PHYSICAL REVIEW E **79**, 060901(R) (2009)

Extreme genetic code optimality from a molecular dynamics calculation of amino acid polar requirement

Thomas Butler and Nigel Goldenfeld
Department of Physics and Institute for Genomic Biology, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, Illinois 61801, USA

Damien Mathew and Zaida Luthey-Schulten
Department of Chemistry and Institute for Genomic Biology, Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, 600 S. Mathews Avenue, Urbana, Illinois 61801, USA

(Received 20 December 2007; revised manuscript received 21 April 2009; published 17 June 2009)

A molecular dynamics calculation of the amino acid polar requirement is used to score the canonical genetic code. Monte Carlo simulation shows that this computational polar requirement has been optimized by the canonical genetic code, an order of magnitude more than any previously known measure, effectively ruling out a vertical evolution dynamics. The sensitivity of the optimization to the precise metric used in code scoring is consistent with code evolution having proceeded through the communal dynamics of statistical proteins using horizontal gene transfer, as recently proposed. The extreme optimization of the genetic code therefore strongly supports the idea that the genetic code evolved from a communal state of life prior to the last universal common ancestor.

DOI: 10.1103/PhysRevE.79.060901 PACS number(s): 87.14.G−, 87.23.Kg

I. INTRODUCTION

The genetic code is one of life's most ancient and universal features [1,2]. It summarizes how RNA transcripts are translated into amino acids to form proteins and is shared by all known cells across the three domains of life with only a very few minor variations [3,4]. Almost immediately after its elucidation, attempts were made to explain the assignment of codons to amino acids. It was noticed that amino acids with related properties were grouped together, which would have the effect of minimizing translation errors [5–7]. In order to determine whether or not this was a genuine correlation or simply a fluctuation reflecting the limited size of the codon table, the canonical genetic code was compared to samples of randomly generated synthetic codes, starting with early but inconclusive Monte Carlo work of Alff-Steinberger [8], and compellingly revisited with larger sample sizes by Haig and Hurst [9]. Depending on the measure used to characterize or score the sampled codes, high degrees of optimality have been reported. For example, using an empirical measure of amino acid differences referred to below as the "experimental polar requirement" (EPR) [10,11], Freeland and Hurst [12] calculated that the genetic code is "one in a million" [9,13]. More recently, it has been shown that when coupled to known patterns of codon usage, the canonical code (and the codon usage) is simultaneously optimized with respect to point mutations and to the rapid termination of peptides that are generated with frame shift errors [14].

These results are generally interpreted to imply that the canonical genetic code had to have undergone a period of evolution and was not simply a frozen accident [15,16]. While it was long assumed that code evolution would be lethal, it has been recently shown how a genetic code can evolve along with a dynamic refinement of the precision of translation [17,18]. The results show clearly that vertically dominated evolution is only capable of a relatively weak degree of optimization, failing to find global extrema, and neither strongly optimized nor converged to a unique code. On the other hand, only if the evolutionary dynamics is horizontally dominated, with genes shared between organisms (as is the case with contemporary microbes [19]), modularity of structures such as the translation apparatus and the genome emerges naturally [20], and optimization is strong, rapid, and convergent to a universal genetic code [18]. Thus, the structure of the genetic code and translation apparatus reflects the evolutionary dynamics from which the code emerged. Although it is already clear that the code's known optimality strongly suggests that it did evolve, the dynamics which dominated during its evolution has not yet so clearly been determined. Thus, it is of essential interest to determine accurately the extent of optimality of the canonical genetic code because the greater the level of optimization the more likely it is that the genetic code evolved when life was communal in character.

