An iterated learning approach to the origins of the standard genetic code can help to explain its sequence of amino acid assignments

Tom Froese [1,2*], Jorge I. Campos [3,2] and Nathaniel Virgo [4]

[1]Institute for Applied Mathematics and Systems Research, National Autonomous University of Mexico, Mexico City, Mexico
[2]Center for the Sciences of Complexity, National Autonomous University of Mexico, Mexico City, Mexico
[3]Faculty of Higher Education Aragon, National Autonomous University of Mexico, Nezahualcoyotl City, Mexico
[4]Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo, Japan
*Corresponding author: t.froese@gmail.com

Abstract

Artificial life has been developing a behavior-based perspective on the origins of life, which emphasizes the adaptive potential of agent-environment interaction even at that initial stage. So far this perspective has been closely aligned to metabolism-first theories, while most researchers who study life's origins tend to assign an essential role to RNA. An outstanding challenge is to show that a behavior-based perspective can also address open questions related to the genetic system. Accordingly, we have recently applied this perspective to one of science's most fascinating mysteries: the origins of the standard genetic code. We modeled horizontal transfer of cellular components in a population of protocells using an iterated learning approach and found that it can account for the emergence of several key properties of the standard code. Here we further investigated the diachronic emergence of artificial codes and discovered that the model's most frequent sequence of amino acid assignments overlaps significantly with the predictions in the literature. Our explorations of the factors that favor early incorporation into an emerging artificial code revealed two aspects: an amino acid's relative probability of horizontal transfer, and its relative ease of discriminability in chemical space.

Introduction

The origin of life is one of the major open questions faced by science, and it requires broad interdisciplinary collaborations for systematic answers to begin to be formulated (Smith & Morowitz, 2016). Part of the problem is that the origin of life is closely related to another major unresolved issue, namely the origin of the genetic code (Koonin & Novozhilov, 2017).

There is a small but growing number of researchers who propose that making progress on the first open question requires us to take seriously the role of life's embodiment: it is likely that even the very first living beings were spatially individuated and interacted with their environment as basic agents (Di Paolo, Buhrmann, & Barandiaran, 2017). This enactive perspective emphasizes a strong continuity between life and mind, and it suggests that movement, interaction, and adaptive behavior could have already played a crucial role at the origin and initial evolution of life (Egbert, Barandiaran, & Di Paolo, 2012; Froese, Virgo, & Ikegami, 2014; Hanczyc & Ikegami, 2010). For instance, it is a common finding that even the most minimal forms of dissipative structures can behave adaptively: they tend to move (or, rather, grow) up chemical gradients that favor their own self-production (e.g. Suzuki & Ikegami, 2009). And when these structures are capable of spontaneous movement in environments without gradients, as a population they become more resilient against adverse conditions such as unstable environments and lower nutrient levels (Virgo, Froese, & Ikegami, 2013).

This behavior-based approach to the origin of life has also opened up new perspectives on the second unresolved issue, namely the origin of the genetic code. Froese (2015) proposed that its regular arrangement and compositional structure could be an emergent outcome of repeated horizontal interactions between protocells, akin to the iterated learning approach to the origins of language. In other words, if there is support for strong life-mind continuity, perhaps it is worthwhile to take seriously the possibility of life-sociality continuity, too.

This proposal fits well with Woese's (2002) intuition that Darwinian evolution by vertical descent had to be preceded by communal evolution of small populations of protocells by means of rampant horizontal transfer of cellular components. Existing simulation models have confirmed that allowing for horizontal transfer indeed aids population convergence on a regular and universal genetic code (e.g. Aggarwal, Bandhu, & Sengupta, 2016; Goldenfeld, Biancalani, & Jafarpour, 2017; Vetsigian, Woese, & Goldenfeld, 2006). However, despite their emphasis on horizontal transfer, these models continue to take code optimization via evolution by vertical descent for granted. To some extent they have simply rediscovered the effects of adding a crossover operator to a genetic algorithm. Accordingly, applying a behavior-based optimization process without relying on Darwinian genetic evolution, such as the iterated learning paradigm, offers an alternative approach to the origin of the genetic code, and one that actually may be more suitable given that optimization by vertical descent to some extent already presupposes an optimized genetic code.

We recently tested Froese's proposal with a population-based iterated learning model, and the results are encouraging: the emerging artificial genetic codes tend to be characterized by all of the most prevalent regularities that are known in the literature (Froese, Campos, Fujishima, Kiga, & Virgo, 2018). For instance, the codes tend to be robust against coding errors because amino acids with similar chemical properties tend to be associated with similar codons. This property would have

