Measuring Biological Complexity in Digital Organisms

Wei Huang, Charles Ofria, and Eric Torng

Department of Computer Science and Engineering
Michigan State University, East Lansing, MI 48824
Article Correspondence: ofria@msu.edu

Abstract

We define biological complexity as the genetic information that an organism has about its environment. We have significantly improved methods to measure complexity based on Shannon Information Theory and the principle of mutation-selection balance from population genetics. The previous method of Adami et al. was a population-based measure; it examined the information content of all genomes corresponding to the same phenotype. This population-based method had inherent limitations such as requiring a full population at equilibrium to be able to approximate complexity, for genomes to be fixed-length, and for the environment to have only a single niche. Our new method overcomes these difficulties because it is genome-based rather than population-based. We approximate the total information in a genome as the sum of the information at each locus. The information content of a position is calculated by testing all of the possible mutations at that position and measuring the expected frequencies of potential genes in the mutation-selection equilibrium state. We discuss how this method reveals the way information is embedded in the organism during the evolutionary process.

Introduction

Our goal is to understand the relationship between biological complexity and evolution. For example, does a detrimental mutation always imply a decrease in complexity? Does a beneficial mutation always increase complexity? How does new complexity arise? Is it always an uphill climb? The answer to these questions depends on the nature of the complexity definition that you use. Many definitions have been proposed in the past, but almost all of them have serious flaws, and few use a rigorously mathematical approach.

We first review some previously used complexity measures and their limitations. We then discuss the concept of physical complexity developed by Adami and Cerf (1) and refined by Adami, Ofria, and Collier (2). To facilitate this discussion, we briefly review Shannon Information Theory and describe the Avida digital life platform. Finally, we discuss our new approach, its advantages, and initial results.

Complexity Measures

The most used metric for the complexity of a sequence is KCS (Kolmogorov-Chaitin-Solomonoff) complexity, defined as the size of shortest algorithm that can generate that sequence. This definition works in many intuitive cases, but has serious problems: some apparently complex structures can be coded in short programs such as fractals and cellular automata (3), while a long sequence with no pattern, and no meaning (effectively random) needs a long program to generate it; one that just lists the entire sequence. Thus, the KCS definition fails to be a useful measure of biological complexity.

A related complexity definition is logical depth (4). Bennett defines the logical depth of a sequence as the running time of the shortest program that computes it. Thus, it overcomes some of the problems with KCS complexity because more complex structures may take a while to generate, but there are still problems when it comes to random sequences with no meaning behind them.

A count of the number of "parts" in an organism is perhaps the simplest definition of complexity, as suggested by Hinegardner and Engelberg (5). This, of course, depends on what we recognize as parts. Hinegardner and Engelberg suggest that at root, organisms are composed of molecules, but they do not take the differences in the complexity of those molecules into account. This definition may provide a useful approximation of complexity, but it neglects any complexity inherent in gene regulation or other interactions.

In 2000 Adami and Cerf (1) developed physical complexity as a method to compute the complexity of symbolic strings. One of the authors (Ofria) then worked with Adami (2) to translate this concept to study the biological evolution of complexity.

Conceptually, the physical complexity of an organism is the amount of information that is stored in its genome about its environment. A genome stores information that is expressed into the functional capabilities of the organism in a given environment. Thus, the physical complexity of a genome or organism should mirror its functional capabilities (phenotype). They relate phenotype and physical complexity by building on some basic concepts from Shannon Information Theory. We review these concepts now.

Shannon Information Theory

The field of Information Theory uses quantitative mathematics to formally define measures of disorder and uncertainty, which are used, in turn, to define the information content of a message as the reduction of uncertainty attributed to that message. Originally, information theory was designed for telecommunications to maximize information transmittal over a noisy channel. In evolutionary biology, we can consider the replication of a genome from parent to child as a channel that genetic information is passed through. Mutations are the noise in this channel, and the quality of the resulting message will determine if a mutation is detrimental, neutral, or even beneficial.

Information theory defines *uncertainty* (also called entropy) as the number of bits we expect to need to fully specify a situation, given a set of probabilities. Uncertainty is maximized when all probabilities are equal—we have no idea what the outcome will be. Uncertainty is defined as:

$$H = -\sum p_i log_2 p_i \qquad (1)$$

In the context of information transmission, we are primarily concerned with how much information the output symbol of a channel tells us about the input symbol of the channel. If the output symbol is random, it thus provides no information about the input symbol. If, however, we have an error-free channel, then the output symbol provides complete information about the input symbol.

