RESEARCH ARTICLE

Genetic codes optimized as a traveling salesman problem

Oliver Attie[1], Brian Sulkow[1], Chong Di[1], Weigang Qiu[1,2,3]*

1 Department of Biological Sciences, Hunter College, City University of New York, New York, United States of America, **2** Graduate Center, City University of New York, New York, United States of America, **3** Department of Physiology and Biophysics & Institute for Computational Biomedicine, Weil Cornell Medical College, New York, New York, United States of America

* weigang@genectr.hunter.cuny.edu

Abstract

The Standard Genetic Code (SGC) is robust to mutational errors such that frequently occurring mutations minimally alter the physio-chemistry of amino acids. The apparent correlation between the evolutionary distances among codons and the physio-chemical distances among their cognate amino acids suggests an early co-diversification between the codons and amino acids. Here we formulated the co-minimization of evolutionary distances between codons and physio-chemical distances between amino acids as a Traveling Salesman Problem (TSP) and solved it with a Hopfield neural network. In this unsupervised learning algorithm, macromolecules (e.g., tRNAs and aminoacyl-tRNA synthetases) associating codons with amino acids were considered biological analogs of Hopfield neurons associating "tour cities" with "tour positions". The Hopfield network efficiently yielded an abundance of genetic codes that were more error-minimizing than SGC and could thus be used to design artificial genetic codes. We further argue that as a self-optimization algorithm, the Hopfield neural network provides a model of origin of SGC and other adaptive molecular systems through evolutionary learning.

OPEN ACCESS

Citation: Attie O, Sulkow B, Di C, Qiu W (2019) Genetic codes optimized as a traveling salesman problem. PLoS ONE 14(10): e0224552. https://doi.org/10.1371/journal.pone.0224552

Editor: Dapeng Wang, University of Leeds, UNITED KINGDOM

Received: March 26, 2019

Accepted: October 16, 2019

Published: October 28, 2019

Data Availability Statement: Source code is available as Open Source on Github at https://github.com/weigangq/code-by-tsp, including Perl/Python/R codes for simulated annealing, Perl codes for random permutation of genetic codes, and Perl codes for calculating fitness of a genetic code.

Funding: We thank Dr Lia Di for graphic assistance. This work was supported by Public Health Service grants AI139782 (to WQ) from the National Institute of Allergy and Infectious Diseases (NIAID) and the grant MD007599 (to Hunter College) from the National Institute on Minority

Introduction

Limits of natural selection

Discoveries by Alfred Wallace and Charles Darwin in the 19th century established natural selection as the primary mechanism of evolutionary adaptation [1,2]. The Modern Synthesis and the more recent Neutral Theory of molecular evolution reaffirm the supremacy of natural selection as the only evolutionary mechanism capable of deterministic, directional changes against a backdrop of stochastic, directionless evolutionary forces including mutation, genetic drift, and recombination [3–5]. Yet theories of natural selection and population genetics offer limited explanations on the apparent evolutionary trend towards increasing organizational complexity of living systems, as evident [6] in the repeated emergence of complex and robust structures and molecular processes including molecular pathways, subcellular structures, multicellularity, sexual reproduction, embryonic development, sociality, and self-regulating

Health and Health Disparities (NIMHD) of the National Institutes of Health (NIH), and grant 1517002 from the National Science Foundation (to Joao Xavier and WQ) of the United States of America The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

ecosystems [7,8]. A major weakness of Darwinian natural selection and population genetic analysis is its inability to specify testable algorithmic steps, to replicate with simulation, or to predict the future outcomes of organic complexity [9,10].

In recent years, computational and algorithmic learning emerged as a major contender to bridge the epistemological gap between the feasibility of organismal complexity foretold by the theory of natural selection and its algorithmic plausibility [9–12]. Algorithmic learning, defined as the process of creating internal representations (e.g., as memories or genomes) of external regularities through "concrete computation that takes a limited number of steps", applies equally well to understand the biological origins of cognition and adaptation [10]. For example, the Multiplicative Weights Update learning algorithm has been shown to be equivalent to natural selection on genotypes in asexual populations and on individual alleles in sexual populations [13,14]. By quantifying the degree of computational complexity of the problems imposed by environments, computational learning has not only the potential to generate complex adaptations but also to predict the limit and adverse consequences of adaptations [10].

Evolutionary connectionism

Computational learning is at the heart of a new evolutionary paradigm termed "evolutionary connectionism", which posits a theoretical equivalence between learning algorithms and evolutionary processes leading to the emergence of complex adaptations [11,15]. Central to evolutionary connectionism is the concept of correlational or associative learning first proposed by Donald Hebb in understanding the spontaneous origin of neural networks capable for memory [16]. The Hebbian rule, known colloquially as "cells that fire together, wire together", is now postulated as a general pattern-generating mechanism in living systems beyond neural systems. For example, the origin of developmental pathways and other gene regulatory networks may be a consequence of following genomic analogs of the Hebbian rule that "genes that are selected together wire together" and "genes that function together junction together" [17]. Bacterial operons, consisting of tandemly arranged and co-transcribed genes, may be a physical manifestation of the Hebbian rule in shaping genome structure ("genes that function together locate together").

Despite (or perhaps because of) its simplicity, the Hebbian correlational learning has considerable power in generating adaptive and robust features in living systems beyond neural networks. First, as an unsupervised learning algorithm, Hebbian learning is efficient in finding locally optimal solutions without the need to search through a prohibitively large number of possible solutions and test these solutions individually, as is necessary in the process of natural selection. Second, as a distributed algorithm, Hebbian learning occurs locally between directly interacting entities (neurons or genes) while nonetheless leading to emergence of system-level optimal structures (neural or gene networks). It is not necessary to hypothesize that the system-level phenotypes (e.g., developmental processes or gene networks) are themselves direct targets of natural selection. Third, by discovering and encoding trait correlations, Hebbian learning results in networks with foresightedness and robustness [11].

For the time being, however, evolutionary connectionism has primarily been a conceptual framework. It remains unclear how extensively Hebbian learning–relative to natural selection and other learning algorithms–operates in shaping molecular and cellular structures beyond neural systems. It is desirable and indeed necessary to test claims of evolutionary learning in the context of a specific molecular system with computational, statistical and biological evidence.

Origin of genetic code

Here we test the hypothesis that the standard genetic code may have been evolved predominantly through a self-learning rather than a natural-selective process. The standard genetic code (SGC, Fig 1A), with which 64 possible combinations of triple nucleotides ("codons") encode 20 canonical amino acids and the translational stop signal, underlines much of the robustness of organisms against deleterious mutations. For example, single-nucleotide mutations occurring at the 3rd codon position typically result in no change of amino acids. Such properties of SGC are called "error-minimization", referring to the fact that SGC is non-randomly structured to minimize disruptions by DNA mutations to protein sequences, structures, and functions.

Although there is a consensus on the adaptiveness of SGC [20–25], the molecular processes that led to its origin remain controversial [20,22,24,24–31]. Natural selection, in conjunction with chemical processes during early history of life, has been the most frequently argued hypothesis to explain the origin of the highly optimized and nearly universal SGC [26,30,32]. However, the selective mechanism tends to be slow and inefficient because it implies that SGC emerges from competition among cells equipped with random genetic codes [33]. More contentiously, because the protein translation machinery is a complex subcellular system consisting of, among others, tRNAs, aminoacyl-tRNA synthetases (aaRSs) and the ribosome, the selective hypothesis raised the question whether (and how) natural selection operates at the level of individual genes, the subcellular system, or the cell system as a whole. For example, it has been argued that the apparent adaptiveness of SGC may have risen as a by-product of incremental evolution through codon capture, during which structurally similar aaRSs tend to

