PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Cite this article: CG N, LaBar T, Hintze A, Adami C. 2017 Origin of life in a digital microcosm. *Phil. Trans. R. Soc. A* **375**: 20160350. http://dx.doi.org/10.1098/rsta.2016.0350

Accepted: 31 May 2017

One contribution of 18 to a theme issue 'Re-conceptualizing the origins of life from a physical sciences perspective'.

Subject Areas:
astrobiology, computational biology

Keywords:
origin of life, Avida, digital life, information theory

Author for correspondence:
Christoph Adami
e-mail: adami@msu.edu

Origin of life in a digital microcosm

Nitash C G[1,2], Thomas LaBar[2,3,4], Arend Hintze[1,2,4,5] and Christoph Adami[2,3,4,6]

[1]Department of Computer Science and Engineering, [2]BEACON Center for the Study of Evolution in Action, [3]Department of Microbiology and Molecular Genetics, [4]Program in Ecology, Evolutionary Biology and Behavior, [5]Department of Integrative Biology, and [6]Department of Physics and Astronomy, Michigan State University, East Lansing, MI 48824, USA

CA, 0000-0002-2915-9504

While all organisms on Earth share a common descent, there is no consensus on whether the origin of the ancestral self-replicator was a one-off event or whether it only represented the final survivor of multiple origins. Here, we use the digital evolution system Avida to study the origin of self-replicating computer programs. By using a computational system, we avoid many of the uncertainties inherent in any biochemical system of self-replicators (while running the risk of ignoring a fundamental aspect of biochemistry). We generated the exhaustive set of minimal-genome self-replicators and analysed the network structure of this fitness landscape. We further examined the evolvability of these self-replicators and found that the evolvability of a self-replicator is dependent on its genomic architecture. We also studied the differential ability of replicators to take over the population when competed against each other, akin to a primordial-soup model of biogenesis, and found that the probability of a self-replicator outcompeting the others is not uniform. Instead, progenitor (most-recent common ancestor) genotypes are clustered in a small region of the replicator space. Our results demonstrate how computational systems can be used as test systems for hypotheses concerning the origin of life.

This article is part of the themed issue 'Re-conceptualizing the origins of life from a physical sciences perspective'.

THE ROYAL SOCIETY PUBLISHING

1. Introduction

There is perhaps no topic in biology more fascinating—and yet more mysterious—than the origin of life. With only one example of organic life to date, we have no way of knowing whether the appearance of life on Earth was an extraordinarily rare event, or it if was a commonplace occurrence that was unavoidable given Earth's chemistry. Were we to replay Earth's history one thousand times [1], how often would it result in a biosphere? And among the cases where life emerged, how different or how similar would the emergent biochemistries be?

The role of historical contingency has been studied extensively in the evolution of life (e.g. [2] and references therein). Here, we endeavour to ask an even more fundamental question: What is the role of historical contingency in the origin of life? The best evidence suggests that the first self-replicators were RNA-based [3,4], although other first self-replicators have been proposed [5]. Given the large number of uncertainties concerning the possible biochemistry that would lead to the origin of self-replication and life, either on Earth or other planets, researchers have begun to study the process of emergence in an abstract manner. Tools from computer science [6–11], information theory [12–15] and statistical physics [16,17] have been used in an attempt to understand life and its origins at a fundamental level, removed from the peculiarities of any particular chemistry. Investigations along those lines may reveal to us general laws governing the emergence of life that are obscured by the $n = 1$ nature of our current evidence, point us to experiments that probe such putative laws and get us closer to understand the inevitability—or perhaps the elusiveness—of life itself [18].

At the heart of understanding the interplay between historical contingency and the origin of life lies the structure of the fitness landscapes of these first replicators, and how that landscape shapes the biomolecules' subsequent evolution. While the fitness landscapes of some RNA-based genotypes have been mapped [19,20] (and other RNA replicators have been evolved experimentally [21]), in all such cases evolution already had the chance to shape the landscape for these organisms and 'dictate', as it were, the sequences most conducive for evolution.

The structure of primordial fitness landscapes, in comparison, is entirely unknown. While we know, for example, that in realistic landscapes highly fit sequences are genetically close to other highly fit sequences (this is the essence of Kauffman's 'Central Massif' hypothesis [22]; see also [23]), we suspect that this convenient property—which makes fitness landscapes 'traversable' [23]—is an outcome of evolution, in particular, the evolution of evolvability. What about primordial landscapes not shaped by evolution? How often are self-replicators in the neighbourhood of other self-replicators? Are self-replicators evenly distributed among sequences, or are there (as in the landscapes of evolved sequences) vast areas devoid of self-replicators and rare (genetic) areas that teem with life? Can evolution easily take hold on such primordial landscapes?

These are fundamental questions, and they are central to our quest to understand life's origins. If the fitness landscape consists of isolated fitness networks, as found in some modern RNA fitness landscapes [19,20], then one may expect the effects of historical contingency to be strong, and the future evolution of life to depend on the characteristics of the first replicator. However, if there exist 'neutral networks' that connect genotypes across the fitness landscape (as found in computational RNA landscapes [24]), then the effect of history may be diminished. Here, we test these hypotheses explicitly using primordial fitness landscapes for digital self-replicators.

Recently, we have used the digital evolution platform Avida as a model system to study questions concerning the origin of life [25]. In Avida, a population of self-replicating computer programs undergo mutation and selection, and are thus undergoing Darwinian evolution explicitly [26]. Because the genomic content required for self-replication is non-trivial, most avidian genomes are non-viable, in the sense that they cannot form 'colonies' and thus propagate information in time. Thus, viable self-replicators are rare in Avida, with their exact abundance dependent on their information content [13,14]. Further work on these rare self-replicators showed that while most of them were evolvable to some degree, their ability to improve in replication speed or evolve complex traits greatly varied [27]. Furthermore, the capability of

avidian self-replicators to evolve greater complexity was determined by the *algorithm* they used for replication, suggesting that the future evolution of life in this digital world would be highly contingent on the original self-replicator [28]. However, all of this research was performed without a complete knowledge of the underlying fitness landscape, by sampling billions of sequences of a specific genome-size class, and testing their capacity to self-replicate.

Sequences used to seed evolution experiments in Avida are usually handwritten [29,30], for the simple reason that it was assumed that they would be impossible to find by chance. Indeed, a typical handwritten ancestral replicator of length 15 instructions is so rare—were it the only replicator among sequences of that length—that it would take a thousand processors, executing a million sequences per second each in parallel, about 50 000 years of search to find it [14]. However, it turns out that shorter self-replicators exist in Avida. An exhaustive search of all 11 881 376 sequences of length $L = 5$, as well as all 308 915 776 sequences of length $L = 6$ previously revealed no self-replicators [14]. However, in that investigation six replicators of length $L = 8$ turned up in a random search of a billion sequences of that length, suggesting that perhaps there are replicators among the 8 billion or so sequences of length $L = 7$.

Here, we confirm that the smallest replicator in Avida must have eight instructions by testing all $L = 7$ sequences, but also report mapping the entirety of the $L = 8$ landscape ($26^8 \approx 209 \times 10^9$ sequences) to investigate the fitness landscape of primordial self-replicators of that length. Mapping all sequences in this space allows us to determine the relatedness of self-replicators and study whether they occur in clusters or evenly in sequence space, all without the usual bias of studying only sequences that are among the 'chosen' already. Of the almost 209 billion possible genomes, we found that precisely 914[1] could undergo self-replication and reproduction, and thus propagate their information forward in time in a noisy environment.