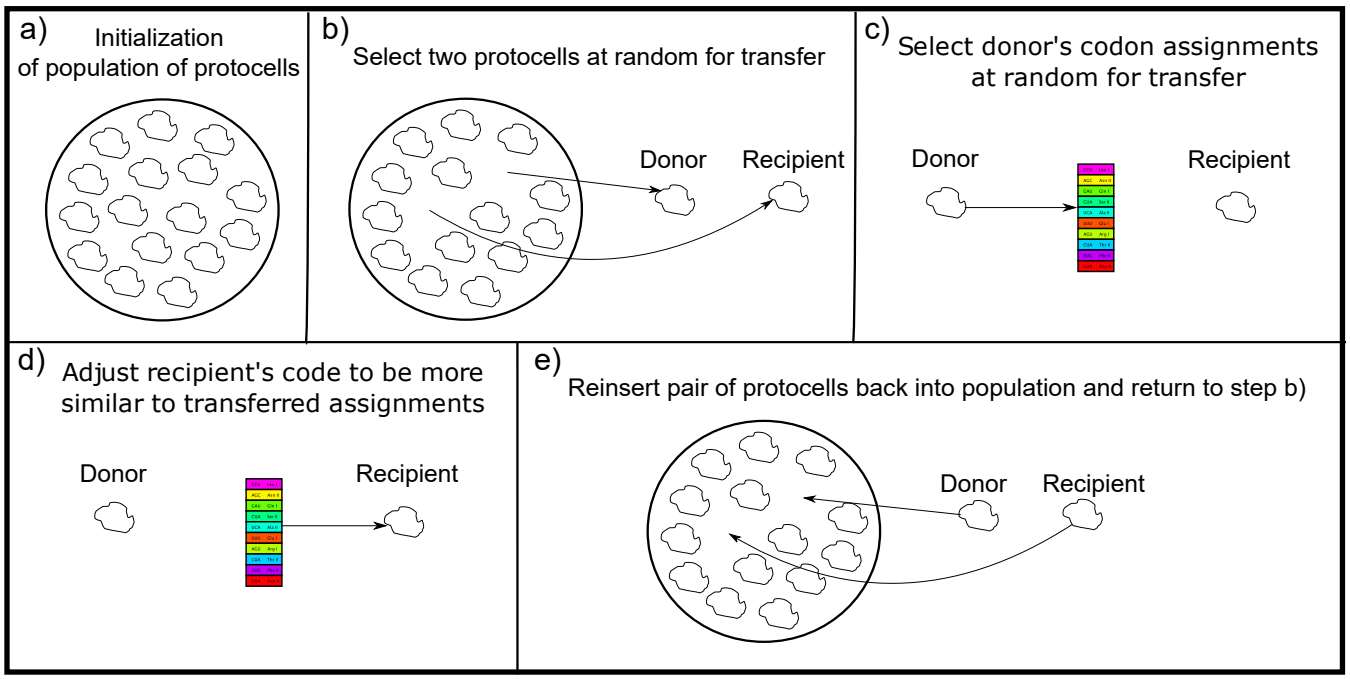

Figure 1: A population-based iterated learning approach to genetic code evolution. The model was inspired by Woese's (2002) idea that at the origins of life there still was not a clear genotype-phenotype distinction, and that the first forms of life must have therefore been subject to a form of communal evolution dominated by the dynamics of horizontal exchange of cellular components, rather than by Darwinian evolution based on vertical descent. The basic mechanism is reminiscent of the iterated learning approach to language evolution, and so we designed our model following a population-based model of iterated learning (Brace et al., 2015).

shielded genetic material from high rates of point mutations during replication and from frequent mistranslations during protein synthesis, both of which a primitive genetic system would have somehow had to cope with before becoming more optimized by Darwinian evolution.

Another example of a regularity of the standard genetic code that also emerged in our model is the positive correlation between an amino acid's relative frequency in proteins and the number of codon assigned it. We mention it here because we had found this regularity to be based on the 20 standard amino acids' relative probability of horizontal transfer, a mechanism that will also turn out to be important for this new analysis.

In the original report we focused our analysis on the final artificial genetic codes, but the same model can also be used to investigate how they emerged diachronically. In particular, we can compare the relative likelihood of each of the standard 20 amino acids becoming incorporated into a code, and verify if this matches the expected order of amino acid fixation in the literature. An advantage of working with a simulation model, compared to studying the unique event of life on earth, is that we can easily rerun the process of code emergence in order to separate consistent correlations from contingent outcomes. As we will explain in more detail later, we used this advantage of working with a simulation model to disambiguate the factors that favor the incorporation of amino acids.

Various theories of the origin of the standard genetic code have been proposed (for a recent review of the state of the art, see Koonin and Novozhilov (2017)). One prominent theory is so-called "coevolution theory" (Wong, 2005), which has proposed two distinct phases of genetic code evolution: *Phase 1* incorporated 10 amino acids into the code that were directly available in the prebiotic environment, namely:

Val, Ala, Leu, Thr, Glu, Asp, Ile, Ser, Gly, Pro

Afterwards, during *Phase 2*, the genetic code is proposed to have incorporated 10 additional amino acids that were mainly or perhaps even only produced by biosynthesis:

Arg, His, Met, Trp, Asn, Gln, Lys, Phe, Tyr, Cys

We can therefore ask whether the artificial codes emerging in our model tend to adhere to this predicted division between early (Phase 1) and late (Phase 2) incorporations.

This distinction between prebiotic and biotic is a binary one and therefore hides some of the complexity of code evolution. We can go further and compare our modeling results with a continuous ranking of prebiotic prevalence. A review of the empirical literature by Higgs and Pudritz (2009), including all kinds of lab experiments and meteorite studies, produced the following ranking of prebiotic amino acid abundance, ordered from high to low relative abundance:

[Gly, Ala, Asp, Glu, Val, Ser, Ile, Leu, Pro, Thr], Lys, Phe, (His, Arg), (Gln, Asn, Tyr, Trp, Met, Cys)

The amino acids in square brackets are the top 10 ranked, which are the same as the Phase 1 amino acids proposed by coevolution theory. Amino acids in curved brackets share the same rank. The last 6 amino acids have not been found in any prebiotic contexts so far.