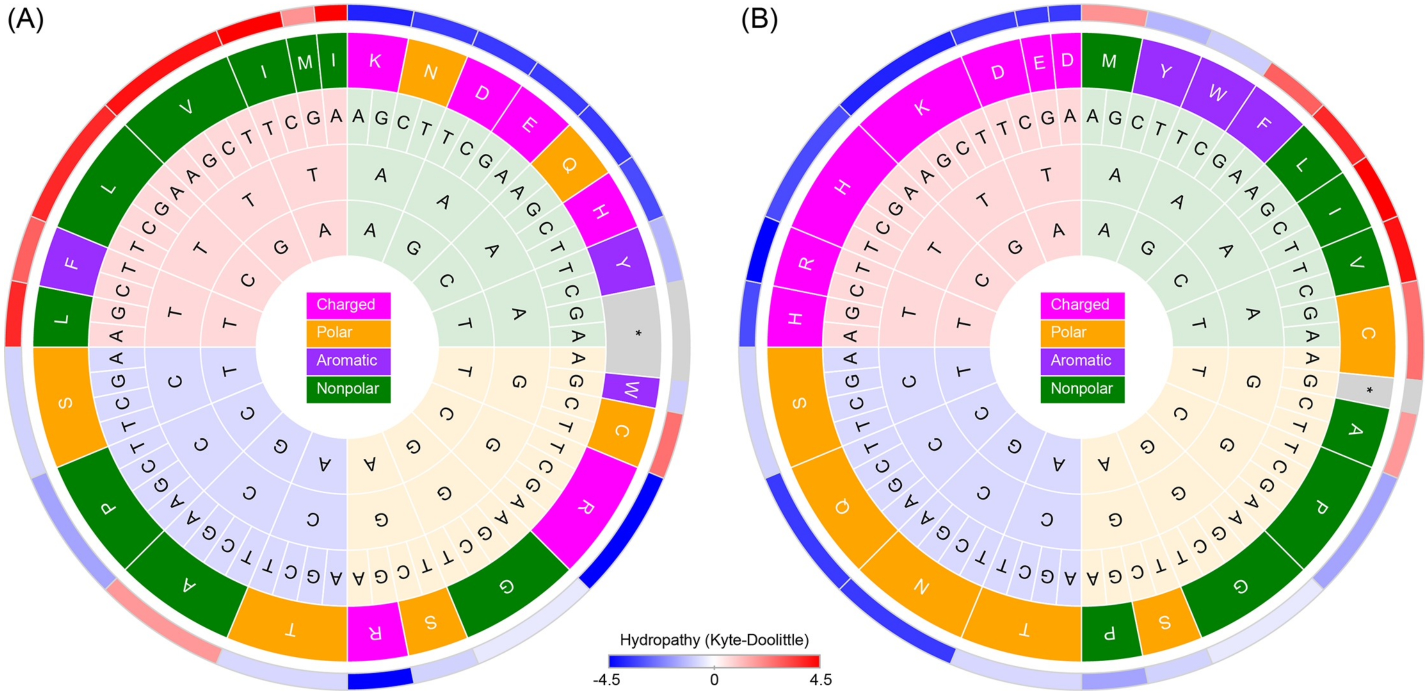

Fig 1. Codon wheels showing (A) the standard genetic code (SGC) and (B) a Hopfield-optimized genetic code. Bases are arranged (from center outward) in the order of 1st, 2nd, and 3rd codon positions. Within each of these three rings, the four bases are cycled clockwise in the order of "AGCT" to minimize the number of transversions. The codon wheel represents the shortest mutation path and is modeled as the time dimension in the Traveling Salesman Problem (TSP) representation of the genetic codes. In the 4th ring, amino acids are colored according to a physio-chemical classification [18]. In the 5th ring, amino acids are colored by the Kyte-Doolittle measure of hydrophobicity [19].

https://doi.org/10.1371/journal.pone.0224552.g001

recognize phylogenetically similar tRNAs and physio-chemically similar amino acids [6,12,34–36]. Regardless, neither the selective hypothesis nor the incremental evolution hypothesis directly addresses the algorithmic processes by which error-minimizing capacity of SGC may have evolved. Besides being of considerable evolutionary interest, the design principle and algorithmic origin of SGC are of practical importance for reengineering the genetic code to create synthetic therapeutic proteins using non-canonical amino acids [37,38].

Here, we explore the possibility of an evolutionary origin of SGC through self-optimization. The search for optimal genetic codes by using learning algorithms is not new and studies have concluded that SGC is far from being globally or even locally optimal [20,24–26,31,39–41]. Our specific objective is to test if evolutionary connectionism (and the Hebbian learning rule in specific) could lead to error-minimizing genetic codes.

Models & methods

Hopfield network & traveling salesman problem

Hopfield neural network is an algorithmic model of associative memory that implements the Hebbian learning rule [42]. A Hopfield neural network, consisting of symmetrically connected neurons, is capable of storing, retrieving, and restoring memories when activities of individual neurons are determined by the Hebbian learning rule [42,43]. For example, in a binary Hopfield network where each neuron's activity takes the value of either 1 or -1, the connection weight ($w_{i,j}$) between a pair of neurons is increased if their activities (x_i and x_j) are positively correlated (both 1 or both -1) and decreased if negatively correlated (one 1 and the other -1). Once connection weights are specified, a random initial state the network would evolve towards a locally stable state where the following definition of the network energy is at the minimum [43]:

$$E = -\frac{1}{2}\sum_{i,j} w_{i,j} x_i x_j \tag{1}$$

As such, while the neuron activities are determined at local levels through interactions with each other, the neural network as a system displays collective, emergent behavior mimicking associative memory. The network is able to, for example, retrieve a complete state from incomplete inputs or recover a correct state from inputs with errors. The system is also robust in the sense that the network functions well even when some of the neurons are removed, mimicking a damaged brain and indicating encoded information redundancy [43].

Besides being a powerful model of associative memory, Hopfield network is an efficient algorithm for solving combinatorial optimization problems, such as the traveling salesman problem (TSP), which is to find the shortest tours to cover N cities while visiting each city once [44]. Instead of being a memory device, here the Hopfield network was used as a computational tool to search for combinatorial optimal states. To solve the TSP, for example, a Hopfield network is simulated with N^2 neurons, each representing the probability of a city being visited at a tour position. An energy function is defined to reflect both the constraints (e.g., each city to be visited once and the salesman can visit only one city at a time) and a measurement to be minimized (e.g., the tour length). When initialized with arbitrary activities, the network progressively reaches a minimum representing a locally shortest path as the energy function converges to a stable local minimum [44,45].

The original Hopfield-Tank algorithm was found difficult to replicate as the number of cities increased and a more efficient algorithm using neural normalization and simulated annealing was proposed [46]. The mean field of a neuron representing city X visited at tour position i

is defined as:

$$E_{Xi} = d_p \sum\nolimits_{Y \neq X} v_{Y_i} + \sum\nolimits_{Y \neq X} d_{XY} (v_{Y,i+1} + v_{Y,i-1}) \tag{2}$$

, where d_p is a penalty experimentally adjusted to ensure that only one city can occupy a tour position and v_{Xi} is the probability of city X at tour position i. Values of v_{xi}'s are assumed to obey a Boltzman distribution at any given simulated temperature T: $v_{Xi} \propto e^{-E_{Xi}/T}$. At each iterative updating step, the neuron outputs are normalized to sum up to one so that each value represents a true probability:

$$v_{Xi} = \frac{e^{-E_{Xi}/T}}{\sum_j e^{-E_{Xj}/T}} \tag{3}$$

Finding a valid path depends on the temperature T and the penalty d_p. As the temperature increases, neuron outputs become increasingly uniform ($v_{Xi} \rightarrow 1/n$). As the temperature drops to a critical value (T_0), the neurons anneal to a steady low-energy state representing a stable, locally minimal-energy mapping between the cities and tour positions. Both the critical temperature T_0 and d_p that lead to valid tours were experimentally determined by simulations as in all other Hopfield solutions of the TSP [46]. The initial values of v_{Xi} influence whether the network converges. To ensure convergence, we used initial voltage values of 0.5 plus a random number in between -0.1 and +0.1 and d_p = 0.7.

Algorithm to generate optimal genetic codes using TSP

TSP model of SGC origin. We modeled a genetic code as a tour in the Traveling Salesman Problem (TSP) so that it could be optimized by using a Hopfield neural network. We hypothesized that the optimal genetic code minimized the total distance of a tour of 20 amino acids from one codon to next, analogous to a tour of N cities by a salesman from one time point to the next. To make the codons equivalent to the linearly ordered time points in a TSP, it was necessary to generate a linearly ordered sequence of codons that minimizes the cumulative mutational distance between codons (see "codon wheel" below).

To represent genetic codes as a TSP, we first constructed a Hopfield network consisting of 21 x 21 neurons, the activity of each of which representing the probability of an amino acid ("cities") at a tour position. We measured distances between two amino acids (X and Y) as the Euclidean distance: $d_{XY} = \sqrt{\sum_i (X_i - Y_i)^2}$, where i stands for one of the physio-chemical indices (either hydrophobicity, polarity, volume, or isoelectric point; Table 1). To remove effects of difference in magnitudes when combining multiple indices, we rescaled each index by normalizing the values to a mean of zero while maintaining the standard deviation. Values for stop codons were arbitrarily assigned to be an outlier (greater or less than two standard deviations from the mean).

Second, the neural network was initialized with uniformly distributed random activities centered at 0.5 [44]. The network was then optimized by following the simulated annealing algorithm with preset values of d_p (e.g., d_p = 0.7) and T (e.g., T = 0.1), determined experimentally to maximize the proportion of valid tour paths (Eqs 2 & 3) [46]. Each resulting optimal tour path was checked for validity to ensure that all amino acids were covered and each amino acid was visited only once. Invalid paths were discarded and valid paths were saved for further analysis.

Third, we mapped the amino acids from an optimal path produced by the Hopfield network to the 64 possible codons. Because there are more codons than amino acids it was necessary to assign multiple codons to a single amino acid.

Table 1. Amino acid indices.

1-letter code	3-letter code	Polarity[a]	Hydrophobicity[a]	Volume[a]	Iso-electric point[a]
A	Ala	7.0 (-0.1672)	1.8 (0.7638)	31 (-1.067)	6.00 (-0.02972)
C	Cys	4.8 (-1.043)	2.5 (1.002)	55 (-0.5477)	5.07 (-0.5561)
D	Asp	13.0 (2.221)	-3.5 (-1.039)	54 (-0.5694)	2.77 (-1.858)
E	Glu	12.5 (2.022)	-3.5 (-1.039)	8.3 (-1.559)	3.22 (-1.603)
F	Phe	5.0 (-0.9631)	2.8 (1.104)	132 (1.120)	5.48 (-0.3241)
G	Gly	7.9 (0.1910)	-0.4 (0.01531)	3 (-1.674)	5.97 (-0.0467)
H	His	8.4 (0.3900)	-3.2 (-0.9373)	96 (0.3402)	7.59 (0.8703)
I	Ile	4.9 (-1.003)	4.5 (1.682)	111 (0.6651)	6.02 (-0.0184)
K	Lys	10.1 (1.067)	-3.0 (-0.8693)	119 (0.8384)	9.74 (2.087)
L	Leu	4.9 (-1.027)	3.8 (1.444)	111 (0.6651)	5.98 (-0.04104)
M	Met	5.3 (-0.8437)	1.9 (0.7978)	105 (0.5352)	5.74 (-0.1769)
N	Asn	10.0 (1.027)	-3.5 (-1.039)	56 (-0.5261)	5.41 (-0.3637)
P	Pro	6.6 (-0.3264)	-1.6 (-0.3930)	32.5 (-1.035)	6.30 (0.1401)
Q	Gln	8.6 (0.4696)	-3.5 (-1.039)	85 (0.102)	5.65 (-0.2278)
R	Arg	9.1 (0.6686)	-4.5 (-1.380)	124 (0.9466)	10.76 (2.665)
S	Ser	7.5 (0.03184)	-0.8 (-0.1208)	32 (-1.046)	5.68 (-0.2108)
T	Thr	6.6 (-0.3264)	-0.7 (-0.08676)	61 (-0.4178)	6.16 (0.06085)
V	Val	5.6 (-0.7243)	4.2 (1.580)	84 (0.08035)	5.96 (-0.05236)
W	Trp	5.2 (-0.8835)	-0.9 (0.1548)	170 (1.943)	5.89 (-0.09198)
X[b]	Stop	(-1.5993)	(1.672)	(-1.5833)	(-2.6562)
Y	Tyr	5.4 (-0.8039)	-1.3 (0.2909)	136 (1.207)	5.66 (-0.2222)

[a] Raw values were obtained from[47] Haig D, Hurst Land normally scaled indices are in parenthesis. Stop signal was excluded from normalization.

