# Optimality properties of a proposed precursor to the genetic code

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*

(Received 18 May 2009; revised manuscript received 24 July 2009; published 30 September 2009)

We calculate the optimality score of a doublet precursor to the canonical genetic code with respect to mitigating the effects of point mutations and compare our results to corresponding ones for the canonical genetic code. We find that the proposed precursor is much less optimal than that of the canonical code. Our results render unlikely the notion that the doublet precursor was an intermediate state in the evolution of the canonical genetic code. These findings support the notion that code optimality reflects evolutionary dynamics, and that if such a doublet code originally had a biochemical significance, it arose before the emergence of translation.

DOI: 10.1103/PhysRevE.80.032901 PACS number(s): 82.39.Pj, 87.23.Kg, 87.10.Rt

It is now well-established that the canonical genetic code is not a frozen accident, but exhibits a pattern of amino acid-codon correspondences that has the effect of making the code insensitive to certain classes of point mutation or translation error [1–7]. A variety of schemes [8], including ones invoking evolutionary dynamics [9] and stereochemistry [10,11], have been put forward to explain this pattern and others [12] in the genetic code (for recent reviews, see [13,14]).Additionally, it has been shown recently that the genetic code has extreme error-minimizing optimality, being more optimal (resistant to the effects of point mutations) than all but one or two random codes generated in sets of ten million [7]. It is important to stress that while the code exhibits some optimality with respect to several measures, such as hydrophobicity [4], the code exhibits extreme optimality with respect to only one particular class of amino acid attributes, related to the free amino acid polar requirement [15,16], and this suggests the code is a very ancient part of the cell's machinery, functioning either in its present role of translation, or in some earlier unknown function. This result lends strong support to the suggestion that the code's evolutionary dynamics was dominated by collective mechanisms arising from horizontal gene transfer [9]. Computational evidence shows that core chemical affinities in the genetic code are fully compatible with, and independent from, evolutionary dynamics that lead to error minimizing optimality [17], suggesting that error-minimizing optimality is not a by-product of chemistry but arises from the evolutionary dynamics.

In this Brief Report, we attempt to ascertain to what extent, if any, error-minimizing optimality can be used to constrain a proposed scenario for the evolution of the genetic code. If the optimality with respect to polar requirement was a feature of the code from very early times, then precursor code proposals must respect error-minimizing optimality to a significant degree. Alternatively, proposed precursor codes may claim to date prior to any code evolution, and to be the product of other factors alone. Such precursors would not be expected to display a significant level of error-minimizing optimality, assuming that it is indeed the case that optimality is primarily a reflection of evolutionary dynamics. Here we show that a specific biochemically motivated precursor code does not show evidence for significant error-minimizing optimality, even though it is a projection of the canonical code; these results support the notion that error-minimizing optimality primarily reflects evolutionary dynamics, and imply that this type of precursor code, if it ever existed, would have arisen prior to the emergence of translation.

Copley *et al.* suggested that first and, to a lesser extent, second base assignments in the canonical code would arise if the code has its origin in amino acid synthesis channels embedded in dinucleotide complexes prior to the emergence of translation [18]. The proposal exploits the strong constraints such a theory imposes on the first two bases of the genetic code to generate a specific precursor doublet code based on a projection of the canonical genetic code to a doublet code. For most of the projection, the third codon is sufficiently redundant that the first two bases are sufficient to define the amino acid coded for by doublet. In the event that the third bases associated with a doublet codon code for multiple amino acids, the proposal favors the simpler of the amino acids (Table I). They further refine the proposal by incorporating possible precursor amino acids motivated by their study of the biosynthetic pathways for amino acids (not shown) [18].

