Vol 465|17 June 2010|doi:10.1038/nature09105

LETTERS

Sequence space and the ongoing expansion of the protein universe

Inna S. Povolotskaya[1] & Fyodor A. Kondrashov[1]

The need to maintain the structural and functional integrity of an evolving protein severely restricts the repertoire of acceptable amino-acid substitutions[1–4]. However, it is not known whether these restrictions impose a global limit on how far homologous protein sequences can diverge from each other. Here we explore the limits of protein evolution using sequence divergence data. We formulate a computational approach to study the rate of divergence of distant protein sequences and measure this rate for ancient proteins, those that were present in the last universal common ancestor. We show that ancient proteins are still diverging from each other, indicating an ongoing expansion of the protein sequence universe. The slow rate of this divergence is imposed by the sparseness of functional protein sequences in sequence space and the ruggedness of the protein fitness landscape: ~98 per cent of sites cannot accept an amino-acid substitution at any given moment but a vast majority of all sites may eventually be permitted to evolve when other, compensatory, changes occur. Thus, $\sim 3.5 \times 10^9$ yr has not been enough to reach the limit of divergent evolution of proteins, and for most proteins the limit of sequence similarity imposed by common function may not exceed that of random sequences.

Proteins evolve slowly because amino-acid substitutions are often deleterious owing to their effects on protein structure, expression or function[1–4]. The selective constraint imposed on evolving proteins by these factors can be very strong, as evidenced by conservative proteins from distant organisms that retain substantial similarity after $\sim 3.5 \times 10^9$ yr of evolution[5–7]. From the perspective of fitness landscapes, this constraint applies because an evolving protein must navigate ridges of high fitness in an enormous imaginary sequence space that encompasses all possible amino-acid sequences[8–14]. However, only a fraction of the sequence space can conform to the specific structural and functional characteristics of a particular protein or protein family[9,15–20].

Selection within sequence space on a small scale has been studied by characterizing negative selection against individual amino-acid substitutions[11,20,21]. However, the global constraint on evolution that a specific structure and function impose by establishing some territory within the total sequence space has not been explored. The constraint on the evolution of conservative ancient proteins is particularly strong. For example, orthologous L14 ribosomal proteins show a remarkable 40% sequence identity between the bacterium *Escherichia coli* and the archaeon *Metallosphaera sedula*, possibly indicating that the broadly defined function of these L14 ribosomal proteins can be performed only by a tiny fraction of all possible sequences and that these proteins have reached the functional limit of their divergence. Proteins that have been structurally and functionally conserved since the last universal common ancestor (LUCA) are the most likely candidates to have already reached the limit of their possible divergence by fully exploring the sequence territory that is available to them. Here we address this issue by comparing ancient homologous proteins from both similar and distant species.

Studies of protein evolution usually involve reconstruction of ancestral sequences. However, such reconstructions are unreliable for all but closely related proteins, and alternative approaches must be sought to study whether or not ancient proteins are still diverging from their common ancestral sequence. The divergence of homologous proteins from a common ancestor bears a strong similarity to the recession of galaxies in the physical universe[15,22–24]. Edwin Hubble interpreted a positive correlation between the distance to a galaxy and the rate of its recession from the Earth[25] as evidence of expansion of stellar bodies from a single point of origin: galaxies are moving away from each other and a thought experiment in which time is reversed suggests that they must have originated from the same point. We copy Hubble's approach by investigating the divergence of protein sequences from each other, which allows us to investigate the limits of protein divergence unaffected by the biases of deep ancestral state reconstructions. We relate the rate of divergence of distant homologues, which is analogous to speed in Hubble's analysis, to the protein distance (one minus the sequence identity) between them, which is analogous to physical distance (Fig. 1). In analogy with Hubble's work, a correlation between the rate of divergence of modern sequences from each other and the protein distance between them reflects the rate of their divergence from a common ancestor.

The rate of divergence of sequences, whether similar or distant, from each other can be determined by using ancestral state reconstruction of protein sequences in closely related species and relating this reconstruction to another, reference, sequence from a more distant species. The directionality of a substitution (polarization) can be determined from a multiple alignment of sequences from two sister species and one or more outgroup species, assuming that the outgroup sequences reflect the ancestral state of the sister species. When such a cluster of similar sequences is related to a reference sequence of a more distant orthologue, two types of informative site can be identified in the resulting cluster-reference alignment: (1) those where the ancestral amino acids for the two sister species are identical to the corresponding amino acid in the reference sequence, and (2) those where one or both of these two amino acids is different from the reference amino acid. A substitution that occurred in either of the sister species can lead away from (if it occurred at a site of type 1), towards or neither away from nor towards the reference sequence (at some sites of type 2) (Fig. 2a). Thus, from a cluster-reference alignment we can obtain the numbers of substitutions in the sister sequences that moved the sequence away from (N_a) and towards (N_t) the reference sequence.