[b] Stop signal, arbitrarily assigned values that are the least polar, most hydrophobic, largest in volume, and most negatively charged so that its distances from amino acids are greater than the distance between any two amino acids.

https://doi.org/10.1371/journal.pone.0224552.t001

Codon wheel. We constructed a linear order of codons with minimal evolutionary distances by following known molecular evolutionary principles. First we gave preference to mutations at the 3rd codon position and then to those at the 1st codon position, followed by those at the 2nd codon position, reflecting increasing evolutionary distances of nucleotide substitutions from the 3rd, to the 1st, and to the 2nd codon positions. Second, for mutations introduced to the same codon positions, we gave preference to transitions over transversions, reflecting the fact that transitions occur more frequently than transversions [22]. Following these two rules, a linear sequence of codons, created by cycling the four bases at each codon position in the order of, e.g., "AGCTTCGA", to minimize the total number of transversions, were uniquely defined and shown as a circular codon wheel (Fig 1). Note that the codon wheel could alternatively be defined by any of the other three possible transversion-minimizing base-cycling sequences: "AGTCCTGA", "GACTTCAG", or "GATCCTAG". The resulting optimal codes would vary but be similar in having the shortest cumulative amino acid distances. Fig 1B shows one of such optimal codes.

To assign multiple codons to the same amino acid, we started with an arbitrary codon in a codon wheel (e.g., "AAA") and traveled clockwise through all codons, while labeling each codon with an integer determined by the order of distinct amino acid to which this codon is assigned according to SGC (Fig 1). In other words, we assigned each codon an "SGC address". For example, starting from "AAA" in a codon wheel shown in Fig 1, the codons were labeled as "AAA" (1), "AAG" (1), "AAC" (2), "AAT" (2), "GAT" (3), "GAC" (3), "GAG" (4), "GAA" (4), and so forth and ended with "ATT" (19), "ATC" (19), "ATG" (20), and "ATA" (21). This way, an optimal amino-acid path generated by a Hopfield network could be assigned to 64 codons based on tour positions. For example, if Lysine (K) has a tour position of 4 in a TSP path, it will be assigned to two codons "GAG" and "GAA", both of which have an "SGC address" of 4 (although Lysine codons are "AAA" and "AAG" in SGC). Note that the SGC addresses of codons were arbitrarily assigned as long as they follow the order of a codon wheel. If tour positions of amino acids were random (i.e., not optimized by Hopfield network), this scheme is equivalent to a random permutation of amino acids among synonymous codon blocks [26]. A Hopfield-optimized tour path of amino acids, on the other hand, is expected to maximize the chance that similar amino acids are assigned to similar codons. In all cases, the choice of the initial codon determines the identity but not the error rate of the evolved codes.

Statistical analysis of genetic codes

Randomized genetic codes. To test optimality of SGC and simulated genetic codes, we generated random codes as statistical controls by permuting the 20 amino acids and the stop signal among the 21 synonymous codon blocks [26]. This randomization scheme is a stringent test of code optimality. It maintains the same codon degeneracy as in SGC while removing any correlation that might exist between amino acids and codon blocks [26].

Code optimality measured by mutational error. We quantified optimality of each code by calculating errors (i.e., changes in an index value) caused by single-nucleotide substitutions [47]. The mutational error of a code, a measure of overall code fitness, was the average error across all pairs of single-mutation codons [26]:

$$\Delta = \frac{\sum_{i=1}^{61}\sum_{j=1}^{9} w|X_i - X_j|}{n} \tag{4}$$

, where i was the source codon, j was the destination codon (stop codons excluded) differing from the source codon by a single nucleotide, w was the transition/transversion ratio, which was set to 5 (Fig 2), X_i and X_j were the physio-chemical values of the two amino acids associated with the two codons, and n was the total number of one-nucleotide neighboring codon pairs. Statistical significance of the error of a code was assessed by the proportion of random codes with an equal or smaller error. Errors were also calculated individually at the three codon positions and for transitions and transversions separately as a way to estimate if errors were minimized at individual codon positions and for transition or transversion.

Phylogenetic analysis

To explore biological basis of the self-learning algorithm, we inferred early evolutionary events during the origin of SGC using the tRNA sequences of *Pyrococcus furiosus* strain DSM_3638, a model hyperthermophilic archaebacterium. We downloaded the structurally aligned *P. furiosus* tRNA gene sequences from tRNAdb [48]. Redundant sequences were removed and an approximate maximum likelihood phylogenetic tree was obtained with FastTree using default settings [49]. Using the BpWrapper BIOTREE utility, branches with low bootstrap support (<0.7) were collapsed and the tree was rooted at the midpoint [50]. The phylogenetic tree was plotted using the APE package on the R/RStudio platform [51].We used phylogenetic autocorrelation to test co-diversification of tRNA sequences with their cognate amino acids. Phylogenetic autocorrelation is a measure of association of a variable with a phylogeny, in the same way as spatial autocorrelation being a measures of the influence of a variable by geographic distances [52,53]. We use the *gearymoran* function in the ADE4 package to calculate Moran's I with an amino acid index (e.g., polarity) and obtained its statistical significance with Geary's randomization protocol [54].

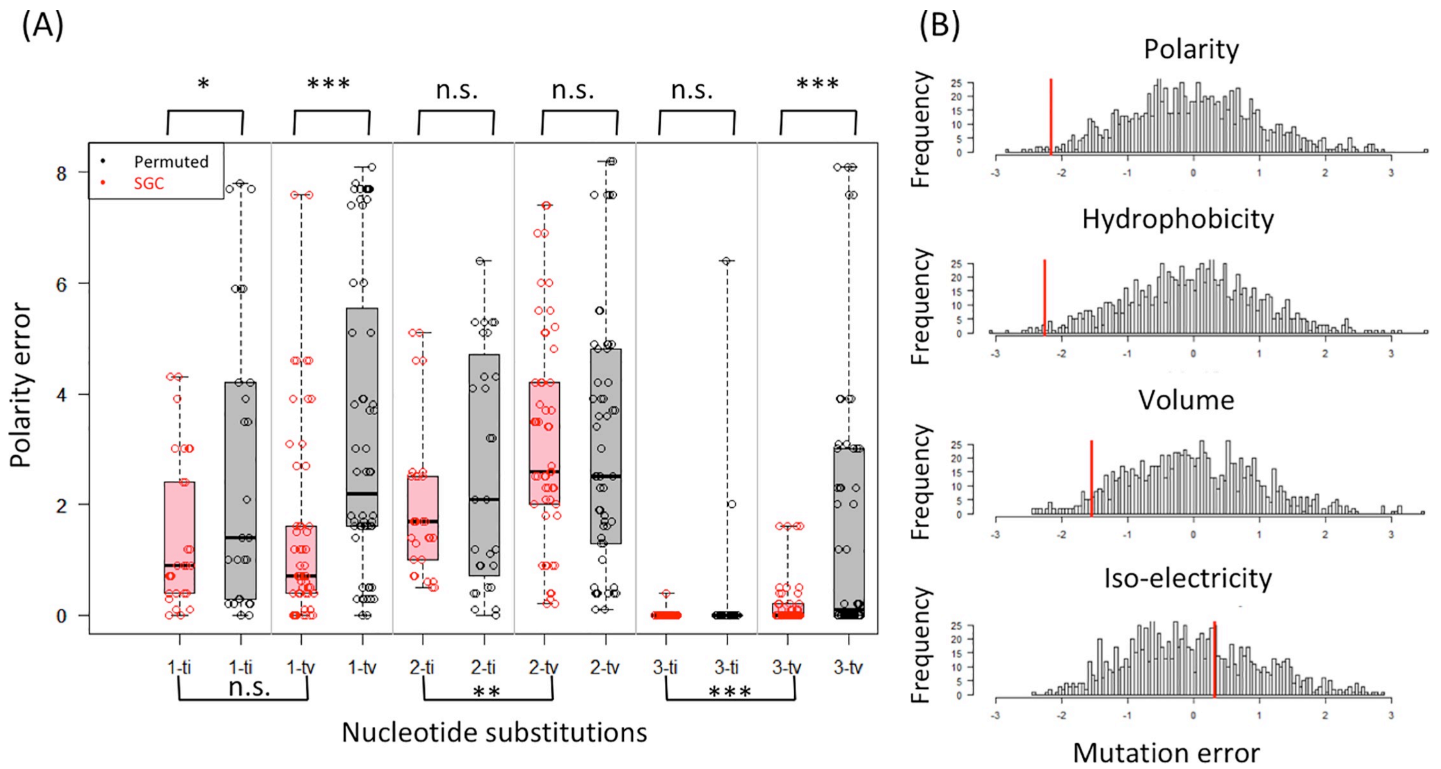

Fig 2. Patterns of mutation-error minimization in SGC. **(A)** On the x-axis, single-nucleotide mutations between pairs of codons are categorized by codon positions (1, 2, and 3) and by transitions (ti) or transversions (tv). The y-axis shows the errors caused by such mutations as quantified by the polarity index (Table 1). Points in red (n = 263) were derived from SGC; points in black (n = 258) by a single round of random permutation of amino acids among the synonymous codon blocks. **(B)** Distributions of amino-acid errors caused by single-nucleotide substitutions (calculated by Eq 4 with ti/tv = 5) from 1000 permuted codes. These error distributions show SGC significantly reduces mutation errors for polarity (p = 0.010, *t*-test), hydrophobicity (p = 0.011), and volume (p = 0.048), but not for isoelectric points (p = 0.648).

