LETTER

doi:10.1038/nature11510

Epistasis as the primary factor in molecular evolution

Michael S. Breen[1], Carsten Kemena[1], Peter K. Vlasov[1], Cedric Notredame[1] & Fyodor A. Kondrashov[1,2]

The main forces directing long-term molecular evolution remain obscure. A sizable fraction of amino-acid substitutions seem to be fixed by positive selection[1–4], but it is unclear to what degree long-term protein evolution is constrained by epistasis, that is, instances when substitutions that are accepted in one genotype are deleterious in another. Here we obtain a quantitative estimate of the prevalence of epistasis in long-term protein evolution by relating data on amino-acid usage in 14 organelle proteins and 2 nuclear-encoded proteins to their rates of short-term evolution. We studied multiple alignments of at least 1,000 orthologues for each of these 16 proteins from species from a diverse phylogenetic background and found that an average site contained approximately eight different amino acids. Thus, without epistasis an average site should accept two-fifths of all possible amino acids, and the average rate of amino-acid substitutions should therefore be about three-fifths lower than the rate of neutral evolution. However, we found that the measured rate of amino-acid substitution in recent evolution is 20 times lower than the rate of neutral evolution and an order of magnitude lower than that expected in the absence of epistasis. These data indicate that epistasis is pervasive throughout protein evolution: about 90 per cent of all amino-acid substitutions have a neutral or beneficial impact only in the genetic backgrounds in which they occur, and must therefore be deleterious in a different background of other species. Our findings show that most amino-acid substitutions have different fitness effects in different species and that epistasis provides the primary conceptual framework to describe the tempo and mode of long-term protein evolution.

The study of the factors determining the tempo and mode of molecular evolution continues to be at the forefront of evolutionary biology. Since the inception of the neutral theory[5], many studies of the rate of molecular evolution have focused on the relative role of selection versus genetic drift in the fixation of amino-acid substitutions[6]. It now seems certain that both of these factors, selection and genetic drift, contribute to a substantial fraction of all amino-acid substitutions in the course of evolution[1–4]. However, the neutral-versus-selective debate on the nature of molecular evolution has primarily focused on the short-term effects of substitutions that may not provide the framework necessary to understand the differences between the functional and selective effects of amino-acid substitutions that accumulate in the course of long-term evolution.

An amino-acid substitution that is neutral or beneficial in one genetic context may be deleterious in another[7–11]. Such a situation, when the fitness effect of one allele state depends on the allele states at other loci, is called epistasis[9,10]. Both the neutral and selective theories of protein evolution provide an accurate framework for understanding long-term protein evolution only if amino-acid states in different genetic contexts have the same effect on fitness, that is, if epistasis is rare. In the absence of epistasis, when the fitness effects of all amino-acid states are independent of one another, substitutions in different species are expected to have similar effects on fitness except in cases where these substitutions enable differences in adaptation to environmental conditions. In that case, if an amino-acid state were found in one species in a protein sequence that is not directly involved in environmental adaptation, such as a housekeeping protein, then the same amino-acid state should be acceptable in an orthologous site in a different species. However, if epistasis is common then amino-acid substitutions that were beneficial or neutral in one species should often be deleterious in another. Therefore, unravelling the extent and basis of epistasis may be crucial to understanding differences in protein sequences between species and long-term protein evolution[5–15]. At present, studies of the differences in the fitness of substitutions in different genetic contexts consider specific genes or events[11,16–24], and it is unknown what fraction of amino-acid substitutions that occur in one species would also be acceptable in another species if they were to occur in orthologous sites (but see ref. 11). Here we develop an approach to quantifying the impact of epistasis in protein evolution and show that the fitness effects of most amino-acid substitutions must depend on the genetic context in which they occur.

We obtained sequence data for some of the most widely sequenced genes for organelle and nuclear-encoded proteins. The choice of genes selected for this study was dictated by four considerations: the availability of at least 1,000 unique orthologous sequences from different species; a well-defined, conserved housekeeping function; the absence or a low rate of gene duplication; and conserved sequence with few insertions and deletions (indels), leading to an unequivocal multiple alignment of the orthologous sequences. Our final data set included orthologous sequences from Metazoa, including 13 mitochondrial genes and 2 nuclear genes, and the large subunit of ribulose 1,5-bisphosphate carboxylase–oxygenase (Rubisco) encoded by the *rbcL* gene in chloroplast genomes of Viridiplantae. The number of unique sequences from different species ranged between 949 and 13,912, with proteins encoded in the organelles having more sequences in the alignment (Table 1). The sequences were aligned using a version of the T-Coffee algorithm[25] adapted to align large data sets of up to 100,000 sequences (Methods). For each of these alignments, we calculated the amino-acid usage (u), defined as the number of different amino acids observed per site (Table 1). We found that the average usage across all genes in our data set was ~9, meaning that in the course of long-term evolution an average site accepted approximately half of all possible amino acids. The distribution of u reveals few invariant sites (those where a single amino acid was observed), indicating that the high usage was not caused by an average of invariant and extremely variable sites (Supplementary Fig. 1).