If $N_t/N_a < 1$ then the sister sequences are evolving away from the reference in sequence space, and, conversely, if $N_t/N_a > 1$ then the sister sequences are evolving towards the reference. If $N_t/N_a = 1$ then the distance in sequence space between them is at an evolutionary

[1]Bioinformatics and Genomics Programme, Centre for Genomic Regulation, Calle Dr Aiguader 88, Barcelona Biomedical Research Park Building, 08003 Barcelona, Spain.

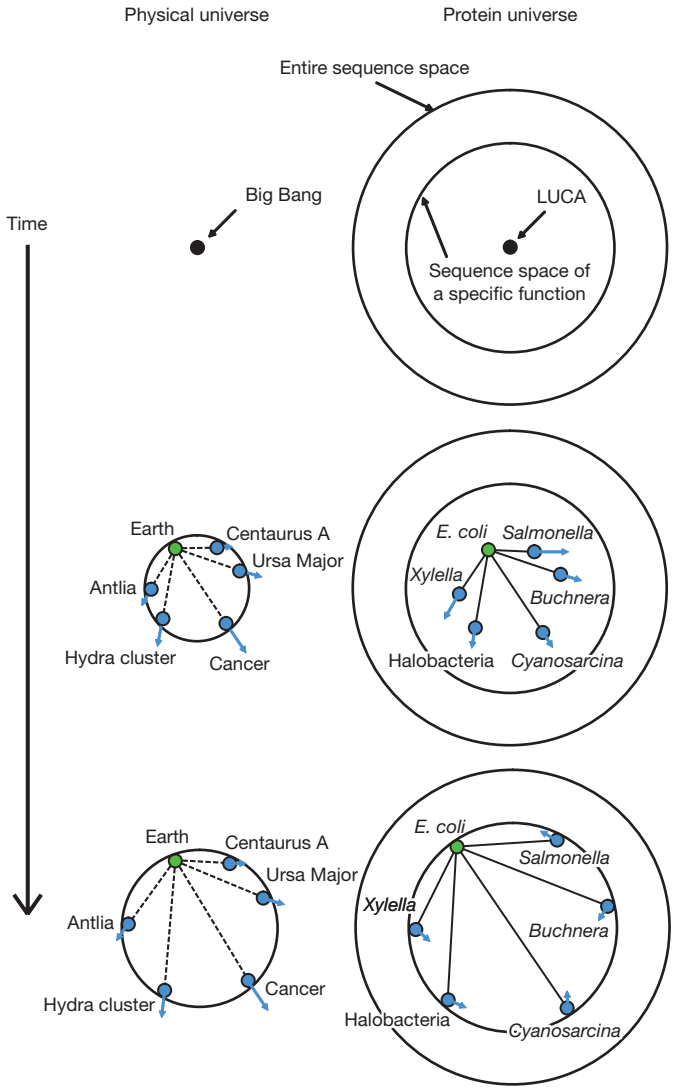

Figure 1 | Expansion of the physical and the protein universes. Physical space continues to expand after the Big Bang, whereas sequence space is a pre-defined abstract entity a limited subspace of which corresponds to sequences with a specific function. In both cases, the correlation between distance from one point of observation (green) to reference points in space (blue) and the relative rates of divergence (blue arrows) reveals whether or not these objects (proteins or galaxies) recede from a common point of origin (middle diagram) or have reached the limit of their divergence (shown only for proteins in the bottom-right diagram).