https://doi.org/10.1371/journal.pone.0224552.g002

Results

Two sources of error minimization in SGC

There are a total of 263 pairs of amino-acid encoding codons that differ by a single base. These codon pairs could be categorized into six types according to the position of the differing base and whether it is a transition (A/G or C/T) or transversion (A/C, A/T, G/C, or G/T). Transitions occur more frequently than transversions, known as the transition/transversion mutational bias. When the error magnitude (measured by e.g., amino-acid polarity; Table 1) between two codons in SGC were plotted for each of the six categories, it was apparent that mutations at the 3rd codon position caused the least errors, followed by the 1st codon position and then by the 2nd codon position (red boxes in Fig 2A). This was largely due to codon degeneracy by which multiple codons code for the same amino acid, which decreased in the order of the 3rd, 1st, and 2nd codon positions and was over-represented by transitions. Codon degeneracy, however, is not itself adaptive because some form of degeneracy is inevitable for any genetic code consisting of more codons than amino acids.

If SGC does not minimize errors between neighboring codon blocks, one would expect the mutational errors similar between SGC and a randomized code. In reality, errors in SGC (measured by, e.g., errors in polarity) were significantly reduced relative to the permuted codes for transitions and transversions at the 1st codon position (p = 2.9e-2 and 8.7e-5 by t-tests) and transversions at the 3rd codon position (p = 1.4e-5), while there was no significant error reduction at the 2nd codon position (Fig 2A). This pattern of error reduction indicates that SGC minimizes errors caused by mutations at the 1st and 3rd codon positions in addition to codon degeneracy at these positions. At the 2nd codon position, although there was neither codon degeneracy nor significant error reduction relative to the random code, transitions caused significantly less errors than transversions (p = 1.3e-3).

In sum, two sources of error minimization in SGC were identified: (*i*) SGC significantly reduces mutation errors at the 1st codon position and transversion errors at the 3rd codon position and (*ii*) at the 2nd and the 3rd codon positions, errors by transitions are significantly minimized relative to those by transversions. These two rules were used to construct a codon wheel that minimizes evolutionary distances between codons (Fig 1).

SGC co-minimizes errors in amino acid polarity, hydrophobicity, and volume

We tested the overall errors of SGC relative to random codes for the four major amino acid indices (Table 1). Results indicated that SGC significantly minimizes errors in polarity (p = 0.010, by t-test), hydrophobicity (p = 0.011), and volume (p = 0.048), but not iso-electricity (p = 0.648) (Fig 2B). The co-minimization of polarity and hydrophobicity was not surprising because these two indices were significantly anti-correlated with correlation coefficient -0.81, according to a principle component analysis of the four indices (Fig 3A).

Hopfield-optimized genetic codes

There are 4 x 10^{84} possible genetic codes [55]. We used Hopfield networks to search for optimal genetic codes with the goal of assessing the possibility that SGC may have originated by a similar self-learning process. None of these codes was the SGC, but was comparable to the SGC in terms of its optimality. First, we obtained optimal codes by minimizing the total distance measured by a single parameter. For example, a total of 856 valid paths were generated from a run of 2500 rounds of simulations using amino acid distances determined by polarity. Indeed, the polarity errors of the simulated codes decreased significantly relative to the random codes and the mean error rate was close to that of SGC (-2.0 standard deviation from the mean error of random codes) (Fig 4A). The errors for hydrophobicity, volume, and iso-electricity of simulated codes were incidentally reduced. Adding volume as an extra distance parameter had similar error-reducing efforts to all four indices, while the decrease in volume was the largest (-3.0 standard deviation from the mean error) (Fig 4B). Continuing by adding iso-electricity as the 3rd parameter, the simulated codes improved upon the random codes for all three measures of errors (Fig 4C). Finally, including all four parameters resulted in simulated codes optimized for all parameters (Fig 4D).

It could be concluded from these simulations that Hopfield network optimized genetic code in a highly sensitive manner depending on which and how many indices were included in the distance function. Further, considering the large iso-electricity error rate of SGC (0.3 standard deviation greater than the random error, Fig 2B), we rejected the hypothesis that SGC evolved by minimizing errors in isoelectric point. Indeed, it is most parsimonious to conclude that SGC evolved by minimizing the polarity error alone, with reduced errors in hydrophobicity and volume as incidental consequences (Fig 4A). This hypothesis was supported by the fact that optimization with respect to hydrophobicity or volume alone resulted in poorer match of errors between SGC and the simulated codes.

Next, we searched for simulated codes that were more optimal than SGC by plotting the error rates grouped by individual codes (Fig 5). One such code with all four errors less than -2.0 standard deviation away from the mean random errors was identified and its codon

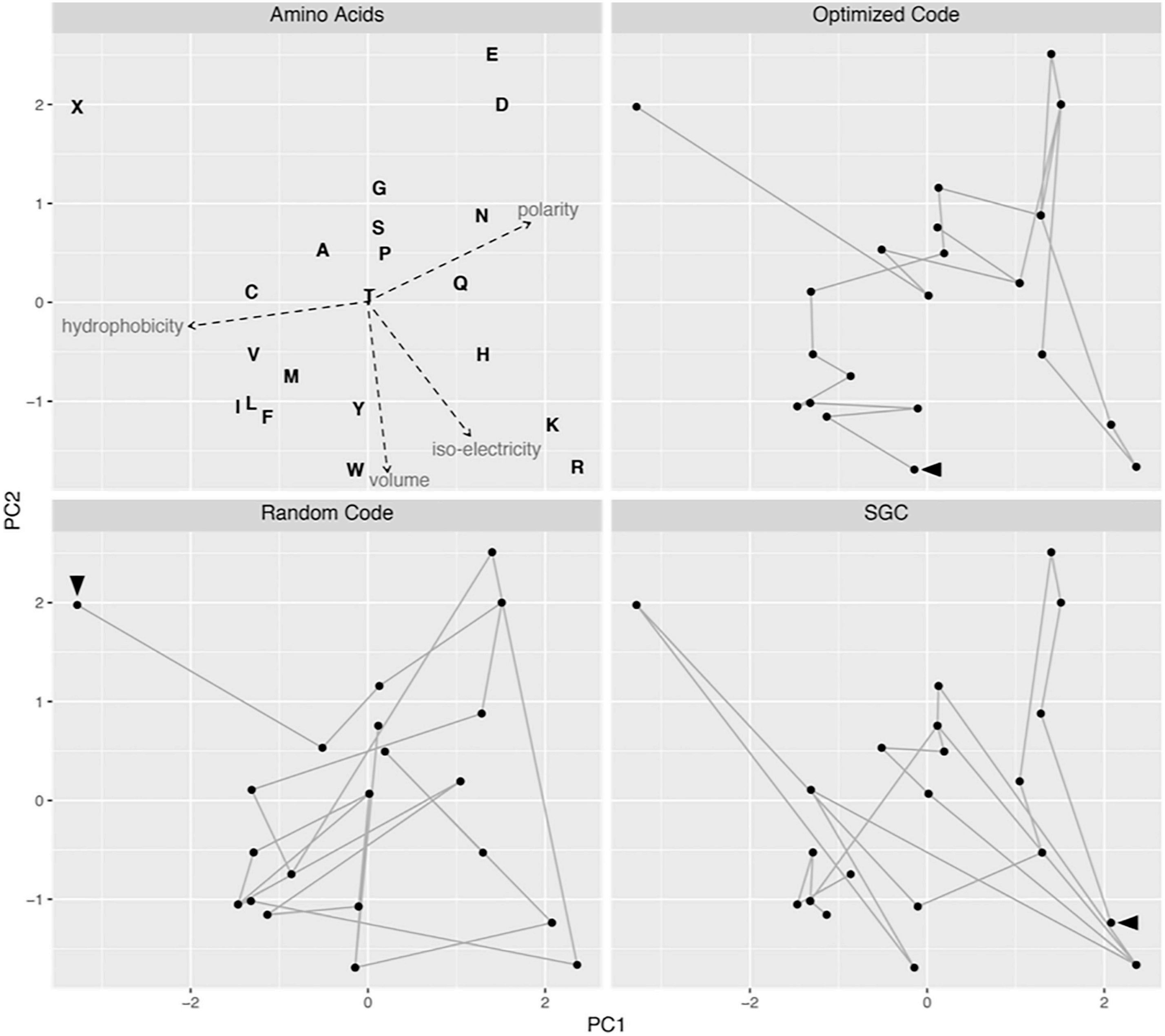

Fig 3. Hopfield network minimizes amino acid tour lengths. (**Top left**) First two principal components of the four amino acid indices (Table 1). The variances explained by the 1st and 2nd principal components are 48.73% and 37.96% respectively. (**Top right**) Amino acid path of a Hopfield-simulated code optimized for polarity and volume (tour lengths 32.2±4.4 in polarity error). (**Bottom left**) Amino acid path of a randomly permuted code (tour lengths 56.3±4.6, sample size 1000). (**Bottom right**) Amino acid path of SGC (tour length 41.0).

https://doi.org/10.1371/journal.pone.0224552.g003

assignment was visualized with a codon wheel (Fig 1B). This Hopfield-optimized code was more optimal than SGC in that, e.g., all non-polar amino acids were mapped to codons with the 2nd codon position being a purine (`A` or `G`). Furthermore, all positively or negatively charged amino acids were mapped to codons with the 2nd codon position being a thymine (`T`). Most significantly, genetic codes similarly or more optimal than SGC were not a rarity but emerged readily from a Hopfield network (Fig 5), suggesting that SGC was suboptimal in error minimization.