Amino-acid usage can be used to predict the expected short-term rates of protein evolution without epistasis. Without epistasis, an amino-acid state observed in one species should also be acceptable in the orthologous site of another species. Therefore, substitutions in a protein sequence leading to amino-acid states observed in orthologous sites should not be inhibited by negative selection. For example, if one-half of amino acids were accepted in a protein sequence in the long term then, in a non-epistatic model, the same fraction of amino acids should be acceptable to this protein in the course of short-term evolution. This prediction can be expressed in terms of the ratio of the per-site rates of non-synonymous and synonymous evolution[6] (dN/dS): if on average ~0.5 of all amino acids were found per site then the expected non-epistatic dN/dS ratio between two closely related orthologues of this gene should be ~0.5 (Methods).

[1]Bioinformatics and Genomics Programme, Centre for Genomic Regulation and Universitat Pompeu Fabra, 88 Dr Aiguader, Barcelona 08003, Spain. [2]Institució Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluís Companys, Barcelona 08010, Spain.

Table 1 | Amino-acid usage, frequency of non-fixed states and rates of evolution

Gene	Number of species with non-redundant sequences	Average amino-acid usage	Frequency of non-fixed state (p)	Average amino-acid usage after Poisson correction	Expected, non-epistatic dN/dS ratio from Poisson correction	Observed average dN/dS ratio	Fraction of epistatic evolution
ATP6	3,021	9.83	5.41×10^{-4}	9.39	0.44	0.056	0.87
ATP8	1,244	12.82	8.24×10^{-4}	11.61	0.56	0.224	0.60
COX1	4,450	6.87	7.13×10^{-4}	6.33	0.28	0.015	0.95
COX2	4,204	9.70	7.64×10^{-4}	9.12	0.43	0.025	0.94
COX3	2,191	9.43	2.00×10^{-4}	7.15	0.32	0.036	0.89
CYTB	7,954	10.70	1.17×10^{-3}	10.68	0.51	0.039	0.92
ND1	2,056	9.76	4.99×10^{-4}	8.41	0.39	0.040	0.90
ND2	5,963	10.67	1.64×10^{-3}	10.66	0.51	0.067	0.87
ND3	2,852	10.54	6.8×10^{-4}	10.24	0.49	0.069	0.86
ND4	2,041	10.08	5.31×10^{-4}	9.00	0.42	0.045	0.89
ND4L	1,785	11.85	4.97×10^{-4}	10.24	0.49	0.076	0.84
ND5	949	8.97	4.07×10^{-4}	7.10	0.32	0.057	0.82
ND6	1,015	11.28	2.88×10^{-4}	8.97	0.42	0.073	0.83
EEF1A1	1,743	3.81	5.31×10^{-4}	3.05	0.11	0.020	0.82
H3.2	1,228	5.18	8.97×10^{-4}	3.59	0.14	0.037	0.74
rbcL	13,912	8.65	1.39×10^{-3}	8.65	0.40	0.072	0.82

Under the simplifying assumption that at a site in a large multiple alignment all permissible amino-acid states are equally likely to be observed at least once, dN/dS should be equal to $(u-1)/19$, where 19 is the number of possible substitutions possible at a site. However, to apply amino-acid usage as a measure for estimating the expected non-epistatic dN/dS value, it is necessary to take into account potential biases that are inherent to usage. Specifically, u is sensitive to erroneous inclusions of rare amino-acid states because an amino acid that is found only once at a site in an alignment makes the same contribution to the measure as an amino acid that is found thousands of times. Indeed, the majority of the observed amino-acid states in our alignments are rare (Supplementary Fig. 2), which at first glance seems to suggest that sequencing errors or deleterious polymorphisms had a large influence on u. However, the frequency distribution of fixed amino-acid states is also expected to follow an exponential decay function. On a realistic phylogeny with many species, the majority of evolutionary time is found in recent branches, and many amino-acid states are therefore expected to be rare because they would have evolved recently (Supplementary Fig. 3). Thus, a majority of fixed events are also expected to be rare as a direct consequence of the phylogenetic nature of evolution and, therefore, a sophisticated method to take into account the impact of rare erroneous states is required.

