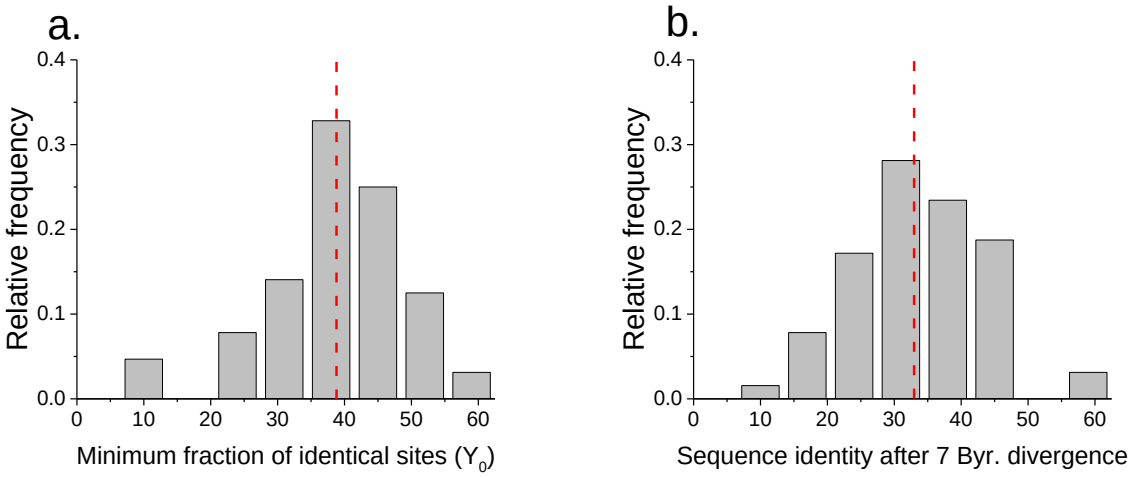


Supplementary Figure S1. The global pairwise sequence identities between pairs of orthologous enzymes are shown as a function of divergence times between the corresponding species. The colored lines indicate fits to the data using the three models of amino acid substitution (Equations 1-3). The dashed 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

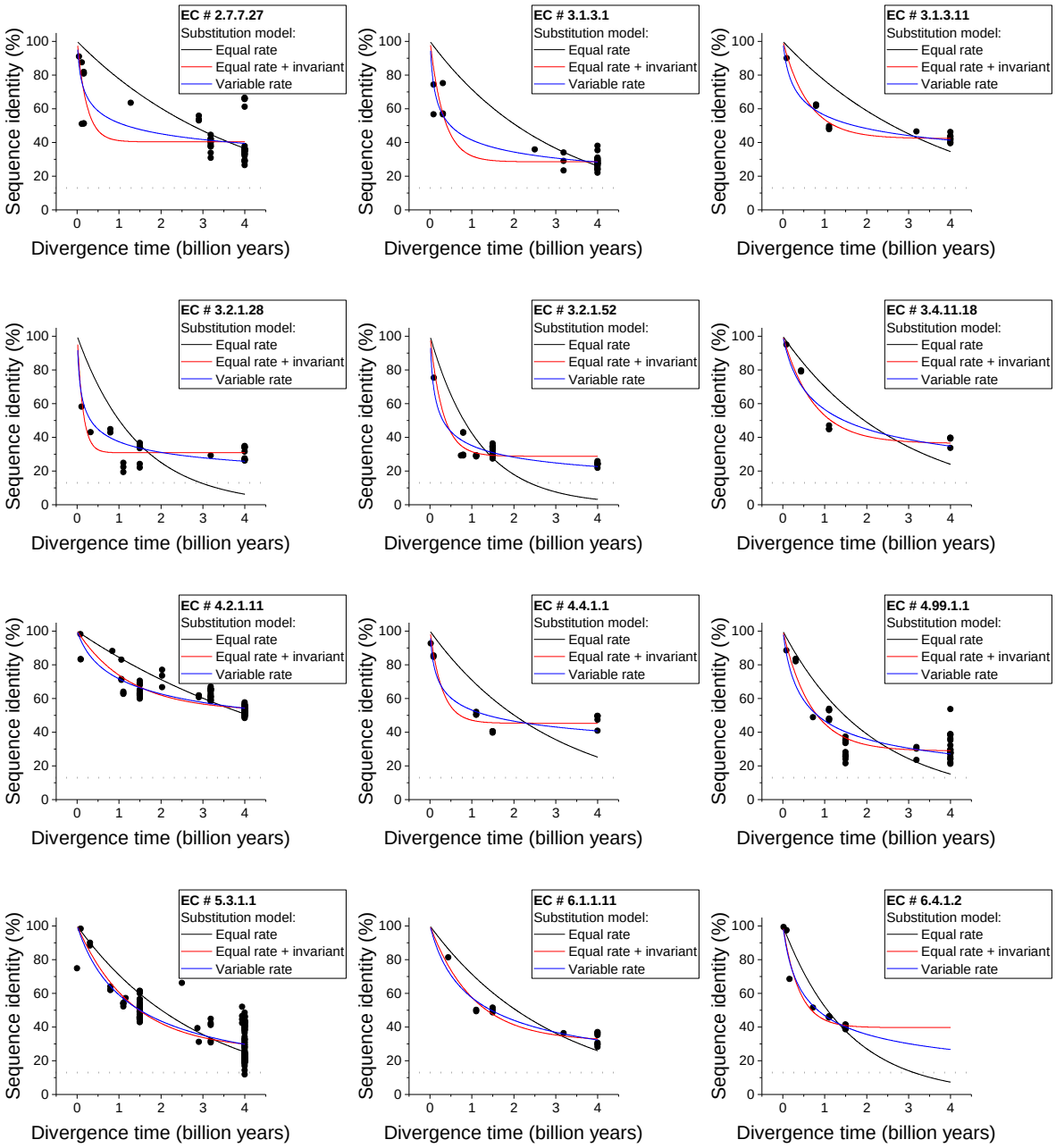
[a] P-values <0.05 indicate Y_0 is significantly higher than 13.5%, the average identity between unrelated sequences