The purpose of this Rapid Communication is to provide two pieces of evidence for the collective evolution of the genetic code. First, we set a lower bound on the level of optimality of the canonical genetic code by using molecular dynamics (MD) to construct a measure of code optimality, the "computational polar requirement" (CPR) without any input from experiment. We then use Monte Carlo simulation to determine the level of code optimality and find that the level is so high that a new and detailed error analysis is required to ensure statistically significant assessment of very small probabilities. Second, we explore the dependence of our results on the scale of code variations. Our results indicate a level of optimization that would only be attainable from some form of collective dynamics [18] and a dependence on scale that indicates that the dynamics involved the refinement over evolutionary time of an ambiguous primitive translation machinery. Ambiguous translation generates a statistical ensemble of related proteins ("statistical proteins")

[7,21] rather than a unique protein, as is now the case, and is exploited in the coevolutionary mechanism [17] of collective code evolution [18].

II. MOLECULAR DYNAMICS OF THE POLAR REQUIREMENT

The experimental polar requirement is a chromatographic measure of amino acid affinity to a water-pyridine solution that was originally motivated by a simple stereochemical theory of the origin of the genetic code [7,10,11,22]. This measure is related to and strongly correlated with several other amino acid measures, such as hydrophobicity and Grantham polarity [23]. In the EPR experiments, water/dimethylpyridine (DMP) ratios ranging from 40–80 % mole fraction water were used for chromatographic separations of each amino acid measured. When the chromatographic factor, R_m was plotted as a function of mole fraction water in log-log scale, a linear trend was observed for each amino acid. The slope of the corresponding best fit line was taken to be the amino acid's EPR.

The methods used for obtaining the CPR numbers are reported elsewhere [24] and are summarized here. The distribution of solute molecules across the water/DMP interface is related to the equilibrium solvent environment surrounding the molecules in a binary solution similar to that used in the experiments. Trends in the local water density of a solvated amino acid in water/DMP solutions were found to be linear functions of mole fraction water. The slopes of these linear trends were used to obtain a set of computed CPR values. To quantitatively measure the differences in local solvent environment, MD calculations were performed using NAMD2 software with a number, pressure, and temperature (*NPT*) ensemble [25] and the CHARMM 27 force field [26,27]. Standard pressure and temperature were maintained for the simulations. The systems consisted of a single amino acid molecule in a box of water and randomly placed DMP molecules of a determined water/DMP ratio. For each amino acid at least four systems, each with a different water/DMP ratio, were simulated. Radial distribution functions (RDFs) of water relative to the amino acid side chains were calculated from the equilibrated MD trajectories using visual molecular dynamics (VMD) [28]. The RDFs were calculated by a time average over the equilibrated portion of a trajectory [29].

The most distant atom of the amino acid side chain was used as a reference atom, and the oxygen or hydrogens (as appropriate) from the water molecules were used as a selection in calculating the RDFs. Calculated in this manner, the maximum value of the first peak in an RDF is related to the relative density of water in the first solvation layer of the amino acid side chain. It was found that these maxima varied linearly with water/DMP ratios for each amino acid and that the slopes of the corresponding lines were strongly correlated with the EPR (R^2=0.92) (Fig. 1). We confirmed that tyrosine's large deviation from the experimental value was not due to a weak signal in the RDF.

III. OPTIMALITY ANALYSIS OF THE CANONICAL GENETIC CODE

To analyze the CPR, we used the point mutation code analysis algorithms described in [9,12] along with an analytical realization of bootstrap error analysis to assess the statistical significance of the results. The algorithms treat the genetic code as a mapping GC^i:*Codons*$\rightarrow$*Amino Acids*, where i indexes a particular set of assignments of codons to amino acids and with GC^1 as the canonical code. *Codons* is the set of codons excepting the termination codons, and *Amino Acids* is the set of amino acids, i.e., $GC^1(UUU)$=*Phe*. New versions $GC^{i\neq 1}$ of the mapping are generated by randomly permuting amino acid labels, leaving termination codons fixed. This preserves the degeneracy structure of the genetic code. The optimality of a given realization of the genetic code GC^i is assessed by evaluating the sum

$$O_i^{-1} = \sum_{\langle c,c'\rangle \neq Ter} W_{c,c'} d^q[GC^i(c), GC^i(c')], \qquad (1)$$

where $\langle c,c'\rangle \neq Ter$ denotes a sum over the nearest-neighbor codons with the nearest neighbors of a codon defined by its single point mutations, with all mutations to or from a termination codon excluded. The matrix $W_{c,c'}$ weighs transition and transversion biases differently for different positions in the codon according to a toy model of typical transversion and transition biases in real translation. In our calculations, we used the values from [12] as listed in Table I. Finally, $d^q(x,y)$ is a metric on the space of amino acids. For the polar requirement, the metric is taken to be $d^q(x,y)=|x-y|^q$ over the polar requirement values corresponding to the given amino acids.