We found that these 914 primordial replicators are not uniformly distributed across genetic space, but instead cluster into two broad groups (discovered earlier in larger self-replicators [28]) that form 13 main clusters. By analysing how these groups (and clusters) evolve, we are able to study how the primordial landscape shapes the evolutionary landscape, and how chance events early in evolutionary history can shape future evolution.

2. Methods

(a) Avida

We used Avida (v. 2.14) as our computational system to study the origin of self-replication. Avida is a digital evolution system in which a population of computer programs compete for the system resources needed to reproduce (see [25] for a full description of Avida). Each of these programs is self-replicating and consists of a genome of computer instructions that encode for replication. During this asexual reproduction process, mutations can occur, altering the speed at which these programs reproduce. As faster replicators will out-reproduce slower replicators, selection then leads to the spread of faster replicators. Because avidian populations undergo Darwinian evolution, Avida has been used to explore many complex evolutionary processes [31–37].

The individual computer programs in Avida are referred to as avidians. They consist of a genome of computer instructions and different containers to store numbers. Each genome has a defined start point and instructions are sequentially executed throughout the avidian's lifetime. Some of these instructions allow the avidian to start the replication process, copy their genome into a new daughter avidian, and divide into two avidians (see appendix A and [28] for the full Avida instruction set). During this replication process, mutations can occur, causing the daughter avidian's genome to differ from its parent. These mutations can have two broad phenotypic outcomes. First, mutations can alter the number of instruction executions required for replication; these mutations can increase or decrease replication speed and thus fitness. Second, the fixation of multiple mutations can lead to the evolution of complex traits in Avida. These traits are the

[1]The sequences of all replicators can be downloaded from http://dx.doi.org/10.6084/m9.figshare.4551559.

ability to input binary numbers from the Avida environment, perform Boolean calculations on these numbers and then output the result of those calculations. In the experiments described here, avidians could evolve any of the nine one- and two-input logic functions (Not, Nand, OrNot, And, Or, AndNot, Nor, Xor and Equals). This is usually referred to as the 'logic-9' environment [38].

The ability to perform the above Boolean logic calculations (to possess any of these nine traits) increases its bearer's replication speed by increasing the number of genome instructions the bearer can execute per unit of time. The more instructions an avidian can execute during a unit of time, the fewer the units of time that are required for self-replication. These units of time are referred to as updates (they are different from generations). During each update, the entire population will execute $30N$ instructions, where N is the current population size. The unit of energy necessary to execute one instruction is called a 'single instruction processing' unit, or SIP. If the population is monoclonal, each avidian will receive, on average, 30 SIPs. However, every avidian also has a *merit* which determines how many SIPs they receive per update. The greater the merit, the more SIPs that individual receives. The ability to perform the nine calculations multiply an individual's merit by the following values: Not and Nand: 2; OrNot and And: 4; AndNot and Or: 8; Nor and Xor: 16; and Equals: 32.

The Avida world consists of a fixed-size toroidal grid of cells. The total number of cells sets the maximum population size. Each cell can be occupied by at most one avidian. After successful reproduction, a new avidian is placed into one of the world's cells. In a well-mixed population, any cell in the population may be chosen, rendering the toroidal structure meaningless. In a population with spatial structure, the new avidian is placed into one of the nine cells neighbouring the parent avidian (including the cell occupied by the parent). If there are empty cells available, the new avidian occupies one of these cells. If all possible cells are occupied, a cell is chosen at random, its occupant removed from the population and the new avidian then occupies this cell. This random removal implements a form of genetic drift in Avida. For the experiments performed here, the population structure was spatial.

(b) Experimental design

To map the entire Avida fitness landscape, we constructed all $26^8 \approx 2.09 \times 10^{11}$ genomes and analysed whether they could self-replicate. This operation was performed by running these genomes through Avida's *Analyze Mode* (described in the Data analysis section) and checking whether these genomes gave their bearer non-zero fitness, and whether they were *viable*. Next, we described the fitness landscape by looking for the presence of genotype clusters among the discovered self-replicators. We constructed a network of the fitness landscape where each genotype is a node and the length between two nodes is the square of the Hamming distance between the genotypes. We also examined the frequency of single instruction motifs (monomers), as well as double instruction motifs (dimers).

To test the evolvability of the 914 self-replicators, we evolved 10 monoclonal populations of each replicator with 3600 individuals for 2×10^4 updates in the logic-9 environment (see above). Point mutations occurred at a rate of 7.5×10^{-3} mutations per copied instruction, while single-instruction insertion and deletion mutations both occurred at a rate of 5×10^{-2} mutations per division. At the end of each population's evolution, we analysed the most abundant genotype from each population.

To test the role of historical contingency when the appearance of self-replicators was frequent, we ran experiments where we evolved all 914 self-replicators in the same population (a 'primordial soup' of replicators). In each population, we placed 10 individuals of each self-replicator. The ancestral population then had 9140 individuals and could expand to 10^4 individuals at maximum capacity. These populations evolved for 5×10^4 updates in the logic-9 environment. Mutation rates were the same as in the previous evolvability experiments. This experiment was performed in 200 replicates. To identify the ancestral genotype that outcompeted all of the other genotypes, we isolated the most abundant genotype at the end of the experiment and traced its evolutionary history back to its original ancestor.

(c) Data analysis

Statistics on different avidians were calculated using Avida's *Analyze Mode*. In Analyze Mode, a single genotype is examined in isolation as it executes the instructions in its genome, runs through its life cycle and possibly creates an offspring. This confers on experimenters the ability to calculate the fitness for an individual avidian (number of offspring generated per unit time) and examine other characteristics, such as whether it can reproduce perfectly (all offspring are genetically identical to each other and the mother genome) or which traits this avidian possesses. Analyze Mode was also used to calculate quantities such as genome size. Avida's Analyse Mode code is available along with the entire Avida software at https://github.com/devosoft/avida.

Across-population means and standard errors were calculated using the NumPy [39] Python software package. The clusters of replicators were rendered using Neato, which is an undirected graph embedder that creates a layout similar to that of multi-dimensional scaling [40]. Figures were plotted using the Matplotlib Python package [41].

3. Results

(a) Structure of the fitness landscape

Of the 26^8 (approx. 209 billion) genomes with eight instructions, we found 914 that could self-replicate. We also searched for self-replicators with seven-instruction genomes but found none, establishing that $L = 8$ is the minimal self-replicator length in Avida. By discovering all self-replicators in this fitness landscape, we can now calculate the precise information content required for self-replication in Avida, using previously established methods [13], as $-\log_{26}(914/26^8) \approx 5.9$ mers (a 'mer' is a unit of entropy or information, normalized by the number of states that each instruction can take on; see [42]). Our previous estimate [14] of the information content of length-8 replicators, based on finding eight replicators among a billion random samples, was 5.81 ± 0.13 mers.