There are three potential sources of error in our estimate of amino-acid usage: alignment errors, sequencing errors and sequence polymorphisms. The latter two we treat as a single confounding factor of non-fixed amino-acid states. We used different approaches to deal with potential alignment errors and the contribution of non-fixed states. To minimize the effect of misaligned sequences, we post-processed the alignments by removing the sequences most likely to be misaligned. In a progressive alignment procedure, such as the one used here, mistakes usually arise through inaccurate indel accumulation. We therefore recursively removed the sequences inducing the largest fraction of indels and realigned the remaining sequences (Supplementary Table 1). We found that the resulting alignments contained few indels, and different alignment methods yielded virtually identical amino-acid usage measures (Supplementary Table 2). Thus, we do not expect alignment errors to contribute to the usage measure of our data set.

Next we used a probabilistic approach to estimate the contribution of non-fixed amino-acid states to the amino-acid usage measure. First we estimated the probability of occurrence of a non-fixed state at one site in one sequence, p, by calculating the divergence of sequences from the same species (Supplementary Methods). This approach combines the estimate of the frequency of occurrence of sequencing errors and segregating polymorphisms into one measure. Using multiple sequences from the same species from a large subset of those species that were used in the original multiple alignment (Supplementary Table 3), we estimated the amino-acid diversity of non-fixed states, π_a, in the same manner as the commonly used nucleotide diversity measure. From π_a we then estimated p as $\pi_a/6$, where 6 is the average number of non-synonymous mutations per codon (Methods). We found that p was generally small, with $0.0002 < p < 0.007$ across the genes in our data set (Table 1). We then calculated the probability of multiple occurrences of the same non-fixed state at one site in the multiple alignment using the Poisson approximation of a binomial distribution. The probability that a non-fixed state with probability of occurrence p is observed k times in an alignment of N sequences is $m = (pN)^k e^{-pN}/k!$. The probability that an amino-acid state is observed at least once out of k times in a fixed rather than non-fixed state is $r = 1 - m$. Thus, at each site the amino-acid usage measure that takes into account the probability that some of the states were non-fixed is $\sum_{i=1}^{u} r_i$, where i is the amino-acid state and u is the amino-acid usage at that site. By applying this approach to our data, we obtained the amino-acid usage corrected for the contribution of non-fixed states. This was slightly smaller than the uncorrected amino-acid usage (Table 1), indicating that the impact of non-fixed states is small. Two different approaches, one probabilistic and one empirical, confirm the low impact of non-fixed states on our measurement of amino-acid usage (Supplementary Tables 4 and 5).

Our data show that a site in a protein can, on average, accept about eight different amino acids in different species over the course of long-term evolution. This estimate is undoubtedly a lower bound owing to the restricted phylogenetic breadth and quantity of sequences included in the alignment. In our data, u tends to plateau as the number of sequences included in the alignment is increased; however, the actual plateau of this function is apparently beyond the number of sequences that were available for our study (Supplementary Fig. 4), confirming our estimate of u as a lower bound. The non-epistatic model predicts that an amino acid that is acceptable to one species should be acceptable to another, such that the expected non-epistatic dN/dS value of orthologues from the same alignment should equal $(u-1)/19$; in our case, for amino-acid usage of eight, the expected dN/dS value is 7/19, or $\sim$0.37. We compared this prediction with the observed dN/dS ratio estimated for each of the 16 genes in our data set using many of the species present in the multiple alignment (Supplementary Table 6). On average, the observed dN/dS ratio was about seven times smaller than that predicted by the non-epistatic model (Table 1).

The $(u-1)/19$ approximation is a crude estimate of the expected dN/dS value because it relies on the assumption that in the multiple alignment all permissible amino acids are equally likely to be observed. However, some amino-acid states may be more than one mutational step away from a sequence in which dN/dS is measured, and such states should therefore not be taken into account when comparing the expected non-epistatic dN/dS ratio with the observed ratio for this particular sequence. We thus estimated the expected dN/dS value using only those amino-acid states that are one mutational step away

from the sequence for which dN/dS was measured. We estimated the non-epistatic dN/dS ratio for these sequences as the average of u_m/n_m across all sites, where n_m is the total number of possible amino-acid states one mutational step away from the state in a given site and u_m is the number of such states observed in the entire alignment. The u_m/n_m measure of epistatic evolution was larger than that predicted by our crude measure $(u-1)/19$, resulting in a difference between the expected non-epistatic dN/dS value and the observed value of more than one order of magnitude (Table 2). The higher expected dN/dS value in this case can be explained to be the result of closely related sequences exploring only a limited part of the sequence space[15] or by the physicochemical similarity of amino acids that are located in the mutational vicinity in the genetic code.