To further assess and characterize the proposed precursor code in [18], we analyze the degree to which it contains error-minimizing optimality. As noted above, the proposed precursor code is based primarily on arguments about biosynthetic pathways rather than evolutionary considerations. Additionally, it explicitly dates to prior to translation [18]. All mechanisms of which we are aware for code evolvability explicitly require translation machinery (see, for example,

TABLE I. Proposed precursor code from Ref. [18]. Row is first base, column is second base.

| Proposed Precursor Code | | | | |
|---|---|---|---|---|
| | **G** | **C** | **A** | **U** |
| **G** | Gly | Ala | Asp | Val |
| **C** | Arg | Pro | Gln | Leu |
| **A** | Ser | Thr | Asn | Ile |
| **U** | Cys | Ser | Tyr | Leu |

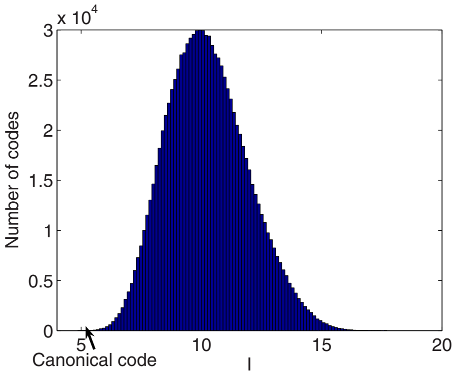

FIG. 1. (Color online) Histogram of average impact $I$ per point mutation for randomly generated codes with the same degeneracy structure as the canonical genetic code.

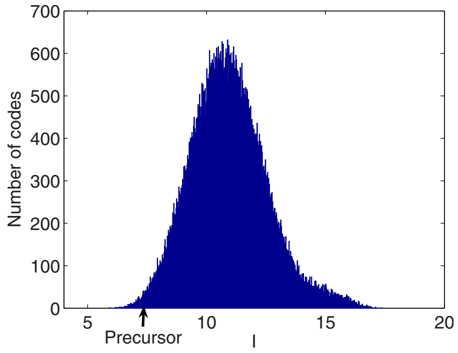

FIG. 2. (Color online) Histogram of average impact $I$ per point mutation for randomly generated codes with the same degeneracy structure as the proposed precursor. There is more noise relative to the canonical code case due to the smaller ensemble of random codes required to calculate $P_b$ for the precursor.

[8,9,19–22]). Thus we anticipate that the proposed precursor code should contain little, if any, evidence for optimality.

We have analyzed the former of these proposed precursor doublet codes (see table) for error-minimizing optimality using the "experimental polar requirement" (EPR) [2,15,16,23] derived originally by Woese and co-workers. We have also analyzed the precursor using a modern computational update of the polar requirement (CPR) [24]. Analysis with the CPR is of particular interest, because it is the measure of amino acid difference that when applied in code optimality analysis algorithms to the canonical genetic code gives rise to the extreme optimality cited above [7]. Thus, the CPR can be considered to capture some essential aspect of amino acid chemistry of particular relevance during the evolution of the genetic code. Analysis of the more refined version of the proposed precursor code is difficult due to the fact that the polar requirements for the proposed precursor amino acids are unknown. This problem can be partially solved by sensitivity analysis, and is discussed in greater detail below.

To analyze the error-minimizing optimality in the proposed precursor code, we used the point mutation code analysis algorithm described in [4,5] to calculate a measure of the average impact of a point mutation of a given code indexed by $i$. Calculating the impact per point mutation allows direct comparison of the optimality of the canonical and precursor codes, because the different size of the set of point mutations for the doublet versus the canonical code has been divided out. With this convention, the optimality distributions for doublet codes and triplet codes are similar (see Figs. 1 and 2).

The presentation of this algorithm in [7] considers an ensemble of random genetic codes genetic code as mappings from the set of codons (minus the termination codons) to the set of amino acids, $GC^i$:Codons → Amino Acids, where $i$ indexes a particular set of assignments of codons to amino acids, with $GC^1$ as the precursor code. Versions $GC^{i\neq 1}$ are generated by randomly permuting amino acid labels, again excluding termination codons. A measure of the average impact $I$ per point mutation for a given code $i$, can then be calculated as

$$I_i = \frac{\Sigma_{\langle c,c'\rangle \neq Ter}[GC^i(c) - GC^i(c')]^2}{\Sigma_{\langle c,c'\rangle \neq Ter}1}, \qquad (1)$$

where $\langle c,c'\rangle \neq Ter$ denotes a sum over 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.

