scientific reports

Check for updates

OPEN

Phylogenetic analysis of mutational robustness based on codon usage supports that the standard genetic code does not prefer extreme environments

Ádám Radványi[1]✉ & Ádám Kun[2,3,4]

The mutational robustness of the genetic code is rarely discussed in the context of biological diversity, such as codon usage and related factors, often considered as independent of the actual organism's proteome. Here we put the living beings back to picture and use distortion as a metric of mutational robustness. Distortion estimates the expected severities of non-synonymous mutations measuring it by amino acid physicochemical properties and weighting for codon usage. Using the biological variance of codon frequencies, we interpret the mutational robustness of the standard genetic code with regards to their corresponding environments and genomic compositions (GC-content). Employing phylogenetic analyses, we show that coding fidelity in physicochemical properties can deteriorate with codon usages adapted to extreme environments and these putative effects are not the artefacts of phylogenetic bias. High temperature environments select for codon usages with decreased mutational robustness of hydrophobic, volumetric, and isoelectric properties. Selection at high saline concentrations also leads to reduced fidelity in polar and isoelectric patterns. These show that the genetic code performs best with mesophilic codon usages, strengthening the view that LUCA or its ancestors preferred lower temperature environments. Taxonomic implications, such as rooting the tree of life, are also discussed.

The origin of translational apparatus and the genetic code is amongst the greatest conundrums of Life[1]. Its fundamental challenge is to uncover the constraints, historical accidents, and evolutionary driving forces that could have shaped the standard codon table. The current views propose the general mechanisms of (1) stereochemical affinity between codons and attributed amino acids (stereochemical theory[2,3]), (2) coevolution between the biosynthetic paths of amino acids and cognate codons (coevolution theory[4,5]), and (3) minimization of translation errors (adaptive, physicochemical or error minimization theory[6,7]) as possible explanations for the overall structure of the standard genetic code. Here, we focus and expand on the third one.

Although these existing hypotheses for the development of the genetic code are still hotly debated (including other theories, see other reviews[8,9] for a recent overview of the field), the scientific community tends to agree on that the code is robust to mutations because its structure reduces the deleterious effects of translational errors. The error minimization theory proposes that the universal genetic code was shaped under selective forces that made the code at least partly optimized for fidelity in the physicochemical properties of mutated amino acids (or such aspects played a role during its development). With a few notable exceptions showing that the code can be locally improved with codon reassignments[10,11], several studies have pointed out that the overwhelming portion of random alternative genetic code structures have inferior error capacities, leading to the argument

[1]Department of Plant Systematics, Ecology and Theoretical Biology, Institute of Biology, Eötvös Loránd University, Budapest, Hungary. [2]Evolutionary Systems Research Group, Centre for Ecological Research, Institute of Evolution, Budapest, Hungary. [3]Parmenides Centre for the Conceptual Foundation of Science, Pullach, Germany. [4]MTA-ELTE Theoretical Biology and Evolutionary Ecology Research Group, Budapest, Hungary. ✉email: adamradvanyi117@gmail.com

of "optimality" of the standard genetic code[6,7,12]. The basis for these comparisons is the average fitness cost of replacing one amino acid with another due to mutation, measured on a scale of some physicochemical property (e.g. hydrophobicity or polarity).

Unfortunately, this is only one piece of the puzzle, because the majority of analyses operate under the implicit assumption that codons occur in a uniform distribution, which fails to address the possible variance in codon usage and its effect on robustness[13,14]. In contrast to this, a large body of studies has shown that adaptions to special environments and lifestyles, such as high salinity or extremely high temperatures, select for characteristic amino acid and codon compositions due to their special physicochemical requirements with regards to protein structure[15–25]. Also, there is an obvious correspondence between codon usage frequencies and GC-content[21,26,27]. The latter is associated with specific environments and lifestyles[28,29] and mutational bias[30–32].

Being aware of these facts, one can narrow the question further: Supposing that the universal genetic code is optimized for mutational robustness, in what condition does it have maximal efficiency? We could argue that it was the most likely environment to witness the final mappings between amino acids and their respective codons.

The information theoretic measure of distortion[33,34] is a suitable concept to study the structure and "environmental behaviour" of the standard genetic code. In this framework, distortion is the expected average effect of mutations on the level of amino acid physicochemical properties, given a distribution of codon usage. It contains the same essentials as previous measures of code performance: (1) the estimate for the cost of faulty translation (one amino acid replaced by another), usually based on some physicochemical trait (e.g. hydrophobicity), and (2) the estimated probability of such translation errors being the result of a code in question. However, unlike conventional error definitions, distortion builds in a third term (3) by weighting for codon usage. Then, supplemented with a simple "background mutation model" (see "Methods" section), one can use distortion to compare the mutational robustness of codon usage profiles associated with different environments.

A previous study (Radványi and Kun, submitted) has suggested that extremophile codon usages might lead to diminished mutational robustness compared to that of mesophiles. Consequently, in thermophiles, the average effect of mutations, that is the distortion in physicochemical properties on the level of amino acids, was estimated to be greater than in mesophiles. Similar were the implications with regards to the GC-content of the coding region, implying a universal AT-bias. In that analysis, phylogenetic bias was not accounted for. Here we show that our result remains robust to taxon sampling.