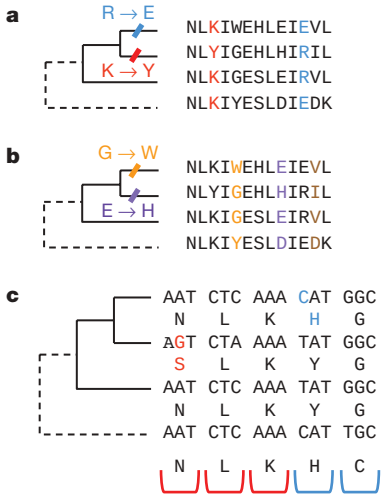

Figure 2 | Measuring the rate of divergence of distant protein sequences. We infer the directionality of a substitution by the use of closely related outgroups (solid lines) and then relate them to an arbitrarily distant reference sequence (broken line). **a**, A substitution leads 'away' from the reference sequence (red) when the ancestral amino acid is identical to that at the orthologous site of the reference sequence, and leads 'towards' the reference sequence (blue) when the derived amino acid is identical to the amino acid in the reference sequence. **b**, Use of amino-acid similarity to infer the directionality of substitutions where neither the derived nor the ancestral amino acid is identical to the reference amino acid. The substitution shown in purple leads the evolving sequence away from the reference sequence (His–Asp and Glu–Asp BLOSUM62 similarity scores are −1 and 2, respectively), whereas the orange one leads the evolving sequence towards the reference sequence (Trp–Tyr score, 2; Gly–Tyr score, −3). Directionality cannot be determined when the BLOSUM62 scores of the ancestral and the derived amino acids are equal to the score of the reference amino acid (brown). **c**, Divergent sites (red brackets) and convergent sites (blue brackets).

equilibrium. To investigate the dynamics and the limits of protein divergence, we relate N_t/N_a to D, the protein distance between the ancestral state of the sister sequences and the reference sequence.

We applied this approach to alignments obtained from 572 clusters of orthologous groups[26] (COGs) that have been previously inferred to have been present in the LUCA[5] encompassing a total of 836 different bacterial and archaeal genomes. These COGs included 28 quadruplets of closely related genomes yielding a total of 13,589,431 cluster-reference alignments (Methods). We obtained N_t, N_a and D (Supplementary Fig. 1) for each alignment and plotted the average N_a/N_t value across different COGs as a function of D. Figure 3 shows that, regardless of their similarity, ancient proteins are still diverging from each other and therefore have not yet reached the limit of their sequence divergence.

A large fraction of substitutions at type-2 sites are ignored by this approach because they neither make a sister sequence identical to the reference nor move it away from it, but instead replace one difference with another. To deal with such cases, we used BLOSUM62, an amino-acid similarity matrix. If the similarity of the reference and the derived amino acids was greater than that of the reference and the ancestral amino acids, then the substitution was counted as leading towards the reference sequence. Conversely, if the reverse was true then the substitution was counted as leading away from the reference sequence (Fig. 2b). This approach uses a great majority of substitutions at type-2 sites of the cluster-reference alignment, and the N_t/N_a ratio that includes such substitutions shows the same dynamics (Fig. 3).

Our use of outgroups as a proxy for the ancestral state, or our assumption that the evolution of the reference sequence on the timescale of sister species divergence is negligible, may lead to a systematic underestimation of N_t/N_a. We explore this possibility by measuring N_t/N_a for nucleotide substitutions at fourfold-synonymous sites (Fig. 3). Divergence at fourfold-synonymous sites rapidly reaches an equilibrium, ruling out the possibility of the continuing divergence of protein sequences being an artefact of our method. Another potential source of bias is the accumulation of slightly deleterious amino-acid substitutions and polymorphisms in the terminal branches of the phylogenetic trees leading up to the sister species[27]. However, the exclusion of substitutions from terminal branches does not affect our results (Supplementary Figs 2 and 3).

Our data reveal an ongoing expansion of the protein universe, such that most extant protein sequences are still diverging from each other and from the ancestral LUCA sequence, and have not yet reached the structural and functional limits in sequence space. The linear relationship between N_t/N_a and D suggests that, in contrast to the accelerating expansion of physical space[28], the expansion of proteins in sequence space is relatively constant (Fig. 3). However, 3.5×10^9 yr of constant evolution was not enough for ancient proteins to reach the limits of functional sequence divergence, and the ongoing divergence must therefore be extremely slow.

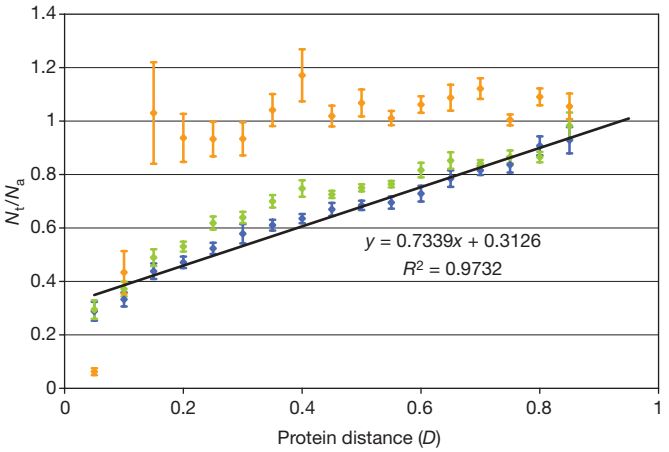

Figure 3 | The rate of expansion of the protein sequence universe. N_t/N_a values for sites where the derived or the ancestral amino acid was identical to the reference amino acid (blue) and for all type-1 and type-2 sites (green), including those where the directionality of the substitution was determined from the BLOSUM62 matrix. The N_t/N_a ratio for fourfold-synonymous sites is shown in orange. Least-squares regression with correlation coefficient R is shown for the N_t/N_a values calculated without the BLOSUM62 correction (blue). Error bars, s.e.m.

