Sequence Complexity in Darwinian Evolution

CHRISTOPH ADAMI[1,2]
[1]*Digital Life Laboratory 136-93, California Institute of Technology, Pasadena, California 91125;*
e-mail: christoph.adami@jpl.nasa.gov
[2]*Jet Propulsion Laboratory 126-347, California Institute of Technology, Pasadena, California 91109*

Received August 20, 2002; revised January 13, 2003; accepted January 13, 2003

Whether or not Darwinian evolution leads to an increase in complexity depends crucially on what we mean by the term. Physical complexity *is a measure based on automata theory and information theory that turns out to be a simple and intuitive measure of the amount of information that an organism stores, in its genome, about the environment in which it evolves. It can be shown that the physical complexity of the genomes of clonal organisms must increase in evolution, if they occupy a single niche and if the environment does not change. This law of increasing complexity is a consequence of natural selection only and can be violated in co-evolving systems as well as at high mutation rates, in sexual populations, and in time-dependent landscapes. Yet, co-evolution, because it can be viewed as creating an increase in physical complexity* across niches, *is likely the agent of a* global *increase in complexity.*

Key Words: evolution; complexity; entropy; information; digital life

The controversial nature of discussions about a trend in the evolution of complexity can be traced back to a lack of agreement on the definition of complexity. For some, it is obvious that complexity has increased, whereas others claim that there is not enough evidence to argue for or against an increase, and others still deny that "progress characterizes the history of life as a whole, or even represents an orienting force in evolution at all" [1]. Often, these camps disagree not only about the existence of a trend, but also on what type of complexity measure to use and whether maximum or average complexity is pertinent. When quizzed directly, however, by and large everybody agrees that nobody really knows what is meant by the word "complexity" when referring to a biological organism. Indeed, although complexity measures abound (many of them invented by physicists [2]) their relationship to biology is not always clear. In particular, complexity can be understood to refer either to form, function, or to the sequence that codes for it. When we consider animals, we usually think of *structural complexity*, but this seems to be the hardest measure to define. McShea [3] has studied several measures of structural complexity, based on number of cell types, different limb-pair types, and even the fractal dimension of sutures in ammonoids and found some evidence for a trend in these indicators, but nothing as conclusive as one might have anticipated. Also, although a trend can often be observed in the maximum, it tends to erode in the mean. *Functional complexity* can be understood as a measure of the capacity of an organism to operate successfully on input data, an *information processing capacity* therefore, which could just as well be applied to computing machinery. Another way to define functional complexity would be as the number of different functions that an organism can perform [4], which suffers from the problem of unambiguously identifying functions and separating them into nonoverlapping classes. Finally, *sequence complexity* focuses on the properties of the underlying program, which gives rise to the complex organism, only. If these three types of complexity were amenable to a mathematical characterization, we would expect to find relations between them. For example, if an organism is viewed as a complicated computing machine, it is plausible that the code that generates the computing machine would reflect the complexity of the machine itself (this is a consequence of the existence of universal Turing machines). Furthermore, we expect the complication of this machine (necessary for processing high-bandwidth data) to be reflected in its structural complexity. However, only sequence complexity has an unambiguous mathematical definition at this point, so we shall focus on this measure here.

Many of the complexity measures introduced in Ref. 2 are in fact sequence complexities. Most of them, however, do not appear satisfactory from an intuitive point of view. One of the measures most often put forward as a candidate of sequence complexity, the Kolmogorov complexity (see, e.g., [1]), turns out to be a measure of the regularity, rather than complexity, of a sequence. This implies that a random sequence is accorded maximum Kolmogorov complexity, clearly not anything we would be interested in as biologists, because random sequences do not give rise to organisms.

The concept of *physical complexity,* introduced in Ref. [5], is different because it appears to correspond exactly to what biologists think is increasing when, as Bennett [6] says, "self-organizing systems organize themselves." Physical complexity applies only to symbolic sequences that describe, and operate on, their environment. It turns out to be a particular case of the "effective complexity" concept independently developed by Gell-Mann and Lloyd [7], thus illustrating its use in evolution.