To study the genetic structure of these replicators, we obtained the distribution of instructions (monomers) across the replicators' genomes (figure 1*a*). This distribution is biased, as every single replicator contained at least the three instructions required for replication: h-copy, h-alloc and h-divide (denoted by `v`, `w` and `x`, respectively; see the mapping between instructions and the letter mnemonic in table 1 in appendix A). In addition, 75% of replicators have a `b` (nop-B), an `f` (if-label) and a `g` (mov-head) instruction, while 25% have a `c` (nop-C), an `h` (jmp-head) and an `r` (swap) instruction in their sequence. We also analysed the distribution of sequential instruction pairs (dimers) and found that while most dimers do not occur in any self-replicators, the dimers `fg` and `gb` occur in approximately 70% of the replicators (figure 1*b*) and are highly over-represented. Other dimers such as `rc`, `hc`, and dimers containing `f`,`g`,`b`,`c`,`v`,`w`, and `x` occur in approximately 20–30% of replicators.

If there were no constraint on the genetic architecture, we would expect self-replicators to be distributed uniformly across the fitness landscape. However, we found instead that self-replicators are not distributed uniformly in the landscape, but are grouped into 41 distinct genotype clusters, shown in figure 2.

The dimer distribution function we analysed above separates primordial self-replicators into two major categories: those that carry `fg`/`gb` motifs ('fg-replicators' for short), as opposed to those carrying `hc`/`rc` motifs (hc-replicators) instead. This separation into two classes was noted earlier from a smaller sample of the landscape [27,28], which we corroborate here. By scanning the entire landscape, we can confirm that these two types are the only types of self-replicators in the landscape, and the clusters of genotypes are homogeneous in the sense that fg-replicators and hc-replicators do not intermix (figure 2). Figure 3 shows four examples of clusters pulled from the landscape, showing that they are tightly interconnected.

Many self-replicators are isolated and 20 of these clusters consist of only one genotype. However, most self-replicators are located in large clusters. Almost 75% of the self-replicators

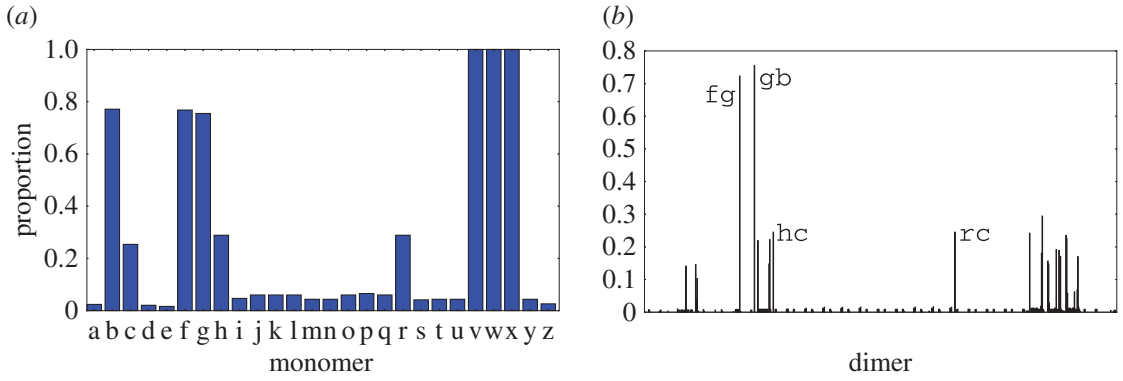

Figure 1. (*a*) Distribution of monomers/single instructions (i.e. proportion of self-replicators containing a given monomer). (*b*) Distribution of dimers (pairs of instructions). Dimers are ordered lexicographically on the *x*-axis (the proportions of `fg`, `gb`, `rc` and `hc` dimers are labelled). (Online version in colour.)

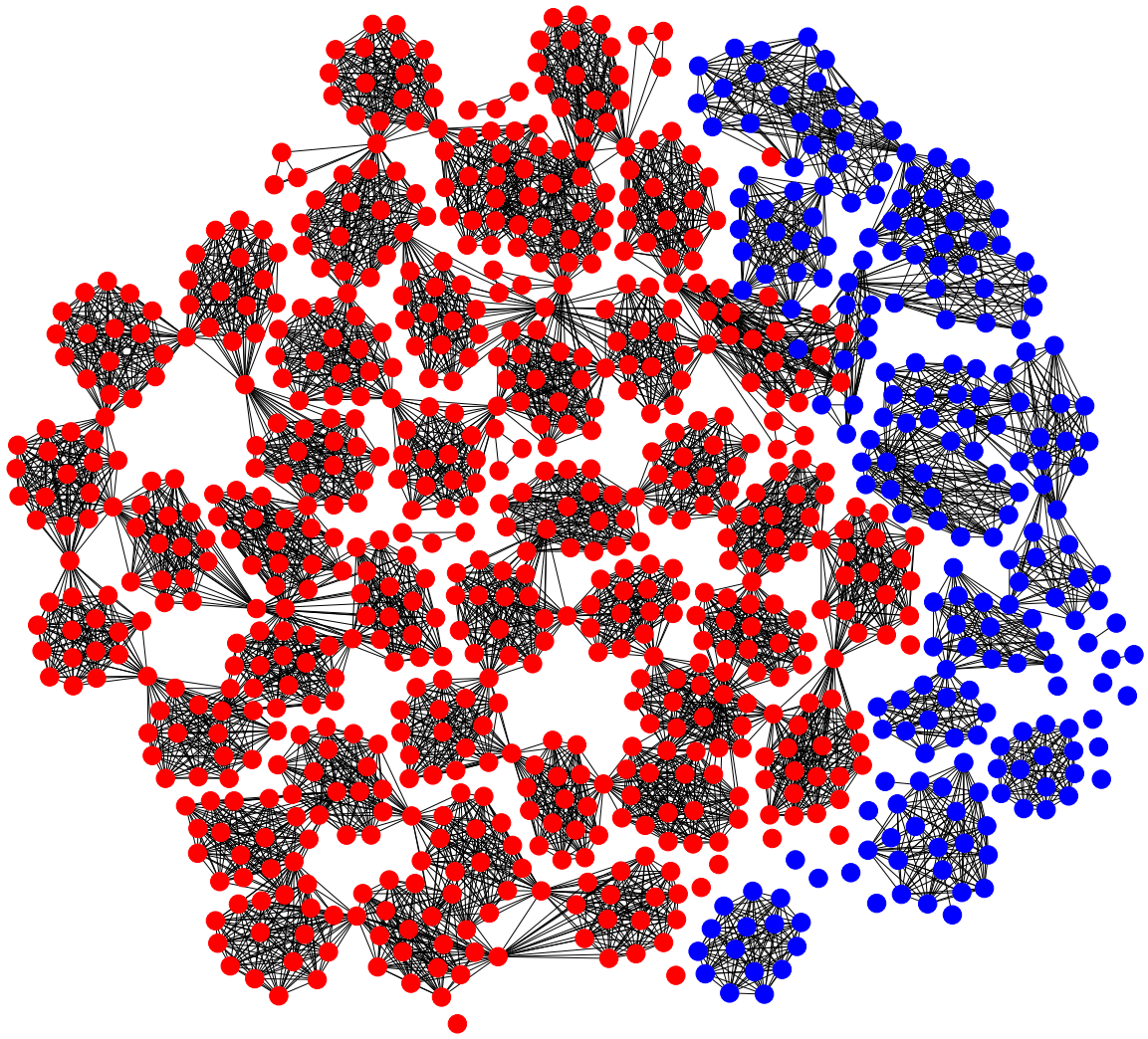

Figure 2. The complete fitness landscape of all 914 length-8 replicators. The replicators are coloured by the class of motifs they contain (fg-replicators are coloured in red, while hc-replicators are coloured in blue). The relative position between any pair of nodes reflects their distance in Hamming space, displayed via multi-dimensional scaling. As a consequence, it appears as if blue and red clusters are linked, which is not the case. One isolated fg-replicator (red) is close to an hc-replicator cluster (blue), but is not connected to it. All visible edges are between nodes that have a Hamming distance of 1 (i.e. they are a point mutation away from each other). (Online version in colour.)