The rate of evolution may be inhibited by two factors: non-epistatic negative selection, where the fitness impact of every allele is independent of genetic context and environmental factors; or a high prevalence of sign epistasis. Sign epistasis is a property of fitness landscapes such that an amino acid at a specific site confers high fitness in one genetic background and low fitness in another[12]. If such situations are common, the fitness landscape can be described as rugged[8] and the order in which substitutions accumulate is restricted[10], with the result that only a small fraction of the substitutions that are possible in at least one genetic background can occur at any given moment[10–14]. Also, the rate of evolution will be slow[10,14] because fitness ridges will consist of fewer and longer evolutionarily accessible paths in sequence space (Fig. 4), and evolving proteins may meander through sequence space on such a rugged fitness landscape for a long time before reaching the limit of their divergence.

Two sequences diverging under non-epistatic selection are expected to reach their asymptotic level of divergence sooner or on the same timescale as selectively neutral diverging sequences[29]. In the absence of epistasis, the relative fitnesses of alleles at all sites are independent of each other, selection against specific alleles is present in all genetic contexts and entire regions of the sequence space confer low fitness (Fig. 4c). As a result, the limit of sequence divergence for proteins under non-epistatic negative selection is much closer, in terms of protein distance, than that for neutrally evolving protein sequences, and such selection reduces the time to reach this limit[29]. Since LUCA, ≥100 of substitutions have occurred per synonymous site[30] and sequences under non-epistatic selection must already have reached the limit of divergence. Thus, the ongoing divergence of ancient proteins (Fig. 3) is inconsistent with non-epistatic negative selection being the main factor responsible for their slow evolution.

To determine the extent of epistasis in ancient proteins, we compared their divergence across different types of site. In the cluster-reference alignments, we identified 'divergent' sites, those of type 1, and 'convergent' sites, those of type 2 where the ancestral amino acid could be changed to the reference amino acid with a single nucleotide substitution. At divergent sites any substitution moves a sister species away from the reference, whereas at convergent sites a fraction of substitutions move a sister species towards the reference. For divergent sites we recorded the total rate of non-synonymous evolution (K_d), and for convergent sites we recorded the rate of those substitutions that moved a sister species towards the reference (K_c), disregarding other substitutions (Fig. 2c). These measures are analogous