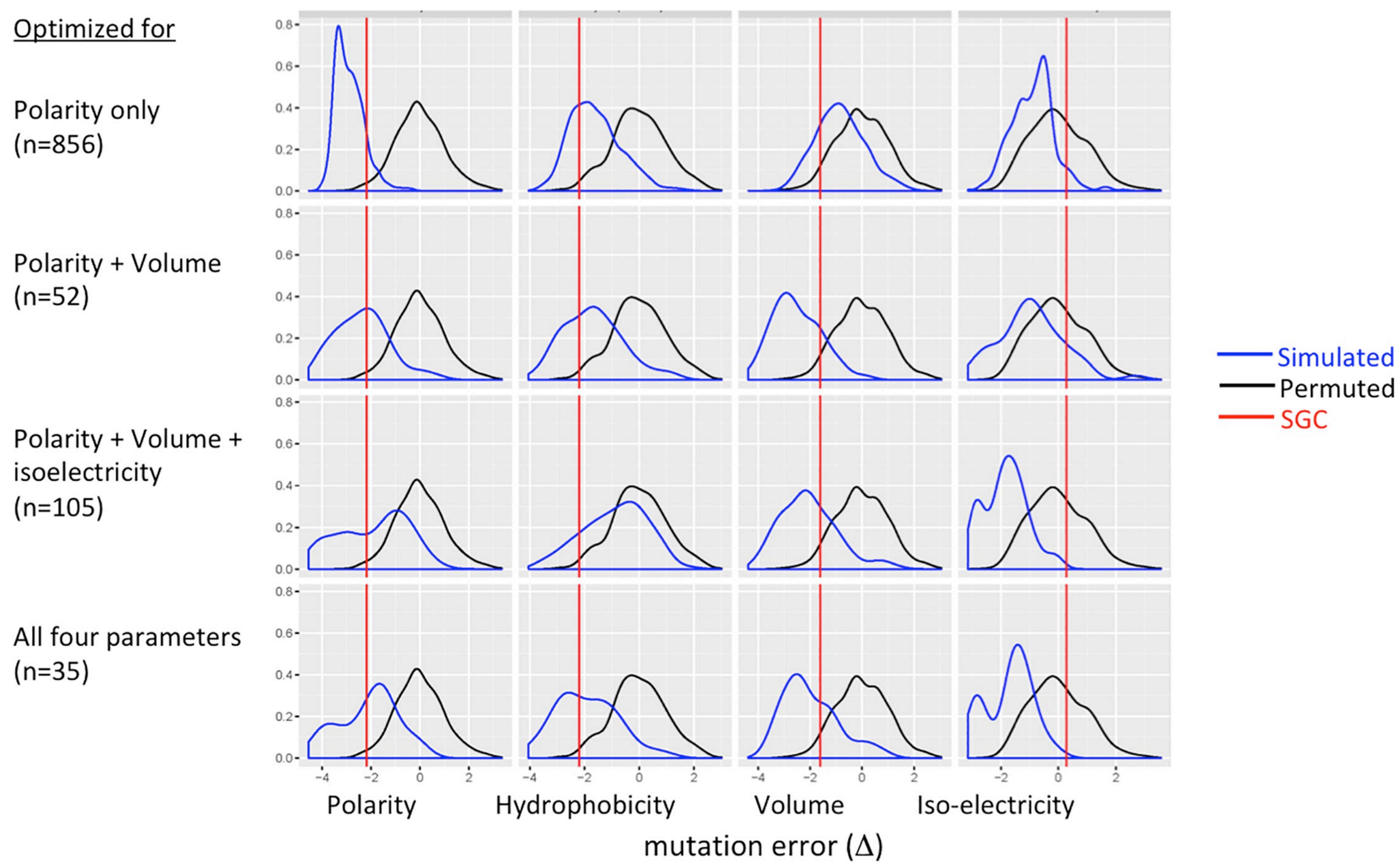

Fig 4. Hopfield network minimizes mutation errors. We constructed a 21-by-21 TSP and searched for the shortest paths traversing the 20 amino acids and the stop codon (21 "cities") using Hopfield neural networks (see Methods). Distances between the "cities" were determined by either polarity alone (**top row**), or a combination of 2, 3, and all 4 parameters (**other rows**). Valid paths (number in parenthesis) were mapped to the codon wheel resulting in simulated genetic codes. Each simulated code was calculated for average single-mutation error rates based on each parameters (Eq 4). All error rates were normalized according to the mean and standard deviation of random codes (as in Fig 2B) and their distribution were shown as density plots. While it is possible to obtain codes optimized for all four parameters (**bottom row**), the most parsimonious codes with errors similar to SGC errors were generated with the Hopfield network that optimized for polarity alone (**top row**). An alternative way to visualize reduced errors in Hopfield-optimized codes is to show each code as an amino acid tour in a 2-dimensional principal component space, which shows significantly shorter tour lengths of Hopfield-optimized codes than those of permuted code and SGC (Fig 3).

https://doi.org/10.1371/journal.pone.0224552.g004

Phylogenetic autocorrelations

So far we have shown that SGC was an unsurprising (and indeed suboptimal) outcome of a Hebbian learning process when operating under the same set of biological constraints as in SGC. Here we explored the possible biological basis of Hebbian learning during the origin of SGC by examining footprints of early evolutionary events left in the gene sequences of the contemporary protein-translation system from the archaebacterium *Pyrococcus furiosus*.

Recognition of amino acids by codons–is mediated by tRNAs, each of which is ligated by an aminoacyl-tRNA synthase (aaRS) at the acceptor stem to a specific amino acid according to the anticodon sequence [56]. Unlike the aaRSs, tRNAs are structurally homologous, suggesting a single common origin. Phylogenies of contemporary tRNA gene sequences show a general monophyly of iso-accepting tRNAs (i.e., tRNAs recognizing the same amino acids) although codon recruitments from different tRNA clades have occurred (Fig 6) [57]. Further, significant phylogenetic autocorrelation of tRNA with the physicochemical properties of cognate amino acids supports that expansion of iso-accepting tRNA groups to all twenty amino acids involved

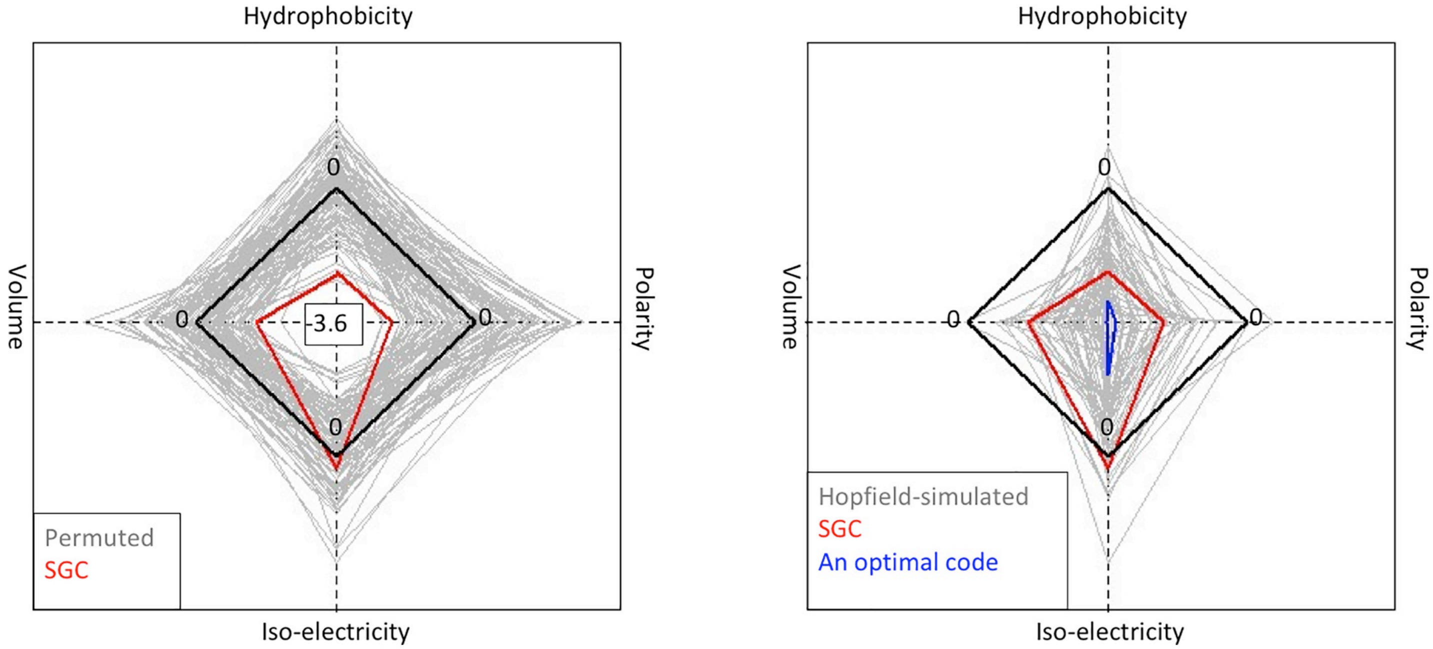

Fig 5. Hopfield network yielded genetic codes more optimal than SGC. (**Left Panel**) In this alternative display of error rates of genetic codes, each code is represented by a quadrilateral with vertices defined by the four errors. The black diamond represents the average errors of the permuted codes (n = 100). Note that the minimum error (-3.6 standard deviation from the mean) is placed at the center for all four parameters so that the smaller a quadrilateral condenses towards the center the more optimal a code is. (**Right Panel**) Gray quadrilaterals represent Hopfield-simulated codes (n = 52) optimized for polarity and volume (the same output used for Fig 4, 2nd row). The simulated codes have generally less errors than randomized codes. Many are more optimal than SGC, one of which is highlighted in blue and its coding pattern is shown with a codon wheel (Fig 1B).