Physical complexity is carefully defined from an automata-theoretic point of view, but it has a very simple relationship to information theory and turns out to be very intuitive. Rather than starting with the mathematical definition, I will instead describe the intuitive notion and connect it with the mathematical definition later. The latter is important to clarify the circumstances under which physical complexity can be measured and to outline the assumptions and errors going into such an estimate. Finally, I show that physical complexity must increase in molecular evolution under certain circumstances [8] and illustrate the trend with experiments conducted with digital organisms. Because the circumstances under which the law holds exactly seem so restrictive as to rule out all realistic situations, I discuss how the law of increasing complexity is manifested in nature and point out the role of co-evolution. Even though the law can be broken (as we know that it must, and has been), we expect it to be responsible for the general trend that has led us from pools of replicating molecules, through prokaryotes, to eukaryotes and multicellular organisms.

PHYSICAL COMPLEXITY

Roughly speaking, the physical complexity of a sequence can be understood as the amount of information that is stored in that sequence about a particular environment. For evolving genomes, this environment is the one in which the genomes replicate and in which their hosts live (in other words, the organism's niche). The definition of physical complexity must be distinguished from *mathematical* (or algorithmic or Kolmogorov) complexity, which is only concerned with the intrinsic regularity (or, in this case, irregularity) of a sequence. The regularity of a sequence is a reflection of the unchanging laws of mathematics, but not of the physical world in which such a sequence may mean something. Information, on the other hand, which as I will show can be used as a proxy for physical complexity, is always *about something*, in particular something physical. According to this measure, a sequence may embody information about one environment (niche) while being random with respect to another. This makes the measure relative or conditional on the environment. In other words, what is complex here may not be complex there, and it is precisely this feature that brings a number of important observations (that seem incompatible with a universal monotonous increase in complexity) in line with a law of increasing physical complexity.

Information is a statistical form of correlation [9] and thus requires, mathematically and intuitively, a reference to the system that the information is about. The sequence of symbols on an information-filled tape enables predictions about the state of the system the sequence is information about. This predictive capability implies that the sequence and the system have "something in common," that they are correlated. Such an information-laden sequence will possibly *not* make predictions about any other environment (unless they are very similar). If it is not known which system or environment the sequence refers to, then the symbols in the sequence cannot be considered information. Instead, they are *potential information.* (Another word for potential information is *entropy.*) Thus, to recapitulate, sequences of symbols acquire the status of information only if we can identify the system within which the sequence is useful (i.e., about which it makes predictions). Otherwise, we must consider them random, and our measure of the sequence's randomness is its entropy.

Let us now proceed to the mathematical definition of physical complexity. Such a definition is important because it immediately suggests how complexity can be measured in real adapting populations. Technically, physical complexity is defined as the “shared Kolmogorov complexity” between the sequence under consideration, and a description of the environment in which that sequence is to be interpreted [4]. The details of this definition are not relevant to us here, in particular because this definition is not practical, because it does not allow the unambiguous determination of sequence complexity from available data. When physical complexity is averaged over an ensemble of sequences, on the other hand, it does become practical, because average mutual (or *shared*) Kolmogorov complexity is, in the limit of *perfect coding*, simply equal to the amount of information the ensemble has about the environment to which it adapts. Perfect coding, in information theory, refers to the limit in which information is coded without loss or waste into a sequence. If this limit is achieved, information is perfectly compressed. This limit is rarely (if ever) achieved in nature, and we will be considering the consequences of imperfect coding (in the form of *epistasis*) below. Because (average) physical complexity is not strictly equal to information, we will often use information simply as a proxy for the sequence’s complexity.

At this juncture, it is sufficient to think of the physical complexity of a sequence as the *amount of information that is coded in an adapting population of such sequences, about the environment to which it is adapting*. This information is given by the difference between the entropy of the population in the absence of selection, and the entropy of the population given the environment, that is, given the selective forces that the environment engenders.