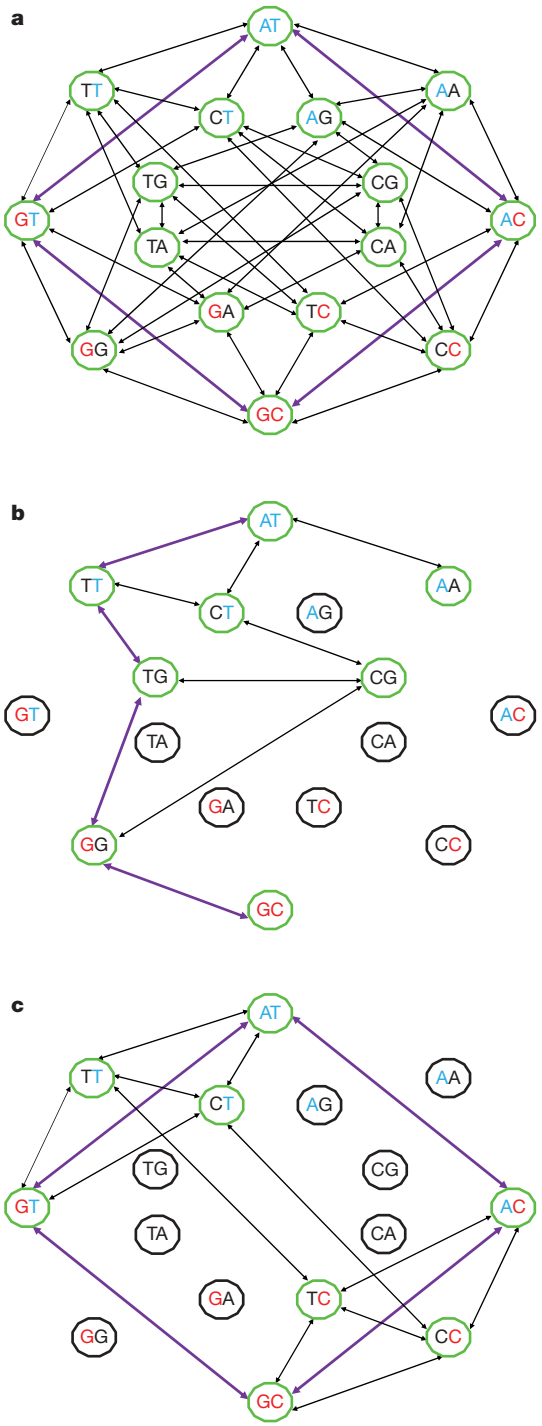

Figure 4 | Sequence space of two nucleotide sites. Fitness landscapes as modified from fig. 1 of ref. 9 as graphs in which vertices correspond to unique sequences and edges correspond to mutational steps. **a**, When all sequences confer high fitness (green vertices), there are two shortest paths (purple arrows), two steps long, that connect two arbitrary points in sequence space (AT and GC). **b**, When half of all sequences confer low fitness (black vertices), the number of evolutionarily accessible paths is reduced and their lengths are increased. **c**, When alleles G and A in the second site are deleterious independently of the alleles in the first site, the fitness landscape is a simple subset of the total sequence space and the availability of short paths between vertices of high fitness is unaffected. The numbers of vertices conferring high fitness are the same in the fitness landscapes shown in **b** and **c**; however, **b** represents a highly rugged fitness ridge whereas **c** shows a fitness ridge without any epistatic interactions.

to K_a, the rate of non-synonymous evolution, and differ from it only by the types of site that are considered, such that K_d and K_c are different components of K_a. We related K_d and K_c to the rate of fourfold-synonymous divergence (K_4) between the sister species and plotted K_d/K_4 and K_c/K_4 as functions of D (Fig. 5).

Non-epistatic and epistatic selection lead to different predictions for the relationships between K_d/K_4 and D and K_c/K_4 and D. If evolution proceeds on a non-epistatic fitness landscape (Fig. 4c), K_d/K_4 should decrease as D increases, owing to the rapid initial accumulation of nearly neutral changes, whereas K_c/K_4 should remain roughly constant in D. In contrast, for proteins evolving on an epistatic fitness landscape (Fig. 4b), K_d/K_4 should remain approximately constant but K_c/K_4 should decrease as D increases, starting from $K_c/K_4 \approx 1$ when $D = 0$, because relative fitnesses at orthologous sites become progressively less similar in diverging proteins. Our data clearly favour an epistatic fitness landscape of ancient proteins (Fig. 5).

As first recognized by John Maynard Smith,[9] "functional proteins must form a continuous network which can be traversed by unit mutational steps without passing through nonfunctional intermediates". This network can be represented by a graph in which vertices represent unique sequences connected by edges representing single mutational steps (Fig. 4a). Although some aspects of the protein fitness landscape (that is, the network) have been probed, we remain largely ignorant of its global structure spanning sequence space. Our data indicate high incidence of sign epistasis, such that selection associated with a substitution at one site depends on other sites, and only a small fraction of all possible sequences confer high fitness, leading to slow divergence of ancient proteins. In principle, gradual environmental changes, or evolution of the function itself, can also contribute to the slow and continuous divergence of ancient proteins. However, the functions of ancient proteins assigned to the same COG are similar[5–7,26] and the functional changes are therefore unlikely to be a major cause of their continuing divergence. In contrast, compensatory interactions between different sites in the protein structure are a well-documented phenomenon[11–14,16–21] that is expected to result in complex, multidimensional sign epistasis and to contribute substantially to the ruggedness of protein fitness landscapes[11–13].

As a protein evolves along a rugged fitness ridge, some previously forbidden amino-acid substitutions at a site become acceptable and some previously acceptable substitutions become forbidden, owing to compensatory substitutions at other sites of the same protein or its interaction partners. Thus, the ruggedness of the fitness landscape in sequence space can be roughly described by the change in the probability that a particular amino acid is tolerated at a site as the entire protein sequence evolves. A general estimate of this probability can be obtained from the dependence of K_d/K_4 and K_c/K_4 on D. We find that $K_d/K_4 \approx 0.02$ and is approximately independent of D, showing that in ancient proteins only ~2% of amino-acid substitutions are tolerated at any given moment. Thus, for a specific ancient protein an orthologue of any similarity has the same power to predict the fitness impacts of substitutions at identical sites.

