SCIENTIFIC REPORTS
nature research

OPEN

Adaptive Properties of the Genetically Encoded Amino Acid Alphabet Are Inherited from Its Subsets

Received: 8 January 2019
Accepted: 8 July 2019
Published online: 28 August 2019

Melissa Ilardo[1], Rudrarup Bose[2], Markus Meringer[3], Bakhtiyor Rasulev[4], Natalie Grefenstette[5], James Stephenson[6,7], Stephen Freeland[8], Richard J. Gillams[9,10], Christopher J. Butch[9,11,12] & H. James Cleaves II[9,12,13]

Life uses a common set of 20 coded amino acids (CAAs) to construct proteins. This set was likely canonicalized during early evolution; before this, smaller amino acid sets were gradually expanded as new synthetic, proofreading and coding mechanisms became biologically available. Many possible subsets of the modern CAAs or other presently uncoded amino acids could have comprised the earlier sets. We explore the hypothesis that the CAAs were selectively fixed due to their unique adaptive chemical properties, which facilitate folding, catalysis, and solubility of proteins, and gave adaptive value to organisms able to encode them. Specifically, we studied *in silico* hypothetical CAA sets of 3–19 amino acids comprised of 1913 structurally diverse α-amino acids, exploring the adaptive value of their combined physicochemical properties relative to those of the modern CAA set. We find that even hypothetical sets containing modern CAA members are especially adaptive; it is difficult to find sets even among a large choice of alternatives that cover the chemical property space more amply. These results suggest that each time a CAA was discovered and embedded during evolution, it provided an adaptive value unusual among many alternatives, and each selective step may have helped bootstrap the developing set to include still more CAAs.

There is mounting evidence that the modern genetically coded amino acid (CAA) alphabet, used nearly universally by all living organisms on Earth, is highly optimized for a number of features including codon mapping and coverage of chemical space (*cf.* refs[1–3], and references therein). A "chemical space" is defined as a set of compounds, hypothetical or actual, which fulfill a given set of criteria, such as molecular formula, chemical property or chemical substructure (*e.g.*, molecules containing the α-amino acid substructure)[4–8].

The CAAs and their corresponding codon mapping are generally believed to have been fixed by the time of origin of the Last Universal Common Ancestor (LUCA)[9]. However, the complete set of 20 likely represents the product of step-wise growth from an earlier, simpler alphabet. For example, the lower molecular weight and structurally simpler amino acids (*e.g.*, glycine (G), alanine (A) and serine (S)) have been argued to have been made

[1]University of Utah Hematology, UC Berkeley Integrative Biology, George and Dolores Eccles Institute of Human Genetics, 15 N 2030 E, Room: 3240, Salt Lake City, UT, 84112, USA. [2]National Institute of Science Education and Research Bhubaneswar, P.O. Jatni, Khurda, 752050, Odisha, India. [3]German Aerospace Center (DLR), Earth Observation Center (EOC), Münchner Straße 20, 82234, Oberpfaffenhofen-Wessling, Germany. [4]Department of Coatings and Polymeric Materials, North Dakota State University, Fargo, ND, 58108, USA. [5]Department of Chemistry, University College London, 20 Gordon street, London, WC1H 0AJ, UK. [6]European Molecular Biology Laboratory–European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. [7]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. [8]University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD, 21250, USA. [9]Earth-Life Science Institute, Tokyo Institute of Technology, 2-12-1-IE-1 Ookayama, Meguro-ku, Tokyo, 152-8550, Japan. [10]Electronics and Computer Science, Institute for Life Sciences, University of Southampton, Southampton, SO17 1BJ, UK. [11]Department of Chemistry, Emory University, 1515 Dickey Drive, Atlanta, GA, USA. [12]Blue Marble Space Institute for Science, 1001 4th Ave, Suite 3201, Seattle, WA, 98154, USA. [13]Institute of Advanced Study, 1 Einstein Drive, Princeton, NJ, 08540, USA. Correspondence and requests for materials should be addressed to H.J.C. (email: henderson.cleaves@gmail.com)

Set size (no. AAs)	# CAA Set Combinations[a]	# CAA Maximal Sets[b]	% Better XAA Sets in 10^8 Trials
3	1,140	509	0.499
4	4,845	1,250	0.242
5	15,504	2,151	0.161
6	38,760	2,875	4.65×10^{-2}
7	77,520	3,044	1.68×10^{-2}
8	125,970	3,177	7.52×10^{-3}
9	167,960	2,787	3.58×10^{-3}
10	184,756	2,160	1.23×10^{-3}
11	167,960	1,566	3.77×10^{-4}
12	125,970	1,181	1.59×10^{-4}
13	77,520	799	1.75×10^{-4}
14	38,760	504	1.68×10^{-4}
15	15,504	289	2.10×10^{-4}
16	4,845	165	3.24×10^{-4}
17	1,140	74	2.07×10^{-4}
18	190	28	6.47×10^{-5}
19	20	8	1.09×10^{-5}
20[c]	1	1	6×10^{-6}

Table 1. Summary of results of better XAA sets as a function of set size. [a]This number is derived from the formula for binomial coefficients, see SI Section 1. [b]This is the number of maximal sets, see methods section. [c]There is only one possible set, which is also maximal, of the 20 CAAs, and only 6 better XAA sets were found in a previous study, of which several contained CAAs[3].

available from abiotic synthesis[10,11], setting the stage for their incorporation into nascent biological systems[11–13]. Additionally, by most assessments, tryptophan (W) is likely to have been the last amino acid fixed into the coded set[11,14–16], and there are computational and physical chemical data suggesting that some of the last additions were those with the greatest redox activity, which became fixed as Earth's atmosphere became oxygenated[17].

As there are so many possible metrics of optimality, we investigated, for the first time the optimality of the extraordinarily large number of possible coded and alternative amino acid sets (see Table 1, which is explained in more detail below) according to the metrics on which it has been claimed the amino acid set may have been selected. The 20 CAAs' repertoire seems to reflect the requirements of providing enough structural diversity that proteins derived from these amino acids are able to define unique three-dimensional shapes[18], and able to produce more adaptive proteins (*e.g.*, those whose catalytic properties improve the function of the cell which hosts them, whether it be by improving turnover number, tuning it to the flux of intracellular intermediates, or by other means) when the repertoire of amino acids is enlarged[19]. These general ideas have been subsequently refined into mounting evidence for the detailed claim that a combination of size, pK_a and hydrophobicity seem to combine as a good first approximation of the CAA's value to natural selection[2,3,20].

Biology produces several α-amino acids in the process of synthesizing the CAAs that are not themselves encoded (for example ornithine, diaminopimelic acid and homocysteine). Further, other amino acids are produced in the process of synthesizing various biochemicals that are also not included in the coded set (*e.g.*, β-alanine and canavanine). It seems reasonable that several mechanisms led to the evolutionary selection of the CAAs as they were made available during the emergence of metabolism, and their incorporation into the genetic code was determined based on factors besides prebiotic availability[14,21–23]. From a different perspective, biological engineering has shown that a wide variety of amino acids can be removed, replaced or substituted in coded proteins[24,25], though to our knowledge a complete organism-wide replacement of a CAA has not yet proven possible (*e.g.*, compare refs[26,27]).