The appropriate quantity to compute is the probability $P_b=Pr(O>O_1)$ that a random realization is more optimal than the canonical code. To compute P_b, we count the number of randomly generated codes that are more optimal than the canonical code and divide by the total number of random codes generated. P_b is invariant to uniform linear rescaling of the amino acid polar requirement data and is smaller for more optimal codes while including the effects of the large

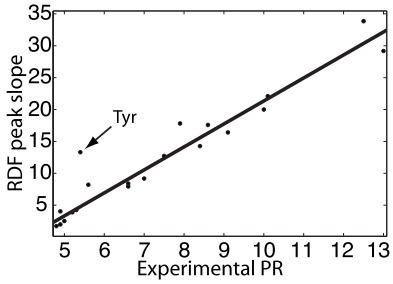

FIG. 1. Scatter plot showing the relationship between RDF peak slope and experimental polar requirement for all amino acids. The straight line is a guide to the eyes.

TABLE I. The matrix $W_{c,c'}$ of transition and transversion biases taken from [12].

	First base	Second base	Third base
Transitions	1	0.5	1
Transversions	0.5	0.1	1

number of codes that can be explored rather than the simple linear scale provided by the bare optimality score.

The error in the computed P_b can be estimated using an analytical realization of bootstrap resampling. Simulated data sets for bootstrap are created by randomly sampling optimality scores from the original data set. When the samples are drawn from the original set, there are only two alternatives: a more or less optimal code can be sampled with the probability $P_b = N_{O>O_1}/N_{total}$ of drawing a random code better than the canonical code. Since the number of better codes in a sample is the number whose error we wish to estimate, we can regard drawing a better code as a step to the right with probability P_b in a one-dimensional random walk. The known formulas for the asymmetric one-dimensional random walk allow us to compute the bootstrap error estimate in the limit of infinitely many resampled sets, i.e., the exact bootstrap estimate. For metrics under which $P_b \ll 1$ holds, we obtain the variance in P_b to be

$$var[P_b] = var\left[\frac{N_{O>O_1}}{N_{total}}\right] = \frac{P_b(1-P_b)}{N_{total}} \approx \frac{N_{O>O_1}}{N_{total}^2}. \quad (2)$$

To obtain a reasonable estimate of error or to compare the results of different metrics on the space of amino acids, the number of more optimal codes $N_{O>O_1}$ from the random sample must be sufficiently large ($\sqrt{N_{O>O_1}} \ll N_{O>O_1}$ or about $N_{O>O_1} = 10$ as a reasonable minimum).

When the computational polar requirement difference squared is used in the amino acid metric $P_b = (19 \pm 4.36) \times 10^{-8}$. In contrast, with the experimental polar requirement, $P_b = (26.5 \pm 1.63) \times 10^{-7}$, an order of magnitude improvement. To assess the impact of tyrosine (which had the largest variation between the CPR and EPR values) on these results, we redid the calculation of P_b for the CPR but with tyrosine replaced with the value from the EPR. The result is $[P_b = (9.3 \pm 1.0) \times 10^{-7}]$. To test the sensitivity of the results for the CPR, we varied each element of $W_{c,c'}$ independently by $\pm 0.1 \times W_{c,c'}$ and repeated the calculation of P_b. This led to the results that were statistically indistinguishable from the results reported above. Shorter computations (justified by the faster convergence due to decreased optimality) for the EPR indicate a similar level of robustness. With a $W_{c,c'}$ uniform among nearest neighbors we saw substantial increases in P_b in agreement with [9]. However, the CPR continued to be superior to the EPR, with the CPR yielding $P_b = (3.7 \pm .61) \times 10^{-5}$ and the EPR yielding $P_b = (11.8 \pm 1.1) \times 10^{-5}$.