MEASURING COMPLEXITY

Because entropies of populations can be measured, the average physical complexity is a practical measure. The entropy of an ensemble (i.e., a population) of sequences X, in which sequences s_i occur with probabilities p_i, is denoted by the symbol $H(X)$ and calculated as

$$H(X) = -\sum_{i=1} p_i \log p_i. \tag{1}$$

The sum in (1) goes over all the different genotypes i in ensemble X. Whether or not selection acts on sequences of the ensemble is crucial for the entropy. When selection does not act, all sequences are equally probable in ensemble X (because in the absence of selection no sequence has an advantage over another). In this case, the probabilities p_i are each equal to the inverse population size, and the entropy takes on its maximal value $H_{\max}(X) = \log N$. In an infinite population, the number of all possible genotypes is given by the size of the monomer alphabet, D, to the power of the length of the sequence, L, i.e., $N = D^L$. If we agree to take logarithms to the base of the alphabet size, then the *unconditional* entropy of a population of sequences (that is, the entropy in the absence of selection) is just equal to the sequence length: $H_{\max}(X) = L$. This result is intuitively simple: the amount of information that can potentially be stored in a sequence of length L is just equal to the sequence length.

In the presence of selection, the probabilities to find particular genotypes i in the population are highly nonuniform: most sequences do not appear (either because they have not yet been discovered by the process of mutation or because their fitness in the particular environment vanishes), whereas a few sequences are over-represented. As described above, the amount of information that a population X stores about the environment E in which it evolves is then given by the difference:

$$I(X{:}E) = H_{\max} - H(X|E) = L + \sum_{i=1} p_i \log p_i. \tag{2}$$

Here, I use the standard notation $I(A : B)$ for the entropy shared between A and B (i.e., the information that A has about B), and the symbol $H(A \mid B)$ for the *conditional* entropy of A given B. Note that although X in the above formulae represents an ensemble of sequences, E stands for one particular environment, not an ensemble of environments. (Because E is not an ensemble but a particular instance, $I(X : E)$ is strictly speaking a difference of entropies rather than information in the sense of Shannon [9], but I will use the term information anyway.) Now it has become clear why I referred to the maximal entropy of the ensemble, $H_{\max}(X) = L$, as the *unconditional* entropy of the population. This entropy does not refer to any environment, thus it is not conditional on any particular one.

The probabilities p_i that go into the calculation of the conditional entropy in (2) are in fact *conditional* probabilities, because the probability to find genotype i in environment E is not equal to the probability to find the same sequence in, say, environment E'. These probabilities can in principle be estimated by simply counting the abundance of each genotype i in the population, n_i, so that

$$p_i \approx \frac{n_i}{N},$$

where N is the population size. Unfortunately, the error committed by approximating the probabilities by the relative abundance gives rise to a sizable error in the entropy of Eq. (1), so large in fact that the estimated entropy is only meaningful for essentially infinite population sizes [10, 11]. Because we need the entropy Eq. (1) in order to estimate the physical complexity, we approximate it instead by summing up the entropy at every site along the sequence. This is done by aligning all sequences in the population and obtaining the substitution probabilities at each site. In this manner, we can obtain the *per-site* entropy

$$H(j) = -\sum_{i=G,C,A,T} p_i(j) \log p_i(j) \quad (3)$$

for site j by compiling the probabilities $p_i(j)$ to find nucleotides i at position j. The (conditional) entropy, Eq. (1), is then approximated by summing over all sites j in the sequence, i.e.,

$$H(X) \approx \sum_{j=1}^{L} H(j), \quad (4)$$

so that an approximation for the physical complexity of a population of sequences of length L is

$$C_1(H) = L - H(X), \quad (5)$$

with $H(X)$ given by Eq. (4) above.