It has so far not been proven possible to construct an organism that can survive with less than the 20 CAAs in its total proteome[28], and thus still not yet possible to experimentally explore metabolism(s) based on fewer than 20 amino acids, due to the high degree of interconnectivity of biological structures, components and processes[29,30]. While many modern proteins do not contain all 20 CAAs (see Figures SI1, SI2), single amino acid types have been systematically removed from specific proteins (*e.g.*, ref.[31]), and non-biological amino acids incorporated into others (*cf.* ref.[32]). Though it appears that it would be difficult to explore the biological implications of reduced CAA sets in modern engineered organisms, chemoinformatics offers methods to explore the possible consequences of unique CAA set composition trajectories during pre-LUCA biochemical evolution. We herein used computational approaches to estimate the adaptive value of potential coded α-amino acid subsets which biology might have explored during biochemical evolution.

The CAAs are distinguished from theoretical alternatives by their properties as a set, specifically their coverage of chemical space[3,33]. We therefore assumed that, in the growth of the amino acid alphabet, individual α-amino acids were selected based not only on their own intrinsic physicochemical properties but also on the way that complemented a pre-existing set. We then investigated an exhaustive set of possible subsets, evaluating them in terms of their coverage of chemical space relative to theoretical α-amino acid sets sampled from a larger virtual library of 1913 amino acids[3] constructed using molecular structure generation software[34–36]. Although

other factors, including biosynthetic availability, could have contributed to shaping the growth of the modern near-universal CAA alphabet, we sought to establish whether the CAAs can be distinguished as optimal using a minimal set of parameters, namely size, hydrophobicity and pK_a[3].

Materials and Methods

We used a reference library consisting of the 20 CAAs and 1893 other theoretically possible, computationally constructed α-amino acids from a previous study[3]. We refer to this set of molecules as the *xeno amino acids* (XAAs). Given that the library from which the XAA sets were selected includes the 20 CAAs, some XAA sets also contain CAAs. To describe the properties of the molecules in the computed sets, we used the same descriptors previously reported: van der Waals volume (V_{vdw}), logarithmic acid dissociation constant (pK_a, considered specifically over the range from 2–14 here) and partition coefficient (logP), which were selected based on their ability to characterize the functional chemistry space of α-amino acids (*cf.* refs[3,37]). V_{vdw} is simply a measure of the volume of space occupied by the amino acid, which is expected to play a role in mediating steric interactions. LogP describes the affinity of a molecule to a hydrophilic or hydrophobic solvent. In the context of protein structure, this is an important factor in protein folding, and is essential for the heterogeneous nature of protein surface potentials, which is essential for catalysis. pK_a describes the pH at the mid-point of a proton transfer by the amino acid side chain, which influences the charged state of an amino acid residue. This in turn affects a host of interactions within and among proteins and their substrates.

While thousands of additional descriptors exist[38], we selected what we considered to be fundamental properties that define molecular interactions. This selection was made to minimize bias introduced by considering instead the properties through which we functionally characterize amino acids in modern biochemical contexts. Additionally, these properties are reliably predicted and quantified through chemical property prediction software, an important consideration given the theoretical and computational nature of our data set. This analysis is exploratory, requiring broad, and simplified, choices of molecular descriptors. A previous publication[20] offers a specific investigation that justifies the descriptors relating to size, pK_a and hydrophobicity, particularly in the light of older work[39] that has been built on productively[40,41]. It is striking that such simple metrics are able to classify amino acid chemical space so effectively[2,3].

We used the definition of optimality that was previously introduced to test the CAA alphabet, as described in Ilardo *et al.*[3], namely that more "optimal" sets are those that have *broader range and/or evenness* of coverage of chemical space. "Better" or "more optimal" sets are those that cover chemical space in the three descriptor categories (V_{vdw}, pK_a and logP) both more broadly and evenly than a comparison set. For calculation of range, evenness and coverage, please see Section 1 of the supporting information (SI) and for further discussion see refs[2,34].

We first generated a comprehensive list of all possible *k*-subsets of the CAAs ranging in set size from three (the smallest size that allows calculation of evenness) to 19 amino acids (as 20 was a previously calculated benchmark[3]) and their range and evenness values in the dimensions of size, pK_a and hydrophobicity. A part of this list showing the computed values for set size 19 is given in Table SI1. From these, we computed sets of maximal coverage (hereafter referred to as *maximal sets*), defined as sets for which there exists no other set of the same size that is "better," with respect to range and evenness in coverage of chemical property space with respect to V_{vdw}, pK_a and logP. These maximal sets are maxima of the partial order introduced by the six dimensions (range and evenness in size, pK_a, and hydrophobicity) of the *k*-subsets ($k > 2$) of CAAs[42]. Partial orders are a frequently used concept in chemistry whenever there is no total order available[43,44]. Section 2 of the SI gives a formal definition of our partial order and a brief introduction to Hasse diagrams, which can be used as graphical representations of partially ordered sets. Figure SI3 shows a Hasse diagram of the 19-subsets of the CAAs.

To quantify the optimality of the maximal subsets of CAAs, we needed an appropriate comparison. As the number of combinatorially possible sets of the XAAs that can be constructed grows exponentially with set size (see Figure SI4), we used a statistical sampling approach and generated 10^8 random sets for each set size selected from the XAA reference library. By repeatedly sampling different set sizes (10^7–10^9), we were able to establish that a sample size of 10^8 sets provides asymptotic results (*i.e.*, larger sample sizes do not produce different results) that are computationally tractable (*e.g.*, returning results on the timescale of ~ 1 day using our available computational resources; sampling 10^9 sets took on the order of a few weeks). Each of these XAA sets was then compared to the maximal CAA subsets, which are again those for which the range and/or evenness in one of the three optimality criteria are not surpassed by another CAA set for that set size. In each comparison the set with more optimal chemical space coverage (as previously described) was designated a "better set". We recorded the number and composition of each better set identified from the XAA sets. In order to gain high certainty on the computational correctness of our results, the algorithms were implemented independently once in Matlab and once in Python. They were run on a Hewlett-Packard Pavilion Windows PC with an Intel(R) core(™) i7-7700HQ 2.8 GHz CPU 32 GB RAM, using up to eight cores in parallel.

Z-values offer a way to compare results to a normal population. A z-value is **a** measure of how many standard deviations below or above the population mean a raw score is. Negative Z-values fall to the left of the normal distribution curve and positive ones to the right. The Z-value formula for a sample is:

$$z = (x - \mu)/\sigma$$

where x is the measured value, μ is the mean and σ the standard deviation. Z-values shown were computed using Microsoft Excel.

For the extant protein analysis, we considered the 554,514 proteins annotated in the manually curated Swissprot database[45] as of 2017, which included a total of 198,500,435 unique amino acid residues (Figure SI1). Of the proteins containing less than 20 unique CAAs (184,929) we identified those CAAs that were absent. Proteins with multiple absences contribute to each of the relevant counts. Figure SI2 shows the proportion of database protein sequences lacking each CAA.