More formally, for a given channel, let X represent the input symbol, Y the output symbol, and m the possible symbols that can go through this channel. $H(X)$ is our base uncertainty of the input without knowing the output Y. We define our uncertainty about X after receiving Y as:

$$H(X|Y) = \sum_{y=1}^{m} \sum_{x=1}^{m} -p(y)p(x|y) \log p(x|y) \qquad (2)$$

Finally, we define the *information* (also called mutual information or mutual entropy) that goes through this channel with each symbol as the difference between these two uncertainties.

$$I(X:Y) = H(X) - H(X|Y) \qquad (3)$$

Living organisms are a special case when we study their replication as an information transmission process. The information contained within a genome determines how the organism behaves; in particular, it determines whether or not the organism can replicate. This becomes a self-reinforcing process since if required information is destroyed no further copies can be made, while if unimportant positions in the genome are mutated, this has no bearing on further replication. Thus, outside of an adaptive event, only changes in the non-informative portions of the genome will persist over time.

Population-Based Complexity

Adami et al. use Shannon Information Theory relate phenotype to physical complexity. If we know nothing about the organism, then we have maximal uncertainty about its genome (any genome is possible). On the other hand, if we know the organism's phenotype, we have less uncertainty about what the organism's genome is (only a small fraction of possible genomes corresponds to any specific phenotype). The difference between these uncertainties represents the information stored in the genome about its environment, and thus its physical complexity.

Unfortunately, it is difficult to define the entropy or uncertainty of a genome given its phenotype. To approximate this uncertainty, Adami proposed that most encodings of a phenotype would be similar to each other – typically only differing by neutral mutations. Given this assumption, a large population with the same phenotype should contain the distribution of genomes needed to calculate physical complexity. Unfortunately, it is difficult to get a large enough population, so Adami showed that in most cases, it is sufficient to calculate the entropy of a population of genomes site by site. If there is no epistasis (non-linear interactions between genome positions), this will give us the same result.

To illustrate this technique, suppose we wish to approximate the physical complexity of DNA-based organisms. We first need a population in an equilibrium state that have identical phenotypes. Without any information about the genome, we must assume that each of the four nucleotides has equal probability of occurring at any site i leading to a maximum site entropy of $H_{\max} = 2$. Let the frequencies for each nucleotide at site i within the actual population be $p_C(i)$, $p_G(i)$, $p_A(i)$, $p_T(i)$. The population entropy of this site is then

$$H_i = -\sum_{j}^{C,G,A,T} p_j(i) \log p_j(i) \tag{4}$$

The information content at site i would then be:

$$I(i) = H_{\max} - H_i = 2 - H_i \tag{5}$$

Finally, the physical complexity for this phenotype is approximated by applying this equation to each site and summing them together.

The Avida Platform

To apply this population-based physical complexity, we must have a population of genomes with the same phenotype. Adami, Ofria and Collier used the Avida Platform (8) to generate populations of evolved digital organisms. The Avida software maintains a population of self-replicating computer programs (similar to computer viruses) that evolve subject to natural selection in a complex environment. The phenotype of an organism corresponds to the set of actions it can perform and related information such as timing. Organisms receive energy for performing specific computations. The fitness of the organism is then its total energy intake divided by the energy required to produce an offspring. The genome of an organism is composed of a Turing-complete programming language; that is they can perform any computable mathematical function—no explicit limitations are imposed on what can be evolved. Indeed, we have witnessed a wide variety of unexpected and seemingly clever adaptations arise through evolution in Avida.

Measurement Limitations

In their previous study, Adami et al. used the Avida platform to examine the evolution of physical complexity in digital organisms. Avida was setup in single-niche, mass-action mode. The single-niche aspect means that the organisms are in direct competition against each other and the species with the highest fitness phenotype will dominate. The fact that the population is mass-action means that there is no local structure so if a higher fitness species evolves it can to take over rapidly due to an exponential growth rate. Finally, they forced all organisms to have the same length genomes (100 sites) so that sequence alignment would be unnecessary. During each experiment, they calculated the frequency of each instruction at each site by counting the number of organisms with that instruction.