Technically, this is only a good approximation if there are no correlations between sites in a sequence. Such correlations manifest themselves by epistatic interactions (epistasis) between mutations. It is well known that such epistasis exists (see Ref. [12] for a review), in particular in populations that are not well equilibrated. Fortunately, as described in the Appendix of Ref. [8], it is possible to correct for epistatic correlations if mutations of the gene under consideration can be obtained and their fitness evaluated. In the following, we are going to assume that epistatic effects are sufficiently weak that the corrections can be ignored. In fact, Epistasis is usually more problematic for clonal organisms (and at low mutation rates) because asexuals are at maximal linkage disequilibrium, and therefore strong epistasis in a gene that could be coded in a much shorter fashion can prevent this compression from happening (perhaps because it would take too many mutations to arrive to a state at which the gene could be compressed). Recombination can be thought of as a way to improve coding efficiency, as it breaks up linkage disequilibrium. In any case, misestimates of complexity due to epistasis can be corrected for by the formula in the Appendix of Ref. 8.

EVOLUTION OF COMPLEXITY IN DIGITAL ORGANISMS

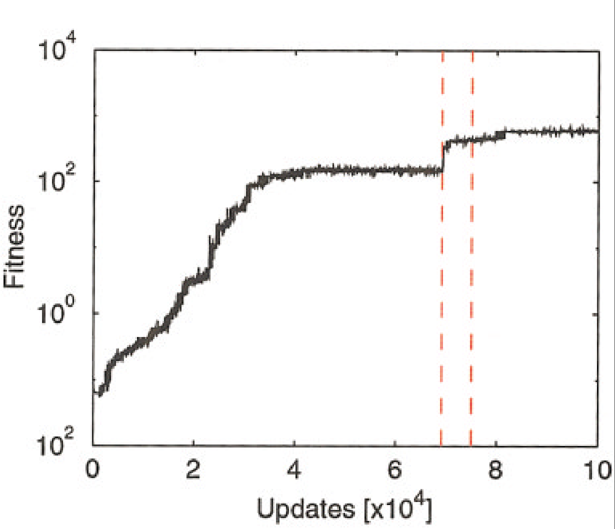
The increase in complexity that is the object of debate refers to the emergence of novelty in macroevolution. Because macroevolution takes place on geological timescales, it is difficult to witness an increase in complexity in conventional experimental populations of animals, plants, or even bacteria. This obstacle disappears if we have access to a form of life with a very short generation time. Digital organisms are just such a form of life: they are computer programs that self-replicate, mutate, and compete for resources [13–21]. Because digital organisms must copy their entire genome to survive within the computer's memory and compete for space and computer time with other programs to which they are related by descent, experiments with populations of digital organisms are to be contrasted with more conventional numerical simulations of the evolutionary process. These organisms, because they are defined by the sequence of instructions that constitute their genome, are not simulated. They are physically present in the computer's memory, and physically live there. The world to which these creatures adapt, on the other hand, is simulated, which allows the digital experimenter unparalleled precision in the planning, execution, and analysis of his experiments.

In order to survive in their world, digital organisms must replicate fast and use the available resources efficiently. The efficient use of resources concerns chiefly the utilization of the primary "energy source" for digital organisms: CPU (central processing unit) time. Without CPU time, digital organisms cannot thrive because they need to copy themselves to survive, and without the code being executed, no copying takes place.

Using random numbers that the organisms can read into their CPU with an appropriate instruction, programs can perform computations. Clearly, only very particular sequences of instructions perform meaningful computations on input numbers. In this sense, we can view such a sequence as the equivalent of a nucleotide sequence coding for an enzyme that catalyzes a reaction, involving two input chemicals, producing the energy-rich "output" chemical. In the evolutionary experiments described below, the rewarded computations are logical operations (such as AND, OR, NOR, etc.) performed on binary input strings. During adaptation, many of these *computational reactions* evolve among the digital organisms, and are used in a coordinated manner to accelerate their reproduction. In that sense, it can be said that these computational genes play the role of a *computational metabolism*, quite analogous to the enzyme-based biochemical metabolisms. The "monomers" from which these programs are constructed (the instruction set) is custom-built for their virtual CPU. For these experiments [8], the alphabet has 28 possible instructions, one of which is a logical primitive: NAND (the "not-and" operation). The experiments described below are performed by running the Avida software [15] on a standard computer. For more details on the biology of digital organisms, see Ref. [20].