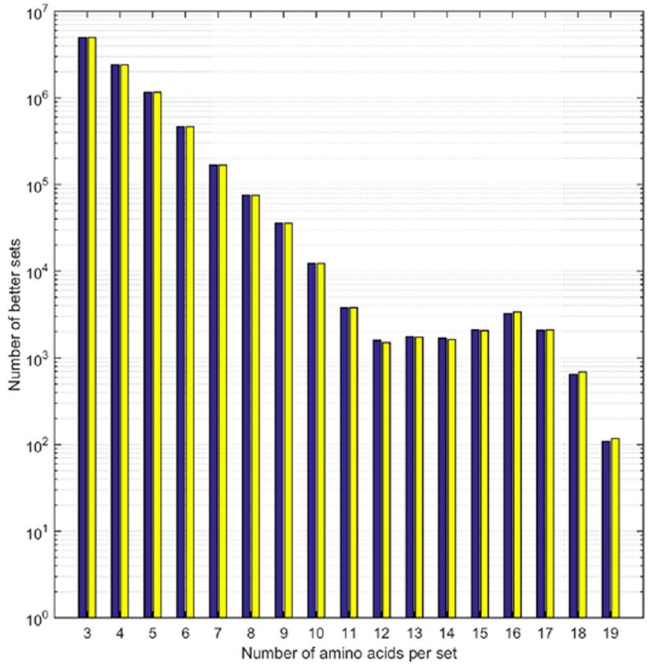

Figure 1. Semi-log plot showing the results of two 10^8 samplings (yellow and blue bars) for better XAA sets of a given set size (shown on the x-axis) from the XAA library. The number of better XAA sets decreases approximately logarithmically with the exception of sets of size 13 to 18.

Results

In order to test the adaptive value of the CAAs at smaller set sizes, we first measured the number of better sets, as described in the methods section, for each set size. The results, listed in Table 1, highlight the significant adaptive properties of the CAAs at even the smallest set size.

For example, we find that already for a set size of three amino acids only about 5% of random XAA sets are "better" than at least one of the maximal CAA sets. Interestingly, the number of better sets roughly follows a trend of logarithmic decrease with set size (Fig. 1), consistent with the results reported in Ilardo *et al.*[3].

There is a slight anomaly between set sizes of 13–18. This deviation from monotonic exponential decline in better sets was intriguing, and we thus conducted further tests to distinguish between the possibility that it was an artifact of the 10^8 sample size (*i.e.*, a sampling error since there are generally vastly more combinations possible than tested) versus other possible explanations (*e.g.*, the chemical similarity of some CAA's such as Ile and Leu, which make the "landscape of the better" relatively flat in places).

Specifically, to do this we repeated the calculation with the extremely rare better 20-amino acid member XAA sets 1 and 2 from Fig. 3 of Ilardo *et al.*[3]. The shape of the deviation was consistent (Figure SI5), though interestingly the better XAA sets were found to have fewer better-set combinations than strict CAA sets in random samples, suggesting there may be theoretical CAA better sets which evolution did not find. Regardless, this result suggests that the deviation is indeed a function of the properties of the CAAs (and those of the members of the better XAA sets) *as ensembles* relative to the properties of the entire XAA library's properties.

We further examined whether the adaptive contribution of the CAAs to XAA better sets could be detected at the level of individual amino acids. We measured the frequency of representation of CAAs in better sets also containing XAAs and compared this to the frequency of non-CAA XAAs in those better sets. The distribution of these values are shown in the box plots in Fig. 2.

Some of the CAAs are noticeably statistically overrepresented even at the smallest set sizes, and this adaptive advantage increases dramatically as set size grows. In Fig. 2, Met, Pro and Phe consistently rank among the most frequent CAAs in better sets, with Gly being abundant in set sizes of 3 and 4. By set sizes of 17, at least one CAA is represented in almost 30% of better sets, and many contain more than one (Figure SI6).

Lastly, we looked at the individual CAAs in the maximal and better sets to determine, whether particular amino acids exert a larger influence than others on our interpretation of adaptive properties for the subsets. Figure 3 shows the distribution of the CAAs in the better and maximal sets.

Discussion

It is unknown which set of amino acids was used before the advent of LUCA, in which the modern CAA set had already been adopted. However, by computing all combinations, we can explore the adaptive potential of many possible subsets. Assuming that early in CAA selection, each CAA addition created, in effect, a "negative chemical property space," in analogy with the term "negative space" as used in the art world – each occupation of the real implies how the occupation of the potential could be occupied. This negative space is then more likely to be filled by a new amino acid that would make the resulting new CAA set more adaptive, then somewhere in the CAA maximal sets (as defined in the Methods section) computed here there are likely to exist representations of

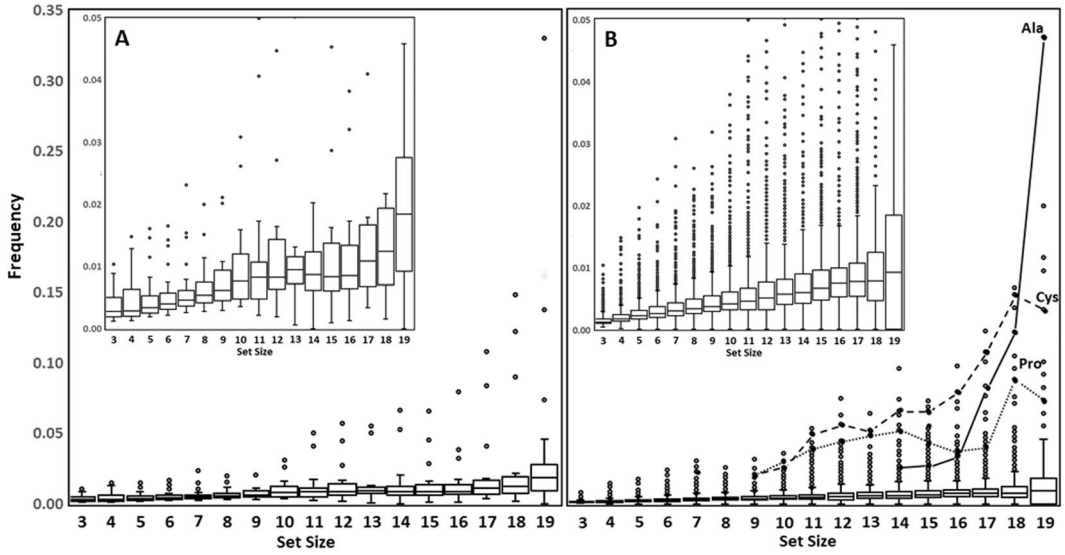

Figure 2. Box plots showing the relative frequency of (**A**) CAAs and (**B**) XAAs in better sets. Boxes extend from the lower to upper quartile values of the data, with a line at the median, and whiskers extending to contain 95% of the data. Zoomed insets help to show that the comparison of (**A**,**B**) reveals that the median values for maximal CAA sets are always higher than those of the corresponding XAA set sizes. In (**A**) all top outlier data points represent Met in set sizes 16–19. The connected data points in Figure B highlight the anomalously high frequency of the CAAs, Ala, Cys and Pro, in better XAA sets of larger set size.

true earlier (however weakly encoded) alphabets. The number of XAA sets identified that appear more adaptive than even one such maximal CAA set is therefore strikingly low. This suggests the CAAs contain members that are directionally optimal with respect to their chemical properties, regardless of the biosynthetic pathways that produce them.