The large difference between the dN/dS value predicted on the basis of amino-acid usage and that observed across a large fraction of the same sequences (Supplementary Table 6) indicates that the vast majority of amino-acid states are acceptable in a species- or clade-specific manner. There are two ways to explain this pattern. First, most amino-acid substitutions may have been subject to positive selection for an environmental adaptation specific to the environment of a particular species. Thus, an amino acid that was beneficial to one species because of a specific environmental adaptation may be detrimental to a species that does not live in the same environment. The second explanation is that the fitness of most amino-acid substitutions could have depended on the genetic context, or internal cellular environment, of the species in which the substitution occurs, such that a substitution beneficial in one background would be expected to be detrimental in a different background. In other words, epistatic interactions are the norm and not the exception when we consider amino-acid substitutions in protein sequences. Two considerations indicate that widespread epistasis is the most likely explanation for the observed data. First, the McDonald–Kreitman test[1] showed that positive selection was not common in the evolution of the proteins in our data set (Table 3), as would have been expected if the majority of substitutions conferred an environmental adaptation. Second, the nature of the functions of the proteins considered here, especially that of mitochondrial proteins, implies that interaction with the external environment was limited. Additionally, numerous examples of genetic interactions within proteins, through structural interactions within the same molecule[8–11,16–20,26] or interaction of cellular components[21–24,27–30], support the hypothesis that epistatic interactions may affect a large majority of all amino-acid substitutions.

We identify epistasis as a powerful factor affecting long-term protein evolution and one that must necessarily be invoked to explain why the vast majority of amino-acid substitutions that occur in one species cannot occur in another regardless of whether or not positive selection plays the dominant role in the course of fixation of amino-acid substitutions in specific genetic contexts. An epistatic perspective of molecular evolution leads to the formulation of several fundamental questions, in addition to the largely unanswered questions posed by John Maynard Smith in 1970 (ref. 12). First, given a specific site, substitutions in how many other sites in the same gene or in the entire genome could change the strength of selection associated with substitutions at this site? Second, out of the entire network of pairwise epistatic interactions between sites across the genome, are there many non-overlapping epistatic subnetworks or are most sites interconnected within the entire network of epistatic interactions? Third, what is the ratio of intergenic to intragenic epistatic interactions? Fourth, what is the molecular basis of epistatic interactions within the genome? Finally, pervasive epistasis in long-term protein evolution raises the possibility that similar epistatic interactions may be prevalent in short-term evolution[20,21] and that situations when a polymorphism is benign or beneficial to one individual but deleterious to another individual within the same population may be more common than is thought at present.

Table 2 | Fraction of epistatic evolution measured by u_m/n_m

Gene	Number of species with dN/dS estimate	Average expected non-epistatic dN/dS ratio estimated by u_m/n_m	Observed average dN/dS ratio	Average fraction of epistatic evolution (s.d.)
ATP6	1,300	0.77	0.056	0.93 (0.063)
ATP8	781	0.52	0.224	0.57 (0.303)
COX1	1,123	0.81	0.015	0.98 (0.027)
COX2	1,214	0.81	0.025	0.97 (0.029)
COX3	622	0.81	0.036	0.96 (0.038)
CYTB	3,992	0.84	0.039	0.95 (0.034)
ND1	569	0.81	0.040	0.95 (0.039)
ND2	3,210	0.79	0.067	0.92 (0.042)
ND3	989	0.69	0.069	0.90 (0.069)
ND4	510	0.83	0.045	0.95 (0.033)
ND4L	441	0.65	0.076	0.88 (0.121)
ND5	370	0.87	0.057	0.93 (0.033)
ND6	406	0.70	0.073	0.90 (0.098)
EEF1A1	1,343	0.77	0.020	0.97 (0.018)
H3.2	670	0.72	0.037	0.95 (0.090)
rbcL	13,546	0.88	0.072	0.92 (0.093)

Table 3 | McDonald–Kreitman test on select species for each gene

Gene	Average α using polymorphisms with >1% frequency in the population	Average α using polymorphisms with >5% frequency in the population	Number of species with polymorphism and divergence data
ATP6	−0.7 (0.06)	−0.4 (0.08)	52
ATP8	−0.1 (0.16)	−0.1 (0.15)	18
COX1	−0.8 (0.08)	−0.8 (0.08)	25
COX2	−0.7 (0.06)	−0.4 (0.09)	54
COX3	−0.5 (0.18)	−0.6 (0.17)	13
CYTB	−0.8 (0.02)	−0.4 (0.04)	255
ND1	−0.6 (0.15)	−0.5 (0.18)	16
ND2	−0.6 (0.06)	−0.4 (0.06)	77
ND3	−0.4 (0.09)	−0.3 (0.09)	38
ND4	−0.6 (0.11)	−0.6 (0.09)	22
ND4L	−0.6 (0.11)	−0.4 (0.12)	20
ND5	−0.6 (0.17)	−0.7 (0.16)	9
ND6	−0.4 (0.12)	−0.5 (0.11)	25
EEF1A1	NA	NA	NA
H3.2	−1.0 (0.04)	−1.0 (0.03)	11
rbcL	−0.3 (0.70)	−1.0 (0.00)	2