Methods

Data. For codon usage profiles, we used the Uniprot Reference Proteome database within UniProtKB[35]. A Python script was used on the corresponding mRNAs to calculate nucleobase and codon distributions along with the distortion measures for each organism. This data was cross-referenced with optimal environmental conditions via NCBI Taxonomy ID-s[36]. The environmental data for optimal growth temperature, pH, and salt concentration were obtained from the BacDive database[37]. S16 rRNAs were retrieved from the 16S RefSeq Nucleotide[38] sequence records. Sequences were aligned using MUSCLE[39]. The final dataset contained 64 taxa (8 archaeal and 56 bacterial), representative of the molecular diversity in each domain.

Phylogenetic tree construction. We used Beast v1.10.4[40] for a Bayesian analysis. The S16 rRNA phylogenetic tree was built from the 64-sequence alignment with a GTR[41] model, a gamma law with eight categories and an estimated proportion of invariant sites, using default priors, and an uncorrelated relaxed clock. Chains were run for 50,000,000 generations and samples were collected in each 1000 generations. The analysis in Tracer[42] demonstrated a good mixing of the chains. A burn-in of 5,000,000 samples was discarded, and a maximum clade credibility tree was computed from the remaining samples (Supplementary Data S1 online).

Distortion as a measure of mutational robustness. In order to provide a practical measure for the error-rate of mistranslation and point-mutations, the information theoretic concept of distortion (Eq. 1)[33,34] is used to estimate the average effect (cost per symbol) of mutations given a source distribution of codons and the uncertainty of the code resulting from noise, i.e. the probability of codon c_i mutating into c_j (see next section about background mutation model). Another important element of distortion is the distortion matrix, which reflects the underlying genetic code and corresponding physicochemical properties. This distortion matrix is essentially identical with matrices widely used in other studies, summarizing the errors of one amino acid mutating into another[6,7]. Distortion matrix with elements $d(aa_i,aa_j)$ specifies the distortion associated with mistaking the encoded symbol aa_i (amino acid) in the source (X) and reproducing it as aa_j in the reproduced copy (Y). We define $d\left(aa_i, aa_j\right) = 0$ if $aa_i = aa_j$, that is, c_i and c_j codes for the same amino acid.

$$D = \sum_{i,j} P(c_i) \times P\left(Y = c_j | X = c_i\right) \times d(aa_i, aa_j) \tag{1}$$

Distinct distortion matrices were defined to provide different physicochemical measures of robustness. We decided to use properties made available by Haig and Hurst[7]. These include polar requirement, hydropathy, molecular volume and isoelectric point, yielding four different measures of code performance, denoted as D_{Hyd}, D_{Pol}, D_{Vol}, D_{pI} (Supplementary Data S2 online).

Background mutation model. The conditional probabilities $P(Y=c_j \mid X=c_i)$ are the result of random mutations appearing in the genome and describe the chance of codon c_i mutating into codon c_j. In order to approximate these probabilities, a simple background mutation model is required describing the generalized mechanism for spontaneous DNA mutations. However, we must estimate the raw, a priori performance of the

genetic code without natural selection introducing additional bias. We introduce a simplified model of random amino acid mutations (Eqs.2–4), which essentially invokes Kimura's two parameter model[43]. Here, κ denotes the transition/transversion rate ratio, otherwise known as ti/tv ratio; p_{ti} is the probability of transition, p_{tv} is the probability of transversion, and μ is the mutation rate. The inherent structure of the genetic code defines the probability of which codon i mutates into codon j given that a transition or transversion occurs; these are denoted by terms $P(c_i \rightarrow c_j \mid ti)$ and $P(c_i \rightarrow c_j \mid tv)$, respectively.

$$\kappa = \frac{p_{ti}}{p_{tv}} \tag{2}$$

$$P(Y = c_i|X = c_j) = \mu \times \left[\frac{\kappa}{(1+\kappa)} \times P(c_i \rightarrow c_j|ti) + \frac{1}{(1+\kappa)} \times P(c_i \rightarrow c_j|tv)\right] \tag{3}$$

$$P(Y = c_i|X = c_i) = 1 - \mu \tag{4}$$

Expected proteomic distortions were then calculated for each taxon's codon composition. Since our goal is a comparative analysis between taxa, the effect of μ is unimportant. Our preliminary work showed that although ti/tv ratio has a quantitative impact on distortion, the qualitative outcomes remain robust (Radványi and Kun, submitted). Our calculations concerning the distortion values were restricted to $\kappa = 2.5$ (roughly 71% of mutations are transitions), approximating ratios encountered by studies of genome-wide and intronic sequences[30,44]. Such regions ought to represent a more relaxed state of selection against mutations, hence providing a closer estimate of the background ratio of ti/tv prior to selection.

Comparative analysis. To correct for non-independence because of common ancestry of species, we performed Phylogenetic Generalized Least Squares (PGLS) models on the combined data. The analyses were carried out with the *ape*[45], *phytools*[46], *caper*[47] and *geiger*[48] packages in RStudio v3.5[49]. Response variables D_{Hyd}, D_{Pol}, D_{Vol} and D_{pI} were modelled separately using the available environmental variables as predictors. The genomic content of guanine and cytosine (GC-content) was also used as predictor since preliminary studies have shown its predominant effect on the codon and amino acid composition of proteomes[21,26,27]. Model assumptions were checked; no violations were apparent. Based on Akaike information criterion (AIC) values, we apply a Brownian model; further branch length transformations and correlation structures did not result in generally better fits.