We can presently only infer the composition of amino acid alphabets that life used pre-LUCA. Therefore, we also examined the frequency of individual CAAs in better XAA sets, as this can give an indication of which amino acids contribute to the adaptive properties exhibited by smaller ($<$20) amino acid alphabets of different sizes. We examined this frequency in two contexts, the frequency of particular amino acids in maximal sets, which only contain CAAs (Figs 3A and 3B), or better sets, where the possible set members (and therefore the context in which they appear adaptive) include all XAAs (Fig. 3C). Interestingly, there are differences in how adaptive the CAAs appear relative to the context in which they are examined (Fig. 3D). For example, Trp occurs infrequently in small size sets ($n \leq 10$) selected from all XAAs, but occurs frequently in small-sized maximal sets selected from the CAAs. One interpretation is that although W expands the chemical space of CAAs, many other, theoretically possible amino acids could have provided a similar advantage. Conversely, Met appears to be advantageous in small set sizes compared to the possibilities within the XAAs. However, it is not until larger set sizes that Met begins to become adaptive when looking at the coded set. This could indicate that Met does not offer an enormous adaptive advantage to the encoded set as a whole, however, there are very few possibilities within the XAA library that have similar chemical properties to Met, and therefore Met is uniquely advantageous to encoded sets.

For the contribution of the CAAs to making sets maximal (Fig. 3A), several interesting trends are evident. For example, Gly, Asn, Arg and Trp play large roles in making sets maximal at small set sizes, while several CAAs (*e.g.*, Ser, Pro, Val and Lys) play large roles in intermediary set sizes, and others in relatively larger ones (Cys, Gln, Met, His, Phe and Tyr). The last cohort roughly corresponds to various explanations offered for the order of incorporation of the CAAs (*e.g.*,[11,13,17]) to the modern CAA set. For the contribution of CAAs to constructing better XAA sets (Fig. 3C), some of the same trends are evident, *e.g.*, that of Gly for small sets and that of Met for larger ones, as well as the general utility of Pro, Asn and Phe.

The difference of these contributions is shown in Fig. 3D, which highlights the importance of the occurrence of other CAAs to make any given CAA appear to be a large contributor to set adaptiveness by these criteria. Here, white and dark blue regions highlight instances where another amino acid is likely required to bring out the adaptive value of an added CAA. These likely occur when a given CAA has some property that is an outlier in property space, which underscores our general model.

The structures of the XAAs represented most frequently in better sets, accumulated over all set sizes, are shown in Figure SI7 together with their frequencies. Some appear frequently independent of set size, in particular the two dehydroproline analogues (Figure SI7a,c). Others, like 2-amino-3,5-hexadienoic acid (Figure SI7e), are only frequent at larger set size (in this case 14–19), while still others like 3-aminoleucine are only frequent among small (3–9) set sizes. The nine structures shown in Figure SI7 are presented using an arbitrary frequency cut-off simply for the sake of exposition, but the frequency with which certain structures appear may point to the potential use of these similarity metrics to introduce novel XAAs into coded proteins.

A Frequency of CAAs in Maximal Sets

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
G	0.18	0.28	0.31	0.36	0.39	0.36	0.36	0.38	0.39	0.44	0.44	0.51	0.63	0.67	0.76	0.82	0.88
A	0.16	0.24	0.28	0.33	0.40	0.45	0.50	0.58	0.67	0.68	0.70	0.72	0.72	0.75	0.80	0.82	0.88
S	0.17	0.22	0.27	0.35	0.44	0.57	0.67	0.76	0.87	0.88	0.89	0.90	0.90	0.92	0.91	0.93	1.00
P	0.17	0.28	0.41	0.49	0.56	0.66	0.74	0.82	0.86	0.90	0.96	1.00	1.00	1.00	1.00	1.00	1.00
V	0.14	0.23	0.32	0.40	0.54	0.69	0.78	0.84	0.91	0.94	0.93	0.91	0.92	0.93	0.91	0.86	0.88
T	0.12	0.13	0.13	0.16	0.17	0.20	0.18	0.18	0.21	0.29	0.41	0.53	0.66	0.78	0.84	0.93	1.00
C	0.15	0.20	0.26	0.34	0.41	0.47	0.55	0.62	0.67	0.74	0.81	0.91	0.97	1.00	1.00	1.00	1.00
I	0.12	0.14	0.18	0.20	0.22	0.25	0.27	0.25	0.21	0.20	0.20	0.27	0.38	0.56	0.68	0.75	0.88
L	0.11	0.12	0.14	0.16	0.16	0.20	0.20	0.19	0.20	0.27	0.35	0.42	0.50	0.59	0.65	0.71	0.88
N	0.19	0.25	0.33	0.37	0.38	0.41	0.46	0.51	0.55	0.60	0.61	0.63	0.71	0.78	0.85	0.96	1.00
D	0.12	0.14	0.17	0.19	0.22	0.24	0.26	0.25	0.24	0.27	0.29	0.31	0.29	0.28	0.43	0.64	0.88
E	0.15	0.18	0.21	0.26	0.30	0.33	0.35	0.39	0.43	0.46	0.49	0.56	0.65	0.74	0.86	0.96	1.00
K	0.14	0.15	0.25	0.34	0.44	0.53	0.64	0.73	0.82	0.85	0.91	0.95	0.97	0.99	0.97	0.96	0.88
Q	0.14	0.20	0.22	0.27	0.31	0.35	0.39	0.44	0.50	0.58	0.68	0.71	0.81	0.90	0.95	1.00	1.00
M	0.12	0.21	0.24	0.32	0.38	0.45	0.49	0.57	0.66	0.74	0.82	0.86	0.89	0.95	0.99	1.00	1.00
H	0.15	0.15	0.18	0.22	0.26	0.28	0.33	0.39	0.46	0.58	0.70	0.80	0.87	0.92	0.92	0.93	1.00
F	0.14	0.17	0.20	0.23	0.28	0.37	0.46	0.55	0.66	0.71	0.74	0.76	0.76	0.81	0.89	0.96	1.00
R	0.21	0.27	0.31	0.38	0.48	0.53	0.61	0.66	0.69	0.74	0.77	0.80	0.81	0.81	0.91	0.96	1.00
Y	0.14	0.16	0.24	0.28	0.31	0.33	0.40	0.48	0.55	0.64	0.77	0.87	0.96	0.98	1.00	1.00	1.00
W	0.19	0.26	0.33	0.35	0.36	0.33	0.35	0.40	0.44	0.49	0.52	0.58	0.61	0.65	0.70	0.79	0.88