https://doi.org/10.1371/journal.pone.0224552.g005

similar tRNAs recognizing physio-chemically similar amino acids (Fig 6). This autocorrelation is consistent with an early co-diversification for tRNA sequences and cognate amino acids.

Discussion

This study was motivated by the proposition of evolutionary connectionism that algorithmic learning could lead to self-optimized, adaptive, and robust living systems [10,11]. Using the origin of genetic code as a test case, we show that it is indeed plausible for an error-minimizing genetic code to emerge through a Hebbian learning process without natural selection playing a role at system levels.

TSP representations of genetic code

Our algorithm for finding optimal genetic codes consisted of two main steps. In the first step, we formulated code optimization as a combinatory optimization problem of finding the shortest paths traversing the 20 amino acids and the stop signal. This Traveling Salesman Problem (TSP) formulation of the genetic code naturally lend itself to be solved with the Hopfield neural network, which is an implementation of the Hebbian learning rule [43,44]. In the second step, we mapped the optimal tour positions emerged from the Hopfield network to a circular sequence of codon–a codon wheel (Fig 1)–that are based on biological constraints present in SGC (e.g., the codon degeneracy decreasing in the order of 3rd, 1st, and 2nd codon positions and the size distribution of synonymous codon blocks). Note that there was no guarantee that optimized genetic codes emerged from the Hopfield network would be as optimal as SGC. This could be seen in randomly permuted codes, most of which showed higher error rates

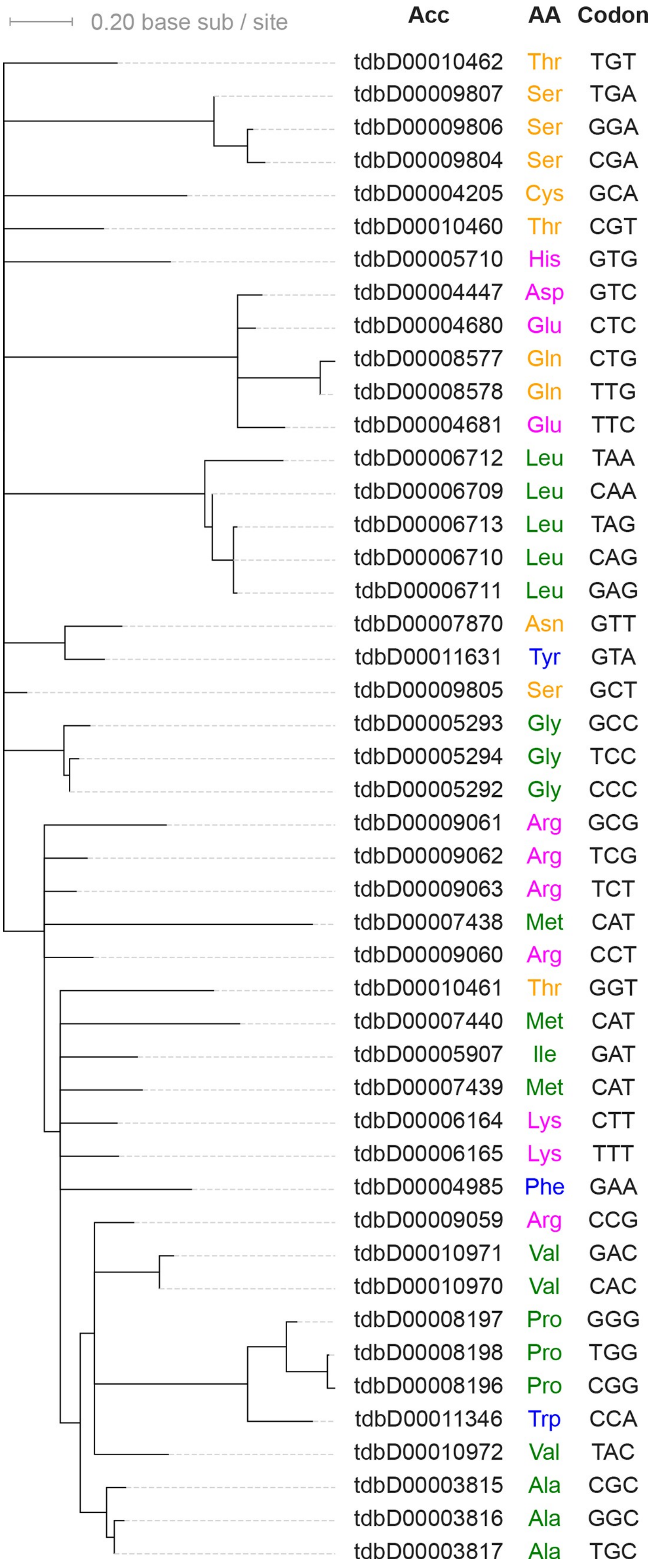
0.20 base sub / site
Acc AA Codon
tdbD00010462 Thr TGT
tdbD00009807 Ser TGA
tdbD00009806 Ser GGA
tdbD00009804 Ser CGA
tdbD00004205 Cys GCA
tdbD00010460 Thr CGT
tdbD00005710 His GTG
tdbD00004447 Asp GTC
tdbD00004680 Glu CTC
tdbD00008577 Gln CTG
tdbD00008578 Gln TTG
tdbD00004681 Glu TTC
tdbD00006712 Leu TAA
tdbD00006709 Leu CAA
tdbD00006713 Leu TAG
tdbD00006710 Leu CAG
tdbD00006711 Leu GAG
tdbD00007870 Asn GTT
tdbD00011631 Tyr GTA
tdbD00009805 Ser GCT
tdbD00005293 Gly GCC
tdbD00005294 Gly TCC
tdbD00005292 Gly CCC
tdbD00009061 Arg GCG
tdbD00009062 Arg TCG
tdbD00009063 Arg TCT
tdbD00007438 Met CAT
tdbD00009060 Arg CCT
tdbD00010461 Thr GGT
tdbD00007440 Met CAT
tdbD00005907 Ile GAT
tdbD00007439 Met CAT
tdbD00006164 Lys CTT
tdbD00006165 Lys TTT
tdbD00004985 Phe GAA
tdbD00009059 Arg CCG
tdbD00010971 Val GAC
tdbD00010970 Val CAC
tdbD00008197 Pro GGG
tdbD00008198 Pro TGG
tdbD00008196 Pro CGG
tdbD00011346 Trp CCA
tdbD00010972 Val TAC
tdbD00003815 Ala CGC
tdbD00003816 Ala GGC
tdbD00003817 Ala TGC

Fig 6. Molecular phylogeny of *Pyrococcus furiosus* tRNAs. Leaf nodes are labeled with the cognate amino acid names and the anticodon sequences. Amino acid names are colored according to an amino acid physio-chemical classification as in Fig 1 [18]. Only significantly supported branches (bootstrap value > = 0.7) are shown (see Methods). Phylogenetic autocorrelation with amino acid indices are all significant (Moran's I = 0.1973 with p = 8.1e-7 for polarity, I = 0.21517 with p = 1.6e-07 for hydrophobicity, I = 0.1593 with p = 6.0e-05 for volume, I = 0.1116 with p = 2.4e-3 for isoelectricity), suggesting early co-diversification between tRNAs and physio-chemistry of cognate amino acids.

https://doi.org/10.1371/journal.pone.0224552.g006

than SGC although generated with the same set of SGC constraints including codon-degeneracy, transition/transversion bias, and the size distribution of synonymous codons (Fig 2). More tellingly, plenty of codes optimized by the Hopfield network and mapped to the same codon wheel showed higher error rates than SGC (Fig 4).

Also using simulated annealing, DiGiulio *et al* [58] was able to achieve a code with an optimality ratio of 51.7% with respect to polarity, whereas we found a code with an optimality ratio of 48% with respect to polarity. Here we use DiGiulio's optimality ratio: $\{\Delta(Mean)-\Delta(SGC)\}/\{\Delta(Mean)-\Delta(Code)\}\times 100$, where Δ(Mean) is the average error in polarity associated with a set of random code, Δ(SGC) is the error in polarity in the SGC, and Δ(Code) is the error in polarity the given code. DiGiulio *et al* used simulated annealing to minimize error directly without a neural network. On average 89.2% of our codes optimized for polarity were better than SGC with respect to polarity. The method however does not always produce optimal codes, as the simulated annealing method to solve the traveling salesman problem only finds local minima, which sometimes improve on SGC and sometimes do not.