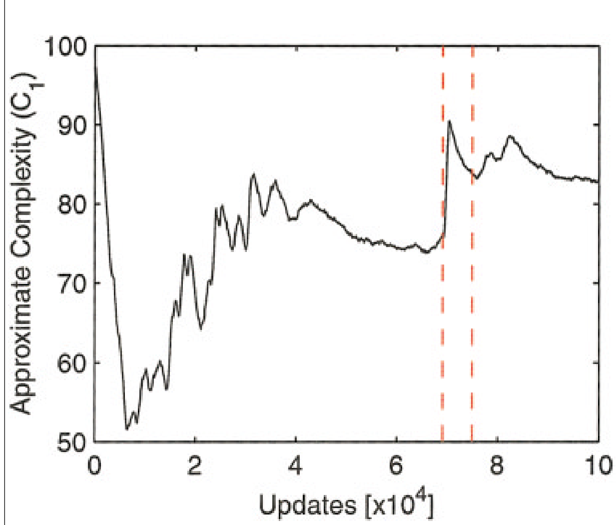
Consider the behavior of fitness over time (depicted below is the replication rate of the fastest replicator in a population of 3600 adapting programs whose sequence length is kept fixed at 100, and seeded with a single simple replicator) in Figure 1. Time is measured here in arbitrary units called *updates*, defined as the time it takes to execute

FIGURE 1



Replication rate of fastest replicator in a population of 3600 adapting digital organisms.

FIGURE 2



Approximate complexity according to Eq. (5) for a population adapting to a complex world. The dashed lines indicate the times chosen as "pre" and "post"-transition, at which the genotypes analyzed in Figure 3 were extracted.

an average of 30 instructions for each of the 3600 programs in the population. One generation corresponds to between 10 and 100 updates in such populations. Note the sudden increase in fitness around update 70,000. At this point in time, a mutation must have created a new genotype that proves to be much superior to all others. After our discussion, we expect this increase in fitness to be associated with an increase in information, so this genotype is a good candidate to look for an increase in physical complexity.

A plot of the approximate complexity [calculated according to Eq. (5)] is shown in Figure 2, where it is apparent that the complexity steadily increases, except for a period at the beginning and shortly after each transition. Both observations can easily be explained. During the initial growth of the population, most instructions appear fixed in the population (meaning, they are the same at that position across the population) because mutations have not had sufficient time to randomize *noncoding* instructions (instructions that do not contribute to the survival of the sequence). Also, evolution may struggle with a (hand-written) genome that, although extremely ill-suited to the environment, is also difficult to "re-code." Such code may simply be badly compressed, and it can take evolution a while to find a better way to represent the same information. (This hypothesis could be tested by measuring the amount of epistasis in the genome during this initial period and compare it to the epistasis after the population has relaxed, because bad compression must be associated with interactions between mutations.)