B Z-Value Distribution of CAAs in Maximal Sets

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
G	1.06	1.56	0.85	0.73	0.32	-0.29	-0.52	-0.58	-0.69	-0.73	-0.92	-0.88	-0.60	-0.70	-0.64	-0.72	-1.19
A	0.26	0.70	0.48	0.40	0.43	0.38	0.29	0.38	0.54	0.37	0.22	0.08	-0.15	-0.27	-0.36	-0.72	-1.19
S	0.55	0.37	0.30	0.54	0.78	1.16	1.30	1.31	1.40	1.26	1.08	0.94	0.75	0.63	0.38	0.26	0.80
P	0.70	1.61	2.21	2.20	1.84	1.83	1.67	1.60	1.37	1.32	1.38	1.38	1.25	1.09	1.03	0.92	0.80
V	-0.53	0.49	0.95	1.14	1.71	2.02	1.92	1.68	1.58	1.50	1.26	0.99	0.87	0.73	0.38	-0.40	-1.19
T	-0.96	-1.35	-1.64	-1.60	-1.59	-1.44	-1.54	-1.59	-1.50	-1.40	-1.08	-0.78	-0.44	-0.13	-0.08	0.26	0.80
C	-0.17	0.06	0.19	0.44	0.50	0.49	0.56	0.61	0.55	0.62	0.72	0.96	1.09	1.09	1.03	0.92	0.80
I	-1.04	-1.15	-0.94	-1.10	-1.13	-1.05	-1.02	-1.24	-1.49	-1.77	-2.00	-2.01	-1.86	-1.29	-1.19	-1.38	-1.19
L	-1.62	-1.48	-1.48	-1.60	-1.65	-1.39	-1.44	-1.52	-1.52	-1.49	-1.32	-1.32	-1.25	-1.13	-1.38	-1.71	-1.19
N	1.56	0.90	1.17	0.76	0.23	0.05	0.06	0.04	0.02	0.01	-0.19	-0.33	-0.22	-0.13	0.01	0.59	0.80
D	-1.04	-1.13	-1.16	-1.26	-1.12	-1.09	-1.11	-1.26	-1.36	-1.45	-1.58	-1.80	-2.29	-2.82	-2.86	-2.37	-1.19
E	-0.17	-0.37	-0.51	-0.50	-0.46	-0.50	-0.56	-0.52	-0.55	-0.61	-0.72	-0.65	-0.51	-0.33	0.10	0.59	0.80
K	-0.24	-0.89	0.03	0.40	0.80	0.93	1.08	1.14	1.18	1.10	1.15	1.17	1.11	1.03	0.84	0.59	-1.19
Q	-0.39	0.08	-0.38	-0.35	-0.38	-0.36	-0.34	-0.30	-0.23	-0.09	0.12	0.06	0.31	0.56	0.66	0.92	0.80
M	-1.04	0.26	-0.11	0.21	0.24	0.33	0.25	0.32	0.49	0.64	0.75	0.74	0.71	0.83	0.93	0.92	0.80
H	-0.03	-0.90	-0.91	-0.92	-0.83	-0.83	-0.70	-0.56	-0.42	-0.10	0.22	0.48	0.57	0.63	0.47	0.26	0.80
F	-0.24	-0.49	-0.76	-0.74	-0.61	-0.20	0.04	0.26	0.46	0.50	0.40	0.30	0.07	0.03	0.29	0.59	0.80
R	2.14	1.28	0.82	0.88	1.19	0.93	0.93	0.81	0.63	0.62	0.54	0.46	0.28	0.03	0.38	0.59	0.80
Y	-0.39	-0.75	-0.18	-0.25	-0.35	-0.49	-0.27	-0.10	0.00	0.20	0.54	0.78	1.02	0.96	1.03	0.92	0.80
W	1.56	1.19	1.06	0.62	0.07	-0.51	-0.59	-0.50	-0.47	-0.48	-0.56	-0.57	-0.70	-0.83	-1.01	-1.05	-1.19

C Z-Value Distribution of CAAs in Better Sets

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
G	2.15	1.07	-0.01	-0.72	-0.75	-1.03	-0.95	-0.59	-0.91	-1.02	-1.14	-0.87	-0.94	-0.89	-0.45	-0.53	-0.02
A	-0.49	-0.78	-0.71	-0.71	-0.63	-0.72	-0.68	-0.78	-0.53	-0.61	-0.30	-0.30	-0.70	-0.49	-0.51	-0.28	0.36
S	-0.07	-0.05	-0.07	0.09	-0.02	0.37	0.36	0.18	-0.01	0.56	-0.13	-0.09	-0.09	-0.17	-0.02	-0.12	-0.40
P	1.17	1.49	1.61	1.84	1.63	2.32	2.30	1.87	1.16	1.45	1.05	0.52	0.93	1.02	0.51	0.39	0.36
V	-0.50	-0.85	-0.83	-0.76	-0.61	-0.78	-0.80	-0.72	-0.75	-0.61	-0.24	-0.43	-0.50	-0.60	-0.45	-0.17	0.36
T	-0.60	-0.54	-0.62	-0.56	-0.45	-0.13	-0.19	-0.07	-0.20	-0.27	-0.64	-0.43	-0.25	-0.34	-0.36	-0.53	-0.40
C	-0.23	-0.08	-0.25	-0.21	-0.22	-0.29	0.28	-0.07	-0.56	-0.68	-0.64	-0.78	-0.86	-0.80	-0.69	-0.53	-0.77
I	-0.84	-0.95	-0.90	-0.90	-0.84	-0.92	-0.93	-0.85	-0.72	-0.54	-0.47	-0.39	-0.33	-0.37	-0.45	-0.48	-0.77
L	-0.81	-0.95	-0.85	-0.83	-0.79	-0.85	-0.91	-0.92	-0.61	-0.47	-0.58	-0.18	-0.33	-0.51	-0.38	-0.38	-0.40
N	0.70	0.89	0.68	0.99	0.87	1.22	1.50	0.67	1.08	0.76	1.73	1.03	1.10	1.17	0.67	0.24	0.74
D	-0.46	-0.40	-0.38	-0.30	-0.34	0.13	0.22	0.03	0.31	-0.34	-0.07	0.04	0.16	0.07	-0.07	-0.28	-0.02
E	-0.45	-0.29	-0.52	-0.53	-0.63	-1.05	-0.99	-1.13	-0.83	-1.09	-1.09	-0.82	-0.99	-0.99	-0.74	-0.63	-0.40
K	-0.35	-0.36	-0.12	0.08	0.36	0.74	0.66	0.91	0.91	0.83	1.45	0.52	0.85	1.05	0.42	0.24	-0.40
Q	-0.75	-0.61	-0.54	-0.51	-0.51	-0.38	-0.43	-0.32	-0.34	0.01	-0.30	-0.30	0.08	-0.03	-0.07	-0.17	-0.77
M	1.16	2.51	2.59	2.63	3.16	1.91	1.51	1.60	2.16	2.82	2.52	3.71	3.30	3.28	3.89	3.97	3.38
H	-0.98	-0.72	-0.51	-0.45	-0.44	-0.58	-0.63	-0.45	-0.58	-0.27	-0.41	-0.22	-0.25	-0.01	-0.29	-0.12	-0.77
F	0.99	1.47	1.85	1.42	0.83	1.22	1.40	2.42	2.35	1.38	0.94	0.13	0.65	0.26	0.22	0.70	1.49
R	2.09	0.70	0.88	0.68	0.43	0.03	-0.25	-0.45	-0.61	-0.75	-0.69	-0.61	-0.82	-0.87	-0.69	-0.63	-0.77
Y	-0.91	-0.67	-0.45	-0.32	-0.24	-0.34	-0.46	-0.49	-0.50	-0.61	-0.07	0.08	-0.41	-0.20	-0.16	-0.28	-0.02
W	-0.83	-0.88	-0.86	-0.92	-0.81	-0.87	-1.00	-0.85	-0.83	-0.54	-0.92	-0.61	-0.58	-0.58	-0.40	-0.38	-0.77