Results

In all four cases of physicochemical distortions, the PGLS resulted in significant models. In the obtained phylogenetic tree, deeper phylogenetic topologies are well supported by posterior values (Fig. 1). The results of the PGLS regressions are shown on Fig. 2, including partial regression lines. The standardized partial coefficients (where variables were Z-transformed prior to analysis) can be found as Supplementary Table S1 online. The highest proportion of explained variances is found in the case of distortion of hydrophobic properties ($R^2 = 0.633$; $F_{4,59} = 25.492$; $p = 2.726 \times 10^{-12}$). The second highest proportion is explained for the distortion calculated for isoelectric points ($R^2 = 0.397$; $F_{4,59} = 9.689$; $p = 4.311 \times 10^{-6}$), followed by that of the volumetric distortion ($R^2 = 0.361$; $F_{4,59} = 8.314$; $p = 2.180 \times 10^{-5}$). The least amount of variance was found in the case of distortion in polar requirements ($R^2 = 0.249$; $F_{4,59} = 4.880$; $p = 1.816 \times 10^{-3}$).

The effect of GC-content. GC-content has a significant positive effect on hydrophobic distortion ($\beta = 0.237$; $t = 9.464$; $p = 1.945 \times 10^{-13}$). A less dominant, but significant negative effect is encountered in the distortion of isoelectric properties ($\beta = -0.078$; $t = -2.147$; $p = 0.036$). In other words, increasing the GC-content of the coding region results in lowered accuracy for hydrophobic traits, but the maintenance of molecular patterns related to isoelectric points becomes easier.

Environmental effects. The effect of optimal growth temperature was positive and significant on hydrophobic ($\beta = 4.510 \times 10^{-4}$; $t = 2.690$; $p = 0.009$), volumetric ($\beta = 0.015$; $t = 5.433$; $p = 1.100 \times 10^{-6}$) and isoelectric distortion ($\beta = 0.001$; $t = 4.467$; $p = 3.645 \times 10^{-5}$); its effect on distortion in polar requirement was not significant. As for optimal NaCl concentration, significant positive effects were encountered with regards to distortions in polar requirement ($\beta = 0.001$; $t = 2.479$; $p = 0.016$) and isoelectric point ($\beta = 0.001$; $t = 2.491$; $p = 0.016$).

These effects translate to a general decrease in the expected physicochemical fidelity both in thermophiles and halophiles. In other words, such extremophilic codon usages could decrease the chance of preserving the respective physicochemical patterns with an occurring mutation, diminishing their mutational robustness.

The effect of ambient pH remains less conclusive. Although there is a significant positive effect on distortion in polar requirements ($\beta = 0.006$; $t = 2.206$; $p = 0.031$), other properties are not shown to be significantly influenced. Evidence of substantial selection on proteins in extreme acidophiles or alkaliphiles is sparse. This may be attributed to the relative invariance of intracellular pH regardless of the ambient environment[50]. We note, however, that our sample provides only a narrow pH range, and bias in codon usage have been recently noted[16].

Model robustness to ti/tv-ratio. To verify the robustness of these effects, we extended our analysis to a wider range of ti/tv-ratios ($\kappa = 2.5$–10; Supplementary Figure S1 online). The quantitative coefficients of the tested factors change asymptotically and remain significant. The signs of the significant coefficients do not change. The only exception is the GC-content where its negative impact on isoelectric properties becomes non-significant