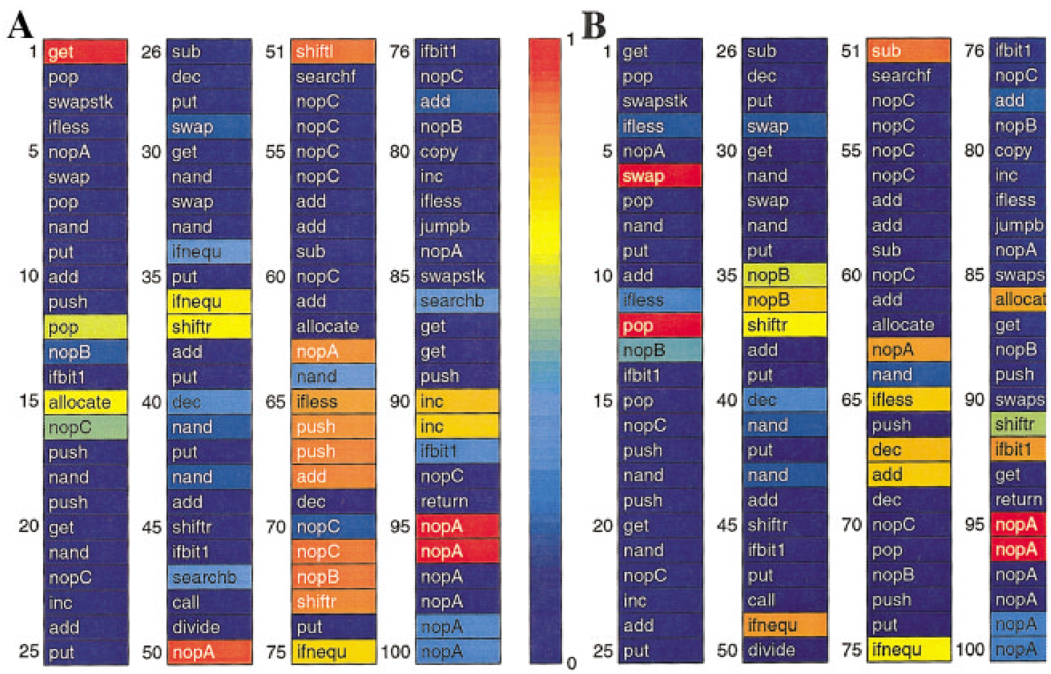
After each transition, the estimated complexity overshoots its equilibrium value due to the *hitchhiking* effect: neutral instructions hitchhiking on beneficial ones appear fixed, until mutations can randomize them again. This is particularly clear in the transition around 70,000 updates in Figure 2, to which we now turn our attention.

Because of the hitchhiking effect just mentioned, the amount of information gained in the transition highlighted in Figure 2 is not measured very accurately, simply because equilibration (required for an accurate estimate) takes longer than the time until the next transition. To get a more accurate estimate of the per-site entropy Eq. (4), we can extract dominant genotypes just before and after the transition. In order to determine whether an instruction is entropy or information, we create all possible one-point mutants of the pair of organisms and obtain their fitness in isolation. In a sense, this is equivalent to building virtual, fully equilibrated populations. If a mutation does not change the fitness or increases it, it is deemed viable, whereas all deleterious mutations are classified together with the lethal ones, because they have a low probability of appearing in subsequent generations. After this has been done for each locus, the per-site entropy at locus x_i can be estimated as

$$H(x_i) \approx \log_D(N_{\text{viable}}), \tag{6}$$

where N_{viable} is the number of neutral or beneficial substitutions at that locus. In Eq. (6), the logarithm is taken to the base of the alphabet size, thus ensuring that our measure for the randomness (entropy) at each location is normalized to

FIGURE 3



Each instruction in the two genomes in (A) and (B) is colored according to its per-site entropy (scale in the middle). An instruction that is fixed in the population has entropy close to zero (blue), implying that a mutation of that locus produces a nonfunctional organism. On the contrary, loci that can be mutated with impunity have entropy one (red). The genome (A) was extracted from the population after 2991 generations (the left of the two red-dashed lines in Figure 2), whereas genome (B) was extracted just after the transition at 3194 generations (right dashed line in Figure 2).

lie between zero and one. If we do this for the organisms before and after the transition, we obtain the per-site entropies of Figure 3. It is interesting to observe the changes in substitution pattern between these two genomes.

The most radical change seems to have taken place in the region between instructions 66 and 73, where about seven instructions that were moderately variable (in the virtual population) seemed to have turned "cold," i.e., they have turned vulnerable to mutations. This is precisely the mechanism by which natural selection creates complexity: entropy is transformed into information. There are other places in the genome where "hot" instructions turned "cold" and vice versa. The net gain in information is about six instructions, which is close to the number that we arrive at if we take into account corrections for epistasis [8].

RISE AND FALL OF COMPLEXITY

The most obvious origin of a complexity catastrophe is a drastically changing environment. As discussed above, physical complexity is a quantity defined with reference to an environment. If the changes in the environment are fast and extreme, organisms inhabiting this environment will most likely be maladapted to it, and their measured physical complexity will have decreased commensurately. In small populations, high mutation rates can also lead to a decrease in complexity, because Muller's ratchet can lead to information loss through the accidental loss of the fittest sequence. In the extreme case (critically high mutation rates) natural selection can become inactive, a phenomenon known as the *error catastrophe* in the macromolecular evolution literature [22].