In the following section we briefly describe the methods we used to run the original iterated learning model, and what we did in the current study to determine the most likely sequence of amino acid assignments in the emerging genetic codes. We then present and analyze the results. It turns out that the model is able to generate sequences of amino acid incorporations

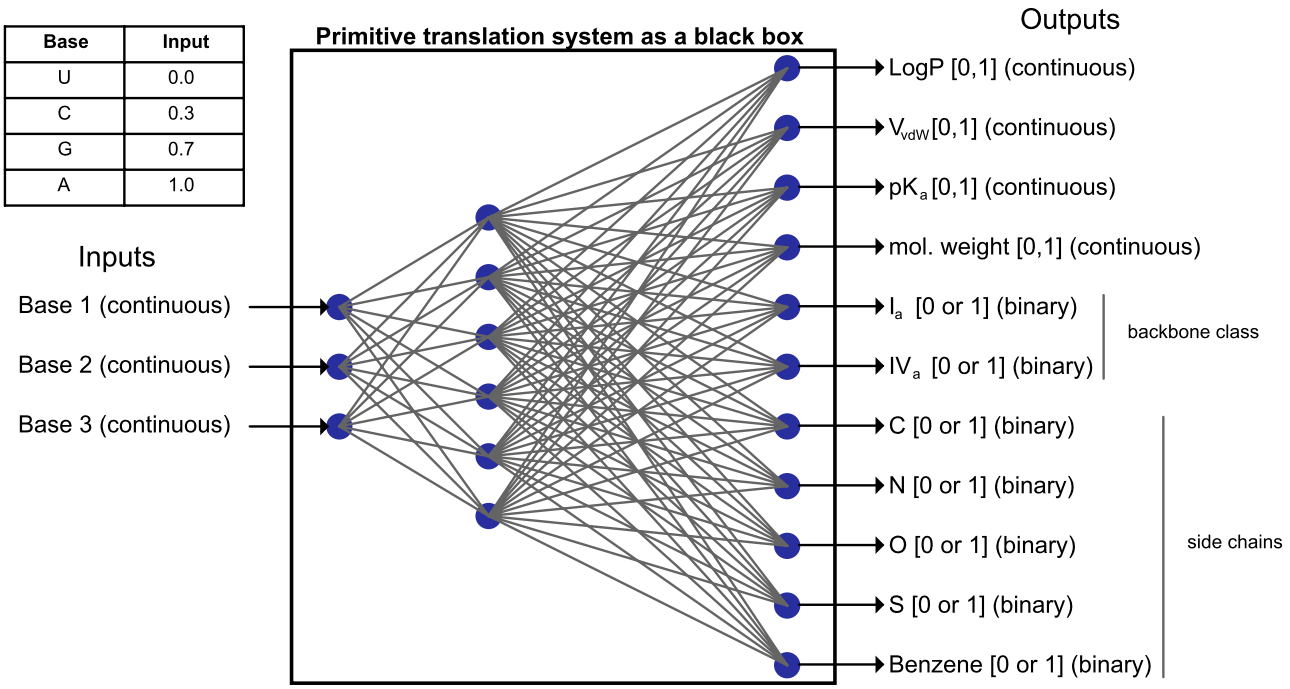

Figure 2: The primitive genetic translation system modeled as a 'black box'. Woese (2002) envisioned the first forms of life as "supramolecular aggregates", consisting of a self-sustaining metabolic network that was incorporated into a modular higher-order architecture formed around nucleic acid components. For the purposes of our model we ignored the metabolic network, and we treat the primitive translation system as a 'black box' system that is capable of nonlinearly mapping from an input in nucleotide space (a codon) to an output in chemical space (an amino acid). For simplicity, and following previous iterated learning models (e.g. Kirby & Hurford, 2002), this 'black box' was implemented as a multilayer perceptron network capable of backpropagation learning. This particular implementation of the translation system is admittedly not a realistic architecture for protocells, but backpropagation is at least within the possibilities of chemical networks in principle (Blount et al., 2017).

into the artificial genetic codes that significantly overlap with what is expected from the literature.

Given that we make all 20 standard amino acids available for potential encoding in the simulation model right from the start, and yet we still tend to find similar sequences to the empirically predicted ones, it suggests that other factors than relative prebiotic abundance should be considered. There is something about Phase 2 amino acids that makes them less likely to be incorporated into emerging codes other than just their rare environmental presence and/or high energetic costs of synthesis. As we will see, compared to Phase 1 amino acids the Phase 2 amino acids also do not provide notably improved coverage of chemical space. Our initial explorations of the factors favoring early incorporation move us beyond these main concerns of current theories by highlighting two other potential factors: differences in rates of horizontal transfer and discriminability in chemical space.

Method

We first ran our simulation model using the same settings as reported in the original article by Froese et al. (2018). Briefly, we applied an iterated learning approach, typically employed to study the evolution of language (Kirby, Griffiths, & Smith, 2014), to the evolution of the genetic code. We implemented a population-based model of communal evolution of the genetic code that consisted of several steps (Figure 1).

(a) A small group of protocells (N = 16) is initialized such that each of their 'black box' primitive translation systems (see Figure 2 for details) encodes a random genetic code, which therefore initially is only capable of specifying a few amino acids. Then the 'iterative learning' cycle begins:

(b) Two protocells are randomly selected to engage in a horizontal transfer of a fragment of the donor's genetic code to the recipient.

(c) A small subset of the donor's codon assignments is randomly chosen (10 out of 64 codon assignments) and then transferred; occasionally, codon assignment inaccuracies can occur in the transferred components.