To extract a measure of optimality that restricts optimality comparisons of the precursor codes to other doublet codes, we compute the probability $P_b$=Prob$(I<I_1)$ that a random realization is less impacted by point mutations (more optimal) than the proposed precursor code. This can be achieved by calculating the percentage of random doublet codes that are more optimal than the precursor code. If we are computing the optimality of the canonical code, $P_b$ is calculated strictly from an ensemble of triplet codes. The fact that $P_b$ is based on strict comparison to the appropriate ensemble of random codes will allow us to compare $P_b$ from the proposed precursor to that of the canonical code.

The error in the computed $P_b$ can be estimated using an analytical realization of bootstrap resampling derived from an exact correspondence with the statistics of the asymmetric one-dimensional random walk [7]. This correspondence shows that if $N$ codes are sampled, and $N_{I<I_1}$ are more optimal than the code being tested, then $P_b$ with standard error is given by the expression

$$P_b = (N_{I<I_1} \pm \sqrt{N_{I<I_1}})/N. \qquad (2)$$

While this is in line with naive expectations for the form of error, the problem of sampling more optimal random codes is a problem of rare event sampling, which is frequently unstable and prone to nonstandard large errors. This makes a rigorous derivation of the exact error a key result essential for robust interpretation of optimality calculation results. The form of the error also informs the computations. It is clear from Eq. (2) that the relevant sample size for a statistically sound analysis is not $N$, but the number of more optimal codes sampled, $N_{I<I_1}$ [7]. A reasonable minimum is, perhaps, 20 more optimal codes sampled to get a statistical estimate. Much larger samples would be preferable, but in many ap-

plications may be hard to obtain due to computational limitations encountered when analyzing highly optimized codes.

When applied to the proposed precursor code, we calculated $P_b = (1.44 \pm 0.038) \times 10^{-2}$ with the experimental polar requirement, or $P_b = (7.95 \pm 0.282) \times 10^{-3}$ with the computational polar requirement [24]. To compare, we applied this simplified code analysis algorithm to the canonical genetic code. The canonical genetic code has optimality of $P_b = (1.18 \pm 0.109) \times 10^{-4}$ or $P_b = (4.7 \pm 0.686) \times 10^{-5}$ with the EPR and CPR, respectively, (the extreme optimality discussed above included transition and transversion biases for each base position in the calculation [5,7]). Thus the optimality of the precursor is, with either the EPR and the CPR, two orders of magnitude less optimal than the canonical genetic code evaluated with the equivalent algorithm. The absolute $I$ for both codes can also be compared because they are calculated per point mutation (see discussion above). For the canonical code, $I = 5.293$, which we know from $P_b$ to highly optimal. For the precursor, $I = 7.498$. Given that the mean of the $I$ distribution is near 10 for both the doublet and the triplet case (Figs. 1 and 2), the optimality is substantially reduced for the doublet, consistent with the results from $P_b$.

As discussed above, the derivation of the doublet code in Table I depended on projecting the third base onto the doublets by favoring the simplest amino acid coded for by the triplet codons associated with a given doublet. We repeated the optimality analysis for versions of the doublet code that favored more complex amino acids at individual doublets (such as substituting Arg for Ser at the AG position). None of the modified doublet codes displayed a significant increase in optimality over the version in Table I.

We also note that the version of the precursor code we studied used some amino acids that are regarded as late additions [25].While it seems unlikely that the later amino acids would have substantially different polar requirements than their predecessors in the same synthesis path, to assess the impact of possible changes in polar requirement values as these amino acids (Arg, Gln, Asn, Ile, and Cys) were introduced, we varied their polar requirement values ±20%, and redid the optimality calculation. In all cases, the optimality of the precursor declined, or showed such small improvement that the error bars overlapped with the primary calculation, leaving our basic conclusions about the optimality of the precursor code unchanged. This analysis shows that our results are unlikely to be changed when analyzed with all of the polar requirements for precursor amino acids proposed in [18]. Since varying individual amino acid polar requirement values did not enhance the optimality properties of the precursor, a version of the precursor code which is highly optimal and respects the underlying biosynthesis theory would differ in several positions from the proposal by Copley *et al.* [18].