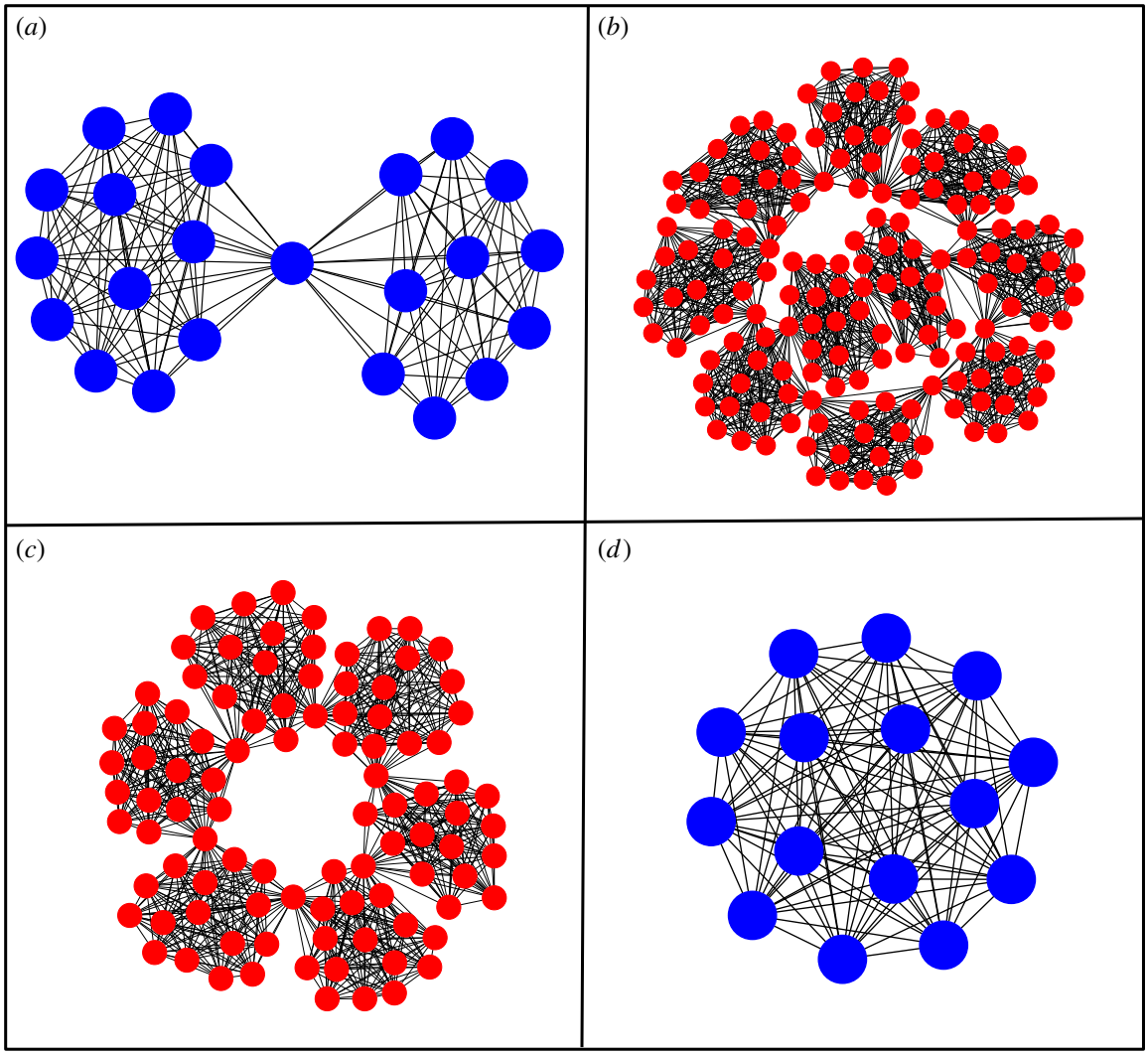

Figure 3. Four clusters from the full landscape of self-replicators of $L = 8$. (*a*) A 23-node cluster of hc-replicators. (*b*) The third-largest cluster in the network: an fg-replicator cluster with 165 members. (*c*) Another large fg-replicator cluster with 96 genotypes. (*d*) A 15-node hc-replicator cluster. (Online version in colour.)

are located in four major clusters with 212, 199, 165 and 95 genotypes each, and almost 96% are contained within the 13 clusters that have at least 14 members. There is thus a distinct gap in the cluster size distribution, with small clusters ranging from 1 to 3 connected members, while the next largest size class is 14.

We find that clusters of replicators are highly connected among each other, with a degree distribution that is sharply peaked around the mean degree of a cluster (figure 4), which is similar to what is seen in neutral networks of random RNA structures [43]. We find that fg-replicators form the denser clusters.

The 914 self-replicators we found vary in fitness, but consistently we find that the fittest self-replicators contain the `fg/gb` motifs and many of the lowest fitness self-replicators contain the `hc/rc` motifs. In figure 5, we show the fitness as a function of the multi-dimensional scaling coordinate. In that figure, colour denotes fitness according to the scale on the right. The highest peaks and plateaus all belong to fg-replicators. The hc-replicators appear as a valley (dark blue) bordering the group of fg-replicators.

(b) Self-replicator evolvability

To explore the subsequent role of historical contingency after the emergence of life, we tested the evolvability of all 914 self-replicators. First, we evolved each replicator separately. Almost

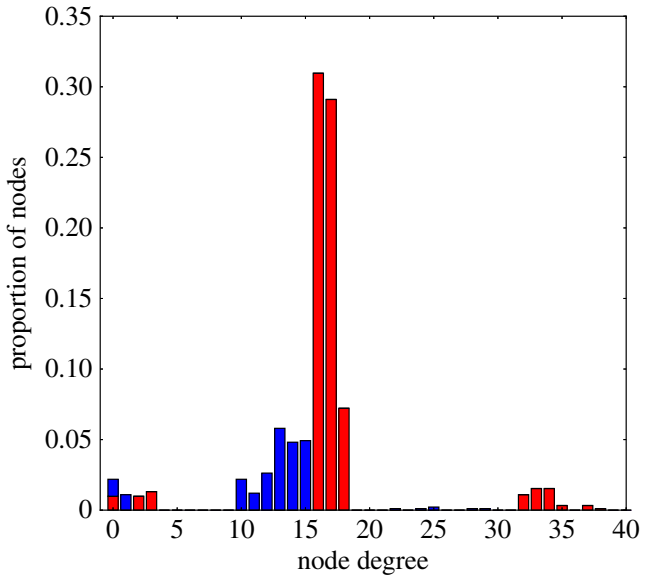

Figure 4. Edge distribution of all replicators in the fitness landscape of $L = 8$. As each cluster has a particular edge distribution, the distributions of the two different kinds of replicators (fg-types and hc-types) do not overlap. Red: fg-replicators; blue: hc-replicators. (Online version in colour.)