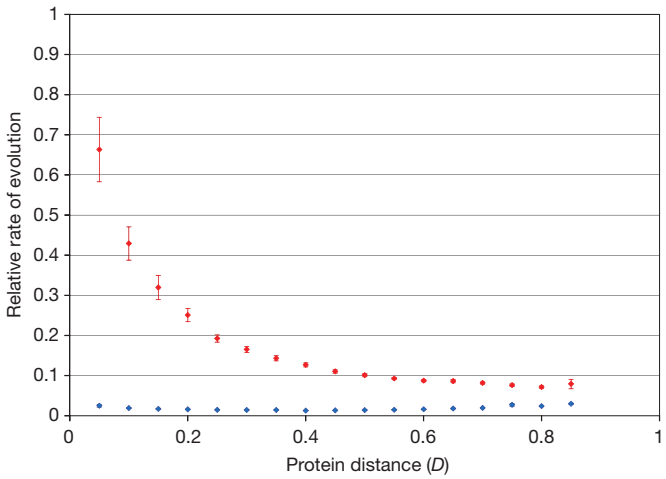

Figure 5 | Divergent and convergent evolution. The relative rates of divergent evolution, K_d/K_4 (blue), and of convergent evolution, K_c/K_4 (red), at divergent and convergent sites, respectively. Error bars, s.e.m.

The dependence of K_c/K_4 on D reflects the rate at which a substitution that was previously tolerated becomes forbidden as the proteins evolve. The K_c/K_4 ratio is a monotonically decreasing function of D that indicates a rapid loss of reversibility of substitutions, with half of the reversals becoming forbidden after ~10% protein divergence (Fig. 5). As D increases, K_c/K_4 approaches an asymptotic value, such that in only 10% of cases can amino acids at non-matching sites in two sequences be interchanged without a substantial loss in fitness, regardless of whether the sequences have diverged by 40% or 80% (Fig. 5). Thus, only similar orthologues are strong predictors of which amino acids can be tolerated in another protein, and all orthologues with divergence above 40% perform equally poorly. Taken together, these data indicate that very few substitutions can be accepted at any time and that amino acids which have been accepted at a site in the past quickly become deleterious as the rest of the protein keeps evolving.

The following picture of the protein sequence space emerges from our analysis. Ridges of high fitness corresponding to specific ancient proteins occupy a tiny fraction of the entire volume of the sequence space. However, these ridges are long and thin and can be more accurately visualized as a wide-mesh net spanning a large part of sequence space, rather than as a small volume within the space. Such fitness ridges imply that sign epistasis and compensatory evolution in ancient proteins must be common. Our data show that >90% of the sites in any protein can eventually accept a substitution given the right combination of amino acids at other sites, although it is not clear whether such substitutions are predominantly neutral or beneficial. Regardless of the importance of positive selection in protein divergence, it seems that many sites are conserved because there has not been enough time to create the right combination of amino acids at other sites to allow them to evolve, which may take billions of years.

Many biological phenomena, such as the evolution of sex[10] and compensatory evolution[11–14], depend on the degree and nature of epistasis. Our data indicate that protein fitness landscapes cannot be described outside the framework of multidimensional epistasis[10], which can reflect non-trivial compensatory interactions, and that simple forms of epistasis, such as synergistic or antagonistic epistasis, do not provide an adequate theoretical framework for understanding protein evolution[9–14].

Sequence similarity between functionally related proteins in the Eukaryota, Archaea and Bacteria is one of the strongest arguments supporting the common ancestry of all of life and is used to identify proteins likely to have been present in LUCA[6,7,23]. Our analysis shows that it is conceivable that many more proteins were present in LUCA but have since diverged beyond our ability to detect their homology, and that given enough time some of the currently identifiable orthologues among the major kingdoms of life will diverge beyond recognition. Finally, our observation of receding protein sequences provides novel evidence of the common ancestry of life.

METHODS SUMMARY

All available completely sequenced prokaryote genomes from GenBank were combined with the COG database[26], with 572 COGs that where predicted to have been present in LUCA[5] being used in the final analysis. We obtained 28 quadruplets of closely related species that were previously assigned to COGs such that the sister species showed an average synonymous distance of between 0.1 and 0.8 and the average protein divergence between the closest outgroup and the sister species was <0.15. Orthologues from the 28 quadruplets species were constructed using the two-directional best-BLAST-hit approach. The cluster-reference alignments were selected from the 572 COGs present in LUCA and consisted of orthologue quadruplets with other homologous reference sequences from species that were phylogenetically more distant from the sister species than the most distant outgroup.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 November 2009; accepted 19 April 2010.
Published online 19 May 2010.