The propensity for positive selection, α, was estimated as $1-(D_s/D_n)/(P_n/P_s)$, where D_s and P_s correspond to the number of synonymous fixed differences and polymorphisms, respectively, and D_n and P_n are the corresponding values for non-synonymous substitutions and polymorphisms. A negative α value indicates that there were fewer non-synonymous substitutions in evolution than expected given the number of non-synonymous polymorphisms. Deleterious polymorphisms with a frequency >5% are expected to be rare, and the presented data therefore suggest a minor role for positive selection in the evolution of these proteins. The reported values represent averages for all species with polymorphism data, with s.e.m. shown in parentheses. Such data were not available (NA) for one gene in our data set.

METHODS SUMMARY

Orthologous sequences for 15 genes from Metazoa and 1 from Viridiplantae were aligned using an adapted version of the T-Coffee multiple-sequence aligner. Each sequence was encoded as a 20^3-vector where each component was the frequency of amino-acid words of size 3. The vectors were clustered in 200 groups using the *k*-means algorithm. Each resulting group containing fewer than 200 sequences was aligned with the default T-Coffee algorithm, and larger clusters were further divided by reapplying the *k*-means algorithm. The resulting alignments were treated as profiles and realigned using the multiple-profile alignment procedure supported by the default T-Coffee algorithm. A small subset of sequences, ~14% on average across genes, contributed a substantial fraction of gaps to the alignment and were removed. Amino-acid usage was measured as the number of different amino-acid states across all sites in the final multiple alignment of orthologues. We used pairwise sequence comparisons to estimate dN/dS values across genes with the codeml program in the PAML package for all pairwise comparisons in the data set with $0.05 < dS < 0.5$. We then averaged the dN/dS estimates across all species with $dS < 0.5$ and then averaged the dN/dS values across such clusters.

Full Methods and any associated references are available in the online version of the paper.

Received 6 March; accepted 14 August 2012.
Published online 14 October 2012.

1. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351,** 652–654 (1991).

2. Andolfatto, P. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437,** 1149–1152 (2005).
3. Charlesworth, J. & Eyre-Walker, A. The rate of adaptive evolution in enteric bacteria. *Mol. Biol. Evol.* **23,** 1348–1356 (2006).
4. Boyko, A. R. *et al.* Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* **4,** e1000083 (2008).
5. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217,** 624–626 (1968).
6. Li, W. H. *Molecular Evolution* 419–429 (Sinauer, 1997).
7. Kimura, M. The role of compensatory neutral mutations in molecular evolution. *J. Genet.* **64,** 7–19 (1985).
8. Weinreich, D. M., Watson, R. A. & Chao, L. Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* **59,** 1165–1174 (2005).
9. Lehner, B. Molecular mechanisms of epistasis within and between genes. *Trends Genet.* **27,** 323–331 (2011).
10. de Visser, J. A., Cooper, T. F. & Elena, S. F. The causes of epistasis. *Proc. R. Soc. Lond. B* **278,** 3617–3624 (2011).
11. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA* **99,** 14878–14883 (2002).
12. Maynard Smith, J. Natural selection and the concept of a protein space. *Nature* **225,** 563–564 (1970).
13. Fitch, W. M. & Markowitz, E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4,** 579–593 (1970).
14. Gillespie, J. H. Natural selection and the molecular clock. *Mol. Biol. Evol.* **3,** 138–155 (1986).
15. Povolotskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465,** 922–926 (2010).
16. Poon, A. F. & Chao, L. Functional origins of fitness effect-sizes of compensatory mutations in the DNA bacteriophage phiX174. *Evolution* **60,** 2032–2043 (2006).
17. Bridgham, J. T., Ortlund, E. A. & Thornton, J. W. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461,** 515–519 (2009).
18. Meer, M. V., Kondrashov, A. S., Artzy-Randrup, Y. & Kondrashov, F. A. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature* **464,** 279–282 (2010).
19. Costanzo, M. S. & Hartl, D. L. The evolutionary landscape of antifolate resistance in Plasmodium falciparum. *J. Genet.* **90,** 187–190 (2011).
20. Salverda, M. L. *et al.* Initial mutations direct alternative pathways of protein evolution. *PLoS Genet.* **7,** e1001321 (2011).
21. Woods, R. J. *et al.* Second-order selection for evolvability in a large *Escherichia coli* population. *Science* **331,** 1433–1436 (2011).
22. Osada, N. & Akashi, H. Mitochondrial-nuclear interactions and accelerated compensatory evolution: evidence from the primate cytochrome C oxidase complex. *Mol. Biol. Evol.* **29,** 337–346 (2012).
23. Kvitek, D. J. & Sherlock, G. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet.* **7,** e1002056 (2011).
24. Silva, R. F. *et al.* Pervasive sign epistasis between conjugative plasmids and drug-resistance chromosomal mutations. *PLoS Genet.* **7,** e1002181 (2011).
25. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302,** 205–217 (2000).
26. Lunzer, M., Golding, G. B. & Dean, A. M. Pervasive cryptic epistasis in molecular evolution. *PLoS Genet.* **6,** e1001162 (2010).
27. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327,** 425–431 (2010).
28. Tokuriki, N. & Tawfik, D. S. Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature* **459,** 668–673 (2009).
29. Poelwijk, F. J., de Vos, M. G. & Tans, S. J. Tradeoffs and optimality in the evolution of gene regulation. *Cell* **146,** 462–470 (2011).
30. Burga, A. Olivia Casanueva, M. & Lehner, B. Predicting mutation outcome from early stochastic variation in genetic interaction partners. *Nature* **480,** 250–253 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements The work was supported by Plan Nacional grants from the Spanish Ministry of Science and Innovation, to F.A.K. and C.N. C.K. was supported by the European Union FP7 project Quantomics (KBBE2A222664). F.A.K. is a European Molecular Biology Organization Young Investigator and Howard Hughes Medical Institute International Early Career Scientist. We thank B. Lehner and T. Warnecke for input and a critical reading of the manuscript.