Our results show that the proposed precursor code has weak error-minimizing optimality with respect to the polar requirement, compared to the canonical genetic code. This result is surprising in one respect, because the doublet code is a projection of the canonical code. A number of interpretations are possible. (1) The doublet precursor code is not an intermediate evolutionary stage from some earlier precursor code; this is consistent with the basis for the original proposal of this code as a biosynthetic pathway, but is puzzling because the latter canonical triplet code is optimized with respect to the free amino acid polar requirement. (2) The precursor has no biological significance at all, and did not evolve from an earlier precursor, which exhibits free amino acid polar requirement optimality. (3) The precursor doublet code predates evolution for error minimization, and if the amino acid synthesis scheme is correct, then modifications to the doublet code during its evolution to today's canonical code are responsible for its observed error-minimizing optimality. The relatively large $P_b$ value (i.e., small amount of observed optimality) in the precursor is an artifact of deriving the doublet code from the highly evolved canonical code.

Our analysis does not address the question of whether or not the detailed biochemical theory proposed is correct, because presumably optimal precursor codes that are both consistent with the biochemical theory and uncorrupted by evolution could be constructed.

We gratefully acknowledge discussions with Carl Woese, Rob Knight, Shelley Copley, Eric Smith, and Harold Morowitz. This material is based on work supported by the National Science Foundation under Grant No. NSF-EF-0526747.

[1] T. Sonneborn, in *Evolving Genes and Proteins*, edited by V. Bryson and H. J. Vogel (Academic Press, New York, 1965), pp. 277–297.

[2] C. R. Woese, Proc. Natl. Acad. Sci. U.S.A. **54**, 71 (1965).

[3] C. Alff-Steinberger, Proc. Natl. Acad. Sci. U.S.A. **64**, 584 (1969).

[4] D. Haig and L. D. Hurst, J. Mol. Evol. **33**, 412 (1991).

[5] S. J. Freeland and L. D. Hurst, J. Mol. Evol. **47**, 238 (1998).

[6] S. Freeland, T. Wu, and N. Keulmann, Origins Life Evol. Biosphere **33**, 457 (2003).

[7] T. C. Butler, N. Goldenfeld, D. Mathew, and Z. Luthey-Schulten, Phys. Rev. E **79**, 060901(R) (2009).

[8] R. Knight, S. Freeland, and L. Landweber, Nat. Rev. Genet. **2**, 49 (2001).

[9] K. Vetsigian, C. Woese, and N. Goldenfeld, Proc. Natl. Acad. Sci. U.S.A. **103**, 10696 (2006).

[10] R. Knight, L. Landweber, and M. Yarus, *Translation Mechanisms* (Kluwer Academic, New York, 2003) p. 115.

[11] M. Yarus, J. Gregory Caporaso, and R. Knight, Annu. Rev. Biochem. **74**, 179 (2005).

[12] S. Itzkovitz and U. Alon, Genome Res. **17**, 405 (2007).

[13] R. Knight, Ph.D. thesis, Princeton University, 2001 (unpublished).

[14] E. Koonin and A. Novozhilov, IUBMB Life **61**, 99 (2009).

[15] C. R. Woese, D. H. Dugre, W. C. Saxinger, and S. A. Dugre, Proc. Natl. Acad. Sci. U.S.A. **55**, 966 (1966a).

[16] C. R. Woese, D. H. Dugre, S. A. Dugre, M. Kondo, and W. C. Saxinger, Cold Spring Harbor Symp. Quant. Biol. **31**, 723

(1966).

[17] J. G. Caporaso, M. Yarus, and R. Knight, J. Mol. Evol. **61**, 597 (2005).

[18] S. D. Copley, E. Smith, and H. J. Morowitz, Proc. Natl. Acad. Sci. U.S.A. **102**, 4442 (2005).

[19] S. Osawa and T. H. Jukes, J. Mol. Evol. **28**, 271 (1989).

[20] D. W. Schultz and M. Yarus, J. Mol. Biol. **235**, 1377 (1994).

[21] R. Knight, S. J. Freeland, and L. F. Landweber, Trends Biochem. Sci. **24**, 241 (1999).

[22] D. H. Ardell and G. Sella, J. Mol. Evol. **53**, 269 (2001).

[23] C. R. Woese, Proc. Natl. Acad. Sci. U.S.A. **54**, 1546 (1965).

[24] D. C. Mathew and Z. Luthey-Schulten, J. Mol. Evol. **66**, 519 (2008).

[25] E. N. Trifonov, Gene **261**, 139 (2000).