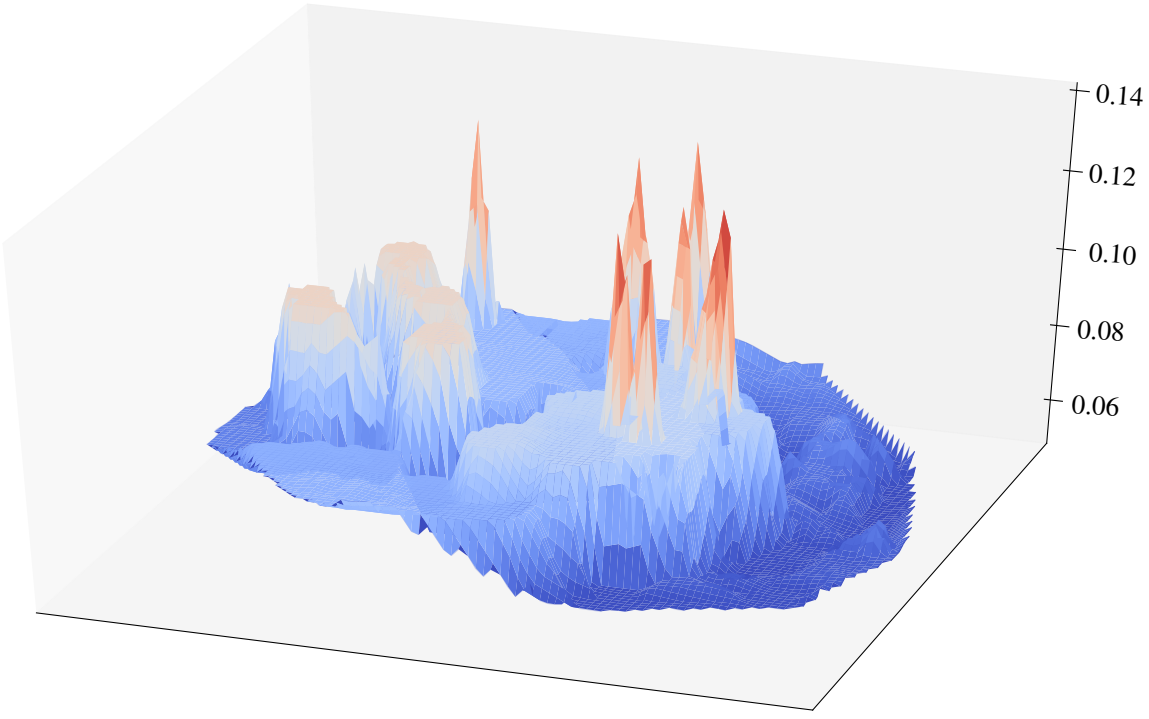

Figure 5. Ancestral fitness of all primordial self-replicators of $L = 8$, where x–y coordinates are the same as the network in figure 2. (Online version in colour.)

all self-replicators could evolve increased fitness (figure 6*b*). However, there was a wide range of mean relative fitness; fg-replicators clearly undergo more adaptation than hc-replicators. To explain why fg-replicators were more evolvable, we first looked at the evolution of genome size. Replicators with the `fg/gb` motifs grew larger genomes than replicators with the `hc/rc` motifs (figure 6*c*). As larger genomes can allow for the evolution of novel traits in Avida, and thus fitness increases, we next checked whether the fg-replicators had evolved more *computational traits* than

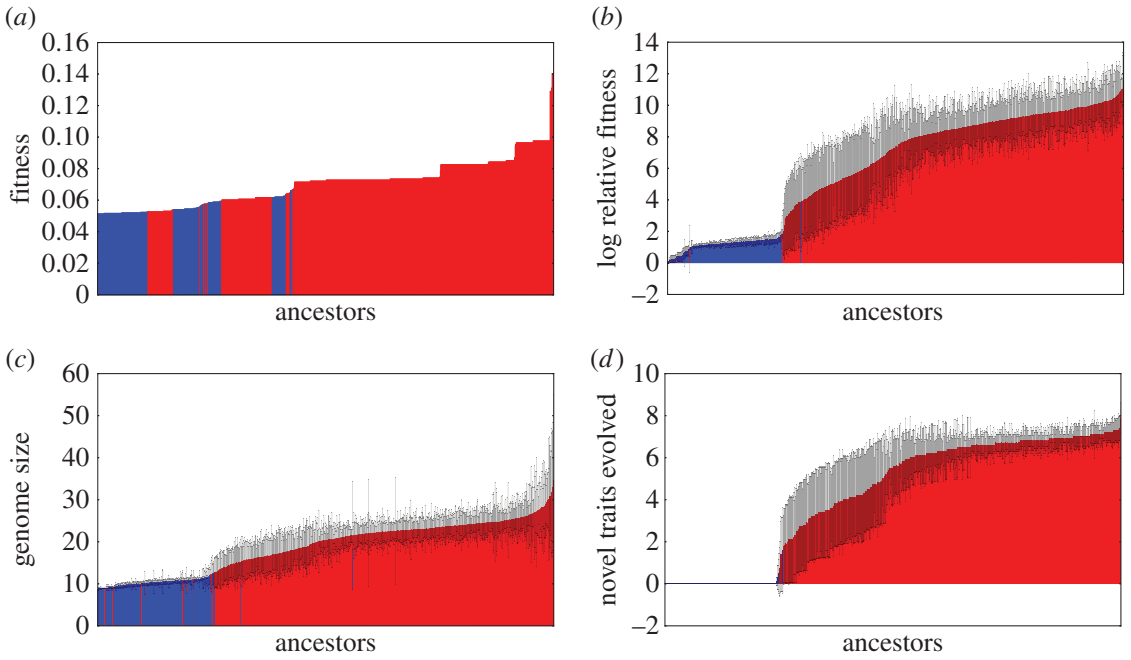

Figure 6. Fitness and other characteristics of all $L = 8$ self-replicators before and after evolution. (*a*) Ancestral fitness of all replicators. (*b*) Log mean relative fitness after 2×10^4 updates of evolution. (*c*) Final genome size after 2×10^4 updates of evolution. (*d*) Number of evolved traits after 2×10^4 updates of evolution. In all plots, fg-replicators are in red and hc-replicators are in blue. Error bars (black) are twice the standard error of the mean. All plots are sorted in increasing order. (Online version in colour.)

the hc-replicators. In Avida, traits are snippets of code that allow the avidian to gain energy from the environment, by performing logic operations on binary numbers that the environment provides (see Methods). Replicators with the `fg/gb` motifs did evolve more novel traits than replicators with the `hc/rc` motifs (figure 6*d*). In fact, only fg-replicators evolved traits in these experiments. Finally, we looked at the effect of historical contingency when all 914 replicators were competed against each other in one population. After 50 000 updates, we identify the most abundant genotype in 200 replicate experiments and reconstruct the line-of-descent to determine which of the replicators gave rise to it (we call that replicator the 'progenitor').

Most replicators did not emerge as the progenitor of life in these experiments (figure 7). Three genotypes, `vvwfgxgb`, `vwvfgxgb` and `wvvfgxgb`, outcompete the other genotypes in 37, 49 and 45 populations out of 200, respectively, or in about 65% of the competitions. The other progenitors of life were not distributed randomly among the other self-replicators either; most of them were present in the same clusters as the three genotypes from above. Thus, while history is a factor in which of the replicators becomes the seed of all life in these experiments, more than half the time the progenitor is one of the three highest fitness sequences. Thus, life predominantly originates from the highest peaks of the primordial landscape.