Author Contributions M.S.B., C.K., P.K.V. and F.A.K. participated in obtaining and quality-testing the data; C.K. and C.N. participated in the design of the alignment algorithm; and F.A.K. designed the study and wrote the manuscript with the participation of all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.A.K. (fyodor.kondrashov@crg.es).

METHODS

Sequences for our study. We obtained amino-acid and nucleotide sequences of the protein-coding genes from GenBank[31] and applied a two-directional best BLAST hit approach[32] for the initial identification of orthologous sequences. The phylogenetic breadths of orthologues that were included in our alignments were purposefully restricted to improve the reliability of the multiple alignments. We took sequences only from Viridiplantae for Rubisco and only from Metazoa for all other proteins. We focused on mitochondrial proteins for four reasons. First, mitochondrially encoded proteins are very rarely, if at all, duplicated. Second, these proteins are expected to have a highly conserved function across the Metazoa taxon. Third, mitochondrial genes have been widely sequenced in many phylogenetic studies, so we expected to find sequences from thousands of different species. Finally, owing to the constraints on the mitochondrial genome and on mitochondrial gene function, we expected to find few insertions and deletions in these genes across the Metazoan phylogeny. The same line of reasoning applied to the inclusion of Rubisco in our data set. Finally, we have undertaken a survey of nuclear-encoded genes that were expected to show similar qualities of functional conservation and lack of duplications, with at least 1,000 sequences from different species. We have specifically mostly avoided sequences from unicellular organisms, because environmental adaptation may be relevant to a higher fraction of genes in such organisms. In multicellular organisms the function of many genes takes place in more isolated intracellular environments. Only two nuclear-encoded proteins passed our selection criteria, elongation factor 1-alpha 1 (EEF1A1) and histone (H3.2).

Alignment method and verification. Sequences were aligned using KM-Coffee, an adapted version of the T-Coffee multiple-sequence aligner[25]. In KM-Coffee, each sequence is encoded as a 20^3-vector in which each component is the frequency of amino-acid words of size 3. The vectors are then clustered in 200 groups using the *k*-means algorithm. Each resulting group containing fewer than 200 sequences is aligned with the default T-Coffee algorithm or further divided into smaller clusters until each cluster can be aligned by T-Coffee. The resulting multiple sequence alignments are then treated as profiles and aligned using the multiple-profile alignment procedure supported by the default version of T-Coffee. The main advantage of the procedure is its capacity to resolve very large data sets while incorporating a consistency-based approach.

To remove the sequences most likely to be misaligned, we designed a post-processing filtering procedure. For each sequence we estimated the indel contribution index, defined as $\sum_c (N-NR)/NR$, where N is the number of sequences, NR is the number of aligned residues in every column and c is the number of columns in the alignment. In effect, this measure shares the cost of any gap across all aligned sequences. We then sorted the sequences by their indel contribution index and removed the top fraction of sequences, seeking a trade-off between retaining the maximum number of sequences and maximizing the agreement between the three large-scale alignment methods. Overall, by applying the indel contribution index we removed 13.5% of sequences across all data sets (Supplementary Table 1). Remaining sequences were then realigned and the resulting alignments used for all subsequent analysis.