D Absolute Differences Between Distributions A and B

	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
G	1.09	-0.49	-0.86	-1.45	-1.07	-0.74	-0.43	-0.01	-0.22	-0.29	-0.23	0.01	-0.35	-0.20	0.19	0.19	1.17
A	-0.75	-1.48	-1.19	-1.11	-1.05	-1.10	-0.97	-1.16	-1.07	-0.98	-0.51	-0.38	-0.55	-0.22	-0.15	0.45	1.55
S	-0.62	-0.42	-0.36	-0.45	-0.79	-0.79	-0.95	-1.13	-1.41	-0.71	-1.21	-1.02	-0.83	-0.80	-0.40	-0.39	-1.19
P	0.47	-0.11	-0.61	-0.37	-0.21	0.48	0.63	0.26	-0.21	0.13	-0.33	-0.86	-0.31	-0.07	-0.52	-0.53	-0.44
V	0.03	-1.34	-1.77	-1.90	-2.33	2.80	-2.72	-2.41	-2.33	-2.11	-1.50	-1.43	-1.36	-1.33	-0.83	0.22	1.55
T	0.37	0.80	1.02	1.03	1.14	1.30	1.35	1.52	1.30	1.13	0.44	0.34	0.19	-0.21	-0.27	-0.80	-1.19
C	-0.06	-0.15	-0.44	-0.65	-0.73	-0.78	-0.28	-0.68	-1.10	-1.30	-1.35	-1.74	-1.95	-1.89	-1.72	-1.45	-1.57
I	0.20	0.20	0.05	0.20	0.29	0.12	0.10	0.38	0.77	1.23	1.53	1.61	1.53	0.93	0.75	0.90	0.42
L	0.80	0.54	0.63	0.76	0.86	0.54	0.53	0.60	0.91	1.02	0.74	1.15	0.92	0.62	1.00	1.33	0.80
N	-0.87	-0.01	-0.50	0.23	0.63	1.17	1.44	0.63	1.05	0.76	1.92	1.37	1.32	1.30	0.66	-0.36	-0.06
D	0.58	0.73	0.78	0.96	0.78	1.22	1.33	1.29	1.67	1.11	1.51	1.84	2.45	2.89	2.79	2.09	1.17
E	-0.28	0.07	-0.02	-0.03	-0.17	-0.55	-0.42	-0.61	-0.28	-0.48	-0.37	-0.18	-0.47	-0.66	-0.84	-1.23	-1.19
K	-0.11	0.53	-0.15	-0.32	-0.44	-0.19	-0.42	-0.23	-0.27	-0.27	0.30	-0.65	-0.26	0.02	-0.42	-0.36	0.80
Q	-0.36	-0.68	-0.16	-0.16	-0.13	-0.02	-0.09	-0.02	-0.11	0.10	-0.42	-0.36	-0.24	-0.59	-0.73	-1.10	-1.57
M	2.20	2.25	2.69	2.42	2.92	1.58	1.26	1.28	1.68	2.18	1.77	2.97	2.59	2.45	2.96	3.05	2.59
H	-0.95	0.19	0.40	0.47	0.39	0.25	0.07	0.11	-0.17	-0.16	-0.63	-0.70	-0.82	-0.64	-0.76	-0.39	-1.57
F	1.23	1.96	2.60	2.16	1.45	1.42	1.36	2.15	1.89	0.88	0.54	-0.17	0.57	0.22	-0.07	0.10	0.70
R	-0.06	-0.59	0.06	-0.20	-0.76	-0.90	-1.18	-1.25	-1.24	-1.37	-1.23	-1.06	-1.10	-0.90	-1.07	-1.23	-1.57
Y	-0.52	0.08	-0.26	-0.07	0.11	0.15	-0.19	-0.38	-0.50	-0.81	-0.61	-0.69	-1.44	-1.16	-1.18	-1.20	-0.81
W	-2.40	-2.07	-1.93	-1.54	-0.88	-0.37	-0.41	-0.35	-0.36	-0.06	-0.35	-0.04	0.13	0.25	0.61	0.68	0.42

Figure 3. Relative frequency at which the individual CAAs are found in maximal CAA and better XAA sets. (**A**) shows the raw relative frequency of occurrence of the CAAs in maximal sets. (**B**) shows the Z-value for the frequency distribution shown in (**A**). In (**C**), green corresponds to a particular CAA occurring at high frequency relative to the other CAAs in sets, while red corresponds with low frequency. In (**D**), the absolute difference between the relative frequencies (Z-values) shown in (**B**) and (**C**) highlight areas where the relative frequencies of a particular amino acid vary between maximal and better sets, possibly highlighting CAAs having different importance depending on the context in which they are compared. Dark blue indicates a large difference between the frequency of a particular CAA in maximal CAA vs. better XAA (*e.g.*, those selected from the total XAA pool) sets. The direction of the bias can be determined by referring to panels B and C. Rounded raw values are shown for reference in each data cell.

As can be seen in Figure SI5, the probability of finding random better sets depends on sampling size and becomes asymptotic in the region around 10^7 sampled sets. At low-end sampling (10^6 trials and below), frequencies greater or lesser than the true average can be found. Frequencies for 10^7 and 10^8 sample sizes are almost indistinguishable. The general logarithmic decline in frequency of better sets, as well as the transient deviations in this trend at larger set sizes, is evident for both the CAAs (see Fig. 1), as well as for XAA sets 1 and 2 from Ilardo *et al.*[3] We believe this deviation comes simply from the nature of the possible set compositions as sets grow in size, *i.e.* the intrinsic structural possibilities and the properties they define, as well as the criteria by which optimality is defined here. These deviations may represent instances where there is the greatest opportunity for the incorporation of novel functionality, though by this point such opportunities are severely circumscribed by previous selections.

That Ile and Leu should appear to have a cooperative effect on fitness is at first glance surprising given their structural similarity. Indeed, these two CAAs score as the most similar pairwise CAAs according to many metrics[46–49], and among those most subsitutable for each other in modern proteins according to the work of Yampolsky and Stoltzfus[50].

Nevertheless, their V_{vdw}, pK_a, and logP values all vary slightly, therefore although we would expect them to have very similar adaptive value according to our metric, we also expect them to be different. This is in fact what is reflected in the heat maps shown in Fig. 3. We further tested why these two might be dissimilar enough to be co-adaptive numerically by simply subtly altering the CAA set. We found that despite their similarity, both Ile and Leu improve the evenness of the natural set, *e.g.*, removing either lowers the evenness score. We also checked doubling Ile and Leu individually (*e.g.*, having two instances of each, to allow them to weight their local numerical space) as psuedo controls, and this gave worse evenness scores than the both together or the omission of either. Although neither Ile nor Leu affects the range in any property dimension, the space is such that even the small amount of calculated difference in their descriptor values improves over all evenness, mainly on the basis of the LogP dimension coverage. This subtle difference is supported by solubility data, for example the Merck index[51] lists the solubility of L-Leu in water as 22.7 g/L at 0 C while while that of L-Ile is 37.9 g/L. Thus, there are subtle but significance in the chemical properties of these two amino acids that may allow them to be collectively adaptive despite their apparent structural similarity.

Ile and Leu are thus not identical (to think so is perhaps an example of the sort of human bias one needs to be wary of), and the presence of two similar CAAs in a coded set may be the sort of legacy one would expect from a fuzzy primordial code. Further, it is possible that in an earlier, more promiscuous preiod of biochemical evolution several there were "isomeric twins" or otherwise similar set members that may have buffered against damaging mutations to physico-chemically dissimilar amino acids. For example, changing the first position of four of the seven Leu codons converts it to Ile. Once protein complexity increased past a certain threshhold, selection operated on finer structural features.