As is well known, sexual recombination can also lead to an accumulation of deleterious mutations that compromise the acquired information. Although asexual populations can purge deleterious mutations with certainty (as long as the mutation rate is not too high and the population size too small, as described above), populations of sexual organisms are at risk of gene loss at any mutation rate if deleterious mutations interact antagonistically [23]. Finally, co-evolution between species occupying different niches is a special case of a changing environment (for each of the interacting species) and thus opens up the possibility of escaping the inexorable growth of complexity dictated by natural selection. Paradoxically however, there are good reasons to assume that, for the most part, co-evolution will aid, rather than hinder, the evolution of complexity, because co-evolution is a slow rather than drastic environmental change, creating new niches that provide new opportunities for adaptation.

As we saw, because niches can change and because many niches of differing *potential information* coexist at the same time, we cannot expect that a trend in one niche will persist forever, nor that the same trend will be observable in all currently existing niches. In one niche, for example, its inhabitants may have incorporated all of its potential information into their genome (such as some prokaryotes), whereas another niche may just have been invaded so that its inhabitants show rapid gene turnover. The coexistence of niches with different entropy (different potential complexity) explains the coexistence of organisms with differing complexity.

Should we not expect an overall trend if evolution produces more and more diverse niches with more and more potential information? This question addresses the issue of co-evolution and whether this process indeed produces niches with more and more entropy (which could then host, in turn, organisms with more and more complexity). This question is complicated by the fact that co-evolution necessarily produces changes in an organism's niche, which can potentially reduce an organism's complexity. In general, a change in niche will almost always first produce a decrease in physical complexity, because only in the most rare circumstances will the change be "just so" that it converts an entropic sequence into an informational one. However, if the change in niche makes it richer (i.e., produces features that are awaiting discovery), then following the initial decline in complexity the organism can enter a period of adaptation that can take it into realms of complexity hitherto unattainable. Note that the invasion of a simple environment by adaptive radiation (a species that has shed genes not necessary for survival in the simple world) would not lead to a decrease in total complexity as long as the ancestral species still exists. Therefore, even if new niches are created via co-evolution that are in equal amounts more simple and more complex than the currently existing ones, because the invasion of the more complicated niches increases total complexity, whereas the invasion of the simples ones do not decrease it, we can be confident that the process of co-evolution and its capacity to create more complicated environments may be the possible unifying process that could give rise to an overall trend.

Unfortunately, the mathematics of information in co-evolving environments has not yet been developed, so it is premature to make a prediction about whether this is the case or not. It seems plausible to me, but it is clear that counterexamples can be manufactured where co-evolution gives rise to catastrophic extinctions (via the annihilation of the existing niches), which reduce the environment's entropy and necessarily the physical complexity of its inhabitants at the same time. In such a formalism, the total complexity of an ecosystem would have to be defined as the mutual entropy of all organisms, about each other and the world they live in. This is an expression that is not difficult to write down, but it is a quantity that is certainly difficult to measure.

CONCLUSIONS

In order to be able to speak about complexity, we must define it. I have presented a mathematical definition of sequence complexity that has a very intuitive interpretation for biological genomes, as the amount of information a population stores about the environment in which it lives. With this definition, we can address the issue of a trend in the evolution of complexity. By showing that natural selection in a niche creates information about that niche, it is possible to show that physical complexity within that niche must increase if the environment does not change.

Although natural selection can fail to maintain the acquired information, it is highly likely that the mechanism of interacting niches in an ecosystem will ultimately lead not only to a trend within each niche, but also to a trend in the overall (total) complexity of an ecosystem. Physical complexity increases if selection acts properly, and decreases if it fails. Still, this measure of complexity does not translate to adaptation. An organism well-adapted to a simple niche can have a lower physical complexity than an organism badly adapted to a complicated niche. Thus, adaptation reflects only the degree to which the potential complexity of the niche is reflected in the physical complexity of the organism, and certainly does not allow complexity comparisons across niches.