This prior, population-based technique allowed for good estimates of physical complexity over time in many instances, but suffers from a number of limitations. First, the technique only produces accurate measurements if the population has reached an equilibrium state. For example, if a beneficial mutation causes a new species (and thus new phenotype) to take over a population, all otherwise neutral sites in the genome hitchhike to fixation, and it will take time before equilibrium is reached where most genotypes of the new phenotype are represented. It is often the case that a new beneficial mutation will arise before this equilibrium preventing us from determining the true complexity of that phenotype. We can see these effects in Figure 1, the upper line of which displays the physical complexity over time using the population-based technique for a typical Avida experiment. Notice that each time complexity increases, it overshoots its mark and then gradually comes down again, typically to a higher resting level than it started.

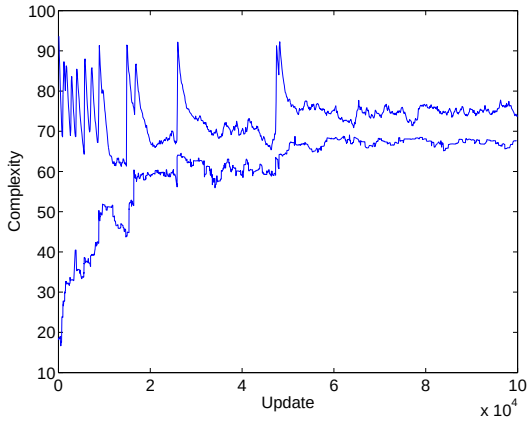

Figure 1: A comparison between the population-based method of calculating physical complexity (upper line) and our new method (lower line). The new method is applied to the lineage of the most abundant organism at the end of the experiment.

The second problem with population-based physical complexity is the constraints that must be placed due to computational concerns. The population size must be small (typically 10,000 or fewer) and the length of their genomes is fixed to prevent alignment problems. The finite population size limits the possible range of genotypes compounding the population diversity limitation noted earlier. If there are too many neutral sites, it is unlikely for them all to be represented in such a limited population even at equilibrium. The fixed length genome puts an inherent cap on complexity growth (there is only so much "blank tape" to write information into) and precludes many powerful forms of mutations such as gene duplications.

A final problem is ensuring that all organisms possess the same phenotype. This can only be achieved by using a single niche environment, and even with a single niche, a single phenotype only occurs at equilibrium. This limits the the range of interesting experiments that can be studied using population-based physical complexity, and excludes the possibility of studying ecosystem complexity.

Our Approach

Here, we demonstrate a new method for calculating physical complexity that has the ability to transcend the limitations listed above. We calculate the complexity of a single genome by examining its local mutational landscape by testing the fitness effects of all possible single-step mutants and using the principles of mutation-selection balance to approximate an infinite landscape. Since we calculate the complexity one organism at a time, we never have to worry about overshooting the correct complexity due to a biased sample, we do not impose any genome size limitations, and we can allow the environment to vary as long as we always test an organism using the state of the environment during its lifetime. In essence, we shift the complexity measure from the phenotype level to the genotype level.

This method does, however, suffer from limitations of its own. First, we only consider single-step mutants. In the future we plan to refine this method by examining multiple sites at once in an attempt to decipher epistatic interactions and improve our complexity measure. Second, a significant amount of extra processing power is required to generate all possible single point mutations from a genome and to test the fitness of each. In a computer this may be feasible, but in a natural system it is nearly impossible given our current technology. We must therefore limit ourselves to applying this physical complexity measurment technique to computational systems for the moment. This is not as severe a problem as it may seem since the main goal of quantifying complexity is to study its origin. In the natural world, evolution progresses too slowly to see significant changes to species in time spans shorter than centuries, so experimental macro-evolution already has this restriction placed on it. Furthermore, as we improve our techniques for calculating complexity in the digital world, we can use this to determine the quality of other complexity approximation algorithms that can more easily be applied to natural systems.

The Technique

As our first step in calculating the physical complexity of Avida genomes, we developed a test environment that organisms can be inserted into. The test environment is initialized to the exact same conditions as the environment that the population is evolving in, but only one organism is tested at a time. That organism is processed until either it gives birth and we can measure its fitness, or else it dies of old age (indicating a zero fitness).