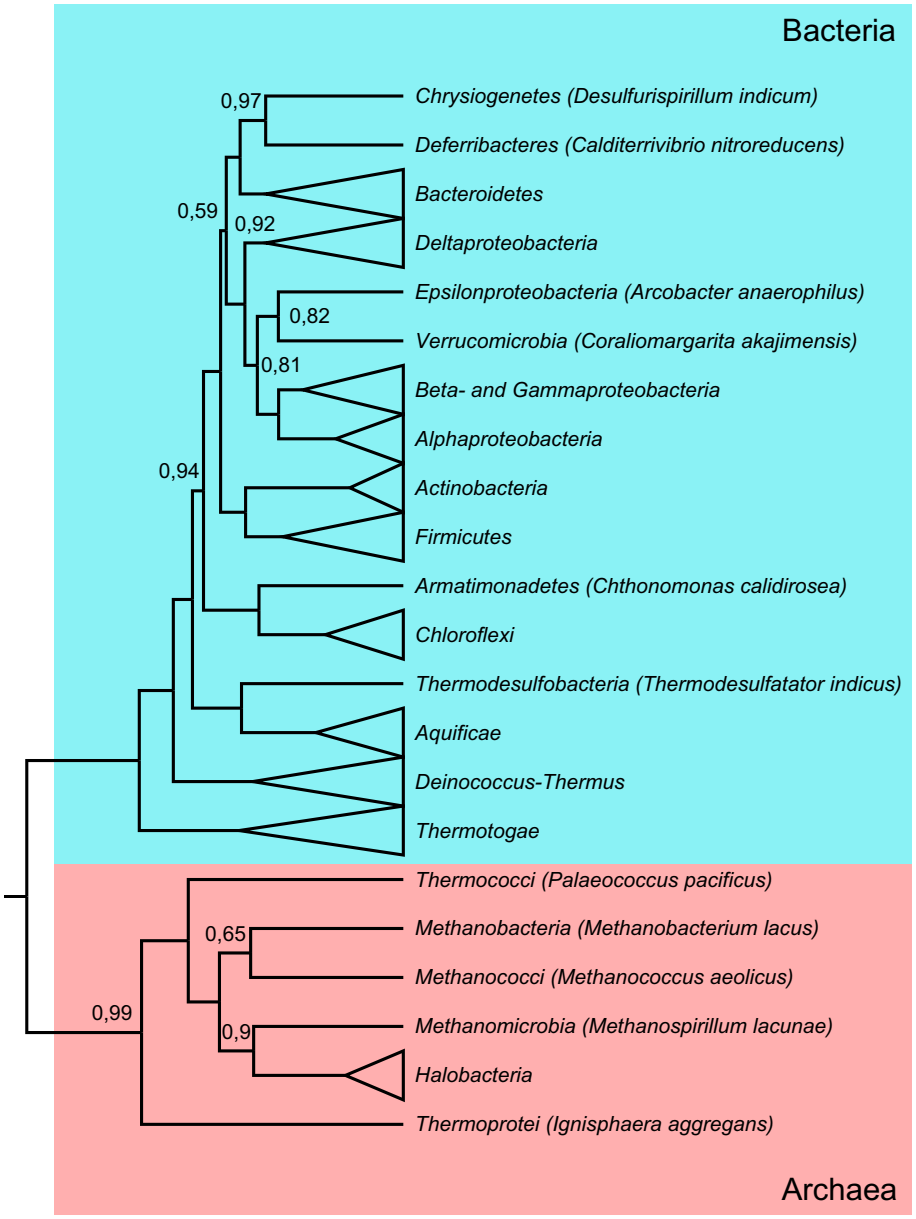

Figure 1. The large-scale hierarchy of the Bayesian phylogenetic tree based on S16 rRNAs sequences. The Bayesian tree was constructed using Beast v1.10.4[40]. The numbers at each node represent posterior probability values (only values < 1.00 are shown). Collapsed nodes contain multiple species (see Supplementary Data S1 online for more details).

($\kappa > 2.5$); therefore, only its positive effect on hydrophobic distortion is confirmed. Thus, we may conclude that the effects of environmental selection are robust on a broader range of ti/tv-ratios, and our interpretations concerning the impact of environmental selection, especially the distortive effect at high temperature, remains valid.

Discussion

The optimality of the genetic code should be discussed in the context of different gradients, such as environmental selection or GC-content. Their possible repercussions on codon distribution will fundamentally impact the expected errors made by the code. In order to study this question, we have applied a previously developed minimalistic background mutation model for the generalised mechanism of emerging point-mutations. Then, we calculated distortions for codon distributions encountered in different taxa with known environmental requirements. Distortion measures were based on different physicochemical properties: hydropathy, polar requirement, volume, and isoelectric point.

Next, in order to account for phylogenetic non-independence, we performed four distinct Phylogenetic Generalized Least Squares (PGLS) regressions on these physicochemical distortion measures using a 16S rRNA tree (Fig. 1). One of the putative predictors was the GC-composition of the coding region of the genome, based on a