(d) The recipient adjusts its genetic code to be more like the donor's code according to the received codon assignments. This takes the form of gradient descent.

(e) The process of horizontal transfer is completed. Then the cycle starts again by going back to (b).

For simplicity, and following previous work on the iterated learning model, we used a fully interconnected feed-forward multi-layer perceptron network to model the translational mapping from a codon to its corresponding amino acid. There are three input nodes, one for each of a codon's bases. The network is not spatially embedded and so the order of the three base positions is arbitrary and interchangeable (i.e. we

did no model Crick's (1966) third base 'wobble' hypothesis). There are six hidden nodes. Output is an 11-dimensional vector that specifies an amino acid in terms of properties by which it can be uniquely distinguished in chemical space.

All 20 amino acids of the standard genetic code were made available for potential incorporation into the artificial codes. This is a simplifying assumption; not all of these 20 amino acids were available at the origin of life, given that some of them are dependent on biological synthesis. However, at least in this way we also avoided biasing the model's sequences of amino acid assignments. In other words, if some amino acids tend not to get incorporated at the start, this is not because the model prevented their incorporation into a code by excluding them in an *a priori* manner.

The probability of a particular amino acid being included in a horizontal transfer was based on an estimate of the cellular relative amino acid abundance of modern organisms (Moura, Savageau, & Alves, 2013). It is currently unknown whether this estimate can be projected back to the early stages of life, but in any case, as we will show, the precise probabilities only have a small effect on the model's outcomes.

For the current study we conducted 100 independent runs of our original simulation model. For each run we recorded the amino acids that were incorporated into the first 10 codes that had managed to specify at least 10 amino acids. This gave us a set of 1,000 early artificial codes specifying at least 10 amino acids each. We then calculated the frequency of each amino acid's incorporation as the percentage of early codes in which at least one codon had been assigned to the amino acid. We then varied the probabilities of transfer and reran the model to determine the effects. First, we directly inverted the probabilities of the original model by ordering the 20 amino acids in terms of their probability of horizontal transfer and reassigned the probabilities in reverse order. We conducted 100 independent runs of this inverted scenario. Second, we tested the model for another 100 runs but this time with neutral probabilities such that all amino acids have the same frequency of horizontal transfer. Finally, as a further control condition, we arbitrarily selected 10 out of the total 20 amino acids a 1,000 times in order to determine what kind of codes would be generated by pure chance alone.

Results

The first two batches of results of this study, consisting of the averages of 100 independent runs, are summarized in Table 1. Using the parameters of the original model we found that, on average, the 10 amino acids that are most likely to be found in the emerging codes that have managed to incorporate at least 10 distinct amino acids are (ordered from most likely to least likely to be incorporated into an emerging code):

Val, Ala, Thr, Leu, Glu, Asp, Lys, Ser, Gly, Gln

A comparison with the amino acids that were proposed by coevolution theory to have been incorporated into the standard genetic code in Phase 1 reveals a significant overlap: 8 out of 10 are the same. The two amino acids that were predicted by the model to be part of Phase 1 but did not match expectations of the literature, Lys and Gln, were low ranked at places 7 and 10, respectively. And of the two amino acids that were not among the 10 most likely initial incorporations (Ile and Pro), it is notable that Pro was regarded as only a "marginal phase 1" amino acid in the literature (Wong, 2005). The model thus does much better than chance, given that arbitrary subsets of 10 of the 20 amino acids would on average only overlap by 5 out of 10 and only at around 50% frequency (see Table 2).

Original transfer bias				**Inverted transfer bias**			
Amino acid (AA)	***Ranking***	***Frequency of appearance in the first 10 encoded AAs (in %)***	***Probability of horizontal transfer (in %)***	***Amino acid (AA)***	***Ranking***	***Frequency of appearance in the first 10 encoded AAs (in %)***	***Probability of horizontal transfer (in %)***
Val	1	97.6	7.0	**Leu**	1	91.2	1.1
Ala	2	95.9	8.8	Lys	2	87.2	4.2
Thr	3	85.0	5.2	**Val**	3	83.8	2.1
Leu	4	85.0	10.2	**Asp**	4	75.2	5.2
Glu	5	81.1	6.4	**Thr**	5	73.7	5.3
Asp	6	77.3	5.3	**Glu**	6	70.3	3.6
Lys	7	77.0	5.8	Phe	7	70	5.8
Ser	8	70.7	6.1	**Ala**	8	52.9	1.2
Gly	9	63.7	7.0	Trp	9	52	8.8
Gln	10	62.6	3.6	Gln	10	43.7	6.4
Asn	11	51.1	4.1	Asn	11	41.3	6.1
His	12	36.0	2.1	Tyr	12	40.6	6.7
Phe	13	27.5	4.2	His	13	38.9	7.0
Arg	14	26.3	5.4	Met	14	36.6	7.0
Tyr	15	20.6	3.1	Arg	15	36.2	4.2
Ile	16	18.2	6.7	Cys	16	34.9	10.2
Pro	17	17.3	4.2	**Ser**	17	24.3	4.1
Trp	18	3.7	1.2	**Pro**	18	21	5.4
Met	19	1.9	2.5	**Ile**	19	17.2	3.1
Cys	20	1.5	1.1	**Gly**	20	9	2.5

Table 1: **Biased horizontal transfer condition**. Amino acids (AAs) are ranked according to their frequency of appearing among the first 10 AAs to be encoded by emerging artificial genetic codes. The model's predicted Phase 1 AAs, which are the top 10 most frequent incorporations, are highlighted with a grey background. The AAs predicted to be part of Phase 1 in the literature are in bold. *Original transfer bias*: These results were generated by using the same model described in Froese et al. (2018), in which horizontal transfer of AAs is biased according to relative abundance in proteins. *Inverted transfer bias*: These results were generated by running the model with an inverted order of probability of horizontal transfer.