As in the previous method, we calculate the complexity of the whole genome by summing the complexities of the individual sites in that genome. To determine the complexity of site i, we start by mutating this site to all other possible states and then use the test environment to calculate the fitness of each. In the case of Avida, there are 26 instructions in the genetic alphabet, so we need to generate 25 new genomes to represent each possible mutation at site i. We then run each of the resulting 26 genomes through the test environment to determine the fitness of each. With these fitnesses and a mutation rate, we can predict the abundance of each instruction at this site were a population at equilibrium.

Intuitively, it is clear that if a genome has equal fitness no matter which instruction is at site i, then we would expect all possible instructions to appear with about equal frequency. Further, this would translate to a maximal entropy for that site, and thus a zero complexity. On the other hand, if only the original instruction has a non-zero fitness, then we expect that instruction to dominate in an equilibrium population (the others would persist at a small frequency due to detrimental mutations creating them.) In this case, the population would have a low entropy at this genomic position, and

it would contribute maximally to complexity. It is slightly more complicated to calculate the expected abundance at sites with mixed fitness levels; our techniques are discussed below.

We show a sample sub-sequence from a genome in Table 1 where a single site is mutated throughout. Its original state was 'm', but all others are tested as well and their fitness recorded.

Sequence	Fitness
...akapbkawbjbo**a**cpbnaqblafpq...	0
...akapbkawbjbo**b**cpbnaqblafpq...	6.46734
...akapbkawbjbo**c**cpbnaqblafpq...	0
...akapbkawbjbo**d**cpbnaqblafpq...	5.94
...	...
...akapbkawbjbo**m**cpbnaqblafpq...	6.46734
...	...
...akapbkawbjbo**z**cpbnaqblafpq...	3.23367

Table 1: Samples for genome sequences with all single-site mutations and the resulting fitness of each. The site being changed is marked in bold.

We use the mutation-selection balance principle from population genetics to take the fitness values and determine the portion of the population that we expect each genotype to fill at equilibrium. Fisher, Haldane, and Wright, pioneers of population genetics, developed mathematical models quantifying the relative importance of selection and mutation in maintaining genetic variation. We simplify and specialize these equations to Avida, which has populations that are asexual, haploid, and have overlapping generations, and where we only consider the possibility of site i mutating, since we are not considering interactions between sites.

Let p_j denote the percentage of the population occupied by genotype j at equilibrium, ω_j the fitness of genotype j, D the alphabet size (in our case $D = 26$), and μ the per-site mutation rate. Furthermore, we assume all mutations are equally probable. The average fitness is defined by

$$\bar{\omega} = \sum_{k=1}^{D} p_k \omega_k \tag{6}$$

At equilibrium, the following equation must hold:

$$p_j = (p_j \omega_j / \bar{\omega})(1 - \mu) + \sum_{k=1}^{D} (p_k \omega_k / \bar{\omega}) \mu (1/D) \tag{7}$$

In this equation, $p_j \omega_j / \bar{\omega}$ is the relative replication rate of genotype j, and $1 - \mu$ is the probability that genotype j replicates without mutation at site i. These two factors are multiplied together to give us the rate of perfect replication within genotype j. For the second part of the equation, $\mu(1/D)$ is the probability that any genotype (including j) mutates to genotype j. We then multiply this by the relative replication rate for each genotype to determine the rate that each genotype mutates to genotype j. These two factors summed together represent the rate at which genotype j enters the population. Since all organisms leave the population with equal probability, at equilibrium, the rate at which genotype j enters the population must be the same as p_j, the percentage of the population occupied by genotype j at equilibrium.

We use equation 6 to simplify equation 7:

$$p_j = (p_j \omega_j / \bar{\omega})(1 - \mu) + \mu(1/D) \tag{8}$$

To determine the final abundance of each of the 26 genotypes at equilibrium, we generate the 26 equations and solve them. This will always provide us with a unique solution that will predict the abundance of each possible instruction at this site, were we at equilibrium in an infinite population—exactly what we need. We can then calculate the physical complexity of this site and repeat this process for each other site in the genome, summing them up to determine the physical complexity of the genome as a whole.