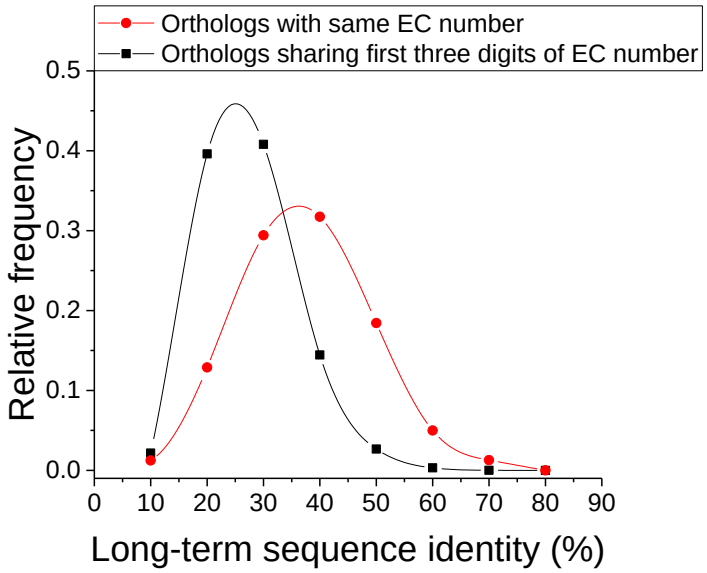
Supplementary Figure S2. a. The distribution of Y_0 parameter values across 64 EC numbers for Model 2 fits (Equation 2). The Y_0 parameter represents the minimum fraction of protein sites that remain identical at long divergence times. **b.** The distribution across 64 EC numbers of the projected global 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



Supplementary 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).



Supplementary Figure S4. The distribution of long-term global sequence identities between orthologous pairs of enzymes annotated in UniProt with the same EC number (red), and only sharing the first three digits of an EC number (black). The long-term sequence identities were obtained using pairs of orthologs from species that diverged over two billion years ago. Results are based on COGs for all possible enzymatic activities and the same 22 species used to analyze global sequence divergence (Table S1). To avoid functional biases, the distribution for orthologs with the same EC number (red) were restricted to the same set of COGs as the distribution for orthologs sharing the first three digits of an EC number (black). The two distributions are significantly different from each other (Mann-Whitney P-value: $<10^{-20}$). Points in different sequence identity bins were connected by a spline function (solid lines).