ACKNOWLEDGMENTS

I am grateful to Murray Gell-Mann for explaining to me the relationship between physical complexity and his effective complexity, and to Charles Ofria and Travis Collier for collaboration in the experimental work reported here. I also thank Richard Lenski for guidance in biological evolution, and for valuable discussions. This work was supported by the National Science Foundation Biocomplexity program under contract No. DEB-9981397. Part of this work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

REFERENCES

1. Gould, S.J. Full House. Harmony Books: New York, 1996, p 3.
2. Badii, R.; Politi, A. Complexity—Hierarchical Structures and Scaling in Physics. Cambridge University Press: Cambridge, 1997.
3. McShea, D.W. Metazoan complexity and evolution:Is there a trend? Evolution 1996, 50, 477–492.
4. McShea, D.W. Functional complexity in organisms: Parts as proxies. Biol Philos 2000, 15, 641–668.

5. Adami, C.; Cerf, N.J. Physical complexity of symbolic sequences. Physica D 2000, 137, 62–69.
6. Bennett, C.H. Universal computation and physical dynamics. Physica D 1995, 86, 268–273.
7. Gell-Mann, M.; Lloyd, S. Information measures, effective complexity, and total information. Complexity 1996, 2, 44–52.
8. Adami, C.; Ofria, C.; Collier, T.C. Evolution of biological complexity. Proc. Natl. Acad. Sci. USA 2000, 97, 4463–4468.
9. Shannon, C.E.; Weaver, W. The Mathematical Theory of Communication. University of Illinois Press: Urbana, 1949.
10. Basharin, G.P. On a statistical estimate for the entropy of a sequence of independent random variables. Theor Probab Appl 1959, 4, 333–336.
11. Schneider, T.D.; Stormo, G.D.; Gold, L.; Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. J Mol Biol 1986, 188, 415–431.
12. Wolf, J.; Brodie, E.; Wade, M. Epistasis and the Evolutionary Process. Oxford University Press: Oxford, 2000.
13. Ray, T.S. An approach to the synthesis of life. In: Proc. Artificial Life II; Langton, C.G.; Taylor, C.; Farmer, J.D.; Rasmussen, S., Eds.; Addison Wesley: Redwood City, 1991, pp 371–408.
14. Adami, C. Learning and complexity in genetic auto-adaptive systems. Physica D 1995, 80, 154–170.
15. Adami, C. Introduction to Artificial Life. Springer Verlag: New York, 1998.
16. Lenski, R.E.; Ofria, C.; Collier, T.C.; Adami, C. Genome complexity, robustness, and genetic interactions in digital organisms. Nature 1999, 400, 661–663.
17. Wagenaar, D.; Adami, C. Influence of chance, history, and adaptation on evolution in *Digitalia*. In: Proc. Artificial Life VII; Bedau, M.A.; McCaskill, J.S.; Packard, N.H.; Rasmussen, S., Eds.; MIT Press: Cambridge, 2000; pp 216–220.
18. Yedid, G.; Bell, G. Microevolution in an electronic microcosm. Am Nat 2001, 157, 465–487.
19. Wilke, C.O.; Wang, J.L.; Ofria, C.; Lenski, R.E.; Adami C. Evolution of digital organisms at high mutation rate leads to survival of the flattest. Nature 2001, 412, 331–333.
20. Wilke, C.O.; Adami, C. The biology of digital organisms. Trends Ecol Evol 2002, 17, 528–532.
21. Yedid, G; Bell, G. Macroevolution simulated with autonomously replicating computer programs, Nature 2002, 420, 810–812.
22. Eigen, M. Natural selection: A phase transition? Biophys Chem 2000, 85, 101–123.
23. Kondrashov, A.S. Deleterious mutations and the evolution of sexual reproduction. Nature 1988, 336, 435–440.