1. Aravind, L., Mazumder, R., Vasudevan, S. & Koonin, E. V. Trends in protein evolution inferred from sequence and structure analysis. *Curr. Opin. Struct. Biol.* **12,** 392–399 (2002).
2. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Rev. Genet.* **6,** 678–687 (2005).
3. Camps, M., Herman, A., Loh, E. & Loeb, L. A. Genetic constraints on protein evolution. *Crit. Rev. Biochem. Mol. Biol.* **42,** 313–326 (2007).
4. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19,** 596–604 (2009).
5. Mirkin, B. G., Fenner, T. I., Galperin, M. Y. & Koonin, E. V. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3,** 2 (2003).
6. Koonin, E. V. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Rev. Microbiol.* **1,** 127–136 (2003).
7. Ranea, J. A., Sillero, A., Thornton, J. M. & Orengo, C. A. Protein superfamily evolution and the last universal common ancestor (LUCA). *J. Mol. Evol.* **63,** 513–525 (2006).
8. Wright, S. in *Proc. Sixth Int. Congr. Genet.* Vol. 1 (ed. Jones, D. F.) 356–366 (Genetics Society of America, 1932).
9. Maynard Smith, J. Natural selection and the concept of a protein space. *Nature* **225,** 563–564 (1970).
10. Kondrashov, F. A. & Kondrashov, A. S. Multidimensional epistasis and the disadvantage of sex. *Proc. Natl Acad. Sci. USA* **98,** 12089–12092 (2001).
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. Weinreich, D. M., Watson, R. A. & Chao, L. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* **59,** 1165–1174 (2005).
13. Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312,** 111–114 (2006).
14. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445,** 383–386 (2007).
15. Koonin, E. V., Wolf, Y. I. & Karev, G. P. The structure of the protein universe and genome evolution. *Nature* **420,** 218–223 (2002).
16. Lesk, A. M. & Chothia, C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136,** 225–230 (1980).
17. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science* **247,** 1306–1310 (1990).
18. Heger, A. & Holm, L. Towards a covering set of protein family profiles. *Prog. Biophys. Mol. Biol.* **73,** 321–337 (2000).
19. Taylor, S. V., Walter, K. U., Kast, P. & Hilvert, D. Searching sequence space for protein catalysts. *Proc. Natl Acad. Sci. USA* **98,** 10596–10601 (2001).
20. Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proc. Natl Acad. Sci. USA* **101,** 9205–9210 (2004).
21. Huang, W., Petrosino, J., Hirsch, M., Shenkin, P. S. & Palzkill, T. Amino acid sequence determinants of beta-lactamase structure and activity. *J. Mol. Biol.* **258,** 688–703 (1996).
22. Holm, L. & Sander, C. Mapping the protein universe. *Science* **273,** 595–602 (1996).
23. Doolittle, W. F. The nature of the universal ancestor and the evolution of the proteome. *Curr. Opin. Struct. Biol.* **10,** 355–358 (2000).
24. Dokholyan, N. V., Shakhnovich, B. & Shakhnovich, E. I. Expanding protein universe and its origin from the biological Big Bang. *Proc. Natl Acad. Sci. USA* **99,** 14132–14136 (2002).
25. Hubble, E. A relation between distance and radial velocity among extra-galactic nebulae. *Proc. Natl Acad. Sci. USA* **15,** 168–173 (1929).
26. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).
27. Golding, B. & Felsenstein, J. A maximum likelihood approach to the detection of selection from a phylogeny. *J. Mol. Evol.* **31,** 511–523 (1990).
28. Guzzo, L. *et al.* A test of the nature of cosmic acceleration using galaxy redshift distortions. *Nature* **451,** 541–544 (2008).
29. Kondrashov, A. S., Povolotskaya, I. S., Ivankov, D. N. & Kondrashov, F. A. Rate of sequence divergence under constant selection. *Biol. Direct* **5,** 5 (2010).
30. Jordan, I. K. *et al.* A universal trend of amino acid gain and loss in protein evolution. *Nature* **433,** 633–638 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. Koonin, Y. Wolf, A. Lobkovsky, D. Petrov, D. Ivankov, J. Sharpe, B. Lehner, Y. Jaeger, P. Vlasov, M. Ptitsyn and M. Roytberg for discussions and A. Kondrashov for extensive feedback on our manuscript. We thank D. Tawfik for inspiring us to start the investigation of the functional limits in sequence space.