We can also compare this artificial ranking produced by our original model with the continuous empirical ranking made by Higgs and Pudritz. In this case we again find a strong overlap: of the two that are part of the model's Phase 1 but absent from the empirical Phase 1 (Lys and Gln), Lys appears immediately afterwards at rank number 11 in the empirical ranking. The other mismatching amino acid, Gln, however, appears among the 6 amino acids sharing last place in the empirical ranking. Looking at the predicted ranking from this other end, we still find a strong overlap: 4 out of the 6 amino acids that have not been found in any prebiotic contexts are also among the 6 least likely to be initially incorporated in the model (Tyr, Trp,

Met, Cys). The other amino acid that is missing from these 6, i.e. apart from Gln, is Asn, which the model at least correctly placed in Phase 2, albeit only marginally so as the 11th most frequent incorporation.

The reasons for this striking consistency between the model and coevolution theory and the estimates of relative prebiotic abundance are not immediately evident. The model does not explicitly simulate prebiotic environmental abundances of amino acids (to the contrary, it makes all the complex amino acids available from the start even though they presumably would only have appeared during later phases of evolution), nor does it include considerations of biosynthetic pathways or energetic costs. However, it does vary the relative probability of each one of the 20 standard amino acids to be included in a horizontal transfer. We therefore expected that probability of transfer would strongly influence the relative likelihood of early incorporation. Indeed, the average probability of transfer of the 10 most and least frequently incorporated amino acids is 6.5% and 3.5%, respectively. In other words, it may be significant that on average those amino acids that are most likely to be incorporated among the first 10 of an emerging code are also twice as likely to be present in a horizontal transfer compared to the other amino acids.

To test this positive correlation between an amino acid's frequency of incorporation and its probability of horizontal transfer, we re-ran the experiment, but this time with the order of amino acid probabilities inverted. The results of the model with this inverted transfer bias are shown in the right-hand side of Table 1. To our surprise, we did not find a simple inversion of the ranking of the original model. Some amino acids shifted their rank in the expected direction, e.g. Gly shifted from rank 9 down to 20 and Trp shifted from rank 18 up to 9. However, the rank of some amino acids was also left strangely unaffected by the inverted probabilities, e.g. Val only descended from rank 1 to 3. Moreover, Leu even moved in the opposite way from what we had expected, namely from rank 4 up to 1, even though the inversion in its probability of transfer had been the most extreme: from 10.2% to 1.1%.

In fact, the previously observed general correlation between frequency of incorporation and probability of transfer was not recovered: unexpectedly, the average probability of transfer of the 10 most and least frequently incorporated amino acids was now almost completely inverted, namely 4.37% and 5.63%, respectively. Nevertheless, despite the reduced probability of transfer, we found that 8 out of the originally predicted top 10 amino acids still remain in the top 10, including 6 out of the 8 that had overlapped with the Phase 1 amino acids proposed by coevolution theory.

In sum, apart from significantly decreasing the frequency of Gly and Ser, inverting the probability of horizontal transfer only had a small effect on the model's predicted ranking. This implies that there must be another essential factor influencing the likelihood of early incorporation into the emerging genetic codes, at least for some of the amino acids.

In order to better understand the nature of this additional influence we reran the model again, but this time without any biased horizontal transfer. In other words, all 20 amino acids are equally likely to be transferred. If probability of horizontal transfer (or, indirectly, relative abundance) is the only or main factor at work, then all of the 20 amino acids should have an equal likelihood of appearing among the first 10. But as we suspected this is not what was found, as shown in Table 2.

No transfer bias				**Random selection**			
Amino acid (AA)	***Ranking***	***Frequency of appearance in the first 10 encoded AAs (in %)***	***Probability of horizontal transfer (in %)***	***Amino acid (AA)***	***Ranking***	***Frequency of appearance in 10 random AAs (in %)***	***Probability of horizontal transfer (in %)***
Val	1	95.4	5.0	Tyr	1	52.1	N/A
Leu	2	92.8	5.0	Gln	2	52.1	N/A
Lys	3	85.0	5.0	Trp	3	51.6	N/A
Ala	4	76.8	5.0	**Glu**	4	51.6	N/A
Thr	5	76.4	5.0	Asn	5	51.5	N/A
Asp	6	68.6	5.0	Lys	6	51.1	N/A
Glu	7	66.8	5.0	**Val**	7	50.9	N/A
Phe	8	49.2	5.0	**Leu**	8	50.7	N/A
Asn	9	41.2	5.0	Arg	9	50.3	N/A
Arg	10	38.4	5.0	**Pro**	10	50.2	N/A
Gln	11	38.4	5.0	**Ile**	11	50.1	N/A
Ser	12	37.8	5.0	His	12	49.7	N/A
Trp	13	37.4	5.0	**Ala**	13	49.7	N/A
His	14	35.8	5.0	**Asp**	14	49.6	N/A
Tyr	15	35.0	5.0	**Thr**	15	49.2	N/A
Met	16	27.6	5.0	**Gly**	16	48.6	N/A
Gly	17	26.2	5.0	Phe	17	48.2	N/A
Ile	18	24.6	5.0	**Ser**	18	47.8	N/A
Pro	19	24.0	5.0	Met	19	47.8	N/A
Cys	20	20.6	5.0	Cys	20	47.2	N/A

Table 2: **Unbiased control conditions.** Amino acids (AAs) are ranked according to their likelihood of appearance in a subset of 10 out of 20 AAs. *No transfer bias:* These results were generated by running the model with all AAs having an equal probability of horizontal transfer. AAs are ranked according to their frequency of appearance among the first 10 AAs to be encoded by the emerging genetic codes. *Random selection:* These control results were not generated with the model. Instead we repeatedly randomly selected 10 out of the 20 AAs. AAs are ranked according to their frequency of inclusion in 1,000 of these subsets.