There are of course likely other determining factors that were involved in the selection of life's biochemical toolkit beyond those that can be addressed in this type of numerical evaluations. For example, shape and steric properties are important for protein folding. Nevertheless, the number of amino acids with similar size and hydrophobicity values is fairly small, depending of course on the cutoff used for evaluating similarity.

It is noteworthy that for XAA set 1 no better sets are found beyond a set size of 16 for 10^6 trials nor beyond a set size of 18 for 10^7 trials, and for XXA set 2 none beyond a set size of 17 for samplings from 10^6 to 10^8 trials, which actually outperforms the actual CAA set at certain set sizes. We believe the idea that there might be many more optimal sets than the CAAs should perhaps be tempered first by the fact that one of these better XAA sets contains two CAAs, and second that neither of them has been analyzed with respect to their potential biosynthetic relationship or various other factors. For example, Gly, Ser, Ala and Cys belong to a fairly tight biosynthetic clique, as do Phe, Tyr and Trp, Asp and Asn, Glu and Gln, and to an extent the branched chain AAs.

This analysis does not take into account biosynthetic pathways and that "network closure," the notion that all processes inside a cell should be linked and share common resources for efficient coordination of metabolism, may have been important in the adaptive evolutionary construction of biochemistry[52]. It would be an interesting exercise to see which types of hypothetical metabolic networks can be constructed among the XAA better sets. However, as XAA better sets are already rather rare among the entire cohort, it seems likely that the addition of such a constraint would only make the emergence of the actual CAA set appear more adaptive and predisposed.

As the full diversity of possible enzymatic transformations using CAA comprised proteins is unknown, and that of potential XAA comprised peptide polymers completely unknown, it is impossible to make very strong statements about the completeness of coverage of catalytic mechanism space by modern biochemistry. However these sorts of relationships also could have affected the search space biology explored during the development of coded sets.

It should also be noted there is some debate about the mechanism of genetic code evolution and the role of selection in its origin (see for example refs[53–59]), nevertheless since it seems unlikely all of the CAAs were available from prebiotic synthesis[14], and thus some may have been adopted into the code after organisms developed the ability to biosynthesize them. Their addition, or non-addition, would then be selectable. There are of course alternative hypotheses one could entertain, for instance that organisms were able to biosynthesize all of the CAAs before any part of the code was canonized, but the stepwise expansion and rewiring of a more primitive code remains a compelling possibility.

One last note should be added here. The XAA library used here is the simplest computed by Meringer *et al*.[34], however, the larger libraries computed therein cover chemical property space based on ways we think justify the use of this smaller library in the present study. There are always more structural variations of larger amino acids than of smaller ones, and the coverage of chemical space becomes very dense for larger molecules because the nuances of structural diversity become subtle. The population of the search space with respect to the properties of the targets could alter the landscape of these results, but we do not feel it would do so qualitatively. The smaller amino acids, for which there are fewer isomers, would still occupy a corner of property space; and the larger ones would occupy regions that are highly redundantly occupied. Using the exhaustive list of alternatives would thus also tend to weight the smaller amino acids, and de-weight amino acids of higher molecular weight. This would tend to underscore the point that the CAAs are positioned in property space in such a way that they are likely selectional outcomes regardless of search blank.

Conclusions

Our analysis suggests that stepwise expansion of the CAA repertoire proceeding through the present CAAs represents a trajectory that became increasingly adaptive relative to hypothetical sets that can be constructed from XAAs. These results pose a number of other questions: why then did biology (with the two known infrequently occurring exceptions of selenocysteine and pyrrololysine) stop exploring chemical space? Is it because property space was already well-explored based on these principles? Our results suggests that this is indeed the case - the possible chemical space both limits and directs the explorable space. This analysis suggests that once evolving organisms acquired one or more amino acids from the modern CAA set, the organisms encoding them would also be poised on a trajectory to incorporate still other modern CAAs, *i.e.*, the fitness landscape would have steered organisms to fill amino acid selection in roughly the same way. We might then expect organisms on other Earth-like planets to use similar amino acids.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request. The compound library of 1913 amino acid structures as SD file can be downloaded from www.molgen.de/data/AACLBR.sdf.zip.

References

1. Freeland, S. J. & Hurst, L. D. The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248 (1998).
2. Philip, G. K. & Freeland, S. J. Did evolution select a nonrandom "alphabet" of amino acids? *Astrobiology* **11**, 235–240 (2011).
3. Ilardo, M., Meringer, M., Freeland, S., Rasulev, B. & Cleaves, H. J. Extraordinarily adaptive properties of the genetically encoded amino acids. *Sci. Rep.* **5** (2015).
4. Dobson, C. M. Chemical space and biology. *Nature* **432**, 824 (2004).
5. Eberhardt, L., Kumar, K. & Waldmann, H. Exploring and exploiting biologically relevant chemical space. *Current Drug Targets* **12**, 1531–1546 (2011).