Together, these two algorithmic steps (i.e., the use of a Hopfield network and a codon wheel) allowed us to establish that genetic code as optimal as SGC emerge quickly and without natural selection. It is conceivable that these two steps be combined into a single TSP algorithm. We have represented the genetic code as a 64 amino acids (with repetition)-by-64 codons TSP and set the distances between the same amino acids as zero. In practice, however, optimal codes emerged from such a Hopfield network turned out to be not strictly comparable to SGC (and its random permutations) because the size distribution of synonymous codon blocks was not preserved. One way of preserving the codon block size distribution in SGC is to number the codon blocks sequentially according to a codon wheel (Fig 1) so that a 21 amino acid– 21 codon block TSP could be constructed. This representation would be equivalent to the two-step algorithm we used because the "SGC address" of a codon was precisely the position of the codon block containing this codon in a codon wheel.

Coevolving molecular network resembles a Hopfield network. TSP is perhaps the most studied combinatorial optimization problem for which Hopfield network is one of numerous heuristic searching algorithms [45,59]. Our choice of Hopfield network is motivated by its embodiment of the Hebbian learning rule, a major self-organizing principle in evolutionary learning [15]. Similar to Hopfield network being a computational implementation of the Hebbian rule (Eq 1), the molecular machinery associated with the genetic code may well be a biological analog of the Hopfield network. Mirroring a Hopfield neuron associating a city with a specific tour position, each macromolecule (e.g., an tRNA) associates a codon with a amino acid. Further resembling Hopfield neurons inter-connected with varying degrees of weights, members of a macromolecule gene family (e.g., tRNAs and aaRSs) are related to each other with varying degrees of phylogenetic distances (Fig 6). Quantitatively, the rate of increase in the weight (w_{ij}) of connection between two Hopfield neurons (i and j) with correlated (r) activities x_i and x_j could be expressed as: $dw_{ij}/dt{\sim}r(x_i,x_j)$ [43]. Similarly, a history of co-diversification among tRNAs and amino acids creates phylogenetic auto-correlation between gene sequences and amino acid physio-chemistry (Fig 6). The phylogenetic autocorrelation could be expressed as: $dL_{ij}/dt \sim d(x_i, x_j)$, where L_{ij} is the phylogenetic distance (i.e., tree length) between two paralogs and $d(x_i, x_j)$ is the codon or amino acid distance associated with the pair of paralogs. Our Hopfield and Hebbian interpretations of the origin of SGC are consistent with the codon capture hypothesis, which posits that primordial codon expansion followed a phylogenetic path of minimum changes in amino acid physio-chemistry or exchangeability [60–62]. Indeed, before developing his namesake network, Hopfield himself had proposed an origin of SGC through co-diversification of tRNA molecules and their binding specificities [63].

To summarize, the phylogenies of macromolecule gene families resemble Hopfield networks and may thus be considered a recapitulation of the self-optimizing process of the SGC.

Self-learning in evolution of multi-gene families

Molecular systems consisting of co-diversifying paralogs are not limited to the genetic code. In fact, genome and gene duplications coupled with neo-functionalization or sub-functionalization are a pervasive and predominant mechanism of evolutionary innovation [64–66]. For example, duplications of the and the α- and β-hemoglobin genes have led to novel capacities for binding oxygen and carbon dioxide [65]. Duplication and sub-functionalization of the homeobox genes contributed to body plan diversification in bilaterian animals [67]. Vertebrate olfactory sensing system consists of rapidly evolving members of the odorant receptors gene family [68]. Hopfield himself proposed a neural computational model akin to his namesake network that distributes odor recognition among a large number of sensory cells [69]. The Major Histocompatibility Complex (MHC) loci responsible for adaptive immunity in vertebrates consists of multi-gene families [70]. For the parasites, genomes of microbial pathogens such as *Trypanosoma* and *Borrelia* are enriched with duplicated surface antigen genes for defense against host immunity [71,72].

In each of these cases, a complex adaptive molecular system has evolved from phylogenetic co-diversification between genes and gene functions. The present work shows that Hopfield network offers a way to simulate and perhaps to design artificial self-optimizing genetic systems.

Acknowledgments

We thank Dr Lia Di for graphic assistance. This work was supported by Public Health Service grants AI139782 (to WQ) from the National Institute of Allergy and Infectious Diseases (NIAID) and the grant MD007599 (to Hunter College) from the National Institute on Minority Health and Health Disparities (NIMHD) of the National Institutes of Health (NIH), and grant 1517002 from the National Science Foundation (to Joao Xavier and WQ) of the United States of America The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceptualization: Weigang Qiu.

Data curation: Brian Sulkow.

Formal analysis: Oliver Attie, Brian Sulkow.

Funding acquisition: Weigang Qiu.

Investigation: Oliver Attie, Brian Sulkow, Chong Di, Weigang Qiu.

Methodology: Brian Sulkow, Weigang Qiu.

Project administration: Weigang Qiu.

Resources: Weigang Qiu.

Software: Oliver Attie, Brian Sulkow, Chong Di.

Validation: Weigang Qiu.

Writing – original draft: Oliver Attie, Weigang Qiu.

References

1. Darwin C. The Origin of Species. P. F. Collier & Son; 1909.
2. Wallace AR. Contributions to the Theory of Natural Selection. Macmillan and Co.; 1871.
3. Charlesworth D, Barton NH, Charlesworth B. The sources of adaptive variation. Proc Biol Sci. 2017; 284. https://doi.org/10.1098/rspb.2016.2864 PMID: 28566483
4. Huxley J. Evolution: the modern synthesis. Allen and Unwin; 1974.
5. Kimura M. The Neutral Theory of Molecular Evolution. Cambridge University Press; 1984.
6. Pak D, Du N, Kim Y, Sun Y, Burton ZF. Rooted tRNAomes and evolution of the genetic code. Transcription. 2018; 9: 137–151. https://doi.org/10.1080/21541264.2018.1429837 PMID: 29372672
7. Pigliucci M. Do we need an extended evolutionary synthesis? Evol Int J Org Evol. 2007; 61: 2743–2749. https://doi.org/10.1111/j.1558-5646.2007.00246.x PMID: 17924956
8. Smith JM, Szathmary E. The Major Transitions in Evolution. OUP Oxford; 1997.
9. Chaitin G. Proving Darwin: Making Biology Mathematical. Vintage Books; 2013.
10. Valiant L. Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World. Basic Books; 2013.
11. Watson RA, Mills R, Buckley CL, Kouvaris K, Jackson A, Powers ST, et al. Evolutionary Connectionism: Algorithmic Principles Underlying the Evolution of Biological Organisation in Evo-Devo, Evo-Eco and Evolutionary Transitions. Evol Biol. 2016; 43: 553–581. https://doi.org/10.1007/s11692-015-9358-z PMID: 27932852
12. Pak D, Kim Y, Burton ZF. Aminoacyl-tRNA synthetase evolution and sectoring of the genetic code. Transcription. 2018; 9: 205–224. https://doi.org/10.1080/21541264.2018.1467718 PMID: 29727262
13. Chastain E, Livnat A, Papadimitriou C, Vazirani U. Algorithms, games, and evolution. Proc Natl Acad Sci U S A. 2014; 111: 10620–10623. https://doi.org/10.1073/pnas.1406556111 PMID: 24979793
14. Papadimitriou C. Algorithms, complexity, and the sciences. Proc Natl Acad Sci U S A. 2014; 111: 15881–15887. https://doi.org/10.1073/pnas.1416954111 PMID: 25349382
15. Watson RA, Szathmáry E. How Can Evolution Learn? Trends Ecol Evol. 2016; 31: 147–157. https://doi.org/10.1016/j.tree.2015.11.009 PMID: 26705684
16. Hebb DO. The Organization of Behavior: A Neuropsychological Theory. Taylor & Francis; 2002.
17. Vey G. Gene coexpression as Hebbian learning in prokaryotic genomes. Bull Math Biol. 2013; 75: 2431–2449. https://doi.org/10.1007/s11538-013-9900-z PMID: 24078338
18. Nelson DL, Cox MM. Lehninger Principles of Biochemistry. 4th ed. Macmillan; 2005.
19. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982; 157: 105–132. https://doi.org/10.1016/0022-2836(82)90515-0 PMID: 7108955
20. Novozhilov AS, Wolf YI, Koonin EV. Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. Biol Direct. 2007; 2: 24. https://doi.org/10.1186/1745-6150-2-24 PMID: 17956616
21. 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. 2019; 464: 21–32. https://doi.org/10.1016/j.jtbi.2018.12.030 PMID: 30579955
22. Santos J, Monteagudo A. Simulated evolution applied to study the genetic code optimality using a model of codon reassignments. BMC Bioinformatics. 2011; 12: 56. https://doi.org/10.1186/1471-2105-12-56 PMID: 21338505
23. Wnętrzak M, Błażej P, Mackiewicz D, Mackiewicz P. The optimality of the standard genetic code assessed by an eight-objective evolutionary algorithm. BMC Evol Biol. 2018; 18: 192. https://doi.org/10.1186/s12862-018-1304-0 PMID: 30545289

24. Błażej P, Wnętrzak M, Mackiewicz P. The role of crossover operator in evolutionary-based approach to the problem of genetic code optimization. Biosystems. 2016; 150: 61–72. https://doi.org/10.1016/j.biosystems.2016.08.008 PMID: 27555085

25. Błażej P, Wnętrzak M, Mackiewicz D, Mackiewicz P. Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm. PloS One. 2018; 13: e0201715. https://doi.org/10.1371/journal.pone.0201715 PMID: 30092017