Varying the value of q in the metric [30] provides a further probe to explore the optimization of the genetic code. Increasing the value of q is equivalent to emphasizing the role of larger and larger differences between the amino acid intended and the one generated by point mutation. Thus, if P_b reduces for increased values of q, the code (along with $W_{c,c'}$) evolved to suppress the effects of rarer, possibly catastrophic errors that may be generated by point mutations. This may happen primarily by evolving small elements of $W_{c,c'}$, where $c \to c'$ is catastrophic or vice versa. Conversely, if P_b reduces for smaller values of q, the code evolved to both mitigate the possibility of these catastrophic errors and to minimize the effects of frequent small errors. Varying q we find that the canonical genetic code is most optimal for q between one and two with significant increases outside this regime in either direction (Fig. 2). This indicates that the genetic code is optimized for minimizing errors according to their size with no undue emphasis to larger or smaller errors. Given the relative weakness of the code when emphasizing large errors, the evolution must have favored organisms that discarded or edited fatally flawed proteins over evolving the code to make them less likely at the cost of reducing its ability to minimize the more frequent moderate and minor errors. The weakness of the canonical code when minor errors are emphasized ($q < 1$) suggests that while the code was still evolving minor errors were on the whole less important biologically as would be expected in evolutionary dynamics [17,18] that utilized ambiguity tolerance in early proteins [7,21].

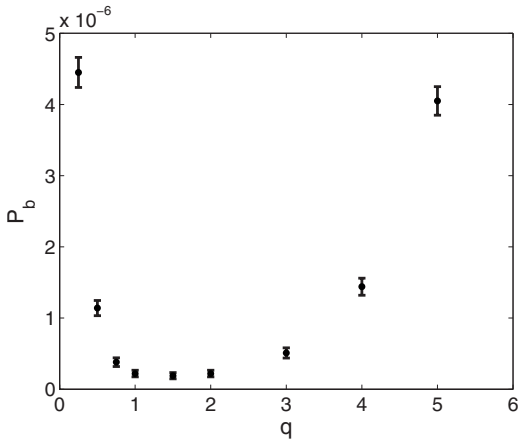

FIG. 2. P_b as a function of the exponent q in the amino acid metric.

IV. OPTIMALITY ANALYSIS OF ALTERNATIVE CODES AND MEASURES

A selection of variant codes was also analyzed using the CPR. Our findings, displayed in Table II, were consistent with the previous findings of Knight [13] in that the alternative codes did not show marked improvements in optimality over the canonical code. This is consistent with our expectation that evolutionary pressure to optimize the code with respect to the polar requirement was eased after the last universal ancestral state.

TABLE II. P_b for several naturally occurring variant codes.

Code	P_b
Canonical	$(19 \pm 4.36) \times 10^{-8}$
Yeast mitochondrial	$(11 \pm 3.32) \times 10^{-8}$
CDH nuclear code	$(21 \pm 4.58) \times 10^{-8}$
Ascidian mitochondrial	$(583 \pm 24.15) \times 10^{-8}$
Echinoderm mitochondrial	$(51 \pm 7.14) \times 10^{-8}$

We also tested Grantham polarity [23], which has been argued in a survey of genetic code optimality under different amino acid measures to be the amino acid measure most optimized by the genetic code [13]. The results yield $P_b=(285\pm16.88)\times10^{-8}$ or an order of magnitude higher than with the CPR metric, leading to the conclusion that the CPR is the most effective known metric for optimization of the genetic code. Previous computations evaluated P_b by generating 100 000 random codes [13]. Scaling our results to the size of these original simulations, we see that the EPR and the Grantham polarity have virtually identical scores. Scaling the errors for the CPR and the Grantham polarity to errors assessed from only 100 000 codes, we get for the CPR, $P_b=(0.19\pm0.44)\times10^{-5}$ and for the Grantham polarity, $P_b=(2.85\pm1.69)\times10^{-5}$. These results are within a standard deviation and a half of each other and are therefore not different in a statistically meaningful way.