6. Drew, K. L., Baiman, H., Khwaounjoo, P., Yu, B. & Reynisson, J. Size estimation of chemical space: how big is it? *J. Pharm. Pharmacol.* **64**, 490–495 (2012).
7. Polishchuk, P., Madzhidov, T. & Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comp.-Aided Mol. Design* **27**, 675–679 (2013).
8. Virshup, A. M., Contreras-García, J., Wipf, P., Yang, W. & Beratan, D. N. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* **135**, 7296–7303 (2013).
9. Cleaves, H. J. The origin of the biologically coded amino acids. *J. Theor. Biol.* **263**, 490–498 (2010).
10. Miller, S. L. A production of amino acids under possible primitive Earth conditions. *Science* **117**, 528–529 (1953).
11. Trifonov, E. Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **261**, 139–151 (2000).
12. Ikehara, K. Possible steps to the emergence of life: The [GADV]-protein world hypothesis. *The Chemical Record* **5**, 107–118 (2005).
13. Higgs, P. G. & Pudritz, R. E. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology* **9**, 483–490 (2009).
14. Wong, J. T. & Bronskill, P. M. Inadequacy of prebiotic synthesis as origin of proteinous amino acids. *J. Mol. Evol.* **13**, 115–125 (1979).
15. Wong, J. Coevolution theory of the genetic code at age thirty. *BioEssays* **27**, 416–425 (2005).
16. Fournier, G. P. & Alm, E. J. Ancestral reconstruction of a pre-LUCA aminoacyl-tRNA synthetase ancestor supports the late addition of Trp to the genetic code. *J. Mol. Evol.* **80**, 171–85 (2015).
17. Granold, M., Hajieva, P., Toşa, M. I., Irimie, F. D. & Moosmann, B. Modern diversification of the amino acid repertoire driven by oxygen. *Proc. Nat. Acad. Sci. USA* **115**, 41–46 (2018).
18. Hinds, D. A. & Levitt, M. From structure to sequence and back again. *J. Mol. Biol.* **258**, 201–209 (1996).
19. Weberndorfer, G., Hofacker, I. L. & Stadler, P. F. On the evolution of primitive genetic codes. *Origins of Life Evol. Biosphere* **33**, 491–514 (2003).
20. Stephenson, J. D. & Freeland, S. J. Unearthing the root of amino acid similarity. *J. Mol. Evol.* **77**, 159–69 (2013).
21. Higgs, P. G. A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. *Biol. Direct* **4**, 16 (2009).
22. Di Giulio, M. The coevolution theory of the origin of the genetic code. *J. Mol. Evol.* **48**, 253–5 (1999).
23. Wong, J. T. A co-evolution theory of the genetic code. *Proc. Nat. Acad. Sci. USA* **72**, 1909–12 (1975).
24. Hohsaka, T. & Sisido, M. Incorporation of non-natural amino acids into proteins. *Curr. Opinion Chem. Biol.* **6**, 809–815 (2002).
25. Lang, K. & Chin, J. W. Cellular incorporation of unnatural amino acids and bioorthogonal labeling of proteins. *Chem. Rev.* **114**, 4764–4806 (2014).
26. Wong, J. T. Membership mutation of the genetic code: loss of fitness by tryptophan. *Proc. Nat. Acad. Sci. USA* **80**, 6303–6 (1983).
27. Bacher, J. M. & Ellington, A. D. Selection and characterization of Escherichia coli variants capable of growth on an otherwise toxic tryptophan analogue. *J. Bacteriol.* **183**, 5414–25 (2001).
28. Yu, A. C. *et al*. Mutations enabling displacement of tryptophan by 4-fluorotryptophan as a canonical amino acid of the genetic code. *Genome Biol. & Evol.* **6**, 629–641 (2014).
29. Heylighen, F. The growth of structural and functional complexity during evolution in *The Evolution of Complexity: The Violet Book of "Einstein Meets Magritte"* (Eds Heylighen, F., Bollen, J. & Riegler, A.) 8:17–44, (VUB University Press 1999).
30. Hazen, R. M., Griffin, P. L., Carothers, J. M. & Szostak, J. W. Functional information and the emergence of biocomplexity. *Proc. Nat. Acad. Sci. USA* **104**, 8574–8581 (2007).
31. Fujishima, K. *et al*. Reconstruction of cysteine biosynthesis using engineered cysteine-free enzymes. *Sci. Rep.* **8**, 1776 (2018).
32. Xie, J. & Schultz, P. G. A chemical toolkit for proteins—an expanded genetic code. *Nature Rev. Mol. Cell Biol.* **7**, 775 (2006).
33. Ilardo, M. A. & Freeland, S. J. Testing for adaptive signatures of amino acid alphabet evolution using chemistry space. *J. Sys. Chem.* **5**, 1–9 (2014).
34. Meringer, M., Cleaves, H. J. & Freeland, S. J. Beyond terrestrial biology: Charting the chemical universe of α-amino acid structures. *J. Chem. Inf. Model.* **53**, 2851–2862 (2013).
35. Meringer, M. & Cleaves, H. J. Exploring astrobiology using *in silico* molecular structure generation. *Phil. Trans. R. Soc. A* **375**, 20160344 (2017).
36. Gugisch, R. *et al*. MOLGEN 5.0, A molecular structure generator. *Adv. Math. Chem. and Applications: Revised Edition* **1**, 113–138 (2016).
37. Lu, Y. & Freeland, S. Testing the potential for computational chemistry to quantify biophysical properties of the non-proteinaceous amino acids. *Astrobiology* **6**, 606–624 (2006).
38. Todeschini, R. & Consonni V. Handbook of Molecular Descriptors. Volume 11 of Methods and Principles in Medicinal Chemistry. John Wiley & Sons (2008).
39. Lu, Y., Bulka, B., desJardins, M. & Freeland, S. J. Amino acid quantitative structure property relationship database: a web-based platform for quantitative investigations of amino acids. *Protein Engineering, Design & Selection* **20**, 347–51 (2007).
40. Bywater, R. P. Why twenty amino acid residue types suffice(d) to support all living systems. *PLoS One* **13** (2018).
41. Doig, A. J. Frozen, but no accident–why the 20 standard amino acids were selected. *FEBS J.* **284**, 1296–1305 (2017).
42. Bose, R., Meringer, M., Ilardo, M. & Cleaves, H. J. Adaptive properties of the amino acid alphabet and its subsets. The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE): 459–460 (2018).
43. Klein, D. J. & Babić, D. Partial orderings in chemistry. *J. Chem. Inf. Comp. Sci.* **37.4**, 656–671 (1997).
44. Brüggemann, R. & Lars, C. eds *Partial order in environmental sciences and chemistry*. Berlin: Springer, (2006).
45. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. *Methods Mol. Biol.* **406**, 89–112 (2007).
46. Sneath, P. Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* **12**, 157–195 (1966).
47. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
48. Epstein, C. Non-randomness of Amino-acid changes in the evolution of homologous proteins. *Nature* **215**, 355–359 (1967).
49. Miyata, T., Miyazawa, S. & Yasunaga, T. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **12**, 219–236 (1979).
50. Yampolsky, L. & Stoltzfus, A. The exchangeability of amino acids in proteins. *Genetics* **170**, 1459–1472 (2005).
51. O'Neil, M. (ed.) *The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals*. Royal Society of Chemistry, Great Britain (2013).
52. Letelier, J.-C., Cárdenas, M. L. & Cornish-Bowden, A. From L'homme machine to metabolic closure: steps towards understanding life. *J. Theor. Biol.* **286**, 100–113 (2011).
53. Massey, S. E. A neutral origin for error minimizaton in the genetic code. *J. Mol. Evol.* **67**, 510 (2008).
54. Di Giulio, M. A non-neutral origin for error minimization in the origin of the genetic code. *J. Mol. Evol.* **86**, 593–597 (2018).
55. Koonin, E. V. & Novozhilov, A. S. Origin and Evolution of the Universal Genetic Code. *Ann. Rev. Genet.* **51**, 45–62 (2017).
56. Fournier, G. P. & Alm, E. J. Ancestral Reconstruction of a Pre-LUCA Aminoacyl-tRNA Synthetase Ancestor Supports the Late Addition of Trp to the Genetic Code. *J. Mol. Evol.* **80**, 171–85 (2015).
57. Bernhardt, H. S. & Patrick, W. M. Genetic code evolution started with the incorporation of glycine, followed by other small hydrophilic amino acids. *J. Mol. Evol.* **78**, 307–9 (2014).
58. Wong, J., Ng, S.-K., Mat, W.-K., Hu, T. & Hong, X. Coevolution theory of the genetic code at age forty: pathway to translation and synthetic life. *Life* **6**, 12 (2016).
59. Fitch, W. M. & K. Upper, K. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. In *Cold Spring Harbor symposia on quantitative biology*, 52, 759–767. Cold Spring Harbor Laboratory Press (1987).

Acknowledgements

The authors would like to thank the Earth-Life Science Institute (ELSI) for support during the preparation of this work, and EON for generous travel support for RB, NG, MI, MM and BR. This work was also supported by JSPS KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas "Hadean Bioscience", Grant Number JP26106003, and by the ELSI Origins Network (EON), which is supported by a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation.

Author Contributions

M.I., R.B., H.C., M.M., B.R., N.G., R.G. and C.B. helped design the code used in this work. J.S. conducted the protein database searches. M.I., R.B., H.C., M.M., B.R., N.G., R.G., S.F. and C.B. helped craft the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-47574-x.

Competing Interests: The authors declare no competing interests.

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