Experiments and Results

Our first experiments test how accurately our models predict the abudance of single-step mutants. We initiate Avida experiments with a population size of 3600 where only a single site is allowed to mutate. Given our instruction set of 26, there can only be 26 possible genotypes in the population. We then compare our predicted abundances to the observed abundances once an equilibrium is reached as shown in Figure 2. We have performed over 30 such comparisons, and all performed similarly well.

Our second set of experiments highlight the improved accuracy of our new method when a population is not at equilibrium compared with the previous population-based method for calculating complexity. In order to be able to calculate the population-based complexity, we perform Avida experiments with large (3600), single-niche populations with fixed-length genomes. According to the Natural Maxwell's Demon proposed by Adami et

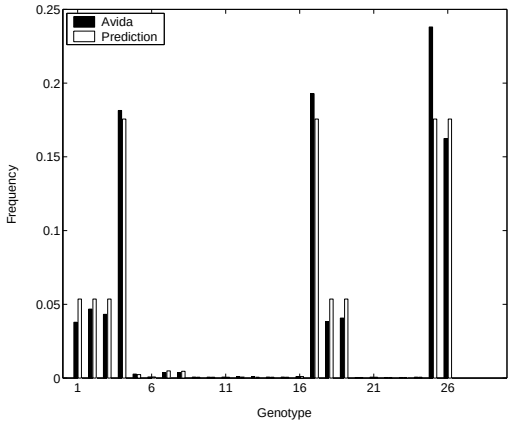

Figure 2: Mutation-selection balance at site 53 of a genome. White bars are mathematical predictions of instruction frequency, black bars are experimental results.

al. (2), complexity should increase over time in such populations.

Within each experiment, we first identify the lineage of the most abundant genotype at the end of the experiment. At each Avida update, we then compute both the population-based complexity (using all genotypes alive at that update) and our single-step mutant complexity of the current genotype on the isolated lineage. Figure 1 contains a plot of both complexity measures at each Avida update for one Avida experiment.

As we can see from Figure 1, when the population is not at equilibrium, the population-based complexity is inaccurate for many of the updates as it suffers from hitch-hiking effects. On the other hand, the proposed growth in complexity over time (with minor fluctuations) is clearly visible in the single-step mutant complexity (lower line in Figure 1). The minor fluctuations in complexity are expected; there are occasional decreases due to detrimental mutations, drift, or (occasionally) evolution of a more compressed way to code for a phenotypic trait. A detrimental mutation can be an important step on the way to significant fitness improvements(9). Some fluctuations may also be caused by inaccurate approximations of the actual complexity due to the fact that we do not yet account for epistasis. At equilibrium, both methods provide a qualitatively similar result, though the population-based complexity measure is always higher than the single-step mutant complexity.

Our third set of experiments highlight the improved accuracy of our new method for calculating complexity when a population is at equilibrium compared with the previous population-based method. We perform 30 sets of Avida experiments (identical to those performed in our second set of experiments) with three different population sizes: 900, 3600, and 14,400, and we focus on the comparison between the computed complexities once the population reaches a final equilibrium. As noted in Figure 1, the population-based complexity is higher than the single-step mutant complexity at equilibrium. The mean difference between complexity

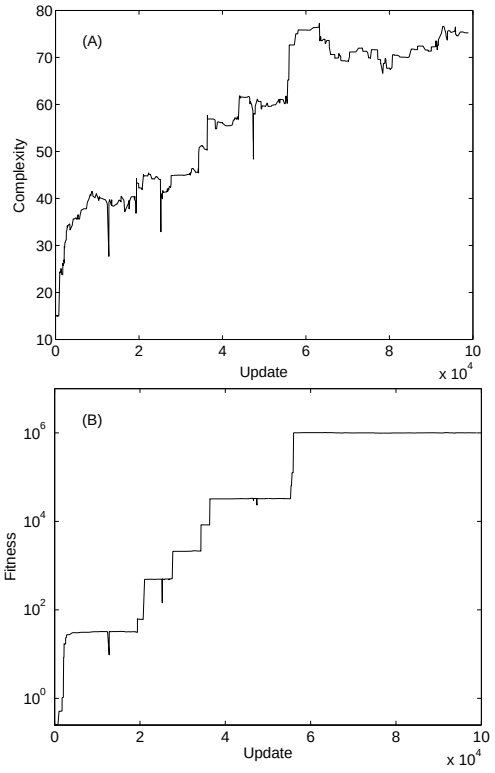

Figure 3: (A) Physical complexity over time for a lineage from an Avida experiment where genome length is allowed to change. (B) Fitness over time for the same lineage from the same experiment.

measures at equilibrium in the size 900 population experiments was 21.35 bits. The mean difference between complexity measures at equilibrium in the size 3600 population experiments was 10.04 bits. Finally, the mean difference between complexity measures at equilibrium in the size 14,400 population experiments was a nearly negligible 1.61 bits.