Surprisingly, neutralizing the probabilities did not have the effect of making the rankings more similar to those produced with the original transfer bias. We find that the model again robustly converges on codes that first incorporate the same 6 amino acids of Phase 1 (Val, Ala, Leu, Thr, Glu, Asp) that had also remained part of the top 10 even under the condition of inverted probabilities. This implies that the early inclusion of these 6 amino acids happens relatively independently of their probabilities of transfer, whereas the inclusion of Gly and Ser does depend on their higher than average probabilities. This is a tantalizing finding because in the standard 20 amino acids Gly, at least, is ranked as the most prebiotically abundant.

However, what about the 6 amino acids that were largely unaffected by changes in their probability of transfer? We suspected that these amino acids must have a configuration of

chemical properties that make them more discriminable or learnable compared to the others, which would help to explain their early fixation in the codes via gradient descent.

We therefore analyzed Phase 1 and Phase 2 amino acids in terms of their respective distributions in a chemical space that is defined by three of their most important properties (Ilardo, Meringer, Freeland, Rasulev, & Cleaves II, 2015; Philip & Freeland, 2011): *size* (V_{vdW}, total volume of the molecule enclosed by the van der Waals surface), *charge* (pK_a), and *hydrophobicity* (logP, the partition coefficient). The two distributions are shown in Figure 3.

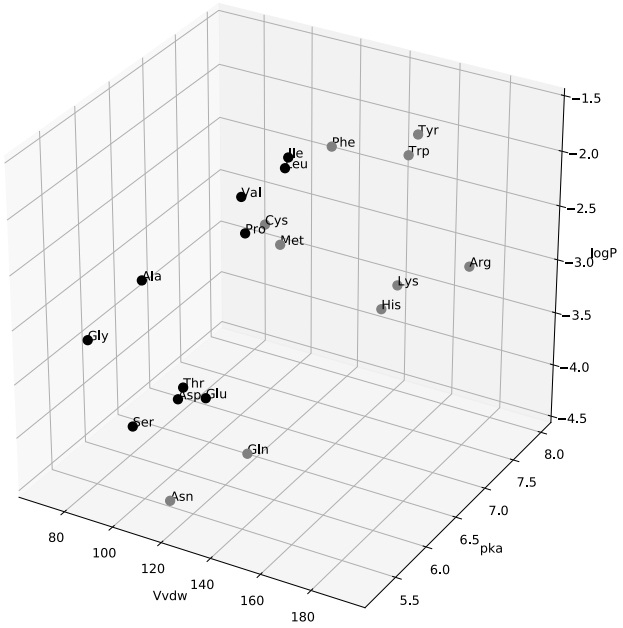

Figure 3: Distribution of 20 amino acids of the standard genetic code in chemical space. The basic chemical space is defined by three properties: size (Vvdw), charge (pKa), and hydrophobicity (logP). *Black points*: 10 amino acids that are most likely to be first incorporated according to the literature (Phase 1). *Grey points*: remaining 10 amino acids that are less easily incorporated into the standard genetic code (Phase 2).

Visual inspection of the distributions of the 10 most likely and 10 least likely incorporated amino acids reveals them to be actually rather similar. It seemed reasonable to assume that the 10 amino acids that take longer to incorporate are either distant outliers or cluster too tightly, but this is not necessarily the case. Ser and Gly could be considered outliers, and Ile is very close to Leu, but these considerations do not explain why Pro did not make into the top 10 predicted by the neutral transfer bias model, or why Ala was included but not Ser or Gly (Table 2). In general, it seems that there is a tendency for Phase 2 amino acids to be larger in size and to be more widely spread in terms of charge and hydrophobicity.

We therefore decided to compare the distributions of the neutral model in a more systematic manner. The measure of *coverage* proposed by Philip and Freeland (2011) allows us to calculate whether one set provides a more adaptive selection of amino acids than another. They define coverage in terms of the combination of two components: *breadth* is the statistical range of a property, i.e. the difference between the maximum and minimum values, and *evenness* is calculated as the statistical variance of the intervals between pairs of an ordered list of the property's values, whereby less variance means more evenness. A set of amino acids is said to provide better coverage of a given dimension of chemical space if (and only if) a chemical property provides *both* more breadth and more evenness at the same time. In order to better understand the robust rankings generated by the model even without transfer bias (Table 2), we calculated coverage of the basic chemical properties, size, charge, and hydrophobicity, of the 10 most and least frequently incorporated amino acids (Table 3). The values of the properties were published by Froese et al. (2018) in a Supplementary Information file.