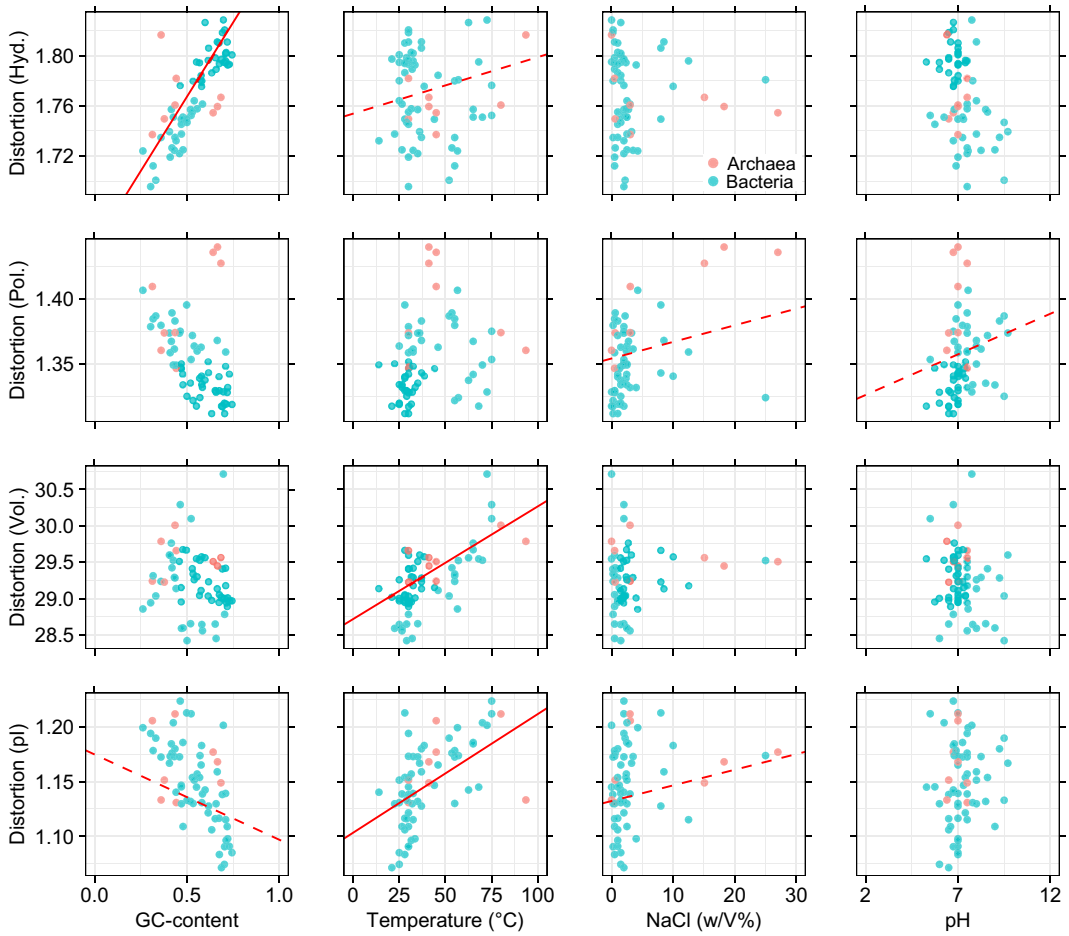

Figure 2. The effect of GC-content and environmental variables (temperature, salt concentration and pH) on the expected distortion measures ($\kappa = 2.5$) of different amino acid physicochemical properties. The effects were estimated by PGLS models using the inferred phylogenetic relationships shown on Fig. 1. Partial regression lines show significant trends (continuous line: $p < 0.001$, dashed line: $p < 0.05$). In each case, the remaining independent variables were set as [GC = 0.5; T = 30 °C; $cc_{NaCl} = 2.5$ w/V%; pH = 7].

large number of studies supporting its predominant effect on the amino acid composition of proteomes[26,51]. The other group of predictors included environmental variables, which can select for characteristic codon or amino acid compositions of proteomes: temperature, salt concentration, and pH optima.

The key insight provided in this study is that adaptations to certain extreme environments and GC-bias seem to have drastic effects on the physicochemical fidelity of translation and the severity of incidental mutations, and these putative effects are not the artefacts of phylogenetic bias.

While we have included only a limited number of taxa based on the availability of their proteomes and the environment they live in, they can be considered representative. Any phylogeny is constrained by what we know at the moment about the diversity of life on Earth. Current diversity is only a subset of the diversity that ever existed (which we need to keep in mind when we want to infer past events based on characteristics of current species), and metagenomics has repeatedly demonstrated that we know only a fraction of the current diversity. Metagenome studies discovered a previously unknown diversity of microbes[52,53]. Quite some of the microbial dark matter[54] were first assigned to novel clades distinct from the established great groups of Bacteria and Archaea. Nowadays it seems, that the truly novel clades are fewer[54,55]. A recent catalogue of microbes, incorporating more than 50 thousand new metagenome-assembled genomes, concluded that the majority of deep-branching lineages (lineages that would be represented on the level of phylum) are represented by current genome sequences[56]; the Candidate Phyla Radiation[57] could be merged into one monophyletic phylum[58]. Consequently, even a limited sample of taxa from all great branches of microbes can be considered representative.

Substitution biases might be caused by aversions of certain physicochemical distortions. We have shown that high GC-content is expected to increase hydropathic distortion. Only this effect of nucleotide composition is robust to ti/tv-ratio, which would predict an AT-bias for the highest fidelity in hydrophobic attributes, pointing towards a general substitution trend that resembles preliminary observations of AT-biased substitution patterns[30–32]. Hydrophobic patterns are regarded as a primary force of protein folding[59–62], further

supported by the fact that secondary structures can be described and predicted along these properties[63]. Therefore, A/T-mutations should be more likely to fix due to their lower risk of jeopardizing the structure.

Environmental selection influences the mutational robustness of the genetic code. With regards to environmental gradients, we show that selection in halophiles and thermophiles results in codon usage profiles generally worse for maintaining physicochemical patterns. Despite the observed adjustments against such mutations[16] the conservation of the required physicochemical properties tends to be more unreliable and the average effects of incidental mutations are expected to be more severe and deleterious. Here, we have demonstrated reduced fidelity in hydrophobic, volumetric, and isoelectric patterns of thermophiles, as well as the vulnerability of halophiles to mutations disturbing polar and isoelectric arrangements in proteins. This is a possible explanation of why these extremophiles possess remarkably low mutation and substitution rates[64–68]: it is a straightforward result of avoiding harsh fitness costs, especially if we consider higher importance of hydrophobic interactions and salt bridges in thermophilic proteins[18,69], as well as the central role of polar properties in halophiles[17,70].

We conclude that such low mutation rates and strong selection patterns encountered in extremophiles are caused by the inefficiency of the genetic code, as it seems especially ill-suited for extremophile codon usages with regards to mutational robustness, which also means that evolvability diminishes with the employment of standard genetic code, casting doubts on what role extreme environments could have played at development of the codon mapping.

Implications of the mesophile optimality of the genetic code and codon usage. Along with our estimated phylogeny, the majority of influential phylogenies have also provided an intuitive evidence of an extremophile LUCA, by placing thermophiles as the most basal groups[71–74]. This observation has long facilitated the somewhat overreaching logic that the cradle of life, including the development of the genetic code, was always associated with "infernal" environments of the Hadean Earth. At the same time, the genetic code is usually considered as a near-optimal, robust mapping that is able to partially maximize the fidelity of translation, since the majority, but not all[10,11], of alternative codes falls short of such error capacity[6,7].

Having these facts put together, our study implies a contradiction between these two notions: The fidelity of the genetic code is expected to decrease with higher temperatures. This not only collides with earlier reports supporting the extremophilic nature of the code[75–77], but also points out that claims between its error-minimization and thermophilic origin seem non-compatible: if *"(and oh what a big if)"*[78] the genetic code evolved in order to be optimal for physicochemical properties, then it is more likely to finish its emergence among milder conditions.