In conclusion, earlier estimates of code optimality were understated by a statistically significant amount. The extent of optimality and its dependence on metric revealed here further support the notion that the genetic code must have evolved during an early communal state of life [18].

ACKNOWLEDGMENTS

We gratefully acknowledge discussions with Carl Woese and thank the referees for helpful suggestions that improved the Rapid Communication. This work was based upon work supported by the National Science Foundation under Grant No. NSF-EF-0526747.

[1] M. Niremberg *et al.*, Cold Spring Harb Symp. Quant Biol. **28**, 549 (1963).
[2] C. R. Woese, *The Genetic Code* (Harper and Row, New York, 1967).
[3] S. Osawa, *Evolution of the Genetic Code* (Oxford University Press, Oxford, 1995).
[4] R. Knight, S. Freeland, and L. Landweber, Nat. Rev. Genet. **2**, 49 (2001).
[5] E. Zuckerkandl and L. Pauling, in *Evolving Genes and Proteins*, edited by V. Bryson and H. J. Vogel (Academic Press, New York, 1965), pp. 97–166.
[6] T. Sonneborn, in *Evolving Genes and Proteins*, edited by V. Bryson and H. J. Vogel (Academic Press, New York, 1965), pp. 277–297.
[7] C. R. Woese, Proc. Natl. Acad. Sci. U.S.A. **54**, 1546 (1965).
[8] C. Alff-Steinberger, Proc. Natl. Acad. Sci. U.S.A. **64**, 584 (1969).
[9] D. Haig and L. D. Hurst, J. Mol. Evol. **33**, 412 (1991).
[10] C. R. Woese, D. H. Dugre, W. C. Saxinger, and S. A. Dugre, Proc. Natl. Acad. Sci. U.S.A. **55**, 966 (1966).
[11] C. R. Woese, D. H. Dugre, S. A. Dugre, M. Kondo, and W. C. Saxinger, Cold Spring Harb Symp. Quant Biol. **31**, 723 (1966).
[12] S. J. Freeland and L. D. Hurst, J. Mol. Evol. **47**, 238 (1998).
[13] R. D. Knight, Ph.D. thesis, Princeton University, 2001.
[14] S. Itzkovitz and U. Alon, Genome Res. **17**, 405 (2007).
[15] F. Crick, J. Mol. Biol. **38**, 367 (1968).
[16] G. Sella and D. Ardell, J. Mol. Evol. **63**, 297 (2006).
[17] D. Ardell and G. Sella, Philos. Trans. R. Soc. London, Ser. B **357**, 1625 (2002).
[18] K. Vetsigian, C. Woese, and N. Goldenfeld, Proc. Natl. Acad. Sci. U.S.A. **103**, 10696 (2006).
[19] H. Ochman, J. Lawrence, and E. Groisman, Nature (London) **405**, 299 (2000).
[20] J. Sun and M. W. Deem, Phys. Rev. Lett. **99**, 228107 (2007).
[21] C. Woese, Naturwiss. **60**, 447 (1973).
[22] C. R. Woese, Proc. Natl. Acad. Sci. U.S.A. **54**, 71 (1965).
[23] R. Grantham, Science **185**, 862 (1974).
[24] D. C. Mathew and Z. Luthey-Schulten, J. Mol. Evol. **66**, 519 (2008).
[25] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, J. Comput. Chem. **26**, 1781 (2005).
[26] A. D. MacKerell, Jr. and N. Banavali, J. Comput. Chem. **21**, 105 (2000).
[27] A. D. MacKerell *et al.*, J. Phys. Chem. B **102**, 3586 (1998).
[28] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[29] M. Allen and D. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1987).
[30] S. J. Freeland, R. D. Knight, L. F. Landweber, and L. D. Hurst, Mol. Biol. Evol. **17**, 511 (2000).