26. Freeland SJ, Knight RD, Landweber LF, Hurst LD. Early fixation of an optimal genetic code. Mol Biol Evol. 2000; 17: 511–518. https://doi.org/10.1093/oxfordjournals.molbev.a026331 PMID: 10742043

27. Goldenfeld N, Biancalani T, Jafarpour F. Universal biology and the statistical mechanics of early life. Philos Transact A Math Phys Eng Sci. 2017;375. https://doi.org/10.1098/rsta.2016.0341 PMID: 29133441

28. José MV, Zamudio GS, Morgado ER. A unified model of the standard genetic code. R Soc Open Sci. 2017; 4: 160908. https://doi.org/10.1098/rsos.160908 PMID: 28405378

29. José MV, Govezensky T, García JA, Bobadilla JR. On the evolution of the standard genetic code: vestiges of critical scale invariance from the RNA world in current prokaryote genomes. PloS One. 2009; 4: e4340. https://doi.org/10.1371/journal.pone.0004340 PMID: 19183813

30. Koonin EV. Frozen Accident Pushing 50: Stereochemistry, Expansion, and Chance in the Evolution of the Genetic Code. Life Basel Switz. 2017; 7. https://doi.org/10.3390/life7020022 PMID: 28545255

31. Koonin EV, Novozhilov AS. Origin and Evolution of the Universal Genetic Code. Annu Rev Genet. 2017; 51: 45–62. https://doi.org/10.1146/annurev-genet-120116-024713 PMID: 28853922

32. Di Giulio M. An Autotrophic Origin for the Coded Amino Acids is Concordant with the Coevolution Theory of the Genetic Code. J Mol Evol. 2016; 83: 93–96. https://doi.org/10.1007/s00239-016-9760-x PMID: 27743002

33. Massey SE. The neutral emergence of error minimized genetic codes superior to the standard genetic code. J Theor Biol. 2016; 408: 237–242. https://doi.org/10.1016/j.jtbi.2016.08.022 PMID: 27544417

34. Pak D, Root-Bernstein R, Burton ZF. tRNA structure and evolution and standardization to the three nucleotide genetic code. Transcription. 2017; 8: 205–219. https://doi.org/10.1080/21541264.2017.1318811 PMID: 28632998

35. Kim Y, Kowiatek B, Opron K, Burton ZF. Type-II tRNAs and Evolution of Translation Systems and the Genetic Code. Int J Mol Sci. 2018; 19. https://doi.org/10.3390/ijms19103275 PMID: 30360357

36. Opron K, Burton ZF. Ribosome Structure, Function, and Early Evolution. Int J Mol Sci. 2018; 20. https://doi.org/10.3390/ijms20010040 PMID: 30583477

37. Chin JW. Expanding and reprogramming the genetic code. Nature. 2017; 550: 53–60. https://doi.org/10.1038/nature24031 PMID: 28980641

38. Xue H, Wong JT-F. Future of the Genetic Code. Life. 2017; 7: 10. https://doi.org/10.3390/life7010010 PMID: 28264473

39. de Oliveira LL, de Oliveira PSL, Tinós R. A multiobjective approach to the genetic code adaptability problem. BMC Bioinformatics. 2015; 16: 52. https://doi.org/10.1186/s12859-015-0480-9 PMID: 25879480

40. Santos J, Monteagudo Á. Inclusion of the fitness sharing technique in an evolutionary algorithm to analyze the fitness landscape of the genetic code adaptability. BMC Bioinformatics. 2017; 18: 195. https://doi.org/10.1186/s12859-017-1608-x PMID: 28347270

41. Tlusty T. A colorful origin for the genetic code: information theory, statistical mechanics and the emergence of molecular codes. Phys Life Rev. 2010; 7: 362–376. https://doi.org/10.1016/j.plrev.2010.06.002 PMID: 20558115

42. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci. 1982; 79: 2554–2558. https://doi.org/10.1073/pnas.79.8.2554 PMID: 6953413

43. MacKay DJC. Information Theory, Inference and Learning Algorithms. Cambridge University Press; 2003.

44. Hopfield JJ, Tank DW. Computing with neural circuits: a model. Science. 1986; 233: 625–633. https://doi.org/10.1126/science.3755256 PMID: 3755256

45. Potvin J-Y. State-of-the-Art Survey—The Traveling Salesman Problem: A Neural Network Perspective. ORSA J Comput. 1993; 5: 328–348. https://doi.org/10.1287/ijoc.5.4.328

46. Bout DEV den Miller TK. Improving the performance of the Hopfield-Tank neural network through normalization and annealing. Biol Cybern. 1989; 62: 129–139. https://doi.org/10.1007/BF00203001

47. Haig D, Hurst LD. A quantitative measure of error minimization in the genetic code. J Mol Evol. 1991; 33: 412–417. https://doi.org/10.1007/bf02103132 PMID: 1960738

48. Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 2009; 37: D159–D162. https://doi.org/10.1093/nar/gkn772 PMID: 18957446

49. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. PloS One. 2010; 5: e9490. https://doi.org/10.1371/journal.pone.0009490 PMID: 20224823

50. Hernández Y, Bernstein R, Pagan P, Vargas L, McCaig W, Ramrattan G, et al. BpWrapper: BioPerl-based sequence and tree utilities for rapid prototyping of bioinformatics pipelines. BMC Bioinformatics. 2018; 19: 76. https://doi.org/10.1186/s12859-018-2074-9 PMID: 29499649

51. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinforma Oxf Engl. 2004; 20: 289–290.

52. Gittleman JL, Kot M. Adaptation: Statistics and a Null Model for Estimating Phylogenetic Effects. Syst Biol. 1990; 39: 227–241. https://doi.org/10.2307/2992183

53. Moran PAP. Notes on Continuous Stochastic Phenomena. Biometrika. 1950; 37: 17–23. https://doi.org/10.2307/2332142 PMID: 15420245

54. Dray S, Dufour A-B. The ade4 Package: Implementing the Duality Diagram for Ecologists. J Stat Softw. 2007; 22. https://doi.org/10.18637/jss.v022.i04

55. Zamudio GS, José MV. On the Uniqueness of the Standard Genetic Code. Life Basel Switz. 2017; 7. https://doi.org/10.3390/life7010007 PMID: 28208827

56. Rodin SN, Rodin AS. On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases. Heredity. 2008; 100: 341–355. https://doi.org/10.1038/sj.hdy.6801086 PMID: 18322459

57. Saks ME, Sampson JR, Abelson J. Evolution of a transfer RNA gene through a point mutation in the anticodon. Science. 1998; 279: 1665–1670. https://doi.org/10.1126/science.279.5357.1665 PMID: 9497276

58. Massimo DiGiulio M.Rosaria Capobianco, Mario Medugno. On the optimization of the physicochemical distances between amino acids in the evolution of the genetic code. J Theor Biol. 1994; 168: 31–41. https://doi.org/10.1006/jtbi.1994.1085

59. Applegate D, Bixby R, Chvátal V, Cook W. The Traveling Salesman Problem. In: Princeton University Press [Internet]. 2007 [cited 4 Mar 2019]. Available: https://press.princeton.edu/titles/8451.html

60. Davis BK. Evolution of the genetic code. Prog Biophys Mol Biol. 1999; 72: 157–243. PMID: 10511799

61. Osawa S, Jukes TH. Codon reassignment (codon capture) in evolution. J Mol Evol. 1989; 28: 271–278. https://doi.org/10.1007/bf02103422 PMID: 2499683

62. Stoltzfus A, Yampolsky LY. Amino acid exchangeability and the adaptive code hypothesis. J Mol Evol. 2007; 65: 456–462. https://doi.org/10.1007/s00239-007-9026-8 PMID: 17896070

63. Hopfield JJ. Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence, and kinetic proofreading. Proc Natl Acad Sci. 1978; 75: 4334–4338. https://doi.org/10.1073/pnas.75.9.4334 PMID: 279919

64. Dittmar K, Liberles D. Evolution after Gene Duplication. John Wiley & Sons; 2011.

65. Graur D. Molecular and Genome Evolution. Sinauer; 2015.

66. Ohno S. Evolution by Gene Duplication. Springer Science & Business Media; 2013.

67. Holland PWH. Did homeobox gene duplications contribute to the Cambrian explosion? Zool Lett. 2015; 1: 1. https://doi.org/10.1186/s40851-014-0004-x PMID: 26605046

68. Hoover KC. Evolution of olfactory receptors. Methods Mol Biol Clifton NJ. 2013; 1003: 241–249. https://doi.org/10.1007/978-1-62703-377-0_18 PMID: 23585047

69. Hopfield JJ. Odor space and olfactory processing: Collective algorithms and neural implementation. Proc Natl Acad Sci. 1999; 96: 12506–12511. https://doi.org/10.1073/pnas.96.22.12506 PMID: 10535952

70. Naz R, Tahir S, Abbasi AA. An insight into the evolutionary history of human MHC paralogon. Mol Phylogenet Evol. 2017; 110: 1–6. https://doi.org/10.1016/j.ympev.2017.02.015 PMID: 28249742

71. Norris SJ. vls Antigenic Variation Systems of Lyme Disease Borrelia: Eluding Host Immunity through both Random, Segmental Gene Conversion and Framework Heterogeneity. Microbiol Spectr. 2014; 2. https://doi.org/10.1128/microbiolspec.MDNA3-0038-2014 PMID: 26104445

72. Taylor JE, Rudenko G. Switching trypanosome coats: what's in the wardrobe? Trends Genet TIG. 2006; 22: 614–620. https://doi.org/10.1016/j.tig.2006.08.003 PMID: 16908087