4. Discussion

Here, we tested the role of fitness landscape structure and historical contingency in the origin of self-replication in the digital evolution system Avida. We characterized the complete fitness landscape of all minimal-genome self-replicators and found that viable genotypes form clusters in the fitness landscape. These self-replicators can be separated into two replication classes, as we previously found for self-replicators with larger genomes [28]. We also found that one of these replication classes (the fg-replicators) is more evolvable than the other, although the evolvability of each genotype varies. Finally, we show that, when all self-replicators are competed against each

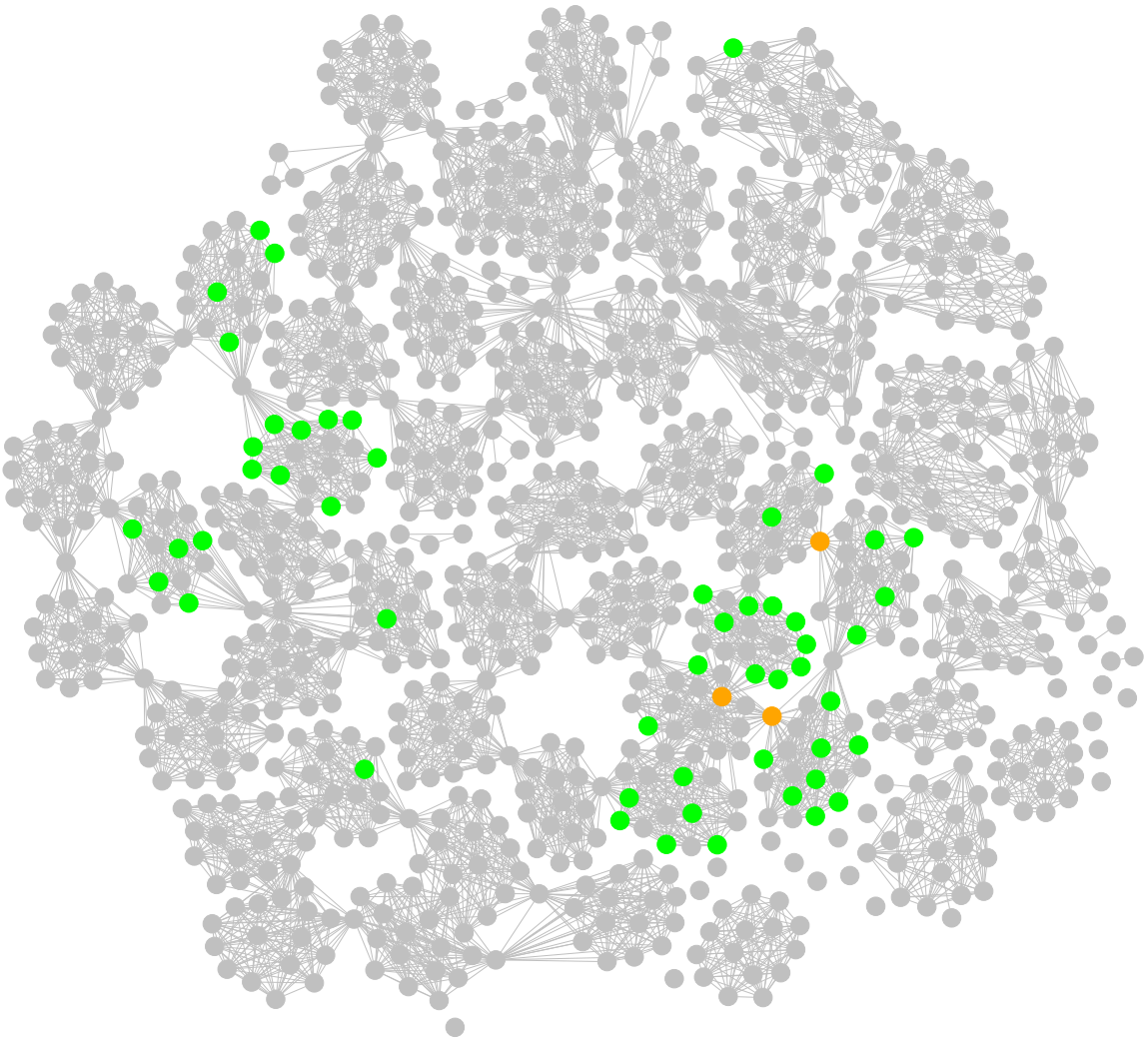

Figure 7. Location of 'progenitors' (ancestral types that were the origin of an evolved population 50 000 updates later) in the primordial landscape. Replicators that were never the ancestor genotype of the entire population are in grey. Those that outcompete all other genotypes in fewer than 6 (out of 200) competitions are coloured in green. The three genomes that eventually become the ancestor of life in over 130 competitions are in orange. (Online version in colour.)

other in a digital 'primordial soup', three genotypes win over 65% of the competitions and many of the other 'winners' come from the same genotype cluster.

In a previous study with Avida, we found that 6 out of 10^9 spontaneously emergent genomes with eight instructions could self-replicate [14]. Here, we found that 914 out of $\approx 2.8 \times 10^{11}$ genomes could replicate, consistent with our previous results. This concordance suggests that the information-theoretic theory of the emergence of life, originally proposed by Adami [13] and tested with Avida by Adami & LaBar [14], can accurately explain the likelihood of the chance emergence of life. Thus, the emergence of self-replication, and life, is dependent on the information required for such life.

By enumerating all of the length-8 self-replicators, we were able to show that self-replicators are not uniformly distributed across the fitness landscape and that viable genotypes cluster together. The size of these clusters varies: there are a few clusters with many genotypes and many clusters with a few genotypes, but the cluster size distribution has a gap. The edge distribution of the clusters is similar to what has been found in random RNA structures, and the mean degree differs between replicator types.

Genotypes with different replication mechanisms were in different clusters with no evolutionary trajectory between the two. Empirical studies of RNA-based fitness landscapes,

biochemical model systems for the origin of life, also show that these landscapes consist of isolated fitness peaks with many non-viable genotypes [19,20]. The fact that both RNA-based landscapes [19,20] and these digital landscapes have similar structures suggests that the evolutionary patterns we see in these Avida experiments may be similar to those one would have seen in the origin of life on Earth. The presence of isolated genotype clusters in both digital and RNA fitness landscapes further suggests that the identity of the first self-replicator may determine life's future evolution, as other evolutionary trajectories are not accessible. However, if populations can evolve larger genomes, non-accessible evolutionary trajectories may later become accessible, as mathematical results on the structure of high-dimensional fitness landscapes suggest [44].

To test for the effects of historical contingency in the origin of self-replication in Avida, we evolved all of the 914 replicators in an environment where they could increase in genome size and evolve novel traits. Previously, we found that the evolvability of spontaneously emergent self-replicators varied and was determined by their replication mechanism [28]. However, those genotypes possessed fixed-length genomes of 15 instructions. Here, we confirmed that the genotype of the first self-replicator, and more specifically the replication mechanism of the first replicator, determine the future evolution of novel traits in Avida. The fg-replicators showed high rates of trait evolution, while hc-replicators failed to evolve novel traits in most populations. However, we did not detect any trade-off in evolvability, as we previously found [28]. This difference is probably due to their differences in capacity to increase in genome size, as genome size increases enhance the evolution of novel traits and fitness increases in Avida [45,46]. Would a similar dynamic occur in a hypothetical population of RNA-based replicators? While experimental evolution of RNA replicators has been performed, the selective environments resulted in genome size decreases [21]. It is unknown how simple RNA replicators vary in their evolvability.