Three methods were available that could align large data sets like the one used for this study: MAFFT[33], Clustal Omega[34] and KM-Coffee. We applied all three to the filtered data sets and estimated the similarity of their output alignments, using the sum-of-pairs method to compare alternative alignments of the same sequences. All three alignment methods show high agreement. The average pairwise similarity between the KM-Coffee alignments and the Clustal Omega alignments is 97.7%, and to the MAFFT alignments it is 97.8%. The correlation between the corrected average amino-acid distributions (Supplementary Table 2) was also high (Pearson coefficient >0.999 for all three pairs).These results indicate significant robustness when comparing KM-Coffee, Clustal Omega and MAFFT. We also measured the similarity of different methods when used on the unfiltered set of alignments and saw an increase in similarity when using KM-Coffee after filtering relative to Clustal Omega (97.7% versus 94.7%) and MAFFT (97.8% versus 93.2%).

Measuring amino-acid usage. To measure the amino-acid usage, we calculated the number of different amino-acid states across all sites in the final multiple alignment of orthologues. We used only sites where more than half of the positions were occupied by an amino acid (not a deletion) and took the average across all sites in the gene. The number of sites where the amino-acid usage was measured is reported in Supplementary Table 2.

Using amino-acid usage to estimate the expected dN/dS ratio. The amino-acid usage measure, u, can be used to obtain a lower-bound estimate of the expected dN/dS ratio under the assumption of no epistasis. The rates of non-synonymous (dN) and synonymous (dS) evolution are measures of the number of substitutions divided by the number of sites[6]. A crude estimate of non-epistatic dN/dS from u is $(u-1)/19$, where $u-1$ is the number of amino-acid states into which the current amino acid can be substituted and 19 is the total number of possible amino-acid substitutions at a site. This estimate is based on the assumption that when dN/dS is measured across many pairwise orthologous comparisons, the likelihood of a substitution between any of the 20 amino acids is equally likely. In two situations, the accuracy of this crude estimate is favourably influenced: when the multiple alignment from which u is calculated contains thousands of diverged sequences from a diverse phylogenetic background, and when dN/dS is measured as an average across pairwise comparisons from a large fraction of sequences in the entire multiple alignment. An estimate that takes into account the mutational neighbourhood of sequences for which dN/dS is measured should be more accurate in all instances. Owing to the limitations of taking into account more than one substitution at a site, dN/dS is typically measured between closely related sequences[6]. In this study, we included dN/dS measurements only for cases with $0.05 < dS < 0.5$. Because the average measurements of dN/dS in our measurements were $\sim$0.05 (Table 1), this implies that $0.0025 < dN < 0.025$. Thus, the probability of two non-synonymous substitutions occurring at the same site is dN^2, which implies that in our measurement of dN/dS we mostly take into account non-synonymous substitutions just one mutational step away. Thus, we also created a measure of the expected non-epistatic dN/dS taking into account only those amino-acid states that are one mutation away from the sequence for which dN/dS is measured. For each sequence for which dN/dS was measured, we calculated n_m, the total number of possible amino-acid states one mutational step away from the sequence, and u_m, the number of such states that were found in the multiple alignment. The estimate of the expected non-epistatic dN/dS ratio across the entire alignment then becomes an average of u_m/n_m across all sequences with observed dN/dS.

Estimating π_a and p. To estimate the expected rate of occurrence of a specific non-fixed amino-acid state, we first measured the average amino-acid diversity, π_a. This measure, analogous to the nucleotide diversity measure[35], is the fraction of amino-acid mismatches in a pairwise alignment of two sequences from the same species. To obtain an average π_a value for each gene in our data set, we first obtained an average π_a value for each species with two or more sequences and then took an average across all species. The number of different species that were used in the estimate of the average π_a value constituted a considerable fraction, $\sim$60% on average, of the total number of species (Supplementary Table 3). There are no reasons to assume that such extensive sampling should be biased towards species with a higher or lower π_a value. The amino-acid diversity is a measure of the expected density of any non-fixed states. However, for an accurate estimate of the contribution of non-fixed states, we required a measure for specific states. The π_a measure was obtained from standing variability, such that most of the segregating polymorphisms were a single mutational step away from the consensus sequence containing the major alleles. Therefore, the probability of observing a specific segregating amino acid is π_a divided by the total number of possible non-synonymous mutations away from an average codon. Because on average a codon is one mutational step away from $\sim$6 codons coding for a different amino acid, we estimated the expected density of specific amino-acid non-fixed states as $\pi_a/6$.