		Vvdw	**pKa**	**logP**
Breadth	*Top 10 AAs*	83	2.77	3.11
	Last 10 AAs	122	2.23	2.91
Evenness	*Top 10 AAs*	48	0.16	0.09
	Last 10 AAs	117	0.02	0.04
More coverage?		Neither	Neither	Neither

Table 3: Comparison of coverage in basic chemical space. A set of amino acids (AAs) is defined as having better coverage if its properties have both more breadth (a bigger range) and more evenness (less variance of intervals). The size, charge, and hydrophobicity of the 10 most frequently incorporated AAs in the model without transfer bias have less, more, and more breadth, as well as more, less, and less evenness, respectively. Neither set of AAs therefore provides better coverage of any of the three basic properties.

Given that coverage has been interpreted as an indication of how adapted a set of amino acids is, it is interesting to note that the initial set of incorporated amino acids already has the same coverage as the later set. In other words, at least in the case of our model, there is no support for the idea that the later amino acids were added to the emerging code because of the better coverage they could provide.

And yet there must be something about the properties of the most frequently incorporated amino acids that makes a subset of them consistently more learnable by the protocells in our model. As an initial step toward determining what makes an amino acid more learnable, we analyzed their discriminability in the model's 11-dimensional chemical space by performing a principal component analysis. For each of the 11 chemical properties we calculated its mean and subtracted it from each of that property's 20 values (one value for each of the 20 amino acids). Then we calculated the standard deviation for each of the 11 chemical properties and divided each of the 20 values of that property by the property's standard deviation. This normalized data was the input to the principal component analysis. For simplicity we focused our analysis on the two most important principal components (Figure 4).

It seems that most of coevolution theory's Phase 1 amino acids are more tightly clustered into a relatively contiguous region, except for Gly and Pro. The 6 amino acids that were frequently incorporated in our model despite inverted transfer probabilities are part of this cluster. On the other hand, Phase 2 amino acids tend to be more widely spread, albeit not as much as Gly and Pro. Nevertheless, this preliminary analysis is consistent with our finding that Gly and Pro, like Phase 2

amino acids, are not so readily incorporated into the emerging codes, although their chances are improved via more frequent horizontal transfer.

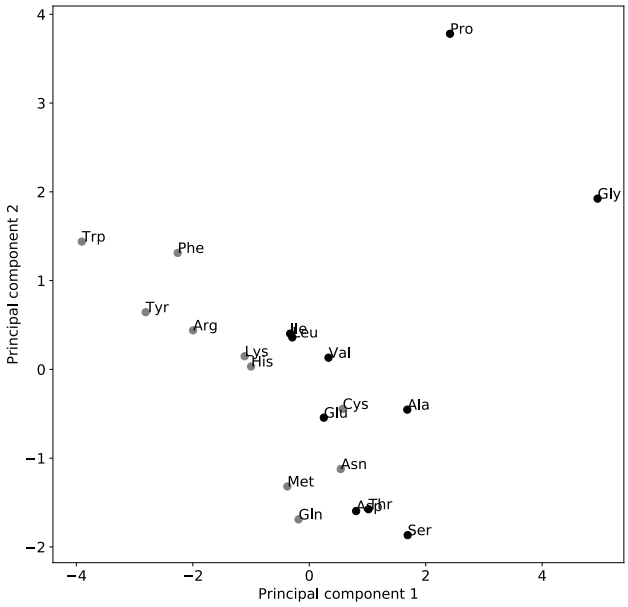

Figure 4: Principal component analysis of the distribution of the 20 amino acids of the standard genetic code. The 11-dimensional chemical space used in the model was reduced to 2 principal components. The Phase 1 amino acids proposed by coevolution theory (black dots) are comparatively more tightly clustered together, except Gly and Pro.

We also did two separate principal component analyses for the 10 most and least frequent incorporations of the original model. We found that the top 10 required fewer components to explain the same amount of variance when compared to the bottom 10. And when we examined the loadings of the first component of each analysis, we found that fewer properties loaded onto the first component of the top 10 compared to the bottom 10. In other words, the least frequent incorporations occupy a relatively more complex configuration in chemical space, and thus require more components that are distributed across more properties to distinguish between them.

In sum, these results suggest that several of Phase 1 amino acids may have become incorporated initially because they are more easily discriminated in chemical space, which would be a feature that is especially desirable at the origin of life when the chemical specificity provided by enzymes was not yet as refined. However, we are also aware that we did not perform a rigorous statistical analysis, and so at this point we simply offer this possibility as a hypothesis to guide future work.

Conclusions

An important contribution of the last decade of research in artificial life has been the development of a novel behavior-based perspective on the origins of life. A growing number of computer-based and chemical studies demonstrate that agent-environment interaction could have already played a crucial role at this early stage, e.g. by spontaneously giving rise to motility and adaptive behavior (Froese et al., 2014), and in facilitating open-ended evolution (Egbert et al., 2012). Our recent work has contributed to this behavior-based perspective by showing that conceiving of the origins of the genetic code in terms of repeated horizontal interactions between protocells leads to the emergence of artificial genetic codes that share several key properties with the standard genetic code (Froese et al., 2018).

Here we have shown the wider applicability of that model: the iterated learning approach also leads to the emergence of sequences of amino acid incorporations that are remarkably similar to the sequences that have been hypothesized in the literature. Moreover, it is notable that significant aspects of these sequences arise robustly even when all amino acids are represented as equally available, i.e. even in the absence of any representation of their relative environmental abundance under prebiotic conditions and of their dependence on certain biosynthetic pathways. We also did not find support for the hypothesis that amino acids were preferentially incorporated in terms of the increase in coverage they provide, i.e. for the breadth and evenness of their distribution in chemical space.