We also performed experiments to test for the role of historical contingency in scenarios where any self-replicator could become the progenitor of digital life. Here, we found that only three self-replicators (or their neighbours in the fitness landscape) became the last common ancestor in the majority of populations. This suggests a lack of contingency in the ancestral self-replicator, but emphasizes the role of the ancestral genotype in determining its future evolution. If life emerges rarely, then its future evolution will be determined by the specific genotype that first emerges, as shown from our first set of evolvability experiments (figure 6). However, if simple self-replicators emerge frequently, then the future evolution is determined by the evolvability of the fittest replicators, a sort of clonal interference [47] among possible progenitors of life. In this case, the self-replicators that most successfully invaded the population happened to also be of the type that evolved the largest genomes and most complex traits. However, it can be imagined that the opposite trend could occur [28], and then the progenitor of life would limit the future evolution of biological complexity.

5. Conclusion

In this work, we have performed the first complete mapping of a primordial sequence landscape in which replicators are extremely rare (about one replicator per 200 million sequences) and found two functionally inequivalent classes of replicators that differ in their fitness as well as evolvability, and that form distinct (mutationally disconnected) clusters in sequence space. In direct evolutionary competition, only the highest fitness sequences manage to repeatedly become the common ancestor of all life in this microcosm, showing that despite significant diversity of replicators, historical contingency plays only a minor role during early evolution.

While it is unclear how the results we obtained in this digital microcosm generalize to a biochemical microcosm, we are confident that they can guide our thinking about primordial fitness landscapes. The functional sequences we discovered here are extremely rare, but probably not as rare as putative biochemical primordial replicators. However, from a purely statistical point of view, it is unlikely that a primordial landscape consisting of sequences that are several

orders of magnitude more rare would look qualitatively different, nor would we expect our results concerning historical contingency to change significantly. After all, random functional RNA sequences (but not replicators, of course) within a computational world [43], chosen only for their ability to fold, show similar clustering and degree distributions as we find here. Follow-up experiments in the much larger $L = 9$ landscape (currently underway) will reveal which aspects of the landscape are specific, and which ones are germane, in this digital microcosm. A comparison between fitness landscapes across a variety of evolutionary systems, both digital [48] and biochemical [19], will further elucidate commonalities expected for simple self-replicators. As the landscapes for these simple self-replicators are mapped, we expect general properties of primordial fitness landscapes to emerge, regardless of the nature of the replicator. As long as primordial self-replicators anywhere in the universe consist of linear heteropolymers that encode the information necessary to replicate, studies with digital microcosms can give us clues about the origin of life that experiments with terrestrian biochemistry cannot deliver.

Data accessibility. The set of self-replicators can be found at http://dx.doi.org/10.6084/m9.figshare.4551559. Avida (v. 2.14) can be downloaded from http://avida.devosoft.org/.

Authors' contributions. N.C.G. and T.L. carried out the experiments and analysed the data. C.A., A.H., N.C.G. and T.L. conceived of and designed the study, and drafted the manuscript. All the authors read and approved the manuscript.

Competing interests. The authors declare that they have no competing interests.

Funding. This work was supported in part by the National Science Foundation's BEACON Center for the Study of Evolution in Action under Cooperative Agreement DBI-0939454.

Acknowledgements. We wish to acknowledge the support of the Michigan State University High Performance Computing Center and the Institute for Cyber Enabled Research (iCER).

Appendix A

Table 1. Instruction set of the avidian programming language used in this study. The notation ?BX? implies that the command operates on a register specified by the subsequent nop instruction (for example, nop-A specifies the AX register, and so forth). If no nop instruction follows, use the register BX as a default. More details about this instruction set can be found in [25].

instruction	description	symbol
nop-A	no operation (type A)	a
nop-B	no operation (type B)	b
nop-C	no operation (type C)	c
if-n-equ	execute next instruction only if ?BX? does not equal complement	d
if-less	execute next instruction only if ?BX? is less than its complement	e
if-label	execute next instruction only if template complement was just copied	f
mov-head	move instruction pointer to same position as flow-head	g
jmp-head	move instruction pointer by fixed amount found in register CX	h
get-head	write position of instruction pointer into register CX	i
set-flow	move the flow-head to the memory position specified by ?CX?	j
shift-r	shift all the bits in ?BX? one to the right	k
shift-l	shift all the bits in ?BX? one to the left	l
inc	increment ?BX?	m
dec	decrement ?BX?	n
push	copy value of ?BX? onto top of current stack	o

(*Continued*.)

Table 1. (*Continued*.)

instruction	description	symbol
pop	remove number from current stack and place in ?BX?	p
swap-stk	toggle the active stack	q
swap	swap the contents of ?BX? with its complement	r
add	calculate sum of BX and CX; put result in ?BX?	s
sub	calculate BX minus CX; put result in ?BX?	t
nand	perform bitwise NAND on BX and CX; put result in ?BX?	u
h-copy	copy instruction from read-head to write-head and advance both	v
h-alloc	allocate memory for offspring	w
h-divide	divide off an offspring located between read-head and write-head	x
IO	output value ?BX? and replace with new input	y
h-search	find complement template and place flow-head after it	z

References

1. Gould SJ. 1990 *Wonderful life: the Burgess Shale and the nature of history*. New York, NY: W.W. Norton.
2. Blount Z, Borland C, Lenski R. 2008 Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **105**, 7899–7906. (doi:10.1073/pnas.0803151105)
3. Gilbert W. 1986 Origin of life: the RNA world. *Nature* **319**, 618. (doi:10.1038/319618a0)
4. Robertson MP, Joyce GF. 2012 The origins of the RNA world. *Cold Spring Harb. Perspect. Biol.* **4**, a003608. (doi:10.1101/cshperspect.a003608)
5. Leslie EO. 2004 Prebiotic chemistry and the origin of the RNA world. *Crit. Rev. Biochem. Mol. Biol.* **39**, 99–123. (doi:10.1080/10409230490460765)
6. Pargellis AN. 1996 The spontaneous generation of digital 'life'. *Phys. D: Nonlinear Phenom.* **91**, 86–96. (doi:10.1016/0167-2789(95)00268-5)
7. Hutton TJ. 2002 Evolvable self-replicating molecules in an artificial chemistry. *Artif. Life* **8**, 341–356. (doi:10.1162/106454602321202417)
8. Pargellis A. 2003 Self-organizing genetic codes and the emergence of digital life. *Complexity* **8**, 69–78. (doi:10.1002/cplx.10095)
9. Dorn ED, Nealson KH, Adami C. 2011 Monomer abundance distribution patterns as a universal biosignature: examples from terrestrial and digital life. *J. Mol. Evol.* **72**, 283–295. (doi:10.1007/s00239-011-9429-4)
10. Walker SI, Davies PC. 2013 The algorithmic origins of life. *J. R. Soc. Interface* **10**, 20120869. (doi:10.1098/rsif.2012.0869)
11. Greenbaum B, Pargellis A. 2017 Self-replicators emerge from a self-organizing prebiotic computer world. *Artif. Life* **23**, 318–342. (doi:10.1162/ARTL_a_00234)
12. Walker SI. 2014 Top-down causation and the rise of information in the emergence of life. *Information* **5**, 424–439. (doi:10.3390/info5030424)
13. Adami C. 2015 Information-theoretic considerations concerning the origin of life. *Orig. Life Evol. Biosph.* **45**, 309–317. (doi:10.1007/s11084-015-9439-0)
14. Adami C, LaBar T. 2017 From entropy to information: biased typewriters and the origin of life. In *From matter to life: information and causality* (eds SI Walker, PCW Davies, GFR Ellis), pp. 130–154. Cambridge, UK: Cambridge University Press.
15. Davies PC, Walker SI. 2016 The hidden simplicity of biology. *Rep. Prog. Phys.* **79**, 102601. (doi:10.1088/0034-4885/79/10/102601)
16. England JL. 2013 Statistical physics of self-replication. *J. Chem. Phys.* **139**, 121923. (doi:10.1063/1.4818538)