Clearly, as population size increases, these measures become dramatically closer. This can be easily explained as our new method approximates an infinite population, and is thus better reflected by larger population sizes.

Our final experiments demonstate the increased flexibility of our new method for calculating physical complexity. In particular, we show that it can be applied to variable length genomes. We perform a sample Avida experiment with a population size of 10,000 organisms where we allow the size of the genome to change. Figure 3A displays our new physical complexity measure of a lineage from this experiment. There are many clear jumps in complexity over time. Figure 3B shows how these complexity jumps correlate with fitness increases. This suggests that we are accurately reflecting the true complexity as at each fitness jump, more information about the environment is encoded into the genome. Downward spikes occassionally occur in both graphs due to detrimental mutations that briefly exist along the lineage.

Discussion and Future Work

Does complexity always have an increasing trend? This is an age-old question. Since Darwinian evolution is a unique and long-term procedure in nature, it is nearly impossible to get a conclusive result from the natural world. In Avida, we are able to perform many experiments and observe macro-evolutionary dynamics to study how this process works. The data we have collected thus far has concurred that complexity does seem to always increase over time, in accordance to the Natural Maxwell's Demon proposed by Adami et. al (2). These experiments, as well as the theory that led to them, assume a single-niche environment. If the environment is in any way unstable this law of increasing complexity breaks down. If, for example, a resource is no longer present in the environment, genomic information about that resource is wasted and is no longer reinforced by selection—it should decay over time or be replaced by more pertinent information.

Initial experiments in environments with multiple, limited resources show that many species can easily co-exist in an Avida population and form primitive eco-systems (10). While theory does not dictate that populations in such naturally fluctuating environments must increase in complexity, we have observed a much more rapid fitness increase in these populations. These new techniques will allow us to examine both the complexity of individuals within an ecosystem, and examine the information shared between organisms as a first step in calculating the complexity of the ecosystem as a whole. It would be impossible to show that any single species must always gain in complexity, but there is much more that may be said about the ecology as a whole.

Acknowledgments

We would like to thank Chris Adami, Richard Lenski, Thomas Schmidt and the Avida Group for helpful discussions. This work was supported by NSF grants EIA-0219229 and DEB-9981397.

References

[1] Chris Adami and Nicholas Cerf. Physical complexity of symbolic sequences. *Physica D*, 137:62–69, 2000.

[2] Chris Adami, Charles Ofria, and Travis C. Collier. Evolution of biological complexity. *Proc. Nat. Acad. Sci*, 97:4463–4468, 2000.

[3] Ben Goertzel. *The Evolving Mind.* Routledge, June 1993.

[4] Charles H. Bennett. Logical depth and physical complexity. In R Herken, editor, *The Universal Turing Machine, A Half-Century Survey*, pages 227–257. Oxford University Press, Oxford, 1988.

[5] R Hinegardner and J Engelberg. Biological complexity. *Journal of Theoretical Biology*, 104:7–20, 1983.

[6] John T. Bonner. *The Evolution of Complexity.* Princeton University Press, 1988.

[7] Daniel W. McShea. Metazoan complexity and evolution: Is there a trend? pages 477–492.

[8] C. Ofria and C. Wilke. Avida: A software platform for research in computational evolutionary biology. *Artificial Life*, 10:191–229, 2004.

[9] Richard E. Lenski, Charles Ofria, Robert T. Pennock, and Christoph Adami. The evolutionary origin of complex features. pages 139–144, May 2003.

[10] T. Cooper and C. Ofria. Evolution of stable ecosystems in populations of digital organisms. In MA Bedau RK Standish and HA Abbass, editors, *Eighth International Conference on Artificial Life*, volume 119, pages 227–232, Boston, MA, 2002. MIT Press.