Estimating the impact of non-fixed states. If non-fixed and fixed states have constant probabilities of occurrence, p and $q = 1-p$, respectively, then the distribution of non-fixed versus fixed states in the multiple alignment is a binary distribution. Thus, the probability that a non-fixed state is observed k times in an alignment of N sequences is $m = C_N^k p^k q^{N-k}$ where C_N^k is the number of possible combinations of k out of N elements. When N is large, as is the case in our data, calculating C_N^k directly is a computationally intensive problem, and we therefore used the Poisson formula, $(pN)^k e^{-pN}/k!$, as a proxy for $C_N^k p^k q^{N-k}$, which gives a good approximation when pN is small. The expected number of times a non-fixed state is expected to be observed at a site, pN, was smaller than one for most genes (Table 1), justifying the use of the Poisson approximation.

Independent verification of the minor impact of non-fixed states. To verify that the corrected amino-acid usage was not strongly biased by non-fixed amino-acid states, we used two approaches that are not based on an estimate of the frequency of specific non-fixed amino-acid states. The first approach used a combination of phylogenetic and probabilistic considerations. Non-fixed states should occur relatively randomly on a phylogenetic tree, and the clustering of rare states on the phylogeny indicates a high likelihood that such states are fixed states. Reducing the number of sequences in the alignment also reduced the probability of occurrence of non-fixed states. Thus, we reduced the number of sequences in the alignment, to reduce the impact of non-fixed states, while keeping pairs of closely related species to increase the probability that each fixed state is found at least twice. Owing to the complications involved with reconstructing an accurate phylogeny for thousands of species, we used sequence divergence as a proxy of phylogenetic distance to select the pairs of sequences from closely related species. By reducing the number of sequences in the alignment while keeping pairs of sequences from closely related species, it was possible to reduce the expected contribution of non-fixed states to almost zero, even for those states that were observed a small number of times in the alignment. For each gene in the total data

set, we selected 200 pairs of sequences that were the most representative of the amino-acid usage of the entire data set. Using blastclust[36] we clustered all of the available sequences into 200 clusters, and for each cluster we selected two sequences. Thus, in the alignment we retained 200 pairs of sequences from phylogenetically diverse species, and within each pair the two sequences were from close relatives. For $N = 400$ and $0.0001 < p < 0.001$ (Table 1), the probability that an amino-acid state that is observed just once is non-fixed is $m = (pN)^1 e^{-pN}/1!$, such that $0.038 < m < 0.27$, and the probability that an amino-acid state that has been observed two or more times is a fixed state is therefore $r = 1 - m$, such that $0.962 > r > 0.73$. Estimating amino-acid usage from this data set taking into account only those states that have been observed more than once yielded lower estimates of average amino-acid usage (Supplementary Table 4), as was expected as a result of there being a much smaller number of sequences in the alignments. However, the measure was still ~6 amino acids per site, which corresponds to the expected dN/dS value being much higher than the observed value (Table 1).

We also used an empirical approach to test the impact of non-fixed states on our amino-acid usage measure. We eliminated rare non-fixed states directly, by producing an alignment using consensus sequences for each species with three or more sequences. The consensus sequence was constructed such that the amino-acid sequence at each site was identical to the most common amino-acid sequence in that species. This approach should effectively eliminate most of the rare states that are not expected to reach fixation (the majority of such states are expected to be rare[37]). The amino-acid usage estimated using consensus sequences from ~5% of all available species was, as expected, lower than the amino-acid usage calculated using all sequences. However, when we estimated the amino-acid usage from the same number of random sequences as there were consensus sequences for each gene, we obtained very similar results (Supplementary Table 5). This provides independent empirical evidence that the contribution of non-fixed states to amino-acid usage is minor.

Measuring dN/dS and the McDonald–Kreitman test. Because of the impossibility of reconstructing accurate phylogenies across our entire data set, we used pairwise sequence comparisons to estimate dN/dS values across genes. We estimated the dN/dS values using the codeml program in the PAML package[38] for all pairwise comparisons in the data set and kept only those comparisons for which $0.05 < dS < 0.5$, to eliminate unreliable estimates. To minimize the impact of clades with many closely related species, we first averaged the dN/dS estimates for all species with $dS < 0.5$ and then averaged the dN/dS values across clusters of pairwise comparisons such that $dS > 0.5$ between any two sequences belonging to different clusters. For Rubisco, the same procedure was followed except the maximum dS cut-off value between sequences was 0.2. The resulting dN/dS estimates came from a wide phylogenetic background from many non-overlapping clusters and show small standard deviations (Supplementary Table 6). We applied the McDonald–Kreitman test[1] in all instances when at least 20 sequences from the same species were available that had at least five polymorphic sites. All values of $\alpha < -1$ were treated as -1.

31. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* **34,** D16–D20 (2006).
32. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).
33. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30,** 3059–3066 (2002).
34. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).
35. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl Acad. Sci. USA* **76,** 5269–5273 (1979).
36. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **40,** D13–D25 (2012).
37. Sunyaev, S. *et al.* Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10,** 591–597 (2001).
38. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).