Author Contributions I.S.P. performed all analyses and obtained all of the data. F.A.K. conceived the study and drafted the manuscript. Both authors participated in the design of the analyses and the interpretation of the results.

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 this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to F.A.K. (fyodor.kondrashov@crg.es).

METHODS

We obtained 836 completely sequenced prokaryote genomes from GenBank and used a two-directional best-BLAST-hit approach[26] to assign protein orthologues for those genomes not already included in the COG database. Individual cluster-reference alignments for each COGs were then aligned using the MUSCLE 3.6 program[31]. For our analyses, we used only the 572 COGs that were predicted to have been present in LUCA[5]. We obtained quadruplets of orthologous genes from 28 quadruplets of closely related genomes from among those we assigned to COGs using data from the ATGC database[32]. The quadruplets were selected such that the average synonymous distance (K_s) between the sister species was >0.1 and <0.8 and the average protein divergence between the closest outgroup and the sister species was <0.15. As one of the tests to eliminate the possibility of unknown biases in our ancestral reconstruction, we selected a subset of 15 quadruplets in which the non-synonymous divergence between the sister species was $K_a < 0.03$ and that between both sister species and the outgroup was $K_a < 0.07$, such that these distances minimize the possibility of potential bias in the ancestral reconstruction affecting our results[33]. These data yielded the same results as the data obtained using all 28 quadruplets (Supplementary Fig. 4). Of these quadruplets, five were selected to measure the rate of divergence of synonymous sites with $K_s < 0.3$ between any of the sister species and the outgroup. Quintuplets with four sister species (Supplementary Fig. 2) with $K_s < 0.1$ within one pair and $K_s < 0.5$ between the pair and the outgroup were selected to test the possibility that substitutions on the terminal branches are the main contributors to protein divergence. A list of species and accession numbers for all sequences used in our study is available in Supplementary Tables 1 and 2. The cluster-reference alignments were selected from the 572 COGs present in LUCA and represented combinations of 28 quadruplets with other orthologous reference sequences from species that were phylogenetically more distant from the sister species than from the outgroup. We reconstructed the ancestral state between the two sister species using three different methods. The data reported in the main text were obtained by a Bayesian approach implemented in version 3.1.2 of MRBAYES[34]. With the Bayesian approach, for each substitution we used the distribution of posterior probabilities of all ancestral states, with N_t and N_a representing the sum of all substitutions multiplied by their posterior probabilities. When taking an average N_t/N_a ratio across different COGs, we discounted all ratios for which the sum of N_t or N_a values for one COG was less than two. Data obtained by a maximum-likelihood approach using the PAML 4.1 program package[35] and a simple parsimony reconstruction using a single closest outgroup yielded the same results as a Bayesian approach (Supplementary Fig. 4).

To eliminate the possibility of horizontal gene transfer from a sister species into a reference genome, we eliminated from consideration individual orthologues that were on average closer to the sister species than the outgroup orthologue. We tested the impact of paralogues on our data by calculating N_t/N_a as a function of D using four methods of selecting different homologues from each COG as a reference sequence: choosing the closest reference homologue, the most distant homologue, the average of all homologues and all homologues from each COG. All four approaches revealed identical relationships between N_t/N_a and D (data not shown), and the data reported in the main text was obtained from 5,824,167 alignments using only the closest homologue from each COG. To test the effects of multiple alignment biases, we replicated all of our data by obtaining N_t, N_a and D from large alignments of all sequences in each COG that were aligned using version 3.6 of the MUSCLE[31] program. We also repeated our analyses with quality-trimmed alignments using the TRIMAL 1.2 program[36]. The different alignment approaches all yielded the same results, demonstrating a high degree of robustness of our results to alignment noise (Supplementary Fig. 5).

31. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
32. Novichkov, P. S., Ratnere, I., Wolf, Y. I., Koonin, E. V. & Dubchak, I. ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. *Nucleic Acids Res.* **37,** D448–D454 (2009).
33. Goldstein, R. A. & Pollock, D. D. Observations of amino acid gain and loss during protein evolution are explained by statistical bias. *Mol. Biol. Evol.* **23,** 1444–1449 (2006).
34. Ronquist, F. & Huelsenbeck, J. P. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574 (2003).
35. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
36. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972–1973 (2009).