Exploring the likelihood of chance emergence of self-replication in a digital system. Thomas LaBar[1,2] and Christoph Adami[1,2,3], [1]Microbiology and Molecular Genetics, [2]BEACON Center for the Study of Evolution in Action [3]Department of Physics and Astronomy, Michigan State University, East Lansing MI 48824. labartho@msu.edu, adami@msu.

Introduction: Little is known about how likely it is that a self-replicating polymer of given length L monomers can occur by chance. While this likelihood must depend on the chemistry of monomers, it is possible to make some statements that depend entirely about the amount of information encoded in the self-replicator. In a recent paper, theoretical arguments suggest that the likelihood can be quantified in information-theoretic terms, so that the probability to find a self-replicator that encodes information I as a polymer derived from monomers with an alphabet size D is [1]

$$P_0 = D^{-L} \quad (1)$$

This formula (where logarithms are taken to the base D so that the units of information is the "mer" [2]) makes a number of assumptions about the rate at which monomers occur by chance, as well as on the distribution of self-replicators in sequence space. Here, we test some of these assumptions by generating polymers of length L within the chemistry of the digital life system Avida [3] that is commonly used to study evolutionary dynamics. Avidians are self-replicating computer programs written in a language with $D=26$ different instructions. The language does not use addresses, and thus the programs can be thought of as symbolic strings just as DNA and proteins. They are capable of near-universal computation, and can encode complex functions akin to the biochemical metabolism, albeit in a computational world. In a typical Avida experiment, a hand-written ancestral self-replicator of length $L=15$ is used to initiate evolution. The information content of this replicator can only be estimated, but appears to be between 5-8 mers. According to this estimate, it should be possible to find self-replicators by chance, as $26^{-5} \approx 8.4\times10^{-8}$, while $26^{-8} \approx 4.8\times10^{-12}$.

Results: We have created one billion avidian strings of length 15 randomly, by giving each monomer an equal chance to appear, and found 58 self-replicators among them, leading to an estimate (according to Eq. (1)) of 5.13 mers of information. We have investigated whether biasing the probability of production of each monomer can increase the probability of random formation of a self-replicator, by analyzing the frequency with which each instruction occurred in an ensemble created by mixing all the instructions present in the set of 58 replicators we found by chance. The entropy (per-site) of this ensemble turned out to be 0.91 (as opposed to the 1 mer for an unbiased ensemble). According to [1], using such a biased production scheme should yield an increase in the probability of finding a self-replicator, as long as the reduced search space (now searching a volume of $D^{0.91L}$ rather than D^{L}) still contains the majority of self-replicators. We found 14,400 self-replicators by chance when testing a billion, corroborating the hypothesis that a biased mutation scheme focuses the search on the part of the fitness landscape containing the replicators.

Equation (1) assumes that replicators are equally distributed in genetic space, and that the probability to find a self-replicator does not depend on the length of the sequence, only on the amount of information encoded within it. To test these assumptions we analyzed the network structure of self-replicators, and while we found no clustering, this does not imply that self-replicators are not evenly distributed (as we did not find all self-replicators). Given that a previous test for generating self-replicators randomly for length 100 avidians did not find any when searching without bias (but did find 10 in 200 million in a biased scheme [4]) we also tested the assumption that the likelihood to find self-replicators is length independent, by searching for self-replicators among $L=8$ as well as $L=30$ sequences. We present some evidence that the likelihood of random emergence of information does depend on the length of the sequence that the information is encoded in, and offer possible ways to understand this deviation from Eq. (1).

Discussion: The information-theoretic predictions of the spontaneous formation of self-replicators given a production rate of monomers and basic assumptions about the fitness landscape allows us to make predictions about the ideal environments within which spontaneous self-replication might emerge. These mathematical assumptions need to be tested, and digital systems in which self-replicators exist (but are rare) are the ideal tool to explore the underlying assumptions.

References:

[1] C. Adami (2015). Information-theoretic considerations concerning the origin of life. Arxiv.org arXiv:1409.0590.

[2] C. Adami (2004). Information theory in molecular biology. Phys. Life Rev. 1, 3-22.

[3] C. Adami (1998). Introduction to Artificial Life. Springer (New York).

[4] M. Rupp et al. (2006) The evolution of structural organization in digital organisms. In ALIFE 10 pp. 268–274, Cambridge, MA: MIT Press.