A mesophile optimality of the genetic code could be still compatible with phylogenies placing thermophiles at basal locations near LUCA. The evolution of the genetic code preceded LUCA. A well-thought-out rooting of the tree of life puts the bacterial clade *Chloroflexi* closest to the root[55], and it has thermophilic members. *Chloroflexi* are photosynthetic bacteria, meaning that LUCA was an autotroph. But, the first cell was, by necessity, heterotrophic, i.e. dependent on the environment for organic building blocks; the fully fledged photosynthesis can evolve only later. This means that inferences about LUCA does not help us understand the environment in which the first cell or the organism inventing the standard genetic code thrived.

But there is no need to accept a thermophilic LUCA. Both rRNA and protein sequences indicate that hyperthermophilic features of Bacteria and Archaea are parallel adaptations, while their ancestors could have been mesophilic or only slightly thermophilic[79–81]. Furthermore, the idea of a thermophile LUCA comes from accepting the root of the universal tree of life to lay between Bacteria and Archaea. Cavalier-Smith has argued for quite some time, that the root lies within Gram negative bacteria, and Archaea and Eukaryotes (compromising the clade Neomura) are derived from Gram positive bacteria[82,83]. Archaea are the exemplars of extremophiles, but if their extremophilic characteristic is derived[84], then there is no need for LUCA to be an extremophile. Indeed, among the *Chloroflexi*, there are mesophilic members, so even that does not contradict the proposition. Our own results presented here also strengthen the view that LUCA was a mesophile[85].

Outlook. Our work supports the expanded view on the optimality of the genetic code by involving codon usage[13,14]. The "mesophilic genetic code" hypothesis demands further research. Psychrophiles and organisms preferring extreme pH conditions could not be sufficiently represented in our current analysis, and albeit linear responses had a good fit in our case, increasing the environmental ranges should lead to more accurate estimations of translational preference. It must be also emphasized that our interpretation of codon usage analysis remains conditional on the assumption of error minimization. It does not weaken the case of other dominant theories of the evolution of the genetic code. The evolution of genetic code is likely to be a result of multiple driving forces and cannot be understood solely by natural selection increasing mutational robustness[86].

Relying on simple codon usage data have disadvantages. Thermophiles and halophiles possess elevated rates of horizontal gene transfer[87–89]. The codon bias of recently acquired, not completely adapted genes originally hosted by non-extremophilic hosts can interfere with the analysis, thus a later focus on core genes is needed. Upcoming studies are also yet to address the effect of mRNA expression patterns. Here, the implicit assumption is that proteins are expressed on the same level in each proteome. However, difference in expression has clear evolutional implications[90] that are also related to thermophilic properties[91] and misfold chance[92]. We believe that these observations can be incorporated into theory.

On the other hand, our result should not be taken as a direct confirmation of the universal genetic code being an inefficient mapping at extreme conditions. To assess that point better, it would demand us to gather data about the robustness of alternative codes and the proportion of variants where codon usage response to extreme environments can increase fidelity.

There is another point where biological codon usage data fails to elaborate. Simply extending the measure of distortion to the domain of alternative codes poses a paramount challenge. The currently known codon usage profiles cover only the standard genetic code (and alternative genetic code variants to some extent[14]). As codon frequencies and their environmental response already carry the inherent effect of the standard genetic code, their variance cannot be directly applied to randomized codes. This warrants further investigation into the physicochemical requirements of extreme habitat conditions as the likely causes of characteristic amino acid compositions that influence codon distribution.