Instead we found indications that several amino acids were frequently incorporated into the incipient codes because they were more easily discriminated in terms of their chemical properties compared to most of the later incorporations. We employed a basic principal component analysis to evaluate this discriminability, but future work could try to apply more sophisticated statistical methods. We also found that amino acids that were less easily discriminated had their chances of incorporation increased by elevated probabilities of transfer. Future work could also try to vary these probabilities more systematically to more accurately quantify their effect.

Nevertheless, although the results of the current study are preliminary and require further confirmation and analysis, they already allow us to hypothesize that discriminability and probability of transfer were likely also important factors that shaped the actual sequence of incorporation of amino acids into the standard genetic code. This hypothesis could help to guide future research into the origins of the genetic code.

Future modeling work could also try to make the immense number of potential amino acids available for encoding in the model, and thereby verify if the emerging sequences would still tend to preferentially incorporate the 20 amino acids of the standard genetic code, thereby potentially replicating key results from another important line of research in the literature (Ilardo et al., 2015; Meringer, Cleaves II, & Freeland, 2013).

Acknowledgments. T.F. and J.I.C. were supported by CONACyT project 221341 and by UNAM-DGAPA-PAPIIT project IA104717. The development of the original model was done at ELSI during a visit by T.F. and J.I.C. that was made possible by an EON Long-Term-Visitor Award.

References

Aggarwal, N., Bandhu, A. V., & Sengupta, S. (2016). Finite population analysis of the effect of horizontal gene transfer on the origin of an universal and optimal genetic code. *Physical Biology, 13*(3), 036007. doi:10.1088/1478-3975/13/3/036007

Crick, F. H. C. (1966). Codon-anticodon pairing: The wobble hypothesis. *Journal of Molecular Biology, 19*, 548-555.

Di Paolo, E. A., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor Life: An Enactive Proposal*. Oxford, UK: Oxford University Press.

Egbert, M. D., Barandiaran, X., & Di Paolo, E. A. (2012). Behavioral metabolution: The adaptive and evolutionary potential of metabolism-based chemotaxis. *Artificial Life, 18*, 1-25.

Froese, T. (2015). Toward a behavior-based approach to the origins of life and the genetic system. In P. Andrews, L. Caves, R. Doursat, S. Hickinbotham, F. Polack, S. Stepney, T. Taylor, & J. Timmis (Eds.), *Proceedings of the European Conference on Artificial Life 2015* (pp. 397). Cambridge, MA: MIT Press.

Froese, T., Campos, J. I., Fujishima, K., Kiga, D., & Virgo, N. (2018). Horizontal transfer of code fragments between protocells can explain the origins of the genetic code without vertical descent. *Scientific Reports, 8*, 3532. doi:10.1038/s41598-018-21973-y

Froese, T., Virgo, N., & Ikegami, T. (2014). Motility at the origin of life: Its characterization and a model. *Artificial Life, 20*(1), 55-76.

Goldenfeld, N., Biancalani, T., & Jafarpour, F. (2017). Universal biology and the statistical mechanics of early life. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences, 375*, 20160341. doi:10.1098/rsta.2016.0341

Hanczyc, M. M., & Ikegami, T. (2010). Chemical basis for minimal cognition. *Artificial Life, 16*, 233-243.

Higgs, P. G., & Pudritz, R. E. (2009). A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology, 9*(5), 483-490.

Ilardo, M., Meringer, M., Freeland, S. J., Rasulev, B., & Cleaves II, H. J. (2015). Extraordinarily adaptive properties of the genetically encoded amino acids. *Scientific Reports, 5*(9414). doi:10.1038/srep09414

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology, 28*, 108-114.

Koonin, E. V., & Novozhilov, A. S. (2017). Origin and evolution of the universal genetic code. *Annual Review of Genetics, 51*, 45-62.

Meringer, M., Cleaves II, H. J., & Freeland, S. J. (2013). Beyond terrestrial biology: Charting the chemical universe of α-amino acid structures. *Journal of Chemical Information and Modeling, 53*, 2851-2862.

Moura, A., Savageau, M. A., & Alves, R. (2013). Relative amino acid composition signatures of organisms and environments. *Plos One, 8*(10), e77319. doi:10.1371/journal.pone.0077319

Philip, G. K., & Freeland, S. J. (2011). Did evolution select a nonrandom "alphabet" of amino acids? *Astrobiology, 11*(3), 235-240.

Smith, E., & Morowitz, H. (2016). *The Origin and Nature of Life on Earth: The Emergence of the Fourth Geosphere*. Cambridge, UK: Cambridge University Press.

Suzuki, K., & Ikegami, T. (2009). Shapes and self-movement in protocell systems. *Artificial Life, 15*(1), 59-70.

Vetsigian, K., Woese, C. R., & Goldenfeld, N. (2006). Collective evolution and the genetic code. *Proceedings of the National Academy of Sciences of the USA, 103*(28), 10696-10701.

Virgo, N., Froese, T., & Ikegami, T. (2013). The positive role of parasites in the origins of life *2013 IEEE Symposium on Artificial Life* (pp. 1-4): IEEE Press.

Woese, C. R. (2002). On the evolution of cells. *Proceedings of the National Academy of Sciences of the USA, 99*(13), 8742-8747.

Wong, J. T.-F. (2005). Coevolution theory of the genetic code at age thirty. *BioEssays, 27*(4), 416-425.