17. Mathis C, Bhattacharya T, Walker SI. 2015 The emergence of life as a first order phase transition. (http://arxiv.org/abs/1503.02776)
18. Cronin L, Walker SI. 2016 Beyond prebiotic chemistry. *Science* **352**, 1174–1175. (doi:10.1126/science.aaf6310)
19. Jiménez JI, Xulvi-Brunet R, Campbell GW, Turk-MacLeod R, Chen IA. 2013 Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proc. Natl Acad. Sci. USA* **110**, 14 984–14 989. (doi:10.1073/pnas.1307604110)
20. Petrie KL, Joyce GF. 2014 Limits of neutral drift: lessons from the in vitro evolution of two ribozymes. *J. Mol. Evol.* **79**, 75–90. (doi:10.1007/s00239-014-9642-z)
21. Mills D, Peterson R, Spiegelman S. 1967 An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl Acad. Sci. USA* **58**, 217–224. (doi:10.1073/pnas.58.1.217)
22. Kauffman SA. 1993 *The origins of order: self-organization and selection in evolution*. New York, NY: Oxford University Press.
23. Østman B, Hintze A, Adami C. 2010 Critical properties of complex fitness landscapes. In *Proc. 12th Int. Conf. on Artificial Life* (eds H Fellerman, M Dörr, MM Hanczyc, L Ladegaard Laursen, S Maurer, D Merkle, PA Monnard, K Stoy, S Rasmussen), pp. 126–132. Cambridge, MA: MIT Press.
24. Huynen MA, Stadler PF, Fontana W. 1996 Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl Acad. Sci. USA* **93**, 397–401. (doi:10.1073/pnas.93.1.397)
25. Ofria C, Bryson DM, Wilke CO. 2009 Avida: a software platform for research in computational evolutionary biology. In *Artificial life models in software* (ed. AA Maciej Komosinski), pp. 3–35. London, UK: Springer.
26. Pennock RT. 2007 Models, simulations, instantiations, and evidence: the case of digital evolution. *J. Exp. Theor. Artif. Intell.* **19**, 29–42. (doi:10.1080/09528130601116113)
27. LaBar T, Adami C, Hintze A. 2015 Does self-replication imply evolvability? In *Proc. European Conf. on Artificial Life 2015*, pp. 595–602. Cambridge, MA: MIT Press.
28. LaBar T, Hintze A, Adami C. 2016 Evolvability tradeoffs in emergent digital replicators. *Artif. Life* **22**, 483–498. (doi:10.1162/ARTL_a_00214)
29. Adami C. 1998 *Introduction to artificial life*. New York, NY: Springer.
30. Adami C. 2006 Digital genetics: unravelling the genetic basis of evolution. *Nat. Rev. Genet.* **7**, 109–118. (doi:10.1038/nrg1771)
31. Lenski RE, Ofria C, Collier TC, Adami C. 1999 Genome complexity, robustness and genetic interactions in digital organisms. *Nature* **400**, 661–664. (doi:10.1038/23245)
32. Adami C, Ofria C, Collier TC. 2000 Evolution of biological complexity. *Proc. Natl Acad. Sci. USA* **97**, 4463–4468. (doi:10.1073/pnas.97.9.4463)
33. Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C. 2001 Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* **412**, 331–333. (doi:10.1038/35085569)
34. Chow SS, Wilke CO, Ofria C, Lenski RE, Adami C. 2004 Adaptive radiation from resource competition in digital organisms. *Science* **305**, 84–86. (doi:10.1126/science.1096307)
35. Covert AW, Lenski RE, Wilke CO, Ofria C. 2013 Experiments on the role of deleterious mutations as stepping stones in adaptive evolution. *Proc. Natl Acad. Sci. USA* **110**, E3171–E3178. (doi:10.1073/pnas.1313424110)
36. Goldsby HJ, Knoester DB, Ofria C, Kerr B. 2014 The evolutionary origin of somatic cells under the dirty work hypothesis. *PLoS Biol.* **12**, e1001858. (doi:10.1371/journal.pbio.1001858)
37. Zaman L, Meyer JR, Devangam S, Bryson DM, Lenski RE, Ofria C. 2014 Coevolution drives the emergence of complex traits and promotes evolvability. *PLoS Biol.* **12**, e1002023. (doi:10.1371/journal.pbio.1002023)
38. Lenski RE, Ofria C, Pennock RT, Adami C. 2003 The evolutionary origin of complex features. *Nature* **423**, 139–144. (doi:10.1038/nature01568)
39. Van Der Walt S, Colbert SC, Varoquaux G. 2011 The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30. (doi:10.1109/MCSE.2011.37)
40. Gansner ER, North SC. 2000 An open graph visualization system and its applications to software engineering. *Software Pract. Exp.* **30**, 1203–1233. (doi:10.1002/1097-024X(200009)30:11<1203::AID-SPE338>3.0.CO;2-N)
41. Hunter JD. 2007 Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95. (doi:10.1109/MCSE.2007.55)

42. Adami C. 2004 Information theory in molecular biology. *Phys. Life Rev.* **1**, 3–22. (doi:10.1016/j.plrev.2004.01.002)
43. Aguirre J, Buldú JM, Stich M, Manrubia SC. 2011 Topological structure of the space of phenotypes: the case of RNA neutral networks. *PLoS ONE* **6**, e26324. (doi:10.1371/journal.pone.0026324)
44. Gavrilets S. 1997 Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol.* **12**, 307–312. (doi:10.1016/S0169-5347(97)01098-7)
45. Gupta A, LaBar T, Miyagi M, Adami C. 2016 Evolution of genome size in asexual digital organisms. *Sci. Rep.* **6**, 25786. (doi:10.1038/srep25786)
46. LaBar T, Adami C. 2016 Different evolutionary paths to complexity for small and large populations of digital organisms. *PLoS Comput. Biol.* **12**, e1005066. (doi:10.1371/journal.pcbi.1005066)
47. Gerrish PJ, Lenski RE. 1998 The fate of competing beneficial mutations in an asexual population. *Genetica* **102**, 127–144. (doi:10.1023/A:1017067816551)
48. Pargellis A, Greenbaum B. 2016 Digital replicators emerge from a self-organizing prebiotic world. In *Proc. Artificial Life Conf. 2016*, pp. 60–67. Cambridge, MA: MIT Press.