Data availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

Received: 4 March 2021; Accepted: 10 May 2021
Published online: 26 May 2021

References

1. Preiner, M. *et al.* The future of origin of life research: bridging decades-old divisions. *Life* **10**, 20 (2020).
2. Yarus, M. The genetic code and RNA-amino acid affinities. *Life* **7**, 13 (2017).
3. Woese, C. R., Dugre, D. H., Dugre, S. A., Kondo, M. & Saxinger, W. C. On the fundamental nature and evolution of the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* **31**, 723–736 (1966).
4. Wong, J., Ng, S.-K., Mat, W.-K., Hu, T. & Xue, H. Coevolution theory of the genetic code at age forty: pathway to translation and synthetic life. *Life* **6**, 12 (2016).
5. Tze-Fei Wong, J. A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. USA* **72**, 1909–1912 (1975).
6. Freeland, S. J. & Hurst, L. D. The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248 (1998).
7. Haig, D. & Hurst, L. D. A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* **33**, 412–417 (1991).
8. Koonin, E. V. & Novozhilov, A. S. Origin and evolution of the universal genetic code. *Annu. Rev. Genet.* **51**, 45–62 (2017).
9. Kun, Á. & Radványi, Á. The evolution of the genetic code: Impasses and challenges. *BioSystems* **164**, 217–225 (2018).
10. Błażej, P., Wnętrzak, M., Mackiewicz, D., Gagat, P. & Mackiewicz, P. Many alternative and theoretical genetic codes are more robust to amino acid replacements than the standard genetic code. *J. Theor. Biol.* **464**, 21–32 (2019).
11. Wong, J. T. Role of minimization of chemical distances between amino acids in the evolution of the genetic code. *Proc. Natl. Acad. Sci.* **77**, 1083–1086 (1980).
12. Di Giulio, M. The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *J. Mol. Evol.* **29**, 288–293 (1989).
13. Zhu, C.-T., Zeng, X.-B. & Huang, W.-D. Codon usage decreases the error minimization within the genetic code. *J. Mol. Evol.* **57**, 533–537 (2003).
14. Wnetrzak, M., Błażej, P. & Mackiewicz, P. Properties of the Standard Genetic Code and Its Alternatives Measured by Codon Usage from Corresponding Genomes. In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies* 44–51 (SCITEPRESS - Science and Technology Publications, 2020). https://doi.org/10.5220/0008981000440051.
15. Goodarzi, H., Torabi, N., Najafabadi, H. S. & Archetti, M. Amino acid and codon usage profiles: adaptive changes in the frequency of amino acids and codons. *Gene* **407**, 30–41 (2008).
16. Khan, M. F. & Patra, S. Deciphering the rationale behind specific codon usage pattern in extremophiles. *Sci. Rep.* **8**, 15548 (2018).
17. Paul, S., Bag, S. K., Das, S., Harvill, E. T. & Dutta, C. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol.* **9**, R70 (2008).
18. Haney, P. J. *et al.* Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. *Proc. Natl. Acad. Sci.* **96**, 3578–3583 (1999).
19. Fukuchi, S., Yoshimune, K., Wakayama, M., Moriguchi, M. & Nishikawa, K. Unique amino acid composition of proteins in halophilic bacteria. *J. Mol. Biol.* **327**, 347–357 (2003).
20. Kreil, D. P. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucl. Acids Res.* **29**, 1608–1615 (2001).
21. Tekaia, F. & Yeramian, E. Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. *BMC Genom.* **7**, 307 (2006).
22. Singer, G. A. C. & Hickey, D. A. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317**, 39–47 (2003).
23. Madern, D., Ebel, C. & Zaccai, G. Halophilic adaptation of enzymes. *Extremophiles* **4**, 91–98 (2000).
24. Wright, D. B., Banks, D. D., Lohman, J. R., Hilsenbeck, J. L. & Gloss, L. M. The effect of salts on the activity and stability of *Escherichia coli* and *Haloferax volcanii* dihydrofolate reductases. *J. Mol. Biol.* **323**, 327–344 (2002).
25. Fukuchi, S. & Nishikawa, K. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *J. Mol. Biol.* **309**, 835–843 (2001).
26. Knight, R. D., Freeland, S. J. & Landweber, L. F. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **2**, RESEARCH0010 (2001).
27. Goncearenco, A. & Berezovsky, I. N. The fundamental tradeoff in genomes and proteomes of prokaryotes established by the genetic code, codon entropy, and physics of nucleic acids and proteins. *Biol. Direct* **9**, 29 (2014).
28. Foerstner, K. U., von Mering, C., Hooper, S. D. & Bork, P. Environments shape the nucleotide composition of genomes. *EMBO Rep.* **6**, 1208–1213 (2005).
29. Mann, S. & Chen, Y.-P.P. Bacterial genomic G+C composition-eliciting environmental adaptation. *Genomics* **95**, 7–15 (2010).
30. Hershberg, R. & Petrov, D. A. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* **6**, e1001115 (2010).
31. Lynch, M. *et al.* A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci.* **105**, 9272–9277 (2008).
32. Albu, M., Min, X. J., Golding, G. B. & Hickey, D. Nucleotide substitution bias within the genus *Drosophila* affects the pattern of proteome evolution. *Genome Biol. Evol.* **1**, 288–293 (2009).
33. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
34. Koonin, E. V. & Novozhilov, A. S. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**, 99–111 (2009).
35. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucl. Acids Res.* **46**, 2699–2699 (2018).
36. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020** (2020).
37. Reimer, L. C. *et al.* BacDive in 2019: bacterial phenotypic data for high-throughput biodiversity analysis. *Nucl. Acids Res.* **47**, D631–D636 (2019).
38. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucl. Acids Res.* **44**, D733–D745 (2016).

39. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32**, 1792–1797 (2004).
40. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, 16 (2018).
41. Tavare, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Some mathematical questions in biology/DNA sequence analysis edited by Robert M. Miura* (1986).
42. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
43. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).
44. Wang, J., Raskin, L., Samuels, D. C., Shyr, Y. & Guo, Y. Genome measures used for quality control are dependent on gene function and ancestry. *Bioinformatics* **31**, 318–323 (2015).
45. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
46. Revell, L. J. Phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
47. Orme, D. *et al.* caper: Comparative Analyses of Phylogenetics and Evolution in R. (2018).
48. Pennell, M. W. *et al.* Geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* **30**, 2216–2218 (2014).
49. R Core Team. R: A Language and Environment for Statistical Computing. (2018).
50. Horikoshi, K. Alkaliphiles: some applications of their products for biotechnology. *Microbiol. Mol. Biol. Rev.* **63**, 735–750 (1999).
51. Li, J., Zhou, J., Wu, Y., Yang, S. & Tian, D. GC-content of synonymous codons profoundly influences amino acid usage. *G3* **5**, 2027–2036 (2015).
52. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
53. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain bacteria. *Nature* **523**, 208–211 (2015).
54. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
55. Cavalier-Smith, T. & Chao, E.E.-Y. Multidomain ribosomal protein trees and the planctobacterial origin of neomura (eukaryotes, archaebacteria). *Protoplasma* **257**, 621–753 (2020).
56. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0718-6 (2020).
57. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
58. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
59. Dyson, H. J., Wright, P. E. & Scheraga, H. A. The role of hydrophobic interactions in initiation and propagation of protein folding. *Proc. Natl. Acad. Sci.* **103**, 13057–13061 (2006).
60. Baldwin, R. L. & Rose, G. D. How the hydrophobic factor drives protein folding. *Proc. Natl. Acad. Sci.* **113**, 12462–12466 (2016).
61. Kauzmann, W. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**, 1–63 (1959).
62. Miller, S., Janin, J., Lesk, A. M. & Chothia, C. Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656 (1987).
63. Cid, H., Bunster, M., Arriagada, E. & Campos, M. Prediction of secondary structure of proteins by means of hydrophobicity profiles. *FEBS Lett.* **150**, 247–254 (1982).
64. Drake, J. W. Avoiding dangerous missense: Thermophiles display especially low mutation rates. *PLoS Genet.* **5**, e1000520 (2009).
65. Friedman, R., Drake, J. W. & Hughes, A. L. Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. *Genetics* **167**, 1507–1512 (2004).
66. Groussin, M. & Gouy, M. Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in Archaea. *Mol. Biol. Evol.* **28**, 2661–2674 (2011).
67. Mackwan, R. R., Carver, G. T., Drake, J. W. & Grogan, D. W. An unusual pattern of spontaneous mutations recovered in the halophilic archaeon *Haloferax volcanii*. *Genetics* **176**, 697–702 (2007).
68. Busch, C. R. & DiRuggiero, J. MutS and MutL are dispensable for maintenance of the genomic mutation rate in the halophilic archaeon *Halobacterium salinarum* NRC-1. *PLoS ONE* **5**, e9045 (2010).
69. Lee, C.-W., Wang, H.-J., Hwang, J.-K. & Tseng, C.-P. Protein thermal stability enhancement by designing salt bridges: a combined computational and experimental study. *PLoS ONE* **9**, e112751 (2014).
70. Kastritis, P. L., Papandreou, N. C. & Hamodrakas, S. J. Haloadaptation: Insights from comparative modeling studies of halophilic archaeal DHFRs. *Int. J. Biol. Macromol.* **41**, 447–453 (2007).
71. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* **87**, 4576–4579 (1990).
72. Stetter, K. O. Hyperthermophilic procaryotes. *FEMS Microbiol. Rev.* **18**, 149–158 (1996).
73. Gaucher, E. A., Kratzer, J. T. & Randall, R. N. Deep phylogeny—how a tree can help characterize early life on Earth. *Cold Spring Harb. Perspect. Biol.* **2**, a002238 (2010).
74. Weiss, M. C. *et al.* The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 16116 (2016).
75. Di Giulio, M. The late stage of genetic code structuring took place at a high temperature. *Gene* **261**, 189–195 (2000).
76. Di Giulio, M. Structuring of the genetic code took place at acidic pH. *J. Theor. Biol.* **237**, 219–226 (2005).
77. Di Giulio, M. The ocean abysses witnessed the origin of the genetic code. *Gene* **346**, 7–12 (2005).
78. Darwin, C. Letter to Joseph Dalton Hooker. (1871).
79. Boussau, B., Blanquart, S., Necsulea, A., Lartillot, N. & Gouy, M. Parallel adaptations to high temperatures in the Archaean eon. *Nature* **456**, 942–945 (2008).
80. Brochier, C. & Philippe, H. A non-hyperthermophilic ancestor for Bacteria. *Nature* **417**, 244 (2002).
81. Galtier, N. A non-hyperthermophilic common ancestor to extant life forms. *Science (80-)* **283**, 220–221 (1999).
82. Cavalier-Smith, T. The Neomuran revolution and phagotrophic origin of eukaryotes and cilia in the light of intracellular coevolution and a revised tree of life. *Cold Spring Harb. Perspect. Biol.* **6**, a016006 (2014).
83. Cavalier-Smith, T. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int. J. Syst. Evol. Microbiol.* **52**, 7–76 (2002).
84. Forterre, P. Thermoreduction, a hypothesis for the origin of prokaryotes. *C. R. Acad. Sci. III* **318**, 415–422 (1995).
85. Forterre, P. The universal tree of life: an update. *Front. Microbiol.* **6**, 717 (2015).
86. Bandhu, A. V., Aggarwal, N. & Sengupta, S. Revisiting the physico-chemical hypothesis of code origin: an analysis based on code-sequence coevolution in a finite population. *Orignal Life Evol. Biosph.* **43**, 465–489 (2013).
87. Aravind, L., Tatusov, R. L., Wolf, Y. I., Walker, D. R. & Koonin, E. V. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* **14**, 442–444 (1998).
88. Mongodin, E. F. *et al.* The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proc. Natl. Acad. Sci.* **102**, 18147–18152 (2005).
89. Rhodes, M. E., Spear, J. R., Oren, A. & House, C. H. Differences in lateral gene transfer in hypersaline versus thermal environments. *BMC Evol. Biol.* **11**, 199 (2011).
90. Pál, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927–931 (2001).

91. Cherry, J. L. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol. Biol. Evol.* **27**, 735–741 (2010).
92. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci.* **102**, 14338–14343 (2005).

Acknowledgements

This work was supported by the National Research, Development and Innovation Office (NKFIH) under the Grant Numbers GINOP-2.3.2-15-2016-00057 and K119347; and by the Volkswagen Stiftung initiative "Leben? – Ein neuer Blick der Naturwissenschaften auf die grundlegenden Prinzipien des Lebens" under project "A unified model of recombination in life".

Author contributions

Á.R. designed the project, performed the simulations, assembled and analysed the data. Á.R. and Á.K. wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-90440-y.

Correspondence and requests for materials should be addressed to Á.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.